IEEE Signal Processing MAGAZINE

VOLUME 31 NUMBER 5 SEPTEMBER 2014

BIG DATA

THEORETICAL AND ALGORITHMIC FOUNDATIONS

OPTIMIZATION AND ESTIMATION OF COMPLEX-VALUED SIGNALS

MULTIUSER COLLABORATIVE VIEWPORT VIA TEMPORAL PSYCHOVISUAL MODULATION

NEW WIRELESS TECHNOLOGIES

IEEE Signal Processing Society

IEEE

[CONTENTS]

[VOLUME 31 NUMBER 5]

[SPECIAL SECTION—BIG DATA]

[FEATURE]

[COLUMNS]

[DEPARTMENT]

Digital Object Identifier 10.1109/MSP.2014.2332376

IEEE SIGNAL PROCESSING magazine

COVER

©ISTOCKPHOTO.COM/PAULFLEET

IEEE PERIODICALS MAGAZINES DEPARTMENT

Jessica Barragué
Managing Editor

Geraldine Krolin-Taylor
Senior Managing Editor

Susan Schneiderman
Business Development Manager
+1 732 562 3946 Fax: +1 732 981 1855

Felicia Spagnoli
Advertising Production Manager

Janet Dudar
Senior Art Director

Gail A. Schnitzer
Assistant Art Director

Theresa L. Smith
Production Coordinator

Dawn M. Melley
Editorial Director

Peter M. Tuohy
Production Director

Fran Zappulla
Staff Director, Publishing Operations

IEEE prohibits discrimination, harassment, and bullying. For more information, visit http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. Individual copies: IEEE Members US$20.00 (first copy only), nonmembers US$201.00 per copy. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright Law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA; 2) pre-1978 articles without fee. Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. For all other copying, reprint, or republication permission, write to IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854 USA. Copyright ©2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals postage paid at New York, NY, and at additional mailing offices. Postmaster: Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188

Printed in the U.S.A.

SCOPE: *IEEE Signal Processing Magazine* publishes tutorial-style articles on signal processing research and applications, as well as columns and forums on issues of interest. Its coverage ranges from fundamental principles to practical implementation, reflecting the multidimensional facets of interests and concerns of the community. Its mission is to bring up-to-date, emerging and active technical developments, issues, and events to the research, educational, and professional communities. It is also the main Society communication platform addressing important issues concerning all members.

Digital Object Identifier 10.1109/MSP.2014.2331751

General Chairs
Vaughan Clarkson
University of Queensland
Jonathan Manton
University of Melbourne

Technical Program Chairs
Doug Cochran
Arizona State University
Doug Gray
University of Adelaide

Finance Chair
Lang White
University of Adelaide

Special Session Chairs
Robert Calderbank
Duke University
Stephen Howard
DSTO
Songsri Sirianunpiboon
DSTO

Tutorials Chair
Daniel Palomar
Hong Kong University of S&T

Local Arrangements Chair
Andrew Bradley
University of Queensland

Registration Chair
Paul Teal
Victoria University of Wellington

Publicity Chair
Matt McKay
Hong Kong University of S&T

Publication Chair
Leif Hanlen
NICTA

Exhibits Chair
Iain Collings
CSIRO

Student Paper Contest Chair
Nikos Sidiropoulos
University of Minnesota

Conference Managers

Registration & Program Enquiries:
Conference Management Services, Inc
3833 S Texas Ave, Ste 221,
Bryan TX 77802, USA

General Enquiries:
arinex pty limited
S3, The Precinct, 12 Browning St
Brisbane QLD 4101, Australia
Ph: +61 2 9265 0700
Email: icassp2015@arinex.com.au

ICASSP
2015
Brisbane Australia

Second Call for Papers

ICASSP 2015

2015 IEEE International Conference on Acoustics,
Speech, and Signal Processing (ICASSP)
Brisbane Convention & Exhibition Centre
April 19 – 24, 2015 • Brisbane, Australia
www.ICASSP2015.org

The 40th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) will be held in the Brisbane Convention & Exhibition Centre, Brisbane, Australia, between April 19th and 24th, 2015. ICASSP is the world's largest and most comprehensive technical conference focused on signal processing and its applications. The conference will feature world-class speakers, tutorials, exhibits, and over 120 lecture and poster sessions. Topics include but are not limited to:

Audio and acoustic signal processing
Bio- imaging and biomedical signal processing
Signal processing education
Speech processing
Industry technology tracks
Information forensics and security
Machine learning for signal processing
Localisation and tracking
Multimedia signal processing
Sensor array & multichannel signal processing
Design /implementation of signal processing systems
Signal processing for communications & networking
Image, video & multidimensional signal processing
Signal processing theory & methods
Spoken language processing
Remote sensing signal processing

Submission of Papers: Prospective authors are invited to submit full-length papers, with up to four pages for technical content including figures and possible references, and with one additional optional 5th page containing only references. A selection of best papers will be made by the ICASSP 2015 committee upon recommendations from the Technical Committees.

Signal Processing Letters: Authors of IEEE Signal Processing Letters (SPL) papers will be given the opportunity to present their work at ICASSP 2015, subject to space availability and approval by the ICASSP Technical Program Chairs. SPL papers published on or after January 1, 2014 and SPL manuscripts accepted on or before November 15, 2014 are eligible for presentation at ICASSP 2015. Because they are already peer-reviewed and published, SPL papers presented at ICASSP 2015 will neither be reviewed nor included in the ICASSP proceedings. Requests for presentation of SPL papers should be made through the ICASSP 2015 website on or before 16 December 2014. Approved requests for presentation must have one author/presenter register for the conference according to the ICASSP 2015 registration instructions.

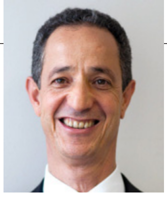
Important Deadlines:

Submission of regular papers	Sunday, October 5th 2014
Early registration opens	Monday, January 12th 2015
Notification of paper acceptance	Wednesday, January 14th 2015
Revised paper upload	Friday, February 13th 2015
Author registration	Friday, February 13th 2015

IEEE

IEEE Signal Processing Society ®

[from the **EDITOR**]



Abdelhak Zoubir
Editor-in-Chief
zoubir@spg.tu-darmstadt.de
http://signalprocessingsociety.org/publications/periodicals/spm

Where Are Today's Signal Processing Heroes?

At the time of this writing, I had just received the electronic version of the July 2014 issue of *IEEE Spectrum*. On the cover page, the headline reads, "Where Are the Heroes? Engineers Created Our Modern World. And Yet Nobody Knows Who They Are." By clicking on that title, I was taken to page 36, where I found the article titled "Engineering Needs More Heroes: There's No Lack of Worthy Characters, So Why Doesn't the Profession Celebrate Them?" [1]. The author of the article is G. Pascal Zachary, a former reporter for *The Wall Street Journal* and a professor at Arizona State University's Walter Cronkite School of Journalism and Mass Communication.

By just reading the title, I immediately thought of my editorial in the July 2014 issue of *IEEE Signal Processing Magazine* (*SPM*) [2], which discusses, in relation to the suggested Society's name change, the means as to how to raise awareness about our profession among the lay people. This is a topic that is gaining much attention within the signal processing community, and it is not surprising that several editorials and articles have addressed this very important topic over the last decade.

I was quite pleased to see the article in *IEEE Spectrum* [1]. Not only is it nice reading, but it in fact also suggests what we urgently need to address in signal processing to achieve our goal mentioned above. As the author of [1] states, "Celebrating heroes is a good way to inspire young people and inform the public, of course…. The hero deficit is in fact bad for engineering because it diminishes the enterprise in the eyes of the public, and it constricts the flow of talent into the field."

The article [1] also appears in [3], where one can find readers' comments. In my view, the question is not whether one agrees with the author of [1] on who is a hero and who is not. Ultimately, each of us decides for himself or herself who the heroes are, and we have our own reasons to believe it. My personal signal processing hero is not a contemporary one; he is Al-Khwarizmi (c. 780–c. 850), earlier transliterated as Algorismi. He introduced the beginnings of algebra, which was revolutionary, moving away from the Greek concept of mathematics that was essentially based on geometry. His contribution is a cornerstone of both science and engineering disciplines and is the foundation of many of contemporary theorems and algorithms, underlying information theory and communications. The term algorithm is derived from a Latin corruption of the name Al-Khwarizmi, which was the technique of performing arithmetic with Hindu–Arabic numerals developed by Al-Khwarizmi [4]. In essence, it is difficult to imagine a special issue of *SPM* without the use of algorithms. (A portrait of Al-Khwarizmi can be downloaded from Don Knuth's home page [5].) The reason why I see him as a signal processing hero is because algorithms development is my profession and passion.

In [1], *IEEE Spectrum* invites its readers to identify today's unsung heroes—exemplars of engineering excellence who, for whatever reason, have not received the recognition they deserve. The invitation also includes the criteria of what makes a hero [1]. The online form can be found in [6]. I encourage you to open the online form and suggest names of signal processing heroes. As mentioned above, it is what you believe that matters, and surely you have good reasons for it.

This special issue of *SPM* is on a highly relevant and timely topic, i.e., signal processing for big data. I would like to thank the guest editors and authors for their contributions. Special thanks go to the lead guest editor, Georgios Giannakis, who, under a very tight deadline, ensured that all articles were secured for a timely production.

I also wish to thank the members of the senior editorial board who support Special Issues Area Editor Fulvio Gini and me in identifying timely and important topics and assessing white proposals for special issues. Their support is instrumental for ensuring that high-quality articles are published in *SPM*. It gives me great pleasure to introduce the new class of 2016 Editorial Board members: Mounir Ghogho, Lina Karam, Stephen McLaughlin, and Erchin Serpedin. With these energetic and dedicated professionals, we shall move *SPM* to an even higher level with more innovations to come.

REFERENCES

[1] *IEEE Spectr.* (2014, July). [Online]. Available: http://online.qmags.com/IEEESM12818947?sessionID=82534985BEF39CF494026C73A&cid=780937&eid=18947#pg1&mode2

[2] A. M. Zoubir, "Signal processing: Is it time to change the name?" *IEEE Signal Processing Mag.*, vol. 31, no. 4, p. 4, July 2014.

[3] G. Pascal Zachary, "Where are today's engineering heroes?" *IEEE Spectr.* [Online]. Available: http://spectrum.ieee.org/geek-life/profiles/where-are-todays-engineering-heroes

[4] C. B. Boyer, "The Arabic hegemony," in *A History of Mathematics,* 2nd ed. New York: Wiley, 1991, ISBN: 0-471-54397-7.

[5] D. Knuth. Home page. [Online]. Available: http://www-cs-faculty.stanford.edu/~uno/graphics.html

[6] *IEEE Spectr.* (2014, July). [Online]. Available: http://spectrum.ieee.org/static/help-us-find-todays-unsung-engineering-heroes

Abdelhak Zoubir

[SP]

Digital Object Identifier 10.1109/MSP.2014.2332377
Date of publication: 19 August 2014

Alex Acero
2014–2015 SPS President
a.acero@ieee.org

Where Does Your Conference Registration Fee Go?

Conferences are a great way to connect with others who share your interests [1]. Many IEEE Signal Processing Society members may wonder how the registration fees to attend one of the Society's conferences are used. Well, they are used for direct expenses, indirect expenses, and membership services. Overall, we strive for a balanced budget since the IEEE is a nonprofit organization with the goal of serving our members.

The direct expenses include the convention center, management company, food and beverages, and Wi-Fi. Convention centers have rental fees for their facilities and require us to use their personnel to support audio/video services. Large conferences such as the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) or IEEE International Conference on Image Processing (ICIP) are so much work that the organizing committee often hires the services of a professional conference organizer company, which takes care of processing the registrations, credit card expenses, manning the registration desk, developing the conference's Web site, handling the paper submissions, assembling the proceedings, producing the USB drives, paying invoices, providing letters to attendees requiring visas, and much more. Conferences often provide Wi-Fi access; a copy of the proceedings, and refreshments during the welcome reception, breakfasts, and coffee breaks.

We ask conference organizers that they budget 80% of the conference registration revenue to cover these direct expenses. The remaining 20% is used to cover Society personnel salaries, travel costs for senior volunteers, student grants, and other membership services including Distinguished Lecturers, educational materials, and awards.

The Society has 2.5 full-time staff members in Piscataway, New Jersey who oversee budgets, negotiate contracts with hotels and convention centers, process travel grants, and support the conference organizers and volunteer committees.

We are fortunate to have many of our members volunteer their time to the Society. They serve on many of our boards, including the Society's Board of Governors, Conference Board, Publications Board, Membership Board, Technical Directions Board, and Awards Board. In addition to numerous conference calls and countless e-mails, these boards meet face to face during ICASSP, and some also meet during the fall meeting (at IEEE Global Conference on Signal and Information Processing (GlobalSIP) or ICIP) in North America. This is a significant time commitment that is highly appreciated, so the Society covers their hotel cost and (coach) airfare.

We also use some of the conference registration fees to provide services to our student members. We provide travel grants for students with accepted papers at ICASSP, ICIP, and the IEEE GlobalSIP. We also provide grants to students attending our summer schools. The Society allocates travel and prize money to the student finalists in our Signal Processing Cup.

The Society prepares a balanced budget for an estimated number of conference attendees based on historical trends. Convention center costs are mostly independent of the number of attendees. If the conference has fewer attendees than planned, it incurs a net loss to the Society. The majority of attendees typically register by the author registration deadline about two months prior to the conference, others by the early registration deadline a month before, and usually less than 10% after that or onsite. If a week or so prior to the conference we believe that we will have more attendees than forecast, we encourage the organizing committee to upgrade the food and drink options (possibly offering complimentary beverages during the welcome reception).

Many of the Society's conferences thus result in a balanced budget. Some workshops lose money. A few large conferences produce a small surplus. Overall, any surplus from conferences goes to the Society's reserves as a "rainy day" fund in case we experience poor economic times, like what we faced in 2007–2009, or unexpected events (ICASSP 2003 in Hong Kong was canceled due to the SARS epidemic), and to fund membership initiatives.

Conferences provide an excellent opportunity for attendees to present their work, learn about the latest trends, and network with each other. The Society is always looking to provide a balance between a good attendee experience and good value. I welcome any suggestions you may have on how to improve the conference experience.

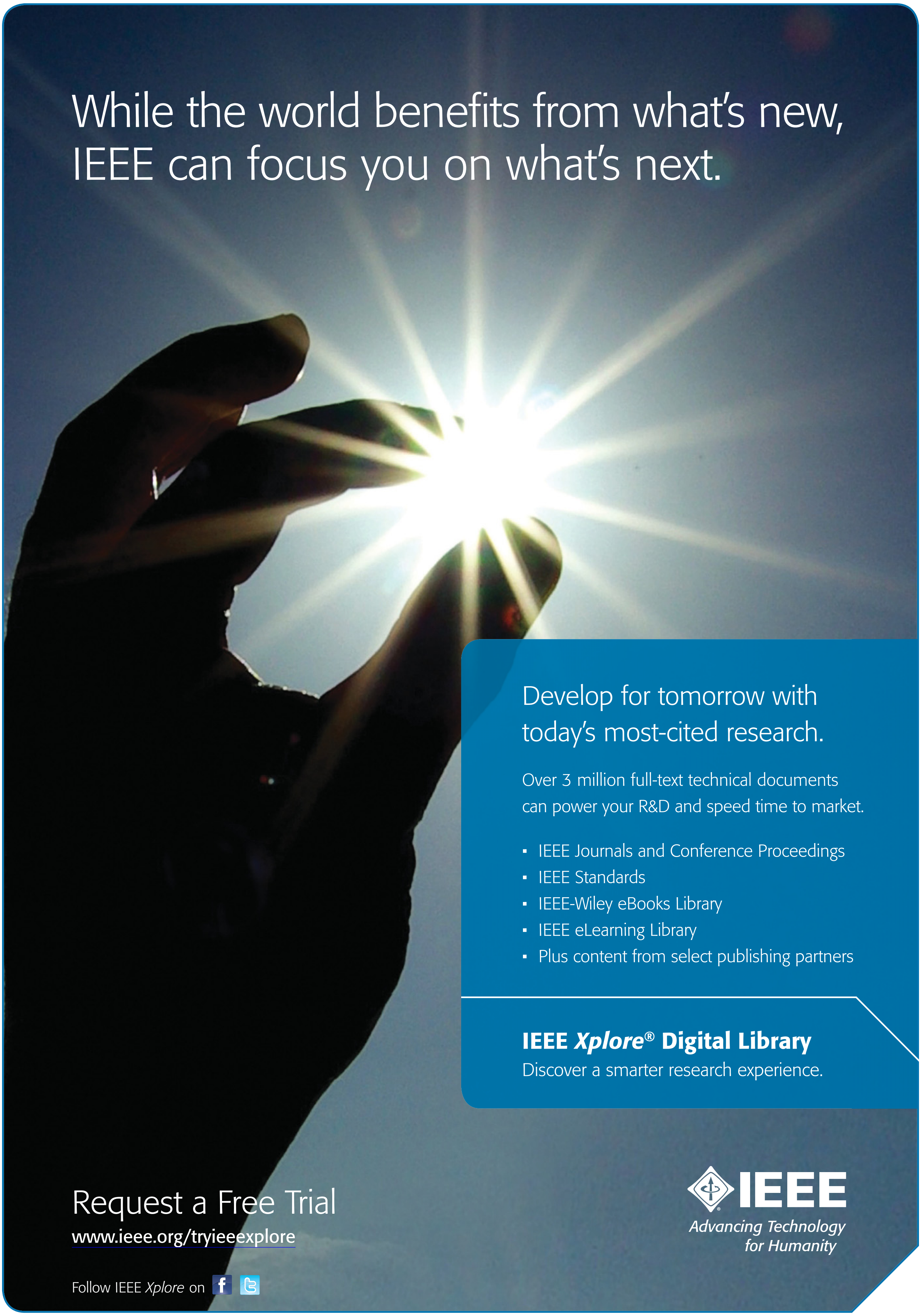
REFERENCE

[1] IEEE.Tv. [Online]. Available: https://ieeetv.ieee.org/player/html/viewer?dl=#why-conferences-matter-the-global-technical-community

AAcero

[SP]

Digital Object Identifier 10.1109/MSP.2014.2336711
Date of publication: 19 August 2014



Top Downloads in IEEE *Xplore*

The "Reader's Choice" column in *IEEE Signal Processing Magazine* contains a list of articles published by the IEEE Signal Processing Society (SPS) that ranked among the top 100 most downloaded IEEE *Xplore* articles. This issue's column is based on download data through March 2014. The table below contains the citation information for each article and the rank obtained in IEEE *Xplore*. The highest rank obtained by an article in this time frame is indicated in bold. Your suggestions and comments are welcome and should be sent to Associate Editor Michael Gormish (gormish@ieee.org).

TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	ABSTRACT	RANK IN IEEE TOP 100						*N* TIMES IN TOP 100 (SINCE JAN 2011)
		MAR 2014	FEB 2014	JAN 2014	DEC 2013	NOV 2013	OCT 2013	
A TUTORIAL ON PARTICLE FILTERS FOR ONLINE NONLINEAR/NON-GAUSSIAN-BAYESIAN TRACKING Arulampalam, M.S.; Maskell,S.; Gordon, N.; Clapp, T. *IEEE Transactions on Signal Processing* vol. 50, no. 2, 2002, pp. 174–188	This paper reviews optimal and suboptimal Bayesian algorithms for nonlinear/non-Gaussian tracking problems, with a focus on particle filters. Variants of the particle filter are introduced within a framework of the sequential importance sampling algorithm and compared with the standard EKF.	9	10	31	8	**6**	25	36
AN INTRODUCTION TO COMPRESSIVE SAMPLING Candes, E.J.; Wakin, M.B. *IEEE Signal Processing Magazine* vol. 25, no. 2, Mar. 2008, pp. 21–30	This article surveys the theory of compressive sampling, also known as compressed sensing or CS, a novel sensing/sampling paradigm that goes against the common wisdom in data acquisition.	21	19	14	**10**	11	**10**	38
IMAGE QUALITY ASSESSMENT: FROM ERROR VISIBILITY TO STRUCTURAL SIMILARITY Zhou W.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. *IEEE Transactions on Image Processing* vol. 13, no. 4, 2004, pp. 600–612	This paper introduces a framework for quality assessment based on the degradation of structural information. Within this framework a structure similarity index is developed and evaluated. MATLAB code available.	31	42	**24**	28	24	33	18
VECTOR-VALUED IMAGE PROCESSING BY PARALLEL LEVEL SETS Ehrhardt, M.J.; Arridge, S.R. *IEEE Transactions on Image Processing* vol. 23, no. 1, pp 8–9	This paper considers the components of an image as a vector. By minimizing large angles parallel level sets are obtained and used for demosaicking.	50	58	**22**	98			4
IMAGE SUPER-RESOLUTION VIA SPARSE REPRESENTATION Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. *IEEE Transactions on Image Processing* vol. 19, no. 11, 2010, pp. 2861–2873	This paper presents an approach to single-image super-resolution, based upon sparse signal representation of low and high-resolution patches.	55	92	**27**	31	44	51	10

Digital Object Identifier 10.1109/MSP.2014.2329991
Date of publication: 19 August 2014

		RANK IN IEEE TOP 100						
TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	**ABSTRACT**	**MAR 2014**	**FEB 2014**	**JAN 2014**	**DEC 2013**	**NOV 2013**	**OCT 2013**	***N* TIMES IN TOP 100 (SINCE JAN 2011)**
SCALING UP MIMO: OPPORTUNITIES AND CHALLENGES WITH VERY LARGE ARRAYS Rusek, F.; Persson, D.; Lau, B.K.; Larsson, E.G.; Marzetta, T.L.; Edfors, O.; Tufvesson, F. *IEEE Signal Processing Magazine* vol. 30, no. 1, 2013, pp. 40–60	The more antennas the transmitter/receiver is equipped with, and the more degrees of freedom that the propagation channel can provide, the better the performance in terms of data rate or link reliability. This article quantifies the reliability and achievable rates.	78		82	**43**	75		12
K-SVD: AN ALGORITHM FOR DESIGNING OVERCOMPLETE DICTIONARIES FOR SPARSE REPRESENTATION Aharon, M.; Elad, M.; Bruckstein, A. *IEEE Transactions on Signal Processing* vol. 54, no. 11, 2006, pp. 4311–4322	K-SVD is an iterative method that alternates between sparse coding of the training examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data and can be used with any pursuit method.	**87**						1
IMAGE QUALITY ASSESSMENT FOR FAKE BIOMETRIC DETECTION: APPLICATION TO IRIS, FINGERPRINT, AND FACE RECOGNITION Galbally, J.; Marcel, S; Fierrez, J. *IEEE Transactions on Image Processing* vol. 23, no. 2, 2014, pp. 710–724	This paper uses 25 general image-quality features extracted from the authentication image to distinguish between legitimate and imposter samples for fingerprint, iris, and two-dimensional face biometrics.		74	**50**				2
SUPER-RESOLUTION IMAGE RECONSTRUCTION: A TECHNICAL OVERVIEW Cheol Park, S; Kyu Park, M.; Gi Kang, M. *IEEE Signal Processing Magazine* vol. 20, no. 3, 2003, pp. 21–36	This article introduces the concept of super-resolutions (SR) algorithms and presents a technical review of various existing SR methodologies and models the low-resolution image acquisition process.			43	**34**	45	90	4
IMAGE PROCESSING USING SMOOTH ORDERING OF ITS PATCHES Ram, I.; Elad, M.; Cohen, I. *IEEE Transactions on Image Processing* vol. 22, no. 7, 2013, pp. 2764–2774	This paper extracts overlapping image patches, orders these patches, and applies one-dimensional filtering to the reordered set of pixels. These techniques are applied to denoising and inpainting.			60	63	90	**36**	8
FINGERPRINT COMPRESSION BASED ON SPARSE REPRESENTATION Guangqui, S; Wu, Y; Yong, A., Liu, X.; Guo, T. *IEEE Transactions on Image Processing* vol. 23, no. 2, 2014, pp. 489–501	Compression using a sparse linear combination of dictionary atoms are used to compress three groups of finger print images and compared with JPEG, JPEG2000, and WSQ.			**66**				1
GLOBAL IMAGE DENOISING Talebi, H.; Milanfar, P. *IEEE Transactions on Image Processing* vol. 23, no. 2, 2014, pp. 755–768	This paper improves on patch similarity denoising methods by using spectral components from all pixels in an image. This global filter can be approximated by sampling a small percentage of pixels in the image.			**67**				1
COMPRESSIVE SENSING [LECTURE NOTES] Baraniuk, R.G. *IEEE Signal Processing Magazine* vol. 24, no. 4, 2007, pp. 118–121	This lecture note presents a new method to capture and represent compressible signals at a rate significantly below the Nyquist rate. This method, called compressive sensing, employs nonadaptive linear projections that preserve the structure of the signal; the signal is then reconstructed from these projections using an optimization process.			73	**39**	58	60	10

[SP]

By John Edwards

Signal Processing Leads to New Wireless Technologies

If wireless technology seems to be everywhere and with everyone these days, we can thank (or blame) signal processing for much of what's happened within the wireless industry over the past several years. From mobile phones to network routers to geolocation devices to an array of different wireless sensors, signal processing has revolutionized the way people and machines communicate and, in the process, changed the world.

Now, as researchers strive to create a fresh generation of wireless technologies as well as refine and improve existing systems, signal processing techniques continue to play a major role in helping designers create innovative features, enhance performance, shrink form factors, and lower costs. Researchers have recently turned to signal processing for assistance in projects aimed at developing an enhanced radio frequency identification (RFID) system, shrinking a frequency modulation (FM) radio transmitter down to atomic size, and replacing conventional space radio communication with a faster and more efficient optical wireless technology.

IMPROVED RFID

RFID is widely used to track everything from global food shipments to components moving down a production line to travelers' passports. Researchers at the University of Cambridge now want to make ultrahigh frequency (UHF) RFID even more ubiquitous by improving the technology's accuracy and increasing its range.

Digital Object Identifier 10.1109/MSP.2014.2330005
Date of publication: 19 August 2014

[FIG1] Sithamparanathan Sabesan, a research fellow at the University of Cambridge's Department of Engineering's Center for Photonic Systems, is collaborating with colleague Michael Crisp, Prof. Richard Penty, and Prof. Ian White on a longer range and more accurate RFID system.

Enhancing today's RFID technology promises to create an almost endless array of new location-oriented applications, such as continuous monitoring of sick and elderly people in their homes and elsewhere, real-time environmental monitoring in areas prone to natural disasters, or paying for goods without the need to visit a conventional point-of-sale system.

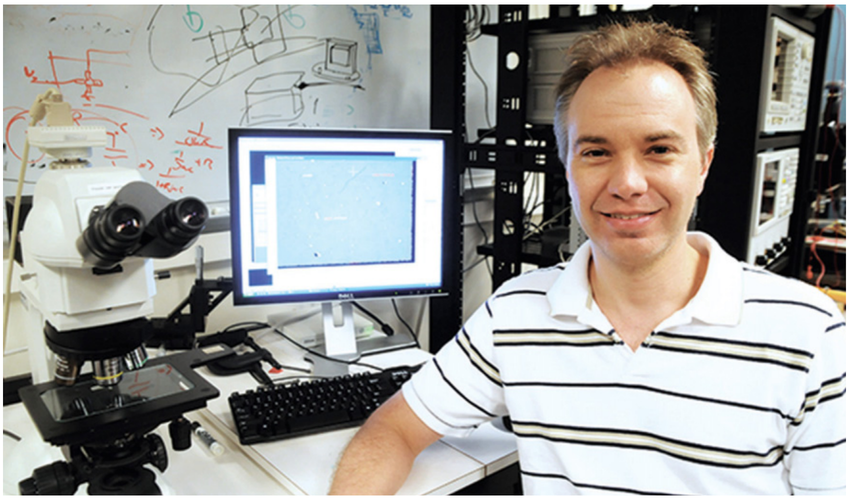
Most UHF RFID systems in place today use a reader to interrogate data that's stored on a passive (unpowered) tag. Yet many existing RFID deployments are plagued by dead spots created by various types of obstacles, such as walls and industrial equipment. "Tag detection accuracy typically degrades at a distance of approximately 2–3 m, and interrogating signals can be canceled due to reflections leading to dead spots within the radio environment," says Sithamparanathan Sabesan (Figure 1), a research fellow at the University of Cambridge's Department of Engineering's Center for Photonic Systems.

A new system created by Sabesan, in collaboration with colleague Michael Crisp, Prof. Richard Penty, and Prof. Ian White, promises to increase the accuracy of passive UHF RFID tag detection from roughly 50% to nearly 100% while boosting the reliable detection range from 2–3 m to approximately 20 m. The technology makes use of a distributed antenna system (DAS), which is similar to the type commonly used to improve Wi-Fi communications within a building.

By multicasting RFID signals over several different transmitting antennas, the researchers say they were able to dynamically move dead spots to nonessential areas to achieve an effectively error-free system. By grouping four transmitting and receiving antenna pairs, the team was able to shrink the number of dead spots in the system from nearly 50% to 0% over a 20×15-m area. Several other methods of expanding passive RFID coverage have been developed over the past few years, but none address the issues of dead spots, Sabesan notes.

The new system actually requires fewer antennas than existing configurations. Most current RFID deployments attempt to ensure accurate tag interrogations by shortening the distance between the antennas and the tags. Such arrangements demand a large number of antennas to reach an acceptable accuracy rate yet still fail to achieve completely accurate detection. "To respond to these challenges, we have developed novel technologies relating to wide-area RFID

1053-5888/14©2014IEEE



[FIG2] Kenneth Shepard, Columbia University professor of electrical engineering, collaborated with James Hone, professor of mechanical engineering, also of Columbia University, to develop the world's smallest FM radio transmitter.

using a DAS and analog signal processing techniques," Sabesan says. "In our work, we did not attempt to remove the nulls; instead, we varied the location of the nulls, moving them away from the tag and facilitating a successful reading."

Sabesan says the researchers were able to nudge the nulls by manipulating the phase between the radio-frequency (RF) signals transmitted at each antenna over a range of RF carrier frequencies in the DAS. "By making a number of read attempts with different combinations of carrier frequency and phase difference between the antennas, the physical locations of constructive and destructive interference between the antennas can be varied," he remarks. "Over a number of read attempts, all locations will experience constructive interference resulting in a 100% probability that all the RFID tags can be read."

A prototype DAS RFID system created by the researchers consisted of a base station containing a reader chip, an RF processing unit, RFID patch antennas, and passive tags. The antennas were distributed over the interrogation area using coaxial cable.

The DAS RFID controller was connected to a centralized server via an Ethernet interface, allowing tag information to be uploaded to the server for processing, analysis, and display. The RF processing unit consisted of phase shifters, switches, splitters, and combiners to perform the frequency and phase shifts, closely synchronized with the tag interrogation attempts by the RFID controller. The RF processing unit was designed to switch antennas so each could perform either transmit or receive operations in successive tag interrogations. Since the number of tags and their locations was unknown and the RF environment (the number and location of nulls) was also unknown, the relative phase and absolute frequency were randomly dithered. Over time, the randomization of the null locations allowed a tag at any location to be read.

According to the researchers, coverage can be further improved by phase hopping the received backscattered signals from each antenna to give constructive interference. Over a number of read attempts, all locations will experience constructive interference resulting in a high probability that all the RFID tags may be read.

Some difficult challenges remain unmet, Sabesan says. "These include high-quality sensing, such as precise three-dimensional tag location and speed/velocity measurement." Sabesan is looking forward to resolving the challenges. "This new RFID system can then be used in many future advanced applications such as intelligent traffic congestion management through real-time RFID-enabled vehicle tracking," he notes.

TINY TRANSMITTER

The world's smallest FM radio transmitter won't directly benefit broadcasters or two-way radio users, but the device does promise to lead to multiple signal processing applications.

Developed by Columbia University researchers, led by Kenneth Shepard, professor of electrical engineering (Figure 2), and James Hone, professor of mechanical engineering, the graphene-based transmitter takes advantage of the substance's unique properties—mechanical strength and electrical conduction—to form a nanomechanical system capable of generating FM signals.

Graphene, a single atomic layer of carbon, is the strongest known material. It also has electrical properties that are superior to the silicon used in chip manufacturing. This combination makes graphene a potentially ideal material for nanoelectromechanical systems (NEMS), scaled-down versions of the microelectromechanical systems (MEMS), widely used as vibration and acceleration sensors. "Our devices are significantly smaller than other radio-signal sources and can be put on the same chip that's used for data processing," Shepard says.

The researchers leveraged graphene's mechanical stretchability to tune the output frequency of their custom oscillator, creating a nanomechanical version of a voltage controlled oscillator (VCO). The team built a graphene NEMS (GNEMS) with a frequency tuned to 100 MHz—near the center of the U.S. FM broadcast radio band (87.7–108 MHz). The team used low-frequency musical signals to modulate the carrier signal and retrieved the signals with an ordinary FM radio receiver.

The atomically thin resonators can be tuned to within a 14% tolerance. The device was fabricated using a small graphene element suspended within a clamp formed from an SU-8 photoresist. SU-8 is a viscous polymer that can be processed with standard contact lithography and patterned into high-aspect-ratio structures.

The device operates as a resonant channel transistor in which the effect of the oscillating capacitance is amplified by the transistor action of the graphene channel. According to the researchers, such a structure is similar to both the resonant gate transistor, which utilizes an oscillating metallic gate electrode, as well as the

[FIG3] Andrew Fletcher of the MIT Lincoln Laboratory is working with counterparts at MIT and several other research facilities to develop a practical optical communications system for space missions.

recently demonstrated resonant body transistor, in which the entire transistor structure oscillates. Graphene devices differ from these CMOS-based structures in that they are orders of magnitude lower in mass and have gate-tunable resonant frequencies. The researchers note that further scaling down of graphene device size and optimization of the structure for lower parasitic capacitance and higher transconductance may enable readout of graphene NEMS in the gigahertz range for use in wireless communication and studies of fundamental physics.

While GNEMS aren't likely to ever replace conventional radio transmitters, they have significant potential applications in wireless signal processing. "Cell phones are becoming smaller, but some devices, particularly components involved in creating and processing RF signals, are very hard to miniaturize," Shepard says. Such "off-chip" components require a great deal of space and power. "Most of these components also can't be easily tuned, requiring multiple copies to cover the range of frequencies used for wireless communication."

Since GNEMs are highly compact, easily integrated, and, thanks to their great mechanical strength, capable of being tuned to a wide range of frequencies, they provide a potential solution to the size/power and tuning challenges. "When you have to do stuff with off-chip components, that adds to the size and it adds cost, and that's been the story of electronics for the last 30 years," Shepard explains. "If you can bring things onto an integrated circuit, it brings you significant advantages in reducing overall system cost but also in overall performance."

The researchers are now working to improve the noise performance of graphene oscillators. They are also planning to demonstrate integration of graphene NEMS with silicon integrated circuits, making the oscillator design even more compact.

OPTICAL IN SPACE

Radio-based communication has been a mainstay of space programs for more than a half-century. Yet today's sophisticated space missions require significantly higher data rates without the drag of extra mass or power demands.

NASA believes that the answer lies in optical communication. Recent advances in optical frequency links promise to make radioless links viable in near-term space applications. "Optical is getting ready to become the next generation of space communications," says researcher Andrew Fletcher (Figure 3) of the Massachusetts Institute of Technology (MIT) Lincoln Laboratory in Lexington. MIT researchers are working with counterparts at several other research facilities on the Laser Communications Relay Demonstration Project (LCRD) to develop a practical optical communications system for space missions. The LCRD program is being led by NASA's Goddard Space Flight Center in Greenbelt, Maryland.

Since laser-light wavelength is far shorter than radio signals, its energy remains much more concentrated as it travels through space. A typical Ka-Band radio signal from Mars, for example,

spreads out so much that its diameter when it reaches Earth is actually larger than the Earth's diameter. An equivalent optical signal, however, would have a footprint equivalent to only a small portion of the continental United States.

According to LCRD researchers, optical communications technology has the ability to achieve bidirectional near-Earth data links at speeds of 10 Gb/s and beyond utilizing differential phase shift keying (DPSK) modulation. Similarly, deep space links with downlinks of up to 1 Gb/s and uplinks up to 100 Mb/s are potentially achievable using photon counting and pulse position monitoring (PPM) modulation techniques. Photon counting PPM is highly photon efficient, but the ultimate data rate is limited due to detector limitations and the need for speedier electronics. LCRD will use both DPSK at 1.25 Gb/s and PPM at 311 Mb/s data rates.

Besides facing the challenge of creating an optical communication system's mechanical design and developing various optical subsystems, the researchers also need to address some serious communication challenges. One major issue is the fact that free-space optical signals are likely to encounter propagation environments far more extreme than the types of RF signals typically experienced.

"What happens when you shine your light through the atmosphere is that it gets bothered by the turbulence," Fletcher remarks. The atmosphere is far from a homogeneous medium, and as light passes through, it goes through varying indexes of refraction, which distorts the beam. "One of the results of that is a fading channel on the receiving end," Fletcher says.

Various techniques have been developed to deal with the challenge posed by optical signal fading. "The approach we use, which can be quite powerful, is forward error correction with some significant data interleaving," Fletcher says. "There's been some real advances over the last 20 years or so in our capabilities to do forward error correction, but it's still hard to do really advanced forward error correction at very high data rates, and it's particularly challenging to be able to do that on a space platform," Fletcher notes.

The project addressed this concern with a DPSK receiver that features an optical preamplifier stage and an optical filter where the light is split between a clock recovery unit and a communications receiver. The receiver uses a delay-line interferometer followed by balanced photodetectors to compare the phases of consecutive pulses, making a hard decision on each channel bit. While coding and interleaving will be applied in the ground terminal to mitigate noise and atmospheric fading, the DPSK receiver does not decode nor de-interleave. The modems instead support a relay architecture where uplink and downlink errors are corrected together in a decoder located at the destination ground station.

"What this allows us to do is have very high-performance signal processing that can happen at the two end points at the ground stations and not have to do it on the relay itself," Fletcher says. "The whole relay architecture is designed with this in mind to be able to leverage the really high-performance error correction that has been developed over the past few decades."

The researchers are looking toward flying and validating a reliable, capable and cost-effective optical communications technology directly applicable to the next generation of NASA's space communications network, serving both near-Earth and deep-space mission requirements. The completed payload will be flown into orbit on a commercial satellite. Mission operators at ground stations in New Mexico and California will test its invisible, near-infrared lasers, beaming data to and from the satellite as they refine the transmission process, studying different encoding techniques and perfect tracking systems.

The researchers also will study the effects of clouds and other disruptions on communications and evaluate mitigating solutions. Ground technology validation testing is set for this year. The satellite payload is scheduled to fly in 2017.

AUTHOR

John Edwards (jedwards@johnedwardsmedia.com) is a technology writer based in the Phoenix, Arizona, area.

[SP]

Georgios B. Giannakis, Francis Bach,
Raphael Cendrillon, Michael Mahoney,
and Jennifer Neville

Signal Processing for Big Data

The information explosion propelled by the advent of online social media, the Internet, and global-scale communications has rendered learning from data increasingly important. At any given time around the globe, large volumes of data are generated by today's ubiquitous communication, imaging, and mobile devices such as cell phones, surveillance cameras, medical and e-commerce platforms, as well as social networking sites. While many find this intrusive and raise legitimately "Big Brother" concerns, there is no denying that tremendous economic growth and improvement in quality of life hinge upon harnessing the potential benefits of analyzing massive data.

The term *big data* was coined to describe this information deluge, and signal processing (SP) tools and applications are clearly well seasoned to play a major role in this data science endeavor. Quoting a recent article published in *The Economist,* "The effect (of Big Data) is being felt everywhere, from business to science, and from government to the arts" [1]. Mining information from unprecedented volumes of data promises to prevent or limit the spread of epidemics and diseases, identifying trends in financial markets, learning the dynamics of emergent social-computational systems, and also protect critical infrastructure including the smart grid and the Internet's backbone network. But great promises come with formidable research challenges; as Google's chief economist explains in the same article, "Data are widely available, what is scarce is the ability to extract wisdom from them." While significant progress has been made in the last decade toward achieving the ultimate goal of "making sense of it all," the consensus is that we are still not quite there.

In this context, this special issue (SI) of *IEEE Signal Processing Magazine* (*SPM*) aims to 1) delineate the theoretical and algorithmic underpinnings along with the relevance of SP tools to the emerging field of big data and 2) introduce readers to the challenges and opportunities for SP research on (massive-scale) data analytics. The latter entails an extended and continuously refined technological wish list, which is envisioned to encompass high-dimensional, decentralized, parallel, online, and robust statistical SP, as well as large, distributed, fault-tolerant, and intelligent systems engineering. The goal of this SI is to selectively sample a diverse gamut of big data challenges and opportunities through surveys of methodological advances, as well as more focused- and application-oriented contributions chosen on the basis of timeliness, importance, and relevance to SP.

The interest in big data-related research from the SP community is evident from the increasing number of papers submitted on this topic to SP-oriented publications, workshops, and conferences. In terms of funding programs, the importance of big data research is also apparent. The White House Office of Science and Technology Policy in concert with several federal departments and agencies announced the Big Data Research and Development Initiative in 2012 [2]. The launch included generous funding in new commitments through the National Science Foundation, National Institutes of Health, Defense Advanced Research Projects Agency, and U.S. Department of Defense (DoD) at large, U.S. Department of Energy (DoE), and the U.S. Geological Survey. The DoD is placing a "Big Bet on Big Data," with two dozen open solicitations. Likewise, the European Union Commission shows increasing interest in big data analytics, e.g., under the Seventh Framework Programme for Research. All these provide ample testament that the theme of this SI is timely, and we hope that it offers something from which the SP readership will benefit.

Our opening article by Slavakis, Giannakis, and Mateos begins with a fairly rich family of models capturing a wide range of SP-relevant data analytic tasks. These include principal component analysis, nonnegative matrix factorization, dictionary learning, compressive sampling, and subspace clustering. Building on these models, the article further offers scalable inference and optimization algorithms for decentralized and online learning problems, while revealing fundamental insights into the various analytic and implementation tradeoffs involved. Generalizations of these encompassing models to timely data-sketching and tensor- and kernel-based learning tasks are also provided. The contribution finally demonstrates how the presented framework applies to several big data tasks, such as network visualization, decentralized and dynamic estimation, prediction, and imputation of network link load traffic, as well as imputation in tensor-based magnetic resonance imaging.

The second article, by Cevher, Becker, and Schmidt, places particular emphasis on recent advances in convex optimization algorithms tailored for big data, having as ultimate goal to markedly reduce the computational, storage, and communication bottlenecks. The valuable overview of this emerging field comprises contemporary approximation techniques such as first-order methods and randomization for

Digital Object Identifier 10.1109/MSP.2014.2330054
Date of publication: 19 August 2014

scalability, as well as parallel and distributed schemes that play an increasingly instrumental role in large-scale computation. The new big data algorithms outlined are based on surprisingly simple principles and attain impressive accelerations even on classical optimization tasks.

As the size of data grows, so does the chance to involve outlying observations. This in turn motivates the need for outlier-resilient learning algorithms scaling to large-scale application settings. In this context, the article by Tajer, Veeravalli, and Poor deals with robust, sequential detection schemes for big data. Outlying sequence detection is particularly important in health, the Internet, energy, telecommunications, and related large-scale problems. The article demonstrates how outlying sequence detection algorithms can be analyzed by viewing them as strategies for hypothesis testing with different outlying recovery objectives. Using this approach allows the effectiveness of outlying sequence detection strategies to be evaluated in the big data regime.

The acquisition modality, information processing, and inference from observations often dictates the need to deal with tensors—often big arrays of data collected in (hyper)cubes, thus generalizing the notion of data matrices. The growth of big data platforms makes it possible to solve large-scale tensor problems, which are encountered in various applications ranging from multiantenna communication transceivers to speech and audio, as well as machine learning from Internet data, to name a few. Sidiropoulos, Papalexakis, and Faloutsos introduce, in their article, interesting identifiability results and a parallel decomposition approach for tensors having low rank. This allows the resultant algorithms to scale nicely to sizes growing inversely proportional to the tensor rank.

High-order tensors and their decompositions are abundantly present in domains such as statistical SP (e.g., high-order moments and sensor arrays), scientific computing (e.g., discretized multivariate functions), and quantum information theory (e.g., for quantum many-body states). Representing the full tensor quickly becomes impractical for modern practical problems as the tensor's order increases. The article by Vervliet, Debals, Sorber, and De Lathauwer focuses on compact multilinear models that enable computational manipulation and estimation of such models from incomplete information.

After overviewing pertinent models and algorithms, two case studies are presented in multidimensional harmonic retrieval and material science to illustrate the potential of these approaches. In addition to matrices and tensors, big data emerge often from large-scale networks and generally graphs that are abundant in SP-relevant applications. The article by Sandryhaila and Moura highlights recent work on developing a paradigm for the analysis of graph-based data based on the so-called discrete signal processing on graphs (DSPG) approach—an effort to extend classical SP notions and techniques to data indexed by general graphs. The motivation should be clear: large data sets that are naturally modeled as graphs are generated and analyzed in a wide range of applications, and extracting valuable information from these data requires innovative approaches. Not surprisingly, some DSPG methods result from a straightforward mapping of time series to spectral graphs, which allows for drawing parallels from the former to the latter in notions as classical as filtering, spectral analysis, and transform theory. Interestingly, this is just the tip of the iceberg, since there are many subtle and fundamental issues that arise in DSPG, as articulated in this article. The discrete Fourier transform (DFT) is one of SP's "workhorses," and its popular implementation relies on the celebrated fast Fourier transform (FFT). The article by Gilbert, Indyk, Iwen, and Schmidt describes recent developments in an alternative, so-called sparse Fourier transform (SFT) implementation, which offers promises in certain large-scale data tasks involving sparse signals. The SFT can compute a compressed Fourier transform using only a subset of the input data in time, considerably shorter than the original data set size. SFT can thus be faster than the FFT when it is hard in large-scale applications to acquire enough data to run the FFT, and/or it is desirable to run DFT in time sublinear in the input size—a welcome attribute in medical imaging, when it is important to reduce the time that the patient spends in the magnetic resonance imaging machine. In addition to an overview of SFT, the article outlines the basic techniques and tradeoffs involved, as well as the connections between the SFT and related methods such as streaming algorithms and compressive sampling.

Given the deluge we experience from video, audio, medical imagery, spectroscopic, geophysical, and seismic data, the models and SP-related tools exposed in this SI promise a significant impact on many traditional but also in various emerging large-scale applications. One such innovative application wraps up this SI and deals with collaborative bike sensing for automatic geographic enrichment. Verstockt, Slavkovikj, De Potter, and Van de Walle put forth in their article a system for automatic annotation of geographical data from cyclists' smartphones. The article describes the effectiveness of this system with large-scale data sets in real-world conditions.

In closing, we would like to express our appreciation to the Editorial Board and staff of *IEEE SPM* (especially SI Area Editor Fulvio Gini) for encouraging, reviewing, welcoming, and facilitating the processing of this SI. And of course, this issue would have not been possible without the high-quality feedback received from the conscientious reviewers whom we wish to thank for their volunteer efforts and timely responses.

REFERENCES

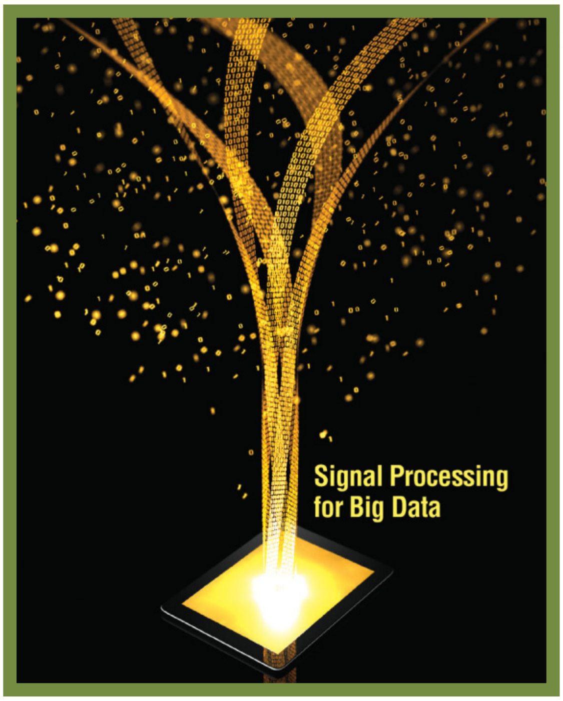
[1] K. Cukier. (2010, Feb. 25). "Data, data everywhere," *The Economist*. [Online]. Available: http://www.economist.com/node/15557443

[2] Office of Science and Technology Policy, "Big data research and development initiative," Executive Office of the President, Mar. 29, 2012. [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

[SP]

Konstantinos Slavakis, Georgios B. Giannakis, and Gonzalo Mateos

Modeling and Optimization for Big Data Analytics



© ISTOCKPHOTO.COM/TA2YO4NORI

(Statistical) learning tools for our era of data deluge

With pervasive sensors continuously collecting and storing massive amounts of information, there is no doubt this is an era of data deluge. Learning from these large volumes of data is expected to bring significant science and engineering advances along with improvements in quality of life. However, with such a big blessing come big challenges. Running analytics on voluminous data sets by central processors and storage units seems infeasible, and with the advent of streaming data sources, learning must often be performed in real time, typically without a chance to revisit past entries. "Workhorse" signal processing (SP) and statistical learning tools have to be re-examined in today's high-dimensional data regimes. This article contributes to the ongoing cross-disciplinary efforts in data science by putting forth encompassing models capturing a wide range of SP-relevant data analytic tasks, such as principal component analysis (PCA), dictionary learning (DL), compressive sampling (CS), and subspace clustering. It offers scalable architectures and optimization algorithms for decentralized and online learning problems, while revealing fundamental insights into the various analytic and implementation tradeoffs involved. Extensions of the encompassing models to timely data-sketching, tensor- and kernel-based learning tasks are also provided. Finally,

Digital Object Identifier 10.1109/MSP.2014.2327238
Date of publication: 19 August 2014

1053-5888/14©2014IEEE

the close connections of the presented framework with several big data tasks, such as network visualization, decentralized and dynamic estimation, prediction, and imputation of network link load traffic, as well as imputation in tensor-based medical imaging are highlighted.

INTRODUCTION

The information explosion propelled by the advent of online social media, Internet, and global-scale communications has rendered data-driven statistical learning increasingly important. At any time around the globe, large volumes of data are generated by today's ubiquitous communication, imaging, and mobile devices such as cell phones, surveillance cameras and drones, medical and e-commerce platforms, as well as social networking sites. The term *big data* is coined to describe this information deluge and, quoting a recent press article, "their effect is being felt everywhere, from business to science, and from government to the arts" [18]. Large economic growth and improvement in the quality of life hinge upon harnessing the potential benefits of analyzing massive data [18], [55]. Mining unprecedented volumes of data promises to limit the spread of epidemics and maximize the odds that online marketing campaigns go viral [35]; to identify trends in financial markets, visualize networks, understand the dynamics of emergent social-computational systems, as well as protect critical infrastructure including the Internet's backbone network [48], and the power grid [26].

BIG DATA CHALLENGES AND SP OPPORTUNITIES

While big data come with "big blessings," there are formidable challenges in dealing with large-scale data sets. First, the sheer volume and dimensionality of data make it often impossible to run analytics and traditional inferential methods using standalone processors, e.g., [8] and [31]. Decentralized learning with parallelized multicores is preferred [9], [22], while the data themselves are stored in the cloud or distributed file systems as in MapReduce/Hadoop [19]. Thus, there is an urgent need to explicitly account for the storage, query, and communication burden. In some cases, privacy concerns prevent disclosing the full data set, allowing only preprocessed data to be communicated through carefully designed interfaces. Due to their possibly disparate origins, big data sets are often incomplete and a sizable portion of them is missing. Large-scale data inevitably contain corrupted measurements, communication errors, and even suffer from cyberattacks as the acquisition and transportation cost per entry is driven to the minimum. Furthermore, as many of the data sources continuously generate data in real time, analytics must often be performed online subject to time constraints so that a high-quality answer obtained slowly can be less useful than a medium-quality answer that is obtained quickly [46], [48], [75].

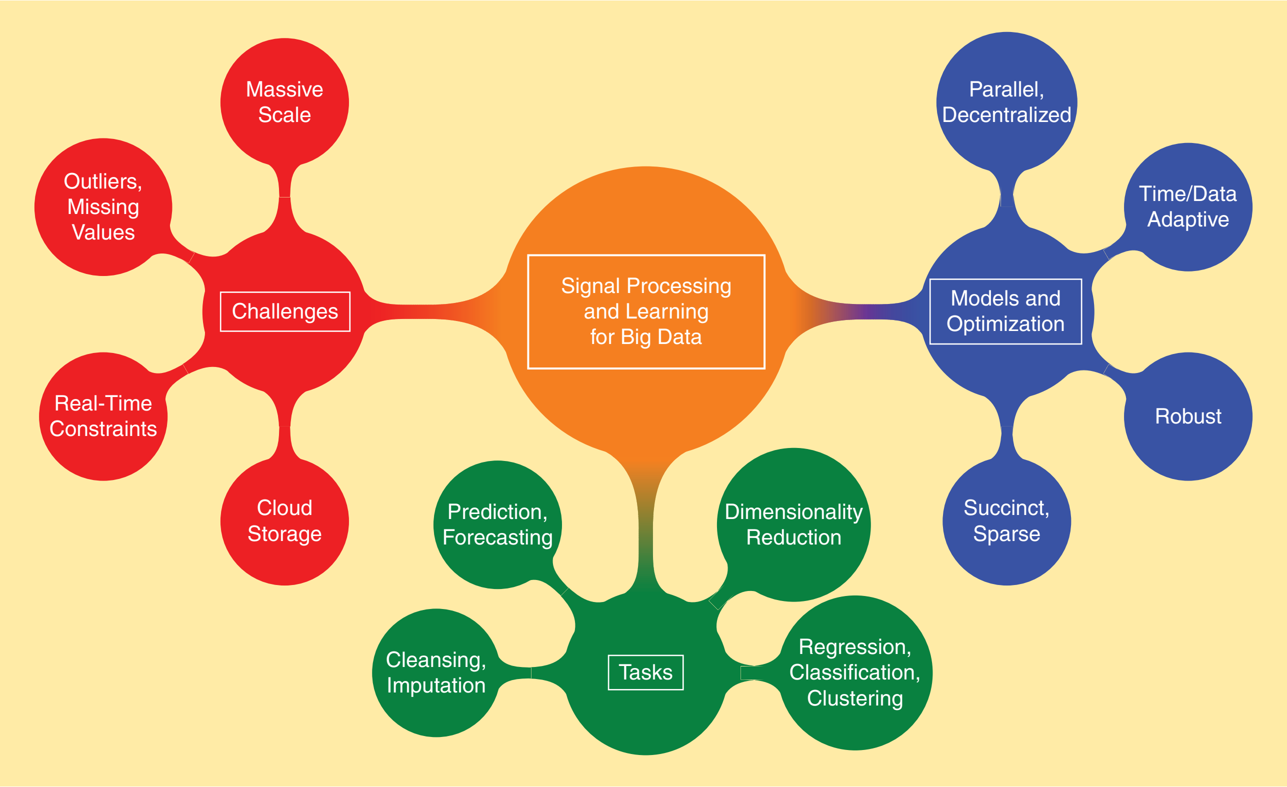
Although past research on databases and information retrieval is viewed as having focused on storage, look-up, and search, the opportunity now is to comb through massive data sets, to discover new phenomena, and to "learn" [31]. Big data challenges offer ample opportunities for SP research [55], where data-driven statistical learning algorithms are envisioned to facilitate distributed and real-time analytics (cf. Figure 1). Both classical and modern SP techniques have already placed significant emphasis on time/data adaptivity, e.g., [69], robustness [32], as well as compression and dimensionality reduction [43]. Testament to this fact is the recent "rediscovery" of stochastic approximation and stochastic-gradient algorithms for scalable online convex optimization and learning [65], oftentimes neglecting Robbins–Monro and Widrow's seminal works that go back half a century [60], [69], [79]. While the principal role of computer science in big data research is undeniable, the nature and scope of the emerging data science field is certainly multidisciplinary and welcomes SP expertise and its recent advances. For example, Web-collected data are often replete with missing entries, which motivates innovative SP imputation techniques that leverage timely (low-rank) matrix decompositions [39], [52], or, suitable kernel-based interpolators [6]. Data matrices gathering traffic values observed in the backbone of large-scale networks can be modeled as the superposition of unknown "clean" traffic, which is usually low-rank due to temporal periodicities as well as network topology-induced correlations, and traffic volume anomalies that occur sporadically in time and space, rendering the associated matrix component sparse across rows and columns [38]. Both quantity and richness of high-dimensional data sets offer the potential to improve statistical learning performance, requiring however innovative models that exploit latent low-dimensional structure to effectively separate the data "wheat from the chaff." To learn these models however, there is a consequent need to advance online, scalable optimization algorithms for information processing over graphs (an abstraction of both networked sources of decentralized data, and multiprocessor, high-performance computing architectures); see, e.g., GraphLab [42] and the alternating direction method of multipliers (ADMM) [9], [10], [51] that enjoy growing popularity for distributed machine learning tasks.

ENCOMPASSING MODELS FOR SUCCINCT BIG DATA REPRESENTATIONS

This section introduces a versatile model to fit data matrices as a superposition of a low-rank matrix capturing correlations and periodic trends, plus a linearly compressed sparse matrix explaining data innovations parsimoniously through a set of (possibly latent) factors. The model is rich enough to subsume various statistical learning paradigms with well-documented merits for high-dimensional data analysis, including PCA [28], DL [56], compressive sampling CS [11], and principal components pursuit (PCP) [12], [14], [52], to name a few.

THE "BACKGROUND" PLUS "PATTERNS AND INNOVATIONS" MODEL FOR MATRIX DATA

Let $\mathbf{L} \in \mathbb{R}^{N \times T}$ denote a low-rank matrix ($\text{rank}(\mathbf{L}) \ll \min\{N, T\}$), and $\mathbf{S} \in \mathbb{R}^{M \times T}$ a sparse matrix with support size considerably smaller than MT. Consider also the large-scale data set $\mathbf{Y} \in \mathbb{R}^{N \times T}$ generically modeled as a superposition of 1) the low-rank matrix $\mathbf{L}$; the "data background or trend," e.g., nominal



[FIG1] **SP-relevant big data themes.**

load curves across the power grid or the background scene captured by a surveillance camera, plus, 2) the “data patterns, (co) clusters, innovations, or outliers” expressed by the product of a (possibly unknown) dictionary $\mathbf{D} \in \mathbb{R}^{N \times M}$ times the sparse matrix $\mathbf{S}$, and 3) a matrix $\mathbf{V} \in \mathbb{R}^{N \times T}$, which accounts for modeling and measurement errors; in short, $\mathbf{Y} = \mathbf{L} + \mathbf{DS} + \mathbf{V}$. Matrix $\mathbf{D}$ could be an overcomplete set of bases or a linear compression operator with $N \leq M$. The aforementioned model offers a parsimonious description of $\mathbf{Y}$, that is welcomed in big data analytics where data sets involve numerous features. Such parsimony facilitates interpretability, model identifiability, and it enhances the model's predictive performance by discarding “noisy” features that bear little relevance to the phenomenon of interest [49].

To explicitly account for missing data in $\mathbf{Y}$ introduce 1) the set $\Omega \subseteq \{1, \ldots, N\} \times \{1, \ldots, T\}$ of index pairs (n, t), and 2) the sampling operator $\mathcal{P}_\Omega(\cdot)$, which nulls entries of its matrix argument not in Ω, leaving the rest unchanged. This way, one can express incomplete and (possibly noise-)corrupted data as

$$\mathcal{P}_\Omega(\mathbf{Y}) = \mathcal{P}_\Omega(\mathbf{L} + \mathbf{DS} + \mathbf{V}). \tag{1}$$

Given $\mathcal{P}_\Omega(\mathbf{Y})$, the challenging goal is to estimate the matrix components $\mathbf{L}$ and $\mathbf{S}$ (and $\mathbf{D}$ if not given), which further entails denoising the observed entries and imputing the missing ones.

An estimator leveraging the low-rank property of $\mathbf{L}$ and the sparsity of $\mathbf{S}$ will be sought to fit the data $\mathcal{P}_\Omega(\mathbf{Y})$ in the least-squares (LS) error sense, as well as minimize the rank of $\mathbf{L}$, and the number of nonzero entries of $\mathbf{S} := [s_{m,t}]$ measured by its ℓ_0-(pseudo) norm. Unfortunately, albeit natural both rank and ℓ_0-norm criteria are in general NP-hard to optimize [53]. With $\sigma_k(\mathbf{L})$ denoting the kth singular value of $\mathbf{L}$, the nuclear norm $\|\mathbf{L}\|_* := \sum_k \sigma_k(\mathbf{L})$, and the ℓ_1-norm $\|\mathbf{S}\|_1 := \sum_{m,t} |s_{m,t}|$ are adopted as surrogates, as they are the closest convex approximants to $\operatorname{rank}(\mathbf{L})$ and $\|\mathbf{S}\|_0$, respectively, e.g., [14] and [48]. Accordingly, assuming known $\mathbf{D}$ for now, one solves

$$\min_{\{\mathbf{L},\mathbf{S}\}} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{Y} - \mathbf{L} - \mathbf{DS})\|_F^2 + \lambda_* \|\mathbf{L}\|_* + \lambda_1 \|\mathbf{S}\|_1, \tag{P1}$$

where $\lambda_*, \lambda_1 \geq 0$ are rank- and sparsity-controlling parameters. Being convex, (P1) is computationally appealing as elaborated in the section “Algorithms,” in addition to being widely applicable as it encompasses a gamut of known paradigms. Notice however that when $\mathbf{D}$ is unknown, one obtains a bilinear model that gives rise to nonconvex estimation criteria. The approaches highlighted next can in fact accommodate more general models than (P1), where data-fitting terms other than the Frobenius-norm one and different regularizers can be utilized to account for various types of a priori knowledge, e.g., structured sparsity or smoothness.

APPLICATION DOMAINS AND SUBSUMED PARADIGMS

Model (1) emerges in various applications, such as 1) network anomaly detection outlined in the section “Inference and Imputation,” where $\mathbf{Y} \in \mathbb{R}^{N \times T}$ represents traffic volume over N links and T time slots; $\mathbf{L}$ captures the nominal link-level traffic (which is low-rank due to temporal periodicities and topology-induced correlations on the underlying flows); $\mathbf{D}$ represents a link × flow binary routing matrix; and $\mathbf{S}$ sparse anomalous flows [47], [48]; 2) medical imaging, where dynamic magnetic resonance imaging separates the background $\mathbf{L}$ from the motion component (e.g., a heart beating) modeled via sparse dictionary representation $\mathbf{DS}$ [25] (see also the section "Inference and Imputation"); 3) face recognition in the presence of shadows and specularities [12]; and 4) acoustic SP for singing voice separation from its music accompaniment [71], to name a few.

In the absence of $\mathbf{L}$ and missing data ($\mathbf{L}=\mathbf{0}$, $\Omega=\{1,\ldots,N\}\times\{1,\ldots,T\}$), model (1) describes an underdetermined sparse signal recovery problem typically encountered with CS [11]. If in addition $\mathbf{D}$ is unknown, (P1) boils down to DL [2], [46], [56], [67], or, to nonnegative matrix factorization (NNMF) if the entries of $\mathbf{D}$ and $\mathbf{S}$ are nonnegative [39]. For $\mathbf{L}=\mathbf{0}$, $\Omega=\{1,\ldots,N\}\times\{1,\ldots,T\}$, and if the columns of $\mathbf{Y}$ lie close to a union of a small number of unknown low-dimensional linear subspaces, then looking for a sparse $\mathbf{S}$ in (1) with $M \ll T$ amounts to subspace clustering [78]; see also [70] for outlier-robust variants with strong performance guarantees. Without $\mathbf{D}$ and with $\mathbf{V}=\mathbf{0}$, decomposing $\mathbf{Y}$ into $\mathbf{L}+\mathbf{S}$ corresponds to PCP, also referred to as *robust PCA* (*R-PCA*) [12], [14]. Even when $\mathbf{L}$ is nonzero, one could envision a variant where the measurements are corrupted with correlated (low-rank) noise [15]. Last but not least, when $\mathbf{S}=\mathbf{0}$ and $\mathbf{V}\neq\mathbf{0}$, recovery of $\mathbf{L}$ subject to a rank constraint is nothing else than PCA—arguably, the workhorse of high-dimensional big data analytics [28]. This same formulation is adopted for low-rank matrix completion—the basic task carried out by recommender systems—to impute the missing entries of a low-rank matrix observed in noise, i.e., $\mathcal{P}_\Omega(\mathbf{Y})=\mathcal{P}_\Omega(\mathbf{L}+\mathbf{V})$ [13]. Based on the maximum likelihood principle, an alternative approach for missing value imputation by expectation-maximization can be found in [73].

ALGORITHMS

As (P1) is jointly convex with respect to (w.r.t.) both $\mathbf{L}$ and $\mathbf{S}$, various iterative solvers are available, including interior point methods and centralized online schemes based on (sub)gradient-based recursions [65]. For big data however, off-the-shelf interior point methods are computationally prohibitive, and are not amenable to decentralized or parallel implementations. Subgradient-based methods are structurally simple but are often hindered by slow convergence due to restrictive step size selection rules. The desiderata for large-scale problems are low-complexity, real-time algorithms capable of processing massive data sets in a parallelizable and/or fully decentralized fashion. The few such algorithms available can be classified as decentralized or parallel schemes, splitting, sequential, and online or streaming.

DECENTRALIZED AND PARALLEL ALGORITHMS

In these divide-and-conquer schemes, multiple agents operate in parallel on disjoint or randomly subsampled subsets of the massive-scale data, and combine their outputs as iterations proceed to accomplish the original learning or inference task [34], [44]. Unfortunately, the nuclear-norm $\|\mathbf{L}\|_*$ in (P1) cannot be easily distributed across multiple learners, since the full singular value decomposition (SVD) of $\mathbf{L}$ has to be computed centrally, prior distributing its set of singular values to each node. In search of a nuclear-norm surrogate amenable to decentralized processing, it is useful to recall that minimizing $\|\mathbf{L}\|_*$ is tantamount to minimizing $(\|\mathbf{P}\|_F^2+\|\mathbf{Q}\|_F^2)/2$, where $\mathbf{L}{=}\mathbf{PQ}^\top$, with $\mathbf{P}{\in}\mathbb{R}^{N\times\rho}$ and $\mathbf{Q}{\in}\mathbb{R}^{T\times\rho}$, for some $\rho{\ll}\min\{N,T\}$, is a bilinear decomposition of the low-rank component $\mathbf{L}$ [47], [72]. In other words, each column vector of $\mathbf{L}$ is assumed to lie in a low ρ-dimensional range space spanned by the columns of $\mathbf{P}$. This gives rise to the following problem:

$$\min_{\{\mathbf{P},\mathbf{Q},\mathbf{S}\}} \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y}-\mathbf{PQ}^\top-\mathbf{DS})\|_F^2+\frac{\lambda_*}{2}(\|\mathbf{P}\|_F^2+\|\mathbf{Q}\|_F^2)+\lambda_1\|\mathbf{S}\|_1. \tag{P2}$$

Unlike (P1), the bilinear term $\mathbf{PQ}^\top$ renders (P2) nonconvex, even if $\mathbf{D}$ is known. Interestingly, [47, Prop. 1] offers a certificate for stationary points of (P2), qualifying them as global optima of (P1).

Thanks to the decomposability of $\|\cdot\|_F^2$ and $\|\cdot\|_1$ across rows, and ignoring for a moment the operator $\mathcal{P}_\Omega$, (P2) can be distributed over a number $\mathcal{V}$ of nodes or processing cores $\mathscr{V}$ with cardinality $|\mathscr{V}|=\mathcal{V}$, where each node $\nu\in\mathscr{V}$ learns from a subset of rows $\mathscr{R}_\nu\subset\{1,\ldots,N\}$. In other words, the N rows of $\mathbf{Y}$ are distributed over a partition of rows $\{\mathscr{R}_\nu\}_{\nu=1}^{\mathcal{V}}$, where by definition $\bigcup_{\nu=1}^{\mathcal{V}}\mathscr{R}_\nu=\{1,\ldots,N\}$, and $\mathscr{R}_{\nu_i}\cap\mathscr{R}_{\nu_j}=\emptyset$, if $i\neq j$. Naturally, (P2) is equivalent to this (modulo $\mathcal{P}_\Omega$) task:

$$\min_{\{\{\mathbf{P}_\nu\}_{\nu=1}^{\mathcal{V}},\mathbf{Q},\mathbf{S}\}} \frac{1}{2}\sum_{\nu=1}^{\mathcal{V}}\|\mathbf{Y}_\nu-\mathbf{P}_\nu\mathbf{Q}^\top-\mathbf{D}_\nu\mathbf{S}\|_F^2 + \frac{\lambda_*}{2}\sum_{\nu=1}^{\mathcal{V}}\|\mathbf{P}_\nu\|_F^2+\frac{\lambda_*}{2}\|\mathbf{Q}\|_F^2+\lambda_1\|\mathbf{S}\|_1, \tag{2}$$

where $\mathbf{Y}_\nu,\mathbf{P}_\nu$, and $\mathbf{D}_\nu$ are submatrices formed by keeping only the $\mathscr{R}_\nu$ rows of $\mathbf{Y},\mathbf{P}$, and $\mathbf{D}$, respectively.

An obstacle in (2) is the coupling of the data-fitting term with the regularization terms via $\{\mathbf{P}_\nu,\mathbf{Q},\mathbf{S}\}$. Direct utilization of iterative subgradient-type methods, due to the nonsmooth loss function, are able to identify local minimizers of (2), at the cost of slow convergence and meticulous choice of step sizes. In the convex analysis setting, successful optimization approaches to surmount this obstacle include the ADMM [10] and the more general Douglas–Rachford (DR) algorithm [5] that split or decouple variables in the nuclear-, ℓ_1-, and Frobenius-norms. The crux of splitting methods, such as ADMM and DR, lies on computing efficiently the proximal mapping of regularizing functions, which for a (non)differentiable lower-semicontinuous convex function g and $\gamma>0$, is defined as $\mathrm{Prox}_{\gamma g}(\mathbf{A}):=\arg\min_{\mathbf{A}'}(1/2)\|\mathbf{A}-\mathbf{A}'\|_F^2+\gamma g(\mathbf{A}')$, $\forall\mathbf{A}$ [5]. The computational cost incurred by $\mathrm{Prox}_{\gamma g}$ depends on g. For example, if g is the nuclear-norm, then $\mathrm{Prox}_{\gamma\|\cdot\|_*}(\mathbf{A})=\mathbf{U}\,\mathrm{Soft}_\gamma(\boldsymbol{\Sigma})\mathbf{V}^\top$, where $\mathbf{A}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the computationally demanding SVD of $\mathbf{A}$, and $\mathrm{Soft}_\gamma(\boldsymbol{\Sigma})$ is the soft-thresholding operator whose (i,j)th entry is $[\text{Soft}_\gamma(\Sigma)]_{ij} = \text{sgn}([\Sigma]_{i,j}) \max\{0, |[\Sigma]_{i,j}| - \gamma\}$. On the contrary, if $g = \|\cdot\|_1$, then $\text{Prox}_{\gamma\|\cdot\|_1}(\mathbf{A}) = \text{Soft}_\gamma(\mathbf{A})$, which is a computationally affordable, parallelizable operation.

Even if (2) is a nonconvex task, a splitting strategy mimicking ADMM and DR is promising also in the current context. If the network nodes or cores can also exchange messages, then (2) can be decentralized. This is possible if e.g., $\nu \in \mathscr{V}$ has a neighborhood $\mathscr{N}_\nu \subset \mathscr{V}$, where $\nu \in \mathscr{N}_\nu$ and all members of $\mathscr{N}_\nu$ exchange information. The decentralized rendition of (P2) becomes

$$\min_{\left\{\substack{\mathbf{P}_\nu, \mathbf{Q}_\nu, \mathbf{S}_\nu \\ \mathbf{P}'_\nu, \mathbf{Q}'_\nu, \mathbf{S}'_\nu}\right\}} \frac{1}{2}\| \mathscr{P}_{\Omega_\nu}(\mathbf{Y}_\nu - \mathbf{P}_\nu \mathbf{Q}_\nu^\top - \mathbf{D}_\nu \mathbf{S}_\nu) \|_F^2$$
$$+ \frac{\lambda_*}{2}(\|\mathbf{P}'_\nu\|_F^2 + \|\mathbf{Q}'_\nu\|_F^2) + \lambda_1 \|\mathbf{S}'_\nu\|_1, \forall \nu \in \mathscr{V}$$
$$\text{s.to} \quad \forall \nu \in \mathscr{V}: \begin{cases} \forall \nu' \in \mathscr{N}_\nu : \begin{cases} \mathbf{Q}_\nu = \mathbf{Q}_{\nu'}, & \mathbf{S}_\nu = \mathbf{S}_{\nu'} \\ \mathbf{Q}'_\nu = \mathbf{Q}'_{\nu'}, & \mathbf{S}'_\nu = \mathbf{S}'_{\nu'}, \end{cases} \\ \mathbf{P}_\nu = \mathbf{P}'_\nu \end{cases} \tag{P3}$$

where consensus constraints are enforced per neighborhood $\mathscr{N}_\nu$, and $\{\mathbf{P}'_\nu, \mathbf{Q}'_\nu, \mathbf{S}'_\nu\}$ are utilized to split the LS cost from the Frobenius- and ℓ_1-norms. Typically, (P3) is expressed in unconstrained form using the (augmented) Lagrangian framework. Decentralized inference algorithms over networks, implementing the previous splitting methodology, can been found in [22], [47], [51], and [62]. ADMM and DR are convergent for convex costs, but they offer no convergence guarantees for the nonconvex (P3). There is, however, ample experimental evidence in the literature that supports empirical convergence of ADMM, especially when the nonconvex problem at hand exhibits "favorable" structure [10], [47].

Methods offering convergence guarantees for (P3), after encapsulating consensus constraints into the loss function, are sequential schemes, such as the block coordinate descent methods (BCDMs) [59], [77]. BCDMs minimize the underlying objective sequentially over one block of variables per iteration, while keeping all other blocks fixed to their most up-to-date values. For example, a BCDM for solving the DL subtask of (2), that is when $\{\mathbf{P}_\nu, \mathbf{Q}\}$ are absent from the optimization problem, is the K-SVD algorithm [2]. Per iteration, K-SVD alternates between sparse coding of the columns of $\mathbf{Y}$ based on the current dictionary and updating the dictionary atoms to better fit the data. For a consensus-based decentralized implementation of K-SVD in the cloud, see [58].

It is worth stressing that (P3) is convex w.r.t. each block among $\{\mathbf{P}_\nu, \mathbf{Q}_\nu, \mathbf{S}_\nu, \mathbf{P}'_\nu, \mathbf{Q}'_\nu, \mathbf{S}'_\nu\}$, whenever the rest are held constant. Recent *parallel* schemes with convergence guarantees take advantage of this underlying structure to speed-up decentralized and parallel optimization algorithms [33], [64]. Additional BCDM examples will be given next in the context of online learning.

ONLINE ALGORITHMS FOR STREAMING ANALYTICS

So far, $\mathbf{Y}$ has been decomposed across its rows corresponding to network agents or processors; in what follows, $\mathbf{Y}$ will be split across its columns. Aiming at online solvers of (P2), with t indexing the columns of $\mathbf{Y} := [\mathbf{y}_1, \ldots, \mathbf{y}_t]$, and $\{\Omega_\tau\}_{\tau=1}^t$ indicating the locations of known data values across time, consider the analytics engine acquiring a stream of vectors $\mathscr{P}_{\Omega_t}(\mathbf{y}_t)$, $\forall t$. An online counterpart of (P2) is the following exponentially weighted LS estimate [48]

$$\min_{\left\{\substack{\mathbf{P} \\ \{\mathbf{q}_\tau, \mathbf{s}_\tau\}_{\tau=1}^t}\right\}} \sum_{\tau=1}^{t} \delta^{t-\tau} \left[\frac{1}{2}\| \mathscr{P}_{\Omega_\tau}(\mathbf{y}_\tau - \mathbf{P}\mathbf{q}_\tau - \mathbf{D}_\tau \mathbf{s}_\tau) \|^2 \right.$$
$$\left. + \frac{\lambda_*}{2\sum_{\tau'=1}^{t} \delta^{t-\tau'}} \|\mathbf{P}\|_F^2 + \frac{\lambda_*}{2}\|\mathbf{q}_\tau\|^2 + \lambda_1 \|\mathbf{s}_\tau\|_1 \right], \tag{P4}$$

where $\mathbf{P} \in \mathbb{R}^{N \times \rho}$, $\{\mathbf{q}_\tau\}_{\tau=1}^t \subset \mathbb{R}^\rho$, $\{\mathbf{s}_\tau\} \subset \mathbb{R}^M$, and $\delta \in (0, 1]$ denotes the so-termed forgetting factor. With $\delta < 1$, past data are exponentially discarded to track nonstationary features. Clearly, $\mathscr{P}_{\Omega_t}$ can be represented by a matrix $\boldsymbol{\Omega}_t$, whose rows are a subset of the rows of the N-dimensional identity matrix.

A provably convergent BCDM approach to efficiently solve a simplified version of (P4) was put forth in [48]. Each time t a new datum is acquired, only $\mathbf{q}_t$ and $\mathbf{s}_t$ are jointly updated via Lasso for fixed $\mathbf{P} = \mathbf{P}_{t-1}$, and then (P4) is solved w.r.t. $\mathbf{P}$ to update $\mathbf{P}_{t-1}$ using recursive LS (RLS). The latter step can be efficiently split across rows $\mathbf{p}_{n,t} = \arg\min_{\mathbf{p}} \sum_{\tau=1}^{t} \delta^{t-\tau} \omega_{n,\tau}(y_{n,\tau} - \mathbf{p}^\top \mathbf{q}_\tau - \mathbf{d}_{n,\tau}^\top \mathbf{s}_\tau)^2 + (\lambda_*/2)\|\mathbf{p}\|^2$—an attractive feature facilitating parallel processing, which nevertheless entails a matrix inversion when $\delta < 1$. Since first introduced in [48], the idea of performing online rank-minimization leveraging the separable nuclear-norm regularization in (P4) has gained popularity in real-time NNMF for audio SP [71], and online robust PCA [21], to name a few examples. In the case where $\mathbf{P}, \{\mathbf{q}_\tau\}_{\tau=1}^t$ are absent from (P4), an online DL method of the same spirit as in [48] can be found in [46], [67].

Algorithms in [48] are closely related to timely robust subspace trackers, which aim at estimating a low-rank subspace $\mathbf{P}$ from grossly corrupted and possibly incomplete data, namely $\mathscr{P}_{\Omega_t}(\mathbf{y}_t) = \mathscr{P}_{\Omega_t}(\mathbf{P}\mathbf{q}_t + \mathbf{s}_t + \mathbf{v}_t)$, $t = 1, 2, \ldots$. In the absence of sparse outliers $\{\mathbf{s}_t\}_{t=1}^\infty$, an online algorithm based on incremental gradient descent on the Grassmannian manifold of subspaces was put forth in [4]. The second-order RLS-type algorithm in [16] extends the seminal projection approximation subspace tracking (PAST) algorithm to handle missing data; see also [50]. When outliers are present, robust counterparts can be found in [15] and [29]. Relative to all aforementioned works, the estimation problem (P4) is more challenging due to the presence of the (compression) dictionary $\mathbf{D}_t$.

Reflecting on (P1)–(P4), all objective functions share a common structure: they are convex w.r.t. each of their variable blocks, provided the rest are held fixed. Naturally, this calls for BCDMs for minimization, as in the previous discussion. However, matrix inversions and solving a batch Lasso per slot t may prove prohibitive for large-scale optimization tasks. Projected or proximal stochastic (sub)gradient methods are attractive low-complexity online alternatives to BCDMs mainly for optimizing convex objectives [65]. Unfortunately, due to their diminishing step-sizes, such first-order solutions exhibit slow convergence even for convex problems. On the other hand, accelerated variants for convex problems offer quadratic convergence of the objective function values, meaning they are optimally “fast” among first-order methods [54], [80]. Although quadratic convergence issues for nonconvex and time-varying costs as in (P4) are largely unexplored, the online, accelerated, first-order method outlined in Figure 2 offers a promising alternative for generally nonsmooth and nonconvex minimization tasks [68].

Let $\mathbf{x}^{(i)}$ be a block of variables, which in (P4) can be $\mathbf{P}$, or $\{\mathbf{q}_\tau\}_{\tau=1}^{t}$, or $\{\mathbf{s}_\tau\}_{\tau=1}^{t}$; that is, $i \in \{1,2,3\}$; and let $\mathbf{x}^{(-i)}$ denote all blocks in $\mathbf{x} := (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(I)})$ except for $\mathbf{x}^{(i)}$. Consider the sequence of loss functions $F_t(\mathbf{x}) := f_t(\mathbf{x}) + \sum_{i=1}^{I} g_i(\mathbf{x}^{(i)})$, where f_t is nonconvex, and Lipschitz continuously differentiable but convex w.r.t. each $\mathbf{x}^{(i)}$, whenever $\{\mathbf{x}^{(j)}\}_{j \neq i}$ are held fixed; $\{g_i\}_{i=1}^{I}$ are convex and possibly nondifferentiable; hence, F_t is nonsmooth. Clearly, the data fit term in (P4) corresponds to f_t, $g_1(\mathbf{x}^{(1)}) := (\lambda_*/2)\|\mathbf{P}\|_F^2$, while g_2 and g_3 describe the other two regularization terms.

The acceleration module Accel of [80], developed originally for offline convex analytic tasks, is applied to F_t in a sequential, per-block (Gauss–Seidel) fashion. Having $\mathbf{x}^{(-i)}$ fixed, unless $\min_{\mathbf{x}^{(i)} \in \mathscr{X}_i} f_t(\mathbf{x}^{(i)} \mid \mathbf{x}_t^{(-i)}) + g_i(\mathbf{x}^{(i)})$ is easily solvable, Accel is employed for $R_i \geq 1$ times to update $\mathbf{x}^{(i)}$. The same procedure is carried over to the next block $\mathbf{x}^{(i+1)}$, until all blocks are updated, and subsequently to the next time instant $t+1$ (Figure 2). Unlike ADMM, this first-order algorithm requires no matrix inversions, and can afford inexact solutions of minimization subtasks. Under several conditions, including (statistical) stationarity of $\{F_t\}_{t=1}^{\infty}$, it also guarantees quadratic-rate convergence to a stationary point of $\mathbb{E}\{F_t\}$, where $\mathbb{E}\{\cdot\}$ denotes expectation over noise and input data distributions [68]. An application of this method to the dictionary-learning context can be found in the “Inference and Imputation” section.

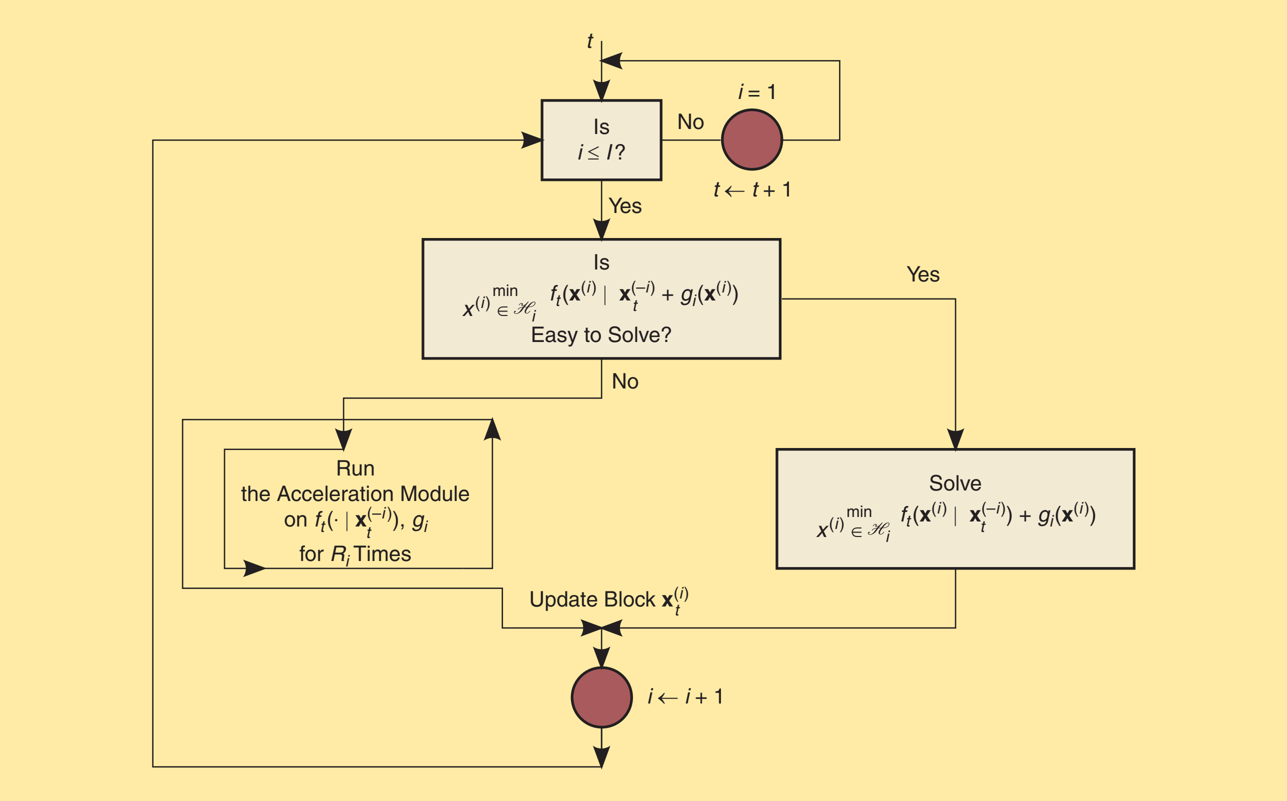
DATA SKETCHING, TENSORS, AND KERNELS

The scope of the “Algorithms” section can be broadened to include random subsampling schemes on $\mathbf{Y}$ (also known as data sketching), as well as multiway data arrays (tensors) and nonlinear modeling via kernel functions.

DATA SKETCHING

Catering to decentralized or parallel solvers, all variables in (P3) should be updated in parallel across learners of individual network nodes. However, there are cases where solving all learning subtasks simultaneously may be prohibitive or inefficient for two main reasons. First, the data size might be so large that computing function values or first-order information over all variables is impossible. Second, the nature and structure of data may prevent a fully parallel operation; e.g., when data are not available in their entirety, but are acquired either in batches over time or where not all of the network nodes are equally responsive or functional.

A recent line of research aiming at obtaining informative subsets of measurements for asynchronous and reduced-dimensionality processing of big data sets is based on (random) subsampling or data



t
$i = 1$
Is $i \leq I$?
No
$t \leftarrow t + 1$
Yes
Is $\min_{x^{(i)} \in \mathscr{X}_i} f_t(\mathbf{x}^{(i)} \mid \mathbf{x}_t^{(-i)}) + g_i(\mathbf{x}^{(i)})$ Easy to Solve?
Yes
No
Run the Acceleration Module on $f_t(\cdot \mid \mathbf{x}_t^{(-i)})$, g_i for R_i Times
Solve $\min_{x^{(i)} \in \mathscr{X}_i} f_t(\mathbf{x}^{(i)} \mid \mathbf{x}_t^{(-i)}) + g_i(\mathbf{x}^{(i)})$
Update Block $\mathbf{x}_t^{(i)}$
$i \leftarrow i + 1$

[FIG2] **The online, accelerated, sequential (Gauss–Seidel) optimization scheme for asymptotically minimizing the sequence $(F_t)_{t \in \mathbb{N}}$ of nonconvex functions.**

sketching (via $\mathcal{P}_\Omega$) of the massive $\mathbf{Y}$ [45]. The basic principles of data sketching will be demonstrated here for the overdetermined $(N \gg \rho)$ LS $\mathbf{q}_* := \mathbf{P}^\dagger \mathbf{y} \in \arg\min_{\mathbf{q}\in\mathbb{R}^\rho} \|\mathbf{y}-\mathbf{P}\mathbf{q}\|^2$ [a task subsumed by (P2) as well], where $\dagger$ denotes pseudo-inverse, and $\mathbf{P}^\dagger = (\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top$, for $\mathbf{P}$ full column-rank. Popular strategies to obtain $\mathbf{q}_*$ include the expensive SVD; the Cholesky decomposition if $\mathbf{P}$ is full column-rank and well conditioned; and the slower but more stable QR decomposition [45].

The basic premise of the subsampling or data sketching techniques is to largely reduce the number of rows of $\mathbf{Y}$ prior to solving the LS task [45]. A data-driven methodology of keeping only the "most" informative rows relies on the so-termed (statistical) leverage scores and is outlined next as a three-step procedure. Given the (thin) SVD $\mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$: (S1) find the normalized leverage scores $\{l_n\}_{n=1}^N$, where $l_n := \rho^{-1}\mathbf{e}_n^\top \mathbf{U}\mathbf{U}^\top \mathbf{e}_n = \rho^{-1}\mathbf{e}_n^\top \mathbf{P}\mathbf{P}^\dagger \mathbf{e}_n$, with $\mathbf{e}_n \in \mathbb{R}^N$ being the nth canonical vector. Clearly, l_n equals the (normalized) nth diagonal element of $\mathbf{P}\mathbf{P}^\dagger$, and since $\mathbf{P}\mathbf{P}^\dagger = \mathbf{U}\mathbf{U}^\top$ is the orthogonal projector onto the linear subspace spanned by the columns of $\mathbf{P}$, it follows that $\mathbf{P}\mathbf{P}^\dagger\mathbf{y}$ offers the best approximation to $\mathbf{y}$ within this subspace. Then, (S2) for an arbitrarily small $\epsilon > 0$, and by using $\{l_n\}_{n=1}^N$ as an importance sampling distribution, randomly sample and rescale by $(rl_n)^{-1}$ a number of $r = \mathcal{O}(\epsilon^{-2}\rho\log\rho)$ rows of $\mathbf{P}$, together with the corresponding entries of $\mathbf{y}$. Such a sampling and rescaling operation can be expressed by a matrix $\boldsymbol{\Psi} \in \mathbb{R}^{r\times N}$. Finally, (S3) solve the reduced-size LS problem $\tilde{\mathbf{q}}_* \in \arg\min_{\mathbf{q}\in\mathbb{R}^\rho} \|\boldsymbol{\Psi}(\mathbf{y}-\mathbf{P}\mathbf{q})\|^2$. With $\kappa(\cdot)$ denoting condition number and $\gamma := \|\mathbf{y}\|^{-1}\|\mathbf{U}\mathbf{U}^\top\mathbf{y}\|$, it holds that [45]

$$\|\mathbf{y}-\mathbf{P}\tilde{\mathbf{q}}_*\| \le (1+\epsilon)\|\mathbf{y}-\mathbf{P}\mathbf{q}_*\| \tag{3a}$$

$$\|\mathbf{q}_*-\tilde{\mathbf{q}}_*\| \le \sqrt{\epsilon}\,\kappa(\mathbf{P})\sqrt{\gamma^{-2}-1}\,\|\mathbf{q}_*\| \tag{3b}$$

so that performance degrades gracefully after reducing the number of equations.

Similar to the nuclear-norm, a major difficulty is that leverage scores are not amenable to decentralized computation [cf. discussion prior (P2)], since the SVD of $\mathbf{P}$ is necessary prior to decentralizing the original learning task. To avoid computing the statistical leverage scores, the following data-agnostic strategy has been advocated [45]: 1) Premultiply $\mathbf{P}$ and $\mathbf{y}$ with the $N\times N$ random Hadamard transform $\mathbf{H}_N\boldsymbol{\Delta}$, where $\mathbf{H}_N$ is defined inductively as

$$\mathbf{H}_N = \frac{1}{\sqrt{N}}\begin{bmatrix}\mathbf{H}_{N/2} & \mathbf{H}_{N/2}\\ \mathbf{H}_{N/2} & -\mathbf{H}_{N/2}\end{bmatrix}, \mathbf{H}_2 := \frac{1}{\sqrt{2}}\begin{bmatrix}+1 & +1\\ +1 & -1\end{bmatrix},$$

and $\boldsymbol{\Delta}$ is a diagonal matrix whose nonzero entries are drawn independently and uniformly from $\{-1,+1\}$, 2) uniformly sample and rescale a number of $r = \mathcal{O}(\rho\log\rho\cdot\log N + \epsilon^{-1}N\log\rho)$ rows from $\mathbf{H}_N\boldsymbol{\Delta}\mathbf{P}$ together with the corresponding components from $\mathbf{H}_N\boldsymbol{\Delta}\mathbf{y}$, and 3) find $\tilde{\mathbf{q}}_* \in \arg\min_{\mathbf{q}\in\mathbb{R}^\rho}\|\boldsymbol{\Psi}\mathbf{H}_N\boldsymbol{\Delta}(\mathbf{y}-\mathbf{P}\mathbf{q})\|^2$, where $\boldsymbol{\Psi}$ stands again for the sampling and rescaling operation. Error bounds similar to those in (1) can be also derived for this preconditioning strategy [45]. Key to deriving such performance bounds is the Johnson–Lindenstrauss lemma, which loosely asserts that for any $\epsilon \in (0,1)$, any set of ρ points in N dimensions can be (linearly) embedded into $r \ge 4(2^{-1}\epsilon^2 - 3^{-1}\epsilon^3)^{-1}\ln\rho$ dimensions, while preserving the pairwise Euclidean distances of the original points up to a multiplicative factor of $(1\pm\epsilon)$.

Besides the previous overdetermined LS task, data sketching has been employed to ease the computational burden of several large-scale tasks ranging from generic matrix multiplication, SVD computation, to k-means clustering and tensor approximation [20], [45]. In the spirit of $\mathbf{H}_N\boldsymbol{\Delta}$, methods utilizing sparse embedding matrices have been also developed for overconstrained LS and ℓ_p-norm regression, low-rank and leverage scores approximation [17]; in particular, they exhibit complexity $\mathcal{O}(|\operatorname{supp}(\mathbf{P})|) + \mathcal{O}(\rho^3\epsilon^{-2}\log^l(\rho^3\epsilon^{-2}))$ for solving the LS task satisfying (3a), where $|\operatorname{supp}(\mathbf{P})|$ stands for the cardinality of the support of $\mathbf{P}$, and $l \in \mathbb{N}_*$. Viewing the sampling and rescaling operator $\boldsymbol{\Psi}$ as a special case of a (weighted) $\mathcal{P}_\Omega$ allows carrying over the algorithms outlined in the "Encompassing Models for Succinct Big Data Representations" and "Algorithms" sections to the data sketching setup as well.

BIG DATA TENSORS

Although the matrix model in (1) is quite versatile and can subsume a variety of important frameworks as special cases, the particular planar arrangement of data poses limitations in capturing available structures that can be crucial for effective interpolation. In the example of movie recommender systems, matrix models can readily handle two-dimensional structures of people × movie ratings. However, movies are classified in various genres and one could explicitly account for this information by arranging ratings in a sparse person × genre × title three-way array or tensor. In general, various tensor data analytic tasks for network traffic, social networking, or medical data analysis aim at capturing an underlying latent structure, which calls for high-order factorizations even in the presence of missing data [1], [50].

A rank-one three-way array $\underline{\mathbf{Y}} = [y_{i_a i_b i_c}] \in \mathbb{R}^{I_a\times I_b\times I_c}$, where the underline denotes tensors, is the outer product $\mathbf{a}\circ\mathbf{b}\circ\mathbf{c}$ of three vectors $\mathbf{a}\in\mathbb{R}^{I_a}, \mathbf{b}\in\mathbb{R}^{I_b}, \mathbf{c}\in\mathbb{R}^{I_c}$: $y_{i_a i_b i_c} = a_{i_a}b_{i_b}c_{i_c}$. One can interpret a_{i_a}, b_{i_b}, and c_{i_c} as corresponding to the people, genre, and title components, respectively, in the previous example. The rank of a tensor is the smallest number of rank-one tensors that sum up to generate the given tensor. These notions readily generalize to higher-way tensors, depending on the application. Notwithstanding, this is not an incremental extension from low-rank matrices to low-rank tensors, since even computing the tensor rank is an NP-hard problem in itself [36]. Defining a convex surrogate for the rank penalty such as the nuclear norm for matrices is not obvious either, since singular values when applicable, e.g., in the Tucker model, are not related to the rank [74]. Although a three-way array can be "unfolded" to obtain a matrix exhibiting latent Kronecker product structure, such an unfolding typically destroys the structure that one looks for.

These considerations, motivate forming a low-rank approximation of tensor $\underline{\mathbf{Y}}$ as

$$\underline{\mathbf{Y}} \approx \sum_{r=1}^{\rho} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r. \qquad (4)$$

Low-rank tensor approximation is a relatively mature topic in multilinear algebra and factor analysis, and when exact, the decomposition (4) is called parallel factor analysis (PARAFAC) or canonical decomposition (CANDECOMP) [36]. PARAFAC is the model of choice when one is primarily interested in revealing latent structure. Unlike the matrix case, low-rank tensor decomposition can be unique. There is deep theory behind this result, and algorithms recovering the rank-one factors [37]. However, various computational and big data-related challenges remain. Missing data have been handled in rather ad hoc ways [76]. Parallel and decentralized implementations have not been thoroughly addressed; see, e.g., ParCube and GigaTensor algorithms for recent scalable approaches [57].

With reference to (4), introduce the factor matrix $\mathbf{A} := [\mathbf{a}_1, \ldots, \mathbf{a}_\rho] \in \mathbb{R}^{I_a \times \rho}$, and likewise for $\mathbf{B} \in \mathbb{R}^{I_b \times \rho}$ and $\mathbf{C} \in \mathbb{R}^{I_c \times \rho}$. Let $\mathbf{Y}_{i_c}, i_c = 1, \ldots, I_c$ denote the i_cth slice of $\underline{\mathbf{Y}}$ along its third (tube) dimension, such that $\mathbf{Y}_{i_c}(i_a, i_b) = y_{i_a i_b i_c}$. It follows that (4) can be compactly represented in matrix form, in terms of slice factorizations $\mathbf{Y}_{i_c} = \mathbf{A}\,\mathrm{diag}(\mathbf{e}_{i_c}^\top \mathbf{C})\,\mathbf{B}^\top, \forall i_c$. Capitalizing on the Frobenius-norm regularization (P2), decentralized algorithms for low-rank tensor completion under the PARAFAC model can be based on the optimization task:

$$\min_{\{\mathbf{A},\mathbf{B},\mathbf{C}\}} \sum_{i_c=1}^{I_c} \| \mathcal{P}_{\Omega_{i_c}}(\mathbf{Y}_{i_c} - \mathbf{A}\,\mathrm{diag}(\mathbf{e}_{i_c}^\top \mathbf{C})\,\mathbf{B}^\top) \|_F^2 + \lambda_* [\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2]. \qquad (5)$$

Different from the matrix case, it is unclear whether the regularization in (5) bears any relation with the tensor rank. Interestingly, [7] asserts that (5) provably yields a low-rank $\hat{\underline{\mathbf{Y}}}$ for sufficiently large λ_*, while the potential for scalable BCDM-based interpolation algorithms is apparent. For an online algorithm, see also (9) in the section "Big Data Tasks" and [50] for further details.

KERNEL-BASED LEARNING

In imputing random missing entries, prediction of multiway data can be viewed as a tensor completion problem, where an entire slice (say, the one orthogonal to the tube direction representing time) is missing. Notice that since (5) does not specify a correlation structure, it cannot perform this extrapolation task. Kernel functions provide the nonlinear means to infuse correlations or side information (e.g., user age range and educational background for movie recommendation systems) in various big data tasks spanning disciplines such as 1) statistics, for inference and prediction [28], 2) machine learning, for classification, regression, clustering, and dimensionality reduction [63], and 3) SP, as well as (non)linear system identification, sampling, interpolation, noise removal, and imputation; see, e.g., [6] and [75].

In kernel-based learning, processing is performed in a high-, possibly infinite-dimensional reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, where function $f \in \mathcal{H}$ to be learned is expressed as a superposition of kernels; i.e., $f(\mathbf{x}) := \sum_{i=1}^{\infty} \varphi_i \kappa(\mathbf{x}, \mathbf{x}_i)$, where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the kernel associated with $\mathcal{H}$, $\{\varphi_i\}_{i=1}^{\infty}$ denote the expansion coefficients, and $\mathbf{x}, \mathbf{x}_i \in \mathcal{X}, \ \forall i$ [63]. Broadening the scope of (5), a kernel-based tensor completion problem is posed as follows. With index sets $\mathcal{X}_a := \{1, \ldots, I_a\}$, $\mathcal{X}_b := \{1, \ldots, I_b\}$, and $\mathcal{X}_c := \{1, \ldots, I_c\}$, and associated kernels $\kappa_{\mathcal{X}_a}(i_a, i_a')$, $\kappa_{\mathcal{X}_b}(i_b, i_b')$ and $\kappa_{\mathcal{X}_c}(i_c, i_c')$, tensor entry $y_{i_a i_b i_c}$ is approximated using functions from the set $\mathcal{F} := \{f(i_a, i_b, i_c) = \sum_{r=1}^{\rho} a_r(i_a) b_r(i_b) c_r(i_c) | a_r \in \mathcal{H}_{\mathcal{X}_a}, b_r \in \mathcal{H}_{\mathcal{X}_b}, c_r \in \mathcal{H}_{\mathcal{X}_c}\}$, where ρ is an upper bound on the rank. Specifically, with binary weights $\{\omega_{i_a i_b i_c}\}$ taking value 0 if $y_{i_a i_b i_c}$ is missing (and 1 otherwise), fitting low-rank tensors is possible using

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i_a, i_b, i_c} \omega_{i_a i_b i_c} [y_{i_a i_b i_c} - f(i_a, i_b, i_c)]^2 + \lambda_* \sum_{r=1}^{\rho} \Big[\|a_r\|_{\mathcal{H}_{\mathcal{X}_a}}^2 + \|b_r\|_{\mathcal{H}_{\mathcal{X}_b}}^2 + \|c_r\|_{\mathcal{H}_{\mathcal{X}_c}}^2 \Big]. \qquad (6)$$

If all kernels are selected as Kronecker deltas, (6) reverts back to (5). The separable structure of the regularization in (6) allows application of Representer's theorem [63], which implies that a_r, b_r, and c_r admit finite dimensional representations given by $a_r(i_a) = \sum_{i_a'=1}^{I_a} \alpha_{r i_a'} \kappa_{\mathcal{X}_a}(i_a, i_a')$, $b_r(i_b) = \sum_{i_b'=1}^{I_b} \beta_{r i_b'} \kappa_{\mathcal{X}_b}(i_b, i_b')$, and $c_r(i_c) = \sum_{i_c'=1}^{I_c} \gamma_{r i_c'} \kappa_{\mathcal{X}_c}(i_c, i_c')$, respectively. Coefficients $\hat{\mathbf{A}} := [\hat{\alpha}_{r i_a'}]$, $\hat{\mathbf{B}} := [\hat{\beta}_{r i_b'}]$, and $\hat{\mathbf{C}} := [\hat{\gamma}_{r i_c'}]$ turn out to be solutions of [cf. (5)]

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) := \arg\min_{\{\mathbf{A},\mathbf{B},\mathbf{C}\}} \sum_{i_c=1}^{I_c} \| \mathcal{P}_{\Omega_{i_c}}(\mathbf{Y}_{i_c} - \mathbf{K}_{\mathcal{X}_a} \mathbf{A}\,\mathrm{diag}(\mathbf{e}_{i_c}^\top \mathbf{K}_{\mathcal{X}_c} \mathbf{C})\,\mathbf{B}^\top \mathbf{K}_{\mathcal{X}_b}) \|_F^2 + \lambda_*\,\mathrm{trace}[\mathbf{A}^\top \mathbf{K}_{\mathcal{X}_a} \mathbf{A} + \mathbf{B}^\top \mathbf{K}_{\mathcal{X}_b} \mathbf{B} + \mathbf{C}^\top \mathbf{K}_{\mathcal{X}_c} \mathbf{C}], \qquad (P5)$$

where $\mathbf{K}_{\mathcal{X}_a} := [\kappa_{\mathcal{X}_a}(i_a, i_a')]$, and likewise for $\mathbf{K}_{\mathcal{X}_b}$ and $\mathbf{K}_{\mathcal{X}_c}$, stand for kernel matrices formed using (cross-)correlations estimated from historical data as detailed in, e.g., [7]. Remarkably, the cost in (P5) is convex w.r.t. any of $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$, whenever the rest of them are held fixed. As such, the low-complexity online accelerated algorithms of the "Algorithms" section carry over to tensors too. Having $\hat{\mathbf{A}}$ available, the estimate $\hat{\alpha}_{r i_a'}$ is obtained, and likewise for $\hat{\beta}_{r i_b'}$ and $\hat{\gamma}_{r i_c'}$. The latter yield the desired predicted values as $\hat{y}_{i_a i_b i_c} := \sum_{r=1}^{\rho} \hat{a}_r(i_a) \hat{b}_r(i_b) \hat{c}_r(i_c) \approx y_{i_a i_b i_c}$.

BIG DATA TASKS

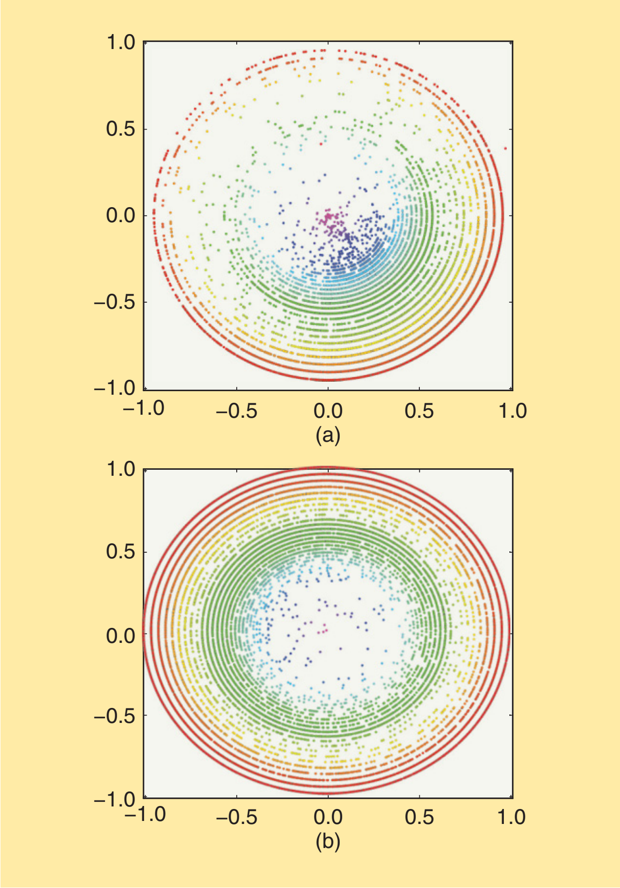
The tools and themes outlined so far will be applied in this section to a sample of big data SP-relevant tasks.

DIMENSIONALITY REDUCTION

NETWORK VISUALIZATION

The rising complexity and volume of networked (graph-valued) data presents new opportunities and challenges for visualization tools that capture global patterns and structural information such as hierarchy, similarity, and communities [3], [27]. Most visualization algorithms tradeoff the clarity of structural characteristics of the underlying data for aesthetic requirements such as minimal edge crossing and fixed internode distance. Although efficient for relatively small networks or graphs (hundreds of nodes), embeddings for larger graphs using these techniques are seldom structurally informative. The growing interest in analysis of big data networks has prioritized the need for effectively capturing structure over aesthetics in visualization. For instance, layouts of metro-transit networks that show hierarchically the bulk of traffic convey a lucid picture about the most critical nodes in the event of a terrorist attack. To this end, [3] captures hierarchy in networks or graphs through well-defined measures of node importance, collectively known as *centrality* in the network science community. Examples are the betweenness centrality, which describes the extent to which information is routed through a specific node by measuring the fraction of all shortest paths traversing it, as well as closeness, eigenvalue, and Markov centrality [3].

THE RISING COMPLEXITY AND VOLUME OF NETWORKED (GRAPH-VALUED) DATA PRESENTS NEW OPPORTUNITIES AND CHALLENGES FOR VISUALIZATION TOOLS THAT CAPTURE GLOBAL PATTERNS AND STRUCTURAL INFORMATION SUCH AS HIERARCHY, SIMILARITY, AND COMMUNITIES.



[FIG3] **The visualization of two snapshots of the large-scale network Gnutella [40] by means of the CC-LLE method. The centrality metric is defined by the node degree. Hence, nodes with low degree are placed far from the center of the embedding. (a) Gnutella-04 (08/04/2012). (b) Gnutella-24 (08/24/2012).**

Consider an undirected graph $\mathscr{G}(\mathscr{V},\mathscr{E})$, where $\mathscr{V}$ denotes the set of vertices (nodes, agents, or processing cores) with cardinality $|\mathscr{V}| = V$, and $\mathscr{E}$ stands for edges (links) that represent pairs of nodes that can communicate. Following (P3), node $\nu \in \mathscr{V}$ communicates with its single- or multihop neighboring peers in $\mathscr{N}_\nu \subset \mathscr{V}$. Given a set of observed feature vectors $\{\mathbf{y}_\nu\}_{\nu\in\mathscr{V}} \subset \mathbb{R}^P$, and a prescribed embedding dimension $p \ll P$ (typically $p \in \{2,3\}$ for visualization), the graph embedding amounts to finding a set of $\{\mathbf{z}_\nu\}_{\nu\in\mathscr{V}} \subset \mathbb{R}^p$ vectors that preserve in the very low-dimensional $\mathbb{R}^p$ the network structure observed via $\{\mathbf{y}_\nu\}_{\nu\in\mathscr{V}}$. The dimensionality reduction module of [3] is based on local linear embedding (LLE) principles [61], which assume that the observed $\{\mathbf{y}_\nu\}_{\nu\in\mathscr{V}}$ live on a low-dimensional, smooth, but unknown manifold, with the objective of seeking an embedding that preserves the local structure of the manifold in the lower dimensional $\mathbb{R}^p$. In particular, LLE accomplishes this by approximating each data point via an affine combination (real weights summing up to 1) of its neighbors, followed by construction of a lower-dimensional embedding that best preserves the weights. If $\mathbf{Y}_\nu := [\mathbf{y}_{\nu'_1}, \ldots, \mathbf{y}_{\nu'_{|\mathscr{N}_\nu|}}] \in \mathbb{R}^{P\times|\mathscr{N}_\nu|}$ gathers all the observed data within the neighborhood of node ν, and along the lines of LLE, the centrality constrained (CC-)LLE method comprises the following two steps:

$$\begin{aligned}
&\mathbf{S1}: \forall \nu \in \mathscr{V}, \mathbf{s}_\nu \in \arg\min_{\mathbf{s}} \|\mathbf{y}_\nu - \mathbf{Y}_\nu \mathbf{s}\|^2 \\
&\qquad\qquad\qquad \text{s. to}\begin{cases}\|\mathbf{Y}_\nu \mathbf{s}\|^2 = h^2(c_\nu)\\ \mathbf{s}^\top \mathbf{1} = 1\end{cases}\\
&\mathbf{S2}: \min_{\{\mathbf{z}_\nu\}_{\nu\in\mathscr{V}}} \sum_{\nu\in\mathscr{V}} \|\mathbf{z}_\nu - \sum_{\nu'\in\mathscr{V}} s_{\nu\nu'} \mathbf{z}_{\nu'}\|^2 \\
&\qquad\qquad\qquad \text{s.to}\, \|\mathbf{z}_\nu\|^2 = h^2(c_\nu), \forall \nu \in \mathscr{V},
\end{aligned} \tag{7}$$

where $\{c_\nu\}_{\nu\in\mathscr{V}} \subset \mathbb{R}$ are centrality metrics, $h(\cdot)$ is a monotone decreasing function that quantifies the centrality hierarchy, e.g., $h(c_\nu) = \exp(-c_\nu)$, and $\mathbf{s}^\top\mathbf{1} = 1$ enforces the local affine approximation of $\mathbf{y}_\nu$ by $\{\mathbf{y}_{\nu'}\}_{\nu'\in\mathscr{N}_\nu}$. In other words, and in the spirit of (P3), $\mathbf{y}_\nu$ is affinely approximated by the "local" dictionary $\mathbf{D}_\nu := \mathbf{Y}_\nu$. It is worth stressing that both objective and constraints in step 1 of (7) can be computed solely by means of the inner-products or correlations $\{\mathbf{Y}_\nu^\top \mathbf{y}_\nu, \mathbf{Y}_\nu^\top \mathbf{Y}_\nu\}_{\nu\in\mathscr{V}}$. Hence, knowledge of $\{\mathbf{y}_\nu\}_{\nu\in\mathscr{V}}$ is not needed in CC-LLE, and only a given set of dissimilarity measures $\{\delta_{\nu\nu'}\}_{(\nu,\nu')\in\mathscr{V}^2}$ suffices to formulate (7), where $\delta_{\nu\nu'} \in \mathbb{R}_{\geq 0}$, $\delta_{\nu\nu'} = \delta_{\nu'\nu}$, and $\delta_{\nu\nu} = 0$, $\forall(\nu,\nu') \in \mathscr{V}^2$; e.g., $\delta_{\nu\nu'} := 1 - |\mathbf{y}_\nu^\top \mathbf{y}_{\nu'}| \|\mathbf{y}_\nu\|^{-1} \|\mathbf{y}_{\nu'}\|^{-1}$ in (7).

After relaxing the nonconvex constraint $\|\mathbf{Y}_\nu \mathbf{s}\|^2 = h^2(c_\nu)$ to the convex $\|\mathbf{Y}_\nu \mathbf{s}\|^2 \leq h^2(c_\nu)$ one, a BCDM approach is followed to solve (7) efficiently, with computational complexity that scales linearly with the network size [3]. Figure 3 depicts the validation of CC-LLE on large-scale degree visualizations of snapshots of the Gnutella peer-to-peer file-sharing network ($|\mathscr{V}| = 26{,}518$,

$|\mathscr{E}| = 65,369$) [40]. Snapshots of this directed network were captured on 4 and 24 August 2002, respectively, with nodes representing hosts. For convenience, undirected renditions of the two networks were obtained by symmetrization of their adjacency matrices. Notice here that the method can generalize to the directed case too, at the price of increased computational complexity. The centrality metric of interest was the node degree, and dissimilarities were computed based on the number of shared neighbors between any pair of hosts. It is clear from Figure 3 that despite the dramatic growth of the network over a span of 20 days, most new nodes had low degree, located thus far from the center of the embedding. The CC-LLE efficiency is manifested by the low running times for obtaining embeddings in Figure 3; 1,684 s for Gnutella-04, and 5,639 s for Gnutella-24 [3].

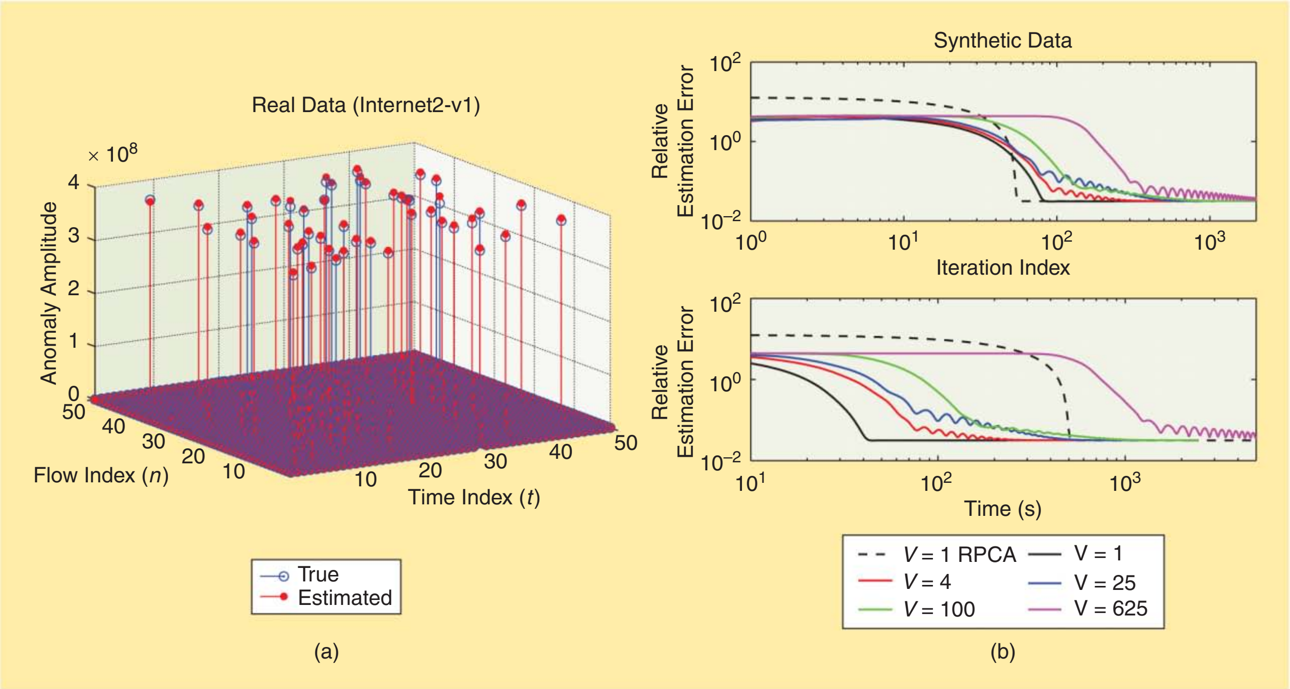
INFERENCE AND IMPUTATION

DECENTRALIZED ESTIMATION OF ANOMALOUS NETWORK TRAFFIC

In the backbone of large-scale networks, origin-to-destination (OD) traffic flows experience abrupt changes that can result in congestion and limit the quality of service provisioning of the end users. These traffic "anomalies" could be due to external sources such as network failures, denial of service attacks, or intruders [38]. Unveiling them is a crucial task in engineering network traffic. This is challenging however, since the available data are high-dimensional noisy link-load measurements, which comprise the superposition of "clean" and anomalous traffic.

Consider as in the section "Dimensionality Reduction" an undirected, connected graph $\mathscr{G}(\mathscr{V},\mathscr{E})$. The traffic $\mathbf{Y} \in \mathbb{R}^{N\times T}$, carried over the edges or links $\mathscr{E}$ ($|\mathscr{E}| = N$) and measured at time instants $t \in \{1,\ldots,T\}$ is modeled as the superposition of unknown "clean" traffic flows $\mathbf{L}_*$, over the time horizon of interest, and the traffic volume anomalies $\mathbf{S}_*$ plus noise $\mathbf{V}$; $\mathbf{Y} = \mathbf{L}_* + \mathbf{S}_* + \mathbf{V}$. Common temporal patterns among the traffic flows in addition to their periodic behavior render most rows (respectively columns) of $\mathbf{L}_*$ linearly dependent, and thus $\mathbf{L}_*$ typically has low rank [38]. Anomalies are expected to occur sporadically over time, and only last for short periods relative to the (possibly long) measurement interval. In addition, only a small fraction of the flows is anomalous at any time slot. This renders matrix $\mathbf{S}_*$ sparse across rows and columns [48].

In the present context, real data including OD flow traffic levels and end-to-end latencies are collected from the operation of the Internet2 network (Internet backbone network across the United States) [30]. OD flow traffic levels were recorded for a three-week operation (sampled per 5 min) of Internet2-v1 during 8–28 December 2003 [38]. To better assess performance, large spikes of amplitude equal to the largest recorded traffic across all flows and time instants were injected into 1% randomly selected entries of the ground-truth matrix $\mathbf{L}_*$. Along the lines of (P3), where the number of links $N = 121$, and $T = 504$, the rows of the data matrix $\mathbf{Y}$ were distributed uniformly over a number of $\mathcal{V} = 11$ nodes. (P3) is solved using ADMM, and a small portion (50×50) of the estimated anomaly matrix $\hat{\mathbf{S}}$ is depicted in Figure 4(a).



[FIG4] Decentralized estimation of network traffic anomalies measured in byte units over 5 min time intervals: (a) only a small portion (50×50) of the sparse matrices $\mathbf{S}_*$ and $\hat{\mathbf{S}}$ entries are shown; (b) relative estimation error versus ADMM iteration index and central processing unit (CPU) time over networks with $\mathcal{V}$ number of nodes. The curve obtained by the centralized R-PCA method [12] is also depicted.

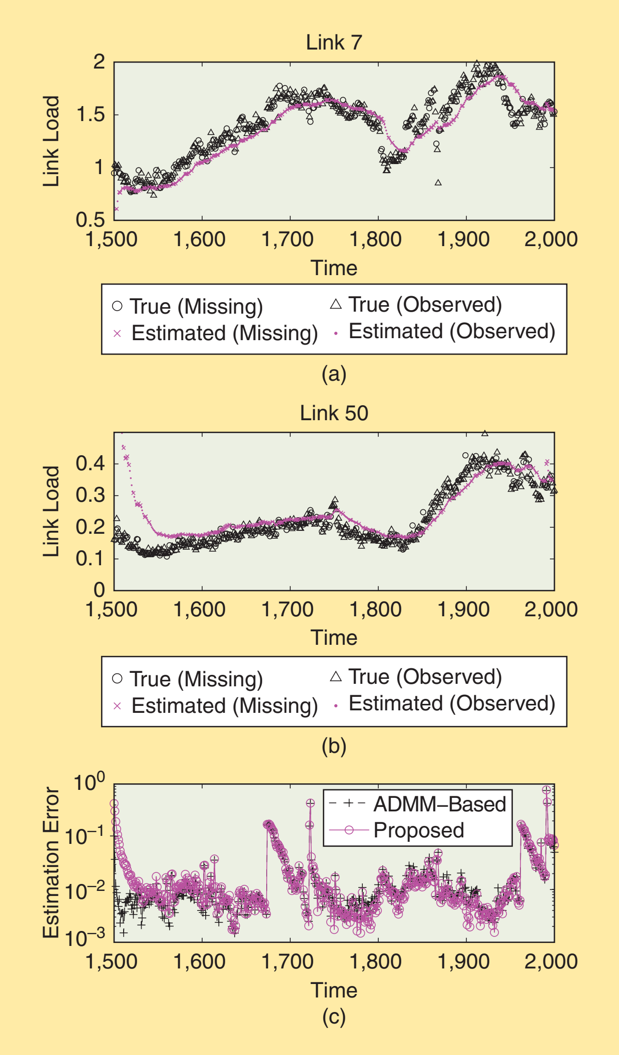
As a means of offering additional design insights, further validation is provided here to reveal the tradeoffs that become relevant as the network size increases. Specifically, comparisons in terms of running time are carried out w.r.t. its centralized counterpart. Throughout, a network modeled as a square grid (uniform lattice) with agents per row/column is adopted. To gauge running times as the network grows, consider a fixed size data matrix $\mathbf{Y} \in \mathbb{R}^{2{,}500 \times 2{,}500}$. The data are synthesized according to the previous model of $\mathbf{Y} = \mathbf{L}_* + \mathbf{S}_* + \mathbf{V}$, details for which can be found in [47, Sec. V]. Rows of $\mathbf{Y}$ are uniformly split among the network nodes. Figure 4(b) illustrates the relative estimation error $\|\hat{\mathbf{S}} - \mathbf{S}_*\|_{\mathrm{F}} / \|\mathbf{S}_*\|_{\mathrm{F}}$ ($\hat{\mathbf{S}}$ stands for the estimate of $\mathbf{S}_*$) versus both iteration index of the ADMM and CPU time over various network sizes.

DYNAMIC LINK LOAD TRAFFIC PREDICTION AND IMPUTATION

Consider again the previous undirected graph $\mathscr{G}(\mathscr{V}, \mathscr{E})$. Connectivity and edge strengths of $\mathscr{G}$ are described by the adjacency matrix $\mathbf{W} \in \mathbb{R}^{\mathcal{V} \times \mathcal{V}}$, where $[\mathbf{W}]_{\nu\nu'} > 0$ if nodes ν and ν' are connected, while $[\mathbf{W}]_{\nu\nu'} = 0$ otherwise. At every $t \in \mathbb{N}_{>0}$, a variable $\chi_{t\nu} \in \mathbb{R}$, which describes a network-wide dynamical process of interest, corresponds to a node $\nu \in \mathscr{V}$. All node variables are collected in $\boldsymbol{\chi}_t := [\chi_{t1}, \ldots, \chi_{tV}]^\top \in \mathbb{R}^V$. A sparse representation of the process over $\mathscr{G}$ models $\boldsymbol{\chi}_t$ as a linear combination of "few" atoms in an $N \times M$ dictionary $\mathbf{D}$, with $M \geq N$; and $\boldsymbol{\chi}_t = \mathbf{D}\mathbf{s}_t$, where $\mathbf{s}_t \in \mathbb{R}^M$ is sparse. Further, only a portion of $\boldsymbol{\chi}_t$ is observed per time slot t. Let now $\boldsymbol{\Omega}_t \in \mathbb{R}^{N' \times N}$, $N' \leq N$, denote a binary measurement matrix, with each row of $\boldsymbol{\Omega}_t$ corresponding to the canonical basis vector for $\mathbb{R}^N$, selecting the measured components of $\mathbf{y}_t \in \mathbb{R}^N$. In other words, the observed data per slot t are $\mathbf{y}_t = \boldsymbol{\Omega}_t \boldsymbol{\chi}_t + \mathbf{v}_t$, where $\mathbf{v}_t$ denotes noise. To impute missing entries of $\boldsymbol{\chi}_t$ in $\mathbf{y}_t$, the topology of $\mathscr{G}$ will be utilized. The spatial correlation of the process is captured by the (unnormalized) graph Laplacian matrix $\boldsymbol{\Lambda} := \operatorname{diag}(\mathbf{W}\mathbf{1}_N) - \mathbf{W}$, where $\mathbf{1}_N \in \mathbb{R}^N$ is the all-ones vector. Following Figure 2 and given a "forgetting factor" $\delta \in (0,1]$, to gradually diminish the effect of past data (and thus account for nonstationarity), define

$$F_t(\mathbf{s}, \mathbf{D}) := \overbrace{\frac{1}{2\Delta_t} \sum_{\tau=1}^{t} \delta^{t-\tau} \|\mathbf{y}_\tau - \boldsymbol{\Omega}_\tau \mathbf{D}\mathbf{s}\|^2 + \frac{\lambda_\Lambda}{2} \mathbf{s}^\top \mathbf{D}^\top \boldsymbol{\Lambda} \mathbf{D}\mathbf{s}}^{f_t(\mathbf{s},\mathbf{D})} + \overbrace{\lambda_1 \|\mathbf{s}\|_1}^{g_1(\mathbf{s})} + \overbrace{\iota_{\mathscr{D}}(\mathbf{D})}^{g_2(\mathbf{D})}, \tag{8}$$

where $\Delta_t := \sum_{\tau=1}^{t} \delta^{t-\tau}$, and $\iota_{\mathscr{D}}$ stands for the indicator function of $\mathscr{D} := \{\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_M] \in \mathbb{R}^{N \times M} \mid \|\mathbf{d}_m\| \leq 1, m \in \{1, \ldots, M\}\}$, i.e., $\iota_{\mathscr{D}}(\mathbf{D}) = 0$ if $\mathbf{D} \in \mathscr{D}$, and $\iota_{\mathscr{D}}(\mathbf{D}) = +\infty$ if $\mathbf{D} \notin \mathscr{D}$ (note that $\forall \gamma > 0$, $\operatorname{Prox}_{\gamma \iota_{\mathscr{D}}}$ is the metric projection onto the closed convex $\mathscr{D}$ [5]). The term including the known $\boldsymbol{\Lambda}$ quantifies the a priori information on the topology of $\mathscr{G}$, and promotes "smooth" solutions over strongly connected nodes of $\mathscr{G}$ [23]. This term is also instrumental for accommodating missing entries in $(\boldsymbol{\chi}_t)_{t \in \mathbb{N}_{>0}}$.



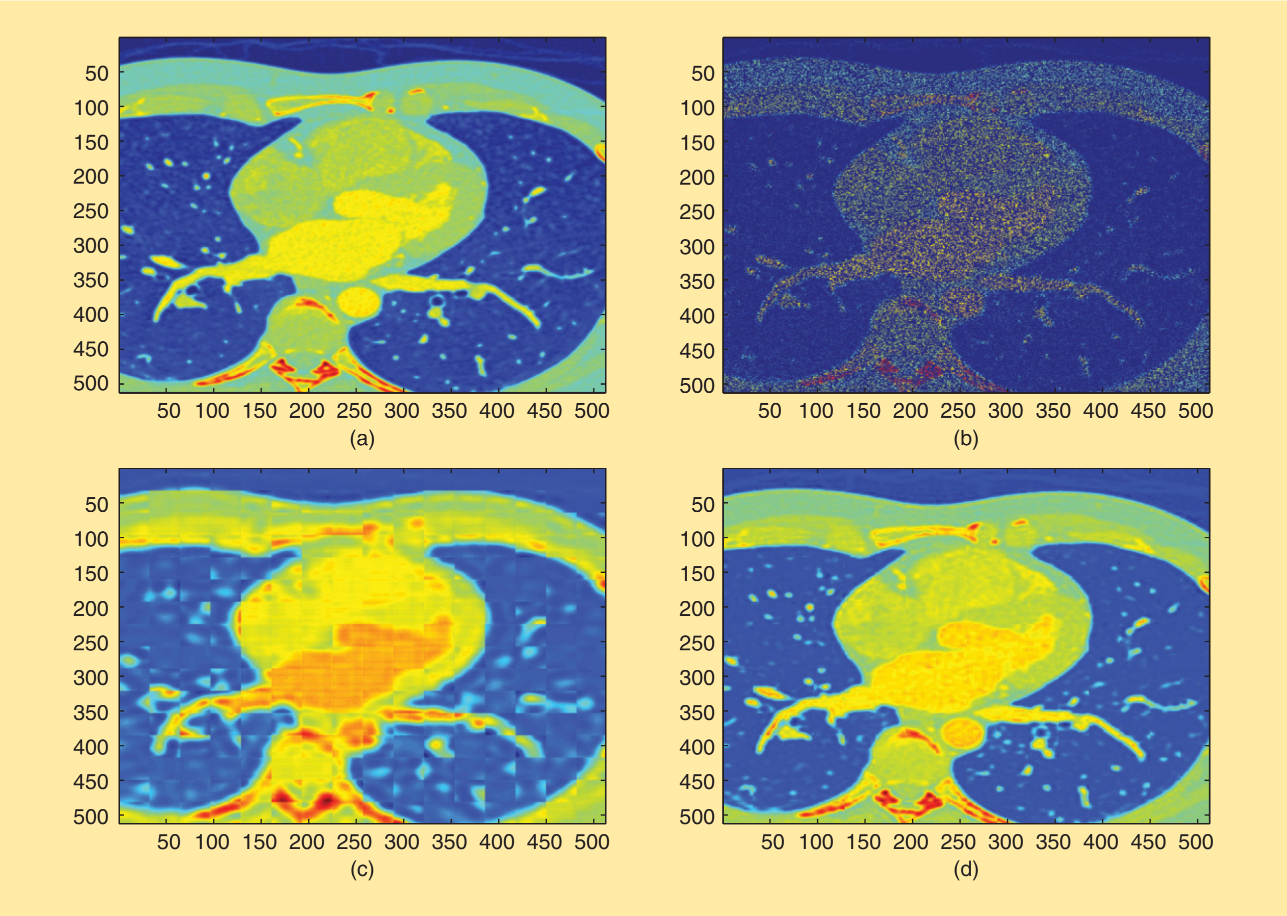
(a)

(b)

(c)

[FIG5] Link load tracking (dots and triangles) and imputation (crosses and circles) on Internet2 [30]. The proposed method is validated versus the ADMM-based approach of [23].

The algorithm of Figure 2 was validated on estimating and tracking network-wide link loads taken from the Internet2 measurement archive [30]. The network consists of $N = 54$ links and nine nodes. Using the network topology and routing information, network-wide link loads $(\boldsymbol{\chi}_t)_{t \in \mathbb{N}_{>0}} \subset \mathbb{R}^N$ become available (in gigabits per second). Per time slot t, only $N' = 30$ of the $\boldsymbol{\chi}_t$ components, chosen randomly via $\boldsymbol{\Omega}_t$, are observed in $\mathbf{y}_t \in \mathbb{R}^{N'}$. Cardinality of the time-varying dictionaries is set to $M = 80$, $\forall t$. To cope with pronounced temporal variations of the Internet2 link loads, the forgetting factor δ in (8) was set equal to 0.5. Figure 5 depicts estimated values of both observed (dots) and missing (crosses) link loads, for a randomly chosen link of the network. The normalized squared estimation error between the true $\boldsymbol{\chi}_t$ and the inferred $\hat{\boldsymbol{\chi}}_t$, specifically $\|\boldsymbol{\chi}_t - \hat{\boldsymbol{\chi}}_t\|^2 \|\boldsymbol{\chi}_t\|^{-2}$, is also plotted in Figure 5 versus time t. The accelerated algorithm was compared with the state-of-the-art scheme in [23] that relies on ADMM, to minimize a cost closely related to (8) w.r.t. $\mathbf{s}$, and uses BCD iterations requiring matrix inversion to optimize (8) w.r.t. $\mathbf{D}$. On the other hand, $R_1 = 1$ and $R_2 = 10$ in the algorithm of Figure 2. It is worth noticing here that ADMM in [23] requires multiple iterations to achieve a prescribed estimation accuracy, and that no matrix inversion



[FIG6] The imputation of missing functional MRI cardiac images by using the PARAFAC tensor model and the online framework of (9). The images were artificially colored to highlight the differences between the obtained recovery results. (a) The original image. (b) The degraded image (75% missing values). (c) The recovered image $(\rho = 10)$ with relative estimation error 0.14. (d) The recovered image $(\rho = 50)$ with relative estimation error 0.046.

was incorporated in the realization of the proposed scheme. Even if the accelerated first-order method operates under lower computational complexity than the ADMM approach, estimation error performance both on observed and missing values is almost identical.

CARDIAC MRI

Cardiac magnetic resonance imaging (MRI) is a major imaging tool for noninvasive diagnosis of heart diseases in clinical practice. However, time limitations posed by the patient's breath-holding time, and thus the need for fast data acquisition degrade the quality of MRI images, resulting often in missing pixel values. In the present context, imputation of the missing pixels utilizes the fact that cardiac MRI images intrinsically contain low-dimensional components.

The FOURDIX data set is considered, which contains 263 cardiac scans with ten steps of the entire cardiac cycle [24]. Each scan is an image of size 512×512 pixels, which is divided into 64 (32×32)-dimensional patches. Placing one after the other, patches form a sequence of slices of a tensor $\underline{\mathrm{Y}} \in \mathbb{R}^{32 \times 32 \times 67{,}328}$. Randomly chosen 75% of the $\underline{\mathrm{Y}}$ entries are dropped to simulate missing data. Operating on such a tensor via batch algorithms is computationally demanding, due to the tensor's size and the computer's memory limitations. Motivated by the batch formulation in (5), a weighted LS online counterpart is [50]

$$\min_{\{\mathrm{A},\mathrm{B},\mathrm{C}\}} \sum_{\tau=1}^{t} \delta^{t-\tau}\Bigg[\| \mathcal{P}_{\Omega}(\mathrm{Y}_{\tau} - \mathrm{A}\,\mathrm{diag}(\mathrm{e}_{\tau}^{\top}\mathrm{C})\,\mathrm{B}^{\top}) \|_{\mathrm{F}}^{2} + \frac{\lambda_{*}}{\sum_{\tau=1}^{t}\delta^{t-\tau}}(\|\mathrm{A}\|_{\mathrm{F}}^{2} + \|\mathrm{B}\|_{\mathrm{F}}^{2}) + \lambda_{*}\|\mathrm{e}_{\tau}^{\top}\mathrm{C}\|^{2}\Bigg], \qquad (9)$$

where $\delta > 0$ is a forgetting factor, and e_{τ} is the τth t-dimensional canonical vector. The third dimension t of $\underline{\mathrm{Y}}$ in (9) indicates the slice number. To solve (9), the variables $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ are sequentially processed; fixing $\{\mathrm{A}, \mathrm{B}\}$, (9) is minimized w.r.t. C, while gradient steepest descent steps are taken w.r.t. each one of A and B, having the other variables held constant. The resultant online learning algorithm is computationally light, with $256\rho^{2}$ operations (on average) per t. The results of its application to a randomly chosen scan image, for different choices of the rank ρ, are depicted in Figure 6 with relative estimation errors, $\|\mathrm{Y}_{\tau} - \hat{\mathrm{Y}}_{\tau}\|_{\mathrm{F}} / \|\mathrm{Y}_{\tau}\|_{\mathrm{F}}$, equal to 0.14 and 0.046 for $\rho = 10$ and 50, respectively.

Additional approaches for batch tensor completion of both visual and spectral data can be found in [41] and [66], whereas the algorithms in [1] and [7] carry out low-rank tensor decompositions from incomplete data and perform imputation as a by-product.

ACKNOWLEDGMENTS

Work in this article was supported by the National Science Foundation grants ECCS 1343248 and Eager 1343860. Moreover, it has been cofinanced by the European Union (European Social Fund and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework-Research Funding Program: Thalis–UoA–Secure Wireless Nonlinear Communications at the Physical Layer. We wish to thank Morteza Mardani and Brian Baingana, from the University of Minnesota, for the fruitful discussions and the numerical tests they provided.

AUTHORS

Konstantinos Slavakis (kslavaki@umn.edu) received his Ph.D. degree from the Tokyo Institute of Technology (TokyoTech), Japan, in 2002. He was a postdoctoral fellow with TokyoTech (2004–2006) and the Department of Informatics and Telecommunications, University of Athens, Greece (2006–2007). He was an assistant professor in the Department of Telecommunications and Informatics, University of Peloponnese, Tripolis, Greece (2007–2012). He is currently a research associate professor with the Department of Electrical and Computer Engineering and Digital Technology Center, University of Minnesota, United States. His current research interests include signal processing, machine learning, and big data analytics problems.

Georgios B. Giannakis (georgios@umn.edu) received his Ph.D. degree from the University of Southern California in 1986. Since 1999, he has been with the University of Minnesota, where he holds the ADC chair in wireless telecommunications in the Department of Electrical and Computer Engineering and serves as director of the Digital Technology Center. His interests are in the areas of communications, networking, and statistical signal processing—subjects on which he has published more than 360 journal and 620 conference papers, 21 book chapters, two edited books, and two research monographs (h-index 108). His current research focuses on sparsity and big data analytics, cognitive networks, renewables, power grid, and social networks. He is the (co) inventor of 22 patents and the (co)recipient of eight best paper awards from the IEEE Communications and Signal Processing Societies. He is a Fellow of the IEEE and EURASIP and has also received technical achievement awards from the IEEE Signal Processing Society and EURASIP.

Gonzalo Mateos (mate0058@umn.edu) received his B.Sc. degree in electrical engineering from Universidad de la Republica, Uruguay, in 2005 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Minnesota, in 2009 and 2012, respectively. Since 2014, he has been an assistant professor with the Department of Electrical and Computer Engineering, University of Rochester. During 2013, he was a visiting scholar with the Computer Science Department, Carnegie Mellon University. From 2003 to 2006, he worked as a systems engineer at ABB, Uruguay. His research interests lie in the areas of statistical learning from big data, network science, wireless communications, and signal processing. His current research focuses on algorithms, analysis, and application of statistical signal processing tools to dynamic network health monitoring, social, power grid, and big data analytics.

REFERENCES

[1] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemomet. Intell. Lab. Syst.*, vol. 106, no. 1, pp. 41–56, 2011.

[2] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[3] B. Baingana and G. B. Giannakis, "Embedding graphs under centrality constraints for network visualization," submitted for publication. arXiv:1401.4408.

[4] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, 2010, pp. 704–711.

[5] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.

[6] J. A. Bazerque and G. B. Giannakis, "Nonparametric basis pursuit via sparse kernel-based learning," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 112–125, July 2013.

[7] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Rank regularization in Bayesian inference for tensor completion and extrapolation," *IEEE Trans. Signal Processing*, vol. 61, no. 22, pp. 5689–5703, Nov. 2013.

[8] T. Bengtsson, P. Bickel, and B. Li, "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems," in *Probability and Statistics: Essays in Honor of David A. Freedman*. Beachwood, OH: IMS, 2008, vol. 2, pp. 316–334.

[9] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1999.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Machine Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[11] E. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 1, pp. 1–37, 2011.

[13] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, June 2009.

[14] V. Chandrasekaran, S. Sanghavi, P. R. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.

[15] Q. Chenlu and N. Vaswani, "Recursive sparse recovery in large but correlated noise," in *Proc. Allerton Conf. Communication, Control, and Computing*, Sept. 2011, pp. 752–759.

[16] Y. Chi, Y. C. Eldar, and R. Calderbank, "PETRELS: Parallel subspace estimation and tracking using recursive least squares from partial observations," *IEEE Trans. Signal Processing*, vol. 61, no. 23, pp. 5947–5959, 2013.

[17] K. L. Clarkson and D. P. Woodruff, "Low rank approximation and regression in input sparsity time," in *Proc. Symp. Theory Computing*, June 1–4, 2013, pp. 81–90. arXiv:1207.6365v4.

[18] K. Cukier. (2010). Data, data everywhere. *The Economist*. [Online]. Available: http://www.economist.com/node/15557443

[19] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Proc. Symp. Operating System Design and Implementation*, San Francisco, CA, 2004, vol. 6, p. 10.

[20] P. Drineas and M. W. Mahoney, "A randomized algorithm for a tensor-based generalization of the SVD," *Linear Algeb. Appl.*, vol. 420, no. 2–3, pp. 553–571, 2007.

[21] J. Feng, H. Xu, and S. Yan, "Online robust PCA via stochastic optimization," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 2013, pp. 404–412.

[22] P. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 11, pp. 1663–1707, May 2010.

[23] P. Forero, K. Rajawat, and G. B. Giannakis, "Prediction of partially observed dynamical processes over networks via dictionary learning," *IEEE Trans. Signal Processing*, to be published.

[24] [Online]. Available: http://www.osirix-viewer.com/datasets/

[25] H. Gao, J. Cai, Z. Shen, and H. Zhao, "Robust principal component analysis-based four-dimensional computed tomography," *Phys. Med. Biol.*, vol. 56, no. 1, pp. 3181–3198, 2011.

[26] G. B. Giannakis, V. Kekatos, N. Gatsis, S. J. Kim, H. Zhu, and B. Wollenberg, "Monitoring and optimization for power grids: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 107–128, Sept. 2013.

[27] L. Harrison and A. Lu, "The future of security visualization: Lessons from network visualization," *IEEE Netw.*, vol. 26, pp. 6–11, Dec. 2012.

[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[29] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, June 2012, pp. 1568–1575.

[30] [Online]. Available: http://www.internet2.edu/observatory/

[31] M. I. Jordan, "On statistics, computation and scalability," *Bernoulli*, vol. 19, no. 4, pp. 1378–1390, 2013.

[32] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, no. 3, pp. 433–481, Mar. 1985.

[33] S.-J. Kim and G. B. Giannakis, "Optimal resource allocation for MIMO ad hoc cognitive radio networks," *IEEE Trans. Info. Theory*, vol. 57, no. 5, pp. 3117–3131, May 2011.

[34] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, "A scalable bootstrap for massive data," *J. Royal Statist. Soc.: Ser. B,* to be published. [Online]. Available: http://dx.doi.org/10.1111/rssb.12050

[35] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York: Springer, 2009.

[36] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[37] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics," *Linear Algeb. Appl.*, vol. 18, no. 2, pp. 95–138, 1977.

[38] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proc. SIGCOMM*, Aug. 2004, pp. 201–206.

[39] D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[40] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Mar. 2007.

[41] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 208–220, Jan. 2013.

[42] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. Hellerstein, "GraphLab: A new framework for parallel machine learning," in *Proc. 26th Conf. Uncertainty in Artificial Intelligence*, Catalina Island: CA, 2010.

[43] Y. Ma, P. Niyogi, G. Sapiro, and R. Vidal, "Dimensionality reduction via subspace and submanifold learning [From the Guest Editors]," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 14–126, Mar. 2011.

[44] L. Mackey, A. Talwalkar, and M. I. Jordan, "Distributed matrix completion and robust factorization," submitted for publication. arXiv:1107.0789v7.

[45] M. W. Mahoney, "Randomized algorithms for matrices and data," *Found. Trends Machine Learn.*, vol. 3, no. 2, pp. 123–224, 2011.

[46] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Machine Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[47] M. Mardani, G. Mateos, and G. B. Giannakis, "Decentralized sparsity-regularized rank minimization: Algorithms and applications," *IEEE Trans. Signal Processing*, vol. 61, no. 11, pp. 5374–5388, Nov. 2013.

[48] M. Mardani, G. Mateos, and G. B. Giannakis, "Dynamic anomalography: Tracking network anomalies via sparsity and low rank," *IEEE J. Sel. Topics Signal Process.*, vol. 8, pp. 50–66, Feb. 2013.

[49] M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Trans. Info. Theory*, vol. 59, no. 8, pp. 5186–5205, Aug. 2013.

[50] M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *IEEE Trans. Signal Processing*, submitted for publication.

[51] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[52] G. Mateos and G. B. Giannakis, "Robust PCA as bilinear decomposition with outlier-sparsity regularization," *IEEE Trans. Signal Processing*, vol. 60, no. 10, pp. 5176–5190, Oct. 2012.

[53] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[54] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$, *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.

[55] Office of Science and Technology Policy. (2012). Big data research and development initiative. *Executive Office of the President.* [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

[56] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[57] E. E. Papalexakis, U. Kang, C. Faloutsos, N. D. Sidiropoulos, and A. Harpale, "Large scale tensor decompositions: Algorithmic developments and applications," *IEEE Data Eng. Bull.*, vol. 36, no. 3, pp. 59–66, Sept. 2013.

[58] H. Raja and W. U. Bajwa, "Cloud K-SVD: Computing data-adaptive representations in the cloud," in *Proc. Allerton Conf. Communication, Control, and Computing*, Oct. 2013, pp. 1474–1481.

[59] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[60] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 327–495, Sept. 1951.

[61] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, Dec. 2003.

[62] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc WSNs with noisy links—Part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Processing*, vol. 56, no. 1, pp. 350–364, Jan. 2008.

[63] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2001.

[64] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Processing*, vol. 62, no. 3, pp. 641–656.

[65] S. Shalev-Shwartz, "Online learning and online convex optimization," *Found. Trends Mach. Learn.*, vol. 4, no. 2, pp. 107–194, 2012.

[66] M. Signoretto, R. V. Plas, B. D. Moor, and J. A. K. Suykens, "Tensor versus matrix completion: A comparison with application to spectral data," *IEEE Signal Process. Lett.*, vol. 18, pp. 403–406, July 2011.

[67] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Trans. Signal Processing*, vol. 58, no. 4, pp. 2121–2130, Apr. 2010.

[68] K. Slavakis and G. B. Giannakis, "Online dictionary learning from big data using accelerated stochastic approximation algorithms," in *Proc. ICASSP,* Florence, Italy, 2014, pp. 16–20.

[69] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[70] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, Dec. 2012.

[71] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," in *Proc. Conf. Int. Society for Music Information Retrieval*, Oct. 2012, pp. 67–72.

[72] N. Srebro and A. Shraibman, "Rank, trace-norm and max-norm," in *Learning Theory*. Berlin/Heidelberg: Germany: Springer, 2005, pp. 545–560.

[73] N. Städler, D. J. Stekhoven, and P. Bühlmann, "Pattern alternating maximization algorithm for missing data in large p small n problems," *J. Mach. Learn. Res.*, to be published. arXiv:1005.0366v3.

[74] J. M. F. ten Berge and N. D. Sidiropoulos, "On uniqueness in CANDECOMP/PARAFAC," *Psychometrika*, vol. 67, no. 3, pp. 399–409, 2002.

[75] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections: A unifying framework for linear and nonlinear classification and regression tasks," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 97–123, Jan. 2011.

[76] G. Tomasi and R. Bro, "PARAFAC and missing values," *Chemom. Intell. Lab. Syst.*, vol. 75, no. 2, pp. 163–180, 2005.

[77] P. Tseng, "Convergence of block coordinate decent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, June 2001.

[78] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

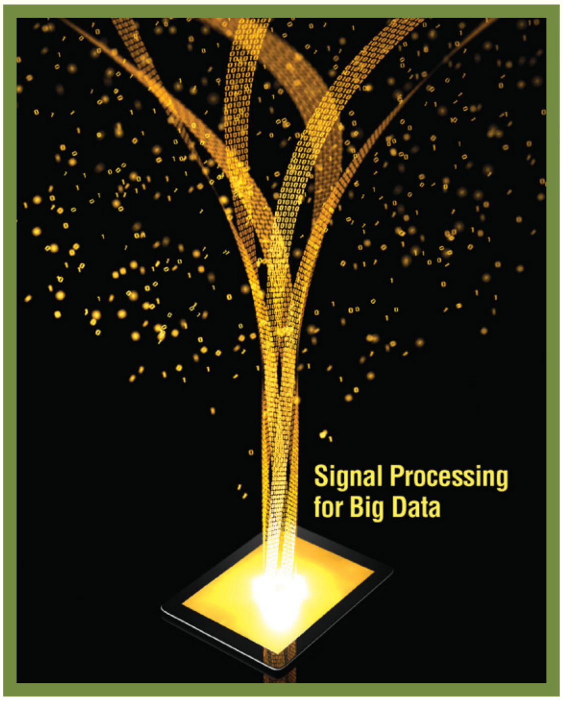
[79] B. Widrow and J. M. E. Hoff, "Adaptive switching circuits," *IRE WESCON Conv. Rec.*, vol. 4, pp. 96–104, Aug. 1960.

[80] M. Yamagishi and I. Yamada, "Over-relaxation of the fast iterative shrinkage-thresholding algorithm with variable stepsize," *Inverse Probl.*, vol. 27, no. 10, p. 105008, 2011.

[SP]

Volkan Cevher, Stephen Becker, and Mark Schmidt

Convex Optimization for Big Data



© ISTOCKPHOTO.COM/TA2YO4NORI

Scalable, randomized, and parallel algorithms for big data analytics

This article reviews recent advances in convex optimization algorithms for big data, which aim to reduce the computational, storage, and communications bottlenecks. We provide an overview of this emerging field, describe contemporary approximation techniques such as first-order methods and randomization for scalability, and survey the important role of parallel and distributed computation. The new big data algorithms are based on surprisingly simple principles and attain staggering accelerations even on classical problems.

Digital Object Identifier 10.1109/MSP.2014.2329397
Date of publication: 19 August 2014

CONVEX OPTIMIZATION IN THE WAKE OF BIG DATA

Convexity in signal processing dates back to the dawn of the field, with problems like least squares (LS) being ubiquitous across nearly all subareas. However, the importance of convex formulations and optimization has increased even more dramatically in the last decade due to the rise of new theory for structured sparsity and rank minimization, and successful statistical learning models such as support vector machines. These formulations are now employed in a wide variety of signal processing applications including compressive sensing, medical imaging, geophysics, and bioinformatics [1]–[4].

There are several important reasons for this explosion of interest, with two of the most obvious being the existence of

1053-5888/14©2014IEEE

efficient algorithms for computing globally optimal solutions and the ability to use convex geometry to prove useful properties about the solution [1], [2]. A unified convex formulation also transfers useful knowledge across different disciplines, such as on sampling and computation, that focus on different aspects of the same underlying mathematical problem [5].

However, the renewed popularity of convex optimization places convex algorithms under tremendous pressure to accommodate increasingly large data sets and to solve problems in unprecedented dimensions. Internet, text, and imaging problems, among a myriad of other examples, no longer produce data sizes from megabytes to gigabytes, but rather from terabytes to exabytes. Despite the progress in parallel and distributed computing, the practical utility of classical algorithms like interior point methods may not go beyond discussing the theoretical tractability of the ensuing optimization problems [3].

In response, convex optimization is reinventing itself for big data, where the data and parameter sizes of optimization problems are too large to process locally, and where even basic linear algebra routines like Cholesky decompositions and matrix–matrix or matrix–vector multiplications that algorithms take for granted are prohibitive. In stark contrast, convex algorithms also no longer need to seek high-accuracy solutions since big data models are necessarily simple or inexact [6].

THE BASICS

We describe the fundamentals of big data optimization via the following composite formulation

$$F^* \overset{\text{def}}{=} \min_x \{F(x) := f(x) + g(x) : x \in \mathbb{R}^p\}, \qquad (1)$$

where f and g are convex functions. We review efficient numerical methods to obtain an optimal solution x^* of (1) as well as required assumptions on f and g. Such composite convex minimization problems naturally arise in signal processing when we estimate unknown parameters $x_0 \in \mathbb{R}^p$ from data $y \in \mathbb{R}^n$. In maximum a posteriori estimation, for instance, we regularize a smooth data likelihood function as captured by f typically with a nonsmooth prior term g that encodes parameter complexity [1].

A basic understanding of big data optimization algorithms for (1) rests on three key pillars:

- *First-order methods*: First-order methods obtain low- or medium-accuracy numerical solutions by using only first-order oracle information from the objective, such as gradient estimates. They can also handle the important nonsmooth variants of (1) by making use of the proximal mapping principle. These methods feature nearly dimension-independent convergence rates, they are theoretically robust to the approximations of their oracles, and they typically rely on computational primitives that are ideal for distributed and parallel computation.
- *Randomization*: Randomization techniques particularly stand out among many other approximation techniques to enhance the scalability of first-order methods since we can control their expected behavior. Key ideas include random partial updates of optimization variables, replacing the deterministic gradient and proximal calculations with cheap statistical estimators, and speeding up basic linear algebra routines via randomization.
- *Parallel and distributed computation*: First-order methods naturally provide a flexible framework to distribute optimization tasks and perform computations in parallel. Surprisingly, we can further augment these methods with approximations for increasing levels of scalability, from idealized synchronous parallel algorithms with centralized communications to enormously scalable asynchronous algorithms with decentralized communications.

The three concepts above complement each other to offer surprising scalability benefits for big data optimization. For instance, randomized first-order methods can exhibit significant acceleration over their deterministic counterparts since they can generate a good quality solution with high probability by inspecting only a negligibly small fraction of the data [3]. Moreover, since the computational primitives of such methods are inherently approximate, we can often obtain near linear speed-ups with a large number of processors [7], [8], which is a difficult feat when exact computation is required.

MOTIVATION FOR FIRST-ORDER METHODS

A main source of big data problems is the ubiquitous linear observation model in many disciplines:

$$y = \Phi x_0 + z, \qquad (2)$$

where x_0 is an unknown parameter, $\Phi \in \mathbb{R}^{n \times p}$ is a known matrix, and $z \in \mathbb{R}^n$ encodes unknown perturbations or noise—modeled typically with zero-mean independent and identically distributed (i.i.d) Gaussian entries with variance σ^2. Linear observations sometimes arise directly from the basic laws of physics as in magnetic resonance imaging and geophysics problems. Other times, (2) is an approximate model for more complicated nonlinear phenomena as in recommender systems and phase retrieval applications.

The linear model (2) along with low-dimensional signal models on x_0, such as sparsity, low total-variation, and low-rankness, has been an area of intense research activity in signal processing. Hence, it is instructive to first study the choice of convex formulations and their scalability implications here. The classical convex formulation in this setting has always been the LS estimator

$$\hat{x}_{\text{LS}} = \operatorname*{argmin}_{x \in \mathbb{R}^p} \left\{ F(x) := \frac{1}{2} \| y - \Phi x \|_2^2 \right\}, \qquad (3)$$

which can be efficiently solved by Krylov subspace methods using only matrix–vector multiplications. An important variant to (3) is the ℓ_1-regularized least absolute shrinkage and selection operator (LASSO), which features the composite form (1)

$$\hat{x}_{\text{LASSO}} = \operatorname*{argmin}_{x \in \mathbb{R}^p} \left\{ F(x) := \frac{1}{2} \| y - \Phi x \|_2^2 + \lambda \| x \|_1 \right\}, \qquad (4)$$

[TABLE 1] A NUMERICAL COMPARISON OF THE DEFAULT FIRST-ORDER METHOD IMPLEMENTED IN TEMPLATES FOR FIRST-ORDER CONIC SOLVERS (TFOCS) (HTTP://CVXR.COM/TFOCS) VERSUS THE INTERIOR POINT METHOD SDPT3 IMPLEMENTED IN CVX (HTTP://CVXR.COM/CVX) FOR THE LASSO PROBLEM (4) WITH $\lambda = 2\sigma\sqrt{2\log p}$. IN THE LINEAR OBSERVATION MODEL (2), THE MATRIX Φ IS A RANDOMLY SUBSAMPLED DCT MATRIX WITH $n = p/2$, THE SIGNAL x_0 HAS $s = p/25$ NONZERO COEFFICIENTS WITH NORM $\|x_0\|_2^2 \approx s$, AND THE NOISE z HAS VARIANCE $\sigma^2 = 10^{-4}$.

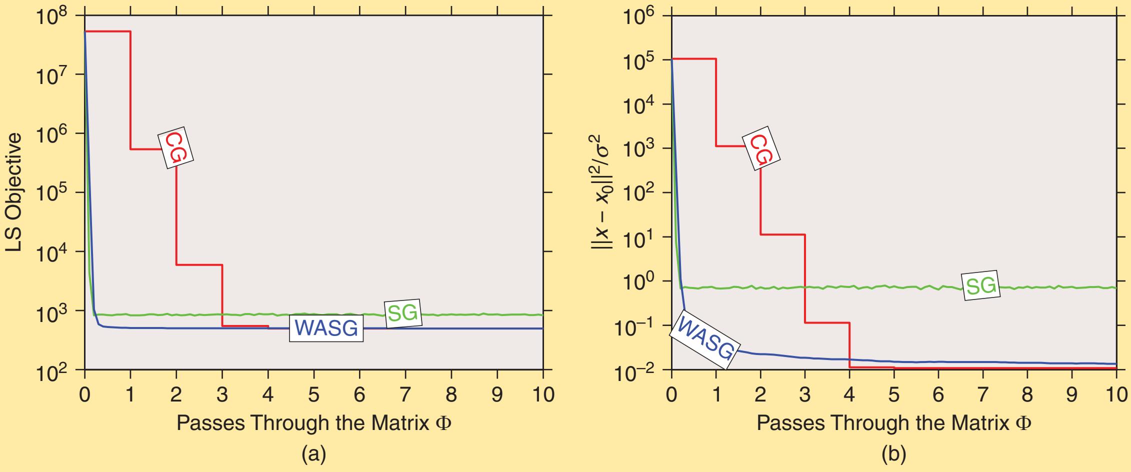
DIMENSION	TIME		ERROR $\|\hat{x} - x_0\|^2/\sigma^2$		ITERATIONS	
	SDPT3	TFOCS	SDPT3	TFOCS	SDPT3	TFOCS
128	0.3 s	0.3 s	1.2	1.2	10	94
512	2.2 s	0.3 s	2.3	2.3	11	121
1,024	16.0 s	0.5 s	2.4	2.4	12	157
2,048	145.0 s	0.7 s	2.8	2.8	12	234
4,096	N/A	1.0 s	N/A	3.3	N/A	281
16,384	N/A	2.9 s	N/A	3.7	N/A	527
131,072	N/A	40.2 s	N/A	4.4	N/A	1,265
1,048,576	N/A	838.5 s	N/A	5.1	N/A	3,440

where λ controls the strength of the regularization. Compared to the LS estimator, the LASSO estimator has the advantage of producing sparse solutions (i.e., $\hat{x}_{\text{LASSO}}$ has mostly zero entries), but its numerical solution is essentially harder since the regularizing term is nonsmooth.

It turns out that the ℓ_1-regularization in sparse signal recovery with the linear model (2) is indeed critical when we are data deficient (i.e., $n < p$). Otherwise, the LASSO formulation imparts only a denoising effect to the solution when $n \geq p$. Theoretical justifications of the LASSO (4) over the LS estimation (3) come from statistical analysis and the convex geometry of (4), and readily apply to many other low-dimensional signal models and their associated composite formulations [1]–[3].

Table 1 illustrates key hallmarks of the first-order methods with the LASSO problem against the classical interior point method: nearly dimension-independent convergence and the ability to exploit implicit linear operators [e.g., the discrete cosine transform (DCT)]. In contrast, interior point methods require much larger space and have near cubic dimension dependence due to the application of dense matrix–matrix multiplications or Cholesky decompositions in finding the Newton-like search directions. Surprisingly, the LASSO formulation possesses additional structures that provably enhance the convergence of the first-order methods [1], making them competitive in accuracy even to the interior point method.

Figure 1 shows that we can scale radically better than even the conjugate gradients (CG) for the LS formulation when $n \gg p$ if we exploit stochastic approximation within first-order methods. We take the simplest optimization method, specifically gradient descent with fixed step-size, and replace its gradient calculations with their cheap statistical estimates (cf. the section “Big Data Scaling via Randomization” for the recipe). The resulting SG algorithm already obtains a strong baseline performance with access to only a fraction of the rows of Φ while the CG method requires many more full accesses.



[FIG1] A numerical comparison of the CG method versus the stochastic gradient (SG) method and the weighted averaged SG (WASG) method for the LS problem (3), showing the objective (a) and the normalized estimation error (b). The matrix Φ has standard normal entries with dimensions $n = 10^5$ and $p = 10^3$. The noise variance is $\sigma^2 = 10^{-2}$ whereas $\|x_0\|_2^2 \approx p$. At a fractional access to the matrix Φ, the stochastic methods obtain a good relative accuracy on the signal estimate. Finally, the SG method has an optimization error due to our step-size choice; cf. the section “Stochastic Gradient Methods” for an explanation.

FIRST-ORDER METHODS FOR SMOOTH AND NONSMOOTH CONVEX OPTIMIZATION

As the LASSO formulation highlights, nonsmooth regularization can play an indispensable role in solution quality. By using the powerful proximal gradient framework, we will see that many of these nonsmooth problems can be solved nearly as efficiently as their smooth counterparts, a point not well understood until the mid-2000s. To this end, this section describes first-order methods within this context, emphasizing specific algorithms with global convergence guarantees. In the sequel, we will assume that the readers have some familiarity with basic notions of convexity and complexity.

SMOOTH OBJECTIVES

We begin our exposition with an important special case of (1), where the objective F only consists of a differentiable convex function f. The elementary first-order technique for this case is the gradient method, which uses only the local gradient $\nabla f(x)$ and iteratively performs the following update:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \tag{5}$$

where k is the iteration count and α_k is an appropriate step-size that ensures convergence.

For smooth minimization, we can certainly use several other faster algorithms such as Newton-like methods. By faster, we mean that these methods require fewer iterations than the gradient method to reach a target accuracy: i.e., $F(x^k) - F^* \leq \varepsilon$. However, we do not focus on these useful methods here since they either require additional information from the function F, more expensive computations, or do not generalize easily to constrained and nonsmooth problems.

Fortunately, the low per-iteration cost of the gradient method can more than make up for its drawbacks in iteration count. For instance, computing the gradient dominates the per-iteration cost of the method, which consists of matrix–vector multiplications with Φ and its adjoint Φ^T when applied to the LS problem (3). Hence, we can indeed perform many gradient iterations for the cost of a single iteration of more complicated methods, potentially taking a shorter time to reach the same level of accuracy ε.

Surprisingly, by making simple assumptions about f, we can rigorously analyze how many iterations the gradient method will in fact need to reach an ε-accurate solution. A common assumption in the analysis that holds in many applications is that the gradient of f is Lipschitz continuous, meaning that

$$\forall x, y \in \mathbb{R}^p, \| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2,$$

for some constant L. When f is twice-differentiable, a sufficient condition is the eigenvalues of its Hessian $\nabla^2 f(x)$ are bounded above by L. Hence, we can trivially estimate $L = \| \Phi \|_2^2$ for (3).

If we simply set the step-size $\alpha_k = 1/L$ or alternatively use a value that decreases f the most, then the iterates of the gradient method for any convex f with a Lipschitz-continuous gradient obey

$$f(x^k) - f^* \leq \frac{2L}{k+4} d_0^2, \tag{6}$$

where $d_0 = \| x^0 - x^* \|_2$ is the distance of the initial iterate x^0 to an optimal solution x^* [9, Cor. 2.1.2]. Hence, the gradient method needs $O(1/\varepsilon)$-iterations for an ε-accurate solution in the worst case.

Unfortunately, this convergence rate does not attain the known complexity lower-bound

$$f(x^k) - f^* \geq \frac{3Ld_0^2}{32\,(k+1)^2},$$

which holds for all functions f with a Lipschitz-continuous gradient. That is, in the worst case any iterative method based only on function and gradient evaluations cannot hope for a better accuracy than $\Omega(1/k^2)$ at iteration k for $k < p$ [9]. Amazingly, a minor modification by Nesterov achieves this optimal convergence by the simple step-size choice $\alpha_k = 1/L$ and an extra-momentum step with a parameter $\beta_k = k/(k+3)$ [9]:

Algorithm 1: Nesterov's accelerated gradient method for unconstrained minimization $(v^0 = x^0)$ [9].

```
1) $x^{k+1} = v^k - \alpha_k \nabla f(v^k)$
2) $v^{k+1} = x^{k+1} + \beta_k (x^{k+1} - x^k)$
```

The accelerated gradient method in Algorithm 1 achieves the best possible worst-case error rate, and hence, it is typically referred to as an *optimal* first-order method.

Many functions also feature additional structures useful for numerical optimization. Among them, strong convexity deserves special attention since this structure provably offers key benefits such as the existence of a unique minimizer and improved optimization efficiency. A function f is called strongly convex if the function $x \mapsto f(x) - \mu/2 \| x \|_2^2$ is convex for some positive value μ. Perhaps not so obvious is the fact that even nonsmooth functions can have strong convexity (i.e., $f(x) = \| x \|_1 + \mu/2 \| x \|_2^2$).

Indeed, we can transform any convex problem into a strongly convex problem by simply adding a squared ℓ_2-regularization term. For instance, when we have $n < p$ in (3), then the classic Tikhonov regularization results in a strongly convex objective with $\mu = \lambda$:

$$\hat{x}_{\text{ridge}} = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ F(x) := \frac{1}{2} \| y - \Phi x \|_2^2 + \frac{\lambda}{2} \| x \|_2^2 \right\}.$$

The solution above is known as the ridge estimator and offers statistical benefits [1]. When f is twice-differentiable, a sufficient condition for strong convexity is that the eigenvalues of its Hessian $\nabla^2 f(x)$ are bounded below by μ for all x. For the LS problem (3), strong convexity simply requires Φ to have independent columns.

For strongly convex problems with Lipschitz gradient, such as the ridge estimator, the gradient method geometrically converges to the unique minimizer when the step-size is chosen as $\alpha_k = 1/L$:

$$\| x^k - x^* \|_2 \leq \left(1 - \frac{\mu}{L} \right)^k \| x^0 - x^* \|_2. \tag{7}$$

[TABLE 2] THE TOTAL NUMBER OF ITERATIONS TO REACH ε-ACCURATE SOLUTIONS FOR FIRST-ORDER OPTIMIZATION METHODS. L AND μ DENOTE THE LIPSCHITZ AND STRONG CONVEXITY CONSTANTS AND $d_0 = \| x^0 - x^* \|_2$.

ALGORITHM	CONVEX	STRONGLY CONVEX
[PROXIMAL]-GRADIENT	$O(Ld_0^2/\varepsilon)$	$O\left(\frac{L}{\mu}\log(d_0^2/\varepsilon)\right)$
ACCELERATED-[PROXIMAL]-GRADIENT	$O(\sqrt{Ld_0^2/\varepsilon})$	$O\left(\sqrt{\frac{L}{\mu}}\log(d_0^2/\varepsilon)\right)$

This convergence improves slightly when we instead use $\alpha_k = 2/(\mu + L)$ [9]. Beside the obvious convergence rate difference, we highlight a subtlety between (6) and (7): guarantees due to the Lipschitz assumption such as (6) does not necessarily imply convergence in iterates x^k, while for strongly convex functions we obtain guarantees on the convergence of both $f(x^k)$ and x^k.

It turns out that the accelerated-gradient method can also benefit from strong convexity with an appropriate choice of the momentum term β_k. For example, if we set $\beta_k = (L - \mu)/(L + \mu)$, the accelerated gradient method obtains a near-optimal convergence rate given its assumptions [9, Th. 2.2.3]. In contrast, the gradient method automatically exploits strong convexity without any knowledge of μ.

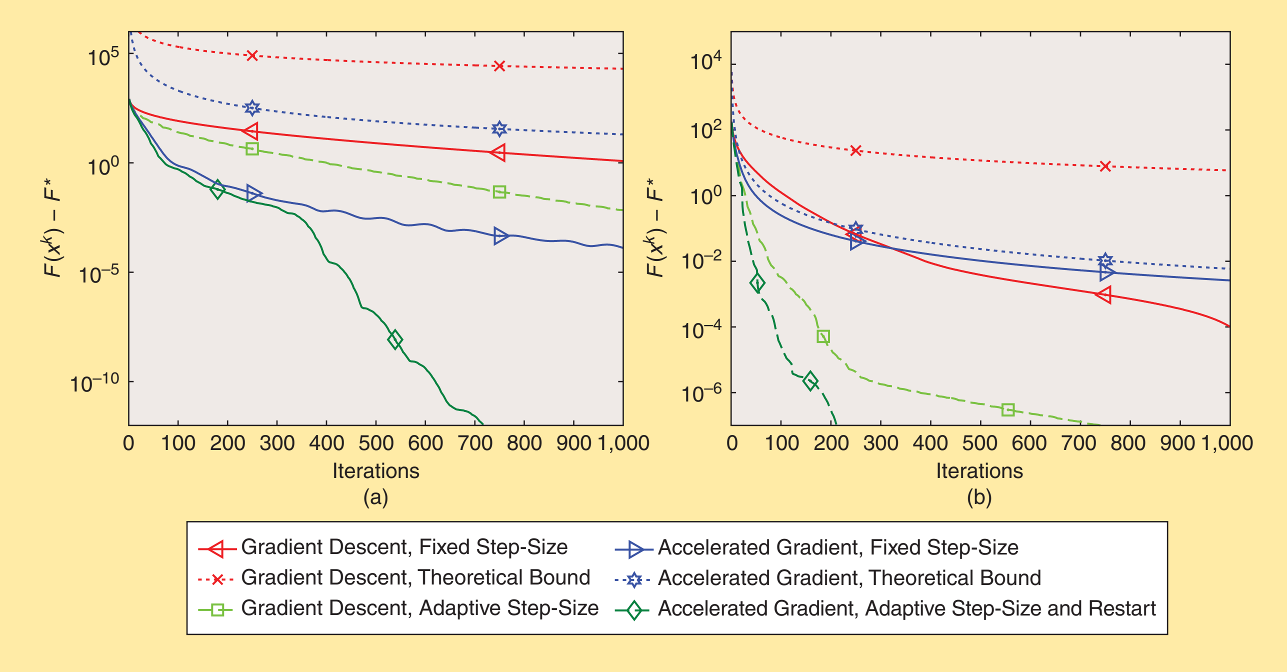
Table 2 summarizes the number of iterations to reach an accuracy of ϵ for the different configurations discussed in this section. Note, however, that there are numerous practical enhancements, such as step-size selection rules for α_k and adaptive restart of the momentum parameter β_k [10] that add only a small computational cost and do not rely on knowledge of the Lipschitz constant L or the strong-convexity parameter μ. While such tricks of the trade do not rigorously improve the worst-case convergence rates, they often lead to superior empirical convergence (cf. Figure 2) and similarly apply to their important proximal counterparts for solving (1) that we discuss next.

Finally, the fast gradient algorithms described here also apply to nonsmooth minimization problems using Nesterov's smoothing technique [9]. In addition, rather than assuming Lipschitz-continuity of the gradient of the objective function, recent work has considered efficient gradient methods for smooth self-concordant functions, which naturally emerge in Poisson imaging, graph learning, and quantum tomography problems [11].

COMPOSITE OBJECTIVES

We now consider the canonical composite problem (1), where the objective F consists of a differentiable convex function f and a nonsmooth convex function g as in (4).

In general, the nondifferentiability of g seems to substantially reduce the efficiency of first-order methods. This was



[FIG2] The performance of first-order methods can improve significantly with practical enhancements. We demonstrate how the objective $F(x^k)$ progresses as a function of iterations k for solving (a) the LS formulation (3) and (b) the LASSO formulation (4), both with $p = 5{,}000$ and $n = 2{,}500$, for four methods: (proximal)-gradient descent with fixed step-size $\alpha = 1/L$ and adaptive step-size, accelerated (proximal)-gradient descent with fixed step-size, and accelerated (proximal)-gradient descent with the adaptive step-size and restart scheme in TFOCS. For the LS formulation, the basic methods behave qualitatively the same as their theoretical upper-bounds predict but dramatically improve with the enhancements. In the LASSO formulation, gradient descent automatically benefits from sparsity of the solution and actually outperforms the basic accelerated method in high-accuracy regime, but adding the adaptive restart enhancement allows the accelerated method to also benefit from sparsity.

indeed the conventional wisdom since generic nonsmooth optimization methods, such as subgradient and bundle methods, require $\mathcal{O}(1/\varepsilon^2)$ iterations to reach ε-accurate solutions [9]. While strong convexity helps to improve this rate to $\mathcal{O}(1/\varepsilon)$, the resulting rates are slower than first-order methods for a smooth objective.

Fortunately, composite objectives are far from generic nonsmooth convex optimization problems. The proximal-gradient methods specifically take advantage of the composite structure to retain the same convergence rates of the gradient method for the smooth problem classes in Table 2 [12]. It becomes apparent that these algorithms are in fact natural extensions of the gradient method when we view the gradient method's iterations (5) as an optimization problem:

$$x^{k+1} = \underset{y \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ f(x^k) + \nabla f(x^k)^T (y - x^k) + \frac{1}{2\alpha_k} \| y - x^k \|^2 \right\}, \tag{8}$$

which is based on a simple local quadratic approximation of f. Note that when $\alpha_k \le 1/L$, the objective function above is a quadratic upper bound on f. Proximal-gradient methods use the same approximation of f, but simply include the nonsmooth term g in an explicit fashion:

$$\begin{aligned} x^{k+1} = \underset{y \in \mathbb{R}^p}{\operatorname{argmin}} & \left\{ f(x^k) + \nabla f(x^k)^T (y - x^k) \right. \\ & \left. + \frac{1}{2\alpha_k} \| y - x^k \|^2 + g(y) \right\}. \end{aligned} \tag{9}$$

For $\alpha_k \le 1/L$, the objective is an upper bound on F in (1).

The optimization problem (9) is the update rule of the proximal-gradient method:

$$x^{k+1} = \operatorname{prox}_{\alpha_k g}(x^k - \alpha_k \nabla f(x^k)),$$

where the proximal map or proximal operator is defined as

$$\operatorname{prox}_g(y) \overset{\text{def}}{=} \underset{x}{\operatorname{argmin}} \left\{ g(x) + \frac{1}{2} \| x - y \|_2^2 \right\}. \tag{10}$$

The accelerated proximal-gradient method is defined analogously:

Algorithm 2: Accelerated proximal gradient method to solve (1) [9], and [13] $v^0 = x^0$.

```
1) $x^{k+1} = \operatorname{prox}_{\alpha_k g}(v^k - \alpha_k \nabla f(v^k))$
2) $v^{k+1} = x^{k+1} + \beta_k (x^{k+1} - x^k)$
```

An interesting special case of the proximal-gradient algorithm arises if we consider the indicator function on a convex set C, which is an elegant way of incorporating constraints into (1)

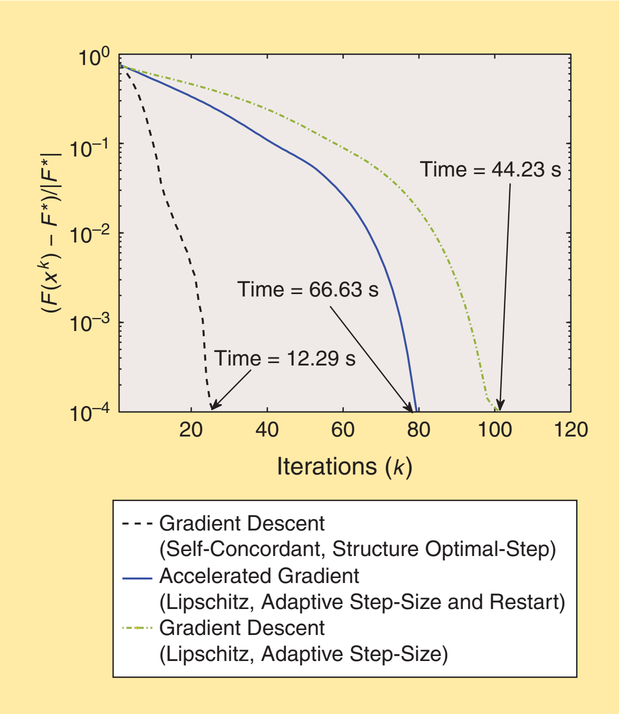
$$g(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}.$$

Then, the proximal-gradient method yields the classic projected-gradient method for constrained optimization.

These methods' fast convergence rates can also be preserved under approximate proximal maps [12]. Proximal operators offer a flexible computational framework to incorporate a rich set of signal priors in optimization. For instance, we can often represent a signal x_0 as a linear combination of atoms $a \in \mathcal{A}$ from some *atomic set* $\mathcal{A} \subseteq \mathbb{R}^p$ as $x_0 = \sum_{a \in \mathcal{A}} c_a a$, where c_a are the representation coefficients. Examples of atomic sets include structured sparse vectors, sign-vectors, low-rank matrices, and many more. The geometry of these sets can facilitate perfect recovery even from underdetermined cases of the linear observations (2) with sharp sample complexity characterizations [2].

To promote the structure of the set $\mathcal{A}$ in convex optimization, we can readily exploit its gauge function: $g_{\mathcal{A}}(x) \overset{\text{def}}{=} \inf\{\rho > 0 \mid x \in \rho \cdot \overline{\operatorname{conv}}(\mathcal{A})\}$, where $\overline{\operatorname{conv}}(\mathcal{A})$ is the convex hull of the set $\mathcal{A}$. The corresponding proximal operator of the gauge function has the following form:

$$\operatorname{prox}_{\gamma g_{\mathcal{A}}}(u) = u - \underset{v \in \mathbb{R}^d}{\operatorname{argmin}} \{ \| u - v \|_2^2 : \langle a, v \rangle \le \gamma, \forall a \in \mathcal{A} \}, \tag{11}$$



[FIG 3] Choosing the correct smoothness structure on f in composite minimization is key to the numerical efficiency of the first-order methods. The convergence plot here simply demonstrates this with a composite objective, called the heteroskedastic LASSO (hLASSO) from [11], where $n = 1.5 \times 10^4$ and $p = 5 \times 10^4$. hLASSO features a self-concordant but not Lipschitz gradient smooth part f, and obtains sparse solutions while simultaneously estimating the unknown noise variance σ^2 in the linear model (2). The simplest first-order method, when matched to correct self-concordant smoothness structure, can calculate its step-sizes optimally, and hence significantly outperforms the standard first-order methods based on the Lipschitz gradient assumption even with the enhancements we discussed. Surprisingly, the accelerated gradient method takes longer than the gradient method to reach the same accuracy since it relies heavily on the Lispchitz gradient assumption in its momentum steps, which lead to costlier step-size adaptation.

which involves a quadratic program in general but can be explicitly calculated in many cases [2]. Intriguingly, (11) also has mathematical connections to discrete submodular minimization.

By and large, whenever the computation of the proximal map is efficient, so are the proximal-gradient algorithms. For instance, when $g(x) = \lambda \| x \|_1$ as in the LASSO formulation (4), the proximal operator is the efficient soft thresholding operator. Against intuition, a set with an infinite number of atoms can admit an efficient proximal map, such as the set of rank-1 matrices with unit Frobenius norm whose proximal map is given by singular value thresholding. On the other hand, a set with a finite number of atoms need not, such as rank-1 matrices with ± 1 entries whose proximal operator is intractable. Numerous other examples exist [2], [4].

When g represents the indicator function of a compact set, the Frank–Wolfe method solves (9) without the quadratic term and can achieve an $O(1/\epsilon)$ convergence rate in the convex case [19]. This linear subproblem may be easier to solve, and each iteration of the Frank–Wolfe method typically only modifies a single element of the atomic set leading to iterations with a controlled degree of sparsity.

Finally, proximal-gradient methods that optimally exploit the self-concordance properties of f are also explored in [11] (c.f., Figure 3 for an example). Interestingly, many self-concordant functions themselves have tractable proximity operators as well—a fact that proves useful next.

PROXIMAL OBJECTIVES

For many applications, the first-order methods we have covered so far are not directly applicable. As a result, we will find it useful here to view the composite form (1) in the following guise:

$$\min_{x,z \in \mathbb{R}^p} \{F(x, z) := h(x) + g(z) : \Phi z = x\}, \qquad (12)$$

and only posit that the proximity operators of h and g are both efficient. For instance, the LASSO problem (4) and many of its generalizations can be written in this fashion via the convex (splitting) technique [4].

This seemingly innocuous reformulation can simultaneously enhance our modeling and computational capabilities. First, (12) can address nonsmooth and non-Lipschitz objective functions that commonly occur in many applications [11], [14], such as robust principal component analysis (RPCA), graph learning, and Poisson imaging, in addition to the composite objectives we have covered so far. Second, we can apply a simple algorithm, called the alternating direction method of multipliers (ADMM) for its solutions, which leverages powerful augmented Lagrangian and dual decomposition techniques [4], [15]:

```
Algorithm 3: ADMM to solve (12); $\gamma > 0$, $z^0 = u^0 = 0$.
  1) $x^{k+1} = \mathrm{argmin}_x\, \gamma h(x) + \frac{1}{2}\| x - \Phi z^k + u^k \|_2^2$
             $= \mathrm{prox}_{\gamma h}(\Phi z^k - u^k)$
  2) $z^{k+1} = \mathrm{argmin}_z\, \gamma g(z) + \frac{1}{2}\| x^{k+1} - \Phi z + u^k \|_2^2$
  3) $u^{k+1} = u^k + x^{k+1} - \Phi z^{k+1}$
```

Algorithm (3) is well suited for distributed optimization and turns out to be equivalent or closely related to many other algorithms, such as Douglas–Rachford splitting and Bregman iterative algorithms [15]. ADMM requires a penalty parameter γ as input and produces a sequence of iterates that approach feasibility and produce the optimal objective value in the limit. An overview of ADMM, its convergence, enhancements, parameter selection, and stopping criteria can be found in [15].

We highlight two caveats for ADMM. First, we have to numerically solve step 2 in Algorithm 3 in general except when $\Phi^T \Phi$ is efficiently diagonalizable. Fortunately, many notable applications support these features, such as matrix completion where Φ models subsampled matrix entries, image deblurring where Φ is a convolution operator, and total variation regularization where Φ is a differential operator with periodic boundary conditions. Second, the naïve extension of ADMM to problems with more than two objective terms no longer has convergence guarantees.

```
Algorithm 4: Primal-dual hybrid gradient algorithm to solve (12); $\gamma > 0$ and $\tau \le 1/\| \Phi \|^2$.
  1) $x^{k+1} = \mathrm{prox}_{\gamma h}(\Phi z^k - u^k)$
  2) $z^{k+1} = \mathrm{prox}_{\gamma\tau g}(z^k + \tau \Phi^T (x^{k+1} - \Phi z^k + u^k))$
  3) $u^{k+1} = u^k + x^{k+1} - \Phi z^{k+1}$
```

Several solutions address the two drawbacks above. For the former, we can update z^{k+1} inexactly by using a single step of the proximal gradient method, which leads to the method shown in Algorithm 4 which was motivated in [16] as a preconditioned variant of ADMM and then analyzed in [17] in a more general framework. Interestingly, when $h(x)$ has a difficult proximal operator in Algorithm 3 but also has a Lipschitz gradient, we can replace $h(x)$ in step 1 with its quadratic surrogate as in (8) to obtain the linearized ADMM [18]. Surprisingly, these inexact update steps can be as fast to converge as the full ADMM in certain applications [16]. We refer the readers to [15], [17], and [18] for the parameter selection of these variants as well as their convergence and generalizations.

For the issue regarding objectives with more than two terms, we can use dual decomposition techniques to treat the multiple terms in the objective of (12) as separate problems and simultaneously solve them in parallel. We defer this to the section "The Role of Parallel and Distributed Computation" and Algorithm 8.

BIG DATA SCALING VIA RANDOMIZATION

In theory, first-order methods are well-positioned to address very large-scale problems. In practice, however, the exact numerical computations demanded by their iterations can make even these simple methods infeasible as the problem dimensions grows. Fortunately, it turns out that first-order methods are quite robust to using approximations of their optimization primitives, such as gradient and proximal calculations [12]. This section describes emerging randomized approximations that increase the reach of first-order methods to extraordinary scales.

To deliver an example of the key ideas, we will focus only on the smooth and strongly convex F as objectives and point out extensions when possible. Many notable big data problems indeed satisfy this assumption. For instance, Google's PageRank problem measures the importance of nodes in a given graph via its incidence matrix $M \in \mathbb{R}^{p \times p}$ and p is on the order of tens of billions. Assuming that more important nodes have more connections, the problem in fact aims to find the top singular vector of the stochastic matrix $\Phi = M\,\mathrm{diag}(M^T \mathbf{1}_p)^{-1}$, where $\mathbf{1}_p \in \mathbb{R}^p$ is the vector of all 1s.

The PageRank algorithm simply solves this basic linear algebra problem (i.e., find $x^* \geq 0$ such that $\Phi x^* = x^*$ and $\mathbf{1}_p^T x^* = 1$) with the power method. However, we can well approximate this goal using a least squares problem when we relax the constraints with a penalty parameter $\gamma > 0$ [20]:

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) \overset{\text{def}}{=} \frac{1}{2} \| x - \Phi x \|_2^2 + \frac{\gamma}{2} (\mathbf{1}_p^T x - 1)^2 \right\}. \qquad (13)$$

Note that we can work with a constrained version of this problem that includes positivity, but since the PageRank formulation itself is not exact model of reality, the simpler problem can be preferable for obvious computational reasons. Clearly, we would like to minimize the number of operations involving the matrix Φ in any solution method.

COORDINATE DESCENT METHODS

Calculating the full gradient for the PageRank problem formulation requires a matrix–vector operation at each iteration. A cheaper vector-only operation would be to pick a coordinate i of x and only modify the corresponding variable x_i to improve the objective function. This idea captures the essence of coordinate descent methods, which have a long history in optimization [21] and are related to classic methods like the Gauss–Seidel cyclic reduction strategy for solving linear systems. The general form of coordinate descent methods is illustrated in Algorithm 5, where e_i is the ith canonical coordinate vector and $\nabla_i F(\cdot)$ is the ith coordinate of the gradient.

Algorithm 5: Coordinate descent to minimize F over $\mathbb{R}^p$.

```
1) Choose an index $i_k \in \{1, 2, \ldots, p\}$ (see the main text for possible selection schemes)
2) $x^{k+1} = x^k - \alpha \nabla_{i_k} F(x^k) e_{i_k}$
```

The key design consideration across all coordinate descent methods is the choice of the coordinate i at each iteration. A simple strategy amenable to analysis is to greedily pick the coordinate with the largest directional derivative $\nabla_i F$. This selection with $\alpha = 1/L_{\max}$ or optimizing the variable exactly leads to a convergence rate of

$$F(x^k) - F(x^*) \leq \left(1 - \frac{\mu}{pL_{\max}}\right)^k (F(x^0) - F(x^*)), \qquad (14)$$

where $L_{\max} := \max_i L_i$ is the maximum across the Lipschitz constants of $\nabla_i F(x)$ [20]. This configuration indeed seeks the best reduction in the objective per iteration we can hope for under this setting.

The example above underlines the fundamental difficulty in coordinate descent methods. Finding the best coordinate to update, the maximum of the gradient element's magnitudes, can require a computational effort as high as the gradient calculation itself. However, the incurred cost is not justified since the method's convergence is provably slower than the gradient method due to the basic relationship $L_i \leq L \leq pL_i$. An alternative proposal is to cycle through all coordinates sequentially. This is the cheapest coordinate selection strategy for which we can hope but it results in a substantially slower convergence rate.

Surprisingly, randomization of the coordinate choice can achieve the best of both worlds. Suppose we choose the coordinate i uniformly at random among the set $\{1, 2, \ldots, p\}$. This selection can be done with a cost independent of p, but surprisingly nevertheless achieves the same convergence rate (14) in expectation [20]. The randomized algorithm's variance around its expected performance is well-controlled.

We also highlight two salient features of coordinate descent methods. First, they are perhaps most useful for objectives of the form $F(Ax)$ with $A \in \mathbb{R}^{n \times p}$, where evaluating the (not necessarily smooth) F costs $O(n)$. By tracking the product Ax^k with incremental updates, we can then perform coordinate descent updates in linear time. Second, if we importance sample the coordinates proportional to their Lipschitz constants L_i, then the convergence rate of the randomized method improves to

$$F(x^k) - F(x^*) \leq \left(1 - \frac{\mu}{pL_{\text{mean}}}\right)^k (F(x^0) - F(x^*)), \qquad (15)$$

where L_{mean} is the mean across the L_i. Hence, this nonuniform random sampling strategy improves the speed by only adding an $O(\log(p))$ importance sampling cost to the algorithm [20].

Finally, accelerated and composite versions of coordinate descent methods have also recently been explored, although accelerated methods often do not preserve the cheap iteration cost of the nonaccelerated versions [3]: cf. [20] for a numerical example of these methods on the PageRank problem (13).

STOCHASTIC GRADIENT METHODS

In contrast to randomized coordinate descent methods, which update a single coordinate at a time with its exact gradient, SG methods update all coordinates simultaneously but use approximate gradients. They are best suited for minimizing decomposable objective functions

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := \frac{1}{n} \sum_{j=1}^{n} F_j(x) \right\}, \qquad (16)$$

where each F_j measures the data misfit for a single data point. This includes models as simple as least squares and also more elaborate models like conditional random fields.

Algorithm 6: SG descent to minimize F over $\mathbb{R}^p$.

```
1) Choose an index $j_k \in \{1, 2, \ldots, n\}$ uniformly at random
2) $x^{k+1} = x^k - \alpha_k \nabla F_{j_k}(x^k)$
```

SG iterations heavily rely on the decomposability of (16) as shown in Algorithm 6. Similar to the coordinate descent methods, the crucial design problem in the SG methods is the selection of the data points j at each iteration. Analogously, we obtain better convergence rates by choosing the j uniformly at random rather than cycling through the data [22]. In contrast, the per iteration cost of the algorithm now depends only on p but not n.

Interestingly, a random data point selection results in an unbiased gradient estimate when we view (16) as the empirical observation of an expected risk function that governs the optimization problem

$$\min_{x\in\mathbb{R}^p}\{F(x) := \mathbb{E}_{\xi}[F_{\xi}(x)]\}, \tag{17}$$

where the expectation is taken over the sampling distribution for the indices ξ. Indeed, if we can sample from the true underlying distribution, then SG methods directly optimize the expected risk minimization problem and result in provable generalization capabilities in machine learning applications [6]. The unbiased gradient estimation idea enables the SG descent method to handle convex objectives beyond the decomposable form (16) [3].

The general SG method has classically used a decreasing sequence of step-sizes $\{\alpha_k\}$. However, this unfortunately leads to the same slow $O(1/\sqrt{\epsilon})$ and $O(1/\epsilon)$ convergence rates of the subgradient method. But interestingly, if we still use a constant step-size at each iteration for the SG method, the algorithm is known to quickly reduce the initial error, even if it has a nonvanishing optimization error [22]. We have observed this for the SG descent example in Figure 1. Indeed, while SG descent methods have historically been notoriously hard to tune, recent results show that using large step-sizes and weighted averaging of the iterates (cf. Figure 1) allows us to achieve optimal convergence rates while being robust to the setting of the step-size and the modeling assumptions [22], [23]. For example, recent work has shown [23] that an averaged SG iteration with a constant step-size achieves an $O(1/\varepsilon)$ convergence rate even without strong convexity under joint self-concordance-like and Lipschitz gradient assumptions. Another interesting recent development has been stochastic algorithms that achieve linear convergence rates for strongly convex problems of the form (16) in the special case where the data size is finite [24].

RANDOMIZED LINEAR ALGEBRA

For big data problems, basic linear algebra operations, such as matrix decompositions (e.g., eigenvalue, singular value, and Cholesky) and matrix–matrix multiplications can be major computational bottlenecks due to their superlinear dependence on dimensions. However, when the relevant matrix objects have low-rank representations (i.e., $M = LR^T$ with $L \in \mathbb{R}^{p\times r}$ and $R \in \mathbb{R}^{p\times r}$ where $r \ll p$), the efficiency of these methods uniformly improves. For instance, the corresponding singular value decomposition (SVD) of M would only cost $O(pr^2 + r^3)$ flops.

The idea behind randomized linear algebra methods is either to approximate $M \approx Q(Q^T M)$ with $Q \in \mathbb{R}^{p\times r}$, or to construct a low-rank representation by column or row subset selection to speedup computation. And indeed, doing this in a randomized fashion gives us control over the distribution of the errors [25], [26]. This idea generalizes to matrices of any dimensions and has the added benefit of exploiting mature computational routines in nearly all programming languages. Hence, they immediately lend themselves well to modern distributed architectures.

We describe three impacts of randomizing linear algebra routines in optimization here. First, we can accelerate computation of the proximity operators of functions that depend on spectral values of a matrix. For instance, the proximity operator of the nuclear norm, used in matrix completion and R PCA problems, requires a partial SVD. This is traditionally done with the Lanczos algorithm which does not parallelize easily due to synchronization and reorthogonalization issues. However, with the randomized approach, the expected error in the computation is bounded and can be used to maintain rigorous guarantees for the convergence of the whole algorithm [27].

Second, the idea also works in obtaining unbiased gradient estimates for matrix objects, when randomization is chosen appropriately, and hence applies to virtually all SG algorithms. Finally, the randomized approach can be used to sketch objective functions, i.e., to approximate them to obtain much cheaper iterations with exact first-order methods while retaining accuracy guarantees for the true objective [26].

Algorithm 7: Randomized low-rank approximation.

```
Require: $M \in \mathbb{R}^{p\times p}$, integer $r$
 1) Draw $\Omega \in \mathbb{R}^{p\times r}$ iid $\mathcal{N}(0,1)$
 2) $W = M\Omega$                         //Matrix multiply, cost is $O(p^2 r)$
 3) $QR = W$   //QR algorithm, e.g., Gram-Schmidt, cost is $O(pr^2)$
 4) $U = M^T Q$                           //Matrix multiply, cost is $O(p^2 r)$
 5) return $\hat{M}_{(r)} = QU^T$                   //Rank $r$ approximation
```

Algorithm 7 is an example of a randomized low-rank approximation, which is simply a single step of the classical QR iteration, using a random initial value. Surprisingly, the error in approximating M is nearly as good as the *best* rank-r approximation, where $\ell = r + \rho$ and ρ is small. Specifically, for $r \ge 2, \rho \ge 2$ and $\ell \le p$, [25] provides the bound

$$\mathbb{E}\|\hat{M}_{(\ell)} - M\|_F \le \sqrt{1 + \frac{r}{\rho - 1}}\,\|M - M_{(r)}\|_F,$$

where the expectation is taken with respect to the randomization, and $M_{(r)}$ is the best rank-r approximation of M, which only keeps the first r terms in the SVD and sets the rest to zero; furthermore, [25] shows a deviation bound showing that the error concentrates tightly around the expectation. Thus, the approximation can be very accurate if the spectrum of the matrix decays to zero rapidly. For additional randomized linear algebra schemes and their corresponding guarantees, including using a power iteration to improve on this bound, we refer the readers to [25].

Figure 4 illustrates the numerical benefits of such randomization over the classical Lanczos method. Since the randomized routine can perform all the multiplications in blocks, it benefits significantly from parallelization.

THE ROLE OF PARALLEL AND DISTRIBUTED COMPUTATION

Thanks to Moore's law of scaling silicon density, raw computational throughput and storage capacity have increased at exponential rates up until the mid-2000s, thereby giving convex optimization algorithms commensurate performance boosts. However, while Moore's law is expected to continue for years to come, transistor efficiencies have plateaued. As dictated by Dennard's law, scaling silicon density now results in unprecedented levels of power consumption. To handle the massive computational and storage resources demanded by big data at reasonable power costs, we must hence increasingly rely on parallel and distributed computation.

> **TO HANDLE THE MASSIVE COMPUTATIONAL AND STORAGE RESOURCES DEMANDED BY BIG DATA AT REASONABLE POWER COSTS, WE MUST HENCE INCREASINGLY RELY ON PARALLEL AND DISTRIBUTED COMPUTATION.**

While first-order methods seem ideally suited for near-optimal performance speed-ups, two issues block us when using distributed and heterogeneous hardware:

- *Communication:* Uneven or faulty communication links between computers and within the local memory hierarchy can significantly reduce the overall numerical efficiency of first-order methods. Two approaches broadly address such drawbacks. First, we can specifically design algorithms that minimize communication. Second, we can eliminate a master vector x^k and instead work with a local copy in each machine that each lead to a consensus x^* at convergence.
- *Synchronization*: To exactly perform the computations in a distributed fashion, first-order methods must coordinate the activities of different computers whose numerical primitives depend on the same vector x^k at each iteration. However, this procedure slows down when even a single machine takes much longer than the others. To alleviate this quintessential synchronization problem, asynchronous algorithms allow updates using outdated versions of their parameters.

In this section we describe several key developments related to first-order methods within this context. Due to lack of space, we will gloss over many important issues that impact the practical performance of these methods, such as latency and multihop communication schemes.

EMBARRASSINGLY PARALLEL FIRST-ORDER METHODS

First-order methods can significantly benefit from parallel computing. These computing systems are typified by uniform processing nodes that are in close proximity and have reliable communications. Indeed, the expression *embarrassingly parallel* refers to an ideal scenario for parallelization where we split the job into independent calculations that can be simultaneously performed in a predictable fashion.

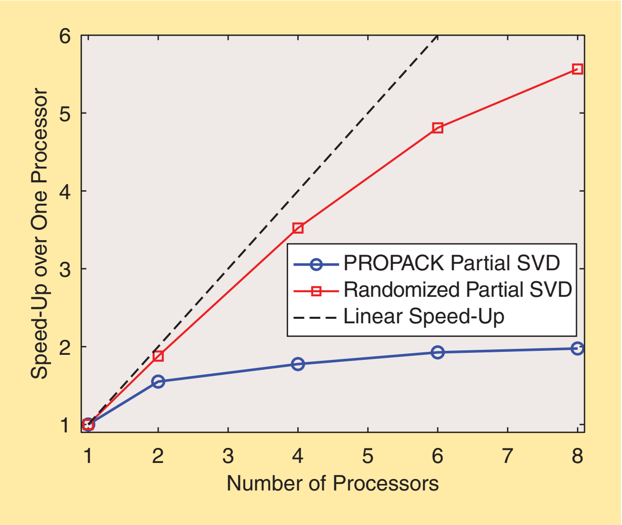
In parallel computing, the formulation of the convex problem makes a great deal of difference. An important embarrassingly parallel example is the computation of the gradient vector when the objective naturally decomposes as in (16). Here, we can process each F_i with one of m computers using only $O(n/m)$ local computation. Each machine also stores data locally with the corresponding $O(n/m)$-data samples since each F_i directly corresponds to a data point. Each processor then communicates with the central location to form the final gradient and achieve the ideal linear speed-up.

Algorithm 8: Decomposition algorithm (aka, consensus ADMM) [28] to solve (18); $\gamma > 0, x_i^0 = 0$ for $i = 1, \ldots, n$.

```
1) $z^{k+1} = (1/n)\sum_{i=1}^{n} \mathrm{prox}_{\gamma F_i}(x_{(i)}^k)$
2) for $i = 1$ to $n$ do
3)     $x_{(i)}^{k+1} = 2z^{k+1} - z^k + x_{(i)}^k - \mathrm{prox}_{\gamma F_i}(x_{(i)}^k)$
4) end for
```

Beyond parallelizing the basic gradient method for smooth problems, an embarrassingly parallel distribute-and-gather framework for nonsmooth problems results from an artificial reformulation of (16) so that we can apply decomposition techniques, such as Algorithm 8:

$$\min_{x, x_{(i)} : i = 1, \ldots, n} \left\{ \frac{1}{n} \sum_{i=1}^{n} F_i(x_{(i)}) : x_{(i)} = x, i = 1, \ldots, n \right\}. \qquad (18)$$



[FIG4] **Computing the top five singular vectors of a 10^9 entry matrix using varying number of computer cores. The matrix is a dense $61,440 \times 17,784$ matrix (8.1-GB RAM) generated from video sequences from http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html. Such partial SVDs are used in the proximity operator for the nuclear norm term that arises in robust PCA formulations of video background subtraction. The randomized factorization happens to be faster than the Lanczos-based SVD from the PROPACK software even with a single core, but more importantly, the randomized method scales better as the parallelism increases. The accuracies of the two methods are indistinguishable.**

Indeed, the decomposition idea above forms the basis of the massively parallel consensus ADMM algorithm, which provides an extremely scalable optimization framework for $n > 2$. See [15], [18], and [28] for convergence analysis and further variants that include additional linear operators.

Fortunately, we have access to many computer programming models to put these ideas immediately into action. Software frameworks, such as MapReduce, Hadoop, Spark, Mahout, MADlib, SystemML, and BigInsights, and corresponding high-level languages such as Pig, Hive, and Jaql, can govern the various optimization tasks in parallel while managing all communications and data transfers within the computing system, and seamlessly provide for redundancy and fault tolerance in communications.

BIG DATA PROBLEMS NECESSITATE A FUNDAMENTAL OVERHAUL OF HOW WE DESIGN CONVEX OPTIMIZATION ALGORITHMS AND SUGGEST UNCONVENTIONAL COMPUTATIONAL CHOICES.

FIRST-ORDER METHODS WITH REDUCED OR DECENTRALIZED COMMUNICATIONS

In large systems, communicating the gradient or its elements to a central location may create a communication bottleneck. In this setting, coordinate descent methods provide a principled approach to reduce communications. There is indeed substantial work on developing parallel versions of these methods, dating back to work on the Jacobi algorithm for solving linear systems. The basic idea is simply to apply several coordinate descent updates at the same time in parallel. The advantage of this strategy in terms of communication is that each processor only needs to communicate a single coordinate update, while it only needs to receive the updates from the coordinates that have changed.

When the objective is decomposable, this is simply an embarrassingly parallel version of the serial algorithm. Furthermore, classical work shows that this strategy is convergent, although it may require a smaller step-size than the serial variant. However, it does not necessarily lead to a speed increase for nonseparable functions. Recent work has sought to precisely characterize the conditions under which parallel coordinate descent methods still obtain a large speed-up [8].

Surprisingly, we can also decentralize the communication requirements of gradient methods for decomposable objectives with only minor modifications [29]. The resulting algorithm performs a modified gradient update to the average of the parameter vectors only among the neighbors with which it communicates. This strategy in fact achieves similar convergence rates to the gradient method with central communications, where the rate degradation depends on the graph Laplacian of the underlying communication network.

ASYNCHRONOUS FIRST-ORDER METHODS WITH DECENTRALIZED COMMUNICATIONS

The gradient and the decomposition methods above still require a global synchronization to handle decomposable problems such as (16). For instance, the gradient algorithm computes the gradient *exactly* with respect to one (or more) examples at x^k and then synchronizes in sequence to update x^{k+1} in a standard implementation. In contrast, SG algorithms that address (16) only use a crude approximation of the gradient. Hence, we expect these algorithm to be robust to outdated information, which can happen in asynchronous settings.

A variety of recent works have shown that this is indeed the case. We highlight the work [7], which models a lock-free shared-memory system where SG updates are independently performed by each processor. While the lock-free SG still keeps a global vector x, processors are free to update it without any heed to other processors and continue their standard motions using the cached x. Under certain conditions this asynchronous procedure preserves the convergence of SG methods, and results in substantial speed-ups when many cores are available. The same memory lock-free model also applies to stochastic parallel coordinate descent methods [8]. Finally, first-order algorithms with randomization can be effective even in asynchronous and decentralized settings with the possibility of communication failures [30].

OUTLOOK FOR CONVEX OPTIMIZATION

Big data problems necessitate a fundamental overhaul of how we design convex optimization algorithms, and suggest unconventional computational choices. To solve increasingly larger convex optimization problems with relatively modest growth in computational resources, this article makes it clear that we must identify key structure-dependent algorithmic approximation tradeoffs.

Since the synchronization and communication constraints of the available hardware naturally dictates the choice of the algorithms, we expect that new approximation tools will continue to be discovered that ideally adapt convex algorithms to the heterogeneity of computational platforms. We also predict an increased utilization of composite models and the corresponding proximal mapping principles in parallel and distributed architectures for nonsmooth big data problems in order to cope with noise and other constraints. For example, the LASSO formulation in (4) has estimation guarantees that are quantitatively stronger than the guarantees of the LS estimator when the signal x_0 has at most k nonzero entries and Φ obeys certain assumptions [1]. That is to say, to get more out of the same data, we must use composite models. This also invites the question of whether we can use composite models to get the same information out but do it faster, an issue which has been discussed [5], [6] but not yet had an impact in practice.

ACKNOWLEDGMENTS

Volkan Cevher's work is supported in part by the European Commission under grants MIRG-268398 and ERC Future Proof and by the Swiss Science Foundation under grants SNF 200021-132548, SNF 200021-146750, and SNF CRSII2-147633.

AUTHORS

Volkan Cevher (volkan.cevher@epfl.ch) received the B.S. (valedictorian) degree in electrical engineering in 1999 from Bilkent University in Ankara, Turkey, and the Ph.D. degree in electrical and computer engineering in 2005 from the Georgia Institute of Technology in Atlanta. He held research scientist positions at the University of Maryland, College Park, from 2006 to 2007 and at Rice University in Houston, Texas, from 2008 to 2009. Currently, he is an assistant professor at the Swiss Federal Institute of Technology Lausanne and a faculty fellow in the Electrical and Computer Engineering Department at Rice University. His research interests include signal processing theory, machine learning, graphical models, and information theory. He received a Best Paper Award at the Signal Processing with Adaptive Sparse Representations Workshop in 2009 and a European Research Council Starting Grant in 2011.

Stephen Becker (srbecker@us.ibm.com) is an assistant professor in the Applied Math Department at the University of Colorado at Boulder. Previously, he was a Goldstine postdoctoral fellow at IBM Research T.J. Watson Lab and a postdoctoral fellow at the Laboratoire Jacques-Louis Lions at Paris 6 University. He received his Ph.D. degree in applied and computational mathematics from the California Institute of Technology in 2011 and bachelor's degrees in math and physics from Wesleyan University in 2005. His work focuses on large-scale continuous optimization for signal processing and machine-learning applications.

Mark Schmidt (mark.schmidt@sfu.ca) is an assistant professor working in the field of machine learning and large-scale optimization in the Department of Computer Science at the University of British Columbia. He previously worked in the Natural Language Laboratory at Simon Fraser University and, from 2011 through 2013, at the École normale supérieure in Paris on inexact and stochastic convex optimization methods. He finished his M.Sc. degree in 2005, at the University of Alberta working as part of the Brain Tumor Analysis Project, and his Ph.D. degree in 2010 at the University of British Columbia working on graphical model structure learning with L1-regularization. He has also worked at Siemens Medical Solutions on heart motion abnormality detection and with Michael Friedlander in the Scientific Computing Laboratory at the University of British Columbia on semistochastic optimization methods.

REFERENCES

[1] M. J. Wainwright, "Structured regularizers for high-dimensional problems: Statistical and computational issues," *Ann. Rev. Stat. Appl.*, vol. 1, no. 1, pp. 233–253, Jan. 2014.

[2] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[3] Y. Nesterov and A. Nemirovski, "On first-order algorithms for ℓ_1/nuclear norm minimization," *Acta Numer.*, vol. 22, pp. 509–575, May 2013.

[4] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. New York: Springer-Verlag, 2011, pp. 185–212.

[5] V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," *Proc. Natl. Acad. Sci.*, vol. 110, no. 13, pp. E1181–E1190, 2013.

[6] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning." in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 20, pp. 161–168, 2008.

[7] F. Niu, B. Recht, C. Ré, and S. J. Wright, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," in *Proc. Advances in Neural Information Processing Systems (NIPS),* vol. 24, pp. 693–701, 2011.

[8] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Math. Programming*, to be published.

[9] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, ser. Applied Optimization. Norwell, MA: Kluwer, 2004, vol. 87.

[10] B. O'Donoghue and E. J. Candès, "Adaptive restart for accelerated gradient schemes," *Fond. Comp. Math.*, July 2013, pp. 1–18.

[11] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, "Composite self-concordant minimization," arXiv:1308.2867v2 [stat.ML].

[12] M. Schmidt, N. L. Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 24, pp. 1458–1466, 2011.

[13] M. Beck and A. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[14] M. B. McCoy, V. Cevher, Q. T. Dinh, A. Asaei, and L. Baldassarre, "Convexity in source separation: Models, geometry, and algorithms" *IEEE Signal Processing Mag.*, vol. 31, no. 3, pp. 87–95, 2014.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[16] E. Esser, X. Zhang, and T. Chan, "A general framework for a class of first order primal-dual algorithms for TV minimization," UCLA, Center for Applied Math, Tech. Rep. 09-67, 2009.

[17] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vision*, vol. 40, no. 1, pp. 120–145, 2010.

[18] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[19] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. Int. Conf. on Machine Learning,* 2013.

[20] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM J. Optim.*, vol. 22, no. 2, pp. 341–362, 2012.

[21] Z.-Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *J. Optim. Theory Appl.*, vol. 72, no. 1, pp. 7–35, 1992.

[22] A. Nedic and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," in *Stochastic Optimization: Algorithms and Applications*. Norwell, MA: Kluwer, 2000, pp. 263–304.

[23] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 26, pp. 773–781, 2013.

[24] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 25, pp. 2663–2671, 2012.

[25] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.

[26] M. Mahoney, "Randomized algorithms for matrices and data," *Found. Trends Mach. Learn.*, vol. 3, no. 2, pp. 123–224, 2011.

[27] S. Becker, V. Cevher, and A. Kyrillidis, "Randomized singular value projection," in *Proc. Sampling Theory and Approximation (SampTA)*, Bremen, Germany, 2013, pp. 364–367.

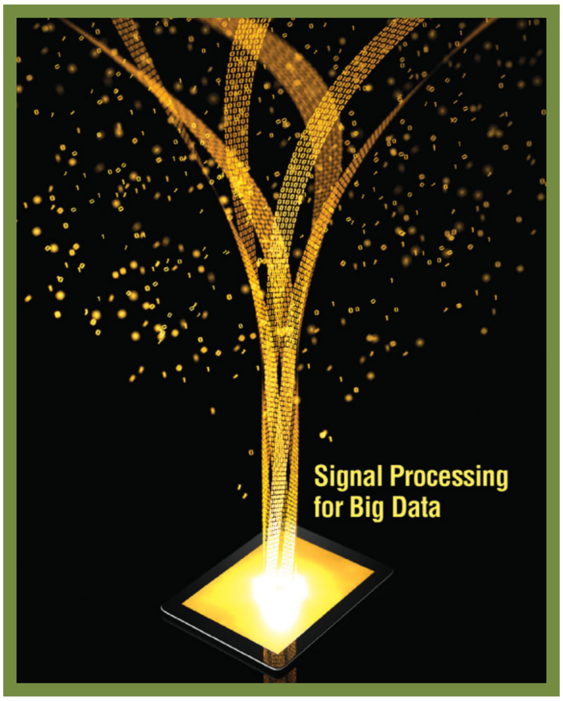
[28] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Prob.*, vol. 24, no. 6, p. 27, 2008.

[29] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," Univ. California Lost Angeles, Rep. 14-34, 2014.

[30] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 24, pp. 873–881, 2011.

Ali Tajer, Venugopal V. Veeravalli, and H. Vincent Poor

Outlying Sequence Detection in Large Data Sets



© ISTOCKPHOTO.COM/TA2YO4NORI

A data-driven approach

Outliers refer to observations that do not conform to the expected patterns in high-dimensional data sets. When such outliers signify risks (e.g., in fraud detection) or opportunities (e.g., in spectrum sensing), harnessing the costs associated with the risks or missed opportunities necessitates mechanisms that can identify them effectively. Designing such mechanisms involves striking an appropriate balance between reliability and cost of sensing, as two opposing performance measures, where improving one tends to penalize the other. This article poses and analyzes outlying sequence detection in a hypothesis testing framework under different outlier recovery objectives and different degrees of knowledge about the underlying statistics of the outliers.

Digital Object Identifier 10.1109/MSP.2014.2329428
Date of publication: 19 August 2014

INTRODUCTION

MOTIVATION

Advances in data acquisition and high-dimensional information processing are rapidly transforming various technological, social, and economic domains, including the Internet, telecommunication, energy grids, social networks, and the health industries, to name a few. Empowered by these advances, such domains are evolving into

1053-5888/14©2014IEEE

complex networked platforms in which high-dimensional and complex data is routinely generated, communicated, stored, and processed for various monitoring, inference, and resource management purposes. Due to the inherent scale of data and complexity of the processes involved, the challenges associated with capturing, curating, searching, and sharing the information are also expected to grow well into the future. Hence, benefiting from the full extent of such enabling technologies is feasible only when appropriate measures are implemented that address these growing challenges while recognizing constraints pertinent to the physical limits of the application domains of interest.

Analyzing large-scale and complex data sets involves multifaceted phases, each of which introduces its own set of challenges. These phases include data acquisition and storage, information extraction, data aggregation, data modeling, and query processing, and the associated challenges include information heterogeneity, processing timeliness, data security and privacy, and human interactions. By capitalizing on the promises of data-driven information processing theories for understanding and addressing these challenges, this article focuses on a particular class of challenges related to information extraction and its associated timeliness requirements.

Extracting information and knowledge from data sets has been studied extensively over the past decade through developing powerful data mining and statistical learning methods. These methods are primarily focused on discovering (inferring) patterns in data sets and have widespread applications. In addition to the ongoing developments in discovering patterns in large data sets, there has also been a growing interest in uncovering outlying observations, which are observations that do not conform to expected patterns in large data sets. Such outlying observations generally refer to observations that are significantly different from the other data set constituents. While defining and identifying outliers are subjective exercises, outlier observations are often abstracted as deviations in the nature of a data set population and are considered to be caused by transient disruptions during data acquisition due to, for instance, a malfunctioning measurement apparatus, noisy data transmission media, or abrupt changes in the nature or behavior of the population. There exists a rich literature on outlier detection for the setting in which outliers are candidates for aberrant data that lead to biased or incorrect inferences. The general approach to cope with outliers in such circumstances is to clean up the data prior to modeling and performing the attendant statistical analysis [1]. Relevant outlier detection methods can be categorized under different taxonomies, the major ones being univariate versus multivariate methods and parametric versus nonparametric methods. Some popular approaches for such outlier detection approaches include Pierce's criterion [2], Chauvenet's criterion [3], and Dixon's test [4].

In contrast to the aforementioned notion of outlier detection that aims to render disturbance-free data, a less-investigated aspect of identifying outliers pertains to searching for rare and at the same time significant anomalies that do not conform to expected patterns and are often manifested as opportunities to be exploited (arising, e.g., in spectrum sensing) or risks to be ameliorated (e.g., network intrusion or fraud detection). In these settings, we can consider the outlying sequence detection problem as one in which a large number of sequences are being monitored simultaneously and the goal is to choose a small subset of sequences that are outliers. We refer to such problems as outlying sequence detection problems to distinguish them from the setting described in the previous paragraph in which a few outlier observations are winnowed out from a single set of data.

Detecting the outliers, especially in large data sets, is often very time-sensitive due to the transient nature of the opportunities that are attractive only when detected quickly, or due to the substantial costs that risks can incur if not managed swiftly. In this article, we focus on the fundamental problems in quick detection of outliers while recognizing different system- and physical-level constraints imposed by various contexts.

BACKGROUND

Outlier detection has immediate application in a broad range of contexts in which large volumes of data are constantly generated and processed. Some of these contexts and their application domains will be reviewed briefly in the section "Application Domains." While outlier observations in all contexts conform in representing unusual changes of the behavior of the underlying physical phenomena over one or more dimensions (e.g., time or space), the broad diversity in the range of the relevant applications necessitates diverse formulations that are customized to capture the specifics of each application domain. The remainder of this subsection focuses on reviewing some of the widespread models for abstracting the outlier detection problem in large data sets. The three major components for modeling the outliers and abstracting the outlier detection problem are the level of available information about the normal and outlying data streams, the type of the outliers, and the figure of merit for identifying the outliers. A comprehensive review of all such abstractions can be found in [5] and [6].

SUPERVISION LEVEL

Availability of information about the models for the data streams governs the modes and approaches for performing outlier detection in large data sets. Specifically, the existing approaches to outlier detection can be broadly categorized into four classes: supervised, semisupervised, unsupervised, and universal approaches, which are distinguished based on the availability of information about the structure of the data streams.

- *Supervised*: In the presence of prior information about the data streams (often acquired through training data) the models of both normal and abnormal (outlying) observations are known, which enables supervised outlier detection. These approaches are appropriate for static data or data models that evolve slowly enough so that tracking and learning the changes in the model are viable. In the statistics and computer science literature, the class of supervised outlier detection is studied extensively under classification-based approaches [7], [8], neural networks [9]–[11], Elman networks [12], naïve Bayes, and support vector machines [13].

- *Semisupervised*: In many practical circumstances acquiring models for both normal and outlying data streams is often infeasible. Based on the availability of information about a model for either normal or outlying sequences, semisupervised outlier detection approaches are developed, which capitalize on the known structure of normal (outlying) data streams to be robust against uncertainty about the structure of outlying (normal) data streams. While there exist scenarios that assume availability of information about the outlying sequences and lack of information about normal data streams [9], [14], these scenarios do not often arise. This is primarily due to the fact that outliers typically have an unpredictable nature and designing learning algorithms that can cover all possible outlying events is difficult. On the other hand, normal behavior is often well defined and thus it is more viable to construct models for normal data streams. Hence, in the majority of the existing literature on semisupervised outlier detection, the normal data streams are assumed to have known models while those of the outliers are unknown.
- *Unsupervised*: Under this category no assumption is made about models for the normal or outlying data streams and, instead, some other assumptions (e.g., parametric) are made about the models. In these approaches the normal observation are those that share a pattern occurring frequently and the outliers are those with rare and distinct patterns. Some representative unsupervised approaches include discriminative approaches [15]–[19], parametric approaches [18], [20]–[24], and online analytical processing (OLAP) approaches [25].
- *Completely universal*: Unlike in the supervised, semisupervised and unsupervised approaches, in the completely universal approach, no training data is available for either the typical or outlier distributions. As we discuss in the section "Universal Outlying Sequence Detection," it is possible to construct decision rules under this completely universal setting, with only the assumption that the typical and outlier distributions are different.

TYPES OF OUTLIERS

A pivotal step toward formulating any outlier detection approach is an abstraction for modeling the outliers. Here we review some of the more common categories of outliers, which are distinguished based on their composition and their relevance to normal observations.

- *Outlying points within a data stream*: This type of outlier occurs in circumstances when we are dealing with one data stream (often modeled as a time series) and one or more isolated elements of the stream do not conform to the common pattern of the data stream. Depending on whether the objective is to perform real-time or in-retrospect (offline) outlier detection, there are two different types of detection procedures. In real-time scenarios, the existing approaches often dynamically provide forecasts for the upcoming observations and, upon collecting the actual observations, a similarity measure between the actual observations and their forecast is computed. This measure determines whether the observation deviates from the expected pattern, and consequently whether it is an outlier or a normal observation [26]. In the offline outlier detection approaches, on the other hand, one popular approach is to cast the outlier detection problem as an in-retrospect change point detection problem [27].
- *Outlying subsequences within a data stream*: In contrast to outlying points, which appear sporadically and in isolation in one data stream, outlying subsequences appear in the form of consecutive outlying points. Similar to outlying points, detecting such outliers can be studied under real-time and offline settings. For the former there exist a body of window-based prediction approaches that form similarity measures for identifying outlying subsequences, and, for the latter, in-retrospect change point detection approaches are applicable.
- *Outlying data streams*: The previous two types of outliers occur within a data stream. Outlying data streams occur when we are given a large group of data streams, most of which follow a common pattern, but a few of which do not conform to this common pattern. Hence, there is no notion of outliers occurring within a stream anymore, but rather, each entire data stream is either normal or outlying. In such circumstances, the objective is to identify a group of sequences that exhibit behaviors different from the common pattern. This setting has been studied extensively in the statistics literature in which several approaches based on autoregression, moving average, and cumulative sum tests have been proposed with details reviewed in depth in [1], [5], and [28].

DECISION MECHANISM

Upon designing the information-gathering process and collecting observations, there are two broad schemes for forming a decision on individual observations or sets of observations and categorizing them as normal or outlying. In one approach, often termed the *labeling technique*, a binary decision about each individual observation is made. The outcome in this approach is a classification of the observations into two sets. The advantage of this approach is its accuracy in labeling every observation with a decision, while its drawback is that when the data volume increases, forming an accurate decision for every single observation is computationally prohibitive. In an alternative approach, often referred to as a *sorting technique*, each observation receives a score that indicates the likelihood of that observation being an outlier. The advantage of this technique is that it is less stringent in reaching an accurate decision for all observations in favor of enhancing the speed of the detection procedure, which makes it more suited for analyzing large data sets. The drawback of this approach, on the other hand, is that there should be a supplementary mechanism deciding about a threshold on the scores to delineate the normal and outlying regions.

APPLICATION DOMAINS

Different combinations of the different types of outliers, supervision level, and decision mechanisms (and other details reviewing, which is not relevant to the scope of this article) create different abstractions for the outlier detection problem, each of which is relevant in certain application domains. Specifically, there exist a wide range of applications in which large volumes of data are constantly generated and the goal is to search for features or to identify anomalies that signify risks or opportunities. These goals can often be cast as outlier detection where the nature of the outliers, supervision level, the attendant decision mechanism, and other assumptions and constraints collectively formulate the underlying outlier detection problem. Examples of the application domains that involve detecting outliers in large data sets include credit card fraud detection [29], clinical trials [30], high-frequency trading [31], voting irregularity analysis [32], spectrum sensing [33], network intrusion [34], severe weather prediction [35], and seismic data analysis [36]. In this subsection we review a few application domains in which the problem of outlying sequence detection has important physical implications.

NETWORK INTRUSION

Network intrusion detection refers to detecting malicious penetrations to data networks. Intrusions exhibit behaviors different from the normal patterns in the network and the measurements associated with them can be modeled as outliers. The major impediment for identifying intrusions in this setting is the large volume of data, which makes the intrusion detection process computationally costly and time-consuming, while agile response to the presence of the intruders is crucial as any delay in detecting them leads to recovery costs for the system. Intrusions can often be modeled as outlying subsequences or sequences for which an observation model is unknown and, consequently, semisupervised or unsupervised approaches are best suited for identifying them. A comprehensive review of the literature on outlier detection approaches for network intrusion detection is available in [37].

FRAUD DETECTION

Fraud detection, which is the practice of identifying deliberately unlawful gains, is widely deployed by commercial entities including financial institutions, telecommunication companies, and insurance agencies. The pivotal step in designing fraud detection algorithms is creating profiles for usage activities of legitimate users and flagging any activity deviant from these profiles as a potential fraud. Hence, fraudulent activities can be modeled as outlying activities that should be identified swiftly to minimize the associated financial losses. A survey of different outlier detection approaches suited for credit card, mobile phone, insurance claim, and insider trading fraud detection is available in [37].

SPECTRUM SENSING

Wireless connectivity is ubiquitous and is constantly growing in scale and complexity to cope with the existing demands (e.g., data communication and sensor networks) and to accommodate the emerging ones (e.g., wireless health and smart grids). All such enabling technologies are viable at the expense of increasing demands for radio spectrum, which is the major commodity in the wireless industry. As reported by the U.S. Federal Communications Commission (FCC), exclusive spectrum access rights lead to underutilization of the spectrum. Driven by this observation and the urgency for higher spectral efficiency, future spectrum access policies are envisioned to provide the flexibility of dynamically granting spectrum access to unlicensed wireless services when the spectrum is underutilized by the license-holding services. Under such envisioned spectrum access policies, unlicensed services compete to make use of shared spectrum opportunities. The underutilized segments of the spectrum, hence, will not be as abundant as they otherwise should be and such reduction in their availability becomes even more severe as wireless sensing and networking grows in size and services. Hence, spectrum holes across wideband spectrum can be modeled as outliers in terms of their occupancy status and the problem of spectrum sensing in congested wideband spectrum can be abstracted as an outlier detection problem [33].

ENVIRONMENTAL MONITORING

The applications of outlier detection in environmental monitoring are multifaceted. Different forms of outlier detection are being used across the globe, e.g., for determining locations with constantly different temperatures from their neighbors, discovering drought areas, positioning fertility loss areas, and detecting hurricanes. A detailed overview of these application domains is available in [6].

DATA-ADAPTIVE OUTLYING SEQUENCE DETECTION

We introduce a general dichotomous hypothesis testing model for the outlying sequence detection problem of interest. This will be a unifying theme for investigating the problem under different settings. In this dichotomous model, we assume that the data set consists of M data streams, each being either a typical or an outlying sequence. Typical sequences exhibit identical statistical behavior, with which the outliers do not comply by exhibiting arbitrarily different known or unknown behaviors. The data volume increases as the number of data streams M increases, and in this article the focus is placed on high-dimensional data by performing the analysis in the asymptote of large values of M (i.e., $M \to \infty$). Furthermore, to emphasize the rarity of the outliers, we assume that the number of outliers grows sublinearly as M increases.

The above dichotomous model is adopted to mainly focus the attention on the discrepancy between the outliers and the typical observations and can be generalized to models that involve multiple statistical behaviors for the typical sequences. Each data stream generates independent and identically distributed (i.i.d.) real observations $\{Y_1^{(i)}, Y_2^{(i)}, \ldots\}$ obeying one of the two models

$$\begin{aligned} \mathrm{H}_0: &\quad Y_t^{(i)} \sim F, \quad t = 1, 2, \ldots \\ \mathrm{H}_1: &\quad Y_t^{(i)} \nsim F, \quad t = 1, 2, \ldots, \end{aligned} \qquad (1)$$

where F denotes a cumulative distribution function (cdf), modeling the statistical behavior of the typical sequences. Designing an optimal outlying sequence detector rests fundamentally on delineating the inherent interplay between two opposing performance measures, one being the frequency of erroneous decisions and the other being the cost of sensing (e.g., the number of measurements taken). To this end, we consider the most general structure for the information-gathering process, which either sequentially, or based on a prespecified rule selects and takes measurements from a subset of the data streams at each time. By denoting the subset of data streams selected at time t by $\mathcal{L}_t \subseteq \{1, \ldots, M\}$, upon collecting the measurements at time t, the outlier detection process takes one of the following actions:

1) *Observation:* due to lack of sufficient information making any decision is deferred and the same set of data streams is retained for more scrutiny, i.e., $\mathcal{L}_{t+1} = \mathcal{L}_t$

2) *Exploration:* the information accumulated is insufficient to identify the outliers, but provides partial information that is sufficient for updating the set of data streams that should be measured more carefully, or possibly ruling out some of the data streams as typical ones, i.e., $\mathcal{L}_t \rightarrow \mathcal{L}_{t+1}$

3) *Detection:* the information gathering process is terminated and the outliers are identified.

The stopping time of the procedure, i.e., the time after which detection is performed, is denoted by τ. Furthermore, a switching function $\psi : \{1, \ldots, \tau\} \rightarrow \{0, 1\}$ is devised to distinguish between observation and exploration actions at time t. The switch is set to $\psi(t) = 0$ if it is decided in favor of performing observation at time t, while $\psi(t) = 1$ indicates a decision in favor of performing exploration. The sequential information-gathering procedure is uniquely determined by its stopping time τ, the sequence of switching functions $\bar{\psi}_\tau = [\psi(1), \ldots, \psi(\tau)]$, and the ordered collection $\overline{\mathcal{L}}_\tau \triangleq \{\mathcal{L}_1, \ldots, \mathcal{L}_{\tau-1}\}$.

The quality of the ultimate decision, which is the output of the detection action, is captured by the frequency of erroneous decisions. To formalize the dependence of such decision quality, on the given set of stopping time τ, switching sequence $\bar{\psi}_\tau$, and observation order $\overline{\mathcal{L}}_\tau$, we denote the frequency of erroneous decisions by $\mathrm{P}_M(\tau, \bar{\psi}_\tau, \overline{\mathcal{L}}_\tau)$. An optimal outlying sequence detection approach can be characterized as a strategy that optimizes a desired balance between this decision quality and the aggregate cost of sensing $\sum_{t=1}^{\tau} |\mathcal{L}_t|$, which incorporates the stopping time and the number of samples taken during the exploration cycles. Such a balance often can be cast as minimizing one of these measures, within a desired constraint on the other, e.g.,

$$\begin{array}{ll} \min_{\tau, \overline{\psi}_\tau, \overline{\mathcal{L}}_\tau} & \mathbb{E}\left[\sum_{t=1}^{\tau} |\mathcal{L}_t|\right] \\ \text{s.t.} & \mathrm{P}_M(\tau, \overline{\psi}_\tau, \overline{\mathcal{L}}_\tau) \leq \rho, \end{array} \tag{2}$$

where ρ controls the decision reliability. In the following sections, we discuss several important topics under which the outlying sequence detection problem has different interpretations and can be cast as a balance between these measures.

Obtaining the optimal strategies for observation, exploration, and detection that strike a desired balance between decision quality and cost of sensing, in its most general form, is an open problem. By imposing certain structures on data or sampling models, however, one can delineate optimal strategies. In the remainder of the article, we discuss different outlying sequence detection approaches with different structures ranging from fully sequential to fully prespecified sampling strategies, and different objectives, ranging from identifying only one outlier to identifying all.

DATA-ADAPTIVE SAMPLING

In this section, we concretize the generic outlying sequence detection problem by focusing the attention on the closed-loop (adaptive) aspects of the sampling process. The extent of data adaptivity of the data-gathering process leads to a wide range of structures for the outlying sequence detection problem. Adaptivity is embedded in the sequential selection of the subset of data streams to be measured at each time, i.e., $\{\mathcal{L}_1, \ldots, \mathcal{L}_\tau\}$. Besides adaptivity in sensing, identifying the outliers can also be performed in either sequential or nonsequential fashion, where in the former the data collected is processed altogether to identify the outliers, whereas in the latter one could identify and remove an outlier and then search for other outliers among the remaining data streams.

QUICKEST SEARCH FOR ALL OUTLIERS

When the objective is to identify all outliers with minimum expected number of aggregate measurements and subject to controlled reliability, the problem is equivalent to forming a decision about the underlying model of all the sequences. Hence, the optimal sampling and decision-making problem can be decomposed into M independent hypothesis testing problems corresponding to the M sequences. The optimal solution to these latter subproblems is the sequential probability ratio test (SPRT), which minimizes the expected number of measurements required for forming a decision for each sequence with prespecified reliability [38] when the underlying distributions for normal and outlying observations are known.

These independent SPRTs can be performed either in parallel or sequentially. When performed in parallel, the sampling procedure is initiated by setting $\mathcal{L}_t = \{1, \ldots, M\}$ and after taking measurements at time t, the set $\mathcal{L}_t$ is refined by discarding the indices of the sequences for which their associated SPRT has reached a decision. In contrast, when the SPRTs are performed sequentially, the sampling strategy focuses on the sequences one at a time. While being effective in forming accurate decisions for individual data streams, performing independent SPRTs becomes computationally prohibitive as the size of the data set grows, and is not a suitable approach for large data sets.

QUICK SEARCH FOR A SUBSET OF OUTLIERS

In certain scenarios, one might be interested in recovering only a fraction of the outliers, especially when the outliers represent rare opportunities of interest, while in certain other scenarios, especially when the outliers model risks, it is imperative to identify all of the outliers.

Shifting the objective from recovering all the outliers to identifying only a fraction of them allows for missing some of the outliers in favor of quickly identifying the fraction of interest. Under this objective, performing SPRTs on all sequences is clearly not optimal as it tends to identify all sequences and does not take advantage of the more relaxed objective. Such a shift of objective and its ensuing flexibility leads to significant reduction in the sensing cost and the delay in reaching a decision.

Obtaining optimal structures of such sequential and data-adaptive experimental designs and finding associated nontrivial performance bounds for the such design are open for most scenarios, with some exceptions discussed in [39]–[41]. Nevertheless, by imposing certain structures on the refinement action, one can ascertain certain optimality properties with provable gains over nonadaptive approaches.

In this subsection, we focus on a specific structure studied in [33] and [42]–[45,] which consists of consecutive rounds of observations and exploration actions, followed by consecutive cycles of observations and satisfies certain optimality properties [45]. Driven by the premise that the outliers (anomalies) occur rarely, this adaptive structure starts by spending the sampling resources conservatively, and as more information about different data streams is accumulated, the sensing resources are progressively allocated to the data streams that are more likely outliers. The central motivation for such progressive allocation of the sensing resources is that while conservative (rough) observations are not accurate enough to identify the outliers, they can be informative enough to discard a considerable fraction of the typical streams. Consecutive cycles of rough observations and exploration, therefore, lead to substantial reduction in the search space, which facilitates using the sensing resources more effectively. Careful design of the exploration actions and the number of exploration actions, can provide sufficient guarantees that the discarded data streams are almost surely typical ones.

In this approach, more specifically, the sampling strategy is initiated by including all the streams for sampling and K consecutive cycles of exploration are performed, where K is determined by the amount of sampling resources and the fraction of the outliers one seeks to identify. The detailed steps of this procedure for identifying T outliers are provided in Table 1. In this procedure the exploration actions are designed such that at least T data streams will be retained after the exploration cycles for the final detection decision.

To assess adaptation gains, we formalize the adaptive experimental design problem as the minimizer of the decision quality under a hard constraint on the sampling budget, i.e.,

$$\mathcal{P}_M(S) \triangleq \begin{cases} \inf_{\tau,\overline{\psi}_\tau,\overline{\mathcal{L}}_\tau} & \mathrm{P}_M(\tau,\overline{\psi}_\tau,\overline{\mathcal{L}}_\tau) \\ \text{s.t.} & \frac{1}{M}\sum_{t=1}^{\tau}|\mathcal{L}_t| \leq S, \end{cases} \quad (3)$$

where S controls the sensing budget. Addressing the sensing problem in this setting sheds light on the ratio of the sensing resources to be allocated to the observation and exploration actions.

To assess the gains of adaptation we investigate the following two settings in which the typical distribution F is Gaussian with known mean and variance and the outliers are also Gaussian with either different mean or different variance values. Specifically, sequence i is generated according to $\mathcal{N}(\mu_i, \sigma_i^2)$. If sequence i is a typical sequence then $\mu_i = \mu$ and $\sigma_i = \sigma$, where μ and σ are known, and if it is an outlier sequence we consider two settings:

[TABLE 1] THE ADAPTIVE OUTLYING SEQUENCE DETECTION ALGORITHM.

```
SET $\mathcal{L}_1 = \{1,\ldots,M\}$
FOR $t = 1$ TO $K$
    TAKE ONE SAMPLE FROM EACH STREAM IN $\mathcal{L}_t$
    SET $\beta_t = (1-\zeta)(|\mathcal{L}_t| - T)$ FOR A PRESPECIFIED CONSTANT $0 < \zeta < 1$
    DISCARD $\beta_t$ STREAMS THAT ARE MOST LIKELY TYPICAL
END FOR
SET $s = \lfloor (S - \sum_{t=1}^{K}|\mathcal{L}_t|)/|\mathcal{L}_K| \rfloor$
TAKE $s$ SAMPLES FROM THE SURVIVING STREAMS
OUTPUT THE $T$ SEQUENCES THAT ARE LEAST LIKELY TYPICAL
```

$$\begin{array}{lll} \text{mean testing:} & \mu_i \neq \mu, & \text{and} \quad \sigma_i = \sigma \\ \text{variance testing:} & \mu_i = \mu, & \text{and} \quad \sigma_i \neq \sigma. \end{array} \quad (4)$$

By defining $\overline{F}$ as the outlier cdf that exhibits the smallest Kullback–Leibler (KL) divergence from F, a necessary and sufficient condition for $\mathcal{P}_M(S) \xrightarrow{M\to\infty} 0$ to successively identify a small fraction of the outliers is presented in the following theorem. Here, a small fraction refers to a fraction that grows with M at a rate dominated by the growth rate of $M\theta_M$, where θ_M is the probability that a stream is an outlier.

THEOREM 1

The decision error probability $\mathcal{P}_M(S)$ tends to zero in the asymptote of large M if and only if [43]

$$\text{mean testing:} \quad \frac{D(F\,\|\,\overline{F})}{\ln M} > \frac{(1-\sqrt{\varepsilon_M})^2}{\hat{S}+K}, \quad (5)$$

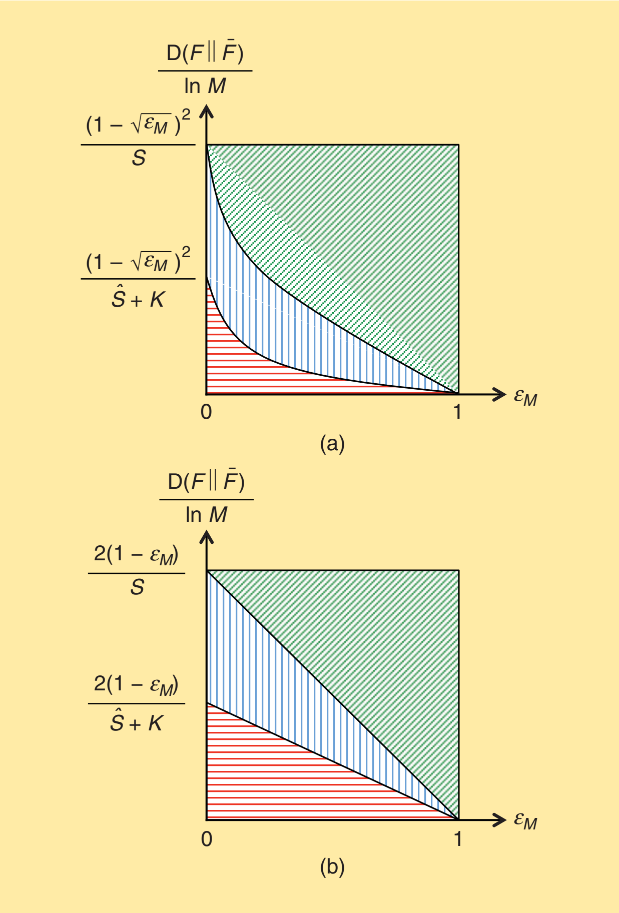
$$\text{variance testing:} \quad \frac{D(F\,\|\,\overline{F})}{\ln M} > \frac{2(1-\varepsilon_M)}{\hat{S}+K}, \quad (6)$$

where $D(\cdot\|\cdot)$ denotes the KL divergence, and $\hat{S}$ is a constant independent of M and determined by the constraints on the cost of sensing $(\hat{S} \approx S\cdot\zeta^{-K})$. Also $\varepsilon_M \in (0,1)$ is defined as

$$\varepsilon_M \triangleq \frac{\ln M\theta_M}{\ln M}, \quad (7)$$

where θ_M is the prior probability that a stream is an outlier.

The necessary and sufficient conditions on $D(F\,\|\,\overline{F})$ in Theorem 1 partition the (D, ε_M) plane into two regions separated by sharp boundaries, as shown in Figure 1. This figure also compares the regions over which the adaptive and nonadaptive procedures are guaranteed to make error-free decisions. Specifically, the diagonally shaded region is the region in which both schemes succeed to detect the T outliers. In the vertically dashed region, however, only the adaptive procedure succeeds and the nonadaptive procedure makes an erroneous decision almost surely, and finally both schemes fail in the horizontally shaded region. It is observed that, depending on the choice of S, the detectability region corresponding to the adaptive



[FIG1] $D(F \| \bar{F})$ versus the prior likelihood ε_M for Gaussian distributions with (a) a different mean and (b) different variance values.

procedure can be substantially larger than that corresponding to the nonadaptive procedure.

It is noteworthy that as long as the objective is to identify a small but prominent fraction of the outliers, the conditions given in (5) and (6) do not depend on the exact number of streams to be identified. This is due to the asymptotic nature of the results, which is dominantly shaped by the regime of interest (small fraction) and the precise number of the outliers has a vanishing effect as M grows. More general necessary and sufficient conditions for identifying any desired fraction of the outliers and with arbitrary distributions for the typical and outlier data streams are provided in [46].

QUICKEST SEARCH FOR ONE OUTLIER

In this subsection we discuss a special scenario of partial recovery of the outliers, in which the objective is to identify only one outlier. While the optimal sequential strategy for solving this problem, as discussed in the section "Data-Adaptive Sampling," is known, by imposing reasonable structures in sensing, some optimality properties can be ensured as $M \to \infty$. Specifically, when the sampling strategy is constrained to

1) observe only one data stream at a time, i.e., $|\mathcal{L}_t| = 1$ for all $t \in \{1, \ldots, \tau\}$
2) once a data stream is discarded after an exploration action, it will be discarded permanently, and the next stream to be examined will be selected randomly from the ones that remain
3) outliers have identical distributions denoted by $\overline{F}$,

the quickest search for detecting an outlier can be restated as

$$\begin{array}{ll} \min_{\tau,\overline{\psi}_\tau,\overline{\mathcal{L}}_\tau} & \mathbb{E}[\tau] \\ \text{s.t.} & \mathrm{P}_M(\tau,\overline{\psi}_\tau,\overline{\mathcal{L}}_\tau) \le \rho. \end{array} \tag{8}$$

The sequential and data-adaptive sampling strategy that optimizes the above tradeoff between the average number of measurements and the decision quality (false alarm probability) is the cumulative sum (CUSUM) test [47]. In this test, one of the sequences is selected at random and measurements are taken from this sequence sequentially. After taking each sample and given all the information accumulated, the likelihood that the sequence under scrutiny is an outlier is updated. If this likelihood exceeds a certain threshold π_U, the sequence is declared an outlier; if it falls below a certain threshold π_L it is discarded permanently and another sequences will be selected to test; and if the likelihood remains within the interval $[\pi_L, \pi_U]$ another sample is taken from the same sequence. By defining π_t as the likelihood that the sequence observed at time t is an outlier given the information accumulated up to time t, the details of the optimal sampling strategy are presented in Table 2 with its optimality established by the following theorem.

THEOREM 2

The optimal stopping time for the quickest search problem in (8) is [48]

$$\tau = \inf\{t : \pi_t > \pi_U\},$$

and the optimal sampling strategy at time t switches to a new sequence if $\pi_t < \pi_L$. The thresholds π_L and π_U are determined uniquely as functions of ζ and the observation cdfs.

GROUP SAMPLING

Motivated by the insights gained from partial recovery of outliers, i.e., rough measurements can be sufficient for eliminating a substantial fraction of the typical streams through the exploration process, we next discuss the idea of group sampling, aiming at basing some of the decisions on even rougher measurements. Group sampling is facilitated by the possibility of taking samples that are combined measurements from multiple sequences. The ultimate objective of such measurements is to expedite the process of exploration and reduce the dimension of the search space with fewer measurements.

A central principle in designing the observation action in the previous section was that at any given time t, one measurement is taken from each data stream included in $\mathcal{L}_t$. In this section, in contrast, we consider two types of samples: coarse and fine samples, which bear information with different qualities. Coarse samples are constructed by linearly combining simultaneous measurements from a group of data streams. While such coarse measurements are not informative for identifying the outliers, they can often be informative enough to discard a group of typical data streams altogether, especially when M is very large and the outliers occur very rarely. When such coarse samples are not sufficient to discard a block, or the block is deemed to contain an outlier, then the data streams constituting the blocks are measured individually via fine samples to refine the information about the status of the data streams within the block. Inclusion of coarse measurements reduces the required sampling budget.

Taking such coarse samples in some applications has a natural interpretation. For instance, in wideband spectrum sensing in which the majority of the channels are occupied and a mobile radio is interested in identifying rare spectrum opportunities (abstracted as outliers), due to the broadcast nature of the wireless channels, any measurement taken by the interested party is a linear superposition of the measurements that it can take from the channels individually via appropriate filtration.

For this purpose, we divide the data streams into blocks of size ℓ and take one sample that is a linear combination of ℓ measurements from the data streams. Such block sampling has, broadly, a twofold effect. On one hand, it takes only one sample for accumulating information about the ℓ sequences and is substantially smaller than the resources needed by the existing approaches that devote at least one sample to each sequence. On the other hand, one combined and aggregated sample is less informative about the status of the individual sequences in comparison to having ℓ different samples. To benefit from the advantage (reduction in sampling rate) and avoid its undesired effects (inaccurate information) these combined samples are used only to obtain some rough confidence about whether the block of data streams include outliers. When a block is deemed to include only typical data streams the entire block is discarded. Alternatively, if the block is deemed to include an outlier, then the block is retained for further scrutiny through more refined (fine) measurements.

QUICK SEARCH FOR A SUBSET OF OUTLIERS VIA GROUP SAMPLING

We define $r \triangleq M/\ell$ to be the number of blocks and without loss of generality we define

$$\mathcal{G}_i \triangleq \{(i-1)\,\ell+1, \dots, i\ell\} \tag{9}$$

to be the set of the data streams grouped in the ith block for $i \in \{1, \dots, r\}$. With the ultimate objective of identifying T outliers the proposed sampling procedure is initiated by taking coarse samples from all groups $\mathcal{G}_1, \dots, \mathcal{G}_r$. Based on these coarse observations a fraction of the groups that are least likely to contain outliers are discarded and the rest are retained for more accurate scrutiny. Repeating this procedure successively refines the search support and progressively focuses the observations on the more promising blocks. More specifically, at each time the sampling procedure selects a subset of the blocks $\{\mathcal{G}_1, \dots, \mathcal{G}_r\}$ and takes one coarse sample from each of these blocks. Upon collecting these measurements, it takes one of the following actions:

[TABLE 2] THE QUICKEST SEARCH FOR ONE OUTLIER.

```
1)   INITIALIZE $t = 0$, $\phi_1 = 1$, $\pi_L$, AND $\pi_U$
2)   $t \leftarrow t+1$
3)   SET $\mathcal{L}_t = \{\phi_t\}$
4)   TAKE ONE SAMPLE FROM $\mathcal{L}_t$
5)   UPDATE $\pi_t$
6)   IF $\phi_t \le \pi_L$
7)      $\phi_{t+1} = \phi_t + 1$; GO TO 2
8)   ELSE IF $\pi_L < \pi_t < \pi_U$
9)      $\phi_{t+1} = \phi_t$; GO TO 2
10)  END IF
11)  SET $\tau = t$; OUTPUT THE SEQUENCE $\phi_t$
```

- *Observation:* Following the spirit of the generic observation action defined earlier, this action is taken in case of lack of sufficient confidence for deciding whether the blocks under scrutiny contain outliers.
- *Exploration:* There is sufficient confidence that some of the blocks are very unlikely to contain an outlier; discard a portion of the groups with the highest likelihoods of containing only typical data streams. This step can be designed similarly to the adaptive sampling procedure in Table 1.
- *Coarse sampling termination:* There is sufficient confidence that the blocks retained contain outliers; stop coarse sampling and start taking fine samples and perform SPRTs on individual sequences until an outlier is identified. If, after performing SPRTs on all sequences in the block, none is identified as an outlier, the sampling procedure resets by moving to the next block and starts taking coarse samples.

After terminating coarse sampling, the retained data streams contain a substantially more condensed proportion of outliers to typical data streams. When the block length $\ell > 1$ and the exploration action are designed carefully, while enjoying the same sensing budgets, adaptive group sampling yields a more reduced dimension for the search space compared with the adaptive procedure of the section "Quick Search for a Subset of Outliers" (i.e., $\ell = 1$). Similar to the mean and variance testing problems for partial recovery of the outliers presented in the section "Quick Search for a Subset of Outliers," the following theorem presents a necessary and sufficient condition for $\mathcal{P}_M(S) \xrightarrow{M \to \infty} 0$, for $\mathcal{P}_M(S)$ defined in (3), to successively identify a small fraction of the outliers. $\overline{F}$ denotes the outlier cdf that minimizes the KL divergence from F.

THEOREM 3

For fixed block size ℓ, the decision error probability $\mathcal{P}_M(S)$ tends to zero in the asymptote of large M if and only if [49]

$$\text{mean testing:} \quad \frac{D(F \,\|\, \overline{F})}{\ln M} > \frac{(1-\sqrt{\varepsilon_M})^2}{\ell(\hat{S}+K)},$$

$$\text{variance testing:} \quad \frac{D(F \,\|\, \overline{F})}{\ln M} > \frac{2(1-\varepsilon_M)}{\ell(\hat{S}+K)},$$

where $\hat{S}$ and $\varepsilon_M \in (0,1)$ are defined in Theorem 1.

This result indicates that as $M \to \infty$, the region of outliers that are undetectable by the adaptive procedure delineated by (5) and

(6) and depicted in Figure 1 is further shrunk by a factor of ℓ through group sampling.

QUICKEST SEARCH FOR ONE OUTLIER VIA GROUP SAMPLING

Similarly to the partial outlier recovery scenario, the quickest search approach of the section "Quickest Search for One Outlier" for identifying one outlier can be further extended by accommodating group sampling into the sampling strategy.

In the simplest scenario, the sequences can be bundled into groups of size $\ell = 2$ and the combined measurements taken will be the sum of two independent samples from each sequence. This leads to three possibilities for the distribution of the combined measurement. The sampling strategy is initiated by selecting a bundle at random and taking a mixed measurement from that sample and follows, in spirit, the same steps as the quickest search procedure in the section "Quickest Search for One Outlier." Specifically, when there is sufficient confidence that the group does not contain an outlier, the block is discarded; when there is a lack of confidence for making any reliable inference about the block, one more mixed sample is taken; and when there exists sufficient confidence that the block contains an outlier, taking combined measurements is terminated, and then the sequences contained in the block are examined individually to identify an outlier.

Designing the optimal sampling strategy involves characterizing two optimal stopping times, one corresponding to the terminal time of taking combined measurements, and the second one corresponding to reaching a decision for individual sequences after taking combined measurements is terminated. An effective approach for identifying these stopping times is proposed in [50], where a CUSUM test is applied to the sequence blocks to find a promising block, and then SPRTs are applied on the individual sequences to reach decisions about their underlying distributions.

UNIVERSAL OUTLYING SEQUENCE DETECTION

Depending on the underlying application, the underlying statistical models of the data streams might or might not be known. Whether the distributions of both typical and outlier sequences are known, only one is known, or both are unknown, outlier detection approaches can take drastically different structures. Representative examples are spectrum sensing in congested wideband channels as a case in which both distributions can be known (spectrum holes are the outliers) and fraud detection as a case in which either the outlier (fraud) or both distributions are unknown. When the statistics are fully known strategies that balance the interplay among different measures optimally can be characterized optimality according to the abstraction given in (2). These optimal strategies can be shown to be exponentially consistent and all the observation, exploration, and detection actions have likelihood-ratio-like structures [43].

When there exist uncertainties associated with the descriptions of the statistical models, the outlying sequence detection problem is related to general composite hypothesis testing problems, for which the generalized likelihood principle, which exhibits certain asymptotic optimality properties [51]–[53], is a popular solution. Universal outlying sequence detection is also closely related to homogeneity testing and classification [51], [54]–[58]. In homogeneity testing, one wishes to decide whether or not two samples come from the same probability law. In classification problems, a set of test data is classified to one of multiple streams of training data with distinct labels.

In this section, we investigate the effects of uncertainties about the statistics of the outliers and discuss a universal approach for identifying outliers in which, besides the premise that the outliers follow a distribution distinct from that governing the typical data streams, no knowledge of their statistics is assumed [59]. To focus the attention on the effects of unknown statistics, we mainly consider a simple setting in which it assumed that

1) only one data stream is an outlier and the remaining $M-1$ ones are typical
2) we have access to n samples from each data stream
3) the samples belong to a finite set $\mathcal{Y}$.

Under the hypothesis that the ith coordinate is the outlier, the joint distribution of all the observations (i.e., the likelihood function) is

$$p_i(y^{Mn}) = p_i(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}) = \prod_{t=1}^{n} \{\bar{f}(y_t^{(i)}) \prod_{j \neq i} f(y_t^{(j)})\}, \tag{10}$$

where

$$\boldsymbol{y}^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)}), i = 1, \ldots, M,$$

and $\bar{f}$ and f denote the probability mass functions (pmfs) of the outlier and typical streams, respectively.

For a universal detection rule $\delta : \mathcal{Y}^{Mn} \rightarrow \{1, \ldots, M\}$, which is not allowed to depend on f and $\bar{f}$, the maximal error probability, which will be a function of the test and $(\bar{f}, f)$, is

$$e(\delta, f, \bar{f}) \triangleq \max_{i=1,\ldots,M} \sum_{y^{Mn}:\, \delta(y^{Mn}) \neq i} p_i(y^{Mn}), \tag{11}$$

with the corresponding error exponent, denoted by

$$\alpha(\delta, f, \bar{f}) \triangleq \lim_{n \rightarrow \infty} -\frac{1}{n} \log e(\delta, f, \bar{f}). \tag{12}$$

We consider the error exponent as n goes to infinity, while M, and hence the number of hypotheses, is kept fixed. Consequently, the error exponent in (12) also coincides with the one for the average probability of error.

A test is termed *universally consistent* if $e(\delta, f, \bar{f}) \rightarrow 0$ for any $(\bar{f}, f)$, $\bar{f} \neq f$ as $n \rightarrow \infty$. It is termed *universally exponentially consistent* if $\alpha(\delta, f, \bar{f}) > 0$.

UNIVERSAL TEST

For each $i = 1, \ldots, M$, denote the empirical distribution of $\boldsymbol{y}^{(i)}$ by γ_i. When f is known and $\bar{f}$ is unknown, we compute the likelihood for outlier hypothesis i by replacing $\bar{f}$ in (10) with its maximum likelihood (ML) estimate $\widehat{\bar{f}}_i \triangleq \gamma_i$, as

$$L_i^{\text{typ}}(y^{Mn}) = \prod_{t=1}^{n} \{\widehat{\bar{f}}_i(y_t^{(i)}) \prod_{j \neq i} f(y_t^{(j)})\}. \tag{13}$$

Similarly, when neither $\bar{f}$ nor f is known, we compute the likelihood for outlier hypothesis i by replacing the $\bar{f}$ and f in (10) with their ML estimates $\hat{\bar{f}}_i \triangleq \gamma_i$, and $\hat{f}_i \triangleq \left(\sum_{j \neq i} \gamma_j\right)/(M-1)$, as

$$L_i^{\text{univ}}(y^{Mn}) = \prod_{t=1}^{n} \{\hat{\bar{f}}_i(y_t^{(i)}) \prod_{j \neq i} \hat{f}_i(y_t^{(j)})\}. \tag{14}$$

Finally, we decide upon the coordinate with the largest likelihood to be the outlier. Using (13) and (14), our universal tests in the two cases can be described respectively as

$$\delta^{\text{typ}}(y^{Mn}) = \underset{i=1,\ldots,M}{\operatorname{argmax}}\, L_i^{\text{typ}}(y^{Mn}), \tag{15}$$

when only f is known, and

$$\delta^{\text{univ}}(y^{Mn}) = \underset{i=1,\ldots,M}{\operatorname{argmax}}\, L_i^{\text{univ}}(y^{Mn}), \tag{16}$$

when neither $\bar{f}$ nor f is known.

RESULTS

Our results will be stated in terms of a distance metric between a pair of pmfs $p, q \in \mathcal{P}(\mathcal{Y})$ called the Bhattacharyya distance, which is related to the Chernoff information (see, e.g., [60]), defined as

$$B(p,q) \triangleq -\log\Big(\sum_{y \in \mathcal{Y}} p(y)^{\frac{1}{2}} q(y)^{\frac{1}{2}}.\Big) \tag{17}$$

Our first theorem for models with one outlier characterizes the optimal exponent for the maximal error probability when both $\bar{f}$ and f are known, and when only f is known.

THEOREM 4

When $\bar{f}$ and f are both known, the optimal exponent for the maximal error probability is equal to [59]

$$2B(\bar{f}, f). \tag{18}$$

Furthermore, the error exponent in (18) is achievable by a test that uses only the knowledge of f. In particular, such a test is our proposed test in (15).

Consequently, in the completely universal setting, when nothing is known about $\bar{f}$ and f except that $\bar{f} \neq f$, and both $\bar{f}$ and f have full supports, it holds that for any universal test δ,

$$\alpha(\delta, f, \bar{f}) \leq 2B(\bar{f}, f). \tag{19}$$

Given the second assertion in Theorem 4, it might be tempting to think that it would be possible to design a test to achieve the optimal error exponent of $2B(\bar{f}, f)$ universally when neither $\bar{f}$ nor f is known. A counterexample given in [59] shows that this is not possible. This motivates us to seek instead a test that yields just a positive (no matter how small) error exponent $\alpha(\delta, f, \bar{f}) > 0$ for every $\bar{f}$ and f, $\bar{f} \neq f$, i.e., a test that achieves universally exponential consistency. Without knowing either $\bar{f}$ or f, it is not clear at the outset that even this lesser objective can be met. One of the main contributions in [59] is to show that the proposed universal test in (16) is indeed universally exponentially consistent for every fixed M.

THEOREM 5

For every pair $\bar{f} \neq f$

$$\alpha(\delta^{\text{univ}}, f, \bar{f}) = \min_{q_1,\ldots,q_M} D(q_1 \,\|\, \bar{f}) + \ldots + D(q_M \,\|\, f),$$

where the minimum is over the set of $(q_1, \ldots, q_M)$ such that

$$\sum_{j \neq 1} D\left(q_j \middle\| \frac{\sum_{k \neq 1} q_k}{M-1}\right) \geq \sum_{j \neq 2} D\left(q_j \middle\| \frac{\sum_{k \neq 2} q_k}{M-1}\right). \tag{20}$$

It can be shown that the solution $\alpha(\delta^{\text{univ}}, f, \bar{f}) > 0$ [59].

Note that for any fixed $M \geq 3$ and $\theta > 0$, regardless of which coordinate is the outlier, it holds that the random empirical distributions $(\gamma_1, \ldots, \gamma_M)$ satisfy

$$\lim_{n \to \infty} \mathbb{P}_i\Big\{ \Big\| \frac{1}{M}\sum_{j=1}^{M} \gamma_j - \Big(\frac{1}{M}\bar{f} + \frac{M-1}{M} f\Big)\Big\|_1 > \theta \Big\} = 0, \tag{21}$$

where $\|\cdot\|_1$ denotes the 1-norm of the argument distribution. Since $(1/M)\bar{f} + (M-1/M)f \to f$ as $M \to \infty$, heuristically speaking, a consistent estimate of the typical distribution can readily be obtained asymptotically in M at the outset from the entire observation set before deciding upon which coordinate is the outlier. This observation and the second assertion of Theorem 4 motivate a study of the asymptotic performance (achievable error exponent) of δ^{univ} when $M \to \infty$ (after having taken the limit as n goes to infinity).

THEOREM 6

For each $M \geq 3$

$$\alpha(\delta^{\text{univ}}, f, \bar{f}) \geq \min_{\substack{q \in \mathcal{P}(\mathcal{Y}) \\ D(q \| f) \leq \frac{1}{M-1}(2B(\bar{f}, f) + C_f)}} 2B(\bar{f}, q), \tag{22}$$

where $C_f \triangleq -\log(\min_{y \in \mathcal{Y}} f(y)) < \infty$ by the fact that f has a full support [59].

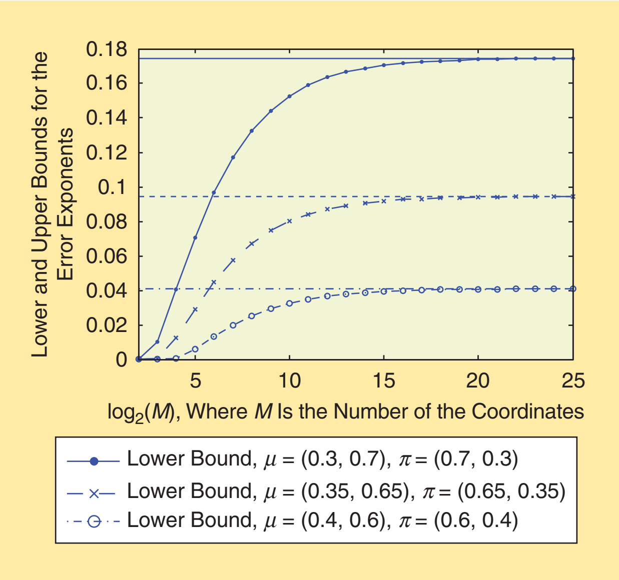
The lower bound on the error exponent in (22) is nondecreasing in $M \geq 3$. Furthermore, as $M \to \infty$, this lower bound converges to the optimal error exponent $2B(\bar{f}, f)$; hence, our test is asymptotically optimal:

$$\lim_{M \to \infty} \alpha(\delta^{\text{univ}}, f, \bar{f}) = 2B(\bar{f}, f), \tag{23}$$

which from Theorem 4 is equal to the optimal error exponent when both $\bar{f}$ and f are known.

Example 1

We now provide some numerical results for an example with $\mathcal{Y} = \{0, 1\}$. Specifically, the three plots in Figure 2 are for three pairs of outlier and typical distributions being $\bar{f} = (p(0) = 0.3, p(1) = 0.7)$, $f = (0.7, 0.3)$; $\bar{f} = (0.35, 0.65)$, $f = (0.65, 0.35)$; and $\bar{f} = (0.4, 0.6)$, $f = (0.6, 0.4)$, respectively. Each horizontal line corresponds to $2B(\bar{f}, f)$, and each curved line corresponds to the lower bound in (22) for the error exponent achievable by δ^{univ}. As



[FIG2] An illustration of the asymptotic optimality of δ^{univ}.

shown in these plots, the lower bounds converge to $2B(\bar{f}, f)$ as $M \to \infty$, i.e., δ^{univ} is asymptotically optimal for all three pairs $\bar{f}, f$.

MODELS WITH AT MOST ONE OUTLIER

A natural question that arises at this point is what would happen if it is also possible that no outlier is present? To answer this question, we now consider models that append an additional null hypothesis with no outlier to the set of possible hypotheses. In particular, under the null hypothesis, the likelihood function is given by

$$p_0(y^{Mn}) = \prod_{t=1}^{n} \prod_{i=1}^{M} f(y_t^{(i)}).$$

A universal test $\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}$ will now also accommodate an additional decision for the null hypothesis. Correspondingly, the maximal error probability is now computed with the inclusion of the null hypothesis according to

$$e(\delta, f, \bar{f}) \triangleq \max_{i=0,1,\ldots,M} \sum_{y^{Mn} : \delta(y^{Mn}) \neq i} p_i(y^{Mn}).$$

With just one additional null hypothesis, contrary to the previous models with one outlier, it becomes impossible to achieve universal exponential consistency even with the knowledge of the typical distribution. This pessimistic result reaffirms that our previous finding that universal exponential consistency is attained for the models with one outlier is indeed quite surprising.

PROPOSITION 1

For the setting with the additional null hypothesis, there cannot exist a universally exponentially consistent test even when the typical distribution is known [59].

In typical applications such as environment monitoring and fraud detection, the consequence of a missed detection of the outlier can be much more catastrophic than that of a false positive. In addition, Proposition 1 tells us that there cannot exist a universal test that yields exponential decays for both the conditional probability of false positive (under the null hypothesis) and the conditional probabilities of missed detection (under all nonnull hypotheses). Consequently, it is natural to look for a universal test fulfilling a lesser objective: attaining universal exponential consistency for conditional error probabilities under only all the nonnull hypotheses, while seeking only universal consistency for the conditional error probability under the null hypothesis. We now show that such a test can be obtained by slightly modifying our earlier test. Furthermore, in addition to achieving universal consistency under the null hypothesis, this new test achieves the same exponent as in (20) in Theorem 5 for the conditional error probabilities under all nonnull hypotheses.

In particular, we modify our previous test in (16) to allow for the possibility of deciding for the null hypothesis as:

$$\delta^{\text{null}}(y^{Mn}) = \begin{cases} \arg\max\limits_{i=1,\ldots,M} \; L_i^{\text{univ}}(y^{Mn}) & \text{if } \max\limits_{j \neq k} \dfrac{L_j^{\text{univ}}(y^{Mn})}{L_k^{\text{univ}}(y^{Mn})} > \lambda_n \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

where $\lambda_n = O(n)$.

THEOREM 7

For every pair of distributions $\bar{f}, f, \bar{f} \neq f$, the test in (24) yields a positive exponent for the conditional probability of error under every nonnull hypothesis $i = 1, \ldots, M$, and a vanishing conditional probability of error under the null hypothesis [59]. In particular, the achievable error exponent under every nonnull hypothesis is the same as that given in (20).

Furthermore, as $M \to \infty$, the test in (24) is asymptotically optimal under each of the nonnull hypotheses, i.e.,

$$\lim_{M \to \infty} \lim_{n \to \infty} -\frac{1}{n} \log(\mathbb{P}_i\{\delta \neq i\}) = 2B(\bar{f}, f), \tag{25}$$

while also yielding that

$$\lim_{n \to \infty} \mathbb{P}_0\{\delta \neq 0\} = 0.$$

EXTENSION TO MULTIPLE OUTLIERS

The aforementioned results on universal outlying sequence detection can be extended to the setting with more than one outlier [59]:

- For the setting with a known number of distinctly distributed outliers, we can construct a universally exponentially consistent test using the generalized likelihood principle as in the section "Universal Test." A key difference from the single outlier case is that the error exponent when both the outlier and typical distributions are known can be larger than that when only the typical distribution is known.
- For the setting with a known number of identically distributed outliers, the error exponent when both the outlier and typical distributions are known is equal to that when only the typical distribution is known, which is equal to $2B(\bar{f}, f)$ (the same as for the case of a single outlier). Furthermore, the universally exponentially consistent test when both the outlier

and typical distributions are unknown is asymptotically optimum as $M \to \infty$ (with the number of outliers fixed) in that its error exponent is also equal to $2B(\bar{f}, f)$.

- For the setting with an unknown number of identically distributed outliers, we construct a test based on modified application of the generalized likelihood principle to achieve a positive error exponent under each nonnull hypothesis, and also consistency under the null hypothesis universally.
- When the outliers can be distinctly distributed (with their total number being unknown), it can be shown that a universally exponentially consistent test cannot exist, even when the typical distribution is known and the null hypothesis is excluded.

CONCLUDING REMARKS

In this article, we have discussed the problem of identifying outlying sequences from a large pool of sequences that is populated by typical sequences. By crafting the problem as a natural dichotomous hypothesis testing problem, we have discussed three general classes of strategies for outlying sequence detection based on different detection objectives and available information about the statistics of the outliers. In this class, we have discussed sequential data-adaptive approaches in which there is no prespecified order for making measurements from the sequences, and the sampling decisions are made dynamically at each time and based on the information accumulated up to that time. Depending on whether one is interested in identifying all outlying sequences, a fraction of them, or only one of them, the data-adaptive sampling strategies exhibit different structures. An important insight one gains from these approaches is that if the objective is not identifying all outliers, incorporating an exploration stage, which uses rough observations to reduce the dimension of the data set with more condensed proportion of outliers, translates into substantial reduction in the cost of sensing. Motivated by this insight, in the second class of approaches we have discussed the notion of group sampling, in which the sequences are split into groups and in the exploration stage the sequences are not measured individually, but instead, rough observations in the form of combined measurements from sequence groups are made. Finally, in the third class, we have investigated the effects of uncertainties about the statistics of the outliers and have discussed a universal approach for identifying outliers, in which besides the premise that the outliers follow a distribution distinct from that governing the typical data streams, no knowledge of their statistics is assumed. Our generalized likelihood approach was based on using the empirical distributions of the data streams. A recent study [61] adopts an alternative kernel-based test, which applies the metric of maximum mean discrepancy that measures the distance between embeddings of distributions into a reproducing kernel Hilbert space. We further note that in our discussion of universal outlying sequence detection, we have restricted attention to the fixed sample size setting in which every sequence is sampled at every time step. Extending the study of universal outlying sequence detection to the sequential and adaptive sampling settings is a challenging open area of research that is worthy of pursuit.

ACKNOWLEDGMENTS

The work of Ali Tajer was supported by the National Science Foundation (NSF) under grant ECCS-1343326. The work of Venugopal V. Veeravalli was supported by the Air Force Office of Scientic Research under grant FA9550-10-1-0458 through the University of Illinois at Urbana-Champaign, by the U.S. Defense Threat Reduction Agency through subcontract 147755 at the University of Illinois from prime award HDTRA1-10-1-0086, and by the NSF under grant NSF CCF 11-11342. The work of H. Vincent Poor was supported by the NSF under grants DMS-1118605 and ECCS-1343210.

AUTHORS

Ali Tajer (tajer@ecse.rpi.edu) received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology in 2002 and 2004, respectively, and the M.A. and Ph.D. degrees from Columbia University in 2010 in statistics and electrical engineering, respectively. He was with Princeton University from 2010 to 2012 as a postdoctoral research associate. He is currently an assistant professor of electrical, computer, and systems engineering at Rensselaer Polytechnic Institute. His research interests include estimation and detection theory, network information theory, wireless communications, and smart grids.

Venugopal V. Veeravalli (vvv@illinois.edu) received the B. Tech. degree (Silver Medal Honors) from the Indian Institute of Technology, Bombay, in 1985, the M.S. degree from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 1987, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, in 1992, all in electrical engineering. He joined the University of Illinois at Urbana-Champaign in 2000, where he is currently a professor in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Information Trust Institute. His research interests include detection and estimation theory, information theory, sensor networks, and wireless communication.

H. Vincent Poor (poor@princeton.edu) is the dean of engineering and applied science at Princeton University, where he is also the Michael Henry Strater University Professor of Electrical Engineering. His interests include the areas of statistical signal processing, stochastic analysis, and information theory, with applications in wireless networks and related fields. Among his publications is the recent book *Mechanisms and Games for Dynamic Spectrum Allocation* (Cambridge, 2014). He is an IEEE Fellow and a member of the National Academy of Engineering, the National Academy of Sciences, and the Royal Society. His recent recognitions include the 2011 IEEE Signal Processing Society Award and the 2014 URSI Booker Gold Medal.

REFERENCES

[1] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Hoboken, NJ: Wiley, 2003.

[2] B. Pierce, "Criterion for the rejection of doubtful observations," *Astron. J.*, vol. 2, no. 21, pp. 161–163, July 1852.

[3] W. Chauvenet, *A Manual of Spherical and Practical Astronomy*, 5th ed. New York: Dover, 1960, vol. II.

[4] R. B. Dean and W. J. Dixon, "Simplified statistics for small numbers of observations," *Anal. Chem.*, vol. 24, no. 4, pp. 636–638, Apr. 1951.

[5] V. Barnett and T. Lewis, *Outliers in Statistical Data*. New York: Wiley, 1978.

[6] M. Gupta, J. Gao, C. Aggarwal, and J. Han, *Outlier Detection for Temporal Data*. San Rafael, CA: Morgan and Claypool, 2014.

[7] P.-N. Tand, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Reading, MA: Addison-Wesley, 2005, ch. 2, pp. 19–96.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2000.

[9] D. Dasgupta and N. Majumdar, "Anomaly detection in multidimensional data using negative selection algorithm," in *Proc. IEEE Conf. Evolutionary Computation*, Hawaii, 2002, pp. 1039–1044.

[10] D. Endler, "Intrusion detection applying machine learning to solaris audit data," in *Proc. 14th Annu. Computer Security Applications Conf.*, 1998, pp. 268–279.

[11] A. K. Gosh, J. Wanken, and F. Charron, "Detecting anomalous and unknown intrusions against programs," in *Proc. 14th Annu. Computer Security Applications Conf.*, 1998, pp. 259–267.

[12] A. Ghosh, A. Schwartzbard, and M. Schatz, "A study in using neural networks for anomaly and misuse detection," in *Proc. 8th Conf. USENIX Security Symp.*, 1999, pp. 12–23.

[13] D.-K. Kang, D. Fuller, and V. Honavar, "Learning classifiers for misuse detection using a bag of system calls representation," in *Proc. 3rd IEEE Int. Conf. Intelligence and Security Informatics*, 2005, pp. 511–516.

[14] D. Dasgupta and F. Nino, "A comparison of negative and positive selection algorithms in novel pattern detection," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, Nashville, TN, 2000, pp. 125–130.

[15] T. Lane and C. Brodley, "Sequence matching and learning in anomaly detection for computer security," in *Proc. AAAI Workshop: AI Approaches to Fraud Detection and Risk Management*, 1997, pp. 43–49.

[16] S. Budalakoti, A. Srivastava, R. Akella, and E. Turkov, "Anomaly detection in large sets of high-dimensional symbol sequences," NASA Ames Research Center, Tech. Rep. TM-2006-214553, 2006.

[17] S. Budalakoti, A. N. Srivastava, and M. E. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Trans. Syst., Man, Cybern. C*, vol. 39, no. 1, pp. 101–113, Jan. 2009.

[18] V. Chandola, V. Mithal, and V. Kumar, "A comparative evaluation of anomaly detection techniques for sequence data," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 743–748.

[19] K. Sequeira and M. Zaki, "ADMIT: anomaly-based data mining for intrusions," in *Proc. 8th ACM Int. Conf. Knowledge Discovery and Data Mining*, 2002, pp. 386–395.

[20] C. Marceau, "Characterizing the behavior of a program using multiple-length N-grams," in *Proc. Workshop New Security Paradigms*, 2000, pp. 101–110.

[21] C. C. Michael and A. Ghosh, "Two state-based approaches to program-based anomaly detection," in *Proc. 16th Annu. Computer Security Applications Conf.*, 2000, pp. 21–30.

[22] E. Eskin, W. Lee, and S. Stolfo, "Modeling system calls for intrusion detection with dynamic window sizes," in *Proc. DARPA Information Survivability Conf. Expo. II*, 2001, pp. 165–175.

[23] G. Florez-Larrahondo, S. M. Bridges, and R. Vaughn, "Efficient modeling of discrete events for anomaly detection using hidden Markov models," in *Proc. 8th Int. Conf. Information Security*, 2005, pp. 506–514.

[24] B. Gao, H.-Y. Ma, and Y.-H. Yang, "HMMs (Hidden Markov Models) based on anomaly intrusion detection method," in *Proc. Int. Conf. Machine Learning and Cybernetics*, 2002, pp. 381–385.

[25] X. Li and J. Han, "Mining approximate top-*K* subspace anomalies in multidimensional time-series data," in *Proc. 33rd Int. Conf. Very Large Databases*, 2007, pp. 447–458.

[26] N. D. Le, R. D. Martin, and A. E. Raftery, "Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models," *J. Am. Stat. Assoc.*, vol. 91, no. 436, pp. 1504–1515, 1996.

[27] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: Optimum tests and applications," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4215–4229, July 2012.

[28] D. M. Hawkins, *Identification of Outliers*. London, U.K.: Chapman & Hall, 1980.

[29] R. J. Bolten and D. J. Hand, "Unsupervised profiling methods for fraud detection," in *Proc. Credit Scoring and Credit Control Conf.*, vol. VII, Edinburgh, Scotland, 2001, pp. 3–5.

[30] K. I. Penny and I. T. Jolliffe, "A comparison of multivariate outlier detection methods for clinical laboratory safety data," *Statistician*, vol. 50, no. 3, pp. 295–308, 2001.

[31] C. Brownlees and G. Gallo, "Financial econometric analysis at ultra-high frequency: Data handling concerns," *Computat. Stat. Data Anal.*, vol. 51, no. 4, pp. 2232–2245, Dec. 2006.

[32] R. M. Alvarez, S. D. Hyde, and T. E. Hall, Eds., *Election Fraud: Detecting and Deterring Electoral Manipulation*, ser. Brookings Series on Election Administration and Reform. Washington, DC: Brookings Institution, 2008.

[33] A. Tajer, R. Castro, and X. Wang, "Adaptive sensing of congested spectrum bands," *IEEE Trans. Inform. Theory*, vol. 58, no. 9, pp. 6110–6125, Sept. 2012.

[34] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proc. 3rd SIAM Int. Conf. Data Mining*, San Francisco, CA, May 2003, pp. 25–36.

[35] M. D. Goldberg, Y. Qu, L. M. McMillin, W. Wolf, L. Zhou, and M. Divakarla, "AIRS near-real-time products and algorithms in support of operational numerical weather prediction," *IEEE Trans. Geosci. Remote Sensing*, vol. 41, no. 2, pp. 379–389, Feb. 2003.

[36] S. Wanga, W. A. Woodward, H. L. Grayb, S. Wiecheckib, and S. R. Sain, "A new test for outlier detection from a multivariate mixture distribution," *J. Computat. Graph. Stat.*, vol. 6, no. 3, pp. 285–299, 1997.

[37] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15.1–15.58, July 2009.

[38] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Stat.*, vol. 19, no. 3, pp. 326–339, 1948.

[39] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problemy Peredachi Informastii*, vol. 10, no. 3, pp. 51–61, 1974.

[40] A. Korostelev, "On minimax rates of convergence in image models under sequential design," *Stat. Probability Lett.*, vol. 43, no. 4, pp. 369–375, July 1999.

[41] R. M. Castro and R. D. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.

[42] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6222–6235, Sept. 2011.

[43] A. Tajer and H. V. Poor, "Quick search for rare events," *IEEE Trans. Inform. Theory*, vol. 59, no. 7, pp. 4462–4481, 2013.

[44] A. Tajer and H. V. Poor, "Adaptive sampling for sparse recovery," in *Proc. 4th Workshop on Information Theoretic Methods in Science and Engineering*, Helsinki, Finland, Aug. 2011.

[45] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, 2013.

[46] A. Tajer and H. V. Poor, "Hypothesis testing for partial sparse recovery," in *Proc. 50th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2012, pp. 901–908.

[47] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, UK: Cambridge Univ. Press, 2009.

[48] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 5375–5386, Aug. 2011.

[49] A. Tajer and H. V. Poor, "Quick search for rare events through adaptive group sampling," in *Proc. 47th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2013, pp. 757–761.

[50] J. Geng, W. Xu, and L. Lai, "Quickest search over multiple sequences with mixed observations," in *Proc. IEEE Int. Symp. Information Theory*, Istanbul, Turkey, July 2013, pp. 2582–2586.

[51] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer, 1994.

[52] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?," *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, Sept. 1992.

[53] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Stat.*, vol. 36, no. 2, pp. 369–401, Apr. 1965.

[54] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, no. 1–2, pp. 250–254, July 1911.

[55] O. Shiyevitz, "On Rényi measures and hypothesis testing," in *Proc. IEEE Int. Symp. Information Theory*, July 31–Aug. 5, 2011, pp. 894–898.

[56] J. Unnikrishnan, "On optimal two sample homogeneity tests for finite alphabets," in *Proc. IEEE Int. Symp. Information Theory*, July 1–6, 2012, pp. 2027–2031.

[57] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 278–286, Mar. 1988.

[58] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[59] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," in *Proc. IEEE Int. Symp. Information Theory*, Istanbul, Turkey, July 2013, pp. 2666–2670.

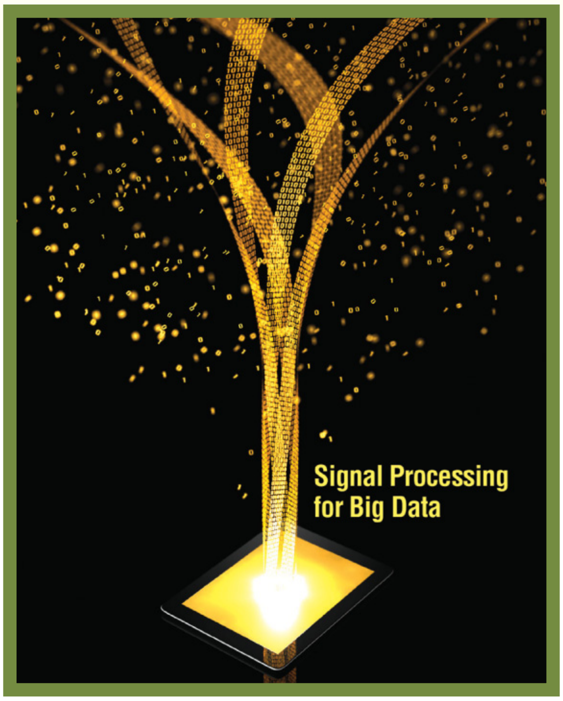
[60] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 2006.

[61] S. Zou, Y. Liang, H. V. Poor, and X. Sh, "Kernel-based nonparametric anomaly detection," in *Proc. IEEE 15th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Toronto, Canada, June 2014.

[SP]

[Nicholas D. Sidiropoulos, Evangelos E. Papalexakis, and Christos Faloutsos]

Parallel Randomly Compressed Cubes



© ISTOCKPHOTO.COM/TA2YO4NORI

[A scalable distributed architecture for big tensor decomposition]

This article combines a tutorial on state-of-the-art tensor decomposition as it relates to big data analytics, with original research on parallel and distributed computation of low-rank decomposition for big tensors, and a concise primer on Hadoop–MapReduce. A novel architecture for parallel and distributed computation of low-rank tensor decomposition that is especially well suited for big tensors is proposed. The new architecture is based on parallel processing of a set of randomly compressed, reduced-size replicas of the big tensor. Each replica is independently decomposed, and the results are joined via a master linear equation per tensor mode. The approach enables massive parallelism with guaranteed identifiability properties: if the big tensor is of low rank and the system parameters are appropriately chosen, then the rank-one factors of the big tensor will indeed be recovered from the analysis of the reduced-size replicas. Furthermore, the architecture affords memory/storage and complexity gains of order (IJ/F) for a big tensor of size $I \times J \times K$ of rank F with $F \leq I \leq J \leq K$. No sparsity is required in the tensor or the underlying latent factors, although such sparsity can be exploited to improve memory, storage, and computational savings.

Digital Object Identifier 10.1109/MSP.2014.2329196
Date of publication: 19 August 2014

INTRODUCTION

Tensors are data structures indexed by three or more indices, say $(i, j, k, \cdots)$, a generalization of matrices, which are data structures indexed by two indices, say (r, c) for (row, column). The term *tensor* has a different meaning in physics, however, it

1053-5888/14©2014IEEE

has been widely adopted across various disciplines in recent years to refer to what was previously known as a *multiway array*. Matrices are two-way arrays, and there are three- and higher-way (or higher-order) tensors.

Tensor algebra has many similarities to but also many striking differences from matrix algebra, e.g., determining tensor rank is NP-hard, and low-rank tensor factorization is unique under mild conditions. Tensor factorizations have already found many applications in signal processing (speech, audio, communications, radar, signal intelligence, and machine learning) and well beyond. For example, tensor factorization can be used to blindly separate unknown mixtures of speech signals in reverberant environments [2], untangle audio sources in the spectrogram domain [3], unravel mixtures of code-division communication signals without knowledge of their spreading codes [4], localize emitters in radar and communication applications [5], detect cliques in social networks [6], and analyze fluorescence spectroscopy data [7], to name a few (see [8] for additional machine-learning applications).

Tensors are becoming increasingly important, especially for analyzing big data, and tensors easily turn really big, e.g., $1{,}000 \times 1{,}000 \times 1{,}000 = 1$ billion entries. Memory issues related to tensor computations with large but sparse tensors have been considered in [9] and [10] and incorporated in the sparse tensor toolbox (http://www.sandia.gov/~tgkolda/TensorToolbox). The main idea in those papers is to avoid intermediate product explosion when computing sequential tensor–matrix (mode) products, but the assumption is that the entire tensor fits in memory (in coordinate-wise representation), and the mode products expand (as opposed to reduce) the size of the core array that they multiply. Adaptive tensor decomposition algorithms for cases where the data is serially acquired (or elongated) along one mode have been developed in [11], but these assume that the other two modes are relatively modest in size. More recently, a divide-and-conquer approach for decomposing big tensors has been proposed in [12]. The idea of [12] is to break the data into smaller boxes that can be factored independently, and the results subsequently concatenated using an iterative process. This assumes that each smaller box admits a unique factorization (which cannot be guaranteed from global uniqueness conditions alone), requires reconciling the different column permutations and scalings of the different blocks, and entails significant communication and signaling overhead.

All of the aforementioned techniques require that the full data be stored in (possibly distributed) memory. Realizing that this is a showstopper for truly big tensors, [6] proposed a random sampling approach, wherein judiciously sampled significant parts of the tensor are independently analyzed, and a common piece of data is used to anchor the different permutations and scalings. The downside of [6] is that it only works for sparse tensors, and it offers no identifiability guarantees—although it usually works well for sparse tensors. A different approach was taken in [13], which proposed randomly compressing a big tensor down to a far smaller one. Assuming that the big tensor admits a low-rank decomposition with sparse latent factors, such a random compression guarantees identifiability of the low-rank decomposition of the big tensor from the low-rank decomposition of the small tensor. This result can be viewed as a generalization of compressed sensing ideas from the linear to the multilinear case. Still, this approach works only when the latent low-rank factors of the big tensor are known to be sparse, and this is often not the case.

This article considers appropriate compression strategies for big (sparse or dense) tensors that admit a low-rank decomposition/approximation, whose latent factors need not be sparse. Latent sparsity is usually associated with membership problems such as clustering and coclustering [14]. A novel architecture for parallel and distributed computation of low-rank tensor decomposition that is especially well suited for big tensors is proposed. The new architecture is based on parallel processing of a set of randomly compressed, reduced-size replicas or the big tensor. Each replica is independently decomposed, and the results are joined via a master linear equation per tensor mode. The approach enables massive parallelism with guaranteed identifiability properties: if the big tensor is indeed of low rank and the system parameters are appropriately chosen, then the rank-one factors of the big tensor will indeed be recovered from the analysis of the reduced-size replicas. Furthermore, the architecture affords memory/storage and complexity gains of order (IJ/F) for a big tensor of size $I \times J \times K$ of rank F with $F \leq I \leq J \leq K$. No sparsity is required in the tensor or the underlying latent factors, although such sparsity can be exploited to improve memory, storage, and computational savings.

This article combines 1) a short tutorial on state-of-the-art tensor decomposition as it relates to big data analytics, 2) novel research results on tensor compression and parallel and distributed tensor decomposition, and 3) a concise primer on Hadoop–MapReduce, starting from a toy signal processing problem, and going up to sketching a Hadoop implementation of a proposed algorithm for tensor decomposition in the cloud. The combination is timely and well motivated given the emerging interest in (and relative scarcity of literature on) signal processing for big data analytics, and in porting/translating and developing new signal processing algorithms for cloud computing platforms.

NOTATION

A scalar is denoted by an italic letter, e.g., a. A column vector is denoted by a bold lowercase letter, e.g., $\mathbf{a}$, whose ith entry is $\mathbf{a}(i)$. A matrix is denoted by a bold uppercase letter, e.g., $\mathbf{A}$, with (i, j)th entry $\mathbf{A}(i, j)$; $\mathbf{A}(:, j)$ $(\mathbf{A}(i, :))$ denotes the jth column (respectively, ith row) of $\mathbf{A}$. A tensor (three-way array) is denoted by an underlined bold uppercase letter, e.g., $\underline{\mathbf{X}}$, with (i, j, k)th entry $\underline{\mathbf{X}}(i, j, k)$. $\underline{\mathbf{X}}(:, :, k)$ denotes the kth frontal $I \times J$ matrix slab of $\underline{\mathbf{X}}$, and similarly for the slabs along the other two modes. Vector, matrix, and three-way array size parameters (mode lengths) are denoted by uppercase letters, e.g., I. $\circ$ stands for the vector outer product; i.e., for two vectors $\mathbf{a}$ $(I \times 1)$ and $\mathbf{b}$ $(J \times 1)$, $\mathbf{a} \circ \mathbf{b}$ is an $I \times J$ matrix with (i, j)th element $\mathbf{a}(i)\mathbf{b}(j)$; i.e., $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T$. For three vectors, $\mathbf{a}$ $(I \times 1)$, $\mathbf{b}$ $(J \times 1)$, $\mathbf{c}$ $(K \times 1)$, $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$ is an $I \times J \times K$ three-way array with (i, j, k)th element $\mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)$. The vec$(\cdot)$ operator stacks the columns of its matrix argument in one tall column; $\otimes$

stands for the Kronecker product; $\odot$ stands for the Khatri–Rao (column-wise Kronecker) product: given $\mathbf{A}$ $(I \times F)$ and $\mathbf{B}$ $(J \times F)$, $\mathbf{A} \odot \mathbf{B}$ is the $JI \times F$ matrix

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{A}(:,1) \otimes \mathbf{B}(:,1) \cdots \mathbf{A}(:,F) \otimes \mathbf{B}(:,F)].$$

For a square matrix $\mathbf{S}$, $\mathrm{Tr}(\mathbf{S})$ denotes its trace, i.e., the sum of elements on its main diagonal. $\|\mathbf{x}\|_2^2$ is the Euclidean norm squared, and $\|\mathbf{A}\|_F^2$, $\|\underline{\mathbf{X}}\|_F^2$ the Frobenious norm squared–the sum of squares of all elements of the given vector, matrix, or tensor.

TENSOR DECOMPOSITION PRELIMINARIES

There is no comprehensive tutorial on tensor decompositions and applications from a signal processing point of view as of this writing, albeit there are several signal processing papers dealing with topics in tensor decomposition that have significant tutorial value. The concise introduction in [15] is still useful, although outdated. An upcoming *IEEE Signal Processing Magazine* tutorial article [8] covers the basic concepts and models well, and touches upon numerous applications. We also refer the reader to [16]–[18] for gentle introductions to tensor decompositions and applications from the viewpoint of computational linear algebra, chemistry, and the social sciences, respectively. Due to space limitations, here we only review essential concepts and results that directly relate to the core of our article.

RANK DECOMPOSITION

The rank of an $I \times J$ matrix $\mathbf{X}$ is the smallest number of rank-one matrices (vector outer products of the form $\mathbf{a} \circ \mathbf{b}$) needed to synthesize $\mathbf{X}$ as

$$\mathbf{X} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f = \mathbf{A}\mathbf{B}^T,$$

where $\mathbf{A} := [\mathbf{a}_1, \cdots, \mathbf{a}_F]$, and $\mathbf{B} := [\mathbf{b}_1, \cdots, \mathbf{b}_F]$. This relation can be expressed element-wise as

$$\mathbf{X}(i,j) = \sum_{f=1}^{F} \mathbf{a}_f(i)\mathbf{b}_f(j).$$

The rank of an $I \times J \times K$ three-way array $\underline{\mathbf{X}}$ is the smallest number of outer products needed to synthesize $\underline{\mathbf{X}}$ as

$$\underline{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f.$$

This relation can be expressed element-wise as

$$\underline{\mathbf{X}}(i,j,k) = \sum_{f=1}^{F} \mathbf{a}_f(i)\mathbf{b}_f(j)\mathbf{c}_f(k).$$

In the sequel we will assume that F is minimal, i.e., $F = \mathrm{rank}(\underline{\mathbf{X}})$, unless otherwise noted. The tensor $\underline{\mathbf{X}}$ comprises K frontal slabs of size $I \times J$; denote them $\{\mathbf{X}_k\}_{k=1}^{K}$, with $\mathbf{X}_k := \underline{\mathbf{X}}(:,:,k)$. Rearranging the elements of $\underline{\mathbf{X}}$ in a tall matrix $\mathbf{X} := [\mathrm{vec}(\mathbf{X}_1), \cdots, \mathrm{vec}(\mathbf{X}_K)]$, it can be shown that

$$\mathbf{X} = (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T \Leftrightarrow \mathbf{x} := \mathrm{vec}(\mathbf{X}) = (\mathbf{C} \odot \mathbf{B} \odot \mathbf{A})\mathbf{1},$$

where, $\mathbf{A}$ $\mathbf{B}$ are as defined for the matrix case, $\mathbf{C} := [\mathbf{c}_1, \cdots, \mathbf{c}_F]$, $\mathbf{1}$ is a vector of all 1s, and we have used the vectorization property of the Khatri–Rao product $\mathrm{vec}(\mathbf{A}\mathbf{D}(\mathbf{d})\mathbf{B}^T) = (\mathbf{B} \odot \mathbf{A})\mathbf{d}$, where $\mathbf{D}(\mathbf{d})$ is a diagonal matrix with the vector $\mathbf{d}$ as its diagonal.

CANDECOMP-PARAFAC

The above rank decomposition model for tensors is known as *parallel factor analysis* (PARAFAC) [19], [20] or *canonical decomposition* (CANDECOMP) [21], or CP (and CPD) for CANDECOMP-PARAFAC (decomposition), or *canonical polyadic decomposition* (CPD, again). CP is usually fitted using an alternating least squares procedure based on the model equation $\mathbf{X} = (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T$. In practice we will have $\mathbf{X} \approx (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T$, due to measurement noise and other imperfections, or simply because we wish to approximate a higher-rank model with a lower-rank one. Fixing $\mathbf{A}$ and $\mathbf{B}$, we solve

$$\min_{\mathbf{C}} \|\mathbf{X} - (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T\|_F^2,$$

which is a linear least squares problem. We can bring any of the matrix factors to the right by reshuffling the data, yielding corresponding conditional updates for $\mathbf{A}$ and $\mathbf{B}$. We can revisit each matrix in a circular fashion until convergence of the cost function, and this is the most commonly adopted approach to fitting the CP model, in good part because of its conceptual and programming simplicity, plus the ease with which one can incorporate additional constraints on the columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ [7].

TUCKER3

CP is in a way the most basic tensor model, because of its direct relationship to tensor rank and the concept of rank decomposition; but other algebraic tensor models exist, and the most notable one is known as *Tucker3*. Like CP, Tucker3 is a sum of outer products model, involving outer products of columns of three matrices, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Unlike CP however, which restricts interactions to corresponding columns (so that the first column of $\mathbf{A}$ only appears in one outer product involving the first column of $\mathbf{B}$ and the first column of $\mathbf{C}$), Tucker3 includes all outer products of every column of $\mathbf{A}$ with every column of $\mathbf{B}$ and every column of $\mathbf{C}$. Each such outer product is further weighted by the corresponding entry of a so-called core tensor, whose dimensions are equal to the number of columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$.

Consider again the $I \times J \times K$ three-way array $\underline{\mathbf{X}}$ comprising K matrix slabs $\{\mathbf{X}_k\}_{k=1}^{K}$, arranged into the tall matrix $\mathbf{X} := [\mathrm{vec}(\mathbf{X}_1), \cdots, \mathrm{vec}(\mathbf{X}_K)]$. The Tucker3 model can be written in matrix form as

$$\mathbf{X} \approx (\mathbf{B} \otimes \mathbf{A})\mathbf{G}\mathbf{C}^T,$$

where $\mathbf{G}$ is the core tensor in matrix form, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ can be assumed orthogonal without loss of generality, because linear transformations of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ can be absorbed in $\mathbf{G}$. The nonzero elements of the core tensor determine the interactions between columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. The associated model-fitting problem is

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{G}} \|\mathbf{X} - (\mathbf{B} \otimes \mathbf{A})\mathbf{G}\mathbf{C}^T\|_F^2,$$

which is usually solved using an alternating least squares procedure. The Tucker3 model can be fully vectorized as $\text{vec}(\mathbf{X}) \approx (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\text{vec}(\mathbf{G})$.

IDENTIFIABILITY

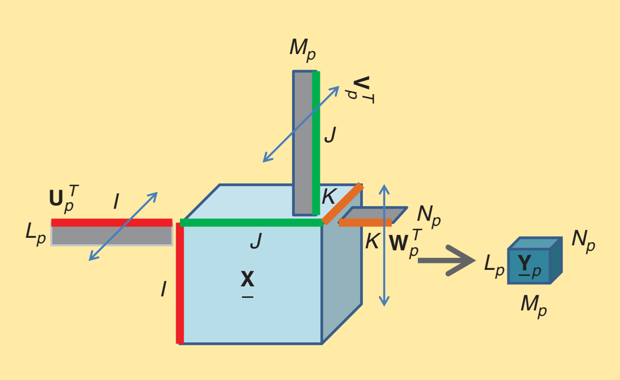
The distinguishing feature of the CP model is its essential uniqueness: under certain conditions, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ can be identified from $\mathbf{X}$ up to a common permutation and scaling/counterscaling of columns [19]–[26]. In contrast, Tucker3 is highly nonunique; the inclusion of all possible outer products of columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ results in overparametrization that renders it unidentifiable in most cases of practical interest. Still, Tucker3 is useful as an exploratory tool and for data compression/interpolation; we will return to this shortly.

Consider an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ of rank F. In vectorized form, it can be written as the $IJK \times 1$ vector $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\mathbf{1}$, for some $\mathbf{A}$ $(I \times F)$, $\mathbf{B}$ $(J \times F)$, and $\mathbf{C}$ $(K \times F)$—a CP model of size $I \times J \times K$ and order F parameterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$. (Notice the slight abuse of notation: we switched from $\mathbf{x} = (\mathbf{C} \odot \mathbf{B} \odot \mathbf{A})\mathbf{1}$ to $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\mathbf{1}$. The two are related via a row permutation, or by switching the roles of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$.) The Kruskal-rank of $\mathbf{A}$, denoted $k_{\mathbf{A}}$, is the maximum k such that any k columns of $\mathbf{A}$ are linearly independent $(k_{\mathbf{A}} \leq r_{\mathbf{A}} := \mathit{rank}(\mathbf{A}))$.

THEOREM 1

Given $\underline{\mathbf{X}}$ $(\Leftrightarrow \mathbf{x})$, $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are unique up to a common column permutation and scaling (e.g., scaling the first column of $\mathbf{A}$ and counterscaling the first column of $\mathbf{B}$ and/or $\mathbf{C}$, so long as their product remains the same), provided that $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2F + 2$ [22]. An equivalent and perhaps more intuitive way to express this is that the outer products $\mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f$ (i.e., the rank-one factors of $\underline{\mathbf{X}}$) are unique.

Note that we can always reshuffle the order of these rank-one factors (e.g., swap $\mathbf{a}_1 \circ \mathbf{b}_1 \circ \mathbf{c}_1$ and $\mathbf{a}_2 \circ \mathbf{b}_2 \circ \mathbf{c}_2$) without changing their sum $\underline{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f$, but this is a trivial and inherently unresolvable ambiguity that we will ignore in the sequel. Theorem 1 is Kruskal's celebrated uniqueness result [22], see also follow-up work in [23]–[25]. Kruskal's result applies to given $(\mathbf{A}, \mathbf{B}, \mathbf{C})$, i.e., it can establish uniqueness of a given decomposition. Recently, more relaxed uniqueness conditions have been obtained, which only depend on the size and rank of the tensor, albeit they cover almost all tensors of the given size and rank, i.e., except for a set of measure zero. Two such conditions are summarized next.



[FIG1] A schematic illustration of tensor compression: going from an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ to a much smaller $L_p \times M_p \times N_p$ tensor $\underline{\mathbf{Y}}_p$ via multiplying (every slab of) $\underline{\mathbf{X}}$ from the I-mode with $\mathbf{U}_p^T$, from the J-mode with $\mathbf{V}_p^T$, and from the K-mode with $\mathbf{W}_p^T$, where $\mathbf{U}_p$ is $I \times L_p$, $\mathbf{V}_p$ is $J \times M_p$, and $\mathbf{W}_p$ is $K \times N_p$.

THEOREM 2

Consider an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ of rank F. If

$$r_{\mathbf{C}} = F \text{ (which implies } K \geq F)$$

and

$$I(I-1)J(J-1) \geq 2F(F-1),$$

then the rank-one factors of $\underline{\mathbf{X}}$ are almost surely unique [27] (see also [24]).

THEOREM 3

Consider an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ of rank F. Order the dimensions so that $I \leq J \leq K$. Let i be maximal such that $2^i \leq I$, and likewise j maximal such that $2^j \leq J$. If $F \leq 2^{i+j-2}$, then the rank-one factors of $\underline{\mathbf{X}}$ are almost surely unique [26]. For I, J powers of 2, the condition simplifies to $F \leq (IJ/4)$. More generally, the condition implies that if $F \leq ((I+1)(J+1)/16)$, then $\underline{\mathbf{X}}$ has a unique decomposition almost surely. Before we proceed to discuss big data and cloud computing aspects of tensor decomposition, we state two lemmas from [13], which we will need in the sequel.

LEMMA 1

Consider $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, where $\mathbf{A}$ is $I \times F$, and let the $I \times L$ matrix $\mathbf{U}$ be randomly drawn from an absolutely continuous distribution (e.g., multivariate Gaussian with a nonsingular covariance matrix). Then $k_{\tilde{\mathbf{A}}} = \min(L, k_{\mathbf{A}})$ almost surely (with probability 1) [13].

LEMMA 2

Consider $\tilde{\mathbf{A}} = \mathbf{U}^T\mathbf{A}$, where $\mathbf{A}$ $(I \times F)$ is deterministic, tall/square $(I \geq F)$ and full column rank $r_{\mathbf{A}} = F$, and the elements of $\mathbf{U}$ $(I \times L)$ are independent and identically distributed (i.i.d.) Gaussian zero mean, unit variance random variables. Then the distribution of $\tilde{\mathbf{A}}$ is absolutely continuous (nonsingular multivariate Gaussian) [13].

TENSOR COMPRESSION

When dealing with big tensors $\underline{\mathbf{X}}$ that do not fit in main memory, a reasonable idea is to try to compress $\underline{\mathbf{X}}$ to a much smaller tensor that somehow captures most of the systematic variation in $\underline{\mathbf{X}}$. The commonly used compression method is to fit a low-dimensional orthogonal Tucker3 model (with low mode-ranks) [17], [18], then regress the data onto the fitted mode-bases. This idea has been exploited in existing CP model-fitting software, such as COMFAC [28], as a useful quick and dirty way to initialize alternating least squares computations in the uncompressed domain, thus accelerating convergence. A key issue with Tucker3 compression of big tensors is that it requires computing singular value decompositions of the various matrix unfoldings of the full data, in an

COMPLEXITY OF MULTIWAY COMPRESSION?

Multiplying a dense $L \times I$ matrix $\mathbf{U}^T$ with a dense vector $\mathbf{a}$ to compute $\mathbf{U}^T\mathbf{a}$ has complexity LI. Taking the product of $\mathbf{U}^T$ and the first $I \times J$ frontal slab $\underline{\mathbf{X}}(:,:,1)$ of the $I \times J \times K$ tensor $\underline{\mathbf{X}}$ has complexity LIJ. Premultiplying from the left all frontal slabs of $\underline{\mathbf{X}}$ by $\mathbf{U}^T$ (computing a mode product) therefore requires $LIJK$ operations, when all operands are dense. Multiway compression as in Figure 1 comprises three mode products, suggesting a complexity of $LIJK + MLJK + NLMK$, if the first mode is compressed first, followed by the second, and then the third mode. Notice that the order in which the mode products are computed affects the complexity of the overall operation; but orderwise, this is $O(\min(L, M, N)IJK)$. Also notice that if I, J, K are of the same order, and so are L, M, N, then the overall complexity is $O(LI^3)$. If $\mathbf{a}$ is sparse with $\mathrm{NZ}(\underline{\mathbf{a}})$ nonzero elements, we can compute $\mathbf{U}^T\mathbf{a}$ as a weighted sum of the columns of $\mathbf{U}^T$ corresponding to the nonzero elements of $\mathbf{a}$. This reduces matrix-vector multiplication complexity to $L\mathrm{NZ}(\underline{\mathbf{a}})$. It easily follows that if $\underline{\mathbf{X}}$ has $\mathrm{NZ}(\underline{\mathbf{X}})$ nonzero elements, the complexity of premultiplying from the left all frontal slabs of $\underline{\mathbf{X}}$ by $\mathbf{U}^T$ can be reduced to $L\mathrm{NZ}(\underline{\mathbf{X}})$. The problem is that, after computing the first mode product, the resulting tensor will be dense, hence subsequent mode products cannot exploit sparsity to reduce complexity. Note that, in addition to computational complexity, memory or secondary storage to save the intermediate results of the computation becomes an issue, even if the original tensor $\underline{\mathbf{X}}$ is sparse.

alternating fashion. This is a serious bottleneck for big data. Another issue is that Tucker3 compression is lossy, and it cannot guarantee that identifiability properties will be preserved. Finally, fitting a CP model to the compressed data can only yield an approximate model for the original uncompressed data, and eventually decompression and iterations with the full data are required to obtain fine estimates.

Consider compressing $\mathbf{x}$ into $\mathbf{y} = \mathbf{S}\mathbf{x}$, where $\mathbf{S}$ is $d \times IJK$, $d \ll IJK$. Sidiropoulos and Kyrillidis [13] proposed using a specially structured compression matrix $\mathbf{S} = \mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T$, which corresponds to multiplying (every slab of) $\underline{\mathbf{X}}$ from the I-mode with $\mathbf{U}^T$, from the J-mode with $\mathbf{V}^T$, and from the K-mode with $\mathbf{W}^T$, where $\mathbf{U}$ is $I \times L$, $\mathbf{V}$ is $J \times M$, and $\mathbf{W}$ is $K \times N$, with $L \le I$, $M \le J$, $N \le K$ and $LMN \ll IJK$; see Figure 1. Such an $\mathbf{S}$ corresponds to compressing each mode individually, which is often natural, and the associated multiplications can be efficiently implemented; see "Complexity of Multiway Compression?" and "Complexity of Multiway Compression–Redux." Due to a fortuitous property of the Kronecker product [29],

$$(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)(\mathbf{A} \odot \mathbf{B} \odot \mathbf{C}) = ((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})),$$

from which it follows that

$$\mathbf{y} = ((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C}))\mathbf{1} = (\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}})\mathbf{1}.$$

i.e., the compressed data follow a CP model of size $L \times M \times N$ and order F parameterized by $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$, with $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$ $\tilde{\mathbf{B}} := \mathbf{V}^T\mathbf{B}$, $\tilde{\mathbf{C}} := \mathbf{W}^T\mathbf{C}$.

This is nice to know, but we are really, naturally, interested in obtaining answers to the following two questions:

1) Under what conditions on $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ are $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$, identifiable from $\mathbf{y}$?
2) Under what conditions, if any, are $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ identifiable from $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$?

We start by answering the first question in the next section.

STEPPING-STONE RESULTS

The following result is a direct consequence of Lemma 1 and Kruskal's uniqueness condition in Theorem 1.

THEOREM 4

Let $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\mathbf{1} \in \mathbb{R}^{IJK}$, where $\mathbf{A}$ is $I \times F$, $\mathbf{B}$ is $J \times F$, $\mathbf{C}$ is $K \times F$, and consider compressing it to $\mathbf{y} = (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{x} = ((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C}))\mathbf{1} = (\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}})\mathbf{1} \in \mathbb{R}^{LMN}$, where the mode-compression matrices $\mathbf{U}\,(I \times L, L \le I)$, $\mathbf{V}\,(J \times M, M \le J)$, and $\mathbf{W}\ (K \times N, N \le K)$ are independently drawn from an absolutely continuous distribution. If

$$\min(L, k_{\mathbf{A}}) + \min(M, k_{\mathbf{B}}) + \min(N, k_{\mathbf{C}}) \ge 2F + 2,$$

then $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ are almost surely identifiable from the compressed data $\mathbf{y}$ up to a common column permutation and scaling.

More relaxed conditions for identifiability of $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ can be derived from Lemma 2, and Theorems 2 and 3.

THEOREM 5

For $\mathbf{x}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{y}$ as in Theorem 4, if $F \le \min(I, J, K)$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are all full column rank (F), $N \ge F$, and

$$L(L-1)M(M-1) \ge 2F(F-1),$$

then $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ are almost surely identifiable from the compressed data $\mathbf{y}$ up to a common column permutation and scaling.

REMARK 1

$F \le \min(I, J, K) \Rightarrow$ full column rank $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ almost surely, i.e., tall matrices are full column rank except for a set of measure zero. In other words, if $F \le \min(I, J, K)$ and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are themselves considered to be independently drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IF}$, $\mathbb{R}^{JF}$, and $\mathbb{R}^{KF}$, respectively, then they will all be full column rank with probability 1.

THEOREM 6

For $\mathbf{x}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{y}$ as in Theorem 4, if $F \le \min(I, J, K)$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are all full column rank (F), $L \le M \le N$, and

$$(L+1)(M+1) \ge 16F,$$

then $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ are almost surely identifiable from the compressed data $\mathbf{y}$ up to a common column permutation and scaling.

COMPLEXITY OF MULTIWAY COMPRESSION–REDUX

In scalar form, the (ℓ, m, n)th element of the tensor $\underline{\mathbf{Y}}$ after multiway compression can be written as

$$\underline{\mathbf{Y}}(\ell, m, n) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbf{U}(i, \ell)\mathbf{V}(j, m)\mathbf{W}(k, n)\underline{\mathbf{X}}(i, j, k).$$

Claim S1

Suppose that $\underline{\mathbf{X}}$ is sparse, with $\mathrm{NZ}(\underline{\mathbf{X}})$ nonzero elements, and suppose that it is stored as a serial list with entries formatted as $[i, j, k, v]$, where v is the nonzero value at tensor position (i, j, k). Suppose that the list is indexed by an integer index s, i.e., $[i(s), j(s), k(s), v(s)]$ is the record corresponding to the sth entry of the list. Then the following simple algorithm will compute the multiway compressed tensor $\underline{\mathbf{Y}}$ in only $LMN\mathrm{NZ}(\underline{\mathrm{X}})$ operations, requiring only LMN cells of memory to store the result, and $IL + JM + KN$ cells of memory to store the matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$.

Algorithm S1: Efficient multiway compression pseudocode

```
Y=zeros(L,M,N);
for s=1:NZX,
 for ell=1:L,
  for m=1:M,
   for n=1:N,
    Y(ell,m,n) = Y(ell,m,n)+ U(i(s),ell)*V(j(s),m)*W(k(s),n)*v(s);
   end
  end
 end
end
```

Notice that, even if $\underline{\mathbf{X}}$ is dense (i.e., $\mathrm{NZ}(\underline{\mathbf{X}}) = IJK$), the above algorithm only needs to read each element of $\underline{\mathbf{X}}$ once, so complexity will be $LMNIJK$ but memory will still be very modest: only LMN cells of memory to store the result, and $IL + JM + KN$ cells of memory to store the matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$. Contrast this to the naive way of serially computing the mode products, whose complexity order is $O(\min(L, M, N)IJK)$ but whose memory requirements are huge for dense $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, due to intermediate result explosion—even for sparse $\underline{\mathbf{X}}$. We see a clear complexity-memory tradeoff between the two approaches for dense data, but Algorithm S1 is a clear winner for sparse data, because sparsity is lost after the first mode product. Notice that the above algorithm can be fully parallelized in several ways—by splitting the list of nonzero elements across cores or processors (paying in terms of auxiliary memory replications to store partial results for $\underline{\mathbf{Y}}$ and the matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, locally at each processor), or by splitting the (ℓ, m, n) loops—at the cost of replicating the data list. As a final word, the memory access pattern (whether we read and write consecutive memory elements in blocks, or make wide strides) is the performance-limiting factor for truly big data, Algorithm S1 makes strides in reading elements of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and writing elements of $\underline{\mathbf{Y}}$. There are ways to reduce these strides, at the cost of requiring more memory and more floating point operations.

MAIN RESULTS

Theorems 4–6 can establish uniqueness of $\tilde{\mathrm{A}}, \tilde{\mathrm{B}}, \tilde{\mathrm{C}}$, but we are ultimately interested in $\mathrm{A}, \mathrm{B}, \mathrm{C}$. We know that $\tilde{\mathrm{A}} = \mathrm{U}^T\mathrm{A}$, and we know U^T, but, unfortunately, it is a fat matrix that cannot be inverted. To uniquely recover A, one needs additional structural constraints. Sidiropoulos and Kyrillidis [13] proposed exploiting column-wise sparsity in A (and likewise B, C), which is often plausible in practice. A need only be sparse with respect to (when expressed in) a suitable basis, provided the sparsifying basis is known a priori. Sparsity is a powerful constraint, but it is not always valid (or a sparsifying basis may be unknown). For this reason, we propose here a different solution, based on creating and factoring a number of randomly reduced replicas of the full data.

Consider spawning P randomly compressed reduced-size replicas $\{\underline{\mathrm{Y}}_p\}_{p=1}^{P}$ of the tensor $\underline{\mathrm{X}}$, where $\underline{\mathrm{Y}}_p$ is created using mode compression matrices $(\mathrm{U}_p, \mathrm{V}_p, \mathrm{W}_p)$; see Figure 2. Assume that identifiability conditions per Theorem 5 or Theorem 6 hold, so that $\tilde{\mathrm{A}}_p, \tilde{\mathrm{B}}_p, \tilde{\mathrm{C}}_p$ are almost surely identifiable (up to permutation and scaling of columns) from $\underline{\mathrm{Y}}_p$. Then, upon factoring $\underline{\mathrm{Y}}_p$ into F rank-one components, we obtain

$$\tilde{\mathrm{A}}_p = \mathrm{U}_p^T\mathrm{A}\boldsymbol{\Pi}_p\boldsymbol{\Lambda}_p, \tag{1}$$

where $\boldsymbol{\Pi}_p$ is a permutation matrix, and $\boldsymbol{\Lambda}_p$ is a diagonal scaling matrix with nonzero elements on its diagonal. Assume that the first two columns of each U_p (rows of U_p^T) are common, and let $\overline{\mathrm{U}}$ denote this common part, and $\overline{\mathrm{A}}_p$ denote the first two rows of $\tilde{\mathrm{A}}_p$. We therefore have

$$\overline{\mathrm{A}}_p = \overline{\mathrm{U}}^T\mathrm{A}\boldsymbol{\Pi}_p\boldsymbol{\Lambda}_p.$$

Dividing each column of $\overline{\mathrm{A}}_p$ by the element of maximum modulus in that column, and denoting the resulting $2 \times F$ matrix $\hat{\mathrm{A}}_p$, we obtain

$$\hat{\mathrm{A}}_p = \overline{\mathrm{U}}^T\mathrm{A}\boldsymbol{\Lambda}\boldsymbol{\Pi}_p.$$

Notice that $\boldsymbol{\Lambda}$ does not affect the ratio of elements in each 2×1 column. If these ratios are distinct (which is guaranteed almost surely if $\overline{\mathrm{U}}$ and A are independently drawn from absolutely continuous distributions), then the different permutations can be matched by sorting the ratios of the two coordinates of each 2×1 column of $\hat{\mathrm{A}}_p$.

In practice, using a few more anchor rows will improve the permutation-matching performance, and is recommended in difficult cases with high noise variance. When S anchor rows are used, the optimal permutation matching problem can be cast as

$$\min_{\boldsymbol{\Pi}} ||\hat{\mathrm{A}}_1 - \hat{\mathrm{A}}_p\boldsymbol{\Pi}||_F^2,$$

where optimization is over the set of permutation matrices. This may appear to be a hard combinatorial problem at first sight; but it is not. Using

$$\|\hat{\mathbf{A}}_1-\hat{\mathbf{A}}_p\boldsymbol{\Pi}\|_F^2=\mathrm{Tr}\left((\hat{\mathbf{A}}_1-\hat{\mathbf{A}}_p\boldsymbol{\Pi})^T(\hat{\mathbf{A}}_1-\hat{\mathbf{A}}_p\boldsymbol{\Pi})\right)$$
$$=\|\hat{\mathbf{A}}_1\|_F^2+\|\hat{\mathbf{A}}_p\boldsymbol{\Pi}\|_F^2-2\mathrm{Tr}(\hat{\mathbf{A}}_1^T\hat{\mathbf{A}}_p\boldsymbol{\Pi})$$
$$=\|\hat{\mathbf{A}}_1\|_F^2+\|\hat{\mathbf{A}}_p\|_F^2-2\mathrm{Tr}(\hat{\mathbf{A}}_1^T\hat{\mathbf{A}}_p\boldsymbol{\Pi}).$$

It follows that we may instead

$$\max_{\boldsymbol{\Pi}}\mathrm{Tr}(\hat{\mathbf{A}}_1^T\hat{\mathbf{A}}_p\boldsymbol{\Pi}),$$

over the set of permutation matrices. This is what is known as the *linear assignment problem* (LAP), and it can be efficiently solved using the Hungarian algorithm.

After this column permutation-matching process, we go back to (1) and permute its columns to obtain $\check{\mathbf{A}}_p$ satisfying

$$\check{\mathbf{A}}_p=\mathbf{U}_p^T\mathbf{A}\boldsymbol{\Pi}\boldsymbol{\Lambda}_p.$$

It remains to get rid of $\boldsymbol{\Lambda}_p$. For this, we normalize each column by dividing it with its norm. This finally yields

$$\check{\mathbf{A}}_p=\mathbf{U}_p^T\mathbf{A}\boldsymbol{\Pi}\boldsymbol{\Lambda}.$$

For recovery of $\mathbf{A}$ up to permutation and scaling of its columns, we then require that the matrix of the linear system

$$\begin{bmatrix}\check{\mathbf{A}}_1\\ \vdots\\ \check{\mathbf{A}}_P\end{bmatrix}=\begin{bmatrix}\mathbf{U}_1^T\\ \vdots\\ \mathbf{U}_P^T\end{bmatrix}\mathbf{A}\boldsymbol{\Pi}\boldsymbol{\Lambda} \qquad (2)$$

be full column rank. This implies that

$$2+\sum_{p=1}^{P}(L_p-2)\geq I$$

i.e.,

$$\sum_{p=1}^{P}(L_p-2)\geq I-2.$$

Note that every submatrix contains the two anchor rows that are common, and duplicate rows clearly do not increase the rank. Also note that once the dimensionality requirement is met, the matrix will be full rank with probability 1, because its nonredundant entries are drawn from a jointly continuous distribution (by design).

Assuming $L_p=L,\ \forall p\in\{1,\cdots,P\}$ for simplicity (and symmetry of computational load), we obtain $P(L-2)\geq I-2$, or, in terms of the number of threads

$$P\geq\frac{I-2}{L-2}.$$

Likewise, from the corresponding full column rank requirements for the other two modes, we obtain

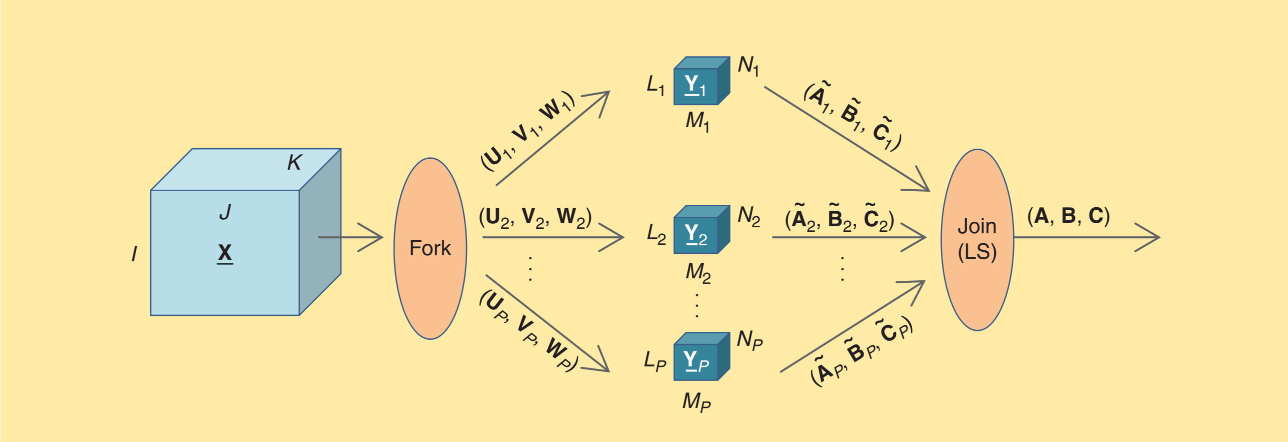
$$P\geq\frac{J}{M},\ \text{and}\ P\geq\frac{K}{N}.$$

Notice that we do not subtract two from numerator and denominator for the other two modes, because the permutation of columns of $\tilde{\mathbf{A}}_p,\tilde{\mathbf{B}}_p,\tilde{\mathbf{C}}_p$ is common, so it is enough to figure it out from one mode, and apply it to other modes as well. In short,

$$P\geq\max\left(\frac{I-2}{L-2},\frac{J}{M},\frac{K}{N}\right).$$

REMARK 2

Note that if, say, $\mathbf{A}$ can be identified and it is full column rank, then $\mathbf{B}$ and $\mathbf{C}$ can be identified by solving a linear least squares problem—but this requires access to the full big tensor data. In the same vein, if $\mathbf{A}$ and $\mathbf{B}$ are identified, then $\mathbf{C}$ can be identified from the full big tensor data even if $\mathbf{A}$ and $\mathbf{B}$ are not full column rank individually—it is enough that $\mathbf{A}\odot\mathbf{B}$ is full column rank, which is necessary for identifiability of $\mathbf{C}$ even from the big tensor, hence not restrictive. Parallel randomly compressed (PARACOMP)-based identification, on the other hand, only requires access to the factors derived from the small



[FIG2] A schematic illustration of the PARACOMP fork-join architecture. The fork step creates a set of P randomly compressed reduced-size replicas $\{\underline{\mathbf{Y}}_p\}_{p=1}^P$. Each $\underline{\mathbf{Y}}_p$ is obtained by applying $(\mathbf{U}_p,\mathbf{V}_p,\mathbf{W}_p)$ to $\underline{\mathbf{X}}$, as detailed in Figure 1. Each $\underline{\mathbf{Y}}_p$ is then independently factored (all P threads can be executed in parallel). The join step collects the estimated mode loading submatrices $(\tilde{\mathbf{A}}_p,\tilde{\mathbf{B}}_p,\tilde{\mathbf{C}}_p)$ from the P threads, and, after anchoring all to a common permutation and scaling, solves a master linear least squares problem per mode to estimate the full mode loading matrices $(\mathbf{A},\mathbf{B},\mathbf{C})$.

replicas. This is clearly advantageous, as the raw big tensor data can be discarded after compression, and there is no need for retrieving huge amounts of data from cloud storage.

One can pick the mode used to figure out the permutation ambiguity, leading to the symmetrized condition $P \geq \min\{P_1, P_2, P_3\}$ with

$$P_1 = \max\left(\frac{I-2}{L-2}, \frac{J}{M}, \frac{K}{N}\right)$$
$$P_2 = \max\left(\frac{I}{L}, \frac{J-2}{M-2}, \frac{K}{N}\right)$$
$$P_3 = \max\left(\frac{I}{L}, \frac{J}{M}, \frac{K-2}{N-2}\right).$$

If the compression ratios in the different modes are similar, it makes sense to use the longest mode for this purpose; if this is the last mode, then

$$P \geq \max\left(\frac{I}{L}, \frac{J}{M}, \frac{K-2}{N-2}\right).$$

We have thus established the following result.

THEOREM 7

In reference to Figure 2, assume $\mathbf{x} := \mathrm{vec}(\underline{\mathbf{X}}) = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\mathbf{1} \in \mathbb{R}^{IJK}$, where $\mathbf{A}$ is $I \times F$, $\mathbf{B}$ is $J \times F$, and $\mathbf{C}$ is $K \times F$ (i.e., the rank of $\underline{\mathbf{X}}$ is at most F). Assume that $F \leq I \leq J \leq K$, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are all full column rank (F). Further assume that $L_p = L$, $M_p = M$, $N_p = N$, $\forall p \in \{1, \cdots, P\}$, $L \leq M \leq N$, $(L+1)(M+1) \geq 16F$, the elements of $\{\mathbf{U}_p\}_{p=1}^{P}$ are drawn from a jointly continuous distribution, and likewise for $\{\mathbf{V}_p\}_{p=1}^{P}$, while each $\mathbf{W}_p$ contains two common anchor columns, and the elements of $\{\mathbf{W}_p\}_{p=1}^{P}$ (except for the repeated anchors, obviously) are drawn from a jointly continuous distribution. Then the data for each thread $\mathbf{y}_p := \mathrm{vec}(\underline{\mathbf{Y}}_p)$ can be uniquely factored, i.e., $(\tilde{\mathbf{A}}_p, \tilde{\mathbf{B}}_p, \tilde{\mathbf{C}}_p)$ is unique up to column permutation and scaling. If, in addition to the above, we also have $P \geq \max(I/L, J/M, (K-2)/(N-2))$ parallel threads, then $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are almost surely identifiable from the thread outputs $\{(\tilde{\mathbf{A}}_p, \tilde{\mathbf{B}}_p, \tilde{\mathbf{C}}_p)\}_{p=1}^{P}$ up to a common column permutation and scaling.

The above result is indicative of a family of results that can be derived, using different CP identifiability results. Its significance may not be immediately obvious, so it is worth elaborating further at this point. On one hand, Theorem 7 shows that fully parallel computation of the big tensor decomposition is possible—the first such result, to the best of our knowledge, that guarantees identifiability of the big tensor decomposition from the intermediate small tensor decompositions, without placing stringent additional constraints. On the other hand, the conditions appear convoluted, and the memory/storage and computational savings, if any, are not necessarily easy to see. The following claim nails down the take-home message.

CLAIM 1

Under the conditions of Theorem 7, if $(K-2)/(N-2) = \max(I/L, J/M, (K-2)/(N-2))$, then the memory/storage and computational complexity savings afforded by the architecture shown in Figure 2 relative to brute-force computation are of order (IJ/F).

Proof 1

Each thread must store LMN elements, and we have $P = (K-2)/(N-2)$ threads in all, leading to a total data size of order LMK versus IJK, so the ratio is (IJ/LM). The condition $(L+1)(M+1) \geq 16F$ only requires LM to be of order F, hence the total compression ratio can be as high as $O(IJ/F)$. Turning to overall computational complexity, note that optimal low-rank tensor factorization is NP-hard, even in the rank-one case. Practical tensor factorization algorithms, however, typically have complexity $O(IJKF)$ (per iteration, and overall if a bound on the maximum number of iterations is enforced). It follows that the practical complexity order for factoring out the P parallel threads is $O(PLMNF)$ versus $O(IJKF)$ for the brute-force computation. Taking into account the lower bound on P, the ratio is again of order (IJ/LM), and since the condition $(L+1)(M+1) \geq 16F$ only requires LM to be of order F, the total computational complexity gain can be as high as $O(IJ/F)$.

REMARK 3

The complexity of solving the master linear equation (2) in the final merging step for $\mathbf{A}$ may be a source of concern—especially because it hasn't been accounted for in the overall complexity calculation. Solving a linear system of order of I equations in I unknowns generally requires $O(I^3)$ computations; but closer scrutiny of the system matrix in (2) reveals interesting features. If all elements of the compression matrices $\{\mathbf{U}_p\}$ (except for the common anchors) are i.i.d. with zero mean and unit variance, then, after removing the redundant rows, the system matrix in (2) will have approximately orthogonal columns for large I. This implies that its left pseudoinverse will simply be its transpose, approximately. This reduces the complexity of solving (2) to I^2F. If higher accuracy is required, the pseudoinverse may be computed offline and stored. It is also important to stress that (2) is only solved once for each mode at the end of the overall process, whereas tensor decomposition typically takes many iterations. In short, the constants are such that we need to worry more about the compression (fork) and decomposition stages, rather than the final join stage.

Theorem 7 assumes $F \leq \min(I, J, K)$ to ensure (via Lemma 2) absolute continuity of the compressed factor matrices, which is needed to invoke almost sure uniqueness per [26]. Cases where $F > \min(I, J, K)$ can be treated using Kruskal's condition for unique decomposition of each compressed replica

$$\min(L, k_{\mathbf{A}}) + \min(M, k_{\mathbf{B}}) + \min(N, k_{\mathbf{C}}) \geq 2F + 2.$$

It can be shown that $k_{\mathbf{A}} = \min(I, F)$ for almost every $\mathbf{A}$ (except for a set of measure zero in $\mathbb{R}^{IF}$); and likewise $k_{\mathbf{B}} = \min(J, F)$, and $k_{\mathbf{C}} = \min(K, F)$, for almost every $\mathbf{B}$ and $\mathbf{C}$. This simplifies the above condition to

$$\min(L, I, F) + \min(M, J, F) + \min(N, K, F) \geq 2F + 2.$$

In other words, if the simplified condition holds, then CP decomposition of each reduced replica is unique for almost every $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and almost every set of compression matrices

$(\mathbf{U}, \mathbf{V}, \mathbf{W})$. Assume $I \geq F$, $J \geq F$, but $K < F$, and pick $L = M = F$, and $N = 3$. Then the condition further reduces to

$$2F + \min(3, K) \geq 2F + 2,$$

which is satisfied for any $K \geq 2$ (i.e., for any tensor). We also need

$$P \geq \max\left(\frac{I}{L}, \frac{J}{M}, \frac{K-2}{N-2}\right),$$

which in this case $(N = 3)$ reduces to

$$P \geq \max\left(\frac{I}{L}, \frac{J}{M}, K-2\right).$$

When $(I/L) = \max(I/L, J/M, K-2)$, then there are (I/L) parallel threads of size $LMN = 3F^2$ each, for total cloud storage $3IF$, i.e., order IF; hence the overall compression ratio (taking all replicas into account) is of order $((IJK)/(IF)) = (JK/F)$. The ratio of overall complexity orders is also $((IJKF)/(IF^2)) = (JK/F)$. This is the same type of result as the one we derived for the case $F \leq \min(I, J, K)$. On the other hand, when $K - 2 = \max(I/L, I/M, K-2)$, there are $K-2$ parallel threads of size $LMN = 3F^2$ each, for total cloud storage $3F^2(K-2)$, i.e., order KF^2; hence the overall compression ratio is $((IJK)/(KF^2)) = (IJ)/F^2$, and the ratio of overall complexity orders is also $((IJKF)/(KF^3)) = (IJ)/F^2$. We see that there is a penalty factor F relative to the case $F \leq \min(I, J, K)$; this is likely an artifact of the method of proof, which we hope to improve in future work. We summarize the result in the following theorem.

THEOREM 8

In reference to Figure 2, assume $\mathbf{x} := \text{vec}(\underline{\mathbf{X}}) = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\mathbf{1} \in \mathbb{R}^{IJK}$, where $\mathbf{A}$ is $I \times F$, $\mathbf{B}$ is $J \times F$, $\mathbf{C}$ is $K \times F$ (i.e., the rank of $\underline{\mathbf{X}}$ is at most F). Assume that $I \geq F$, $J \geq F$ (K can be $< F$), and pick $L_p = L$, $M_p = M$, $N_p = N$, $\forall p \in \{1, \cdots, P\}$, with $L = M = F$, and $N = 3$. The compression matrices are chosen as in Theorem 7. If $P \geq \max(I/L, J/M, K-2)$, then $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is identifiable from $\{(\tilde{\mathbf{A}}_p, \tilde{\mathbf{B}}_p, \tilde{\mathbf{C}}_p)\}_{p=1}^{P}$, for almost every $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and almost every set of compression matrices. When $(I/L) = \max((I/L), (J/M), K-2)$, the total storage and complexity gains are of order (JK/F); whereas if $K-2 = \max((I/L), (J/M), K-2)$, the total storage and complexity gains are of order (IJ/F^2).

LATENT SPARSITY

If latent sparsity is present, we can exploit it to reduce P. Assume that every column of $\mathbf{A}$ $(\mathbf{B}, \mathbf{C})$ has at most n_a (respectively, n_b, n_c) nonzero elements. A column of $\mathbf{A}$ can be uniquely recovered from only $2n_a$ incoherent linear equations [30]. Therefore, we may replace the condition

$$P \geq \max\left(\frac{I}{L}, \frac{J}{M}, \frac{K-2}{N-2}\right),$$

with

$$P \geq \max\left(\frac{2n_a}{L}, \frac{2n_b}{M}, \frac{2n_c-2}{N-2}\right). \tag{3}$$

Assuming

$$\frac{2n_c-2}{N-2} = \max\left(\frac{2n_a}{L}, \frac{2n_b}{M}, \frac{2n_c-2}{N-2}\right),$$

it is easy to see that the total cloud storage and complexity gains are of order $(IJ/F)(K/n_c)$—improved by a factor of (K/n_c). It is interesting to compare this result with the one in Sidiropoulos and Kyrillidis [13], which corresponds to using $P = 1$ in our present context. Notice that (3) implies $L \geq (2n_a/P)$, $M \geq (2n_b/P)$, $N-2 \geq ((2n_c-2)/P) \Rightarrow N \geq (2n_c/P) + 2(1-(1/P)) \Rightarrow N \geq (2n_c/P)$. Substituting $P = 1$ we obtain $L \geq 2n_a$, $M \geq 2n_b$, $N \geq 2n_c$, which is exactly the condition required in [13]. We see that PARACOMP subsumes [13], offering greater flexibility in terms of choosing P to reduce the size of replicas for easier in-memory processing, at the cost of an additional merging step at the end. Also note that PARACOMP is applicable in the case of dense latent factors, whereas [13] is not.

REMARK 4

In practice we will use a higher P, i.e.,

$$P \geq \max\left(\frac{\mu n_a}{L}, \frac{\mu n_b}{M}, \frac{\mu n_c - 2}{N-2}\right),$$

with $\mu \in \{3, 4, 5\}$ instead of 2, and an ℓ_1 sparse underdetermined linear equations solver for the final merging step for $\mathbf{A}$. This will increase complexity from $O(I^2F)$ to $O(I^{3.5}F)$, and the constants are such that the difference is significant. This is the price paid for the reduced memory and intermediate complexity benefits afforded by latent sparsity.

MAPREDUCE IMPLEMENTATION

With the proliferation of large collections of data, as well as big clusters of (usually commodity) computers that were largely underutilized, arose the need for a unified framework of scalable distributed computation in the cloud. In [31], Dean et al. from Google introduced such a framework, called *MapReduce*. MapReduce provides a very versatile level of programming abstraction: it conceals all its inner workings from the programmer, and simply requires the implementation of two functions: `Map` and `Reduce`.

The `Map` function runs in parallel on many machines; each instance reads data serially from the Distributed File System (DFS), performs some sort of parsing or computation on that data, and emits a series of (`key,value`) pairs. (DFS is defined by MapReduce.) Consequently, the `Reduce` function runs in parallel on a set of machines, and each instance of `Reduce` receives as input (`key,value`) pairs with the same `key`; it performs some sort of (user defined) aggregation or computation on these `values`, and then emits a series of (`key',value'`) pairs, which are eventually written to DFS. This way, any task that can be expressed as a combination of a `Map` and a `Reduce` function may be run in a distributed fashion on a cluster of computers, on data that is also stored in the cloud, and much bigger than what a typical personal computer can store or process in memory. The MapReduce framework also deals with machine failures (an issue which arises very often in large clusters of computers) in a way that is transparent to the programmer. Among other safety measures, MapReduce uses

SOLVING A TOY PROBLEM IN HADOOP–MAPREDUCE

Consider a large speech/audio, image, or video signal, stored as a text file, with each line containing a signal value. This file is stored in a distributed fashion, in DFS. To compute its histogram, it suffices to use a single MapReduce job.

- `Map`: Each mapper gets a portion of the file and reads it line-by-line. For each line-entry, n, the mapper sets `key` $= n$ and `value` $= 1$, and emits a $(n, 1)$ pair.
- `Reduce`: As mentioned earlier, each instance of a reducer receives all such (`key,value`) pairs that have the same `key`. In this particular case, all instances of number n will be processed by the same reducer, since the `Map` function set `key`$= n$. As a consequence, each reducer has all the information needed to calculate the exact count of appearances of a given number n. Thus, each reducer simply calculates the total number of $(n, 1)$ pairs (denoted by f), and emits a single tuple (n, f), which contains the number and its corresponding frequency of occurrence.

Finally, when all reducers have terminated, the output of the above MapReduce task will contain lines in the form: (`number, frequency`).

Even though the above example is very simple, the logic that underlies the transformation of an algorithm into a series of MapReduce tasks is the same: decompose the algorithm into self-contained pieces, find a (`key, value`) representation for the intermediate data of each piece, and finally express this computation as a pair of `Map` and `Reduce` functions.

three-way replication of each computation, so that even if one machine fails, there are still two backup machines that are carrying out the same task. This way, the user does not have to deal with the frustrations of machine failures. The original MapReduce implementation is internal to Google; however, there exists a very robust and well-tested open source implementation by Apache, called *Hadoop* [32]. The two primary programming languages that can be used with Hadoop are Java and Python.

Signal processing algorithms are generally not realizable as a single MapReduce task, but it is often possible to break up a given algorithm in parts, each of which may be written as a MapReduce computation. In this way, the overall signal processing algorithm can be implemented as a chain of MapReduce tasks.

The most typical introductory example of a MapReduce task is the `WordCount` application [33], where the goal is to estimate the frequency of occurrence of each word in a corpus. Given that MapReduce was originally developed by Google, a search engine that relies heavily on indexing large collections of text to provide fast and accurate search results, the `WordCount` example fits perfectly in the original context. In "Solving a Toy Problem in Hadoop–MapReduce," we instead use a very simple and common signal processing task to illustrate the way MapReduce works: computing the histogram of a big speech/audio, image, or video signal. The particular kind of signal is not important here, but bear in mind that our motivation is to be able to handle big data, distributed over the cloud. To simplify exposition, we assume that the signal of interest is integer valued.

SKETCH OF PARACOMP IN MAPREDUCE

We now provide a sketch of an implementation of PARACOMP in MapReduce. As in Figure 2, we break the algorithm down to three distinct steps: 1) compression, 2) decomposition, and 3) recovery of factor matrices. Each of the three steps consists of a few MapReduce chain tasks.

COMPRESSION

For the compression step, we first need to create P triplets of random compression matrices $\mathbf{U}_p$, $\mathbf{V}_p$, $\mathbf{W}_p$. This may be carried out simply by a mapper that emits p (the replica index) as key, and the dimensions of the matrices as the value. Thus, each reducer is responsible for creating and storing on DFS all three compression matrices. Depending on how large the compression matrices are, instead of assigning a single reducer the burden of creating an entire batch of $\mathbf{U}_p$, $\mathbf{V}_p$, $\mathbf{W}_p$, we may instead choose to assign each reducer to create a single row of each of the matrices. Taking a closer look at Algorithm S1 in "Complexity of Multiway Compression–Redux," we can devise a MapReduce task for the compression step. Let us assume that the tensor is stored in a text file, in multiple lines (as many as the nonzero values in the tensor), in the form

$$i(s),\ j(s),\ k(s),\ v(s),$$

which is appropriate for sparse tensors. Each mapper reads a segment of that file, processing one line at a time. By inspecting the core equation of Algorithm S1 in "Complexity of Multiway Compression–Redux"

$$\underline{\mathbf{Y}}(\ell, m, n) = \underline{\mathbf{Y}}(\ell, m, n) + \mathbf{U}(i(s), \ell)\mathbf{V}(j(s), m)\mathbf{W}(k(s), n)v(s),$$

we see that for each mapper, it suffices to hold $\mathbf{U}(i(s), :)$, $\mathbf{V}(j(s), :)$, and $\mathbf{W}(k(s), :)$ in memory, so that it calculates the contribution of the current nonzero value of the tensor $v(s)$ to the partial sum that comprises $\underline{\mathbf{Y}}(\ell, m, n)$. Since L, M, N are considerably smaller than I, J, K, we use $O(L+M+N)$ of memory on each mapper. Thus, each mapper emits as key the concatenation of (ℓ, m, n) and as value $\mathbf{U}(i(s), \ell)\mathbf{V}(j(s), m)\mathbf{W}(k(s), n)v(s)$. Finally, each reducer receives all partial values of the sum that builds $\underline{\mathbf{Y}}(\ell, m, n)$ up, sums up all incoming values, and emits a pair with key equal to (ℓ, m, n) and value equal to $\underline{\mathbf{Y}}(\ell, m, n)$, which is eventually written to DFS.

Since we execute multiple repetitions of the compression step, we may concatenate the repetition number p to the key that is emitted by the mapper, as well as the key emitted by the reducer. Thus, at the end, there will be one file containing the nonzero values for each compressed tensor in the form:

$$p, \ell, m, n, \underline{\mathbf{Y}}_p(\ell, m, n).$$

DECOMPOSITION

For the decomposition step, we spawn P parallel processes on different machines, each one fitting the CP decomposition to the respective compressed tensor. To do that in the MapReduce framework, we may use the `Map` function to feed the appropriate data to each reducer. More specifically, each mapper will read portions of the file created by the compression step, and use as a key the repetition index p, and as value the rest of the row, i.e., $(\ell, m, n, \underline{\mathbf{Y}}_p(\ell, m, n))$. Consequently, P reducers will be spawned, each receiving all the data of a single compressed tensor. We assume that the compressed tensor fits in the main memory of a single machine, therefore each reducer simply stores the incoming values in a three dimensional array, and proceeds with in-memory computation of the CP decomposition. In case the reducers are unable to store the compressed tensor in main memory, there exist methods that fit the CP decomposition on MapReduce [34]. However, solving each one of the parallel decompositions on MapReduce would significantly hurt performance, therefore we should aim for compressed tensors that fit in memory.

RECOVERY OF FACTOR MATRICES

The final step involves the recovery of the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ from the partial factors as obtained from the parallel decomposition step. Recovery for each factor matrix is achieved by stacking the partial results on top of each other, as well as the compression matrices in a similar fashion, and solving a least squares problem involving these two matrices. The stacking of both partial factors and compression matrices can be done through a simple MapReduce task: each mapper will be emitting (i, f) (i.e., the indices of each matrix coefficient) as key, and the value will be the coefficient of the matrix at (i, f) (denoted by v) and the index p, indicating the replica number. Then, each reducer will emit tuples of the form

$$i', f, v,$$

where i' will be the original row index adjusted appropriately using p, to account for the stacking.

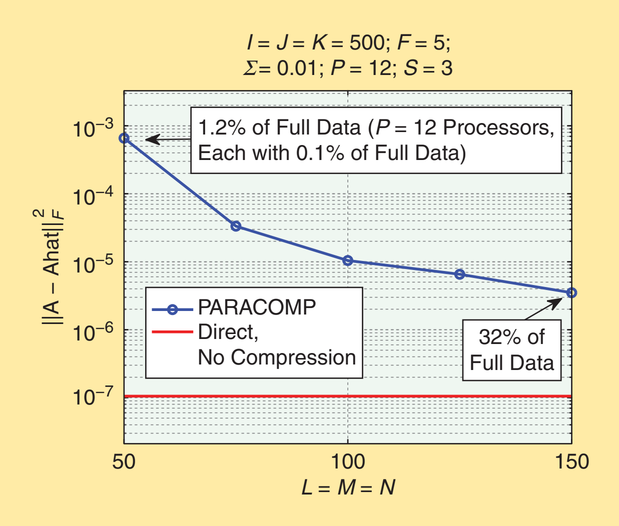
To solve the least squares step, we may use scalable algorithms that implement the Moore–Penrose pseudoinverse on MapReduce [35]. After pseudoinversion, we need to carry out matrix multiplication, a problem that has also been thoroughly studied for MapReduce [36].

ILLUSTRATIVE NUMERICAL RESULTS

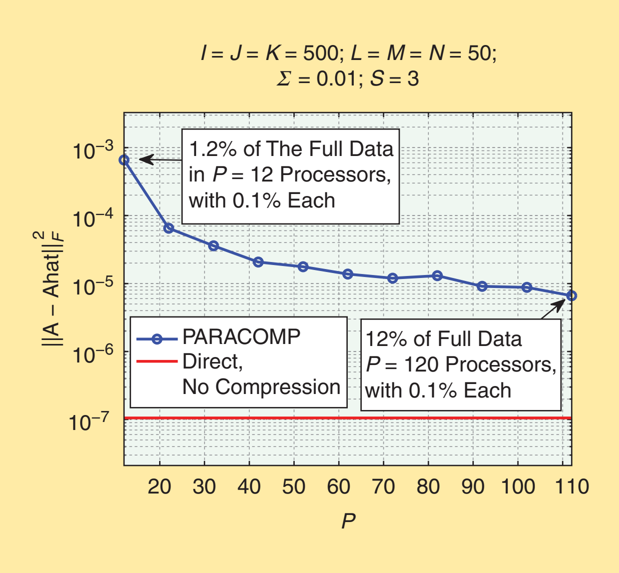
Our theorems ensure that PARACOMP works with ideal low- and known-rank tensors, but what if there is measurement noise or other imperfections, or we underestimate the rank? Does the overall approach fall apart in this case? From "The Color of Compressed Noise" and "Is Component Ordering Preserved After Compression?" we have good reasons to believe that this is not the case, but one cannot be confident without numerical experiments that corroborate intuition. In this section, we provide indicative results to illustrate what can be expected from PARACOMP and the effect of various parameters on estimation performance.

In all cases considered, $I = J = K = 500$, the noiseless tensor has rank $F = 5$, and is synthesized by randomly and independently drawing $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ each from an i.i.d. zero-mean, unit-variance Gaussian distribution (`randn(500,5)` in MATLAB), and then taking their tensor product; i.e., computing the sum of outer products of corresponding columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Gaussian i.i.d. measurement noise is then added to this noiseless tensor to yield the observed tensor to be analyzed. The nominal setup uses $L = M = N = 50$ (so that each replica is 0.1% of the original tensor), and $P = 12$ replicas are created for the analysis (so the overall cloud storage used for all replicas is 1.2% of the space needed to store the original tensor). $S = 3$ common anchor rows (instead of $S = 2$, which is the minimum possible) are used to fix the permutation and scaling ambiguity. These parameter choices satisfy PARACOMP identifiability conditions without much additional slack. The standard deviation of the measurement noise is nominally set to $\sigma = 0.01$.

Figure 3 shows the total squared error for estimating $\mathbf{A}$, i.e., $\|\mathbf{A} - \hat{\mathbf{A}}\|_2^2$, where $\hat{\mathbf{A}}$ denotes the estimate of $\mathbf{A}$ obtained using PARACOMP, as a function of $L = M = N$. The baseline is the total squared error attained by directly fitting the uncompressed $500 \times 500 \times 500$ tensor using a mature tensor decomposition algorithm (COMFAC, available at www.ece.umn.edu/~nikos)—the size of the uncompressed tensor used here makes such direct fitting possible, for comparison purposes. We see that PARACOMP yields respectable accuracy with only 1.2% of the full data, and is just an order of magnitude worse than the baseline algorithm when $L = M = N = 150$, corresponding to 32% of the full data. This is one way we can tradeoff memory/storage/computation versus estimation accuracy in the PARACOMP framework: by controlling the size of each replica. Another way to tradeoff memory/storage/computation for accuracy is through P. Figure 4 shows accuracy as a function of the number of replicas (computation threads) P, for fixed $L = M = N = 50$. Finally, Figure 5 plots accuracy as a function of measurement noise variance σ^2, for $L = M = N = 50$ and $P = 12$.



[FIG3] **MSE as a function of $L = M = N$.**



[FIG4] MSE as a function of P, the number of replicas/parallel threads spawned.

THE COLOR OF COMPRESSED NOISE

Consider a noisy tensor $\underline{\mathbf{Y}} = \underline{\mathbf{X}} + \underline{\mathbf{Z}}$, where $\underline{\mathbf{Z}}$ denotes zero-mean additive white noise. In vectorized form, $\mathbf{y} = \mathbf{x} + \mathbf{z}$, with $\mathbf{y} := \text{vec}(\underline{\mathbf{Y}})$, $\mathbf{x} := \text{vec}(\underline{\mathbf{X}})$, and $\mathbf{z} := \text{vec}(\underline{\mathbf{Z}})$. After multiway compression, one obtains the reduced-size tensor $\underline{\mathbf{Y}}_c$, whose vectorized representation $\mathbf{y}_c := \text{vec}(\underline{\mathbf{Y}}_c) = (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{y} = (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{x} + (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{z}$. Let $\mathbf{z}_c := (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{z}$. Clearly, $E[\mathbf{z}_c] = 0$, and

$$\begin{aligned} E[\mathbf{z}_c \mathbf{z}_c^T] &= E[(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{z}\mathbf{z}^T(\mathbf{U} \otimes \mathbf{V} \otimes \mathbf{W})] \\ &= (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)E[\mathbf{z}\mathbf{z}^T](\mathbf{U} \otimes \mathbf{V} \otimes \mathbf{W}) \\ &= \sigma^2(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)(\mathbf{U} \otimes \mathbf{V} \otimes \mathbf{W}) \\ &= \sigma^2((\mathbf{U}^T\mathbf{U}) \otimes (\mathbf{V}^T\mathbf{V}) \otimes (\mathbf{W}^T\mathbf{W})), \end{aligned}$$

where we have used two properties of the Kronecker product: transposition

$$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T,$$

and the mixed product rule [29]

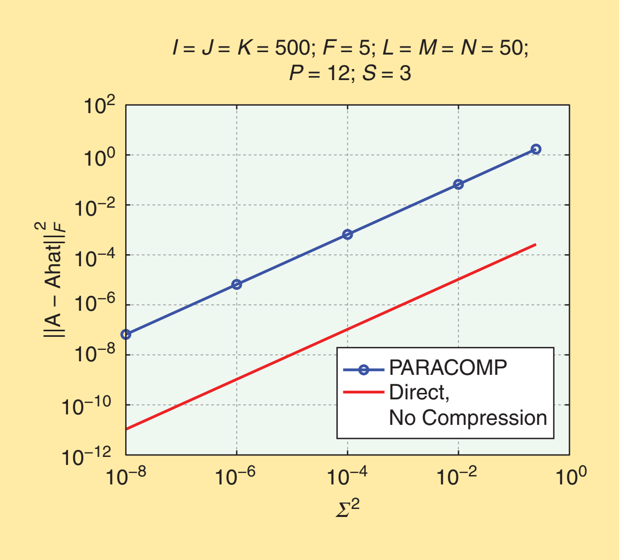
$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD}).$$

We see that, if $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ are orthonormal, then the noise in the compressed domain is white. Note that, for large I and $\mathbf{U}$ drawn from a zero-mean unit-variance uncorrelated distribution, $\mathbf{U}^T\mathbf{U} \approx \mathbf{I}$ by the law of large numbers. Furthermore, even if $\mathbf{z}$ is not Gaussian, $\mathbf{z}_c$ will be approximately Gaussian for large IJK, by the central limit theorem. From these, it follows that least-squares fitting is approximately optimal in the compressed domain, even if it is not so in the uncompressed domain. Compression thus makes least-squares fitting universal!

SUMMARY AND TAKE-HOME POINTS

SUMMARY

We have reviewed the basics of tensors and tensor decomposition, and presented a novel architecture for parallel and distributed computation of low-rank tensor decomposition that is especially well suited for big tensors. It is based on parallel processing of a set of randomly compressed, reduced-size replicas of the big tensor. We have also provided a friendly introduction to Hadoop–MapReduce, starting from a toy signal processing problem, and going up to sketching a Hadoop implementation of tensor decomposition in the cloud.



[FIG5] MSE as a function of additive white Gaussian noise variance σ^2.

MOTIVATION AND IMPACT

There is rapidly growing interest in signal processing for big data analytics, and in porting/translating and developing new signal processing algorithms for cloud computing platforms. Tensors are multidimensional signals that have found numerous applications in signal processing, machine learning, data mining, and well beyond (psychology, chemistry, life sciences, etc.), and they are becoming increasingly important for online marketing, social media, search engines, and many more applications. Tensors easily grow to be really big, as their total size is the product of mode sizes, hence exponential in the number of modes (dimensions in signal processing parlance). Big tensor data will thus be a big part of big data.

TAKE-HOME POINTS

1) PARACOMP enables massive parallelism with guaranteed identifiability properties: if the big tensor is indeed of low rank and the system parameters are appropriately chosen, then the rank-one factors of the big tensor will indeed be recovered from the analysis of the reduced-size replicas.

2) PARACOMP affords memory/storage and complexity gains of order up to (IJ/F) for a big tensor of size $I \times J \times K$ of rank F.

IS COMPONENT ORDERING PRESERVED AFTER COMPRESSION?

Consider randomly compressing a rank-one tensor $\underline{\mathbf{X}} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$, written in vectorized form as $\mathbf{x} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$ (recall that the Kronecker product $\otimes$ and the Khatri–Rao product $\odot$ coincide when all arguments involved are vectors). The compressed tensor is $\underline{\tilde{\mathbf{X}}}$, in vectorized form

$$\begin{aligned} \tilde{\mathbf{x}} &= (\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T)(\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}) \\ &= (\mathbf{U}^T\mathbf{a}) \otimes (\mathbf{V}^T\mathbf{b}) \otimes (\mathbf{W}^T\mathbf{c}), \end{aligned}$$

using the mixed product rule [29]. It follows

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_2^2 &= \tilde{\mathbf{x}}^T\tilde{\mathbf{x}} \\ &= ((\mathbf{a}^T\mathbf{U}) \otimes (\mathbf{b}^T\mathbf{V}) \otimes (\mathbf{c}^T\mathbf{W}))((\mathbf{U}^T\mathbf{a}) \otimes (\mathbf{V}^T\mathbf{b}) \otimes (\mathbf{W}^T\mathbf{c})) \\ &= (\mathbf{a}^T\mathbf{U}\mathbf{U}^T\mathbf{a}) \otimes (\mathbf{b}^T\mathbf{V}\mathbf{V}^T\mathbf{b}) \otimes (\mathbf{c}^T\mathbf{W}\mathbf{W}^T\mathbf{c}) \\ &= \|\mathbf{U}^T\mathbf{a}\|_2^2 \|\mathbf{V}^T\mathbf{b}\|_2^2 \|\mathbf{W}^T\mathbf{c}\|_2^2, \end{aligned}$$

where we have used the transposition and mixed product rules, and that the Kronecker product of scalars is a plain product. Notice that for our choice of $\mathbf{U}$ (i.i.d. zero-mean Gaussian of variance 1, i.e., `randn(I,L)` in Matlab), $\mathbf{U}^T\mathbf{U} \approx \mathbf{I}_{L\times L}$, but $\mathbf{U}\mathbf{U}^T$ is rank-deficient ($L < I$), thus far from $\mathbf{I}_{I\times I}$. However, considering one generic element of $\mathbf{U}^T\mathbf{a}$, say $\mathbf{u}^T\mathbf{a}$, and its magnitude-square, note that $|\mathbf{u}^T\mathbf{a}|^2 = \mathbf{a}^T\mathbf{u}\mathbf{u}^T\mathbf{a}$, so

$$E\left[|\mathbf{u}^T\mathbf{a}|^2\right] = \mathbf{a}^T E\left[\mathbf{u}\mathbf{u}^T\right]\mathbf{a} = \mathbf{a}^T\mathbf{a} = \|\mathbf{a}\|_2^2.$$

Next, it can be shown that

$$\mathrm{Var}\left[|\mathbf{u}^T\mathbf{a}|^2\right] = 2\|\mathbf{a}\|_2^4.$$

So now, looking at $\|\mathbf{U}^T\mathbf{a}\|_2^2$,

$$E\left[\|\mathbf{U}^T\mathbf{a}\|_2^2\right] = L\|\mathbf{a}\|_2^2,$$

and, since the different rows of $\mathbf{U}^T$ are independent, hence variance adds up

$$\mathrm{Var}\left[\|\mathbf{U}^T\mathbf{a}\|_2^2\right] = L2\|\mathbf{a}\|_2^4.$$

So $\|\mathbf{U}^T\mathbf{a}\|_2^2$ has mean^2/variance ('SNR') of $(L/2)$.

Turning to $\|\tilde{\mathbf{x}}\|_2^2 = \|\mathbf{U}^T\mathbf{a}\|_2^2\|\mathbf{V}^T\mathbf{b}\|_2^2\|\mathbf{W}^T\mathbf{c}\|_2^2$, it can be shown that it has mean

$$E\left[\|\tilde{\mathbf{x}}\|_2^2\right] = LMN\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2\|\mathbf{c}\|_2^2,$$

and mean^2/variance ('SNR')

$$\frac{(E[\|\tilde{\mathbf{x}}\|_2^2])^2}{\mathrm{Var}[\|\tilde{\mathbf{x}}\|_2^2]} = \frac{L^2M^2N^2}{(L^2+2L)(M^2+2M)(N^2+2N) - L^2M^2N^2}.$$

Assuming without loss of generality that $L \leq M \leq N$, this SNR is of order $(L/2)$. What this means is that, for moderate L, M, N and beyond, the Frobenious norm of a compressed rank-one tensor component (= Euclidean norm of the corresponding vectorized representation) is approximately proportional to the Frobenious norm of the uncompressed rank-one tensor component of the original tensor. In other words: compression approximately preserves component ordering. This is important because it implies that low-rank least-squares approximation of the compressed tensor approximately corresponds to low-rank least-squares approximation of the big tensor. The result also suggests that it may be possible to match the component permutations across replicas simply by sorting component energies. These are ignored in the permutation-matching procedure discussed in the main text, due to the normalization needed to account for the scaling ambiguity. Including energy in the matching process will enhance robustness to noise. It seems intriguing to try rank (principal component) deflation in this context, but we will pursue this elsewhere due to space limitations in the article.

No sparsity is required, although such sparsity can be exploited to improve memory, storage, and computational savings.

3) We have shown that using white noiselike compression matrices

- approximately preserves component ordering
- ensures that the compressed noise is approximately white if the original measurement noise is white
- makes the compressed noise look Gaussian, rendering classical least-squares CP algorithms well suited for fitting the reduced-size replicas, even if the measurement noise in the big tensor is far from Gaussian.

4) Each replica is independently decomposed, and the results are joined via a master linear equation per tensor mode. The number of replicas and the size of each replica can be adjusted to fit the number of computing nodes and the memory available to each node, and each node can run its own CP software, depending on its computational capabilities. This flexibility is why PARACOMP is better classified as a computational architecture, as opposed to a method or algorithm.

AUTHORS

Nicholas D. Sidiropoulos (nikos@umn.edu) is a professor in the Department of Electrical and Computer Engineering at the University of Minnesota. He has over 15 years of experience in tensor decomposition and its applications. His research interests include topics in signal processing, communications, convex optimization and approximation of NP-hard problems, and cross-layer resource allocation for wireless networks. His current research focuses primarily on signal and tensor analytics, with applications in cognitive radio, big data, and preference measurement. He received the National Science Foundation/CAREER Award in 1998 and the IEEE Signal Processing Society (SPS) Best Paper Award in 2001, 2007, and 2011. He served as SPS Distinguished Lecturer (2008–2009) and as chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007–2008). He was an associate editor for *IEEE Transactions on Signal Processing* (2000–2006), *IEEE Signal Processing Letters* (2000–2002), and was on the editorial board of *IEEE Signal Processing Magazine* (2009–2011). He currently serves as an area editor of *IEEE Transactions on Signal Processing* (2012–present) and as an associate editor of *Signal Processing*. He received the 2010 SPS Meritorious Service Award.

Evangelos E. Papalexakis (epapalex@cs.cmu.edu) is a Ph.D. student in the Computer Science Department at Carnegie Mellon University (CMU). He earned a diploma and M.Sc. degree in electronic and computer engineering at the Technical University of Crete, Chania, Greece. He has considerable experience in tensor decomposition and applications, as well as parallel and distributed computations in Hadoop. He has published in *IEEE Transactions on Signal Processing* and the International Conference on Acoustics, Speech, and Signal Processing, as well as in prime computer science conferences and journals. He is also the liaison that connects the CMU and UMN groups in a joint National Science Foundation project on big tensor data and its applications in automated Web-based language learning and brain data mining.

Christos Faloutsos (christos@cs.cmu.edu) is a professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award from the 2006 IEEE International Conference on Data Mining, the SIGKDD Innovations Award (2010), 19 Best Paper Awards (including two "test of time" awards), and four teaching awards. He is an ACM fellow and has served as a member of the executive committee of SIGKDD. He has published over 200 refereed articles, 11 book chapters, and one monograph. He holds six patents and has given over 30 tutorials and over ten invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bioinformatics data. He has a long-term interest in tensor decompositions and their practical applications in data mining, having published numerous papers in the area.

REFERENCES

[1] N. Sidiropoulos, E. Papalexakis, and C. Faloutsos, "A parallel algorithm for big tensor decomposition using randomly compressed cubes (PARACOMP)," in *Proc. IEEE ICASSP* 2014, May 4–9, Florence, Italy.

[2] D. Nion, K. Mokios, N. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 6, pp. 1193–1207, 2010.

[3] C. Fevotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Exploring Music Contents*, ser. Lecture Notes in Computer Science, S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, Eds. Berlin: Springer, 2011, vol. 6684, pp. 102–115.

[4] N. Sidiropoulos, G. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Processing*, vol. 48, no. 3, pp. 810–823, 2000.

[5] N. Sidiropoulos, R. Bro, and G. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[6] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, "Parcube: Sparse parallelizable tensor decompositions." in *ECML/PKDD (1)*, ser. Lecture Notes in Computer Science, P. A. Flach, T. D. Bie, and N. Cristianini, Eds. Berlin: Springer, 2012, vol. 7523, pp. 521–536.

[7] R. Bro and N. Sidiropoulos, "Least squares regression under unimodality and non-negativity constraints," *J. Chemomet*., vol. 12, no. 4, pp. 223–247, July/Aug. 1998.

[8] A. Cichocki, D. Mandic, C. Caiafa, A.-H. Phan, G. Zhou, Q. Zhao, and L. De Lathauwer, "Multiway component analysis: Tensor decompositions for signal processing applications," *IEEE Signal Processing Mag*., to be published.

[9] B. W. Bader and T. G. Kolda, "Efficient MATLAB computations with sparse and factored tensors," *SIAM J. Sci. Comput*., vol. 30, no. 1, pp. 205–231, Dec. 2007.

[10] T. G. Kolda and J. Sun, "Scalable tensor decompositions for multi-aspect data mining," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM 2008)*, Dec. 2008, pp. 363–372.

[11] D. Nion and N. Sidiropoulos, "Adaptive algorithms to track the PARAFAC decomposition of a third-order tensor," *IEEE Trans. Signal Processing*, vol. 57, no. 6, pp. 2299–2310, 2009.

[12] A. Phan and A. Cichocki, "PARAFAC algorithms for large-scale problems," *Neurocomputing*, vol. 74, no. 11, pp. 1970–1984, 2011.

[13] N. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Processing Lett*., vol. 19, no. 11, pp. 757–760, 2012.

[14] E. Papalexakis, N. Sidiropoulos, and R. Bro, "From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors," *IEEE Trans. Signal Processing*, vol. 61, no. 2, pp. 493–506, 2013.

[15] N. Sidiropoulos, "Low-rank decomposition of multi-way arrays: A signal processing perspective," in *Proc. IEEE SAM Workshop*, Sitges, Barcelona, Spain, 2004, pp. 52–58.

[16] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Rev*., vol. 51, no. 3, pp. 455–500, 2009.

[17] A. Smilde, R. Bro, P. Geladi, and J. Wiley, *Multi-Way Analysis with Applications in the Chemical Sciences*. Hoboken, NJ: Wiley, 2004.

[18] P. Kroonenberg, *Applied Multiway Data Analysis*. Hoboken, NJ: Wiley, 2008.

[19] R. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Pap. Phonet*., vol. 16, pp. 1–84, Dec. 1970.

[20] R. Harshman, "Determination and proof of minimum uniqueness conditions for PARAFAC-1," *UCLA Working Pap. Phonet*., vol. 22, pp. 111–117, 1972.

[21] J. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[22] J. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra Appl*., vol. 18, no. 2, pp. 95–138, 1977.

[23] N. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemomet*., vol. 14, no. 3, pp. 229–239, 2000.

[24] T. Jiang and N. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints," *IEEE Trans. Signal Processing*, vol. 52, no. 9, pp. 2625–2636, 2004.

[25] A. Stegeman and N. Sidiropoulos, "On Kruskal's uniqueness condition for the CANDECOMP/PARAFAC decomposition," *Linear Algebra Appl*., vol. 420, no. 2–3, pp. 540–552, 2007.

[26] L. Chiantini and G. Ottaviani, "On generic identifiability of 3-tensors of small rank," *SIAM. J. Matrix Anal. Appl*., vol. 33, no. 3, pp. 1018–1037, 2012.

[27] A. Stegeman, J. ten Berge, and L. De Lathauwer, "Sufficient conditions for uniqueness in CANDECOMP/PARAFAC and INDSCAL with random component matrices," *Psychometrika*, vol. 71, no. 2, pp. 219–229, 2006.

[28] R. Bro, N. Sidiropoulos, and G. Giannakis. (1999). A fast least squares algorithm for separating trilinear mixtures, in *Proc. ICA99 Int. Workshop on Independent Component Analysis and Blind Signal Separation*, pp. 289–294. [Online]. Available: http://www.ece.umn.edu/~nikos/comfac.m

[29] J. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circuits Syst*., vol. 25, no. 9, pp. 772–781, 1978.

[30] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via minimization," *Proc. Nat. Acad. Sci*., vol. 100, no. 5, pp. 2197–2202, 2003.

[31] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[32] Apache. Hadoop. [Online]. Available: http://hadoop.apache.org/

[33] A. Hadoop. Word count example. [Online]. Available: http://wiki.apache.org/hadoop/WordCount

[34] U. Kang, E. Papalexakis, A. Harpale, and C. Faloutsos, "Gigatensor: Scaling tensor analysis up by 100 times-algorithms and discoveries," in *Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2012, pp. 316–324.

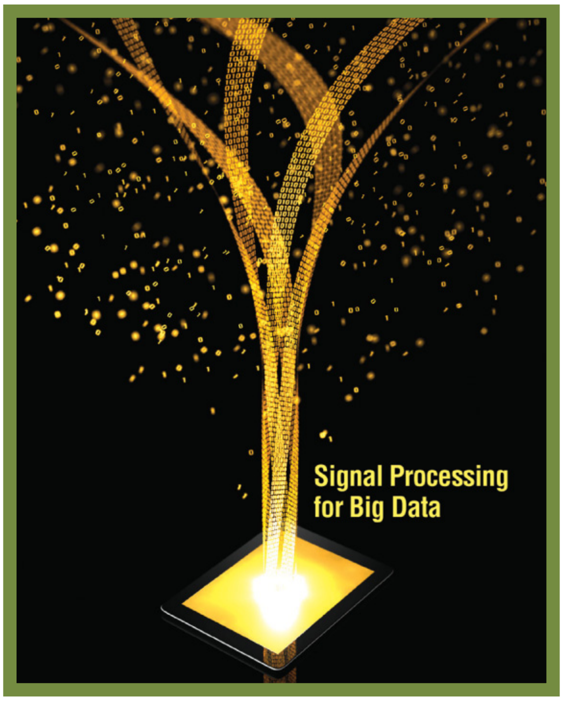
[35] U. Kang, B. Meeder, and C. Faloutsos, "Spectral analysis for billion-scale graphs: Discoveries and implementation," in *Advances in Knowledge Discovery and Data Mining*. New York: Springer, 2011, pp. 13–25.

[36] U. Kang, C. E. Tsourakakis, and C. Faloutsos, "Pegasus: A peta-scale graph mining system implementation and observations," in *Proc. 9th IEEE Int. Conf. Data Mining 2009* (*ICDM'09*), pp. 229–238.

[SP]

[Nico Vervliet, Otto Debals, Laurent Sorber, and Lieven De Lathauwer]

Breaking the Curse of Dimensionality Using Decompositions of Incomplete Tensors



© ISTOCKPHOTO.COM/TA2YO4NORI

[Tensor-based scientific computing in big data analysis]

Higher-order tensors and their decompositions are abundantly present in domains such as signal processing (e.g., higher-order statistics [1] and sensor array processing [2]), scientific computing (e.g., discretized multivariate functions [3]–[6]), and quantum information theory (e.g., representation of quantum many-body states [7]). In many applications, the possibly huge tensors can be approximated well by compact multilinear models or decompositions. Tensor decompositions are more versatile tools than the linear models resulting from traditional matrix approaches. Compared to matrices, tensors have at least one extra dimension. The number of elements in a tensor increases exponentially with the number of dimensions, and so do the computational and memory requirements. The exponential dependency (and the problems that are caused by it) is

Digital Object Identifier 10.1109/MSP.2014.2329429
Date of publication: 19 August 2014

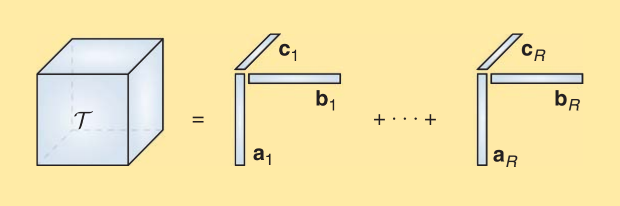
1053-5888/14©2014IEEE

called the *curse of dimensionality*. The curse limits the order of the tensors that can be handled. Even for a modest order, tensor problems are often large scale. Large tensors can be handled, and the curse can be alleviated or even removed by using a decomposition that represents the tensor instead of using the tensor itself. However, most decomposition algorithms require full tensors, which renders these algorithms infeasible for many data sets. If a tensor can be represented by a decomposition, this hypothesized structure can be exploited by using compressed sensing (CS) methods working on incomplete tensors, i.e., tensors with only a few known elements.

In domains such as scientific computing and quantum information theory, tensor decompositions such as the Tucker decomposition and tensor trains (TTs) have been successfully applied to represent large tensors. In the latter case, the tensor can contain more elements than the number of atoms in the universe [8] [estimated at $\mathcal{O}(10^{82})$]. Algorithms to compute these decompositions using only a few mode-n vectors (fibers) of the tensors have been developed to cope with the curse of dimensionality. In this tutorial, we show, on the one hand, how decompositions already known in signal processing [e.g., the canonical polyadic decomposition (CPD) and the Tucker decomposition] can be used for large and incomplete tensors and, on the other hand, how existing decompositions and techniques from scientific computing can be used in a signal processing context. We conclude with a convincing proof-of-concept case study from materials sciences, to our knowledge the first-known example of breaking the curse of dimensionality in data analysis.

NOTATION AND PRELIMINARIES

A general Nth-order tensor of size $I_1 \times I_2 \times \cdots \times I_N$ is denoted by a calligraphic letter as $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ and is a multidimensional array of numerical values $a_{i_1 i_2 \cdots i_N} = \mathcal{A}(i_1, i_2, \ldots, i_N)$. Tensors can be seen as a higher-order generalization of vectors (denoted by a bold, lowercase letter, e.g., $\mathbf{a}$) and matrices (denoted by a bold, uppercase letter, e.g., $\mathbf{A}$). In the same way matrices have rows and columns, tensors have mode-n vectors, which are constructed by fixing all but one index, e.g., $\mathbf{a} = \mathcal{A}(i_1, \ldots, i_{n-1}, :, \ i_{n+1}, \ldots, i_N)$. The mode-1 vectors are the columns of the tensor, and the mode-2 vectors are the rows of the tensor. More generally, an nth-order slice is constructed by fixing all but n indices. Tensors often need to be reshaped. An example is the mode-n matrix unfolding of a tensor $\mathcal{A}$, which arranges the mode-n vectors in a certain order as the columns of a matrix $\mathbf{A}_{(n)}$ [9], [10].



[FIG1] A PD of a third-order tensor $\mathcal{T}$ takes the form of a sum of R rank-1 tensors. If R is the minimum number for the equality to hold, the decomposition is called *canonical*, and R is the rank of the tensor.

A number of products have to be defined when working with tensors. The outer product of two tensors $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ and $\mathcal{B} \in \mathbb{C}^{J_1 \times J_2 \times \cdots \times J_M}$ is given as

$$(\mathcal{A} \otimes \mathcal{B})_{i_1 i_2 \cdots i_N j_1 j_2 \cdots j_M} = a_{i_1 i_2 \cdots i_N} \, b_{j_1 j_2 \cdots j_M}.$$

The mode-n tensor–matrix product between a tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ and a matrix $\mathbf{B} \in \mathbb{C}^{J \times I_n}$ is defined as

$$(\mathcal{A} \cdot_n \mathbf{B})_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum\nolimits_{i_n = 1}^{I_n} a_{i_1 i_2 \cdots i_N} b_{j i_n}.$$

The Hadamard product $\mathcal{A} * \mathcal{B}$ for $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ is the element-wise product. Finally, the Frobenius norm of a tensor $\mathcal{A}$ is denoted by $\| \mathcal{A} \|$ [9], [10].

TENSOR DECOMPOSITIONS

Most tensors of practical interest in applications are generated by some sort of process, such as a partial differential equation, a signal measured on a multidimensional grid, or the interactions between atoms. The resulting structure can be exploited by using decompositions, which approximate the tensor using only a small number of parameters. By using tensor decompositions instead of full tensors, the curse of dimensionality can be alleviated or even removed. We look into three decompositions in this tutorial: the CPD, the Tucker decomposition, and the TT decomposition. We conclude with a more general concept from scientific computing and quantum information theory called *tensor networks*. For more theory and applications, please see the "References" section, especially [4]–[6] and [9]–[11].

CANONICAL POLYADIC DECOMPOSITION

In a polyadic decomposition (PD), a tensor $\mathcal{T}$ is written as a sum of R rank-1 tensors (see Figure 1), each of which can be written as the outer product of N factor vectors $\mathbf{a}_r^{(n)}$:

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \otimes \mathbf{a}_r^{(2)} \otimes \cdots \otimes \mathbf{a}_r^{(N)} \triangleq [\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}]\!]. \qquad (1)$$

The latter notation is a shorthand for the PD, and the factor vectors $\mathbf{a}_r^{(n)}$ are the columns of the factor matrices $\mathbf{A}^{(n)}$ [9]. The PD is called *canonical* (CPD) when R is the minimum number of rank-1 terms needed for (1) to be exact. In this case, R is the CP rank of the tensor. Ignoring the trivial indeterminacies due to scaling and ordering of the rank-1 terms, the CPD is unique under mild conditions [12]. The decomposition has many names, such as the parallel factor model (PARAFAC, chemometrics) and the canonical decomposition (CANDECOMP, psychometrics) or R-term representation (scientific computing) [4], [9].

The number of free parameters in this decomposition is only $R\left(\left(\sum_{n=1}^{N} I_N\right) - N + 1\right)$ (because of a scaling indeterminacy in the decomposition), which is $\mathcal{O}(NIR)$, assuming $I_n = I$, $n = 1, \ldots, N$. More importantly, it is linear in the number of dimensions N. This means the curse of dimensionality can be broken by using a CPD instead of a full tensor if the tensor admits a good CPD [13]. In many practical cases in signal processing, R is small, and $R \ll I$. In cases where the rank R cannot be derived from the problem definition, finding the rank is a hard problem.

In practice, many CPDs will be fitted to the data until a sufficiently small approximation error is attained [13]. However, there is no guarantee that this process yields the CP rank R, as the best rank-R approximation may not exist. This is because the set of rank-R tensors is not closed, which means a sequence of rank-R' tensors with $R' < R$ can converge to a rank-R tensor, while two or more terms grow without bounds. This problem is referred to as *degeneracy* [9], [14], [15]. By imposing constraints such as non-negativity or orthogonality on the factor matrices, degeneracy can be avoided [10], [14], [16].

TUCKER DECOMPOSITION AND LOW MULTILINEAR RANK APPROXIMATION

The Tucker decomposition of a tensor $\mathcal{T}$ is given as a multilinear transformation (see Figure 2) of a typically small core tensor $\mathcal{G} \in \mathbb{C}^{R_1 \times R_2 \times \cdots \times R_N}$ by factor matrices $\mathrm{A}^{(n)} \in \mathbb{C}^{I_n \times R_n}$, $n = 1, \ldots, N$:

$$\mathcal{T} = \mathcal{G} \bullet_1 \mathrm{A}^{(1)} \bullet_2 \mathrm{A}^{(2)} \cdots \bullet_N \mathrm{A}^{(N)} \triangleq [\![\mathcal{G}; \mathrm{A}^{(1)}, \mathrm{A}^{(2)}, \ldots, \mathrm{A}^{(N)}]\!], \qquad (2)$$

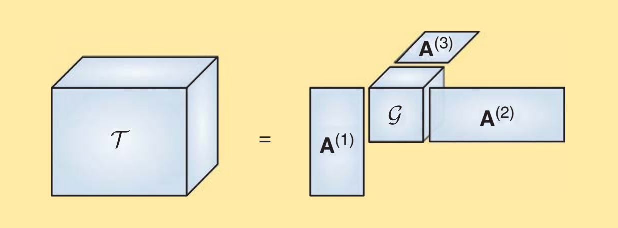
where the latter is a shorthand notation [9]. The N-tuple $(R_1, R_2, \ldots, R_N)$ for which the core size is minimal is called the *multilinear rank*. R_1 is the dimension of the column space, R_2 is the dimension of the row space, and, more generally, R_n is the dimension of the space spanned by the mode-n vectors [9]. In general, the Tucker decomposition (2) is not unique, but the subspaces spanned by the vectors in the factor matrices are, which is useful in certain applications [9], [10]. In its original definition, the Tucker decomposition imposed orthogonality and ordering constraints on the factor matrices and the core tensor. In this definition, the Tucker decomposition can be interpreted as a higher-order generalization of the singular value decomposition (SVD) and can be obtained by reliable algorithms from numerical linear algebra (in particular, algorithms for computing the SVD). In this context, the names *multilinear SVD* and *higher-order SVD* are also used [17].

The number of parameters in the Tucker decomposition is $\mathcal{O}(NIR + R^N)$ when we take $I_n = I$ and $R_n = R$, $n = 1, \ldots, N$. This means the number of parameters in a Tucker decomposition still depends exponentially on the number of dimensions N. The curse of dimensionality is alleviated, however, as typically $R \ll I$. More generally, when a tensor is approximated by (2) where the size of the core tensor is chosen by the user, this decomposition is called a *low multilinear rank approximation* (*LMLRA*). As in PCA, a Tucker decomposition can be compressed or truncated by omitting small multilinear singular values [9], [10]. This reduction in R is beneficial, given the exponential factor $\mathcal{O}(R^N)$ in the number of parameters, as the total number of parameters decreases exponentially. Note that a truncated Tucker decomposition is just one, not necessarily optimal, way to obtain an LMLRA [17].

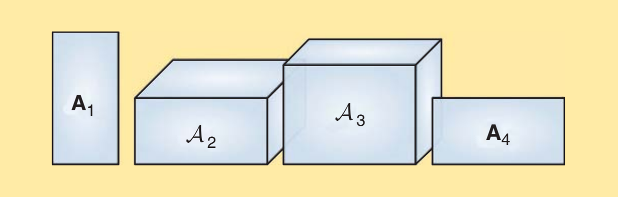
TENSOR TRAINS

TTs are a concept from scientific computing and from quantum information theory, where they are known as *matrix product states* [3], [5]–[7]. Each element in a tensor $\mathcal{T}$ can be written as

$$t_{i_1 i_2 \cdots i_N} = \sum_{r_1, r_2, \ldots, r_{N-1}} a^{(1)}_{i_1 r_1} a^{(2)}_{r_1 i_2 r_2} \cdots a^{(N)}_{r_{N-1} i_N},$$



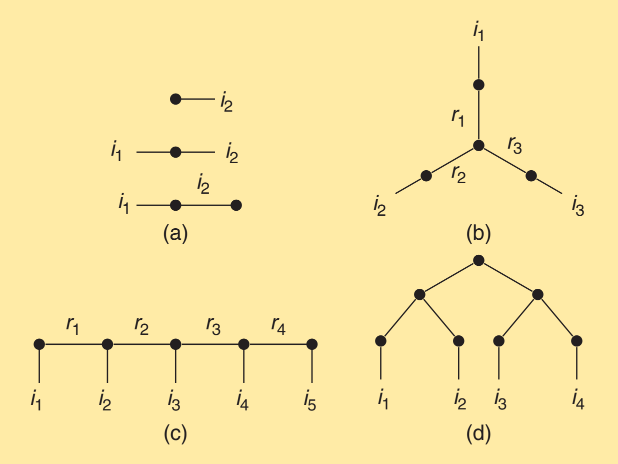
[FIG2] The Tucker decomposition of a third-order tensor $\mathcal{T}$ involves a multilinear transformation of a core tensor $\mathcal{G}$ by factor matrices $\mathbf{A}^{(n)}$, $n = 1, \ldots, N$.



[FIG3] A fourth-order tensor $\mathcal{T}$ can be written as a TT by linking a matrix $\mathbf{A}^{(1)}$; two tensors, $\mathcal{A}^{(2)}$ and $\mathcal{A}^{(3)}$ (the carriages); and a matrix $\mathbf{A}^{(4)}$.

with $r_n = 1, \ldots, R_n$, $n = 1, \ldots, N-1$. The matrices $\mathrm{A}^{(1)} \in \mathbb{C}^{I_1 \times R_1}$ and $\mathrm{A}^{(N)} \in \mathbb{C}^{R_{N-1} \times I_N}$ are the "head" and "tail" of the train; the core tensors $\mathcal{A}^{(n)} \in \mathbb{C}^{R_{n-1} \times I_n \times R_n}$, $n = 2, \ldots, N-1$, are the carriages, as can be seen in Figure 3. The auxiliary indices R_n, $n = 1, \ldots, N-1$ are called the *compression ranks* or the *TT ranks* [3]. It can be proven that the compression ranks are bounded by the CP rank of a tensor [18].

A TT combines the good properties of the CPD and the Tucker decomposition. The number of parameters in a TT is $\mathcal{O}(2IR + (N-2)IR^2)$, assuming $I_n = I$, $R_n = R$, $n = 1, \ldots, N$, which is linear in the number of dimensions, similar to a CPD [3], [6]. This means a TT is suitable for high-dimensional problems, as using it removes the curse of dimensionality. As for the Tucker decomposition, numerically reliable algorithms based on the SVD can be used to compute the decomposition [3], [17].



[FIG4] Different types of tensor networks: (a) a vector, a matrix, and their matrix–vector product (a contraction); (b) a Tucker decomposition; (c) a TT decomposition; and (d) an HT decomposition.

TENSOR NETWORKS

The TT decomposition represents a higher-order tensor as a set of linked (lower-order) tensors and matrices, and it is an example of a linear tensor network. A more general tensor network is a set of interconnected tensors. This can be visualized using tensor network diagrams (see Figure 4) [4], [7]. Each vector, matrix, or tensor is represented as a dot. The order of each tensor is determined by the number of edges connected to it. An interconnection between two dots represents a contraction, which is the summation of the products over a common index. Tensor network diagrams are an intuitive and visual way to efficiently represent decompositions of higher-order tensors. An example is the hierarchical Tucker (HT) decomposition (see Figure 4), which is another important decomposition used in scientific computing [4]–[6]. More complicated tensor networks can also contain cycles, e.g., tensor chains and projected entangled-pair states from quantum physics [5], [7].

COMPUTING DECOMPOSITIONS OF LARGE, INCOMPLETE TENSORS

To compute tensor decompositions, most algorithms require a full tensor and are therefore not an option for large and high-dimensional data sets. The knowledge that the data are structured and can be represented by a small number of parameters can be exploited by sampling the tensor in only a few elements. Then, the decomposition is calculated using an incomplete tensor. There are two important situations in which incomplete data sets are used. In the first case, some elements are unknown, e.g., because of a broken sensor [19], or unreliable, e.g., because of Rayleigh scattering [15], and the matrix or tensor needs to be completed [20]. In the second case, the cost of acquiring a full tensor is too high in terms of money, time, or storage requirements. By sampling the tensor in only a few elements, this cost can be reduced.

CS methods are used to reconstruct signals using only a few measurements taken by a linear projection of the original data set [21]. Many extensions of these methods to tensors have been developed [10], [22], and new methods tailored to tensors have emerged, e.g., [23] and [24]. In this tutorial, we focus on a class of CS methods where decompositions of very large tensors are computed using only a small number of known elements. In particular, we first discuss methods to compute a CPD from a randomly sampled incomplete tensor. Then, we discuss how matrices can be approximated by extracting only a few rows and columns. This idea can be extended to tensors, and we conclude by elaborating on two mode-n vector sampling methods: one for the TT decomposition and the other for the LMLRA.

OPTIMIZATION-BASED ALGORITHMS

Most algorithms to compute a CPD use optimization to find the factor matrices $\mathrm{A}^{(n)}$:

$$\min_{\mathrm{A}^{(1)},\mathrm{A}^{(2)},\ldots,\mathrm{A}^{(N)}} \frac{1}{2}\left\| \mathcal{T} - [\![\mathrm{A}^{(1)}, \mathrm{A}^{(2)}, \ldots, \mathrm{A}^{(N)}]\!] \right\|^2, \tag{3}$$

which is a least-squares problem in each factor matrix separately. The popular alternating least-squares (ALS) method alternately solves a least-squares problem for one factor matrix while fixing the others. This method is easy to implement and works well in many cases but has a linear convergence rate and tends to be slow when the factor vectors become more aligned. It is even possible that the algorithm does not converge at all [25], [26]. CP-OPT uses a nonlinear conjugate gradients method to solve (3) [26]. By using first-order information, the method also achieves linear convergence. Recently, some new methods based on nonlinear least-squares (NLS) algorithms have been developed. These methods exploit the structure in the objective function's approximate Hessian. Because of the NLS framework, the second-order convergence can be attained under certain circumstances [25], [27]. The latter two methods are both guaranteed to converge to a stationary point, which can be a local optimum, however.

Although efficient methods exist, the complexity of all methods working on full tensors is at least $\mathcal{O}(I^N)$, which becomes infeasible for large, high-dimensional tensors. To handle missing data, (3) can be adapted to

$$\min_{\mathrm{A}^{(1)},\mathrm{A}^{(2)},\ldots,\mathrm{A}^{(N)}} \frac{1}{2}\left\| \mathcal{W} * \left(\mathcal{T} - [\![\mathrm{A}^{(1)}, \mathrm{A}^{(2)}, \ldots, \mathrm{A}^{(N)}]\!]\right) \right\|^2,$$

where $\mathcal{W} \in \{0,1\}^{I_1 \times I_2 \times \cdots \times I_N}$ is a binary observation tensor with a 1 for every known element [19]. The popular ALS method has been extended by using an expectation maximization (EM) framework to impute each missing value with a value from the current CP model [15]. Because of the imputation, the ALS-EM method still suffers from the curse of dimensionality. The CP-WOPT (weighted CP-OPT) method is an extension of the CP-OPT method and uses only the known elements, thereby relaxing the curse of dimensionality [19]. Adaptions of the Jacobian and the Gramian of the Jacobian for incomplete tensors can be used in an inexact NLS framework. Second-order convergence can again be attained under certain circumstances, while the computational complexity is still linear in the number of known elements [28].

The distribution of the known elements in the tensor can be random, although performance may decrease in case of missing mode-n vectors or slices [19], [28]. The elements should be known a priori, contrary to the mode-n vector based algorithms. Constraints such as nonnegativity or a Vandermonde structure can easily be added to the factor matrices, which is useful for many signal processing applications [16].

PSEUDOSKELETON APPROXIMATION FOR MATRICES

Instead of randomly sampling a tensor, a more drastic approach can be taken by sampling only mode-n vectors and only using these mode-n vectors in the decomposition. These techniques originate from the pseudoskeleton approximation or the CUR decomposition for matrices. These decompositions state that a matrix $\mathrm{A} \in \mathbb{C}^{I_1 \times I_2}$ of rank R can be approximated using only R columns and R rows of this matrix:

$$\mathrm{A} = \mathrm{CGR}, \tag{4}$$

where $\mathrm{C} \in \mathbb{C}^{I_1 \times R}$ has R columns with indices $\mathcal{J}$ of A, i.e., $\mathrm{C} = \mathrm{A}(:, \mathcal{J})$, $\mathrm{R} \in \mathbb{C}^{R \times I_2}$ has R rows with indices $\mathcal{I}$ of A, i.e., $\mathrm{R} = \mathrm{A}(\mathcal{I}, :)$ and $\mathrm{G} = \hat{\mathrm{A}}^{-1}$, where $\hat{\mathrm{A}}$ contains the intersection of C and $\mathbf{R}$, i.e., $\hat{\mathbf{A}} = \mathbf{A}(\mathcal{I}, \mathcal{J})$. If $\operatorname{rank}(\mathbf{A}) = R$, then (4) is exact when $\mathbf{C}$ has R linearly independent columns of $\mathbf{A}$ and $\mathbf{R}$ has R linearly independent rows of $\mathbf{A}$ (which implies that $\hat{\mathbf{A}}$ is nonsingular) [8]. Usually, we are interested in the case where $\operatorname{rank}(\mathbf{A}) > R$. The best choice for the submatrix $\hat{\mathbf{A}}$ (and consequently $\mathbf{C}$ and $\mathbf{R}$) is, in this case, the $R \times R$ submatrix having the largest volume, which is given by the modulus of its determinant [29].

To determine the optimal submatrix $\hat{\mathbf{A}}$, the determinants of all possible submatrices have to be evaluated. This is computationally challenging, and, moreover, all the elements of the matrix have to be known. A heuristic called *cross approximation* (*CA*) can be used to calculate a quasi-optimal maximal volume submatrix by only looking at a few rows and columns. The following general scheme can be used (based on [30]). An initial column index set $\mathcal{J} \subset \{1, 2, \ldots, I_2\}$ and an initial row index set $\mathcal{I} \subset \{1, 2, \ldots, I_1\}$ are chosen, and $\mathbf{C}$ is defined as $A(:, \mathcal{J})$. Then, the submatrix $\hat{\mathbf{C}} = \mathbf{C}(\hat{\mathcal{I}}, :)$ with (approximately) the largest volume is calculated, e.g., using a technique based on full pivoting [30]. Next, the process is repeated for the rows, i.e., the subset $\hat{\mathcal{J}}$ resulting in the maximal volume submatrix $\hat{\mathbf{R}} = \mathbf{R}(:, \hat{\mathcal{J}})$ in $\mathbf{R} = \mathbf{A}(\mathcal{I}, :)$ is calculated. Next, the index sets are updated as $\mathcal{J} = \mathcal{J} \cup \hat{\mathcal{J}}$ and $\mathcal{I} = \mathcal{I} \cup \hat{\mathcal{I}}$, and the process is repeated until a stopping criterion is met, e.g., when the norm of the residual $\|\mathbf{A} - \mathbf{CGR}\|$ is small enough. To calculate the norm, only the extracted rows and columns are taken into account. To make this more concrete, we give a simple method selecting one column and row at a time [31]:

1) Set $\mathcal{J} = \emptyset$, $\mathcal{I} = \emptyset$, $j_1 = 1$, and $p = 1$.
2) Extract the column $\mathbf{A}(:, j_p)$ and find the maximal volume submatrix in the residual, i.e., the largest element in modulus in the vector $\mathbf{a}_{j_p}$ minus the corresponding elements in the already known rank-1 terms, and set i_p to its location.
3) Extract the row $\mathbf{A}(i_p, :)$ and find the maximal volume submatrix in the residual that is not in the previously chosen column j_p, and set j_{p+1} to its location.
4) Set $\mathcal{J} = \mathcal{J} \cup \{j_p\}$, $\mathcal{I} = \mathcal{I} \cup \{i_p\}$.
5) If the stopping criterion is not satisfied, set $p = p + 1$ and go to step 2.

For more details, we refer to [8], [30], and [31].

CROSS APPROXIMATION FOR TT

Before we outline a CA-based algorithm for the TT decomposition, we first present a simplified version of a TT algorithm for full tensors based on repeated truncated SVD [8]:

1) Set $R_0 = 1$ and $\mathbf{M} = \operatorname{reshape}\left(\mathcal{T}, \left[I_1, \prod_{k=2}^{N} I_k\right]\right)$.
2) For $n = 1, \ldots, N-1$:
   a) Calculate the (truncated) SVD: $\mathbf{M} = \mathbf{U\Sigma V}^{\mathrm{H}}$ (with $^{\mathrm{H}}$ being the conjugated transpose).
   b) Set $\mathcal{A}^{(n)} = \operatorname{reshape}(\mathbf{U}, [R_{n-1}, I_n, R_n])$, and if $n < N-1$, set $\mathbf{M} = \operatorname{reshape}\left(\mathbf{\Sigma V}^{\mathrm{H}}, \left[R_n I_{n+1}, \prod_{k=n+2}^{N} I_k\right]\right)$.
3) $\mathbf{A}^{(N)} = \mathbf{\Sigma V}^{\mathrm{H}}$.

The truncation step 2a) determines the compression rank R_n. In a scientific-computing context, the compression ranks R_n are chosen such that the decomposition approximates the (noise free) tensor with a user-defined accuracy ϵ [3]. In signal processing, the tensor is often perturbed by noise. Therefore, the compression ranks R_n can be determined by using a procedure estimating the noise level.

The use of the SVD in 2a) has two disadvantages: all the elements in the tensor need to be known, and when this tensor is large, calculating the SVD is expensive. In [8], a CA-type method is suggested: the SVD can be replaced by a pseudoskeleton approximation as described above by replacing $\mathbf{U}$ with $\mathbf{CG}$ and $\mathbf{\Sigma V}^{\mathrm{H}}$ with $\mathbf{R}$. The matrix $\mathbf{R}$ does not require extra calculations, and working with $\mathbf{R}$ does not require additional memory as it can be handled implicitly by selecting the proper indices. This algorithm requires the compression ranks to be known in advance. By using a compression or rounding algorithm on the resulting TT decomposition, the compression rank can be overestimated safely (see, e.g., [3] for a compression method based on the truncated SVD). In a signal processing context, this rounding algorithm can be adapted to use a noise-level estimation procedure instead of using a user-defined accuracy ϵ. For practical implementation details, we refer to [8].

The positions of the elements needed by the CA algorithm are unknown a priori but are generated based on information in the mode-n vectors that already have been extracted. In a scientific-computing context, where the tensor is often given as a multivariate function, this is not a problem as sampling an entry is evaluating this function. In a signal processing context, this means that either the elements are sampled while running the CA algorithm, or the full tensor has to be known a priori. This last condition can be relaxed, however, by imputing unknown elements in selected mode-n vector by an estimate of the value of these elements, e.g., the mean value over the mode-n vector. But this only works well if the imputed value effectively is a good estimator of the unknown value. Only $\mathcal{O}(2KNR)$ columns of length $R_{n-1}I_n$ are investigated during the CA algorithm, where K is the number of iterations in the CA algorithm and assuming $R_n = R$, $n = 1, \ldots, N-1$. If the compression ranks and the number of iterations are low, very few elements need to be sampled.

CROSS APPROXIMATION FOR LMLRA

The CA method for TT essentially only replaced the SVD with a pseudoskeleton approximation. In the case of the LMLRA, we look at another generalization of the pseudoskeleton approximation method: $\mathcal{T}$ can be approximated by

$$\hat{\mathcal{T}} = [\![\mathcal{G}; \mathbf{C}^{(1)}\mathbf{G}_{(1)}^{\dagger}, \ldots, \mathbf{C}^{(N)}\mathbf{G}_{(N)}^{\dagger}]\!],$$

where $\mathbf{C}^{(n)}$ contains $\prod_{m \neq n} R_m$ mode-n vectors for $n = 1, \ldots, N$ and where the size of the core tensor $\mathcal{G}$ is $R_1 \times \cdots \times R_N$. The core tensor is the subtensor of $\mathcal{T}$ containing the intersection of the selected mode-n vectors. More concretely, we define the index sets $\mathcal{I}^{(n)} \subset \{1, \ldots, I_n\}$, $n = 1, \ldots, N$. Each column of $\mathbf{C}^{(n)}$ contains a selected mode-n vector defined by an index set in $\times_{m \neq n}\mathcal{I}^{(m)}$, i.e.,

$$\mathcal{T}(i_1, \ldots, i_{n-1}, :, i_{n+1}, \ldots, i_N)$$

with

$$(i_1, \ldots, i_{n-1}, i_{n+1}, \ldots, i_N) \in \times_{m \neq n}\mathcal{I}^{(m)}.$$

$\mathbf{C}^{(n)}$ thus is a matricized $(R_1 \times \cdots \times R_{n-1} \times I_n \times R_{n+1} \times \cdots \times R_N)$ subtensor of $\mathcal{T}$. The intersection core tensor then is defined as

[TABLE 1] THE NUMBER OF PARAMETERS AND TOUCHED ELEMENTS FOR THE THREE DECOMPOSITIONS OF INCOMPLETE TENSORS. THE NUMBER OF TOUCHED ELEMENTS CONCERN THE PRESENTED ALGORITHMIC VARIANTS. IN THE CASE OF A CPD AND TT, THE CURSE OF DIMENSIONALITY CAN BE OVERCOME.

	NUMBER OF PARAMETERS	NUMBER OF TOUCHED ELEMENTS
CPD	$\mathcal{O}(NIR)$	N_{samples}
LMLRA	$\mathcal{O}(NIR + R^N)$	$\mathcal{O}(NIR^{N-1})$
TT	$\mathcal{O}(2IR + (N-2)IR^2)$	$\mathcal{O}(2KNR^2I)$

$\mathcal{G} = \mathcal{T}(\mathcal{I}^{(1)}, \dots, \mathcal{I}^{(N)})$. To determine the index sets $\mathcal{I}^{(n)}$, an adaptive procedure can be used. Each iteration, the index $i^{(n)}$ having the largest modulus of the residual in the mode-n vector through the pivot is added to $\mathcal{I}^{(n)}$. The residual is defined as $\mathcal{E} = \mathcal{T} - \hat{\mathcal{T}}$, where the matrices $\mathrm{C}^{(n)}$, $n = 1, \dots, N$ and the core tensor $\mathcal{G}$ are defined by the current index sets $\mathcal{I}^{(n)}$. A simplified version of this fiber-sampling tensor decomposition algorithm [32] is given as:

1) Choose an initial mode-N vector in $\mathcal{T}$ defined by $i_1^{(n)}$, for $n = 1, \dots, N-1$ and set $i_1^{(N)}$ to the index containing the maximal modulus in this fiber.
2) Set $\mathcal{I}^{(n)} = \{i_1^{(n)}\}$ for $n = 1, \dots, N$ and set the pivot to $(i_1^{(1)}, \dots, i_1^{(N)})$.
3) For $r = 2, \dots, R$:
   a) For each mode $n = 1, \dots, N$:
      i) Select the index $i_r^{(n)}$ of the maximal modulus of the mode-n vector e going through the pivot in the residual tensor $\mathcal{E}$, i.e., in $\mathrm{e} = \mathcal{E}(i_r^{(1)}, \dots, i_r^{(n-1)}, :, i_{r-1}^{(n+1)}, \dots, i_{r-1}^{(N)})$; unless for $r = 2$ and $n = 1$, then we select the maximal modulus in $\mathcal{T}$ [32].
      ii) Set $\mathcal{I}^{(n)} = \mathcal{I}^{(n)} \cup \{i_r^{(n)}\}$ and select $(i_r^{(1)}, \dots, i_r^{(n)}, i_{r-1}^{(n+1)}, \dots, i_{r-1}^{(N)})$ as new pivot.

Each matrix $\mathrm{C}^{(n)}$ contains $|\times_{m \neq n} \mathcal{I}^{(m)}| = \mathcal{O}(R^{N-1})$ columns. The total number of mode-n vectors of length I_n that has to be extracted in this algorithm is then $\mathcal{O}(NR^{N-1})$. Similarly to the pseudoskeleton approximation, an exact decomposition based on fiber sampling can be attained if $\mathcal{T}$ has an exact LMLRA structure and multilinear rank $(R_1, \dots, R_N)$, i.e., $\mathcal{T} = [\![\mathcal{G}; \mathrm{A}^{(1)}, \dots, \mathrm{A}^{(N)}]\!]$. In this case, it can be proven that only R_n mode-n fibers per mode n and $\prod_{n=1}^{N} R_n$ core elements have to be extracted [32]. In both cases, the computational complexity still has an exponential dependence on the number of dimensions [32]. Even with CA, the representation of a tensor as an LMLRA is limited by the curse of dimensionality to low-dimensional problems. We can make the same remarks for this method as for the TT decomposition concerning the fact that the indices of the required elements are only known at runtime. A variation of this algorithm determines the largest element in slices instead of in mode-n vectors [31]. An alternative to the pseudoskeleton approach is to sample mode-n vectors after a fast estimation of probability densities [33].

CASE STUDIES

To illustrate the use of the decompositions and incomplete tensors, two case studies are reported. The first case study shows how the concepts can be applied in a signal processing context. To compare the results with full-tensor methods, moderate-size tensors are used. The second case study gives an example from materials sciences, where a huge tensor is decomposed while using only a very small fraction of the elements.

MULTIDIMENSIONAL HARMONIC RETRIEVAL

Multidimensional harmonic retrieval problems appear frequently in signal processing, e.g., in radar applications and channel sounding [34]. To model a multipath wireless channel, e.g., a broadband wireless channel sounder can be used to measure a (time-varying) channel in the time, frequency, and spatial domains. The measurement data can then be transformed into a tensor:

$$y_{i_1 i_2 \cdots i_D k} = \sum_{r=1}^{R} s_r(k) \prod_{d=1}^{D} \mathrm{e}^{\mathrm{j}(i_d - 1)\mu_r^{(d)}} + n_{i_1 i_2 \cdots i_D k}, \tag{5}$$

where $\mathrm{j}^2 = -1$ and $s_r(k)$ is the kth complex symbol carried by the rth multidimensional harmonic. The parameters $\mu_r^{(d)}$ are, e.g., the direction of departure, the direction of arrival, the Doppler shift, and the delay. (For more information, we refer the reader to [34].) The noise $n_{i_1 i_2 \cdots i_D k}$ is modeled as zero-mean independent and identically distributed additive Gaussian noise. We can rewrite the model as a CPD

$$\mathcal{Y} = [\![\mathrm{A}^{(1)}, \mathrm{A}^{(2)}, \dots, \mathrm{A}^{(D)}, \mathrm{S}]\!] + \mathcal{N}, \tag{6}$$

with the Vandermonde structured factor matrices $\mathrm{A}^{(d)} \in \mathbb{C}^{I_d \times R}$, $a_{i_d, r}^{(d)} = \mathrm{e}^{\mathrm{j}(i_d - 1)\mu_r^{(d)}}$, and $\mathrm{S} \in \mathbb{C}^{K \times R}$ with $s_{kr} = s_r(k)$. In the noiseless case, the CP rank of this tensor is equal to R. The multilinear ranks and the TT ranks are, at most, R. The uniqueness properties of (6) are given in [35].

To estimate the parameters $\mu_r^{(d)}$, a subspace-based approach is used. First, $\mathcal{Y}$ is decomposed using a CPD, an LMLRA or a TT. Then, in the case of the LMLRA and TT, we compute the subspaces $\mathrm{B}^{(d)}$ spanned by the mode-d vectors, $d = 1, \dots, D$. (The parameters can be estimated directly from the factor vectors in case of CPD.) Finally, we use a standard total-least-squares method to estimate the parameters $\hat{\boldsymbol{\mu}}^{(d)}$ from these subspaces (see [36]). Here, we focus on the first two steps, i.e., the approximation of the full or incomplete tensor $\mathcal{Y}$ by a CPD, an LMLRA, or a TT and the computation of the subspaces.

In the case of a CPD, we use the `cpd_nls` method from Tensorlab [25], [37] to get an estimate of the factor matrices $\hat{\mathrm{A}}^{(d)}$, $d = 1, \dots, D+1$. This method works on both full and incomplete tensors. Here, we can estimate the parameters directly as the generators of the noisy Vandermonde vectors $\mathrm{a}_r^{(d)}$, $d = 1, \dots, D$, so the computation of the subspaces is not necessary. The number of sampled entries N_{samples} can be chosen by the user (see Table 1).

In the case of an LMLRA $[\![\hat{\mathcal{G}}; \hat{\mathrm{A}}^{(1)}, \dots, \hat{\mathrm{A}}^{(D+1)}]\!]$, we first compute the decomposition using `lmlra` from Tensorlab [37], which uses an NLS-based optimization method on the full tensor, and using `lmlra_aca`, which implements a fiber-sampling adaptive cross-approximation technique. In the latter case, the choice of the core size $R_1 \times \cdots \times R_{D+1}$ controls the number of touched elements, i.e., the number of elements from the tensor that are used during the algorithm (see Table 1). Recall that R is the number of multidimensional harmonics in (5). The subspaces $\hat{\mathrm{B}}^{(d)}$ are now computed using the first R left singular vectors of the unfolded product $(\hat{\mathcal{G}} \cdot_d \hat{\mathrm{A}}^{(d)})_{(d)}$. (It can be verified that these are the dominant mode-d vectors of $[\![\hat{\mathcal{G}}; \hat{\mathrm{A}}^{(1)}, \ldots, \hat{\mathrm{A}}^{(D+1)}]\!]$ if the factor matrices are normalized to have orthonormal columns.)

Finally, in the case of TT, we compute the TT cores $\hat{\mathrm{A}}^{(1)}$, $\hat{\mathcal{A}}^{(d)}$, $\hat{\mathrm{A}}^{(D+1)}$ using `tt_full`, which uses the truncated SVD of the full tensor (cf. supra), and using `dmrg_cross`, which uses cross approximation and touches only a limited number of mode-n vectors (cf. supra). Both methods are available in the TT-Toolbox (http://spring.inm.ras.ru/osel/). The number of touched elements is controlled by the compression ranks R_n and the number of iterations K (see Table 1). The estimates for the subspaces $\hat{\mathrm{B}}^{(1)}$ and $\hat{\mathrm{B}}^{(d)}$ can be computed using the first R left singular vectors of $\hat{\mathrm{A}}^{(1)}$ and of the mode-2 unfolding $\hat{\mathcal{A}}^{(d)}_{(2)}$, $d = 2, \ldots, D$, respectively.

With two experiments, we show how the number of touched elements and noise influence the quality of the retrieved parameters using the three decompositions. We create an $8 \times 8 \times 8 \times 8 \times 20$ tensor $\mathcal{Y}$ with rank $R = 4$ according to (6). The $D = 4$ parameter vectors are chosen as follows: $\boldsymbol{\mu}^{(1)} = [1.0, -0.5, 0.1, -0.8]$, $\boldsymbol{\mu}^{(2)} = [-0.5, 1.0, -0.9, 1.0]$, $\boldsymbol{\mu}^{(3)} = [0.2, -0.6, 1.0, 0.4]$, and $\boldsymbol{\mu}^{(4)} = [-0.8, 0.4, 0.3, -0.1]$. Each of the $R = 4$ uncorrelated binary phase-shift keying sources takes $K = 20$ values. To evaluate the quality of the estimates, the root-mean-square (RMS) error is used:

$$E_{\mathrm{RMS}} = \sqrt{\frac{1}{RD} \sum_{r=1}^{R} \sum_{d=1}^{D} (\mu_r^{(d)} - \hat{\mu}_r^{(d)})^2}.$$

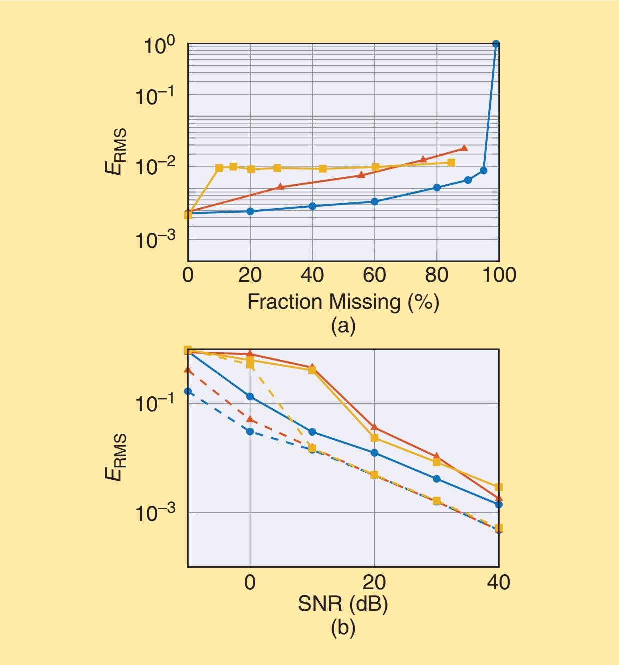
For each experiment, the median value more than 100 Monte Carlo runs is reported.

In the first experiment, the signal-to-noise ratio (SNR) is fixed to 20 dB, while the fraction of missing entries varies [see Figure 5(a)]. When there are no missing entries, we use the corresponding algorithms for full tensors. All methods then attain a similar accuracy. When the fraction of missing entries is increased, the error E_{RMS} also increases, except for TT, where the error remains almost constant but is higher than for CPD and LMLRA. An increase in the error is expected, as there are fewer noisy samples from which to estimate the parameters. For 99% missing entries, the CPD algorithm no longer finds a solution as the number of known entries (820) is close to the number of free parameters (204). The CPD-based method has the best performance. ($\mathcal{Y}$ has a CPD structure from which to start.)

In the second experiment, the number of known elements is kept between 8% and 12% (remember that it is difficult to control the accesses in an adaptive algorithm), and the SNR is varied [see Figure 5(b)]. In case of the full-tensor methods, E_{RMS} is almost equal for all decompositions, except for when there is low SNR. In the case of the incomplete methods, the CPD-based method performs better, especially in the low-SNR cases.

MATERIAL SCIENCES EXAMPLE

When designing new materials, the physical properties of these new materials are key parameters. In the case of alloys, the concentrations of the different constituent materials can be used to model the physical properties. In this particular example, we model the melting point of an alloy, using a data set kindly provided by InsPyro NV, Belgium. The data set contains a small set of random measurements of the melting point in function of the concentrations of ten different constituent materials. This data set can be represented as a ninth-order tensor $\mathcal{T}$. (One concentration is superfluous as concentrations must add up to 100%.) The curse of dimensionality is an important problem for this kind of data, as the number of elements in this tensor is approximately $100^N = 10^{18}$, with $N+1$ being the number of constituent materials. Because measuring and computing all these elements is infeasible, only 130,000 elements are sampled.

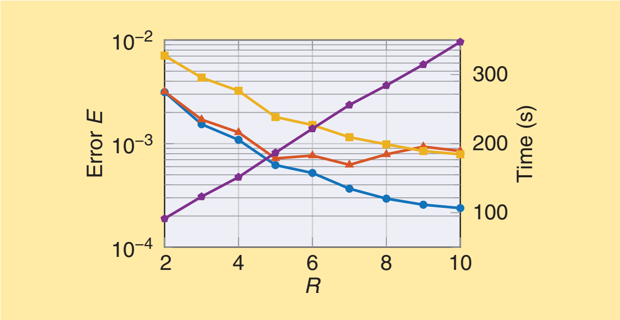


[FIG5] The influence of (a) the number of known elements and (b) the SNR on E_{RMS} for the CPD (—●—), the LMLRA (—▲—), and TT (—■—). The dashed lines give the results for the full-tensor methods.

This case study illustrates how a tensor decomposition algorithm for incomplete tensors can overcome the curse of dimensionality. We use the `cpdi_nls` algorithm [28] because it is suitable for a data set containing only a small fraction of randomly sampled elements. In particular, we approximate the training tensor $\mathcal{T}_{\mathrm{tr}}$, which contains 70% of the data, by a CPD $\hat{\mathcal{T}} = [\![\mathrm{A}^{(1)}, \ldots, \mathrm{A}^{(9)}]\!]$ and repeat this for several ranks R. To evaluate the quality of a rank-R model, the validation error E_{val} of the model is computed using an independent validation tensor $\mathcal{T}_{\mathrm{val}}$ containing the remaining 30% of the data. This error is defined as the weighted relative norm of the error between $\mathcal{T}_{\mathrm{val}}$ and the model

$$E_{\mathrm{val}} = \frac{\| \mathcal{W}_{\mathrm{val}} * (\mathcal{T}_{\mathrm{val}} - [\![\mathrm{A}^{(1)}, \ldots, \mathrm{A}^{(9)}]\!]) \|}{\| \mathcal{W}_{\mathrm{val}} * \mathcal{T}_{\mathrm{val}} \|}.$$

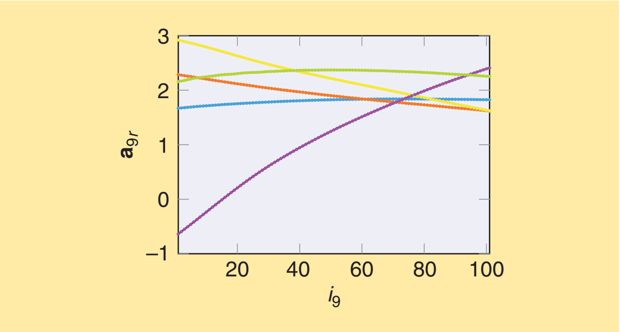
The binary observation tensor $\mathcal{W}_{\mathrm{val}}$ has only ones at the positions of known validation elements. We also report the 99%



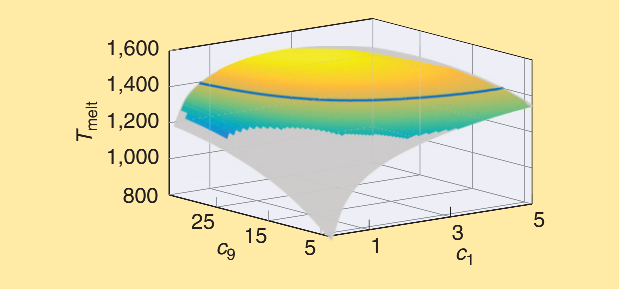
[FIG6] Errors on the training E_{tr} (—●—) and validation E_{val} (—▲—) set and the 99% quantile error E_{quant} (—■—) for different CPDs. The computation time for each model is indicated by (—◆—) on the right y-axis.

quantile of the relative residuals between known elements in the validation set and the model as E_{quant}. The timing experiments are performed on a relatively recent laptop (Intel core i7, quadcore at 2.7 GHz, 8-GB RAM, and MATLAB 2013b).

To compute a CPD from the training tensor, the `cpdi _ nls` method [28] is used. This method is an extension of the `cpd _ nls` method from Tensorlab [25], [37] for incomplete tensors. When choosing the initial factor matrices, we have to take the high-order N into account: from (1), we see that every element in $\hat{\mathcal{T}}$ is the sum of R products of $N = 9$ variables. This means that, if most elements in the factor matrices are close to zero, $\mathcal{T} \approx 0$. Here, we have drawn the elements in the initial factor matrices from a uniform distribution in $(0, 1)$, and we have scaled each factor vector $\mathbf{a}_r^{(n)}$, $n = 1, \ldots, N$ by $\sqrt[N]{\lambda_r}$, where λ_r are the minimizers of $\left\| \mathcal{W}_{\text{tr}} * (\mathcal{T}_{\text{tr}} - \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \otimes \cdots \otimes \mathbf{a}_r^{(N)}) \right\|$. Finally, we use a best-out-of-five strategy, which means that we choose five different optimally scaled initial solutions and keep the best result in terms of error on the training tensor E_{tr}. The result is shown in Figure 6. Both E_{tr} and E_{quant} keep decreasing as R increases, which indicates that outliers are also modeled when more rank-1 terms are added to the model. Starting from $R = 5$, E_{val} and E_{tr} begin to diverge, which can indicate that the data are overmodeled for $R > 5$, although E_{quant} keeps decreasing. For the remainder of this case study, we assume $R = 5$ to be a good choice as rank: the relative error E_{quant} is smaller than $1.81 \cdot 10^{-3}$ for 99% of the validation points, while it only took 3 min to compute the model. (The time rises linearly in R, as can be seen in Figure 6.)



[FIG7] The $R = 5$ factor vectors for the ninth mode $\mathbf{a}_r^{(9)}$ are shown as dots. They clearly follow a smooth function.



[FIG8] A visualization of the continuous surface of melting points when all but two concentrations are fixed. The blue line links all points having a melting temperature of 1,400 °C. The model is only valid in the colored region.

To summarize: by using 10^5 elements we have reduced a data set containing 10^{18} elements to a model having $NIR \approx 4{,}500$ parameters. We can now go one step further by looking at the values in the different factor vectors (see Figure 7). We see that the factor vectors have a smooth, low-degree polynomial-like behavior, a little perturbed by noise. By fitting smooth spline functions to each factor vector, a continuous model for the physical parameter can be created:

$$\mathcal{T} \approx f(c_1, \ldots, c_N) = \sum_{r=1}^{R} \prod_{n=1}^{N} a_r^{(n)}(c_n),$$

where $a_r^{(n)}$ are continuous functions in the concentrations c_n, $n = 1, \ldots, N$. This has many advantages: the high-dimensional model can be visualized more easily, and all elements having a certain melting point can be calculated (see, e.g., Figure 8). Furthermore, the model can be used in further steps in the design of the material.

CONCLUSIONS

Tensor decompositions open up new possibilities in analysis and computation, as they can alleviate or even break the curse of dimensionality that occurs when working with high-dimensional tensors. Decompositions such as the TT decomposition are often used in fields such as scientific computing and quantum information theory. These decompositions can easily be ported to a signal processing context. We have addressed some problems when computing decompositions of full tensors. By exploiting the structure of a tensor, CS methods can be used to compute these decompositions using incomplete tensors. We have illustrated this with random sampling techniques for the CPD, and with mode-n vector sampling techniques originating from scientific computing for the LMLRA and the TT decomposition.

ACKNOWLEDGMENTS

Nico Vervliet is supported by the Research Foundation–Flanders (FWO). The research of Otto Debals and Laurent Sorber is funded by a Ph.D. grant from the Agency for Innovation by Science and Technology. This research is supported by: 1) Research Council KU Leuven: GOA-MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC); 2) FWO: project G.0427.10N, G.0830.14N, G.0881.14N;

3) the Belgian Federal Science Policy Office: IUAP P7 (DYSCO II, Dynamical systems, control and optimization, 2012–2017); and 4) European Union: European Research Council advanced grant number 339804 (BIOTENSORS). This article reflects only the authors' views and the European Union is not liable for any use that may be made of the contained information.

AUTHORS

Nico Vervliet (Nico.Vervliet@esat.kuleuven.be) obtained his M.Sc. degree in mathematical engineering from KU Leuven, Belgium, in 2013. He is a Ph.D. candidate at the STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics of the Electrical Engineering Department, KU Leuven, and is affiliated with iMinds Medical IT. His research interests include decomposition algorithms for large and incomplete tensors and for multiview data.

Otto Debals (Otto.Debals@esat.kuleuven.be) obtained his M.Sc. degree in mathematical engineering from KU Leuven, Belgium, in 2013. He is a Ph.D. candidate at the STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics of the Electrical Engineering Department, KU Leuven, and is affiliated with iMinds Medical IT. His research concerns the tensorization of matrix data.

Laurent Sorber (Laurent.Sorber@cs.kuleuven.be) received his M.Sc. and Ph.D. degrees from the Faculty of Engineering, KU Leuven, Belgium, in 2010 and 2014, respectively. He is affiliated with the STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics of the Electrical Engineering Department; with the Numerical Approximation and Linear Algebra Group of the Computer Science Department, KU Leuven; and with iMinds Medical IT. His research includes the development of numerical algorithms for tensor decompositions and structured data fusion. He is the main developer of the Tensorlab toolbox.

Lieven De Lathauwer (Lieven.DeLathauwer@kuleuven-kulak.be) received his Ph.D. degree from the Faculty of Engineering, KU Leuven, Belgium, in 1997. From 2000 to 2007, he was a research associate with the Centre National de la Recherche Scientifique, France. He is currently a professor with KU Leuven. He is affiliated with the Science, Engineering, and Technology Group of Kulak; with the STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics of the Electrical Engineering Department; and with iMinds Medical IT. He is an associate editor of *SIAM Journal on Matrix Analysis and Applications* and has served as an associate editor of *IEEE Transactions on Signal Processing*. His research concerns the development of tensor tools for engineering applications.

REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. New York: Academic, 2010.

[2] N. Sidiropoulos, R. Bro, and G. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[3] I. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput*., vol. 33, no. 5, pp. 2295–2317, 2011.

[4] B. Khoromskij, "Tensors-structured numerical methods in scientific computing: Survey on recent advances," *Chemomet. Intell. Lab. Syst*., vol. 110, no. 1, pp. 1–19, 2012.

[5] L. Grasedyck, D. Kressner, and C. Tobler. (2013, Feb.). A literature survey of low-rank tensor approximation techniques. [Online]. Available: http://arxiv.org/pdf/1302.7121.pdf

[6] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus* (Springer Series in Computational Mathematics, vol. 42). Heidelberg: Springer, 2012.

[7] R. Orus. (2014, June). A practical introduction to tensor networks: Matrix product states and projected entangled pair states. [Online]. Available: http://arxiv.org/pdf/1306.2164.pdf

[8] I. Oseledets and E. Tyrtyshnikov, "TT-cross approximation for multidimensional arrays," *Linear Algebra Appl*., vol. 432, no. 1, pp. 70–88, 2010.

[9] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Rev*., vol. 51, no. 3, pp. 455–500, 2009.

[10] A. Cichocki, D. Mandic, C. Caiafa, A.-H. Phan, G. Zhou, Q. Zhao, and L. De Lathauwer, "Tensor decompositions for signal processing applications," *IEEE Signal Processing Mag.*, to be published.

[11] P. Comon, "Tensors: A brief introduction," *IEEE Signal Processing Mag*., vol. 31, no. 3, pp. 44–53, May 2014.

[12] I. Domanov and L. De Lathauwer, "On the uniqueness of the canonical polyadic decomposition of third-order tensors—Part II: Uniqueness of the overall decomposition," *SIAM J. Matrix Anal. Appl*., vol. 34, no. 3, pp. 876–903, 2013.

[13] G. Beylkin and M. J. Mohlenkamp, "Numerical operator calculus in higher dimensions," *Proc. Natl. Acad. Sci*., vol. 99, no. 16, pp. 10, 246–10, 251, 2002.

[14] W. Krijnen, T. Dijkstra, and A. Stegeman, " On the non-existence of optimal solutions and the occurrence of "degeneracy" in the CANDECOMP/PARAFAC model," *Psychometrika*, vol. 73, no. 3, pp. 431–439, 2008.

[15] A. Smilde, R. Bro, P. Geladi, and J. Wiley, *Multi-Way Analysis with Applications in the Chemical Sciences*. Chichester, U.K.: Wiley, 2004.

[16] L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," Tech. Rep. 13-177, ESAT-STADIUS, KU Leuven, Belgium, 2013.

[17] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl*., vol. 21, no. 4, pp. 1253–1278, 2000.

[18] I. Oseledets and E. Tyrtyshnikov, "Breaking the curse of dimensionality, or how to use SVD in many dimensions," *SIAM J. Sci. Comput*., vol. 31, no. 5, pp. 3744–3759, 2009.

[19] E. Acar, D. Dunlavy, T. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemomet. Intell. Lab. Syst*., vol. 106, no. 1, pp. 41–56, 2011.

[20] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Prob*., vol. 27, no. 2, p. 025010, 2011.

[21] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Mag*., vol. 25, no. 2, pp. 21–30, 2008.

[22] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Wiley Interdisciplinary Rev.: Data Mining Knowl. Discov*., vol. 3, no. 6, pp. 355–380, 2013.

[23] N. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Process. Lett*., vol. 19, no. 11, pp. 757–760, Nov. 2012.

[24] Q. Li, D. Schonfeld, and S. Friedland, "Generalized tensor compressive sensing," in *Proc. 2013 IEEE Int. Conf. Multimedia and Expo (ICME)*, July 2013, pp. 1–6.

[25] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(L_r, L_r, 1) terms, and a new generalization," *SIAM J. Optim*., vol. 23, no. 2, pp. 695–720, 2013.

[26] E. Acar, D. Dunlavy, and T. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions," *J. Chemomet*., vol. 25, no. 2, pp. 67–86, 2011.

[27] A.-H. Phan, P. Tichavsky, and A. Cichocki, "Low complexity damped Gauss–Newton algorithms for CANDECOMP/PARAFAC," *SIAM J. Appl. Math*., vol. 34, no. 1, pp. 126–147, 2013.

[28] O. Debals and N. Vervliet, "Efficiënte tensorgebaseerde methoden voor modellering en signaalscheiding," (in Dutch), Master's thesis, Departments of Comput. Sci. and Electr. Eng., KU Leuven, 2013.

[29] S. Goreinov, N. Zamarashkin, and E. Tyrtyshnikov, "Pseudo-skeleton approximations by matrices of maximal volume," *Math. Notes*, vol. 62, no. 4, pp. 515–519, 1997.

[30] E. Tyrtyshnikov, "Incomplete cross approximation in the mosaic-skeleton method," *Computing*, vol. 64, no. 4, pp. 367–380, 2000.

[31] I. Oseledets, D. Savostianov, and E. Tyrtyshnikov, "Tucker dimensionality reduction of three-dimensional arrays in linear time," *SIAM J. Matrix Anal. Appl*., vol. 30, no. 3, pp. 939–956, 2008.

[32] C. Caiafa and A. Cichocki, "Generalizing the column-row matrix decomposition to multi-way arrays," *Linear Algebra Appl*., vol. 433, no. 3, pp. 557–573, 2010.

[33] M. Mahoney, M. Maggioni, and P. Drineas, "Tensor-CUR decompositions for tensor-based data," *SIAM J. Matrix Anal. Appl*., vol. 30, no. 3, pp. 957–987, 2008.

[34] X. Liu, N. D. Sidiropoulos, and T. Jiang, "Multidimensional harmonic retrieval with applications in MIMO wireless channel sounding," in *Space-Time Processing for MIMO Communications*, A. Gershman and N. Sidiropoulos, Eds. Hoboken, NJ: Wiley, 2005.

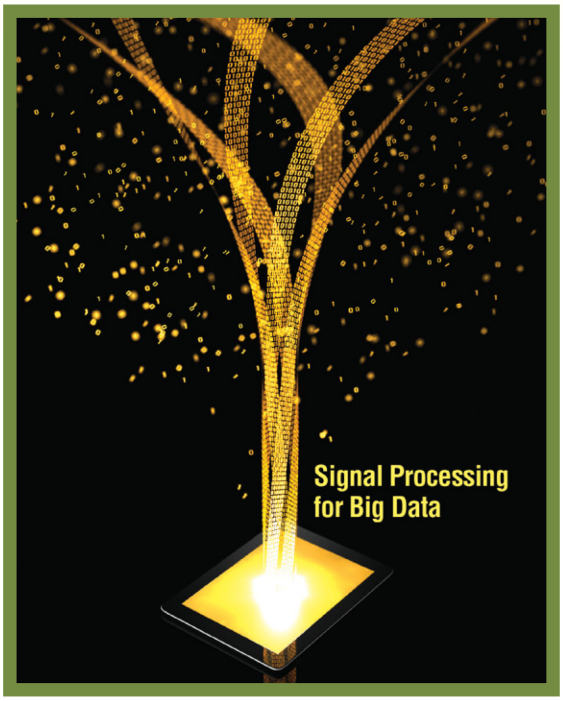
[35] M. Sorensen and L. De Lathauwer, "Multidimensional harmonic retrieval via coupled canonical polyadic decomposition," Internal Rep. 13-240, ESAT-STADIUS, KU Leuven, Belgium, 2013.

[36] S. Van Huffel, H. Chen, C. Decanniere, and P. Vanhecke, "Algorithm for time-domain NMR data fitting based on total least squares," *J. Magn. Reson., Ser. A*, vol. 110, no. 2, pp. 228–237, 1994.

[37] L. Sorber, M. Van Barel, and L. De Lathauwer. (2014, Jan.). Tensorlab v2.0. [Online]. Available: www.tensorlab.net

[Aliaksei Sandryhaila and José M.F. Moura]

Big Data Analysis with Signal Processing on Graphs



© ISTOCKPHOTO.COM/TA2YO4NORI

[Representation and processing of massive data sets with irregular structure]

Analysis and processing of very large data sets, or big data, poses a significant challenge. Massive data sets are collected and studied in numerous domains, from engineering sciences to social networks, biomolecular research, commerce, and security. Extracting valuable information from big data requires innovative approaches that efficiently process large amounts of data as well as handle and, moreover, utilize their structure. This article discusses a paradigm for large-scale data analysis based on the discrete signal processing (DSP) on graphs (DSP$_G$). DSP$_G$ extends signal processing concepts and methodologies from the classical signal processing theory to data indexed by general graphs. Big data analysis presents several challenges to DSP$_G$, in particular, in filtering and frequency analysis of very large data sets. We review fundamental concepts of DSP$_G$, including graph signals and graph filters, graph Fourier transform, graph frequency, and spectrum ordering, and compare them with their counterparts from the classical signal processing theory. We then consider product graphs as a graph model

Digital Object Identifier 10.1109/MSP.2014.2329213
Date of publication: 19 August 2014

1053-5888/14©2014IEEE

that helps extend the application of DSP$_G$ methods to large data sets through efficient implementation based on parallelization and vectorization. We relate the presented framework to existing methods for large-scale data processing and illustrate it with an application to data compression.

INTRODUCTION

Data analysts in scientific, government, industrial, and commercial domains face the challenge of coping with rapidly growing volumes of data that are collected in numerous applications. Examples include biochemical and genetics research, fundamental physical experiments and astronomical observations, social networks, consumer behavior studies, and many others. In these applications, large amounts of raw data can be used for decision making and action planning, but their volume and increasingly complex structure limit the applicability of many well-known approaches widely used with small data sets, such as principal component analysis (PCA), singular value decomposition (SVD), spectral analysis, and others. This problem—the big data problem [1]—requires new paradigms, techniques, and algorithms.

Several approaches have been proposed for representation and processing of large data sets with complex structure. Multidimensional data, described by multiple parameters, can be expressed and analyzed using multiway arrays [2]–[4]. Multiway arrays have been used in biomedical signal processing [5], [6], telecommunications and sensor array processing [7]–[9], and other domains.

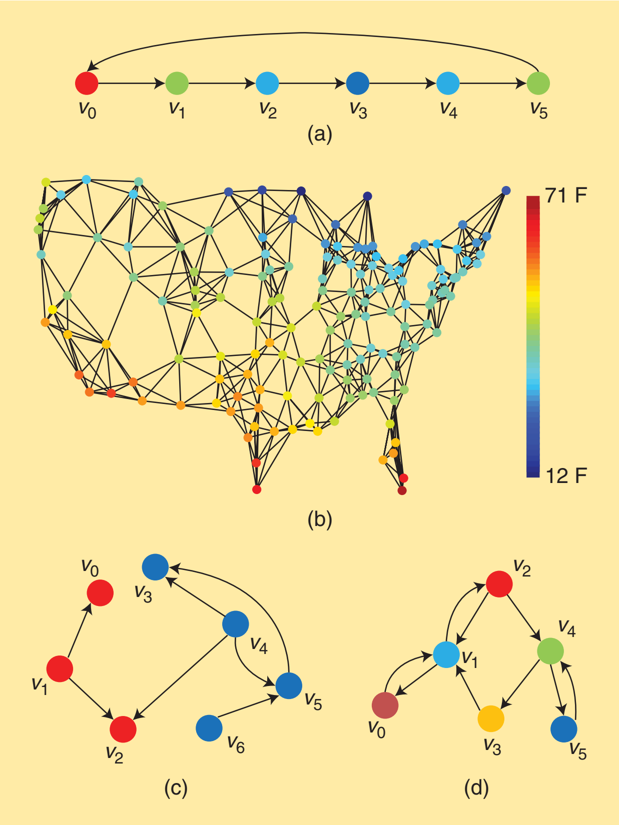
Low-dimensional representations of high-dimensional data have been extensively studied in [10]–[13]. In these approaches, data sets are viewed as graphs in high-dimensional spaces and data are projected on low-dimensional subspaces generated by small subsets of the graph Laplacian eigenbasis.

Signal processing on graphs extends classical signal processing theory to general graphs. Some techniques, such as in [14]–[16], are motivated in part by the works on graph Laplacian-based low-dimensional data representations. DSP$_G$ [17], [18] builds upon the algebraic signal processing theory [19], [20].

This article considers the use of DSP$_G$ as a methodology for big data analysis. We discuss how, for appropriate graph models, fundamental signal processing techniques, such as filtering and frequency analysis, can be implemented efficiently for large data sizes. The discussed framework addresses some of the key challenges of big data through arithmetic cost reduction of associated algorithms and use of parallel and distributed computations. The presented methodology introduces elements of high-performance computing to DSP$_G$ and offers a structured approach to the development of data analysis tools for large data volumes.

SIGNAL PROCESSING ON GRAPHS

We begin by reviewing notation and main concepts of DSP$_G$. For a detailed introduction to the theory, we refer the readers to [17] and [18]. Definitions and constructs presented here apply to general graphs. In the special case of undirected graphs with nonnegative real edge weights, similar definitions can be formulated using the graph Laplacian matrix, as discussed in [14]–[16] and references therein.



[FIG1] **Examples of graph signals. Signal values are represented with different colors. (a) The periodic time series $\cos(2\pi n/6)$ resides on a directed line graph with six nodes; the edge from the last node to the first captures the periodicity of the series. (b) Temperature measurements across the United States reside on the graph that represents the network of weather sensors. (c) Web site topics are encoded as a signal that resides on the graph formed by hyperlinks between the Web sites. (d) The average numbers of tweets for Twitter users are encoded as a signal that resides on the graph representing who follows whom.**

GRAPH SIGNALS

DSP$_G$ studies the analysis and processing of data sets in which data elements are related by dependency, similarity, physical proximity, or other properties. This relation is expressed though a graph $G = (\mathcal{V}, \mathrm{A})$, where $\mathcal{V} = \{v_0, \dots, v_{N-1}\}$ is the set of N nodes and A is the weighted adjacency matrix of the graph. Each data element corresponds to a node v_n (we also say the data element is indexed by v_n). A nonzero weight $A_{n,m} \in \mathbb{C}$ indicates the presence of a directed edge from v_m to v_n that reflects the appropriate dependency or similarity relation between the nth and mth data elements. The set of neighbors of v_n forms its neighborhood denoted as $\mathcal{N}_n = \{m \mid \mathrm{A}_{n,m} \neq 0\}$.

Given the graph, the data set forms a graph signal, defined as a map

$$\mathrm{s}: \mathcal{V} \to \mathbb{C}, \; v_n \mapsto s_n,$$

where $\mathbb{C}$ is the set of complex numbers. It is convenient to write graph signals as vectors

$$\mathbf{s} = [s_0 \;\; s_1 \;\; \ldots \;\; s_{N-1}]^T \in \mathbb{C}^N. \tag{1}$$

One should view the vector (1) not just as a list, but as a graph with each value s_n residing at node v_n.

Figure 1 shows examples of graph signals. Finite periodic time series, studied by finite-time DSP [19], [21], are indexed by directed cyclic graphs, such as the graph in Figure 1(a). Each node corresponds to a time sample; all edges are directed and have the same weight 1, reflecting the causality of time series; and the edge from the last to the first node reflects the periodicity assumption. Data collected by sensor networks is another example of graph signals: sensor measurements form a graph signal indexed by the sensor network graph, such as the graph in Figure 1(b). Each graph node is a sensor, and edges connect closely located sensors. Graph signals also arise in the World Wide Web: for instance, Web site features (topic, view count, relevance) are graph signals indexed by graphs formed by hyperlink references, such as the graph in Figure 1(c). Each node represents a Web site, and directed edges correspond to hyperlinks. Finally, graph signals are collected in social networks, where characteristics of individuals (opinions, preferences, demographics) form graph signals on social graphs, such as the graph in Figure 1(d). Nodes of the social graph represent individuals, and edges connect people based on their friendship, collaboration, or other relations. Edges can be directed (such as follower relations on Twitter) or undirected (such as friendship on Facebook or collaboration ties in publication databases).

GRAPH SHIFT

In DSP, a signal shift, implemented as a time delay, is a basic nontrivial operation performed on a signal. A delayed finite periodic time series of length N is $\tilde{s}_n = s_{n-1 \bmod N}$. Using the vector notation (1), the shifted signal is written as

$$\tilde{\mathbf{s}} = [\tilde{s}_0 \;\; \ldots \;\; \tilde{s}_{N-1}]^T = \mathbf{C}\mathbf{s}, \tag{2}$$

where $\mathbf{C}$ is the $N \times N$ cyclic shift matrix (only nonzero entries are shown)

$$\mathbf{C} = \begin{bmatrix} & & & 1 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}. \tag{3}$$

Note that (3) is precisely the adjacency matrix of the periodic time series graph in Figure 1(a).

DSP$_{\mathrm{G}}$ extends the concept of shift to general graphs by defining the graph shift as a local operation that replaces a signal value s_n at node v_n by a linear combination of the values at the neighbors of v_n weighted by their edge weights:

$$\tilde{s}_n = \sum_{m \in \mathcal{N}_n} \mathrm{A}_{n,m} s_m. \tag{4}$$

It can be interpreted as a first-order interpolation, weighted averaging, or regression on graphs, which is a widely used operation in graph regression, distributed consensus, telecommunications, Markov processes and other approaches. Using the vector notation (1), the graph shift (4) is written as

$$\tilde{\mathbf{s}} = [\tilde{s}_0 \;\; \ldots \;\; \tilde{s}_{N-1}]^T = \mathbf{A}\mathbf{s}. \tag{5}$$

The graph shift (5) naturally generalizes the time shift (2).

Since in DSP$_{\mathrm{G}}$ the graph shift is defined axiomatically, other choices for the operation of a graph shift are possible. The advantage of the definition (4) is that it leads to a signal processing framework for linear and commutative graph filters. Other choices, such as selective averaging over a subset of neighbors for each graph vertex, do not lead to linear commutative filters and hence to well-defined concepts of frequency, Fourier transform, and others.

GRAPH FILTERS AND *z*-TRANSFORM

In signal processing, a *filter* is a system $\mathrm{H}(\cdot)$ that takes a signal (1) as an input and outputs a signal

$$\tilde{\mathbf{s}} = [\tilde{s}_0 \;\; \ldots \;\; \tilde{s}_{N-1}]^T = \mathrm{H}(\mathbf{s}). \tag{6}$$

Among the most widely used filters are linear shift-invariant (LSI) ones. A filter is linear, if for a linear combination of inputs it produces the same combination of outputs: $\mathrm{H}(\alpha\mathbf{s}_1 + \beta\mathbf{s}_2) = \alpha\mathrm{H}(\mathbf{s}_1) + \beta\mathrm{H}(\mathbf{s}_2)$. Filters $\mathrm{H}_1(\cdot)$ and $\mathrm{H}_2(\cdot)$ are commutative, or shift-invariant, if the order of their application to a signal does not change the output: $\mathrm{H}_1(\mathrm{H}_2(\mathbf{s})) = \mathrm{H}_2(\mathrm{H}_1(\mathbf{s}))$.

The z-transform provides a convenient representation for signals and filters in DSP. By denoting the time delay (2) as z^{-1}, all LSI filters in finite-time DSP are written as polynomials in z^{-1}

$$h(z^{-1}) = \sum_{n=0}^{N-1} h_n z^{-n}, \tag{7}$$

where the coefficients $h_0, h_1, \ldots, h_{N-1}$ are called *filter taps*. Similarly, finite time signals are written as

$$s(z^{-1}) = \sum_{n=0}^{N-1} s_n z^{-n}. \tag{8}$$

The filter output is calculated by multiplying its z-transform (7) with the z-transform of the input signal (8) modulo the polynomial $z^{-N} - 1$, [19]:

$$\tilde{s}(z^{-1}) = \sum_{n=0}^{N-1} \tilde{s}_n z^{-n} = h(z^{-1}) s(z^{-1}) \bmod (z^{-N} - 1). \tag{9}$$

Equivalently, the output signal is given by the product [21]

$$\tilde{\mathbf{s}} = h(\mathbf{C})\mathbf{s} \tag{10}$$

of the input signal (1) and the matrix

$$h(\mathbf{C}) = \sum_{n=0}^{N-1} h_n \mathbf{C}^n = \begin{bmatrix} h_0 & h_{N-1} & \ldots & h_1 \\ h_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_{N-1} \\ h_{N-1} & \ldots & h_1 & h_0 \end{bmatrix}. \tag{11}$$

Observe that the circulant matrix $h(\mathbf{C})$ in (11) is obtained by substituting the time shift matrix (3) for z^{-1} in the filter z-transform (7). In finite-time DSP, this substitution establishes a surjective (onto) mapping from the space of LSI filters and the space of $N \times N$ circulant matrices.

DSP$_\mathrm{G}$ extends the concept of filters to general graphs. Similarly to the extension of the time shift (2) to the graph shift (5), filters (11) are generalized to graph filters as polynomials in the graph shift [17], and all LSI graph filters have the form

$$h(\mathbf{A}) = \sum_{\ell=0}^{L-1} h_\ell \mathbf{A}^\ell. \tag{12}$$

In analogy with (10), the graph filter output is given by

$$\tilde{\mathbf{s}} = h(\mathbf{A})\mathbf{s}. \tag{13}$$

The output can also be computed using the graph z-transform that represents graph filters (12) as

$$h(z^{-1}) = \sum_{\ell=0}^{L-1} h_\ell z^{-\ell}, \tag{14}$$

and graph signals (1) as polynomials $s(z^{-1}) = \sum_{n=0}^{N-1} s_n b_n(z^{-1})$, where $b_n(z^{-1})$, $0 \le n < N$, are appropriately constructed, linearly independent polynomials of degree smaller than N (see [17] for details). Analogously to (9), the output of the graph filter (14) is obtained as the product of z-transforms modulo the minimal polynomial $m_\mathbf{A}(z^{-1})$ of the shift matrix $\mathbf{A}$:

$$\tilde{s}(z^{-1}) = \sum_{n=0}^{N-1} \tilde{s}_n b_n(z^{-1}) = h(z^{-1})s(z^{-1}) \bmod m_\mathbf{A}(z^{-1}). \tag{15}$$

Recall that the minimal polynomial of $\mathbf{A}$ is the unique monic polynomial of the smallest degree that annihilates $\mathbf{A}$, i.e., $m_\mathbf{A}(\mathbf{A}) = 0$ [22].

Graph filters have a number of important properties. An inverse of a graph filter, if it exists, is also a graph filter that can be found by solving a system of at most N linear equations. Also, the number of taps in a graph filter is not larger than the degree of the minimal polynomial of $\mathbf{A}$, which provides an upper bound on the complexity of their computation. In particular, since the graph filter (12) can be factored as

$$h(\mathbf{A}) = h_{L-1} \prod_{\ell=0}^{L-1} (\mathbf{A} - g_\ell \mathbf{I}), \tag{16}$$

the computation of the output (13) requires, in general, $L \le \deg m_\mathbf{A}(x)$ multiplications by $\mathbf{A}$.

GRAPH FOURIER TRANSFORM

Mathematically, a Fourier transform with respect to a set of operators is the expansion of a signal into a basis of the operators' eigenfunctions. Since in signal processing the operators of interest are filters, DSP$_\mathrm{G}$ defines the Fourier transform with respect to the graph filters.

For simplicity of the discussion, assume that $\mathbf{A}$ is diagonalizable and its eigendecomposition is

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}, \tag{17}$$

where the columns $\mathbf{v}_n$ of the matrix $\mathbf{V} = [\mathbf{v}_0 \ \ldots \ \mathbf{v}_{N-1}] \in \mathbb{C}^{N\times N}$ are the eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda} \in \mathbb{C}^{N\times N}$ is the diagonal matrix of corresponding eigenvalues $\lambda_0, \ldots, \lambda_{N-1}$ of $\mathbf{A}$. If $\mathbf{A}$ is not diagonalizable, Jordan decomposition into generalized eigenvectors is used [17].

The eigenfunctions of graph filters $h(\mathbf{A})$ are given by the eigenvectors of the graph shift matrix $\mathbf{A}$ [17]. Since the expansion into the eigenbasis is given by the multiplication with the inverse eigenvector matrix [22], which always exists, the graph Fourier transform of a graph signal (1) is well defined and computed as

$$\begin{aligned}\hat{\mathbf{s}} = [\hat{s}_0 \ \ldots \ \hat{s}_{N-1}]^T &= \mathbf{V}^{-1}\mathbf{s} \\ &= \mathbf{F}\mathbf{s},\end{aligned} \tag{18}$$

where $\mathbf{F} = \mathbf{V}^{-1}$ is the graph Fourier transform matrix.

The values $\hat{s}_n$ in (18) are the signal's expansion in the eigenvector basis and represent the graph frequency content of the signal $\mathbf{s}$. The eigenvalues λ_n of the shift matrix $\mathbf{A}$ represent graph frequencies, and the eigenvectors $\mathbf{v}_n$ represent the corresponding graph frequency components. Observe that each frequency component $\mathbf{v}_n$ is a graph signal, too, with its mth entry indexed by the node v_m.

The inverse graph Fourier transform reconstructs the graph signal from its frequency content by combining graph frequency components weighted by the coefficients of the signal's graph Fourier transform:

$$\mathbf{s} = \hat{s}_0\mathbf{v}_0 + \hat{s}_1\mathbf{v}_1 + \cdots + \hat{s}_{N-1}\mathbf{v}_{N-1} = \mathbf{F}^{-1}\hat{\mathbf{s}} = \mathbf{V}\,\hat{\mathbf{s}}. \tag{19}$$

Analogously to other DSP$_\mathrm{G}$ concepts, the graph Fourier transform is a generalization of the discrete Fourier transform from DSP. Recall that the mth Fourier coefficient of a finite time series of length N is

$$\hat{s}_m = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} s_n e^{-j\frac{2\pi}{N}mn},$$

and the time signal's discrete Fourier transform is written in vector form as $\hat{\mathbf{s}} = \mathbf{DFT}_N\mathbf{s}$, where $\mathbf{DFT}_N$ is the $N \times N$ discrete Fourier transform matrix with the (n, m)th entry $1/\sqrt{N}\exp(-j2\pi nm/N)$. It is well known that the eigendecomposition of the time shift matrix (3) is

$$\mathbf{C} = \mathbf{DFT}_N^{-1} \begin{bmatrix} e^{-j\frac{2\pi\cdot 0}{N}} & & \\ & \ddots & \\ & & e^{-j\frac{2\pi\cdot(N-1)}{N}} \end{bmatrix} \mathbf{DFT}_N.$$

Hence, the discrete Fourier transform is the graph Fourier transform for cyclic line graphs, such as the graph in Figure 1(a), and $\lambda_n = \exp(-j2\pi n/N)$, $0 \le n < N$, are the corresponding frequencies. In DSP, the ratio $2\pi n/N$ in the exponent $\lambda_n = \exp(-j2\pi n/N)$ is also sometimes called (angular) frequency.

ALTERNATIVE CHOICES OF GRAPH FOURIER BASIS

In some cases, for example, when eigenvector computation is not stable, it may be advantageous to use other vectors as the graph Fourier basis, such as singular vectors or eigenvectors of the Laplacian matrix. These choices are consistent with DSP$_G$, since singular vectors form the graph Fourier basis when the graph shift matrix is defined as $\mathbf{AA}^*$, and Laplacian eigenvectors form the graph Fourier basis when the shift matrix is defined by the Laplacian. However, the former implicitly turns the original graph into an undirected graph, and the latter explicitly requires that the original graph is undirected. As a result, in both cases the framework does not use the information about the direction of graph edges that is useful in various applications [17], [19], [23]. Examples, where relations are directed and not always reciprocal, are Twitter (if user A follows user B, user B does not necessarily follow user A), and the World Wide Web (if document A links to document B, document B does not necessarily link to document A).

FREQUENCY RESPONSE OF GRAPH FILTERS

In addition to expressing the frequency content of graph signals, the graph Fourier transform also characterizes the effect of filters on the frequency content of signals. The filtering operation (13) can be written using (12) and (18) as

$$\widetilde{\mathbf{s}} = h(\mathbf{A})\mathbf{s} = h(\mathbf{F}^{-1}\boldsymbol{\Lambda}\mathbf{F})\mathbf{s} = \mathbf{F}^{-1}h(\boldsymbol{\Lambda})\mathbf{F}\,\mathbf{s}, \qquad (20)$$

where $h(\boldsymbol{\Lambda})$ is a diagonal matrix with values $h(\lambda_n) = \sum_{\ell=0}^{L-1} h_\ell \lambda_n^\ell$ on the diagonal. As follows from (20),

$$\tilde{\mathbf{s}} = h(\mathbf{A})\mathbf{s} \Leftrightarrow \mathbf{F}\tilde{\mathbf{s}} = h(\boldsymbol{\Lambda})\hat{\mathbf{s}}. \qquad (21)$$

That is, the frequency content of a filtered signal is modified by multiplying its frequency content elementwise by $h(\lambda_n)$. These values represent the graph frequency response of the graph filter (12).

The relation (21) is a generalization of the classical convolution theorem [21] to graphs: filtering a graph signal in the graph domain is equivalent in the frequency domain to multiplying the signal's spectrum by the frequency response of the graph filter.

LOW AND HIGH FREQUENCIES ON GRAPHS

In DSP, frequency contents of time series and digital images are described by complex or real sinusoids that oscillate at different rates [24]. These rates provide an intuitive, physical interpretation of "low" and "high" frequencies: low-frequency components oscillate less and high-frequency ones oscillate more.

In analogy to DSP, frequency components on graphs can also be characterized as "low" and "high" frequencies. In particular, this is achieved by ordering the graph frequency components according to how much they change across the graph; that is, how much the signal coefficients of a frequency component differ at connected nodes. The amount of "change" is calculated using the graph total variation [18]. For graphs with real spectra, the ordering from lowest to highest frequencies is $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{N-1}$. For graphs with complex spectra, frequencies are ordered by their distance from the point $|\lambda_{\max}|$ on the complex plane, where $\lambda_{\max}$ is the eigenvalue with the largest magnitude. The graph frequency order naturally leads to the definition of low-, high-, and band-pass graph filters, analogously to their counterparts in DSP (see [18] for details).

In the special case of undirected graphs with real nonnegative edge weights, the graph Fourier transform (18) can also be expressed using the eigenvectors of the graph Laplacian matrix [16]. In general, the eigenvectors of the adjacency and Laplacian matrices do not coincide, which can lead to a different Fourier transform matrix. However, when graphs are regular, both definitions yield the same graph Fourier transform matrix, and the same frequency ordering [18].

APPLICATIONS

DSP$_G$ is particularly motivated by the need to extend traditional signal processing methods to data sets with complex and irregular structure. Problems in different domains can be formulated and solved as standard signal processing problems. Applications include data compression through Fourier transform or through wavelet expansions; recovery, denoising, and classification of data by signal regularization, smoothing, or adaptive filter design; anomaly detection via high-pass filtering; and many others (see [15] and [16] and references therein).

For instance, a graph signal can be compressed by computing its graph Fourier transform and storing only a small fraction of its spectral coefficients, the ones with largest magnitudes. The compressed signal is reconstructed by computing the inverse graph Fourier transform with the preserved coefficients. When the signal is sparse in the Fourier domain, that is, when most energy is concentrated in a few frequencies, the compressed signal is reconstructed with a small error [17], [25].

Another example application is the detection of corrupted data. In traditional DSP, a corrupted value in a slowly changing time signal introduces additional high-frequency components that can be detected by high-pass filtering of the corrupted signal. Similarly, a corrupted value in a graph signal can be detected through a high-pass graph filter, which can be used, for instance, to detect malfunctioning sensors in sensor networks [18].

CHALLENGES OF BIG DATA

While there is no single, universally agreed upon set of properties that define big data, some of the commonly mentioned ones are volume, velocity, and variety of data [1]. Each of these characteristics poses a separate challenge to the design and implementation of analysis systems and algorithms for big data. First of all, the sheer volume of data to be processed requires efficient distributed and scalable storage, access, and processing. Next, in many applications, new data is obtained continuously. High velocity of new data arrival demands fast algorithms to prevent bottlenecks and explosion of the data volume and to extract valuable information from the data and incorporate it into the decision-making process in real time. Finally, collected data sets contain information in all varieties and forms, including numerical, textual, and visual data. To generalize data analysis techniques to diverse data sets, we need a common representation framework for data sets and their structure.

The latter challenge of data diversity is addressed in DSP$_G$ by representing data set structure with graphs and quantifying data into graph signals. Graphs provide a versatile data abstraction for multiple types of data, including sensor network measurements, text documents, image and video databases, social networks, and others. Using this abstraction, data analysis methods and tools can be developed and applied to data sets of a different nature.

For efficient big data analysis, the challenges of data volume and velocity must be addressed as well. In particular, the fundamental signal processing operations of filtering and spectral decomposition may be prohibitively expensive for large data sets both in the amount of required computations and memory demands.

Recall that processing a graph signal (1) with a graph filter (16) requires L multiplications by a $N \times N$ graph shift matrix $\mathbf{A}$. For a general matrix, this computation requires $O(LN^2)$ arithmetic operations (additions and multiplications) [26]. When $\mathbf{A}$ is sparse and has on average K nonzero entries in every row, graph filtering requires $O(LNK)$ operations. In addition, graph filtering also requires access to the entire graph signal in memory. Similarly, computation of the graph Fourier transform (18) requires $O(N^2)$ operations and access to the entire signal in memory. Moreover, the eigendecomposition of the matrix $\mathbf{A}$ requires additional $O(N^3)$ operations and memory access to the entire $N \times N$ matrix $\mathbf{A}$. Note that graph filtering can also be performed in the spectral domain with $O(N^2)$ operations using the graph convolution theorem (21), but it also requires the initial eigendecomposition of $\mathbf{A}$.

Degree heterogeneity in graphs with heavily skewed degree distributions, such as scale-free graphs, presents an additional challenge. Graph filtering (16) requires iterative weighted averaging over each vertex's neighbors, and for vertices with large degrees this process takes significantly longer than for vertices with small degrees. In this case, load balancing through smart distribution of vertices between computational nodes is required to avoid a computation bottleneck.

For very large data sets, algorithms with quadratic and cubic arithmetic cost are not acceptable. Moreover, computations that require access to the entire data sets are ill suited for large data sizes and lead to performance bottlenecks, since memory access is orders of magnitude slower than arithmetic computations. This problem is exacerbated by the fact that large data sets often do not fit into main memory or even local disk storage of a single machine, and must be stored and accessed remotely and processed with distributed systems.

Fifty years ago, the invention of the famous fast Fourier transform algorithm by Cooley and Tukey [27], as well as many other algorithms that followed (see [28] and [29] and references therein), dramatically reduced the computational cost of the discrete Fourier transform by using suitable properties of the structure of time signals, and made frequency analysis and filtering of very large signals practical. Similarly, in this article, we identify and discuss properties of certain data representation graphs that lead to more efficient implementations of DSP$_G$ operations for big data. A suitable graph model is provided by product graphs discussed in the next section.

PRODUCT GRAPHS

Consider two graphs $G_1 = (\mathcal{V}_1, \mathbf{A}_1)$ and $G_2 = (\mathcal{V}_2, \mathbf{A}_2)$ with $|\mathcal{V}_1| = N_1$ and $|\mathcal{V}_2| = N_2$ nodes, respectively. The product graph, denoted by $\diamond$, of G_1 and G_2 is the graph

$$G = G_1 \diamond G_2 = (\mathcal{V}, \mathbf{A}_\diamond), \tag{22}$$

with $|\mathcal{V}| = N_1 N_2$ nodes and an appropriately defined $N_1 N_2 \times N_1 N_2$ adjacency matrix $\mathbf{A}_\diamond$ [30], [31]. In particular, three commonly studied graph products are the Kronecker, Cartesian, and strong products.

For the Kronecker graph product, denoted as $G = G_1 \otimes G_2$, the adjacency matrix is obtained by the matrix Kronecker product of adjacency matrices $\mathbf{A}_1$ and $\mathbf{A}_2$:

$$\mathbf{A}_\otimes = \mathbf{A}_1 \otimes \mathbf{A}_2. \tag{23}$$

Recall that the Kronecker product of matrices $\mathbf{B} = [b_{mn}] \in \mathbb{C}^{M \times N}$ and $\mathbf{C} \in \mathbb{C}^{K \times L}$ is a $KM \times LN$ matrix with block structure

$$\mathbf{B} \otimes \mathbf{C} = \begin{bmatrix} b_{0,0}C & \dots & b_{0,N-1}\mathbf{C} \\ \vdots & \vdots & \vdots \\ b_{M-1,0}C & \dots & b_{M-1,N-1}\mathbf{C} \end{bmatrix}. \tag{24}$$

For the Cartesian graph product, denoted as $G = G_1 \times G_2$, the adjacency matrix is

$$\mathbf{A}_\times = \mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2. \tag{25}$$

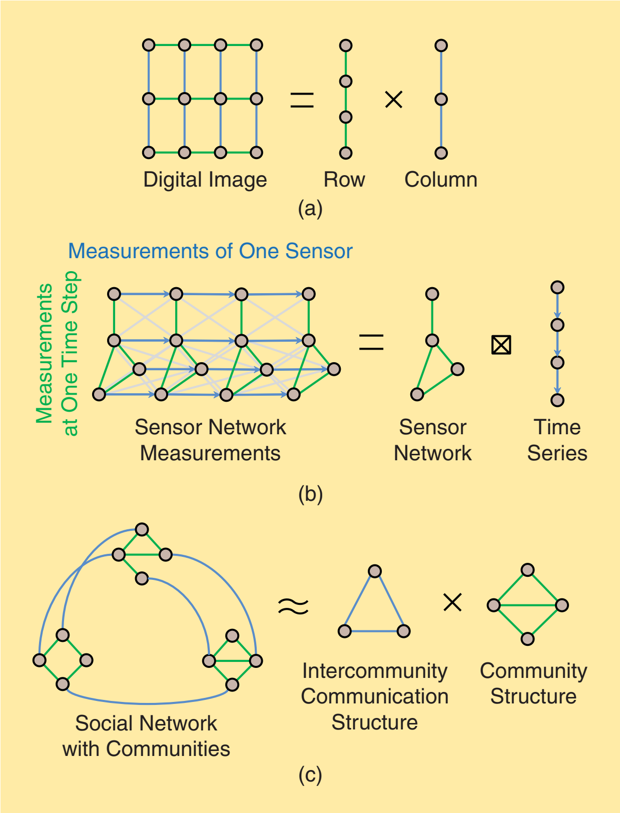
Finally, for the strong product, denoted as $G = G_1 \times G_2$, the adjacency matrix is

$$\mathbf{A}_\boxtimes = \mathbf{A}_1 \otimes \mathbf{A}_2 + \mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2. \tag{26}$$

The strong product can be seen as a combination of the Kronecker and Cartesian products. Since the products (24)–(26) are associative, Kronecker, Cartesian, and strong graph products can be defined for an arbitrary number of graphs.

Product graphs arise in different applications, including signal and image processing [32], computational sciences and data mining [33], and computational biology [34]. Their probabilistic counterparts are used in network modeling and generation [35]–[37]. Multiple approaches have been proposed for the decomposition and approximation of graphs with product graphs [30], [31], [38], [39].

Product graphs offer a versatile graph model for the representation of complex data sets in multilevel and multiparameter ways. In traditional DSP, multidimensional signals, such as digital images and video, reside on rectangular lattices that are Cartesian products of line graphs. Figure 2(a) shows a


Digital Image
Row
Column
(a)
Measurements of One Sensor
Measurements at One Time Step
Sensor Network Measurements
Sensor Network
Time Series
(b)
Social Network with Communities
Intercommunity Communication Structure
Community Structure
(c)

[FIG2] Examples of product graphs indexing various data: (a) digital images reside on rectangular lattices that are Cartesian products of line graphs for rows and columns, (b) measurements of a sensor network are indexed by the strong product of the sensor network graph with the time series graph (the edges of the Cartesian product are shown in blue and green, and edges of the Kronecker product are shown in gray; the strong product contains all edges), and (c) a social network with three similar communities is approximated by a Cartesian product of the community structure graph with the intercommunity communication graph.

two-dimensional (2-D) lattice formed by the Cartesian product of two one-dimensional lattices.

Another example of graph signals residing on product graphs is data collected by a sensor network over a period of time. In this case, the graph signal formed by measurements of all sensors at all time steps resides on the product of the sensor network graph with the time series graph. As the example in Figure 2(b) illustrates, the kth measurement of the nth sensor is indexed by the nth node of the kth copy of the sensor graph (or, equivalently, the kth node of the nth copy of the time series graph). Depending on the choice of product, a measurement of a sensor is related to the measurements collected by this sensor and its neighbors at the same time and previous and following time steps. For instance, the strong product in Figure 2(b) relates the measurement of the nth sensor at time step k to its measurements at time steps $k-1$ and $k+1$, as well as to measurements of its neighbors at times $k-1,\ k,$ and $k+1$.

A social network with multiple communities also may be representable by a graph product. Figure 2(c) shows an example of a social network that has three communities with similar structures, where individuals from different communities also interact with each other. This social graph may be seen as an approximation of the Cartesian product of the graph that captures the community structure and the graph that captures the interaction between communities.

Other examples where product graphs are potentially useful for data representation include multiway data arrays that contain elements described by multiple features, parameters, or characteristics, such as publications in citation databases described by their topics, authors, and venues; or Internet connections described by their time, location, IP address, port accesses, and other parameters. In this case, the graph factors in (22) represent similarities or dependencies between subsets of characteristics.

Graph products are also used for modeling entire graph families. Kronecker products of scale-free graphs with the same degree distribution are also scale free and have the same distribution [35], [40]. K- and ϵ-nearest neighbor graphs, which are used in signal processing, communications, and machine learning to represent spatial and temporal location of data, such as sensor networks and image pixels, or data similarity structure, can be approximated with graph products, as the examples in Figure 2(a) and (b) suggest. Other graph families, such as trees, are constructed using rooted graph products [41], which are not discussed in this article.

SIGNAL PROCESSING ON PRODUCT GRAPHS

In this section, we discuss how product graphs help "modularize" the computation of filtering and Fourier transform on graphs and improve algorithms, data storage, and memory access for large data sets. They lead to graph filtering and Fourier transform implementations suitable for multicore and clustered platforms with distributed storage by taking advantage of such performance optimization techniques as parallelization and vectorization. The presented results illustrate how product graphs offer a suitable and practical model for constructing and implementing signal processing methodologies for large data sets. In this, product graphs are similar to other graph families, such as scale-free and small-world graphs, that are used to model properties of real-world graphs and data sets: while models do not fit exactly to all real-world graphs, they capture and abstract relevant representations of graphs and facilitate their analysis and processing.

FILTERING

Recall that graph filtering is computed as the multiplication of a graph signal (1) by a filter (16). As we discussed in the section "Challenges of Big Data," computation of a filtered signal requires repeated multiplications by the shift matrix, which is in general a computation- and memory-expensive operation for very large data sets.

Now, consider, for instance, a Cartesian product graph with the shift matrix (25). A graph filter of the form (16) for this graph is written as

$$h(\mathbf{A}_\times) = h_L \prod_{\ell=0}^{L-1} (\mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2 - g_\ell \mathbf{I}_{N_1 N_2}). \qquad (27)$$

Hence, multiplication by the shift matrix $\mathbf{A}_\times$ is replaced with multiplications by matrices $\mathbf{A}_1 \otimes \mathbf{I}_{N_2}$ and $\mathbf{I}_{N_1} \otimes \mathbf{A}_2$.

Multiplication by matrices of the form $\mathbf{I}_{N_1} \otimes \mathbf{A}_2$ and $\mathbf{A}_1 \otimes \mathbf{I}_{N_2}$ have multiple efficient implementations that take advantage of modern optimization and high-performance techniques, such as parallelization and vectorization [26], [42], [43]. In particular, the product $(\mathbf{I}_{N_1} \otimes \mathbf{A}_2)\mathbf{s}$ is calculated by multiplying N_1 signal segments $\mathbf{s}_{n,\ldots,n+N_2}$, $0 \le n < N_1$, of length N_2 by the matrix $\mathbf{A}_2$. These products are computed with independent parts of the input signal, which eliminates data dependency and makes these operations highly suitable for a parallel implementation on a multicore or cluster platform [42]. As an illustration, for $N_1 = 3$, $N_2 = 2$, matrix

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} \qquad (28)$$

and a signal $\mathbf{s} \in \mathbb{C}^6$, we obtain

$$(\mathbf{I}_3 \otimes \mathbf{A})\,\mathbf{s} = \begin{bmatrix} \mathbf{A} & & \\ & \mathbf{A} & \\ & & \mathbf{A} \end{bmatrix} \mathbf{s} = \begin{bmatrix} \mathbf{A}\begin{bmatrix} s_0 \\ s_1 \end{bmatrix} \\ \mathbf{A}\begin{bmatrix} s_2 \\ s_3 \end{bmatrix} \\ \mathbf{A}\begin{bmatrix} s_4 \\ s_5 \end{bmatrix} \end{bmatrix}.$$

Here, all multiplications by $\mathbf{A}$ are independent from each other both in data access and computations.

Similarly, the product $(\mathbf{A}_1 \otimes \mathbf{I}_{N_2})\mathbf{s}$ is calculated by multiplying N_2 segments $\mathbf{s}_{n,n+N_1,\ldots,n+(N_2-1)N_1}$, $0 \le n < N_2$, of the input signal by the matrix $\mathbf{A}_1$. These products are highly suitable for a vectorized implementation, available on modern computational platforms, that performs an operation on several input values simultaneously [42]. For instance, for $\mathbf{A}$ in (28), we obtain

$$(\mathbf{A} \otimes \mathbf{I}_3)\,\mathbf{s} = \begin{bmatrix} a_{00}\begin{bmatrix} s_0 \\ s_1 \\ s_2 \end{bmatrix} + a_{01}\begin{bmatrix} s_3 \\ s_4 \\ s_5 \end{bmatrix} \\ a_{10}\begin{bmatrix} s_0 \\ s_1 \\ s_2 \end{bmatrix} + a_{11}\begin{bmatrix} s_3 \\ s_4 \\ s_5 \end{bmatrix} \end{bmatrix}.$$

Here, three sequential signal values are multiplied by one element of matrix $\mathbf{A}$ at the same time. These operations are performed simultaneously by a processor with vectorization capabilities, which respectively decreases the computation time by a factor of three.

In addition to its suitability for parallelized and vectorized implementations, computing the output of the filter (27) on a Cartesian graph also requires significantly fewer operations, since the multiplication by the shift matrix (25) requires N_1 multiplications by an $N_2 \times N_2$ matrix and N_2 multiplications by an $N_1 \times N_1$ matrix, which results in $O(N_1 N_2^2) + O(N_1^2 N_2) = O(N(N_1 + N_2))$ operations rather than $O(N^2)$. (We discuss here operation counts for general graphs with full matrices. In practice, adjacency matrices are often sparse, and their multiplication requires fewer operations. Computational savings provided by product graphs are, likewise, significant for sparse adjacency matrices.) For example, when $N_1, N_2 \approx \sqrt{N}$, this represents a reduction of the computational cost of graph filtering by a factor $\sqrt{N}$. To put this into the big data perspective, for a graph with a million vertices, the cost of filtering is reduced by a factor of 1,000, and for a graph with a billion vertices, the cost reduction factor is more than 30,000.

Furthermore, the multiplication by a matrix of the form $\mathbf{I} \otimes \mathbf{A}$ can be replaced by the multiplication with a matrix $\mathbf{A} \otimes \mathbf{I}$ with no additional arithmetic operations by suitable permutation of signal values [22], [42], [43]. This interchangeability leads to a selection between parallelized and vectorized implementations and provides means to efficiently compute graph filtered signals on platforms with arbitrary number of cores and vectorization capabilities.

The advantages of filtering on Cartesian product graphs also apply to Kronecker and strong product graphs. In particular, using the property [22]

$$\mathbf{A}_1 \otimes \mathbf{A}_2 = (\mathbf{A}_1 \otimes \mathbf{I}_{N_2})(\mathbf{I}_{N_1} \otimes \mathbf{A}_2), \qquad (29)$$

we write the graph filter (16) for the Kronecker product as

$$h(\mathbf{A}_\otimes) = h_L \prod_{\ell=0}^{L-1} ((\mathbf{A}_1 \otimes \mathbf{I}_{N_2})(\mathbf{I}_{N_1} \otimes \mathbf{A}_2) - g_\ell \mathbf{I}_{N_1 N_2}),$$

and for the strong product as

$$h(\mathbf{A}_\boxtimes) = h_L \prod_{\ell=0}^{L-1} (\mathbf{A}_1 \otimes \mathbf{I}_{N_2})(\mathbf{I}_{N_1} \otimes \mathbf{A}_2) + \mathbf{A}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{A}_2 - g_\ell \mathbf{I}_{N_1 N_2}).$$

Similarly to (27), these filters multiply input signals by matrices $\mathbf{I}_{N_1} \otimes \mathbf{A}_2$ and $\mathbf{A}_1 \otimes \mathbf{I}_{N_2}$ and are implementable using parallelization and vectorization techniques. They also lead to substantial reductions of the number of required computations.

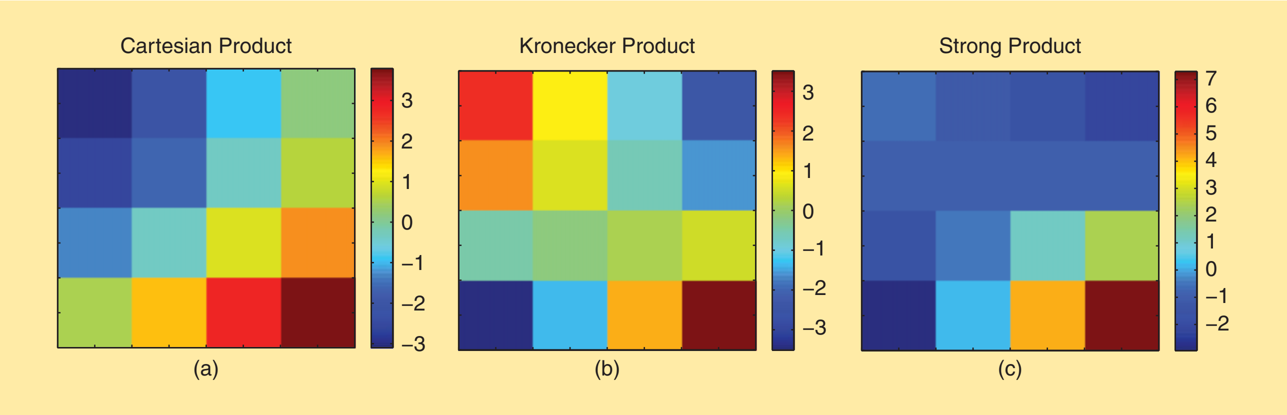
FOURIER TRANSFORM

The frequency content of a graph signal is computed through the graph Fourier transform (18). In general, this procedure has the computational cost of $O(N^2)$ operations and requires access to the entire signal in memory. Moreover, it also requires a preliminary calculation of the eigendecomposition of the graph shift matrix $\mathbf{A}$, which, in general, takes $O(N^3)$ operation.

Let us consider a Cartesian product graph with the shift matrix (25). Assume that the eigendecomposition (17) of the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ is respectively $\mathbf{A}_i = \mathbf{V}_i \mathbf{\Lambda}_i \mathbf{V}_i^{-1}$, $i \in \{1,2\}$, where $\mathbf{\Lambda}_i$ has eigenvalues $\lambda_{i,0}, \ldots, \lambda_{i,N-1}$ on the main diagonal. Similar results can be obtained for nondiagonalizable matrices using Jordan decomposition. The derivation is more involved, and we omit it for simplicity of discussion.

If we denote $\mathbf{V} = \mathbf{V}_1 \otimes \mathbf{V}_2$, then the eigendecomposition of the shift matrix (25) is [22]

$$\mathbf{A}_\times = \mathbf{V}(\mathbf{\Lambda}_1 \otimes \mathbf{I}_{N_2} + \mathbf{I}_{N_1} \otimes \mathbf{\Lambda}_2)\mathbf{V}^{-1}. \qquad (30)$$



[FIG3] The frequency values for the product graphs in Figure 2(b). Frequencies are shown as a color-coded 2-D map, with *x*- and *y*-axis representing frequencies of two factor graphs. Higher values correspond to lower frequencies and vice versa. (a) The Cartesian product, (b) Kronecker product, and (c) strong product.

Hence, the graph Fourier transform associated with a Cartesian product graph is given by the matrix Kronecker product of the graph Fourier transforms for its factor graphs:

$$\mathrm{F}_{\times} = (\mathrm{V}_1 \otimes \mathrm{V}_2)^{-1} = \mathrm{V}_1^{-1} \otimes \mathrm{V}_2^{-1} = \mathrm{F}_1 \otimes \mathrm{F}_2, \qquad (31)$$

and the spectrum is given by the element-wise summation of the spectra of the smaller graphs: $\lambda_{1,n} + \lambda_{2,m}$, $0 \le n < N_1$ and $0 \le m < N_2$.

Reusing the property (29), (31) can be written as $\mathrm{F}_{\times} = \mathrm{F}_1 \otimes \mathrm{F}_2 = (\mathrm{F}_1 \otimes \mathrm{I}_{N_2})(\mathrm{I}_{N_1} \otimes \mathrm{F}_2)$ and efficiently implemented using parallelization and vectorization techniques. Moreover, the computation of the eigendecomposition (30) is replaced with finding the eigendecomposition of the shift matrices A_1 and A_2, which reduces the computation cost from $O(N^3)$ to $O(N_1^3 + N_2^3)$. For instance, when $N_1, N_2 \approx \sqrt{N}$, the computational cost of the eigendecomposition is reduced by a factor $N\sqrt{N}$. Hence, for a graph with a million vertices, the cost of computing the eigendecomposition is reduced by a factor of more than 3×10^4, and for a graph with a billion vertices, the cost reduction factor is over 3×10^{13}.

The same improvements apply to the Kronecker and strong matrix products, since the eigendecomposition of the corresponding shift matrices is

$$\mathrm{A}_{\otimes} = \mathrm{V}(\boldsymbol{\Lambda}_1 \otimes \boldsymbol{\Lambda}_2)\mathrm{V}^{-1},$$
$$\mathrm{A}_{\boxtimes} = \mathrm{V}(\boldsymbol{\Lambda}_1 \otimes \mathrm{I}_{N_2} + \mathrm{I}_{N_1} \otimes \boldsymbol{\Lambda}_2 + \boldsymbol{\Lambda}_1 \otimes \boldsymbol{\Lambda}_2)\mathrm{V}^{-1}.$$

Observe that all three graph products have the same graph Fourier transform. However, the corresponding spectra are different: for Cartesian and strong products, they are, respectively, $\lambda_{1,n}\lambda_{2,m}$ and $\lambda_{1,n}\lambda_{2,m} + \lambda_{1,n} + \lambda_{2,m}$, where $0 \le n < N_1$ and $0 \le m < N_2$. Thus, while all three graph products have the same frequency components, the ordering of these components from lowest to highest frequencies, as defined by DSP$_{\mathrm{G}}$ and discussed in the section "Signal Processing on Graphs," can be different. As an illustration, consider the example in Figure 3. It shows the frequencies (eigenvalues) of the three graph products in Figure 2(b). All product graphs have the same 16 frequency components (eigenvectors), but the frequencies (eigenvalues) corresponding to these components are different and on each graph have a different interpretation as low or high frequency. For example, the values in the upper left corners of Figure 3(a)–(c) correspond to the same frequency component. By comparing these values, we observe that this component represents the highest frequency in the Cartesian product graph, the lowest frequency in the Kronecker product graph, and a midspectrum component in the strong product graph.

FAST GRAPH FOURIER TRANSFORMS

A major motivation behind the use of product graphs in signal processing and DSP$_{\mathrm{G}}$ is derivation of fast computational algorithms for the graph Fourier transform. A proper overview of this topic requires an additional discussion of graph concepts and an algebraic approach to fast algorithms [29], [44], [45] that are beyond the scope of this article.

As an intuitive example, consider a well-known and widely used decimation-in-time fast Fourier transform for power-of-two sizes [27]. It is derived using graph products as follows. We view the $\mathbf{DFT}_N$ as the graph Fourier transform of a graph with adjacency matrix $\mathbf{C}^2$, where $\mathbf{C}$ is the cyclic shift matrix (3). This is a valid algebraic assumption, since the $\mathbf{DFT}N$ is a graph Fourier transform not only for the graph in Figure 1(a), but for any graph with adjacency matrix given by a polynomial $h(\mathbf{C})$. This graph, after a permutation of its vertices at stride two (which represents the decimation-in-time step), becomes a product of a cyclic graph with $N/2$ vertices with a graph of two disconnected vertices. As a result, its graph Fourier transform $\mathbf{DFT}_N$ becomes a product $\mathrm{I}_2 \otimes \mathbf{DFT}_{N/2}$ and additional, sparse matrices that capture the operations of graph restructuring. By continuing this process recursively for $\mathbf{DFT}_{N/2}$, $\mathbf{DFT}_{N/4}$, and so forth, we decompose $\mathbf{DFT}_N$ into a product of sparse matrices with cumulative arithmetic cost of $O(N \log N)$, thus obtaining a fast algorithm for the computation of $\mathbf{DFT}_N$.

[TABLE 1] ERRORS INTRODUCED BY COMPRESSION OF THE TEMPERATURE DATA.

	FRACTION OF COEFFICIENTS USED (*C*/*N*)						
	1/50	1/20	1/15	1/10	1/7	1/5	1/3
ERROR (%)	4.9	3.5	3.1	2.6	2.1	1.6	0.7
PSNR (dB)	71.2	74.1	75.1	76.7	78.5	80.9	87.1

RELATION TO EXISTING APPROACHES

The instantiation of DSP$_G$ for product graphs relates to existing approaches to complex data analysis that are not based on graphs but rather view data as multidimensional arrays [2]–[4]. Given a K-dimensional data set $\mathbf{S} \in \mathbb{C}^{N_1 \times N_2 \times \ldots \times N_K}$, the family of methods called *canonical decomposition* or *parallel factor analysis* searches for K matrices $\mathbf{M}_k \in \mathbb{C}^{N_k \times R}$, $1 \le k \le K$, that provide an optimal approximation of the data set

$$\mathbf{S} = \sum_{r=1}^{R} \mathbf{m}_{1,r} \circ \mathbf{m}_{2,r} \circ \ldots \circ \mathbf{m}_{K,r} + \mathbf{E}, \qquad (32)$$

that minimizes the error

$$\| \mathbf{E} \| = \sqrt{\sum_{n_1=1}^{N_1} \ldots \sum_{n_k=1}^{N_K} | \mathbf{E}_{n_1,n_2,\ldots,n_K} |^2}.$$

Here, $\mathbf{m}_{k,r}$ denotes the rth column of matrix $\mathbf{M}_k$, and $\circ$ denotes the outer product of vectors.

A more general approach, called *Tucker decomposition*, searches for K matrices $\mathbf{M}_k \in \mathbb{C}^{N_k \times R_k}$, $1 \le k \le K$, and a matrix $\mathbf{C} \in \mathbb{C}^{R_1 \times R_2 \times \ldots \times R_K}$ that provide an optimal approximation of the data set as

$$\mathbf{S} = \sum_{r_1=1}^{R_1} \ldots \sum_{r_K=1}^{R_K} \mathbf{C}_{r_1,\ldots,r_K} \mathbf{m}_{1,r_1} \circ \ldots \circ \mathbf{m}_{K,r_K} + \mathbf{E}. \qquad (33)$$

Tucker decomposition is also called a higher-order PCA or SVD, since it effectively extends these techniques from matrices to higher-order arrays.

Decompositions (32) and (33) can be interpreted as signal compression on product graphs. For simplicity of discussion, assume that $K = 2$ and consider a signal $\mathbf{s} \in \mathbb{C}^{N_1 N_2}$ that lies on a product graph (22) and corresponds to a 2-D signal $\mathbf{S} \in \mathbb{C}^{N_1 \times N_2}$, so that $\mathbf{S}_{n_1,n_2} = \mathbf{s}_{n_1 N_2 + n_2}$, where $0 \le n_i < N_i$ for $i = 1,2$. If matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ contain as columns, respectively, R_1 and R_2 eigenvectors of $\mathbf{A}_1$ and $\mathbf{A}_2$, then the decomposition (33) represents a lossy compression of the graph signal in the frequency domain, a widely used compression technique in signal processing [21], [24].

EXAMPLE APPLICATION

As a motivational application example of DSP$_G$ on product graphs, we consider data compression. For the testing data set, we use the set of daily temperature measurements collected by 150 weather stations across the United States [17] during the year 2002. Figure 1(b) shows the measurements from one day (1 December 2002), as well as the sensor network graph. The graph is constructed by connecting each sensor to eight of its nearest neighbors with undirected edges with weights given by [17, eq. (29)]. As illustrated by the example in Figure 2(b), such a data set can be described by a product of the sensor network graph and the time series graphs. We use the sensor network graph in Figure 1(b) with $N_1 = 150$ nodes and the time series graph in Figure 1(a) with $N_2 = 365$ nodes.

The compression is performed in the frequency domain. We compute the Fourier transform (31) of the data set, keep only C spectrum coefficients with largest magnitudes and replace others with zeros, and perform the inverse graph Fourier transform on the resulting coefficients. This is a lossy compression scheme, with the compression error given by the norm of the difference between the original data set and the reconstructed one normalized by the norm of the original data set. Note that, while the approach is tested here on a relatively small data set, it is applicable in the same form to arbitrarily large data sets.

The compression errors for the considered temperature data set are shown in Table 1. The results demonstrate that even for high compression ratios, that is, when the number C of stored coefficients is much smaller than the data set size $N = N_1 N_2$, the compression introduces only a small error and leads to insignificant loss of information. A comparison of this approach with schemes that compress the data only in one dimension (they separately compress either time series from each sensor or daily measurements from all sensors) [17], [25] also reveals that compression based on the product graph is significantly more efficient.

CONCLUSIONS

In this article, we presented an approach to big data analysis based on the DSP on graphs. We reviewed fundamental concepts of the framework and illustrated how it extends traditional signal processing theory to data sets represented with general graphs. To address important challenges in big data analysis and make implementations of fundamental DSP$_G$ techniques suitable for very large data sets, we considered a generalized graph model given by several kinds of product graphs, including the Cartesian, Kronecker, and strong product graphs. We showed that these product graph structures significantly reduce arithmetic cost of associated DSP$_G$ algorithms and make them suitable for parallel and distributed implementation, as well as improve memory storage and access of data. The discussed methodology bridges a gap between signal processing, big data analysis, and high-performance computing, as well as presents a framework for the development of new methods and tools for analysis of massive data sets.

AUTHORS

Aliaksei Sandryhaila (asandryh@andrew.cmu.edu) received a B.S. degree in computer science from Drexel University, Philadelphia,

Pennsylvania, in 2005, and a Ph.D. degree in electrical and computer engineering from Carnegie Mellon University (CMU), Pittsburgh, Pennsylvania, in 2010. He is currently a research scientist in the Department of Electrical and Computer Engineering at CMU. His research interests include big data analysis, signal processing, machine learning, design and optimization of algorithms and software, and high-performance computing. He is a Member of the IEEE.

José M.F. Moura (moura@ece.cmu.edu) is the Philip and Marsha Dowd University Professor at Carnegie Mellon University (CMU). In 2013–2014, he is a visiting professor at New York University with the Center for Urban Science and Progress. He holds degrees from IST (Portugal) and the Massachusetts Institute of Technology, where he has been a visiting professor. At CMU, he manages the CMU/Portugal Program. His interests are in signal processing and data science. He was an IEEE Board director, president of the IEEE Signal Processing Society (SPS), and editor-in-chief of *IEEE Transactions on Signal Processing*. He received the IEEE SPS Technical Achievement Award and the IEEE SPS Society Award. He is a Fellow of the IEEE and the AAAS, a corresponding member of the Academy of Sciences of Portugal, and a member of the U.S. National Academy of Engineering.

REFERENCES

[1] P. Zikopoulos, D. deRoos, and K. P. Corrigan, *Harness the Power of Big Data*. New York: McGraw-Hill, 2012.

[2] M. W. Mahoney, M. Maggoni, and P. Drineas, "Tensor-CUR decompositions for tensor-based data," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 957–987, 2008.

[3] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[4] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Trans. Knowledge Data Eng.*, vol. 21, no. 1, pp. 6–20, 2009.

[5] A. H. Andersen and W. S. Rayens, "Structure-seeking multilinear methods for the analysis of fMRI data," *Neuroimage*, vol. 22, no. 2, pp. 728–739, 2004.

[6] F. Miwakeichi, E. Martinez-Montes, P. A. Valdes-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, "Decomposing EEG data into space-time-frequency components using parallel factor analysis," *Neuroimage*, vol. 22, no. 3, pp. 1035–1045, 2004.

[7] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Processing*, vol. 48, no. 3, pp. 810–823, 2000.

[8] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[9] L. De Lathauwer and J. Castaing, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Processing*, vol. 56, no. 3, pp. 1096–1105, 2008.

[10] J. F. Tenenbaum, V. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[11] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[12] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comp.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[13] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[14] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *J. Appl. Comp. Harm. Anal.*, vol. 30, no. 2, pp. 129–150, 2011.

[15] S. K. Narang and A. Ortega, "Perfect reconstruction two-channel wavelet filter banks for graph structured data," *IEEE Trans. Signal Processing*, vol. 60, no. 6, pp. 2786–2799, 2012.

[16] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs," *IEEE Signal Processing Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[17] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Processing*, vol. 61, no. 7, pp. 1644–1656, 2013.

[18] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Processing*, vol. 62, no. 12, pp. 3042–3054, 2014.

[19] M. Püschel and J. M. F. Moura, "Algebraic signal processing theory: Foundation and 1-D time," *IEEE Trans. Signal Processing*, vol. 56, no. 8, pp. 3572–3585, 2008.

[20] M. Püschel and J. M. F. Moura, "Algebraic signal processing theory: 1-D space," *IEEE Trans. Signal Processing*, vol. 56, no. 8, pp. 3586–3599, 2008.

[21] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing,* 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[22] P. Lancaster and M. Tismenetsky, *The Theory of Matrices,* 2nd ed. New York: Academic, 1985.

[23] A. Sandryhaila and J. M. F. Moura, "Classification via regularization on graphs," in *Proc. IEEE Global Conf. Signal Information Processing*, 2013, pp. 495–498.

[24] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. New York: Academic, 2008.

[25] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Graph Fourier transform," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 6167–6170.

[26] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[27] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Computat.*, vol. 19, no. 9, pp. 297–301, 1965.

[28] P. Duhamel and M. Vetterli, "Fast Fourier transforms: A tutorial review and a state of the art," *J. Signal Process.*, vol. 19, no. 4, pp. 259–299, 1990.

[29] M. Püschel and J. M. F. Moura, "Algebraic signal processing theory: Cooley–Tukey type algorithms for DCTs and DSTs," *IEEE Trans. Signal Processing*, vol. 56, no. 4, pp. 1502–1521, 2008.

[30] W. Imrich, S. Klavzar, and D. F. Rall, *Topics in Graph Theory: Graphs and Their Cartesian Product*. Boca Raton, FL: CRC Press, 2008.

[31] R. Hammack, W. Imrich, and S. Klavzar, *Handbook of Product Graphs*, 2nd ed. Boca Raton, FL: CRC Press, 2011.

[32] D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[33] E. Acar, R. J. Harrison, F. Olken, O. Alter, M. Helal, L. Omberg, B. Bader, A. Kennedy, H. Park, Z. Bai, D. Kim, R. Plemmons, G. Beylkin, T. Kolda, S. Ragnarsson, L. Delathauwer, J. Langou, S. P. Ponnapalli, I. Dhillon, L. Lim, J. R. Ramanujam, C. Ding, M. Mahoney, J. Raynolds, L. Elden, C. Martin, P. Regalia, P. Drineas, M. Mohlenkamp, C. Faloutsos, J. Morton, B. Savas, S. Friedland, L. Mullin, and C. Van Loan, "Future directions in tensor-based computation and modeling," NSF Workshop Rep., Arlington, VA, Feb. 2009.

[34] M. Hellmuth, D. Merkle, and M. Middendorf, "Extended shapes for the combinatorial design of RNA sequences," *Int. J. Comp. Biol. Drug Des.*, vol. 2, no. 4, pp. 371–384, 2009.

[35] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, pp. 985–1042, Feb. 2010.

[36] S. Moreno, S. Kirshner, J. Neville, and S. Vishwanathan, "Tied Kronecker product graph models to capture variance in network populations," in *Proc. Allerton Conf. Communication, Control, and Computing,* 2010, pp. 1137–1144.

[37] S. Moreno, J. Neville, and S. Kirshner, "Learning mixed Kronecker product graph models with simulated method of moments," in *Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2013, pp. 1052–1060.

[38] C. Van Loan and N. Pitsianis, "Approximation with Kronecker products," in *Proc. Linear Algebra for Large Scale and Real Time Applications*, 1993, pp. 293–314.

[39] M. Hellmuth, W. Imrich, and T. Kupka, "Partial star products: A local covering approach for the recognition of approximate Cartesian product graphs," *Math. Comp. Sci.*, vol. 7, no. 3, pp. 255–273, 2013.

[40] J. Leskovec and C. Faloutsos, "Scalable modeling of real graphs using Kronecker multiplication," in *Proc. Int. Conf. Machine Learning*, 2007, pp. 497–504.

[41] C. D. Godsil and B. D. McKay, "A new graph product and its spectrum," *Bull. Aust. Math. Soc.*, vol. 18, no. 1, pp. 21–28, 1978.

[42] F. Franchetti, M. Püschel, Y. Voronenko, S. Chellappa, and J. M. F. Moura, "Discrete Fourier transform on multicore," *IEEE Signal Processing Mag.*, vol. 26, no. 6, pp. 90–102, 2009.

[43] Y. Voronenko, F. de Mesmay, and M. Püschel, "Computer generation of general size linear transform libraries," in *Proc. IEEE Int. Symp. Code Generation and Optimization,* 2009, pp. 102–113.

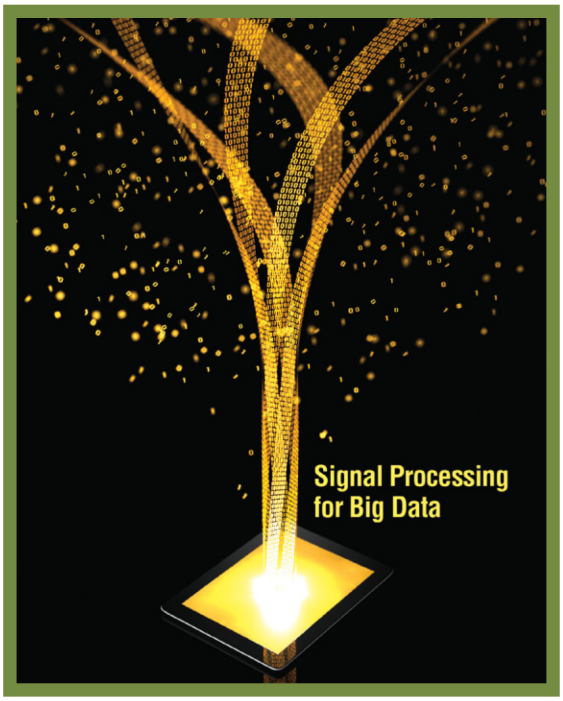
[44] M. Püschel and J. M. F. Moura, "The algebraic approach to the discrete cosine and sine transforms and their fast algorithms," *SIAM J. Comp.*, vol. 32, no. 5, pp. 1280–1316, 2003.

[45] A. Sandryhaila, J. Kovacevic, and M. Püschel, "Algebraic signal processing theory: Cooley–Tukey type algorithms for polynomial transforms based on induction," *SIAM J. Matrix Anal. Appl.*, vol. 32, no. 2, pp. 364–384, 2011.

[SP]

[Anna C. Gilbert, Piotr Indyk, Mark Iwen, and Ludwig Schmidt]

Recent Developments in the Sparse Fourier Transform



© ISTOCKPHOTO.COM/TA2YO4NORI

[A compressed Fourier transform for big data]

The discrete Fourier transform (DFT) is a fundamental component of numerous computational techniques in signal processing and scientific computing. The most popular means of computing the DFT is the fast Fourier transform (FFT). However, with the emergence of big data problems, in which the size of the processed data sets can easily exceed terabytes, the "fast" in FFT is often no longer fast enough. In addition, in many big data applications it is hard to acquire a sufficient amount of data to compute the desired Fourier transform in the first place. The sparse Fourier transform (SFT) addresses the big data setting by computing a compressed Fourier transform using only a subset of the input data, in time smaller than the data set size. The goal of this article is to survey these recent developments, explain the basic techniques with examples and applications in big data, demonstrate tradeoffs in empirical performance of the algorithms, and discuss the connection between the SFT and other techniques for massive data analysis such as streaming algorithms and compressive sensing.

Digital Object Identifier 10.1109/MSP.2014.2329131
Date of publication: 19 August 2014

1053-5888/14©2014IEEE

INTRODUCTION

The DFT is one of the main mathematical workhorses of signal processing. The most popular approach for computing the DFT is the FFT algorithm. Invented in the 1960s, the FFT computes the frequency representation of a signal of size N in $O(N \log N)$ time. The FFT is widely used and was considered to be one of the most influential and important algorithmic developments of the 20th century. However, with the emergence of big data problems, in which the size of the processed data sets can easily exceed terabytes, the FFT is not always fast enough. Furthermore, in many applications it is hard to acquire a sufficient amount of data to compute the desired Fourier transform. For example, in medical imaging, it is highly desirable to reduce the time that the patient spends in a magnetic resonance imaging machine. This motivates the need for algorithms that can compute the Fourier transform in sublinear time (in an amount of time that is considerably smaller than the size of the data), and that use only a subset of the input data. The SFT provides precisely this functionality.

Developed over the last decade, SFT algorithms compute an approximation or compressed version of the DFT in time proportional to the sparsity of the spectrum of the signal (i.e., the number of dominant frequencies), as opposed to the length of the signal. The algorithms use only a small subset of the input data and run in time proportional to the sparsity or desired compression, considerably faster than in time proportional to the signal length. This is made possible by requiring that the algorithms report only the nonzero or large frequencies and their complex amplitudes, rather than a vector containing this information for all frequencies. Since most video, audio, medical images, spectroscopic measurements (e.g., nuclear magnetic resonance), global positioning system (GPS) signals, seismic data, and many more massive data sets are compressible or are sparse, these results promise a significant practical impact in many big data domains.

The first algorithms of this type were designed for the Hadamard transform; i.e., the Fourier transform over the Boolean cube [19], [20] (cf. [7]). Shortly thereafter, algorithms for the complex Fourier transform were discovered as well [1], [5], [8], [22]. The most efficient of those algorithms [8] computed the DFT in time $k \log^{O(1)} N$, where k is the sparsity of the signal spectrum. All of the algorithms are randomized and have a constant probability of error. These developments were covered in [9].

Over the last few years, the topic has been the subject of extensive research, from the algorithmic [3], [4], [6], [10]–[13], [15], [17], [18], [21], [23], [26], implementation [25], [27], and hardware [2], [29] perspectives. These developments include the first deterministic algorithms that make no errors [3], [17], [18], as well as algorithms that, given a signal with k-sparse spectrum, compute the nonzero coefficients in time $O(k \log N)$ [11] or even $O(k \log k)$ [6], [21], [23], [26]. The goal of this article is to survey these developments. We focus on explaining the basic components and techniques used in the aforementioned algorithms, coupled with illustrative examples and concrete applications. We do not cover the analysis of sampling rates and running times of the algorithms (the reader is referred to the original papers for proofs and analysis). However, we present an empirical analysis of the performance of the algorithms. Finally, we discuss the connection between the SFT and other techniques for massive data analysis such as streaming algorithms and compressive sensing.

DEFINITIONS

Let $F \in \mathbb{C}^{N \times N}$ be the DFT matrix of size N, defined entrywise by

$$F_{\omega,j} := \frac{e^{-2\pi i \omega j/N}}{N} \tag{1}$$

for $0 \le \omega, j < N$. The DFT of a vector $\mathbf{f} \in \mathbb{C}^N$ is simply

$$\hat{\mathbf{f}} := F\mathbf{f}. \tag{2}$$

Equivalently, one may define $\hat{\mathbf{f}} \in \mathbb{C}^N$ componentwise by

$$\hat{f}_\omega = \frac{1}{N} \sum_{j=0}^{N-1} f_j \, e^{-2\pi i \omega j/N}. \tag{3}$$

In this section it will also be useful to consider the inverse of the DFT matrix above. Its entries are given by

$$F^{-1}_{j,\omega} := e^{2\pi i \omega j/N} \tag{4}$$

for $0 \le \omega, j < N$. The inverse DFT of a vector $\hat{\mathbf{f}} \in \mathbb{C}^N$ is just

$$\mathbf{f} := F^{-1}\hat{\mathbf{f}} = F^{-1}(F\mathbf{f}). \tag{5}$$

This allows one to write $\mathbf{f} \in \mathbb{C}^N$ in terms of its Fourier components by the formula

$$f_j = \sum_{\omega=0}^{N-1} \hat{f}_\omega \, e^{2\pi i \omega j/N}. \tag{6}$$

TECHNIQUES

In this section, we outline the basic components and techniques used in sparse FFT algorithms. We start from the simple case when the spectrum of the signal consists of, or is dominated by, only a single nonzero frequency. In this case we show that the position and the value of the nonzero frequency can be found using only two samples of the signal (if the signal is a pure tone without any noise) or a logarithmic number of samples (if the signal is contaminated by noise). These techniques are described in the section "Phase Encoding." Second, we address the general case by reducing it to several subproblems involving a single nonzero frequency. This is achieved by grouping subsets of Fourier space together into a small number of bins. In the simplest case, each bin corresponds to a frequency band, but other groupings are also possible. If each bin contains only a single frequency (i.e., if a frequency is isolated), then we can solve the problem separately for each bin using the techniques mentioned earlier. This leads to the sample complexity and the running time proportional to the number of bins, which is lower bounded by the sparsity parameter k. The binning techniques are described in the section "Filtering to Isolate Frequencies." Finally, in the section "Randomly Binning Frequencies," we show how to deal with spectra in which two nonzero frequencies are very close to each other, and thus cannot be easily isolated via binning. Specifically, we show how to permute the spectrum of the signal in a pseudorandom fashion, by pseudorandomly permuting the time domain signal. Since the positions of the nonzero frequencies in the permuted signals are (pseudo)random, they are likely to be isolated, and then recovered by the binning procedure. To ensure that all nonzero coefficients are recovered, the permutation and binning procedure is repeated several times, using fresh randomness every time.

All sparse FFT algorithms follow this general approach, as discussed in more detail in the section "The Prototypical SFT." However, they differ in the implementations of the specific modules, as well as in the methods they use for aggregating the information gathered from different invocations of the permutation and binning procedure.

SINGLE-FREQUENCY RECOVERY

In the simplest possible case, a vector $\mathrm{f} \in \mathbb{C}^N$ contains a single pure frequency. In such a setting an SFT must be able to rapidly determine the single tone much more quickly than the $O(N \log N)$-time FFT. In this section we will illustrate several different techniques for accomplishing this fundamental task. In later sections, we will demonstrate techniques for filtering more general vectors to produce the type of single-frequency vectors considered here. For now, however, we will assume that our vector is of the following form:

$$f_j := \hat{f}_\omega \cdot e^{2\pi i\omega j/N}. \tag{7}$$

for a fixed $\omega \in \{0, 1, \ldots, N-1\}$.

PHASE ENCODING

Given a simple vector defined as in (7), one can quickly calculate ω by choosing $j \in \{0, \ldots, N-1\}$, and then computing the phase of

$$\frac{f_{j+1}}{f_j} = e^{2\pi i\omega/N} = \cos\left(\frac{2\pi\omega}{N}\right) + i\cdot\sin\left(\frac{2\pi\omega}{N}\right). \tag{8}$$

If $j+1 = N$, we set $f_{j+1} = f_0$. More generally, we will assume that indices are always taken modulo N when referring to entries of $\mathrm{f} \in \mathbb{C}^N$ below. Furthermore, once ω is known, $\hat{f}_\omega$ can be calculated by computing $f_j \cdot e^{-2\pi i\omega j/N}$. This procedure effectively finds the DFT of the vector from (7) in $O(1)$-time by inspecting only two entries of f. The procedure, referred to as the *OFDM trick* in [11], was also used in [21]. It can also be seen as a very special case of the Prony method [13].

Although fast, this straightforward technique is not generally very robust to noise. Specifically, if $f_j := \hat{f}_\omega \cdot e^{2\pi i\omega j/N} + \epsilon_j$ for all j, the phases calculated from (8) can fail to yield ω unless $|\epsilon_j|$ is much smaller than $|\hat{f}_\omega|/N$. Hence, it is often necessary to use different techniques to find ω from (7).

A BINARY SEARCH TECHNIQUE

One means of learning ω from (7) in a more noise tolerant fashion is to perform the equivalent of a binary search for ω through the frequency domain. Many variants of such a search can be performed. In this subsection we will illustrate the most basic type of search for the example vector

$$f_j := \hat{f}_\omega \cdot e^{2\pi i\cdot\omega j/8}. \tag{9}$$

Note that this is exactly (7) with $N = 8$. Here, ω is unknown. We begin knowing only that

$$\omega \in \{0, 1, 2, 3, 4, 5, 6, 7\}.$$

Our job is to find ω using three rounds of testing based on at most six entries of f.

Our tests will be based on the following observations: If $0 < \omega < N/2 = 4$, then $e^{2\pi i\omega/8}$ will be closer to $i = e^{2\pi i2/8}$ than to $-i = e^{2\pi i6/8}$ [see Figure 1(a)]. Conversely, if $\omega > N/2 = 4$, then $e^{2\pi i\omega/8}$ will be closer to $-i$ than to i. Similarly, $\omega < N/4 = 2$ or $\omega > 3N/4 = 6$ implies that $e^{2\pi i\omega/8}$ is closer to 1 than to -1, and $2 = N/4 < \omega < 3N/4 = 6$ implies that $e^{2\pi i\omega/8}$ will be closer to -1 than to 1.

During our first round of testing we will choose $j \in \{0, \ldots, 7\}$ and then test whether both

$$\begin{aligned} |f_j|\cdot|i - e^{2\pi i\cdot\omega/8}| &= |i\cdot f_j - f_{j+1}| < |i\cdot f_j + f_{j+1}| \\ &= |f_j|\cdot|i + e^{2\pi i\cdot\omega/8}|, \end{aligned} \tag{10}$$

and

$$\begin{aligned} |f_j|\cdot|1 - e^{2\pi i\cdot\omega/8}| &= |f_j - f_{j+1}| < |f_j + f_{j+1}| \\ &= |f_j|\cdot|1 + e^{2\pi i\cdot\omega/8}| \end{aligned} \tag{11}$$

are true. Note that (10) and (11) will be true if and only if

$$|e^{2\pi i\omega/8} - i| < |e^{2\pi i\omega/8} - (-i)|, \tag{12}$$

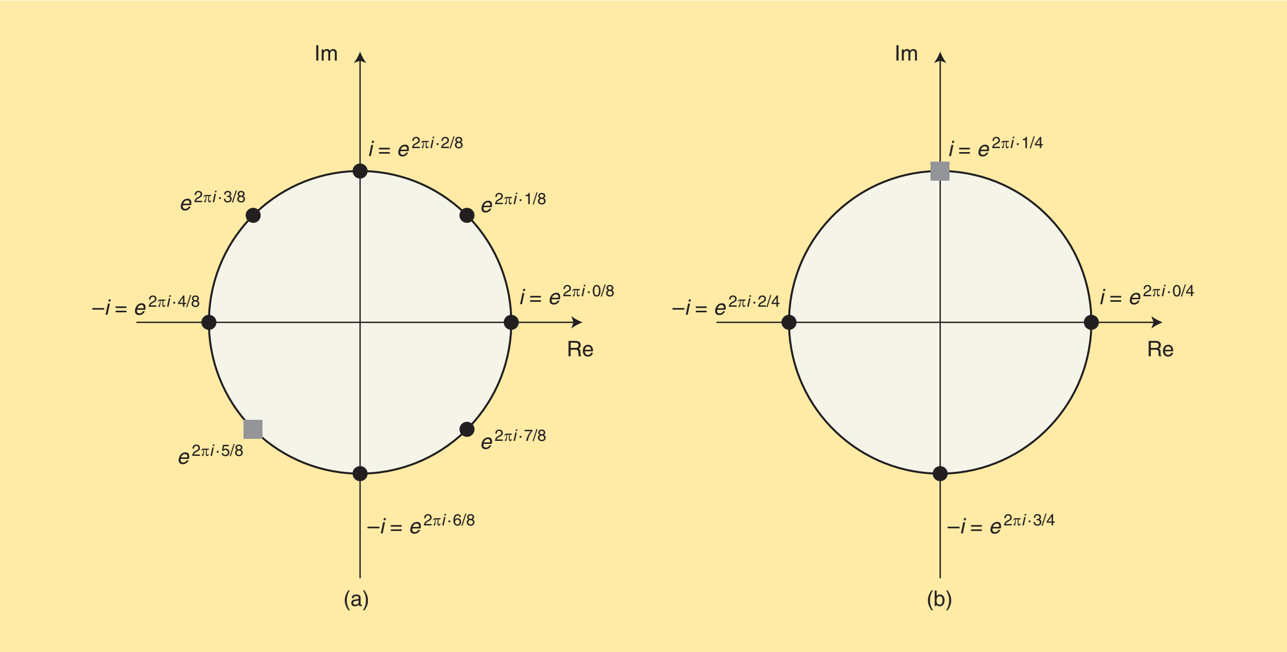
and

$$|e^{2\pi i\omega/8} - 1| < |e^{2\pi i\omega/8} - (-1)| \tag{13}$$

are true, respectively. Hence, (10) and (11) are simply testing which axes of the complex plane are best aligned with $e^{2\pi i\omega/8}$. If (10) holds true we may safely conclude that $\omega \notin \{5, 6, 7\}$ [Figure 1(a)]. Otherwise, if (10) is false, we conclude that $\omega \notin \{1, 2, 3\}$. Similarly, (11) holding true implies that $\omega \notin \{3, 4, 5\}$, while (11) failing to hold implies that $\omega \notin \{0, 1, 7\}$.

Returning to our example (9), suppose that both (10) and (11) fail to hold. The first test (10) failing tells us that $\omega \notin \{1, 2, 3\}$, and the second failure tells us that $\omega \notin \{0, 1, 7\}$. Taken all together, then, the tests tell us that $\omega \in \{4, 5, 6\}$ in this case (i.e., we learn that $e^{2\pi i\cdot\omega/8}$ is in the third quadrant of the complex plane).

Having learned that $\omega \in \{4, 5, 6\}$ allows us to simplify the problem. In particular, we may now implicitly define a new vector


Im
$i = e^{2\pi i \cdot 2/8}$
$e^{2\pi i \cdot 3/8}$
$e^{2\pi i \cdot 1/8}$
$-i = e^{2\pi i \cdot 4/8}$
$i = e^{2\pi i \cdot 0/8}$
Re
$e^{2\pi i \cdot 5/8}$
$e^{2\pi i \cdot 7/8}$
$-i = e^{2\pi i \cdot 6/8}$
(a)
Im
$i = e^{2\pi i \cdot 1/4}$
$-i = e^{2\pi i \cdot 2/4}$
$i = e^{2\pi i \cdot 0/4}$
Re
$-i = e^{2\pi i \cdot 3/4}$
(b)

[FIG1] Recovering a single frequency via a binary search (see the section "A Binary Search Technique"). Both parts show the unit circle in the complex plane. The black dots indicate the potential locations of the dominant, hidden frequency, which is represented by the gray square. (a) The initial search problem. Our goal is to locate the hidden frequency $\omega = 5$, i.e., the gray square at $e^{2\pi i \omega/8}$. In the first stage of the binary search, we determine that the hidden frequency lies in the third quadrant. (b) The simplified search problem associated with f′, which is the second stage of the binary search. The unknown frequency $\omega = 5$ has been mapped to $\omega' = 1$ (i.e., the gray square) within a smaller search space.

$$f'_j := e^{-2\pi i \cdot 4 \cdot (2j/8)} \cdot f_{2j} = \hat{f}_\omega \cdot e^{2\pi i \cdot (\omega - 4)j/4} \tag{14}$$

for all $0 \leq j < 4$. Note that $\mathrm{f}' \in \mathbb{C}^4$ was formed by 1) shifting the possible range for ω into the first quadrant (by multiplying f by $e^{-2\pi i \cdot 4j/8}$), and then 2) discarding the odd entries. This effectively halves our initial problem: $\mathrm{f}' \in \mathbb{C}^4$ is a vector with one frequency, $\omega' = (\omega - 4) \in \{0, 1, 2\}$ [see Figure 1(b)]. Our new goal is to find ω' using two entries of f′ (i.e., two additional entries of f).

Our second round of tests now proceeds exactly as before. We choose $j \in \{0, \ldots, 3\}$ and then consider both

$$|i \cdot f'_j - f'_{j+1}| < |i \cdot f'_j + f'_{j+1}|, \tag{15}$$

and

$$|f'_j - f'_{j+1}| < |f'_j + f'_{j+1}|. \tag{16}$$

As previously shown, these tests will collectively determine the quadrant of the complex plane containing $e^{2\pi i \cdot (\omega - 4)/4}$.

Continuing our example, suppose that (15) is true and (16) is false. This means that $(\omega - 4) \in \{1, 2\}$ (i.e., we have ruled out 0 and 3). We can now implicitly form our last new vector for the third round of tests. In particular, we form $\mathrm{f}'' \in \mathbb{C}^2$ by

$$f''_j := e^{-2\pi i \cdot 1 \cdot (2j/4)} \cdot f'_{2j} = \hat{f}_\omega \cdot (-1)^{(\omega - 5)j} \tag{17}$$

for $j = 0, 1$. Note that we have once again formed our new vector by 1) shifting the possible values of $\omega' = (\omega - 4)$ to the first quadrant of the complex plain, and then 2) discarding all odd entries of f′. We now know that $\omega'' = (\omega - 5) \in \{0, 1\}$, and may decide which it is by testing f″.

In particular, suppose that

$$|f''_j - f''_{j+1}| < |f''_j + f''_{j+1}| \tag{18}$$

holds for a $j \in \{0, 1\}$. Then, we conclude that

$$\omega - 5 = 0 \Rightarrow \omega = 5.$$

This concludes the description of the binary search procedure for identifying the nonzero frequency. Since we learn the position of the frequency bit by bit, the total number of samples used is $O(\log N)$. Furthermore, we note that the binary search is (relatively) robust to noise. Adding small perturbations to each entry of (9) will not stop us from determining that $\omega = 5$. Further details are given, e.g., in [9].

AN ALIASING-BASED SEARCH

We will conclude our discussion of single-frequency recovery techniques with an example of a modified search method that takes advantage of natural aliasing phenomena (see, e.g., [17] and [18]). These ideas are of use when subsampling signals is easy to implement directly, or when N is a product of several smaller relatively prime integers. Suppose, e.g., that

$$N = 70 = 2 \cdot 5 \cdot 7,$$

and let $\mathrm{a} \in \mathbb{C}^2$ be the two-element subvector of f from (7) with

$$a_0 := f_0 = \hat{f}_\omega, \text{ and } a_1 := f_{N/2} = \hat{f}_\omega \cdot (-1)^\omega. \tag{19}$$

Calculating $\hat{\mathrm{a}} \in \mathbb{C}^2$ using (3) we get that

$$\hat{a}_0 = \hat{f}_\omega \cdot \frac{1 + (-1)^\omega}{2}, \tag{20}$$

and

$$\hat{a}_1 = \hat{f}_\omega \cdot \frac{1 + (-1)^{\omega+1}}{2}. \tag{21}$$

Note that since ω is an integer, exactly one element of $\hat{\mathrm{a}}$ will be nonzero. If $\hat{a}_0 \neq 0$ then we know that $\omega \equiv 0$ modulo 2. On the other hand, $\hat{a}_1 \neq 0$ implies that $\omega \equiv 1$ modulo 2.

In this same fashion, we may use several potentially aliased FFTs in parallel to discover ω modulo 5 and 7, since they both also divide N. Once we have collected these moduli we can reconstruct ω via the Chinese remainder theorem (CRT) (see "Theorem 1: Chinese Remainder Theorem").

THEOREM 1: CHINESE REMAINDER THEOREM

Any integer x is uniquely specified modulo N by its remainders modulo m relatively prime integers $p_1, \ldots, p_m$ as long as $\prod_{l=1}^{m} p_l \geq N$.

To finish our example, suppose that we have used four FFTs on subvectors of f of size 2, 5, and 7 to determine that $\omega \equiv 1$ mod 2, $\omega \equiv 4$ mod 5, and $\omega \equiv 3$ mod 7, respectively. Using that $\omega \equiv 1$ mod 2 we can see that $\omega = 2 \cdot a + 1$ for some integer a. Using this new expression for ω in our second modulus we get

$$(2 \cdot a + 1) \equiv 4 \bmod 5 \Rightarrow a \equiv 4 \bmod 5.$$

Therefore, $a = 5 \cdot b + 4$ for some integer b. Substituting for a we get that $\omega = 10 \cdot b + 9$. By similar work we can see that $b \equiv 5$ mod 7 after considering ω modulo 7. Hence, $\omega = 59$ by the CRT. As an added bonus we note that our three FFTs will have also provided us with three different estimates of $\hat{f}_\omega$.

The end result is that we have used significantly fewer than 70 entries to determine both ω and $\hat{f}_\omega$. Using the CRT we read only $2 + 5 + 7 = 14$ entries of f. In contrast, a standard FFT would have had to process all 70 entries to compute $\hat{\mathrm{f}}$. This CRT-based single frequency method also reduces the required computational effort. Of course, a single frequency signal is incredibly simple. Vectors with more than one nonzero Fourier coefficient are much more difficult to handle since frequency moduli may begin to collide modulo various numbers. In the next section we will discuss methods for removing this difficulty.

FILTERING TO ISOLATE FREQUENCIES

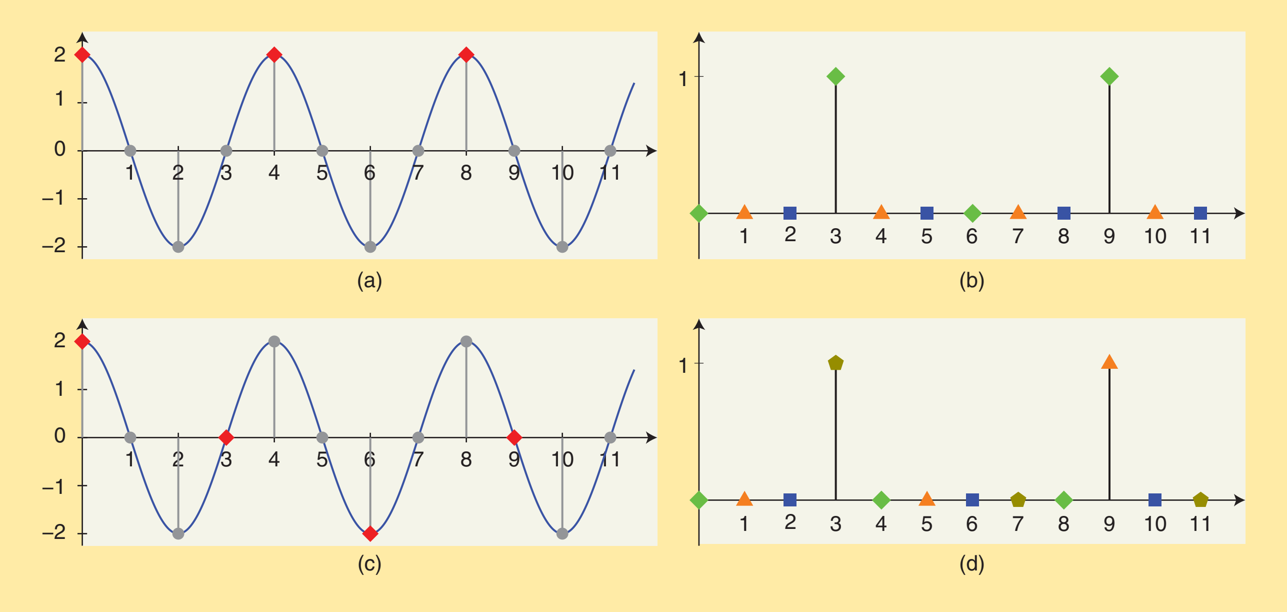
We will begin our discussion of filtering by extending our aliasing-based frequency identification ideas from the last section. In this example we assume that our vector f has length 12, with entries given by $f_j := \cos(2\pi \cdot 3 \cdot j/12)$ for $j = 0, \ldots, 11$. Note that this means $\hat{\mathrm{f}} \in \mathbb{C}^{12}$ has two nonzero entries: $\hat{f}_3 = 1$, and $\hat{f}_9 = 1$. Our objective is to learn the location and Fourier coefficient of each of them by reading fewer than 12 entries of f.

Proceeding according to the last section, we might try to learn the two frequencies by looking at the three-element subvector of f, $\mathrm{a} \in \mathbb{R}^3$, given by $a_j := f_{4j}$ for $j = 0, 1, 2$ [see Figure 2(a)]. Unfortunately, we will fail to learn anything about the individual entries of $\hat{\mathrm{f}}$ this way because both of its nonzero DFT entries are congruent to 0 modulo 3 [see Figure 2(b)]. Recall that $\hat{\mathrm{a}}_\omega$ is the sum of all Fourier coefficients whose indices are congruent to ω modulo 3. In particular, we will only see that $\hat{a}_0 = \hat{f}_0 + \hat{f}_3 + \hat{f}_6 + \hat{f}_9 = 2$, and that $\hat{a}_1 = \hat{a}_2 = 0$. If all we know is that $\hat{\mathrm{f}}$ contains at most two nonzero entries, we are unable to determine its nonzero entries using this information. The problem is that the two nonzero entries of $\hat{\mathrm{f}}$ have collided modulo 3 (i.e., they are both congruent to the same residue modulo 3).

Note, however, that the CRT guarantees that the two nonzero entries of $\hat{\mathrm{f}}$ can not also collide modulo $4 = 12/3$. If a new subarray of f is created using four equally spaced entries [see Figure 2(c)], its DFT will separate the two nonzero entries of $\hat{\mathrm{f}}$ [see Figure 2(d)]. The end result is that two subvectors will always reveal the locations of the nonzero entries of $\hat{\mathrm{f}}$, as long as $\hat{\mathrm{f}}$ has at most two nonzero entries. More generally, one can use similar ideas to learn the k largest magnitude entries of $\hat{\mathrm{f}}$ from the results of a small number of aliased DFTs of subvectors of f.

In the continuous setting, one can view the preceding discussion as a demonstration of how a relatively small set of spike-train filters can be used to separate important frequencies from one another in Fourier space. This turns out to be a fruitful interpretation. This perspective motivates the development of other types of filters which, when modulated (i.e., shifted in Fourier space) a few times, can be used to group different subsets of Fourier space together into a small number of shorter intervals, or "bins" (see Figure 3). If the most important frequencies in a function, f, are uniformly spread over a given interval of Fourier space, one will be likely to isolate them from one another by convolving f with a few different modulations of such a filter. This effectively "bins" the Fourier coefficients of $\hat{f}$ into different frequency bins. Once an important frequency is isolated in a filtered version of f, the methods from the last subsection can then be used to recover it via, e.g., a modified binary search.

Note that a good continuous filter can be periodized and discretized for use as part of a discrete SFT. One generally does so to design a discrete filter that is highly sparse in time (i.e., that is "essentially zero" everywhere, except for a small number of time-domain entries). This allows fast convolution calculations to be performed with the filter during frequency binning. This is crucial since these convolutions are used to repeatedly



[FIG2] A demonstration of aliasing-based filtering. The vector under consideration has entries $f_j := \cos(2\pi \cdot 3 \cdot j/12)$ for $j = 0, \ldots, 11$. Note that $\hat{\mathbf{f}} \in \mathbb{C}^{12}$ contains only two nonzero entries: $\hat{f}_3 = 1$, and $\hat{f}_9 = 1$. (a) marks the entries of a subvector, $\mathbf{a} \in \mathbb{R}^3$, of $\mathbf{f}$. Its DFT, $\hat{\mathbf{a}}$, also has three entries, each corresponding to a different subset of $\hat{\mathbf{f}}$. Each subset is labeled using a different symbol in (b). Note that both nonzero entries of $\hat{\mathbf{f}}$ fall into the same subset—both are labeled with a green diamond. (c) marks the entries of another subvector of $\mathbf{f}$ consisting of four entries. Its DFT partitions the entries of $\hat{\mathbf{f}}$ into four subsets [see (d)]. In this case the two nonzero entries of $\hat{\mathbf{f}}$ fall into different subsets: one is labeled with a pentagon and the other with a triangle. (a) Three equally spaced subsamples (the red diamonds). (b) Three frequency bins, one for each residue modulo 3. (c) Four equally spaced subsamples (the red diamonds). (d) Four frequency bins, one for each residue modulo 4.

compute time samples from filtered versions of $\hat{\mathrm{f}}$ during each modified binary search for an important frequency. Furthermore, the DFT of the discrete filter should also have a special structure to aid in the construction of good, low-leakage, frequency filters. (We say that a frequency filter leaks if it has nonzero values at frequencies other than our desired values.) Suppose, e.g., that one wants to isolate the k-largest entries of $\hat{\mathrm{f}}$ from one another. To accomplish this, one should use a filter whose DFT looks like a characteristic function on $\{0, \ldots, O(N/k)\} \subset [0, N) \cap \mathbb{Z}$ (i.e., on a $O(1/k)$-fraction of the "discrete Fourier spectrum" of f). The filter can then be modulated $O(k)$ times to create a filter bank with $O(k)$ approximate "pass regions," each of size $O(N/k)$, that collectively tile all of $[0, N) \cap \mathbb{Z}$. These pass regions form the frequency bins discussed above [Figure 3(b)]. To date several different types of filters have been utilized in sparse FFTs, including Gaussians [12], indicator functions [5], [8], spike trains [6], [17], [18], [23], and Dolph–Chebyshev filters [11].

RANDOMLY BINNING FREQUENCIES

As mentioned in the previous subsection, a filter function can be used to isolate the most important entries of $\hat{\mathrm{f}}$ from one another when they are sufficiently well separated. Unfortunately, an arbitrary vector $\mathrm{f} \in \mathbb{C}^N$ will not generally have a DFT with this property. The largest magnitude entries of $\hat{\mathrm{f}}$ can appear anywhere in principle. One can compensate for this problem, however, by pseudorandomly permuting $\hat{\mathrm{f}}$ so that it "looks" uniformly distributed. As long as the permutation is reversible, any information gathered from the permuted vector can then be directly translated into information about the original vector's DFT, $\hat{\mathrm{f}}$.

Perhaps the easiest means of randomly permuting $\mathrm{f} \in \mathbb{C}^N$, and therefore $\hat{\mathrm{f}}$, is to use two basic properties of the Fourier transform: the scaling property, stating that for $a_j = f_{cj}$ we have $\hat{a}_j = \hat{f}_{c^{-1}j}$ (where c^{-1} is the inverse of c modulo N, assuming it exists); and the modulation property, stating that for $a_j = e^{2\pi i \cdot b \cdot j/N} \cdot f_j$ we have $\hat{a}_j = \hat{f}_{j-b}$. We proceed by choosing two random integers $b, c \in [0, N)$, and defining $\mathrm{a} \in \mathbb{C}^N$ as

$$a_j := e^{2\pi i \cdot b \cdot j/N} \cdot f_{c \cdot j} \tag{22}$$

for $j = 0, \ldots, N-1$. It can be seen that $\hat{\mathrm{a}}$ is a permuted version of $\hat{\mathrm{f}}$, as the entry $\hat{f}_\omega$ appears in $\hat{\mathrm{a}}$ as entry $(\omega \cdot c + b) \bmod N$. Note that permutations of this form are not fully random, even though b and c were selected randomly. Nevertheless, they are "random enough" for our purposes. In particular, the probability that any two nonzero coefficients land close to each other can be shown to be small.

THE PROTOTYPICAL SFT

In the simplest setting, a sparse FFT is a method that is designed to approximately compute the DFT of a vector $\mathrm{f} \in \mathbb{C}^N$ as quickly as possible under the presumption that the result, $\hat{\mathrm{f}} \in \mathbb{C}^N$, will be sparse, or compressible. Here compressible means that $\hat{\mathrm{f}}$ will have a small number of indices (i.e., frequencies) whose entries (i.e., Fourier coefficients) have magnitudes that are large compared to the Euclidean norm of $\hat{\mathrm{f}}$ (i.e., the energy of $\hat{\mathrm{f}}$). Sparse FFTs improve on the runtime of traditional FFTs for such Fourier sparse signals by focusing exclusively on identifying energetic frequencies, and then estimating their Fourier coefficients. This allows sparse FFTs to avoid "wasting time" computing the Fourier coefficients of many insignificant frequencies.

Although several different sparse FFT variants exist, they generally share a common three-stage approach to computing the sparse DFT of a vector: Briefly put, all sparse FFTs (repeatedly) perform some version of the three following steps:

1) identification of frequencies whose Fourier coefficients are large in magnitude (typically a randomized subroutine)
2) accurate estimation of the Fourier coefficients of the frequencies identified in the first step
3) subtraction of the contribution of the partial Fourier representation computed by the first two steps from the entries of f before any subsequent repetitions.

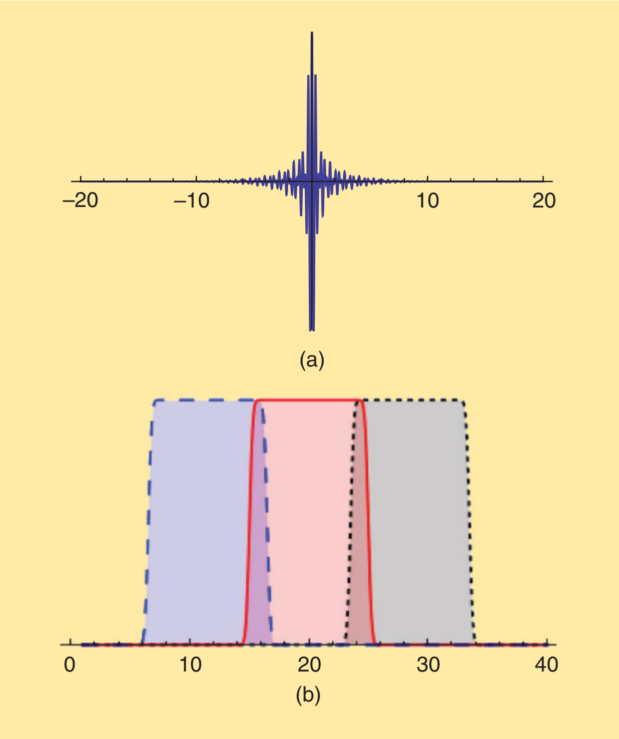
Generally, each repetition of the three stages above is guaranteed to gather a substantial fraction of the energy present in $\hat{\mathrm{f}}$ with high probability. Subtracting the located coefficients from the signal effectively improves the spectral sparsity of the given input vector, f, from one repetition to the next. The end result is that a small number of repetitions will gather (almost) all of the signal energy with high probability, thereby accurately approximating the SFT, $\hat{\mathrm{f}}$, of the given vector f.

Consider, e.g., a vector f whose DFT has 100 nonzero entries (i.e., 100 nonzero Fourier coefficients). The first round of the three stages above will generally find and accurately estimate a large fraction of these entries (e.g., three fifths of them, or 60 terms in this case). The contributions of the discovered terms are then subtracted off of the remaining samples. This effectively reduces the number of nonzero entries in $\hat{\mathrm{f}}$, leaving about 40 terms in the current example. The next repetition of the three stages is now executed as before, but with the smaller effective sparsity of 40. Eventually all nonzero entries of $\hat{\mathrm{f}}$ will be found and estimated after a few repetitions with high probability. We will now consider each of the three stages mentioned above in greater detail.

STAGE 1: IDENTIFYING FREQUENCIES

Stage 1 of each repetition, which identifies frequencies whose Fourier coefficients are large in magnitude, is generally the most involved of the three repeated stages mentioned above. It usually consists of several ingredients, including: randomly sampling f to randomly permute its DFT, filtering to separate the permuted Fourier coefficients into different frequency bands, and estimating the energy in subsets of each of the aforementioned frequency bands. Many of these ingredients are illustrated with concrete examples in the section "Techniques." Our objective now is to understand the general functionality of the identification stage as a whole.

Roughly speaking, stage 1 works by randomly binning the Fourier coefficients of f into a small number of "bins" (i.e., frequency bands), and then performing a single frequency recovery procedure (as described in the section "Phase Encoding") within each bin to find any energetic frequencies that may have been isolated there. The randomness is introduced into the Fourier spectrum of $\mathrm{f} \in \mathbb{C}^N$ by randomly subsampling its entries (see the section "Randomly Binning Frequencies"). This has the effect of randomly permuting the entries of $\hat{\mathrm{f}}$. The resulting "randomized version of $\hat{\mathrm{f}}$" is then binned via a filter bank. (i.e., as discussed in the section "Filtering to Isolate Frequencies"). Because $\hat{\mathrm{f}}$ is approximately sparse, each "frequency bin" is likely to receive exactly one relatively large Fourier coefficient. Each such isolated Fourier coefficient is then identified by using one of the procedures described in the section "Single Frequency Recovery." The collection of frequencies discovered in each different bin is then saved to be analyzed further during the estimation stage 2.



[FIG3] A filter function with small (effective) support, and several frequency bins resulting from different modulations of the filter. The filter is a product of a sinc function with a Gaussian [see (a)]. The Fourier transform of the filter is a characteristic (i.e., box) function convolved with a Gaussian. Three translates of the Fourier transform, each produced by a different modulation of the filter function, are graphed in (b). Note that modulations of the filter can be used to (effectively) regroup the Fourier spectrum into a small number of (essentially) disjoint frequency bins. (a) The filter function. (b) The Fourier space: three modulations of the filter.

STAGE 2: ESTIMATING COEFFICIENTS

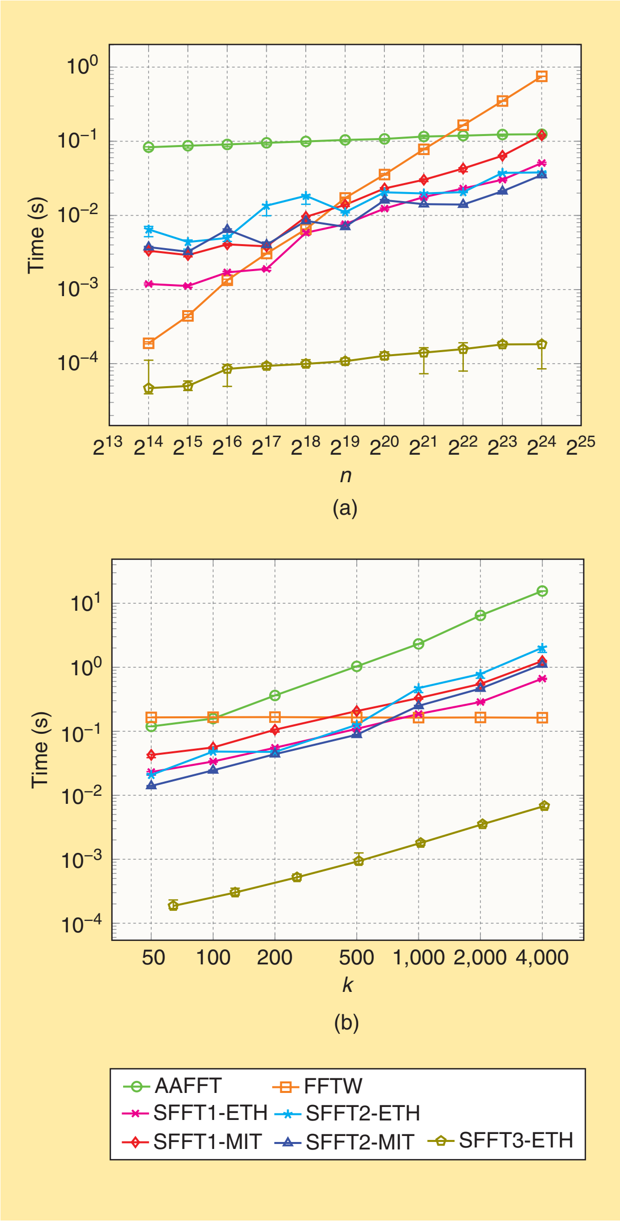
Recall that stage 2 involves estimating the Fourier coefficient, $\hat{f}_\omega$, of each frequency ω identified during stage 1 of the sparse FFT. In the simplest case, this can be done for each such ω by using $L \ll N$ independent and uniformly distributed random samples from the entries of f, $f_{u_1}, \ldots, f_{u_L}$, to compute the estimator

$$\hat{f}'_\omega := \frac{1}{L} \cdot \sum_{l=1}^{L} f_{u_l} e^{-2\pi i \cdot \omega \cdot u_l / N}. \qquad (23)$$

Note that $\hat{f}'_\omega$ is an unbiased estimator for $\hat{f}_\omega$ (i.e., $\mathrm{E}[\hat{f}'_\omega] = \hat{f}_\omega$) whose variance is $O(\| \hat{\mathbf{f}} \|_2^2/L)$. Thus, the estimator will approximate $\hat{f}_\omega$ to high (relative) precision with high probability whenever $|\hat{f}_\omega|^2$ is large compared to $\| \hat{\mathbf{f}} \|_2^2/L$. In slightly more complicated scenarios, the estimates might come "for free" as part of the identification stage (see the section "Filtering to Isolate Frequencies" for an example).

STAGE 3: REPEATING

A naive implementation of stage 3 is even more straightforward than stage 2. Suppose that



[FIG4] Running time plots for several algorithms and implementations of sparse FFT. (a) Runtime as a function of the signal length *N*, for *k* = 50. (b) Runtime as a function of the signal length *k*, for *N* = 2^{22}.

$$\{\hat{f}'_{\omega_m} \mid m = 1, \ldots, k\} \subset \mathbb{C}$$

is the approximate sparse DFT discovered for $\mathbf{f}$ during stages 1 and 2 of the current repetition of our sparse FFT. Here, $\omega_1, \ldots, \omega_k$ are the frequencies identified during stage 1, while $\hat{f}'_{\omega_1}, \ldots, \hat{f}'_{\omega_m}$ are the estimates for their Fourier coefficients found during stage 2 [e.g., via (23)]. In future iterations of stages 1–3, one can simply replace each sampled entry of f, f_j, with

$$f_j - \sum_{m=1}^{k} \hat{f}'_{\omega_m} e^{2\pi i \cdot \omega_m \cdot j/N}. \tag{24}$$

If the entries of $\mathbf{f}$ to be used during each iteration of the three stages have been predetermined, which is often the case, (24) can be used to update them all at once. These "updated samples" are then used in the subsequent repetitions of the three stages.

One of the primary purposes of stage 3 is to avoid mistakenly identifying insignificant frequencies as being energetic. Suppose, e.g., that a small number of erroneous Fourier coefficients are identified during the jth repetition of the first two stages. Then, subtracting their contribution from the original signal samples during the third stage will effectively add them as new, albeit erroneous, energetic Fourier coefficients in $\hat{\mathbf{f}}$. This, in turn, allows them to be corrected in subsequent repetitions of the first two stages. Hence, stage three allows errors (assuming they are rare) to be corrected in later repetitions.

In contrast, some SFT methods [12], [17], [18] perform only stages 1 and 2 without any stage 3. These methods identify all the energetic frequencies in stage 1 and then estimate their Fourier coefficients in stage 2, completely in only one iteration. Of course, such methods can also mistakenly identify significant frequencies as being energetic during stage 1. Such mistaken frequencies are, however, generally discovered as being insignificant by these methods later, during stage 2, when their Fourier coefficients are estimated.

EMPIRICAL EVALUATION

In this section, we compare several existing SFT implementations to the fastest Fourier transform in the West (FFTW), a fast implementation of the standard FFT, to demonstrate the computational gains that recent SFTs can provide over the standard FFT when dealing with Fourier-compressible signals. To this end, we consider the following algorithms and implementations:

- FFTW: base line implementation of the standard FFT
- AAAFT: an implementation of [8]
- SFFT1-MIT, SFFT2-MIT: implementations of the algorithms in [12] by the authors
- SFFT1-ETH, SFFT2-ETH: implementations of the algorithms in [12] given in [25]
- SFFT3-ETH: implementation of a variant of the algorithm in [11] given in [25].

All implementations are freely available. FFTW is available at http://www.fftw.org. AAFFT, as well as several significantly faster sampling-based SFTs, are available at http://sourceforge.net/projects/aafftannarborfa/. The ETH implementations are available at http://www.spiral.net/software/sfft.html. All other SFT implementations are available at http://groups.csail.mit.edu/netmit/sFFT/. The SFT variants considered herein are limited to those which compute the DFT of a vector [i.e., (2)]. Additional experiments involving other existing SFT variants that sample continuous functions, as well as additional experiments demonstrating noise tolerance and sampling complexity, can be found at https://github.com/ludwigschmidt/sft-experiments.

Figure 4 plots the runtimes of these SFTs against FFTW for various sparsity levels, k, and vector lengths, N. The algorithms were run on randomly generated vectors of length N whose DFTs were k-sparse (containing k ones in randomly chosen locations) for varying values of k and N. For each pair of values of k and N, the parameters of the (randomized) algorithms were optimized to minimize the running time while ensuring that the empirical probability of correct recovery was greater than 0.9.

APPLICATIONS

In this section, we give an overview of some of the data-intensive applications of sparse FFT algorithms and techniques that emerged over the last few years. These applications involve, e.g., GPS receivers, cognitive radios, and, more generally, any analog signal that we wish to digitize. It is these applications that we focus on as they highlight the role of sparse FFT algorithms in the signal processing of large data.

GPS SYNCHRONIZATION

In the (simplified) GPS synchronization problem, we are given a (pseudorandom) code, corresponding to a particular satellite. (For simplicity, this description ignores certain issues such as the Doppler shift, etc. See [10] for details.) The satellite repeatedly transmits the code. Furthermore, we are given a signal recorded by a GPS receiver, which consists of a window of the signal generated by the satellite, corrupted by noise. The goal is to align the code to the recorded signal, i.e., identify where the code starts and ends. To this end, the receiver computes the convolution of the code and the received signal and reports the shift that maximizes the correlation. This computation is typically done using the FFT: one applies the FFT to the code and the signal, computes the product of the outputs, and applies the inverse FFT to the product.

The paper [10] uses sparse FFT techniques to speed up the process. The improvement is based on the following observation: since the output of the inverse FFT contains a single peak corresponding to the correct shift, the inverse step can be implemented using the sparse FFT. In fact, since $k = 1$, the algorithm is particularly simple, and relies on a simple aliasing filter. Furthermore, since the sparse inverse FFT algorithm uses only some of the samples of the product, it suffices to compute only those samples. This reduces the cost of the forward step as well. The experiments on real signals show that the new algorithm reduces the median number of multiplications by a factor of 2.2, or more if the value of the Doppler shift is known.

SPECTRUM SENSING

The goal of a spectrum sensing algorithm is to scan the available spectrum and identify the "occupied" frequency slots. In many applications this task needs to be done quickly, since the spectrum changes dynamically. Unfortunately, scanning a GHz-wide spectrum is a highly power-consuming operation. To reduce the power and acquisition time, one can use an SFT to compute the frequency representation of a sparse signal without sampling it at full bandwidth. One such proposal was presented in [28], which uses a method of frequency identification similar to that described in the section "An Aliasing-Based Search." Another approach is presented in [14], which uses a sparse FFT procedure similar to that in [6], [23]. It describes a prototype device using three software radios called Universal Software Radio Peripherals (USRPs), each sampling the spectrum at 50 MHz. The device captures 0.9 GHz, i.e., six times larger bandwidth than the three USRPs combined.

ANALOG TO DIGITAL CONVERTERS

The random binning procedure described in the section "Randomly Binning Frequencies" forms the basis of the pulse-position modulation (PPM) analog to digital converter presented in [29]. A prototype 9-bit random PPM analog to digital converter incorporating a pseudorandom sampling scheme is implemented as proof of concept. The approach leverages the energy efficiency of time-based processing.

OTHER APPLICATIONS

Other applications include 2-D correlation spectroscopy [24].

CONCLUSIONS

It is interesting to note that SFTs have a good deal in common with compressive sensing techniques. The latter generally aim to reduce sampling requirements as much as possible to recover accurate sparse approximations of frequency-compressable functions. SFTs, on the other hand, attempt to recover accurate sparse approximations of frequency-compressible functions as quickly as absolutely possible. By necessity, therefore, an SFT also cannot sample a function many times (i.e., since sampling takes time). As a result, SFTs also utilize a relatively small number of samples and, so, can be considered as compressive sensing algorithms. This puts SFTs into a broader spectrum of compressive sensing strategies that tradeoff additional sampling for decreased computational complexity.

SFTs are also closely related to streaming (or sublinear) algorithms developed in the computer science community. Streaming algorithms aim to run in time considerably smaller than the time required to read the entire original data set, or signal. Hence, the streaming literature contains a rich set of tools for processing and approximating large data sets both quickly and accurately. Many of the techniques employed in existing SFTs are adapted to the Fourier setting from streaming techniques. For an overview of the links between these two topics, see [16].

ACKNOWLEDGMENTS

We would like to thank Haitham Hassanieh, Chinmay Hegde, David Lawlor, Eric Price, and Jörn Schumacher for providing source code and very helpful discussions. Anna C. Gilbert was supported in part by National Science Foundation (NSF) CIF 0910765 and NSF CCF 1161233. Piotr Indyk was supported in part by the Simons Foundation and NSF CCF 1065125. Mark Iwen was supported in part by NSA grant H98230-13-1-0275.

AUTHORS

Anna C. Gilbert (annacg@umich.edu) is a professor of mathematics at the University of Michigan. She was a postdoctoral fellow at AT&T Labs and Yale University in 1997. From 1998 to 2004, she was a member of technical staff at AT&T Labs. Her research interests include randomized algorithms with applications to harmonic analysis, signal and image processing, networking, and massive data sets. She has received several awards for her research on streaming algorithms and sparse Fourier transforms, including a Sloan Research Fellowship (2006), a National Science Foundation CAREER Award (2006), the National Academy of Sciences Award for Initiatives in Research (2008), and the SIAM Ralph E. Kleinman Prize (2013).

Piotr Indyk (indyk@theory.lcs.mit.edu) is a professor of electrical engineering and computer science at the Massachusetts Institute of Technology (MIT). He joined MIT in 2000 after earning a Ph.D. degree from Stanford University. His research interests lie in the design and analysis of efficient algorithms. Specific interests include high-dimensional computational geometry, sketching and streaming algorithms, sparse recovery, and compressive sensing. He received the Sloan Fellowship (2003), the Packard Fellowship (2003), and the Simons Investigator Award (2013). His work on sparse Fourier sampling was named to *MIT Technology Review* TR10 in 2012, while his work on locality-sensitive hashing has received the 2012 ACM Kanellakis Theory and Practice Award.

Mark Iwen (markiwen@math.msu.edu) earned a Ph.D. degree in applied and interdisciplinary mathematics from the University of Michigan in 2008. From 2008 to 2010, he was a postdoctoral fellow at the Institute for Mathematics and its Applications, and then moved to Duke University as a visiting assistant professor from 2010 until 2013. He has been an assistant professor at Michigan State University since the fall of 2013.

Ludwig Schmidt (ludwigs@mit.edu) is a Ph.D. student in the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology (MIT), where he works with Prof. Piotr Indyk in the Theory of Computation group. His interest lies in algorithms, with a focus on applications outside theoretical computer science. Specific interests include signal processing, especially compressive sensing and similarity search. Before coming to MIT, he received his bachelor's degree in computer science from the University of Cambridge in 2011.

REFERENCES

[1] A. Akavia, S. Goldwasser, and S. Safra, "Proving hard-core predicates using list decoding," in *Proc. Annu. IEEE Symp. Foundations of Computer Science*, vol. 44, pp. 146–159, 2003.

[2] O. Abari, E. Hamed, H. Hassanieh, A. Agarwal, D. Katabi, A. P. Chandrakasan, and V. Stojanovic, "A 0.75-million-point Fourier-transform chip for frequency-sparse signals," in *IEEE Int. Solid-State Circuits Conf. Dig. Technical Papers (ISSCC)*, 2014.

[3] A. Akavia, "Deterministic sparse Fourier approximation via fooling arithmetic progressions," in *Proc. 23rd Conf. Learning Theory (COLT)*, 2010, pp. 381–393.

[4] P. Boufounos, V. Cevher, A. C. Gilbert, Y. Li, and M. J. Strauss, "What's the frequency, Kenneth? Sublinear Fourier sampling off the grid," in *Proc. RANDOM/APPROX*, 2012.

[5] A. Gilbert, S. Guha, P. Indyk, M. Muthukrishnan, and M. Strauss, "Near-optimal sparse Fourier representations via sampling," in *Proc. 34th Annu. Symp. Theory of Computing (STOC)*, 2002.

[6] B. Ghazi, H. Hassanieh, P. Indyk, D. Katabi, E. Price, and L. Shi, "Sample-optimal average-case sparse Fourier transform in two dimensions," in *Proc. Allerton*, 2013.

[7] O. Goldreich and L. Levin, "A hard-core predicate for all one-way functions," in *Proc. 21st Annu. Symp. Theory of Computing (STOC)*, 1989, pp. 25–32.

[8] A. Gilbert, M. Muthukrishnan, and M. Strauss, "Improved time bounds for near-optimal space Fourier representations," in *Proc. SPIE Conf. Wavelets*, 2005.

[9] A. C. Gilbert, M. J. Strauss, and J. A. Tropp, "A tutorial on fast Fourier sampling," *IEEE Signal Processing Mag.*, 2008.

[10] H. Hassanieh, F. Adib, D. Katabi, and P. Indyk, "Faster GPS via the sparse Fourier transform," in *Proc. 18th Annu. Int. Conf. Mobile Networking and Computing (MOBICOM)*, 2012.

[11] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Near-optimal algorithm for sparse Fourier transform," in *Proc. 44th Annu. ACM Symp. Theory of Computing (STOC)*, 2012.

[12] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Simple and practical algorithm for sparse Fourier transform," in *Proc. 23rd Annu. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2012.

[13] S. Heider, S. Kunis, D. Potts, and M. Veit, "A sparse Prony FFT," in *Proc. 10th Int. Conf. Sampling Theory and Applications (SAMPTA)*, 2013.

[14] H. Hassanieh, L. Shi, O. Abari, E. Hamed, and D. Katabi, "Ghz-wide sensing and decoding using the sparse Fourier transform," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*, 2014.

[15] P. Indyk, M. Kapralov, and E. Price, "(Nearly) sample-optimal sparse Fourier transform," in *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, 2014.

[16] P. Indyk, "Sketching via hashing: From heavy hitters to compressed sensing to sparse Fourier transform," in *Proc. 32nd Symp. Principles of Database Systems (PODS)*, 2013, pp. 87–90.

[17] M. A. Iwen, "Combinatorial sublinear-time Fourier algorithms," *Found. Computat. Math.*, vol. 10, pp. 303–338, 2010.

[18] M. A. Iwen, "Improved approximation guarantees for sublinear-time Fourier algorithms," *Appl. Computat. Harm. Anal.*, vol. 34, pp. 57–82, 2013.

[19] E. Kushilevitz and Y. Mansour, "Learning decision trees using the Fourier spectrum," in *Proc. 23rd Annu. Symp. Theory of Computing (STOC)*, 1991.

[20] L. A. Levin, "Randomness and non-determinism," *J. Symb. Logic*, vol. 58, no. 3, pp. 1102–1103, 1993.

[21] D. Lawlor, Y. Wang, and A. Christlieb, "Adaptive sub-linear time Fourier algorithms," *Adv. Adapt. Data Anal.*, vol. 5, no. 1, 2013.

[22] Y. Mansour, "Randomized interpolation and approximation of sparse polynomials," in *Proc. 19th Int. Colloq. Automata, Languages and Programming (ICALP)*, 1992.

[23] S. Pawar and K. Ramchandran, "Computing a k-sparse n-length discrete Fourier transform using at most $4k$ samples and $O(k \log k)$ complexity," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2013.

[24] L. Shi, O. Andronesi, H. Hassanieh, B. Ghazi, D. Katabi, and E. Adalsteinsson, "Mrs sparse-fft: Reducing acquisition time and artifacts for in vivo 2D correlation spectroscopy," in *Proc. Int. Society for Magnetic Resonance in Medicine Annu. Meeting & Exhibition (ISMRM)*, 2013.

[25] J. Schumacher, "High performance sparse fast Fourier transform," Master's thesis, Computer Science, ETH Zurich, Switzerland, 2013.

[26] R. Scheibler, S. Haghighatshoar, and M. Vetterli, "A sparse sub-linear Hadamard transform," in *Proc. Allerton*, 2013.

[27] C. Wang, M. Araya-Polo, S. Chandrasekaran, A. St-Cyr, B. Chapman, and D. Hohl, "Parallel sparse FFT," in *Proc. SC Workshop on Irregular Applications: Architectures and Algorithms*, 2013.

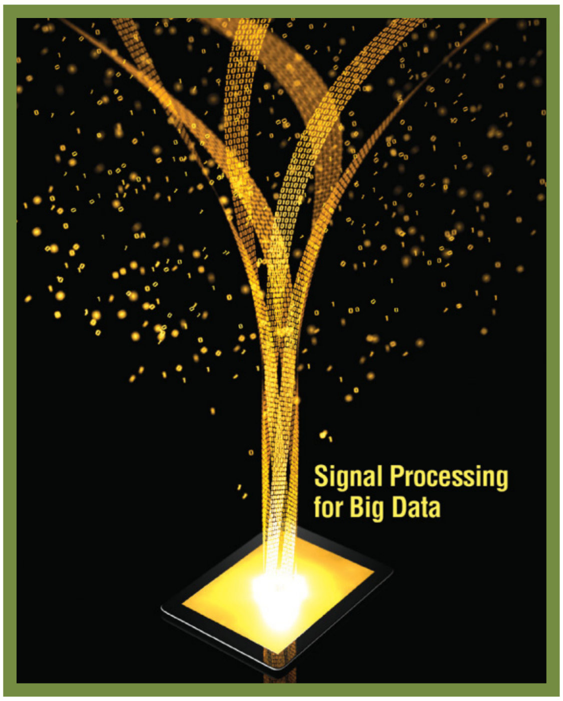
[28] P. Yenduri and A. C. Gilbert, "Compressive, collaborative spectrum sensing for wideband cognitive radios," in *Proc. Int. Symp. Wireless Communication Systems (ISWCS)*, 2012, pp. 531–535.

[29] P. K. Yenduri, A. Z. Rocca, A. S. Rao, S. Naraghi, M. P. Flynn, and A. C. Gilbert, "A low-power compressive sampling time-based analog-to-digital converter," *IEEE J. Emerg. Select. Topics Circuits Syst.*, vol. 2, no. 3, pp. 502–515, 2012.

[SP]

Steven Verstockt, Viktor Slavkovikj, Pieterjan De Potter, and Rik Van de Walle

Collaborative Bike Sensing for Automatic Geographic Enrichment



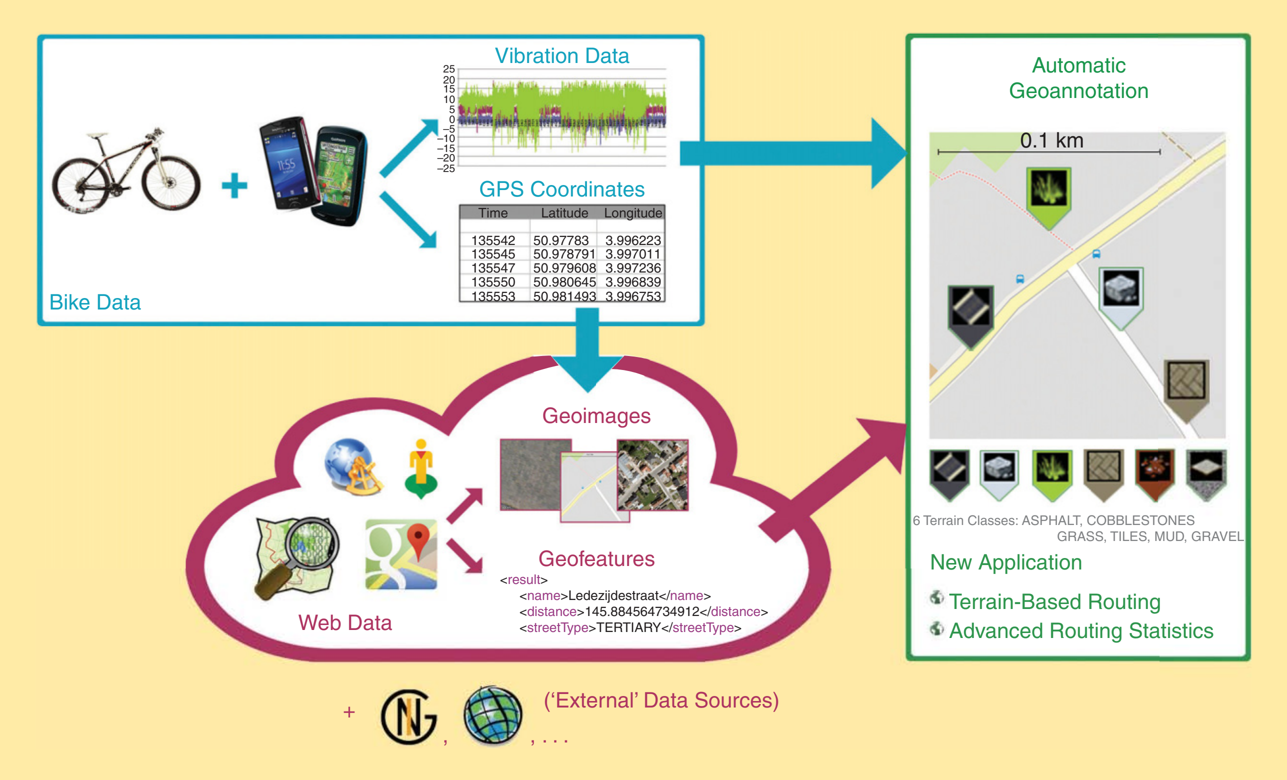
© ISTOCKPHOTO.COM/TA2YO4NORI

Geoannotation of road/terrain type by multimodal bike sensing

In this article, we describe a multimodal bike-sensing setup for automatic geoannotation of terrain types using Web-based data enrichment. The proposed classification system is mainly based on the analysis of volunteered geographic information gathered by cyclists. By using participatory accelerometer and global positioning system (GPS) sensor data collected from cyclists' smartphones, which is enriched with data from geographic Web services, the proposed system is able to distinguish between six different terrain types. For the classification of the Web-based enriched sensor data, the system employs a random decision forest (RDF) (which compared favorably for the geoannotation task against different classification algorithms). The accuracy of the novel bike-sensing system is 92% for six-class road/terrain classification and 97% for two-class on-road/off-road classification. Since the evaluation is performed on large-scale data gathered during real bike runs, these "real-life" accuracies show the feasibility of our novel approach.

Digital Object Identifier 10.1109/MSP.2014.2329379
Date of publication: 19 August 2014

1053-5888/14©2014IEEE



Time	Latitude	Longitude
135542	50.97783	3.996223
135545	50.978791	3.997011
135547	50.979608	3.997236
135550	50.980645	3.996839
135553	50.981493	3.996753

```
<result>
    <name>Ledezijdestraat</name>
    <distance>145.884564734912</distance>
    <streetType>TERTIARY</streetType>
```

[FIG1] A multimodal bike-sensing setup for automatic geoannotation of terrain types.

INTRODUCTION

THE GEOSENSING (R)EVOLUTION

The beginning of the 21st century is characterized by a mobile sensing (r)evolution. Mobile phones have increasingly evolved in functionality, features, and capability and are being used by many for more than just communication. With the continuous improvement in sensor technology built into these devices, and Web services to aggregate and interpret the logged information, people are able to create, record, analyze, and share a huge amount of data about their daily activities or the places they visit. As such, the mobile phone is well on its way to becoming a personal sensing platform [19].

Within this mobile sensing (r)evolution, phone users act as sensor operators, and more people start to contribute their sensor measurements as part of a larger-scale effort to collect data about a population or a geographical area. This is the idea behind participatory or human-centric sensing. By combining mobile data taken from large groups of individuals, it is possible to derive new values for end users in ways that the contributor of the content even did not plan or imagine as well as perform functions that are either difficult to automate or expensive to implement.

Recently, the tendency of participatory data gathering has also started to occur in the domain of geographic information systems (GIS). Where the process of mapping the Earth has been the task of a small group of people (surveyors, cartographers, and geographers) for many years, it has started to become possible now for everyone to participate in several types of collaborative geographic projects, such as OpenStreetMap (OSM) and RouteYou [5]. These projects are built upon user-generated geographic content, so-called volunteered geographic information (VGI). VGI makes it easier to create, combine, and share maps and supports the rapid production of geographic information. One drawback of current VGI approaches, however, is that a lot of the work still involves manual labor. Within our article, we focus on how mobile sensors can help to automate and facilitate the more labor-intensive VGI tasks.

A common task performed by recreational GPS users is to find good routes in an area. From all the route characteristics, the road quality, i.e., the physical condition of the terrain, and the terrain surface showed to have a significant impact on how the users rank their routes [15]. Currently, however, this information is largely unavailable. To bridge this gap, there is a need for automatic road classification. We investigate the ability to determine the current terrain type from onboard mobile sensors (i.e., from a smartphone mounted on a bike), enriched with geographic Web data from the GPS coordinates. Contrarily to manual VGI, our approach facilitates real-time updates/annotation, e.g., when road conditions change or new roads are found. Furthermore, by using common phones, it is not required to buy expensive, specialized sensing equipment, keeping the costs very low. Finally, the novel approach allows creating more advanced route statistics.

MULTIMODAL BIKE SENSING

A general overview of our multimodal bike-sensing setup is shown in Figure 1. Both bike data and Web data are used to extract geographic information of the terrain the user is traversing. The bike data consists of accelerometer/vibration signals and GPS coordinates. Both are collected using the onboard smartphone of the cyclist. Important to remark is that for the collection of the bike data, the device can be placed or stored as wanted by the user. Compared to our camera-based multimodal bike-sensing system proposed in [22], this gives more flexibility and freedom to the user. The Web data consists of a set of geographic images and features centered at the location that corresponds to the bike GPS coordinates. To retrieve this information, we query the Web application programming interfaces (APIs) of online geoservices like Google Maps, Streetview, OSM, and GisGraphy. When available, it is also possible to use external data sources. Based on the bike data and corresponding online geoimages and features, the terrain type is estimated using a multimodal RDF-based classifier, which is fed with a set of discriminative image and accelerometer features. Finally, a geographic map can be annotated automatically using this road/terrain information or advanced route statistics can be generated.

IN THIS ARTICLE, WE INVESTIGATE THE ABILITY TO DETERMINE THE CURRENT TERRAIN TYPE FROM "ONBOARD" MOBILE SENSORS (I.E., FROM A SMARTPHONE MOUNTED ON A BIKE), ENRICHED WITH GEOGRAPHIC WEB DATA FROM THE GPS COORDINATES.

RELATED WORK

MOBILE SENSING: GIS TERRAIN CLASSIFICATION

The majority of mobile-sensing solutions for GIS road/terrain classification either use accelerometer data or visual images. Although they can easily (and successfully) be combined, the combination of both sensor types is only scarcely/marginally investigated. First, we will discuss the state-of-the-art accelerometer- and camera-based single-sensor approaches. Second, we zoom in on some multimodal/multisensor approaches.

Weiss et al. [25] used an accelerometer mounted on a robot to perform vibration-based road classification. To train and classify the vibration signals, they fed a set of distinctive accelerometer features to a support vector machine (SVM). Although they achieved 80% correct classifications, the speed of the vehicle is not realistic (i.e., too slow) and the experiments were performed in a "controlled" environment. The accelerometer features, however, were well chosen and are (partly) used in our setup. A similar approach was presented by Ward and Iagnemma [24], where the algorithm was validated in real-world conditions. They classified multiple terrain types as 89% correct. However, they made use of expensive, specialized sensing equipment, and the classifier was only trained to recognize four very distinctive classes. When the classes' vibration behavior was closer to each other, e.g., when comparing tiles to cobblestones, the confusion of classes is expected to be higher, leading to lower accuracy. By using visual features, in addition to accelerometer data, we are able to tackle this problem.

Tang and Breckon [21] classified urban, rural, and off-road terrains by analyzing several color and texture features. They reported a performance of almost 90% correct SVM classification. A drawback of their method, however, was the genericity of the terrain classes, i.e., too broad for recreational purposes. Similar limitations arise in [14]. What is interesting, however, is that these authors perform a "voting" over small image regions. In this way, conflicting or confusing zones can be detected and eliminated, leading to higher classification accuracy. Furthermore, it is important to mention that the majority of visual approaches use an "unrealistic" setup, i.e., sharp images containing a single terrain type captured from a perpendicular camera angle. Our approach, on the other hand, uses images from online geoservices, containing blurred images with nonsharp terrain boundaries. As such, our 72% for six-class road/terrain classification and 88% for two-class on-road/off-road classification are "real-life" accuracies. Although SVM has shown to perform best in the related work, Khan et al. [11] recently showed that RDF beat SVM in the context of terrain classification. This hypothesis was also confirmed by our experiments.

Recently, some authors also investigated the combination of accelerometer and visual images for terrain classification [18], [23], [25]. However, none of these works was found to have the same level of flexibility/freedom as the proposed mobile approach with online geoenrichment. Both Wang et al. [23] and Smith III [18] made use of an instrumented "calibrated" road vehicle, which limited the practical use of such systems for large-scale collaborative sensing. A similar remark holds for the robot setup of Weiss [25]. Due to the different hardware/sensors used and the differences in terrain types/data sets, direct comparison with these systems is also difficult. Our accuracy of 92% for six-class road/terrain classification, however, is already higher than the reported accuracy of 90% of Wang's four-class terrain classification. In combination with the higher usability, we believe that we may say that our system improves the state of the art in this domain. Furthermore, by focusing on computational low-cost feature extraction and classification, the proposed system is optimized for large-scale collaborative sensing.

DATA ENRICHMENT USING WEB APIs

While the state-of-the-art approaches discussed in the previous section only use their own sensor data to detect the terrain type, we expect it is beneficial to use publicly available geodata from the Internet. With the growing availability of geodata Web services, it is possible to achieve a unique combination of geographic data of different origin coupled to the location's coordinates. In this section, we will briefly discuss the related work in this domain.

Hariharan et al. [7] described several applications that take advantage of existing Web data combined with GPS location measurements. Pinpoint Search, for example, converts the (latitude, longitude) pair of a location into search terms for a search engine, giving Web pages relevant to the user's immediate surroundings. The conversion from the raw (latitude, longitude) pair into a street address or a closely located point-of-interest (POI) is performed using a geocoding API. With some of these geocoders, like the GisGraphy API, it is even possible to find the distance to the closest matching street. In our setup, this feature is used to facilitate the road/off-road terrain classification.

Pannevis and Marx [13] discussed several providers of location-related Web data and list the problems related to each of these services. The two main problems geospatial data may suffer from are the variable quality and the description conflicts [10]. The first one concerns updating, completeness, and accuracy of the data. The second problem concerns inconsistent descriptions provided by different sources for the same location. To cope with geodata from a different origin with different data models, resolution, and types of geometric representations, we extract and weight the geographical features from each geoservice individually and do not merge the data itself.

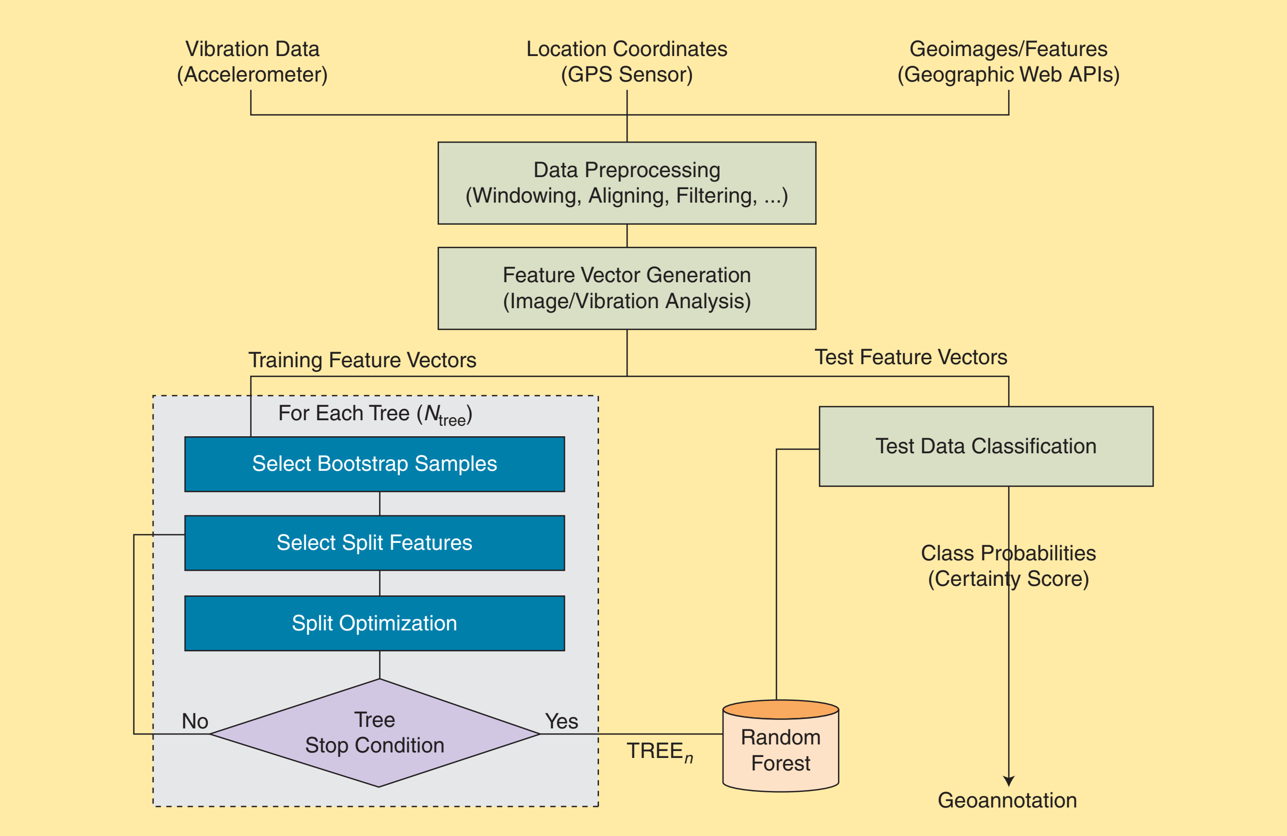
MULTIPLE CYCLES WITH VARYING TERRAIN CONDITIONS (IN TYPE AND FREQUENCY) WERE PERFORMED IN SEVERAL RURAL AND (SUB)URBAN REGIONS ALL OVER BELGIUM.

The works most closely related to our approach are [8] and [9]. Both approaches query OSM data to enrich a location-based mobile application. The first work improves autonomous robot navigation in urban environments using the free-to-use and globally available online geographic OSM data. The latter work presents a mobile application that enables location-based haptic exploration of OSM data for visually impaired users. Both OSM-based approaches show the feasibility of Web-based geodata enrichment. Instead of only focusing on OSM, our approach also uses other geographic data providers to improve the overall classification result.

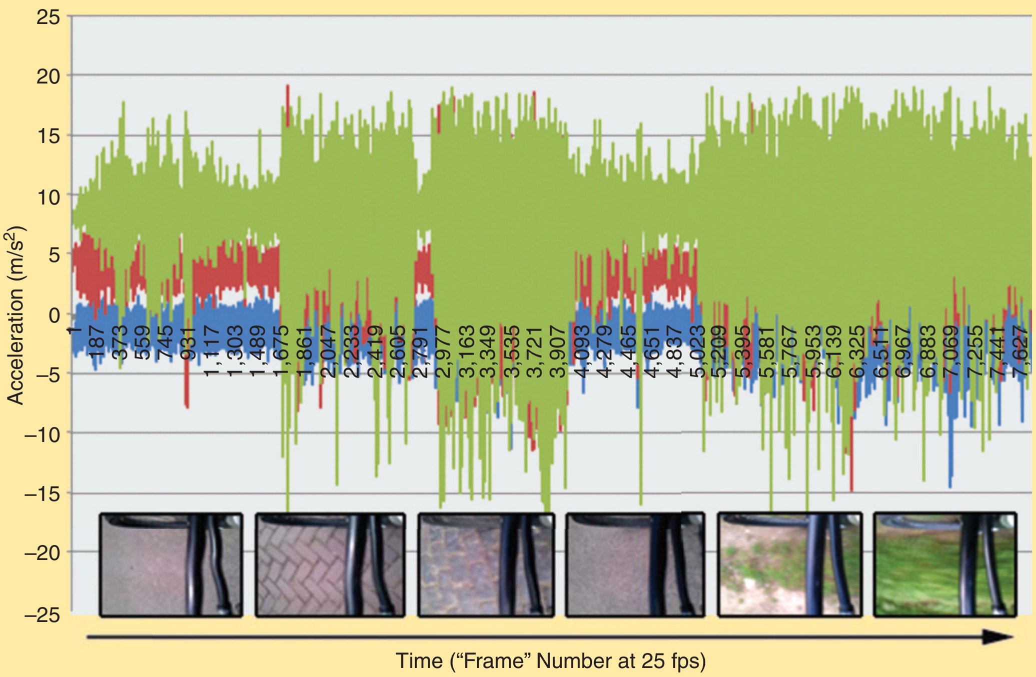
TERRAIN CLASSIFICATION

The multimodal bike-sensing system is built upon two sensing components (an accelerometer and a GPS sensor) and a location-based querier of geographic Web APIs. Each of these "data providers" independently and concurrently captures terrain data. Based on this multimodal data, the proposed terrain classification system estimates which type of terrain (asphalt, cobblestones, tiles, gravel, grass, and mud) the vehicle is currently traversing.

A general scheme of the classification system is shown in Figure 2. First, the raw sensor data is preprocessed. The windowing



[FIG2] A general scheme of the multimodal RDF-based terrain classification.



[FIG3] Exemplary accelerometer data along the X, Y, and Z axes. Visual images of corresponding terrain types are shown below the graph. By visual inspection of the accelerometer data, accelerometer differences between the terrain types can be noticed. Based on these differences, we have constructed our features.

groups the vibration data into overlapping data fragments of 5 s and aligns them onto the corresponding geoimages and the GPS data. Subsequently, we further process/analyze the sensor data to create a set of training and test feature vectors (which is discussed in detail in the next section). Next, the training vectors are used to construct a random forest of binary decision trees (as explained in the section "RDF Classification"). Finally, the test vectors are classified using the trained RDF. Based on the RDF class probabilities and corresponding GPS data, geoannotation of test data can be performed.

FEATURE EXTRACTION

For each of the sensor data segments, i.e., for each 5 s of biking, we extract a set of discriminative Web and vibration features that best describe the terrain conditions. The selection of these features is based on the "State of the Art" (SOTA) study (see the section "Related Work"), and on our test data evaluation (see the section "Experimental Setup and Evaluation Results"). When features show similar behavior, the feature with lowest computational cost is chosen.

ACCELEROMETER/VIBRATION FEATURES

By inspecting the accelerometer readings for the different terrain types, it was found that not every road type has a distinct pattern. Similar "feature equalities" occur when analyzing the geo-Web images, however, not between the same pairs of road/terrain types. As such, by performing a multimodal analysis it is expected that the ambiguities in the vibration data can be compensated by visual data, and vice versa.

The accelerometer of our mobile device(s) detects the vibration along the X, Y, and Z-axes (see Figure 3). It is important to remark that, depending on the position of the device, the tri-axial acceleration values $\{A_x, A_y, A_z\}$ will vary and complicate the classification task. To overcome this obstacle of forcing the user to place the device in a predefined position, the magnitude m of the accelerometer A is calculated in a similar way as in [2] using

$$m = \sqrt{A_x^2 + A_y^2 + A_z^2}. \tag{1}$$

Computing (and analyzing) the features on the vibration magnitude m, instead of on the individual accelerometer data along the X, Y, and Z axes, enables our system to assume an arbitrary and possibly changing orientation for the mobile device, i.e., increases the user's freedom [16].

The set of features that were found to best describe the bike vibrations are a combination of the ones proposed in [25] and [16], and are defined as:

- $\mu(m)$: mean of m–for flatter/smoother surfaces (e.g., asphalt), $\mu(m)$ is low (close to 0)
- $\max(m)$: maximum of m–takes large values for terrain types that contain big bumps, e.g., cobblestones and grass/mud
- $\min(m)$: minimum of m–takes larger values for flat terrains (e.g., asphalt)

- $\sigma(m)$: standard deviation of m–is higher for coarse terrain types (e.g., gravel) than for smoother ones (such as tiles and asphalt)
- $\|m\|$: norm of m–is large if the acceleration is constantly high, as it is for cobblestones
- $E(m)$: energy, i.e., squared fast Fourier transform sum of m [28]–takes larger values for coarse terrains.

It is important to remark that each of these vibration features is calculated over a sliding overlapping time window of 5 s, to align them with the Web-based geoimages/features. A similar windowing approach has demonstrated success in a previous work [29].

THE CURRENT DATA SET CONTAINS OVER 16,000 "REAL-LIFE" TERRAIN SAMPLES IN TOTAL TO TRAIN AND TEST OUR RDF-BASED CLASSIFICATION ALGORITHM.

Table 1 shows exemplary accelerometer feature values for each of the investigated terrain types. This makes clearer the relation between each of the features and the road/off-road terrain types. Each of these features can be calculated in real time on the mobile device or can be generated at the server based on the raw accelerometer data. The former approach is "battery-consuming" and the latter approach is "network-consuming." Due to the low computational cost of our features, the former approach is chosen.

WEB-BASED GEOIMAGES/FEATURES

To enrich the accelerometer features, we have evaluated several geographic Web services based on their ease of use, data type, and accuracy. First we discuss the geofeatures that directly, i.e., without preprocessing, could be fed to the RDF. Next, we go more into detail on the geoimages/maps features.

GISGRAPHY GEOFEATURES

The Gisgraphy World Geocoding API allows for finding address information for a given GPS coordinate pair via a representational state transfer Web service. The two most interesting features for our setup are the distance to the closest matching street and the street type. The street type, however, tends not to be well documented (as revealed by our experiments). As such, only the distance feature is used:

- d(lat,long): distance from current location to closest matching street–higher for off-road.

[TABLE 1] EXEMPLARY ACCELEROMETER FEATURES FOR EACH OF THE INVESTIGATED TERRAIN TYPES.

	ASPHALT	COBBLE-STONES	TILES	GRASS	MUD	GRAVEL
$\mu(m)$	9.811	10.777	10.248	11.274	10.899	9.121
MAX(m)	12.904	27.692	19.193	21.781	18.435	17.036
MIN(m)	6.749	1.909	3.201	1.916	4.022	3.294
$\sigma(m)$	1.139	4.185	3.559	4.537	3.284	3.050
$\|m\|$	108	129	120	135	126	107
$\mathbf{E}(m)$	11,705	16,556	14,463	18,144	15,927	11,461

MAP IMAGE FEATURES

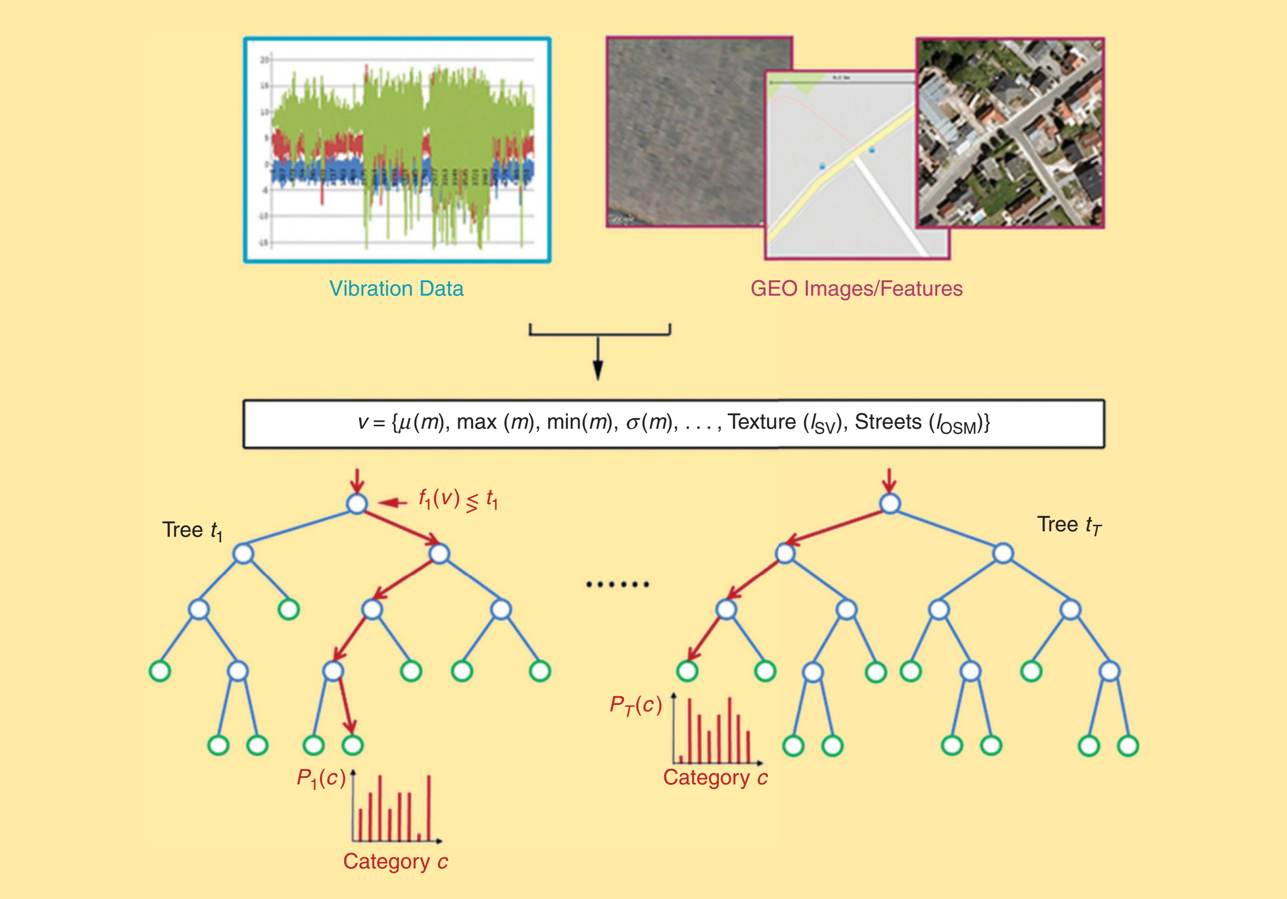
Most of the geographic Web services also allow for images to be queried from the neighborhood of a (lat, long)-pair. Depending on the service, these images may differ in detail, colors, and content or representation. As such, it is necessary to convert each of these image types individually into one or more features representing the image content. In the current setup, we use images from OSM (I_{OSM}), Google Maps (I_{GM}) and Street View (I_{SV}), and NGI (I_{NGI}), i.e., the Belgian National Geographic Institute. From these images, we extracted the following eight features:

- texture (I_{SV}): the number of strong Canny edge pixels of the Google Street View image I_{SV} (which pitch is set to -90 to face down the "camera"). It takes large values for cobblestones and tiles. Since no Street View images exist for off-road locations, texture (I_{SV}) is left blank, facilitating road/off-road classification.
- streets (I_{OSM})/streets (I_{NGI}): the percentage of street-colored pixels in the OSM and NGI image. The street pixels are filtered out using the specific OSM and NGI street color ranges. It is higher for road types (e.g., asphalt and cobblestones) than for off-road types (such as grass).
- grass (I_{OSM})/grass (I_{GM})/grass (I_{NGI}): the percentage grass- or rural-colored pixels in the OSM, Google Maps, and NGI image. Grass pixels are filtered out using the specific OSM, Google Maps, and NGI rural color ranges. It takes large values for off-road terrain types. grass (I_{GM}) only takes large values for grass, and not for mud or gravel. The latter terrain types can be detected using the mud (I_{GM}) feature.
- mud (I_{GM}): the percentage of low-saturated "orange-red" hue-saturation-value pixels (~mud-colored pixels) in the Google Maps image IGM. It is large for mud and some types of gravel.
- urban(I_{OSM}): the percentage of "gray" pixels based on red-green-blue equality. It is larger for road terrain types like asphalt, cobblestones, and tiles.

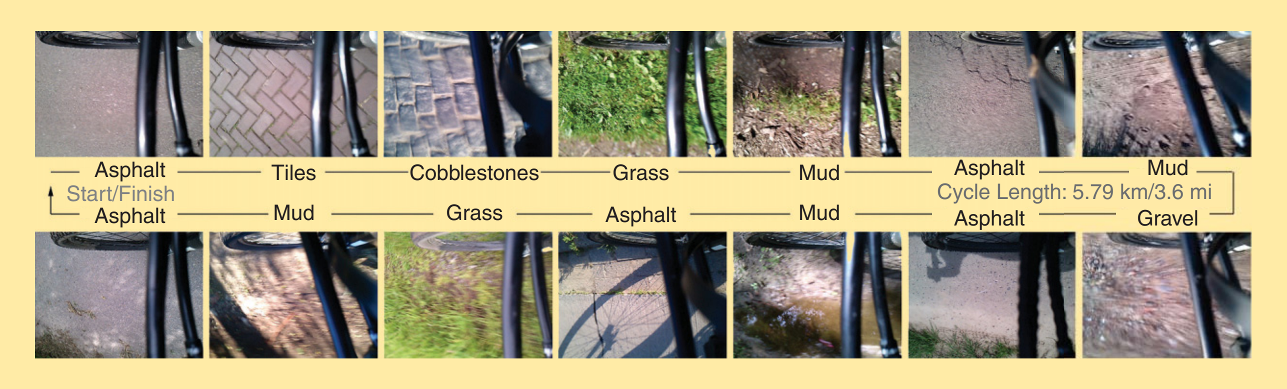
RDF CLASSIFICATION

RDF is a very fast tool for classification and clustering, which has shown to be extremely flexible in the context of computer vision. The most well-known application of RDF is the detection of human body parts in Microsoft's Kinect. The accuracy of RDF is comparable with other classifiers. Other advantages are its simple training and testing, and the fact that it can easily perform multiclass classifications. For a more general discussion on random forests, we refer the reader to [1] and [17].

Random forests are ensembles of randomized decision trees T_n, as illustrated in Figure 4. Each of the N_{tree} trees consists of split nodes and leaves that map the multimodal



[FIG4] The RDF for terrain classification [17].



[FIG5] An exemplary bike cycle (start to finish).

feature vector v to a distribution $P_i(c)$ stored at each leaf. The split nodes evaluate the arriving feature vector, and, depending on the feature values, pass it to the left or right child. Each leaf stores the statistics of the training vectors. For a classification task, it is the probability for each class c, denoted by $P(c|v)$:

$$P(c|v) = \sum_{n=1}^{N_{\text{tree}}} P_n(c|v). \tag{2}$$

EXPERIMENTAL SETUP AND EVALUATION RESULTS

To evaluate the proposed architecture, we have performed several bike tours. During these tours, we collected the training/test data and annotated them with the ground truth (GT). Based on this GT, we evaluated the test data while varying the number of trees (N_{tree}) and the sample ratio r (i.e., the percentage of randomized training vectors used in each tree construction). Furthermore, we have also launched a bike app to extensively test the proposed setup and collect more test data.

DATA COLLECTION

The data collection was performed using standard 26-in and 29-in mountain bikes. Multiple cycles with varying terrain conditions (in type and frequency) were performed in several rural and (sub)urban regions all over Belgium. An exemplary run, in which all six terrain types occurred, is shown in Figure 5. To have varying weather conditions, the cycle runs were spread over the year. Tire pressure and tire types were changed in between several runs to cope with the tire-vibration dependency.

To collect the vibration and GPS data, we used a Sony Ericsson Xperia mini-Android smartphone and a Garmin Edge 800 bike GPS. On the smartphone, we ran an accelerometer data logger and the time lapse Android app, which takes a picture every 5 s. These pictures are used for the GT creation. The bike GPS collected all geographical data and bike statistics. Based on the time stamps, which are stored for each sensor reading, the sensor data is aligned on each other. With our bike-sensing app, all future data will be captured using only the smartphone, increasing the usability of the overall setup.

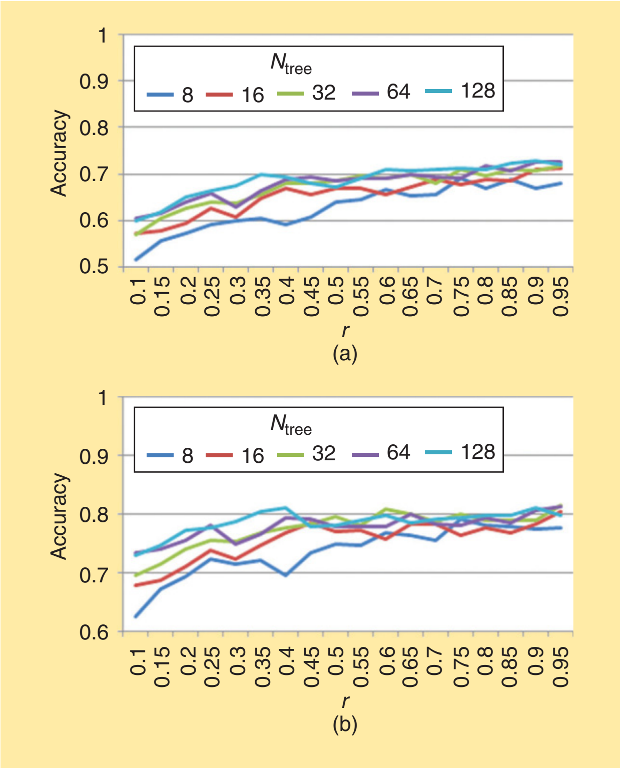
The current data set contains over 16,000 "real-life" terrain samples in total to train and test our RDF-based classification algorithm. Each terrain sample consists of the features of 5 s of accelerometer data and the corresponding Web-based geo-images/features. The distribution of the classes, which is retrieved using the GT creation (discussed in the section "GT Creation"), consists of 26% asphalt, 11% cobblestones, 9% tiles, 12% gravel, 19% grass, and 23% mud. To cope with the class imbalance in the data set, i.e., to have a more "unbiased" classifier, we follow the idea of cost-sensitive learning and use a weighted random forest [3], [25]. We assign a weight to each class, with the minority classes given larger weight, i.e., higher misclassification cost.

THE GT CREATION IS PERFORMED BY VISUAL ANALYSIS OF THE TERRAIN IMAGES USING A CUSTOM BUILT GT MARKING APPLICATION.

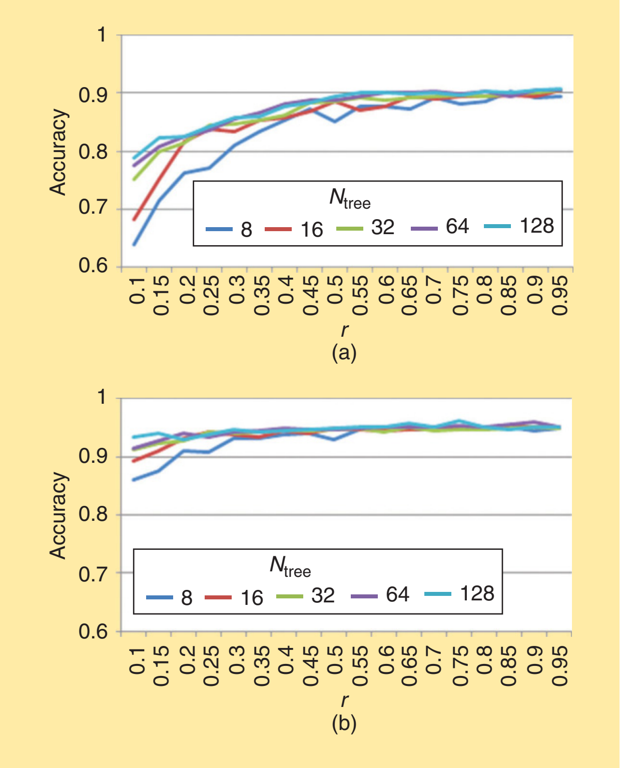
GT CREATION

The GT creation is performed by visual analysis of the terrain images using a custom built GT marking application. In addition to this image-based annotation, we also extend the GT with the available geographic terrain data of online maps. This data can be retrieved by reverse geocoding of the GPS Exchange Format latitude/longitude information of our GPS.

As can be seen in the cycle run in Figure 5, it is not always clear/easy to distinguish between the off-road types. Sometimes the terrain consists of a combination of multiple terrain types, e.g., grass and mud. In these situations, GT annotation



[FIG6] (a) The accuracy of six-class road classification solely based on accelerometer data. (b) The accuracy of two-class road/off-road classification solely based on accelerometer data.



[FIG7] (a) The accuracy of six-class road classification solely based on online geographic data. (b) The accuracy of two-class road/off-road classification solely based on online geodata.

is difficult and can be error prone. A similar kind of GT inaccuracy was also reported in [20]. To cope with this GT issue, we will extend the GT concept to allow multiannotation. Currently, one can also discard these misclassifications from the confusion matrices and other evaluation metrics, which are discussed hereafter.

EVALUATION STRATEGY/METRICS

First, it is important to mention that both six-class and two-class road/off-road classifications are evaluated. This facilitates comparison with SOTA works, which mainly perform two-class classification or do not always use the same set of terrain types. Furthermore, depending the application in which the classification system is used, the degree of specificity will also differ, i.e., for some GIS tools, a road/off-road discrimination is sufficient.

The accuracy of the proposed system is evaluated for an increasing number of RDF trees (N_{tree}) and increasing sample ratio r (which is related to the number of bootstrap samples). We define the accuracy as the proportion of the total number of predictions that were correct, i.e., the ratio of the number of correctly classified test vectors and the total number of test vectors. This accuracy will be calculated for each of the sensors individually, i.e., the accelerometer and geodata accuracy, and also for their multimodal combination. When they are combined, the highest class probability in $P(c|v)$ wins.

Like in the work of Khan et al. [12], the evaluation is performed using tenfold cross-validation. The data collected during our bike cycles is randomly divided into ten equally sized pieces. Each piece is used as the test set with training done on the remaining 90% of the data. The test results are then averaged over the ten cases, i.e., the accuracies that are reported are the average accuracy over ten RDF runs.

To allow a more detailed analysis, we also generated confusion matrices for the optimal RDF $N_{\text{tree}}-r$ combinations. The strength of a confusion matrix is that it identifies the nature of the classification errors, as well as their quantities.

RESULTS

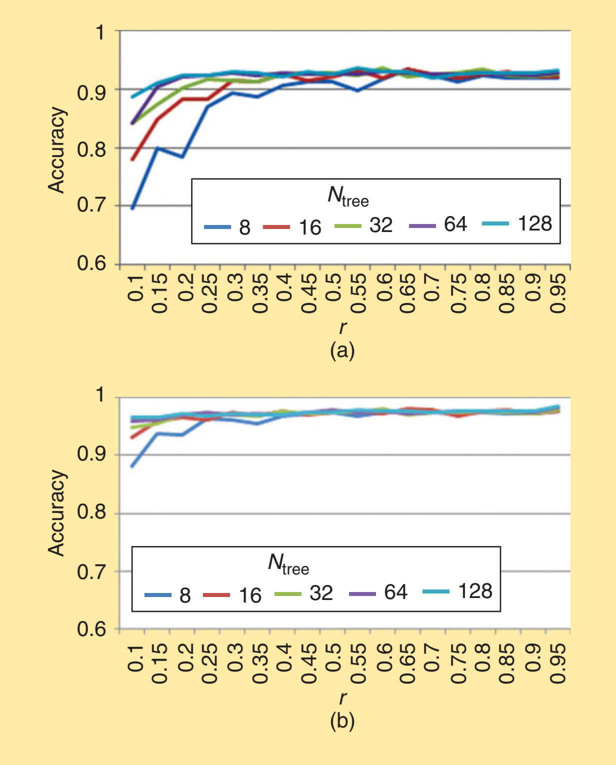
First, we will present the accuracy results for each of the sensors individually, i.e., the accelerometer and geodata accuracy. Subsequently, we will present their multimodal accuracy, based on a simple merging strategy. Figures 6–8 show the accuracy for increasing number of RDF trees (N_{tree}) and increasing sample ratio r. Both six-class and road/off-road two-class accuracy are shown.

ACCELEROMETER/VIBRATION RESULTS

Figure 6(a) shows the accuracy for the six-class terrain classification solely based on accelerometer data. For an optimal RDF configuration ($N_{\text{tree}} \backsimeq 32; \mathrm{r} \backsimeq 0.75$), an accuracy of 72% is achieved. For two-class road/off-road classification, the accuracy is 81%, as can be seen in Figure 6(b).

GEODATA RESULTS

Figure 7(a) shows the accuracy for the six-class terrain classification solely based on online geographic data. For an optimal RDF configuration ($N_{\text{tree}} \backsimeq 64; \mathrm{r} \backsimeq 0.60$), an accuracy of 90% is achieved. For two-class road/off-road classification, the "geo-only" accuracy is 95%, as can be seen in Figure 7(b). Due to the "big" online enrichment, the gain of multimodal analysis is not that high (<2%). However, such an extensive enrichment will not always be possible, e.g., due to a lack of geographic data.



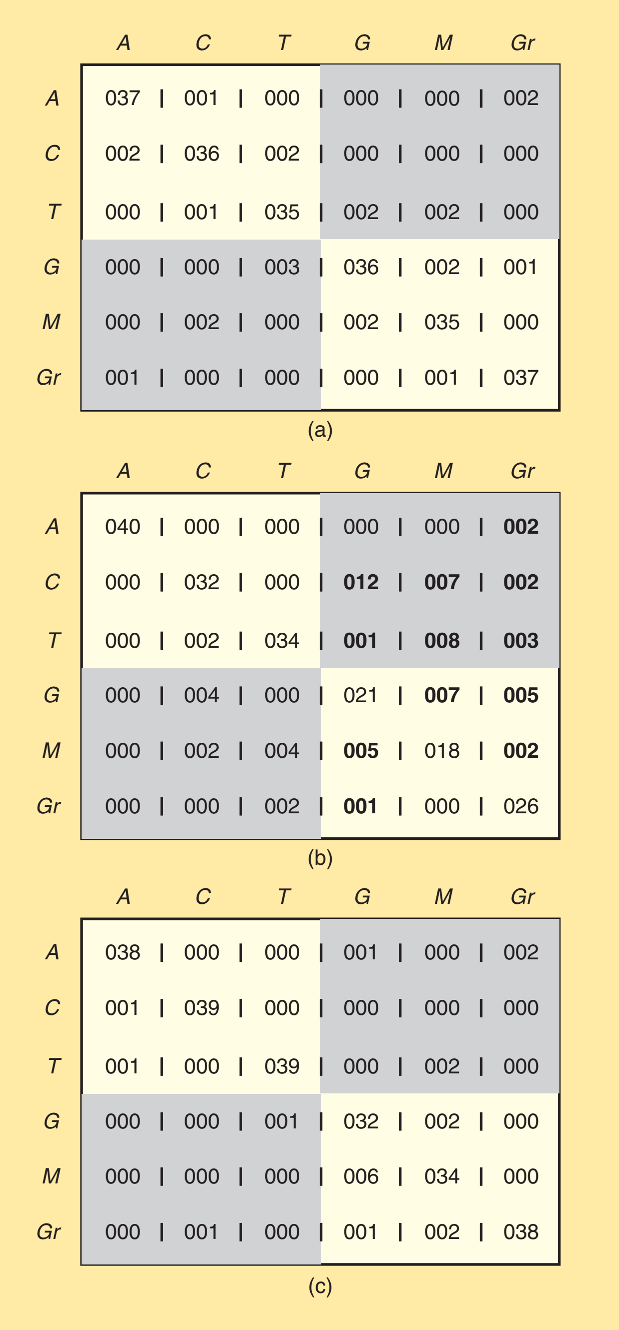
[FIG8] (a) The accuracy of six-class road classification based on multimodal data. (b) The accuracy of two-class road/off-road classification based on multimodal data.

COMBINED "MULTIMODAL" RESULTS

Figure 8(a) shows the accuracy for the six-class terrain classification based on both accelerometer and online geodata. For an optimal RDF configuration ($N_{\text{tree}} \backsimeq 32; \mathrm{r} \backsimeq 0.5$), an accuracy of 92% is achieved. For two-class road/off-road classification, the multimodal accuracy is 97% [see Figure 8(b)]. Both results show that our system outperforms the SOTA work in this domain (see the section "Related Work").

CONFUSION MATRICES

Figure 9 shows the accelerometer, geodata, and multimodal confusion matrices for their optimal RDF $N_{\text{tree}}-r$ combinations on a test set of 240 terrain segments (with 40 segments for each class). These matrices contain information about the actual (~GT) and predicted classifications done by our RDF-based classification system and report the number of true/false positives and true/false negatives. As the geodata confusion matrix in Figure 9(a) shows, each of the terrain types was classified correctly to a high degree. Only a limited number of misclassifications occurred. For the



	A	C	T	G	M	Gr
A	037	001	000	000	000	002
C	002	036	002	000	000	000
T	000	001	035	002	002	000
G	000	000	003	036	002	001
M	000	002	000	002	035	000
Gr	001	000	000	000	001	037

(a)

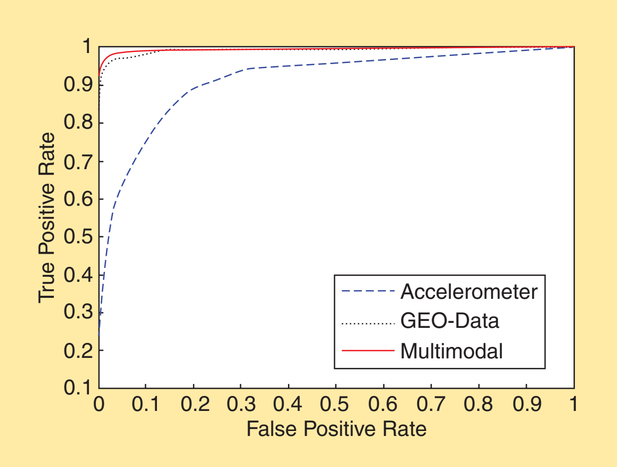
	A	C	T	G	M	Gr
A	040	000	000	000	000	**002**
C	000	032	000	**012**	**007**	**002**
T	000	002	034	**001**	**008**	**003**
G	000	004	000	021	**007**	**005**
M	000	002	004	**005**	018	**002**
Gr	000	000	002	**001**	000	026

(b)

	A	C	T	G	M	Gr
A	038	000	000	001	000	002
C	001	039	000	000	000	000
T	001	000	039	000	002	000
G	000	000	001	032	002	000
M	000	000	000	006	034	000
Gr	000	001	000	001	002	038

(c)

[FIG9] The confusion matrices for geodata, accelerometer, and multimodal classification on a test set of 240 terrain segments (with 40 segments for each class). (a) A geodata confusion matrix. (b) An accelerometer confusion matrix. (c) A multimodal confusion matrix.



[FIG10] The ROC curve for the RDF-based terrain classification using tenfold cross-validation. The high prediction accuracy of our classification method is shown by the AUC, which is 0.9136 for accelerometer only, 0.985 for geodata only, and 0.994 for the multimodal approach.

accelerometer classification, most misdetections occur on off-road terrain types [as shown in bold in Figure 9(b)].

The confusion matrix for the multimodal result is shown in Figure 9(c). As already mentioned in the section "Geodata Results," the accuracy gain of multimodal analysis compared to the geodata result is not that high (only 2%) due to the extensive enrichment in our test setup. This is also reflected by the multimodal confusion matrix, which holds similar results as the confusion matrix of the geodata. However, as can be seen by comparing Figure 9(a) and (c), some improvements are achieved in the road/off-road classification–road/off-road misclassifications and are shown in Figure 9 with a gray background.

RECEIVER OPERATOR CHARACTERISTIC/ AREA UNDER THE CURVE EVALUATION

The performance of the RDF-based classifier in predicting the terrain type is also assessed by plotting the receiver operator characteristic (ROC) curve for the tenfold cross validation results on the evaluation set. To generate the multiclass ROC operating points we treat the multiclass problem as a "one versus all" binary classification problem and calculate the operating points for each class and then average it out for the entire classifier [4], [6]. The ROC curve, shown in Figure 10, shows the tradeoff between prediction sensitivity and specificity. The RDF cutoff is the parameter that is varied along the curve. The area under the curve (AUC) is near the maximum and demonstrates the high prediction accuracy of our multimodal terrain classification algorithm.

CONCLUSIONS

This article focuses on the automatic geoannotation of road/terrain types by collaborative bike sensing and presents the detailed design, implementation, and evaluation of a novel road/terrain classification system. The proposed system shows how mobile sensors can help to automate and facilitate some of the more labor-intensive VGI tasks. Based on the analysis of volunteered geographic information gathered by cyclists, enriched with Web-based geographic data, geographic maps can be annotated automatically with each of the six terrain types. A geographic map can be annotated automatically using this road/terrain information or advanced route statistics can be generated. In the future, it should even be possible to use the collected data to perform terrain-based routing.

It is worth pointing out that the proposed techniques can also be extended to other sensing scenarios. First of all, the concept of combining mobile sensing with online (geographic) enrichment can lead to improvements in many domains. The proposed contributions are not limited to geographic map enrichment, but can easily be adapted to other applications, such as transportation analysis, health/activity monitoring, and robot navigation. Our article shows how computationally low-cost features that are collected by a sensing device can be extended/filtered/improved at large scale using data on the Web. Furthermore, it is shown how an RDF-based classifier can be used to perform multimodal classification tasks at low training and testing time. Finally, it is important to mention that the set of mobile and online terrain features can also be used for other big data classification tasks, such as scene classification and image clustering.

AUTHORS

Steven Verstockt (steven.verstockt@ugent.be) received his master's degree in informatics from Ghent University (Belgium) in 2003. In 2008, he began working on his Ph.D. degree on video fire analysis at the Multimedia Lab of the Electronics and Information Systems Department of Ghent University–iMinds, where he has worked since 2012 as a postdoctoral researcher. His current research focuses on multimodal signal processing and geographic information systems.

Viktor Slavkovikj (viktor.slavkovikj@ugent.be) was awarded a joint Erasmus Mundus master's degree in computer science from University Jean Monnet (France) in 2011. At the end of 2011, he enrolled as a Ph.D. candidate, working on multimodal sensor analysis for automatic object detection and tracking, at the Multimedia Lab of Ghent University–iMinds (Belgium). His research interests include computer vision and machine learning.

Pieterjan De Potter (pieterjan.depotter@ugent.be) received the M.S. degree in engineering from Ghent University (Belgium), in 2008. Between 2008 and 2010, he was involved in the Interdisciplinary Institute for Broadband Technology project Share4Health. He is currently pursuing the Ph.D. degree at the Multimedia Lab, Ghent University–iMinds (Belgium). His research interests include video analytics and semantic Web technologies.

Rik Van de Walle (rik.vandewalle@ugent.be) received his M.S. and Ph.D. degrees in engineering from Ghent University (Belgium) in 1994 and 1998, respectively. After a visiting scholarship at the University of Arizona (United States), he returned to Ghent University, where he became a professor of multimedia systems and applications, and head of the Multimedia Lab. His current research interests include multimedia content delivery, presentation and archiving, coding and description of multimedia data, content adaptation, and interactive (mobile) multimedia applications.

REFERENCES

[1] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[2] A. Bujari, B. Licar, and E. C. Palazzi, "Movement pattern recognition through smartphone's accelerometer," in *Proc. Consumer Electronics and Networking Conf. (CCNC)*, 2012, pp. 502–506.

[3] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Statistics Tech. Rep. 666, Univ. California Berkeley Library, 2004, pp. 1–12.

[4] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[5] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, 2008.

[6] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.

[7] R. Hariharan, J. Krumm, and E. Horvitz, "Web enhanced GPS," in *Proc. LNCS Int. Workshop Location and Context-Awareness*, 2005, pp. 95–104.

[8] M. Hentschel and B. Wagner, "Autonomous robot navigation based on OpenStreetMap geodata," in *Proc. 13th Int. IEEE Conf. Intelligent Transportation Systems*, 2010, pp. 1645–1650.

[9] N. Kaklanis, K. Votis, and D. Tzovaras, "Touching OpenStreetMap data in mobile context for the visually impaired," in *Proc. 3rd Workshop on Mobile Accessibility—ACM SIGCHI Conf. Human Factors in Computing Systems*, 2013, pp. 1–4.

[10] R. Karam and M. Melchiori, "Improving geo-spatial linked data with the wisdom of the crowds," in *Proc. Joint 16th Int. Conf. Extending Database Technology/16th Int. Conf. Database Theory (EDBT/ICDT) 2013 Workshops*, pp. 68–74.

[11] Y. N. Khan, P. Komma, K. Bohlmann, and A. Zell, "Grid-based visual terrain classification for outdoor robots using local features," in *Proc. IEEE Symp. Computational Intelligence in Vehicles & Transportation Systems*, 2011, pp. 16–22.

[12] Y. N. Khan, A. Masselli, and A. Zell, "Visual terrain classification by flying robots," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2012, pp. 498–503.

[13] M. Pannevis and M. Marx, "Using Web-sources for location based systems on mobile phones," in *Proc. Workshop Mobile Information Retrieval* (*MobIR'08*), 2008, pp. 1–8.

[14] D. Popescu, R. Dobrescu, and D. Merezeanu, "Road analysis based on texture similarity evaluation," in *Proc. 7th WSEAS Int. Conf. Signal Processing*, 2008, pp. 47–51.

[15] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava, "Biketastic: Sensing and mapping for better biking," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2010, pp. 1817–1820.

[16] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *Trans. Sens. Netw.*, vol. 6, no. 2, pp. 1–27, 2010.

[17] J. Shotton, T.-K. Kim, and B. Stenger, "Boosting & randomized forests for visual recognition (tutorial)," in *Proc. Int. Conf. Computer Vision*, 2009. [Online]. Available: http://www.iis.ee.ic.ac.uk/icvl/iccv09_tutorial.html

[18] H. Smith III, "Improving the quality of terrain measurement," master's thesis, Dept. Mechanical Engineering, Virginia Polytechnic Institute and State Univ., Blacksburg, VA, 2009.

[19] M. Srivastava, T. Abdelzaher, and B. Szymanski, "Human-centric sensing," in *Philos. Trans. R. Soc.*, vol. 370, no. 1958, 2012, pp. 176–197.

[20] G. Strazdins, A. Mednis, R. Zviedris, G. Kanonirs, and L. Selavo, "Virtual ground truth in vehicular sensing experiments: How to mark it accurately," in *Proc. 5th Int. Conf. Sensor Technologies and Applications* (*SENSORCOMM 2011*), pp. 295–300.

[21] I. Tang and T. P. Breckon, "Automatic road environment classification," *IEEE Trans. Intell. Transport. Syst.*, vol. 12, no. 2, pp. 476-484, 2011.

[22] S. Verstockt, V. Slavkovikj, P. De Potter, J. Slowack, and R. Van de Walle, "Multimodal bike sensing for automatic geo-annotation: Geo-annotation of road/terrain type by participatory bike-sensing," in *Proc. 10th Int. Conf. Signal Processing and Multimedia Applications* (*SIGMAP'13*), 2013, pp. 39–49.

[23] S. Wang, S. Kodagoda, Z. Wang, and G. Dissanayake, "Multiple sensor based terrain classification," in *Proc. Australasian Conf. Robotics and Automation (ACRA)*, 2011, pp. 1–7.

[24] C. C. Ward and K. Iagnemma, "Speed-independent vibration-based terrain classification for passenger vehicles," *Veh. Syst. Dyn.*, vol. 47, no. 9, pp. 1095–1113, 2009.

[25] C. Weiss, H. Frohlich, and A. Zell, "Vibration-based terrain classification using support vector machines," in *Proc. 2006 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 4429–4434.

[26] C. Weiss, H. Tamimi, and A. Zell, "A combination of vision- and vibration-based terrain classification," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS),* 2008, pp. 2204–2209.

[27] G. M. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?," in *Proc. Int. Conf. Data Mining (DMIN)*, 2007, pp. 35–41.

[28] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Proc. 17th Conf. on Innovative Applications of Artificial Intelligence*, 2005, pp. 1541–1546.

[29] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proc. Pervasive Computing*, vol. 3001, 2004, pp. 1–17.

[SP]

Optimization and Estimation of Complex-Valued Signals

[Theory and applications in filtering and blind source separation]

[Tülay Adalı and Peter J. Schreier]

Digital Object Identifier 10.1109/MSP.2013.2287951
Date of publication: 19 August 2014

Complex-valued signals occur in many areas of science and engineering and are thus of fundamental interest. When developing signal processing methods in the complex domain, there are two key issues: making use of the full statistical information and optimization. In this article, we review the necessary tools to address these two key issues and provide examples in filtering and blind source separation (BSS) that utilize these tools.

INTRODUCTION

Complex-valued random signals are essential to a great number of applied research areas, such as communications, radar, sonar, geophysics, oceanography, optics, and electromagnetics. A common assumption when dealing with complex random signals is that they are proper or circular. Most often, this is not explicitly stated but implied by ignoring some aspect of the statistics of a complex signal. A proper complex random variable is uncorrelated with its complex conjugate, and a circular complex random variable has a probability distribution that is rotationally invariant in the complex plane. These assumptions are mathematically convenient because they simplify many computations. Often, they can indeed be justified. However, there are also many situations where proper and circular signals are very poor models of the underlying physics. While this has been known and appreciated by oceanographers since the early 1970s [1], and pioneering works in signal processing date back

1053-5888/14©2014IEEE

to the 1990s [2]–[5], it is only more recently that the signal processing community has started showing an increasing interest in this topic.

Another important issue in the processing of complex-valued signals is related to optimization. Since cost functions are real valued and hence nonanalytic, two approaches have been the common practice in optimization: derivatives are either evaluated with respect to the real and imaginary parts separately and then combined, or optimization is performed in an augmented space by transforming the problem from the complex domain to the real domain of double the dimension. The first approach leads to unnecessarily long expressions, and the second requires finding the appropriate transformation, which is not always straightforward, especially when dealing with nonlinear functions. Wirtinger calculus [6] addresses this issue by relaxing the definition of differentiability and defining a general framework that includes analytic functions as a special case. The development by the Austrian mathematician Wirtinger dates back to 1927. It was rediscovered in the engineering community, without reference to Wirtinger, by Brandwood in 1983 [7], and then further developed for gradient and Hessian formulations by van den Bos [8], [9] using an augmented representation that doubles the dimensionality. It is only recently that a larger fraction of the signal processing community has taken notice of the development. The biggest advantage of using Wirtinger derivatives is that the expressions are kept simple and similar to the real case, and many algorithms and analyses can be readily extended from the real to the complex domain. An additional advantage is that, since the expressions do not become unnecessarily complicated, many of the simplifying assumptions—of which circularity has been a common one—can be avoided.

When a signal is improper or noncircular, accounting for this fact can provide significant performance gains. For instance, in mobile multiuser communications, it can enable an improved tradeoff between spectral efficiency and power consumption. Important examples of digital modulation schemes that produce improper complex baseband signals are: binary phase-shift keying, pulse-amplitude modulation, Gaussian minimum shift keying, and offset quaternary phase-shift keying. A small sample of papers exploiting the impropriety of these signals is [10]–[15]. Improper baseband communication signals can also arise due to an imbalance between their in-phase and quadrature (I/Q) components. I/Q imbalance degrades the signal-to-noise ratio and thus the bit error rate performance. Some papers proposing ways of compensating I/Q imbalance in different types of communication systems include [16]–[19]. Techniques for wideband system identification when the system, e.g., a wideband wireless communication channel, is not rotationally invariant are presented in [20] and [21]. Widely linear beamformers have been considered by [22]–[25].

Data-driven methods for signal processing, particularly BSS, is another area where exploiting impropriety and noncircularity has led to important advances. An important technique for BSS is independent component analysis (ICA), where the multivariate data are decomposed into additive components that are as independent as possible. Under certain conditions, ICA can be achieved by exploiting impropriety [26], [27]. Significant performance gains can be obtained when algorithms explicitly take the noncircular nature of the data into account [28]–[33]. Among the many applications of ICA, medical data analysis and communications have been two of the most active. For example, in [33], noncircularity is exploited for feature extraction in electrocardiograms, and, in [34], noncircularity is shown to improve the estimation of neural activity when analyzing functional magnetic resonance imaging (fMRI) data.

To exploit the improper or noncircular nature of signals, we need to utilize the complete statistical characterization of complex-valued random signals. When restricted to second-order moments, this means that not only the (standard) correlation matters but also the complementary (or pseudo-) correlation, which is the correlation of a complex signal with its complex conjugate. To access the information contained in this correlation, we can employ linear-conjugate linear, or widely linear, transformations, or we can use cost functions that take the full second-order statistics into account with a linear structure [35]. Complementary correlations were first mentioned in the signal processing literature in 1969 by [36], and the optimum widely linear minimum mean-squared error (WLMMSE) filter was derived by [2] and [37]. When higher-order moments are considered, then there is a multitude of complementary correlations [38], [39]. For instance, in the fourth-order moment $E\{x_1^{(*)}x_2^{(*)}x_3^{(*)}x_4^{(*)}\}$, each term may or may not be conjugated. Higher-order statistics (HOS) are often utilized implicitly by optimizing a cost function such as entropy.

In this article, we review some of the basic results on statistical signal processing of complex-valued data and also some of the more recent applications in filtering and blind source separation.

OPTIMIZATION

In the derivation of signal processing algorithms, we often have to compute gradients and Hessians of cost functions, such as quadratic forms or likelihood functions. Cost functions are real-valued but often involve complex-valued parameters. Such functions are not analytic and hence not differentiable. To overcome this basic limitation, two possible approaches have traditionally been adopted in the signal processing literature. The most common approach is the evaluation of separate derivatives with respect to the real and imaginary parts of the nonanalytic function. Another approach has been to define "augmented" vectors by stacking the real and imaginary parts in a vector of twice the original dimension, and then to perform all the evaluations in the real domain. In the end, the solution is converted back to the complex domain. Needless to say, both approaches are cumbersome and might also require additional assumptions, such as circularity, to simplify the expressions.

The framework based on Wirtinger calculus [6], [40]—also called the $\mathbb{CR}$ calculus [41]—provides a simple and straightforward approach to calculating derivatives with respect to complex parameters, in particular for the important case of nonanalytic functions. Wirtinger calculus allows one to perform all the derivations and analysis in the complex domain. This can be done without considering the real and imaginary parts separately and without doubling the dimensionality, which was the approach taken by [9]. Hence, all of the computations can be carried out in a manner very similar to the real-valued case, making many tools and methods developed for the real case readily available for the complex case.

In this section, we introduce the main idea behind Wirtinger calculus for scalar, vector, and matrix optimization and give examples to demonstrate its application. By keeping the expressions and evaluations simple, a key advantage is that common assumptions in complex-valued signal processing—most notably circularity—can be avoided.

OPTIMIZATION: SCALAR CASE

We first consider a complex-valued function $f(z) = u(z_r, z_i) + jv(z_r, z_i)$, where $z = z_r + jz_i$. The classical definition of *complex differentiability* requires that the derivatives defined as the limit

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \tag{1}$$

be independent of the direction in which Δz approaches 0 in the complex plane. This requires that the Cauchy–Riemann equations [40], [42]

$$\frac{\partial u}{\partial z_r} = \frac{\partial v}{\partial z_i} \text{ and } \frac{\partial u}{\partial z_i} = -\frac{\partial v}{\partial z_r} \tag{2}$$

be satisfied. These conditions are necessary for $f(z)$ to be complex-differentiable. If the partial derivatives of $u(z_r, z_i)$ and $v(z_r, z_i)$ are continuous, then they are sufficient as well. A function that is complex-differentiable on its entire domain is called *holomorphic* or *analytic*. Obviously, since real-valued cost functions have $v(z_r, z_i) = 0$, the Cauchy–Riemann conditions do not hold, and hence cost functions are not analytic. The Cauchy–Riemann equations impose a rigid structure on $u(z_r, z_i)$ and $v(z_r, z_i)$ and thus $f(z)$. A simple demonstration of this fact is that either $u(z_r, z_i)$ or $v(z_r, z_i)$ alone suffices to determine the derivatives of an analytic function.

Wirtinger calculus provides a general framework for differentiating nonanalytic functions, and is general in the sense that it includes analytic functions as a special case. It only requires that $f(z)$ be differentiable when expressed as a function $f: \mathbb{R}^2 \to \mathbb{R}^2$. Such a function is called *real differentiable*. If $u(z_r, z_i)$ and $v(z_r, z_i)$ have continuous partial derivatives with respect to z_r and z_i, f is real differentiable. For such a function, we can write

$$\frac{\partial f}{\partial z} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial z_r} - j\frac{\partial f}{\partial z_i}\right) \text{ and } \frac{\partial f}{\partial z^*} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial z_r} + j\frac{\partial f}{\partial z_i}\right), \tag{3}$$

which can be easily derived by writing $z_r = (z + z^*)/2$ and $z_i = (z - z^*)/2j$ and then using the chain rule [43]. Instead of computing the derivatives with respect to z_r and z_i, the complex derivatives (3) can be evaluated by considering f to be a bivariate function $f(z, z^*)$ and treating z and z^* as independent variables. That is, when applying $\partial f/\partial z$, we take the derivative with respect to z, while formally treating z^* as a constant. Similarly, $\partial f/\partial z^*$ yields the derivative with respect to z^*, formally regarding z as a constant. Thus, there is no need to develop new differentiation rules. This was shown in [7] in 1983 without a specific reference to Wirtinger's earlier work [6]. Interestingly, many of the references that refer to [7] and use the generalized derivatives in (3) do evaluate them by computing derivatives with respect to z_r and z_i separately, rather than considering the function in the form $f(z, z^*)$ and directly taking the derivative with respect to z or z^*. This leads to unnecessarily complicated derivations.

When we consider the function in the form $f(z, z^*)$, the Cauchy–Riemann equations can simply be stated as $\partial f/\partial z^* = 0$. In other words, an analytic function cannot depend on z^*. If f is analytic, then the usual complex derivative in (1) and $\partial f/\partial z$ in (3) coincide. Hence, Wirtinger calculus contains standard complex calculus as a special case.

For real-valued $f(z)$, we have $(\partial f/\partial z)^* = \partial f/\partial z^*$, i.e., the derivative and the conjugate derivative are complex conjugates of each other. Because they are related through conjugation, we only need to compute one or the other. As a consequence, a necessary and sufficient condition for real-valued f to have a stationary point is $\partial f/\partial z = 0$. An equivalent necessary and sufficient condition is $\partial f/\partial z^* = 0$ [7].

EXAMPLE

Consider the real-valued function $f(z) = |z|^4 = z_r^4 + 2z_r^2 z_i^2 + z_i^4$. We can evaluate $\partial f/\partial z$ by differentiating separately with respect to z_r and z_i,

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial z_r} - j\frac{\partial f}{\partial z_i}\right) = 2z_r^3 + 2z_r z_i^2 - 2j(z_r^2 z_i + z_i^3), \tag{4}$$

or we can write the function as $f(z) = f(z, z^*) = z^2(z^*)^2$ and differentiate by treating z^* as a constant,

$$\frac{\partial f}{\partial z} = 2z(z^*)^2. \tag{5}$$

The second approach is clearly simpler. It can be easily shown that the two expressions, (4) and (5), are equal. However, while the expression in (4) can easily be derived from (5), it is not quite as straightforward the other way around. Because $f(z)$ is real valued, there is no need to compute $\partial f/\partial z^*$: it is simply the conjugate of $\partial f/\partial z$. ■

Series expansions are a valuable tool in the study of nonlinear functions. For analytic, i.e., complex-differentiable, functions, the Taylor series expansion assumes the same form as in the real case

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k, \tag{6}$$

where $f^{(k)}(z_0)$ denotes the kth-order derivative of f evaluated at z_0. If $f(z)$ is analytic for $|z| \le R$, then the Taylor series given in (6) converges uniformly in $|z| \le R_1 < R$.

As in the case of Taylor series expansions, the desire to have the complex domain representation follow the real-valued case closely has also been the main motivation for defining differentiability in the complex domain using (1). However, the class of functions that admit such a representation is limited, excluding the important group of cost functions. For functions that are real differentiable, Wirtinger calculus can be employed to write the Taylor series of a nonanalytic function as an expansion in z and z^*. We discuss this approach in more detail in the next section, when we introduce vector optimization using Wirtinger calculus. This simple but useful idea for Taylor series expansions of real differentiable functions has been introduced in [38] and formalized in [44] using the duality between $\mathbb{R}^{2N}$ and $\mathbb{C}^N$.

OPTIMIZATION: VECTOR CASE

SECOND-ORDER EXPANSIONS

In the development and study of adaptive signal processing algorithms, i.e., in iterative optimization of a selected cost function and in performance analysis, the first- and second-order expansions prove to be most useful. For an analytic function $f(\mathbf{z}): \mathbb{C}^N \mapsto \mathbb{C}$, we define $\Delta f = f(\mathbf{z}) - f(\mathbf{z}_0)$ and $\Delta \mathbf{z} = \mathbf{z} - \mathbf{z}_0$ to write the second-order approximation to the function in the neighborhood of $\mathbf{z}_0$ as

$$\begin{aligned}\Delta f &\approx \Delta\mathbf{z}^T \nabla_{\mathbf{z}} f + \frac{1}{2}\Delta\mathbf{z}^T \mathbf{H}(\mathbf{z})\Delta\mathbf{z} \\ &= \langle \nabla_{\mathbf{z}} f, \Delta\mathbf{z}^* \rangle + \frac{1}{2}\langle \mathbf{H}(\mathbf{z})\Delta\mathbf{z}, \Delta\mathbf{z}^* \rangle, \qquad (7)\end{aligned}$$

where

$$\nabla_{\mathbf{z}} f = \left.\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}}\right|_{\mathbf{z}_0}$$

is the gradient evaluated at $\mathbf{z}_0$, and

$$\nabla_{\mathbf{z}}^2 f \triangleq \mathbf{H}(\mathbf{z}) = \left.\frac{\partial^2 f(\mathbf{z})}{\partial \mathbf{z}\,\partial \mathbf{z}^T}\right|_{\mathbf{z}_0}$$

is the Hessian matrix evaluated at $\mathbf{z}_0$. As in the real-valued case, the Hessian matrix is symmetric, and it is constant if the function is quadratic.

On the other hand for a cost function, $f(\mathbf{z}): \mathbb{C}^N \mapsto \mathbb{R}$, which is nonanalytic, we can use Wirtinger calculus to expand $f(\mathbf{z})$ in two variables $\mathbf{z}$ and $\mathbf{z}^*$, which are treated as independent

$$\begin{aligned}\Delta f(\mathbf{z}, \mathbf{z}^*) &\approx \langle \nabla_{\mathbf{z}} f, \Delta\mathbf{z}^* \rangle + \langle \nabla_{\mathbf{z}^*} f, \Delta\mathbf{z} \rangle + \frac{1}{2}\left\langle \frac{\partial f^2}{\partial\mathbf{z}\partial\mathbf{z}^T}\Delta\mathbf{z}, \Delta\mathbf{z}^* \right\rangle \\ &+ \left\langle \frac{\partial f^2}{\partial\mathbf{z}\partial\mathbf{z}^H}\Delta\mathbf{z}^*, \Delta\mathbf{z}^* \right\rangle + \frac{1}{2}\left\langle \frac{\partial f^2}{\partial\mathbf{z}^*\partial\mathbf{z}^H}\Delta\mathbf{z}^*, \Delta\mathbf{z} \right\rangle. \qquad (8)\end{aligned}$$

Thus, the series expansion has the same form as one for a real-valued function of two variables, except that these are replaced by $\mathbf{z}$ and $\mathbf{z}^*$. Note that when $f(\mathbf{z}, \mathbf{z}^*)$ is real valued, we have

$$\langle \nabla_{\mathbf{z}} f, \Delta\mathbf{z}^* \rangle + \langle \nabla_{\mathbf{z}^*} f, \Delta\mathbf{z} \rangle = 2\,\mathrm{Re}\{\langle \nabla_{\mathbf{z}^*} f, \Delta\mathbf{z} \rangle\}, \qquad (9)$$

since, in this case, $\nabla f_{\mathbf{z}^*} = (\nabla f_{\mathbf{z}})^*$. Using the Cauchy–Bunyakovskii–Schwarz inequality [45], we have

$$|\Delta\mathbf{z}^H \nabla_{\mathbf{z}^*} f| \le \|\Delta\mathbf{z}\| \|\nabla_{\mathbf{z}^*} f\|,$$

which holds with equality when $\Delta\mathbf{z}$ is in the same direction as $\nabla_{\mathbf{z}^*} f$. Thus, for maximum change in the function value, one needs to calculate and use the gradient with respect to the complex conjugate of the variable, i.e., $\nabla f(\mathbf{z}^*)$.

It is also important to note that when $f(\mathbf{z}, \mathbf{z}^*) = f(\mathbf{z})$, i.e., the function is analytic (complex differentiable), all derivatives with respect to $\mathbf{z}^*$ in (8) vanish, and (8) thus coincides with (7). As noted earlier, the Wirtinger framework includes analytic functions, and when the function is analytic, all the expressions reduce to those for analytic functions.

COMPLEX GRADIENT UPDATES

To derive the expressions for gradient descent and Newton updates in the complex domain, we construct three closely related vectors from two real vectors $\mathbf{w}_r \in \mathbb{R}^N$ and $\mathbf{w}_i \in \mathbb{R}^N$. The first one is the complex vector $\mathbf{w} = \mathbf{w}_r + j\mathbf{w}_i \in \mathbb{C}^N$, and the second is the real composite $2N$-dimensional vector $\mathbf{w}_{\mathbb{R}} = [\mathbf{w}_r^T, \mathbf{w}_i^T]^T \in \mathbb{R}^{2N}$, obtained by stacking $\mathbf{w}_r$ on top of $\mathbf{w}_i$. Finally, the third one is the complex augmented vector $\underline{\mathbf{w}} = [\mathbf{w}^T, \mathbf{w}^H]^T \in \mathbb{C}^{2N}$, obtained by stacking $\mathbf{w}$ on top of its complex conjugate $\mathbf{w}^*$. Augmented vectors are always underlined. The complex augmented vector $\underline{\mathbf{w}}$ is related to the real composite vector $\mathbf{w}_{\mathbb{R}} \in \mathbb{R}^{2N}$ through $\underline{\mathbf{w}} = \mathbf{U}_N \mathbf{w}_{\mathbb{R}}$ and $\mathbf{w}_{\mathbb{R}} = (1/2)\mathbf{U}_N^H \underline{\mathbf{w}}$, where the real-to-complex transformation

$$\mathbf{U}_N = \begin{bmatrix} \mathbf{I} & j\mathbf{I} \\ \mathbf{I} & -j\mathbf{I} \end{bmatrix} \in \mathbb{C}^{2N \times 2N} \qquad (10)$$

is unitary up to a factor of 2, i.e., $\mathbf{U}_N \mathbf{U}_N^H = \mathbf{U}_N^H \mathbf{U}_N = 2\mathbf{I}$. The complex augmented vector $\underline{\mathbf{w}}$ is obviously an equivalent redundant, but convenient, representation of $\mathbf{w}_{\mathbb{R}}$. Consider a function $f(\mathbf{w})$: $\mathbb{C}^N \mapsto \mathbb{R}$ that is real differentiable up to second order. If we write the function as $f(\mathbf{w}_{\mathbb{R}}): \mathbb{R}^{2N} \mapsto \mathbb{R}$ using the augmented vector definition given above, we can easily establish the following two relationships [40], [46]:

$$\frac{\partial f}{\partial \mathbf{w}_{\mathbb{R}}} = \mathbf{U}_N^H \frac{\partial f}{\partial \underline{\mathbf{w}}^*} \qquad (11)$$

$$\frac{\partial^2 f}{\partial \mathbf{w}_{\mathbb{R}} \partial \mathbf{w}_{\mathbb{R}}^T} = \mathbf{U}_N^H \frac{\partial^2 f}{\partial \underline{\mathbf{w}}^* \partial \underline{\mathbf{w}}^T} \mathbf{U}_N. \qquad (12)$$

We can use these relationships to derive the expressions for gradient descent and Newton updates for iterative optimization in the complex domain.

From the real gradient update rule $\Delta\mathbf{w}_{\mathbb{R}} = -\mu(\partial f / \partial \mathbf{w}_{\mathbb{R}})$, we obtain the complex update relationship

$$\Delta\underline{\mathbf{w}} = \mathbf{U}_N \Delta\mathbf{w}_{\mathbb{R}} = -\mu \mathbf{U}_N \frac{\partial f}{\partial \mathbf{w}_{\mathbb{R}}} = -2\mu \frac{\partial f}{\partial \underline{\mathbf{w}}}.$$

The dimension of the update equation can be further reduced as

$$\begin{bmatrix}\Delta\mathbf{w}\\ \Delta\mathbf{w}^*\end{bmatrix} = -2\mu\begin{bmatrix}\dfrac{\partial f}{\partial\mathbf{w}^*}\\ \dfrac{\partial f}{\partial\mathbf{w}}\end{bmatrix} \Rightarrow \Delta\mathbf{w} = -2\mu\frac{\partial f}{\partial\mathbf{w}^*}.$$

Again, we note that the gradient with respect to the conjugate of the parameter gives the direction for maximal first-order change, derived here using the representation equivalent to the real-valued case in $\mathbb{R}^{2N}$.

COMPLEX NEWTON UPDATES

The Newton update in $\mathbb{R}^{2N}$ given by

$$\frac{\partial^2 f}{\partial\mathbf{w}_{\mathbb{R}}\partial\mathbf{w}_{\mathbb{R}}^T}\Delta\mathbf{w}_{\mathbb{R}} = -\frac{\partial f}{\partial\mathbf{w}_{\mathbb{R}}} \tag{13}$$

can be shown to be equivalent to

$$\Delta\mathbf{w} = -(\mathbf{H}_2^* - \mathbf{H}_1^*\mathbf{H}_2^{-1}\mathbf{H}_1)^{-1}\left(\frac{\partial f}{\partial\mathbf{w}^*} - \mathbf{H}_1^*\mathbf{H}_2^{-1}\frac{\partial f}{\partial\mathbf{w}}\right) \tag{14}$$

in $\mathbb{C}^N$, where

$$\mathbf{H}_1 \triangleq \frac{\partial^2 f}{\partial\mathbf{w}\partial\mathbf{w}^T} \text{ and } \mathbf{H}_2 \triangleq \frac{\partial^2 f}{\partial\mathbf{w}\partial\mathbf{w}^H}. \tag{15}$$

To establish this relationship, we can use (11) and (12) to express the real-domain Newton update in (13) as

$$\frac{\partial^2 f}{\partial\underline{\mathbf{w}}^*\partial\underline{\mathbf{w}}^T}\Delta\underline{\mathbf{w}} = -\frac{\partial f}{\partial\underline{\mathbf{w}}^*},$$

which can then be rewritten as

$$\begin{bmatrix}\mathbf{H}_2^* & \mathbf{H}_1^*\\ \mathbf{H}_1 & \mathbf{H}_2\end{bmatrix}\begin{bmatrix}\Delta\mathbf{w}\\ \Delta\mathbf{w}^*\end{bmatrix} = -\begin{bmatrix}\dfrac{\partial f}{\partial\mathbf{w}^*}\\ \dfrac{\partial f}{\partial\mathbf{w}}\end{bmatrix},$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are defined in (15). We can use the formula for the inverse of a partitioned positive definite matrix [47, p. 472], provided that the nonnegative definite matrix $(\partial^2 f/\partial\underline{\mathbf{w}}^*\partial\underline{\mathbf{w}}^T)$ is full rank, to write

$$\begin{bmatrix}\Delta\mathbf{w}\\ \Delta\mathbf{w}^*\end{bmatrix} = -\begin{bmatrix}\mathbf{T}^{-1} & -\mathbf{H}_2^{-*}\mathbf{H}_1^*\mathbf{T}^{-*}\\ -\mathbf{T}^{-*}\mathbf{H}_1\mathbf{H}_2^{-*} & \mathbf{T}^{-*}\end{bmatrix}\begin{bmatrix}\dfrac{\partial f}{\partial\mathbf{w}^*}\\ \dfrac{\partial f}{\partial\mathbf{w}}\end{bmatrix}, \tag{16}$$

where $\mathbf{T} \triangleq \mathbf{H}_2^* - \mathbf{H}_1^*\mathbf{H}_2^{-1}\mathbf{H}_1$ and $(\bullet)^{-*}$ denotes $[(\bullet)^*]^{-1}$. Since $(\partial^2 f)/(\partial\underline{\mathbf{w}}^*\partial\underline{\mathbf{w}}^T)$ is Hermitian, we finally obtain the complex Newton update given in (14). The expression for $\Delta\mathbf{w}^*$ is the conjugate of (14).

In [48], it was shown that the Newton algorithm for N complex variables cannot be written in a form similar to the real-valued case. However, as established here, it can be written as in (16) using the augmented form, which is equivalent to the Newton method in $\mathbb{R}^{2N}$. In $\mathbb{C}^N$, it can be expressed as in (14).

EXAMPLE

A linear filter approximates the desired sequence $x(n)$ through a linear combination of a window of input samples $y(n)$ such that the estimate of the desired sequence is

$$\hat{x}(n) = \mathbf{w}^H\mathbf{y}(n),$$

where the input vector at time n is written as $\mathbf{y}(n) = [y(n)y(n-1)\cdots y(n-N+1)]^T$, and the filter weights are $\mathbf{w} = [w_0 w_1 \cdots w_{N-1}]^T$. Hence, the random vector $\mathbf{y}(n)$ is formed by the current (at time n) and last $n-1$ samples of the discrete time random sequence $y(n)$. The minimum mean-square error (MSE) filter is designed such that the error

$$J_{\mathrm{mse}}(\mathbf{w}) = E\{|e(n)|^2\} = E\{|x(n) - \hat{x}(n)|^2\}$$

is minimized. To evaluate the weights $\mathbf{w}_{\mathrm{opt}}$ given by

$$\mathbf{w}_{\mathrm{opt}} = \arg\min_{\mathbf{w}} J_{\mathrm{mse}}(\mathbf{w}),$$

we can directly take the derivative of the MSE with respect to $\mathbf{w}^*$—by treating the variable $\mathbf{w}$ as a constant—such that

$$\begin{aligned}\frac{\partial E\{e(n)e^*(n)\}}{\partial\mathbf{w}^*} &= \frac{\partial E\{[x(n) - \mathbf{w}^H\mathbf{y}(n)][x^*(n) - \mathbf{w}^T\mathbf{y}^*(n)]\}}{\partial\mathbf{w}^*}\\ &= -E\{\mathbf{y}(n)[x^*(n) - \mathbf{w}^T\mathbf{y}^*(n)]\}\end{aligned} \tag{17}$$

and obtain the complex Wiener–Hopf equation

$$E\{\mathbf{y}(n)\mathbf{y}^H(n)\}\mathbf{w}_{\mathrm{opt}} = E\{x^*(n)\mathbf{y}(n)\}$$

by setting (17) to zero. We assume that the input is a zero-mean wide-sense-stationary (WSS) process and that the desired sequence and input are jointly WSS. We then define the input covariance matrix $\mathbf{C}_{yy} = E\{\mathbf{y}(n)\mathbf{y}^H(n)\}$ and the cross-covariance vector $\mathbf{c}_{xy} = E\{x^*(n)\mathbf{y}(n)\}$, to write

$$\mathbf{w}_{\mathrm{opt}} = \mathbf{C}_{yy}^{-1}\mathbf{c}_{xy} \tag{18}$$

when the input is persistently exciting, i.e., the covariance matrix is nonsingular. ■

Another example would be the derivation of the backpropagation update rule for training a multilayer perceptron (MLP), which involves the optimization of a nonlinear function. As demonstrated in [40], Wirtinger calculus simplifies the derivation considerably by allowing the use of simple tools such as the chain rule for nonanalytic functions, significantly shortening the derivation compared with the derivations given in, e.g., [49]–[51].

MATRIX CASE

Wirtinger calculus extends straightforwardly to functions $f{:}\mathbb{C}^N \to \mathbb{C}^M$ or $f{:}\mathbb{C}^{N\times M} \to \mathbb{C}$. There is no need to develop new differentiation rules for Wirtinger derivatives. All rules for taking derivatives for real functions remain valid. However, care must be taken to properly distinguish between the variables with respect to which differentiation is performed and those that are formally regarded as constants. So all the expressions from the real-valued case given, e.g., in [52], can be straightforwardly applied to the complex case. For instance, for $g(\mathbf{Z},\mathbf{Z}^*) = \mathrm{Trace}(\mathbf{Z}\mathbf{Z}^H)$, we obtain

$$\frac{\partial g}{\partial \mathbf{Z}} = \frac{\partial \mathrm{Trace}\,(\mathbf{Z}(\mathbf{Z}^*)^T)}{\partial \mathbf{Z}} = \mathbf{Z}^* \quad \text{and} \quad \frac{\partial g}{\partial \mathbf{Z}^*} = \mathbf{Z}.$$

Another example is the conjugate derivative of $\log|\det(\mathbf{W})|^2$, a term that acts as a regularizer in the derivation of ICA algorithms using maximum likelihood (ML) cost. It can be simply evaluated as

$$\begin{aligned}\frac{\partial \log|\det(\mathbf{W})|^2}{\partial \mathbf{W}^*} &= \frac{\partial \log[\det(\mathbf{W})\det(\mathbf{W}^*)]}{\partial \mathbf{W}^*}\\ &= \frac{\partial \log[\det(\mathbf{W}^*)]}{\partial \mathbf{W}^*} = \mathbf{W}^{-H}.\end{aligned} \tag{19}$$

There are a number of comprehensive references (e.g., [40], [41], [44], [46], [53], and [54]) on Wirtinger calculus that deal with the chain rule for nonanalytic functions, complex gradients and Hessians, and complex Taylor series expansions.

Complex-valued ICA is an example for the usefulness of Wirtinger calculus for complex-valued matrix derivatives. When the demixing matrix is not constrained to be unitary, it is a matrix-valued parameter that needs to be optimized. In this case, Wirtinger calculus has again proven powerful, both for the derivation of the algorithm and in stability and performance analysis [28], [55], [56].

STATISTICS

How is a complex random vector statistically described using probability distributions and moments? This is the topic of this section, where we pay particular attention to the ever-important Gaussian distribution.

COMPLEX RANDOM VECTORS

An N-dimensional complex random vector $\mathbf{x}$ is defined as $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, where $\mathbf{x}_r$ and $\mathbf{x}_i$ are a pair of N-dimensional real random vectors. The probability distribution (density) of a complex random vector is given by the joint distribution (density) of its real and imaginary parts. If the probability density function (pdf) exists, we write this as

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(\mathbf{x}_r + j\mathbf{x}_i) \triangleq p_{\mathbf{x}_r\mathbf{x}_i}(\mathbf{x}_r, \mathbf{x}_i).$$

If there is no risk of confusion, the subscripts may also be dropped. The expected value of a function $g:\mathbb{D} \to \mathbb{C}^N$ whose domain $\mathbb{D}$ includes the range of $\mathbf{x}$, is given by

$$\begin{aligned}E\{g(\mathbf{x})\} &= E\{\mathrm{Re}[g(\mathbf{x})]\} + jE\{\mathrm{Im}[g(\mathbf{x})]\}\\ &= \int_{\mathbb{R}^{2N}} g(\mathbf{x}_r + j\mathbf{x}_i)p(\mathbf{x}_r, \mathbf{x}_i)d\mathbf{x}_r d\mathbf{x}_i.\end{aligned} \tag{20}$$

Unless otherwise stated, we assume that all random vectors have zero mean. To characterize the second-order statistical properties of $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$, we consider the real composite random vector $\mathbf{x}_{\mathbb{R}} = [\mathbf{x}_r^T, \mathbf{x}_i^T]^T$. Its covariance matrix is

$$C_{x_{\mathbb{R}}x_{\mathbb{R}}} = E\{\mathbf{x}_{\mathbb{R}}\mathbf{x}_{\mathbb{R}}^T\} = \begin{bmatrix}\mathbf{C}_{x_rx_r} & \mathbf{C}_{x_rx_i}\\ \mathbf{C}_{x_rx_i}^T & \mathbf{C}_{x_ix_i}\end{bmatrix}$$

with $\mathbf{C}_{x_rx_r} = E\{\mathbf{x}_r\mathbf{x}_r^T\}$, $\mathbf{C}_{x_rx_i} = E\{\mathbf{x}_r\mathbf{x}_i^T\}$, and $\mathbf{C}_{x_ix_i} = E\{\mathbf{x}_i\mathbf{x}_i^T\}$. The augmented covariance matrix of $\mathbf{x}$ is [57]–[59]

$$\underline{\mathbf{C}}_{xx} = E\{\underline{\mathbf{x}}\,\underline{\mathbf{x}}^H\} = \mathbf{U}_N\mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}}\mathbf{U}_N^H = \begin{bmatrix}\mathbf{C}_{xx} & \widetilde{\mathbf{C}}_{xx}\\ \widetilde{\mathbf{C}}_{xx}^* & \mathbf{C}_{xx}^*\end{bmatrix} = \underline{\mathbf{C}}_{xx}^H, \tag{21}$$

which is related to the real covariance matrix $\mathbf{C}_{x_{\mathbb{R}}x_{\mathbb{R}}}$ through the real-to-complex transformation $\mathbf{U}_N$ from (10). The northwest block of the augmented covariance matrix is the usual (Hermitian and nonnegative definite) covariance matrix

$$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^H\} = \mathbf{C}_{x_rx_r} + \mathbf{C}_{x_ix_i} + j(\mathbf{C}_{x_rx_i}^T - \mathbf{C}_{x_rx_i}) = \mathbf{C}_{xx}^H, \tag{22}$$

and the northeast block is the complementary covariance matrix

$$\widetilde{\mathbf{C}}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{C}_{x_rx_r} - \mathbf{C}_{x_ix_i} + j(\mathbf{C}_{x_rx_i}^T + \mathbf{C}_{x_rx_i}) = \widetilde{\mathbf{C}}_{xx}^T, \tag{23}$$

which uses a regular transpose rather than a Hermitian (conjugate) transpose. Other names for $\widetilde{\mathbf{C}}_{xx}$ include *pseudo-covariance matrix* [60], *conjugate covariance matrix* [2], or *relation matrix* [5]. It is important to note that, in general, both $\mathbf{C}_{xx}$ and $\widetilde{\mathbf{C}}_{xx}$ are required for a complete second-order characterization of $\mathbf{x}$. In the important special case where the complementary covariance matrix vanishes, $\widetilde{\mathbf{C}}_{xx} = 0$, $\mathbf{x}$ is called *proper*, otherwise *improper* [60]. Similar to the augmented vector $\underline{\mathbf{x}}$, the augmented covariance matrix $\underline{\mathbf{C}}_{xx}$ is an equivalent redundant, but convenient, representation of $\mathbf{C}_{xx}$ and $\widetilde{\mathbf{C}}_{xx}$.

Necessary and sufficient conditions for propriety on the covariance and cross-covariance of real and imaginary parts $\mathbf{x}_r$ and $\mathbf{x}_i$ are $\mathbf{C}_{x_rx_r} = \mathbf{C}_{x_ix_i}$ and $\mathbf{C}_{x_rx_i} = -\mathbf{C}_{x_rx_i}^T$. When $x = x_r + jx_i$ is scalar, then having uncorrelated real and imaginary parts is necessary, but not sufficient, for propriety. If x is proper, its Hermitian covariance matrix is

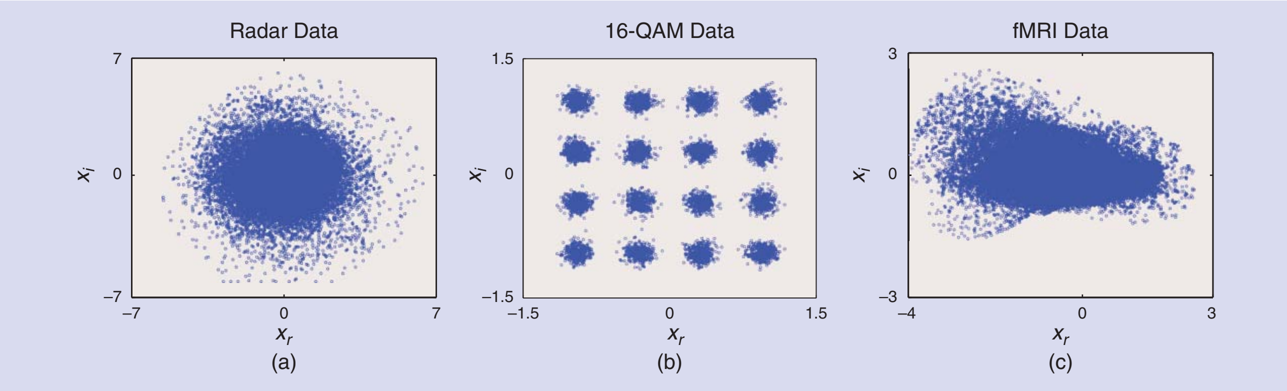
$$\mathbf{C}_{xx} = 2\mathbf{C}_{x_rx_r} - 2j\mathbf{C}_{x_rx_i} = 2\mathbf{C}_{x_ix_i} + 2j\mathbf{C}_{x_rx_i}^T,$$

and its augmented covariance matrix $\underline{\mathbf{C}}_{xx}$ is block-diagonal. If complex x is proper and scalar, then its variance is twice the variance of real and imaginary parts: $\sigma_x{}^2 = 2\sigma_{x_r}{}^2 = 2\sigma_{x_i}{}^2$.

The complex multivariate Gaussian pdf can be written in terms of the augmented covariance matrix as [57], [58]:

$$p(\mathbf{x}) = \frac{1}{\pi^N\sqrt{\det\underline{\mathbf{C}}_{xx}}}\exp\left\{-\frac{1}{2}\underline{\mathbf{x}}^H\underline{\mathbf{C}}_{xx}^{-1}\underline{\mathbf{x}}\right\}. \tag{24}$$

This pdf depends algebraically on $\underline{\mathbf{x}}$, i.e., $\mathbf{x}$ and $\mathbf{x}^*$, but is interpreted as the joint pdf of $\mathbf{x}_r$ and $\mathbf{x}_i$, and can be used for proper or improper $\mathbf{x}$. In the past, the term *complex Gaussian distribution* often implicitly assumed propriety. Therefore, some researchers call an improper complex Gaussian random vector *generalized complex Gaussian*, not to be confused with the complex generalized Gaussian [61]. It is particularly illuminating and instructive to look at the scalar complex improper



[FIG1] Scatter plots for (a) circular, (b) proper but noncircular, and (c) improper (and thus noncircular) data.

Gaussian case; see [46] for a detailed discussion. The simplification that occurs when $\widetilde{\mathbf{C}}_{xx} = 0$ is obvious and leads to the pdf of a complex proper Gaussian random vector $\mathbf{x}$ [62], [63]

$$p(\mathbf{x}) = \frac{1}{\pi^{N}\det\mathbf{C}_{xx}}\exp\{-\mathbf{x}^{H}\mathbf{C}_{xx}^{-1}\mathbf{x}\}.$$

It is also possible to define a stronger version of propriety in terms of the probability distribution of a random vector. A vector is called *circular* if its probability distribution is rotationally invariant, i.e., $\mathbf{x}$ and $\mathbf{x}' = e^{j\alpha}\mathbf{x}$ have the same probability distribution for any given real α. Circularity does not imply any condition on the standard covariance matrix $\mathbf{C}_{xx}$ because

$$\mathbf{C}_{x'x'} = E\{\mathbf{x}'\mathbf{x}'^{H}\} = E\{e^{j\alpha}\mathbf{x}\mathbf{x}^{H}e^{-j\alpha}\} = \mathbf{C}_{xx}. \quad (25)$$

On the other hand,

$$\widetilde{\mathbf{C}}_{x'x'} = E\{\mathbf{x}'\mathbf{x}'^{T}\} = E\{e^{j\alpha}\mathbf{x}\mathbf{x}^{T}e^{j\alpha}\} = e^{j2\alpha}\widetilde{\mathbf{C}}_{xx} \quad (26)$$

can be equal to $\widetilde{\mathbf{C}}_{xx}$ for arbitrary α if and only if $\widetilde{\mathbf{C}}_{xx} = 0$. Because the Gaussian distribution is completely determined by second-order statistics, a complex Gaussian random vector $\mathbf{x}$ is circular if and only if it is zero-mean and proper [64].

Propriety requires that second-order moments be rotationally invariant, whereas circularity requires that the distribution, and thus all moments (if they exist), be rotationally invariant. Therefore, circularity implies zero mean and propriety, but not vice versa, and impropriety implies noncircularity, but not vice versa. By extending the reasoning of (25) and (26) to higher-order moments, we see that if $\mathbf{x}$ is circular, a pth-order moment can be nonzero only if it has the same number of conjugated and nonconjugated terms [4], [38]. In particular, all odd-order moments must be zero. This holds for arbitrary order p.

EXAMPLES

As examples for proper/improper/noncircular signals, we show scatter plots—sample values in the complex plane—of three signals in Figure 1: (a) Ice Multiparameter Imaging X-Band Radar (IPIX) data from http://soma.crl.mcmaster.ca/ipix/, (b) a 16-quadrature amplitude modulated (QAM) signal, and (c) the functional MRI data for a simple box-car type paradigm [40], which is naturally represented as complex valued.

The radar signal in Figure 1(a) is narrowband. Evidently, the gain and phase of the in-phase and quadrature channels are matched, as the data appear circular, and, therefore proper. The uniform phase is due to carrier phase fluctuation from pulse-to-pulse and the amplitude fluctuations are due to variations in the scattering cross section. The 16-QAM signal in (b) has zero complementary covariance function and is therefore proper—second-order circular. However, its distribution is not rotationally invariant and therefore it is noncircular. The fMRI component shown in (c) is highly noncircular as can be easily observed. Of course, classifying signals as circular or improper/noncircular should not be done based on inspection of their scatter plots, but rather on sound statistical arguments. In the next section, we discuss how we can measure the degree of impropriety.

CIRCULARITY COEFFICIENTS AND ENTROPY

We now derive a maximal invariant for the augmented covariance matrix $\underline{\mathbf{C}}_{xx}$ under nonsingular linear transformation. Such a set is given by the canonical correlations between $\mathbf{x}$ and its conjugate $\mathbf{x}^{*}$, which [27] calls the *circularity coefficients* of $\mathbf{x}$. *Maximal invariant* means two things: 1) the circularity coefficients are invariant under nonsingular linear transformation, and 2) if two jointly Gaussian random vectors $\mathbf{x}$ and $\mathbf{y}$ have the same circularity coefficients, then $\mathbf{x}$ and $\mathbf{y}$ are related by a nonsingular linear transformation, $\mathbf{x} = \mathbf{M}\mathbf{y}$.

Assuming $\mathbf{C}_{xx}$ has full rank, the canonical correlations between $\mathbf{x}$ and $\mathbf{x}^{*}$ are determined by starting with the coherence matrix [65]

$$\mathbf{R} = \mathbf{C}_{xx}^{-1/2}\widetilde{\mathbf{C}}_{xx}[(\mathbf{C}_{xx}^{*})^{-1/2}]^{H} = \mathbf{C}_{xx}^{-1/2}\widetilde{\mathbf{C}}_{xx}[\mathbf{C}_{xx}^{-1/2}]^{T}. \quad (27)$$

Since $\mathbf{R}$ is complex symmetric, $\mathbf{R} = \mathbf{R}^{T}$, yet not Hermitian symmetric, i.e., $\mathbf{R} \neq \mathbf{R}^{H}$, there exists a special singular value decomposition, called the *Takagi factorization* [47], which is

$$\mathbf{R} = \mathbf{F}\mathbf{K}\mathbf{F}^{T}. \quad (28)$$

The complex matrix $\mathbf{F}$ is unitary, and $\mathbf{K} = \text{diag}\ (k_1, k_2, \ldots, k_N)$ contains the canonical correlations $1 \geq k_1 \geq\ k_2 \geq \cdots \geq k_N \geq 0$ on its diagonal. The canonical correlations are a maximal invariant for $\underline{\mathbf{C}}_{xx}$ under nonsingular linear transformation of $\mathbf{x}$. Therefore, any function of $\underline{\mathbf{C}}_{xx}$ that is invariant under nonsingular linear transformation must be a function of these canonical correlations only. Following [27], we call these canonical correlations k_n the *circularity coefficients*, and the set $\{k_n\}_{n=1}^{N}$ the *circularity spectrum* of $\mathbf{x}$, however note that they actually measure the degree of impropriety. The asymptotic distribution of the estimated circularity coefficients has been derived by [66].

DIFFERENTIAL ENTROPY

The (differential) entropy of a complex random vector $\mathbf{x}$ is defined to be the entropy of the real composite vector $\mathbf{x}_{\mathbb{R}}$. The entropy of a complex Gaussian random vector $\mathbf{x}$ with augmented covariance matrix $\underline{\mathbf{C}}_{xx}$ is thus

$$\mathrm{H}(\mathbf{x}) = \frac{1}{2}\log[(\pi e)^{2N} \det \underline{\mathbf{C}}_{xx}].$$

Noting that

$$\det \underline{\mathbf{C}}_{xx} = \det{}^2 \mathbf{C}_{xx} \det(\mathbf{I} - \mathbf{K}\mathbf{K}^H) = \det{}^2 \mathbf{C}_{xx} \prod_{n=1}^{N} (1 - k_n^2), \quad (29)$$

we may write the entropy of a complex noncircular Gaussian random vector $\mathbf{x}$ as [27], [67]

$$\begin{aligned} H_{\text{noncircular}} &= \frac{1}{2}\log[(\pi e)^{2N} \det \underline{\mathbf{C}}_{xx}] \\ &= \underbrace{\log[(\pi e)^{N} \det \mathbf{C}_{xx}]}_{H_{\text{circular}}} + \underbrace{\frac{1}{2}\log \prod_{n=1}^{N} (1 - k_n^2)}_{\leq 0}, \quad (30) \end{aligned}$$

where H_{circular} is the entropy of a circular Gaussian random vector with the same Hermitian covariance matrix $\mathbf{C}_{xx}$ (but $\widetilde{\mathbf{C}}_{xx} = 0$). The entropy is maximized if and only if $\mathbf{x}$ is circular. If $\mathbf{x}$ is noncircular, the loss in entropy compared to the circular case is given by the second term in (30), which is a function of the circularity spectrum. This loss in entropy can serve as a measure for the degree of impropriety. The maximally improper case, which is also called rectilinear [23], is given by $\mathbf{K} = \mathbf{I}$.

At this point, we will make a cautionary remark. On the one hand, if signals are indeed noncircular, we would expect a noncircular model to capture their properties more accurately. On the other hand, noncircular models have more degrees of freedom than circular models, and the principle of parsimony says that one should choose simple models to avoid overfitting to noise fluctuations. This means that using a noncircular model for a circular or only slightly noncircular signal is generally detrimental. For instance, it may slow down the convergence speed of iterative algorithms [68]. This can be also addressed as a model selection problem using information theoretic criteria [68], [69] to show that circular models are to be preferred not only when the degree of noncircularity is low but also when the signal-to-noise ratio is low, or the number of samples is small.

Choosing between proper/circular and improper/noncircular models is a question of how to detect noncircularity, for which a number of tests exist [65], [70]–[74]. For the simplest generalized likelihood ratio test (GLRT) for impropriety, the test statistic is the loss of entropy, given by the second term in (30). The more general problem of detecting the number of circular and noncircular component signals in an observed signal has been discussed by [69].

WIDELY LINEAR ESTIMATION

To introduce the main idea behind widely linear estimation, we first discuss correlation coefficients between two complex random variables x and y, and linear and conjugate linear estimation of x from y.

ROTATIONAL AND REFLECTIONAL CORRELATIONS

For a pair of complex zero-mean random variables x and y, as a straightforward extension of the real case, we can define the complex correlation coefficient

$$\rho_{xy} = \frac{E\{xy^*\}}{\sqrt{E\{|x|^2\}}\sqrt{E\{|y|^2\}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad (31)$$

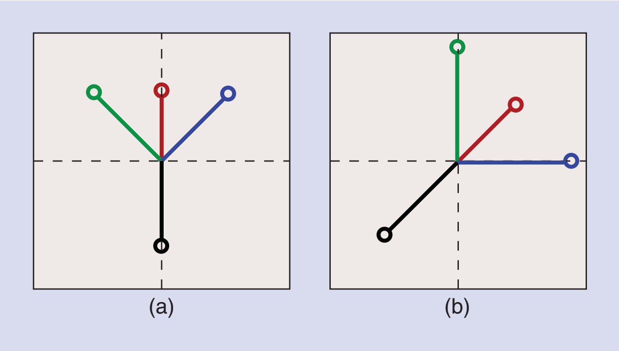
which satisfies $0 \leq |\rho_{xy}| \leq 1$. The linear minimum MSE (LMMSE) estimate of x from y is

$$\hat{x}(y) = \frac{\sigma_{xy}}{\sigma_y^2} y = \frac{|\sigma_{xy}|}{\sigma_y^2} e^{j\measuredangle\sigma_{xy}} y, \quad (32)$$

which achieves the minimum error

$$E\{|\hat{x}(y) - x|^2\} = \sigma_x^2 - \frac{|\sigma_{xy}|^2}{\sigma_y^2} = \sigma_x^2 (1 - |\rho_{xy}|^2). \quad (33)$$

Hence, if $|\rho_{xy}| = 1$, $\hat{x}(y)$ is a perfect estimate of x from y. Figure 2 depicts four sample pairs of two complex random variables x and y with $\rho_{xy} = \exp(-j\pi/4)$, in the complex plane. Plot (a) shows samples of x and (b) the corresponding



[FIG2] Sample pairs of two complex random variables x and y with $\rho_{xy} = \exp(-j\pi/4)$, and $|\sigma_{xy}|/\sigma_y^2 = 1.2$. Plot (a) depicts samples of x and (b) samples of y in the complex plane. Corresponding samples are shown in the same color.

samples of y. We observe that (b) is simply a scaled and rotated version of (a). The amplitude is scaled by the factor $|\sigma_{xy}|/\sigma_y^2$, preserving the aspect ratio, and the rotation angle is $\measuredangle\sigma_{xy} = \measuredangle\rho_{xy}$.

In the section "Statistics," we saw that to characterize the second-order statistics of a complex random variable x, we need to consider both the variance σ_x^2 and the complementary variance $\tilde{\sigma}_x^2$, which is the covariance between x and x^*. This suggests that when considering two random variables x and y, we should not only consider the covariance between x and y, but also the complementary covariance $\tilde{\sigma}_{xy} = E[xy]$, which is the covariance between x and y^*. We can do so by estimating x as a linear function of y^*

$$\hat{x}(y^*) = \frac{E\{x(y^*)^*\}}{E\{|y|^2\}}y^* = \frac{\tilde{\sigma}_{xy}}{\sigma_y^2}y^* = \frac{|\tilde{\sigma}_{xy}|}{\sigma_y^2}e^{j\measuredangle\tilde{\sigma}_{xy}}y^*. \tag{34}$$

We call this estimator the *conjugate linear* minimum MSE (CLMMSE) estimator. The corresponding correlation coefficient is

$$\tilde{\rho}_{xy} = \frac{\tilde{\sigma}_{xy}}{\sigma_x\sigma_y}, \tag{35}$$

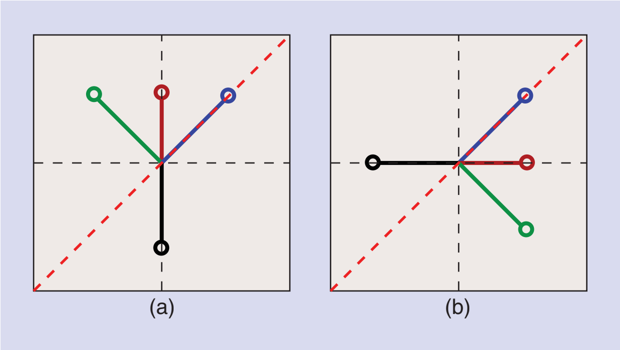
with $0 \le |\tilde{\rho}_{xy}| \le 1$, and the CLMMSE is

$$E\{|\hat{x}(y^*) - x|^2\} = \sigma_x^2 - \frac{|\tilde{\sigma}_{xy}|^2}{\sigma_y^2} = \sigma_x^2(1 - |\tilde{\rho}_{xy}|^2). \tag{36}$$

Hence, if $|\tilde{\rho}_{xy}| = 1$, $\hat{x}(y^*)$ is a perfect estimate of x from y^*. Figure 3 depicts four sample pairs of two complex random variables x and y with $\tilde{\rho}_{xy} = \exp(j\pi/2)$, in the complex plane. Figure 3(a) shows samples of x and (b) the corresponding samples of y. We observe that (b) is a reflected version of (a). The amplitude is unchanged because $|\tilde{\sigma}_{xy}|/\sigma_y^2 = 1$. Since $\measuredangle x = \measuredangle\tilde{\sigma}_{xy} - \measuredangle y$, we have, with probability 1

$$\measuredangle x - \frac{1}{2}\measuredangle\tilde{\sigma}_{xy} = -\left(\measuredangle y - \frac{1}{2}\measuredangle\tilde{\sigma}_{xy}\right). \tag{37}$$

Thus, the reflection axis is given by $\measuredangle\tilde{\sigma}_{xy}/2 = \measuredangle\tilde{\rho}_{xy}/2$, which is the dashed red line in the figure.



[FIG3] Sample pairs of two complex random variables x and y with $\tilde{\rho}_{xy} = \exp(j\pi/2)$ and $|\tilde{\sigma}_{xy}|/\sigma_y^2 = 1$. Plot (a) depicts samples of x and (b) samples of y in the complex plane. Samples of y correspond to reflected samples of x. The reflection axis is the dashed red line, which is given by $\measuredangle\tilde{\sigma}_{xy}/2 = \measuredangle\tilde{\rho}_{xy}/2 = \pi/4$.

Depending on whether rotation or reflection better models the relationship between x and y, $|\rho_{xy}|$ or $|\tilde{\rho}_{xy}|$ will be greater. We note the ease with which the best possible reflection axis is determined as half the angle of the complementary covariance $\tilde{\sigma}_{xy}$ (or half the angle of the correlation coefficient $\tilde{\rho}_{xy}$). This would be significantly more cumbersome with real-valued notation.

Of course, data might exhibit a combination of rotational and reflectional correlation, motivating a WLMMSE

$$\hat{x}(y, y^*) = \alpha y + \beta y^*, \tag{38}$$

where α and β are chosen to minimize $E\{|\hat{x}(y, y^*) - x|^2\}$. We will derive this WLMMSE estimator in the next section. More details about rotational and reflectional correlations are provided in [46].

WIDELY LINEAR MMSE ESTIMATION

Next, we extend the results from the previous subsection to the more general setting of estimating an N-dimensional message (or signal) $\mathbf{x}$ from an M-dimensional measurement $\mathbf{y}$. If a random variable is improper, we should use widely linear estimators rather than linear estimators to achieve the best possible performance. For vector-valued signal and measurement, such a widely linear estimator takes on the form

$$\hat{\mathbf{x}} = \mathbf{W}_1\mathbf{y} + \mathbf{W}_2\mathbf{y}^*. \tag{39}$$

We can simplify the derivation of $\mathbf{W}_1$ and $\mathbf{W}_2$ by working with augmented vectors. Using $\underline{\mathbf{x}} = [\mathbf{x}^T, \mathbf{x}^H]^T$ and $\underline{\mathbf{y}} = [\mathbf{y}^T, \mathbf{y}^H]^T$, we can write (39) equivalently as

$$\underline{\hat{\mathbf{x}}} = \underline{\mathbf{W}}\,\underline{\mathbf{y}}, \tag{40}$$

where we have introduced the augmented matrix

$$\underline{\mathbf{W}} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_2^* & \mathbf{W}_1^* \end{bmatrix}.$$

If we work with augmented matrices, we always need to enforce the block-pattern of $\underline{\mathbf{W}}$, where the northwest block is the conjugate of the southeast block, and the northeast block is the conjugate of the southwest block. We now determine $\underline{\mathbf{W}}$ such that the MSE $E\{\|\hat{\mathbf{x}} - \mathbf{x}\|^2\} = (1/2)E\{\|\underline{\hat{\mathbf{x}}} - \underline{\mathbf{x}}\|^2\}$ is minimized. This can be done by applying the orthogonality principle $(\hat{\mathbf{x}} - \mathbf{x}) \perp \mathbf{y}$ and $(\hat{\mathbf{x}} - \mathbf{x}) \perp \mathbf{y}^*$ [37], or equivalently, $(\underline{\hat{\mathbf{x}}} - \underline{\mathbf{x}}) \perp \underline{\mathbf{y}}$. This says that the error between the augmented estimator and the augmented signal must be orthogonal to the augmented measurement. This leads to $E\{\underline{\hat{\mathbf{x}}}\,\underline{\mathbf{y}}^H\} - E\{\underline{\mathbf{x}}\,\underline{\mathbf{y}}^H\} = 0$ and thus

$$\underline{\mathbf{W}}\,\underline{\mathbf{C}}_{yy} - \underline{\mathbf{C}}_{xy} = 0 \Leftrightarrow \underline{\mathbf{W}} = \underline{\mathbf{C}}_{xy}\underline{\mathbf{C}}_{yy}^{-1}. \tag{41}$$

Thus, $\underline{\hat{\mathbf{x}}} = \underline{\mathbf{C}}_{xy}\underline{\mathbf{C}}_{yy}^{-1}\underline{\mathbf{y}}$, or equivalently, [37]

$$\hat{\mathbf{x}} = (\mathbf{C}_{xy} - \widetilde{\mathbf{C}}_{xy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^*)\mathbf{P}_{yy}^{-1}\mathbf{y} + (\widetilde{\mathbf{C}}_{xy} - \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\widetilde{\mathbf{C}}_{yy})\mathbf{P}_{yy}^{-*}\mathbf{y}^*.$$

In this equation, the Schur complement $\mathbf{P}_{yy} = \mathbf{C}_{yy} - \widetilde{\mathbf{C}}_{yy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^{*}$ is the error covariance matrix for linearly estimating $\mathbf{y}$ from $\mathbf{y}^{*}$. The augmented error covariance matrix $\underline{\mathbf{Q}}$ of the error vector $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$ is

$$\underline{\mathbf{Q}} = E\{\underline{\mathbf{e}}\,\underline{\mathbf{e}}^{H}\} = \underline{\mathbf{C}}_{xx} - \underline{\mathbf{C}}_{xy}\underline{\mathbf{C}}_{yy}^{-1}\underline{\mathbf{C}}_{xy}^{H}.$$

A competing estimator $\underline{\hat{\mathbf{x}}}' = \underline{\mathbf{W}}'\underline{\mathbf{y}}$ will produce an augmented error $\underline{\mathbf{e}}' = \underline{\hat{\mathbf{x}}}' - \underline{\mathbf{x}}$ with covariance matrix

$$\underline{\mathbf{Q}}' = E\{\underline{\mathbf{e}}'\underline{\mathbf{e}}'^{H}\} = \underline{\mathbf{Q}} + (\underline{\mathbf{W}} - \underline{\mathbf{W}}')\,\underline{\mathbf{C}}_{yy}(\underline{\mathbf{W}} - \underline{\mathbf{W}}')^{H}, \qquad (42)$$

which shows that $\underline{\mathbf{Q}}' - \underline{\mathbf{Q}}$ is nonnegative definite, i.e., $\underline{\mathbf{Q}} \le \underline{\mathbf{Q}}'$. As a consequence, all real-valued increasing functions of $\underline{\mathbf{Q}}$ are minimized, in particular, $E\{\|\underline{\mathbf{e}}\|^{2}\} = \mathrm{Trace}\{\underline{\mathbf{Q}}\} \le \mathrm{Trace}\{\underline{\mathbf{Q}}'\} = E\{\|\underline{\mathbf{e}}'\|^{2}\}$ and $\det\{\underline{\mathbf{Q}}\} \le \det\{\underline{\mathbf{Q}}'\}$. These statements hold for the error vector $\mathbf{e}$ as well as the augmented error vector $\underline{\mathbf{e}}$ because $\underline{\mathbf{Q}} \le \underline{\mathbf{Q}}' \Rightarrow \mathbf{Q} \le \mathbf{Q}'$. A particular choice for a generally suboptimum filter is the LMMSE filter

$$\underline{\mathbf{W}}' = \begin{bmatrix} \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1} & 0 \\ 0 & \mathbf{C}_{xy}^{*}\mathbf{C}_{yy}^{-*} \end{bmatrix} \Leftrightarrow \mathbf{W}' = \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1},$$

which ignores complementary covariance matrices.

An important special case is when the signal $\mathbf{x}$ is real. Then, $\widetilde{\mathbf{C}}_{xy} = \mathbf{C}_{xy}^{*}$, which leads to the simplified expression

$$\hat{\mathbf{x}} = 2\,\mathrm{Re}\{(\mathbf{C}_{xy} - \widetilde{\mathbf{C}}_{xy}\mathbf{C}_{yy}^{-*}\widetilde{\mathbf{C}}_{yy}^{*})\,\mathbf{P}_{yy}^{-1}\mathbf{y}\}.$$

While the WLMMSE estimate of a real signal from a complex signal is always real [37], the LMMSE estimate is generally complex.

APPLICATIONS

In this section, we discuss two important signal processing applications to demonstrate the importance of taking full statistical information into account: filtering and BSS. We provide a more detailed discussion on the use of second-order statistics in both filtering and source separation, and a shorter discussion, along with some key references to recent work, on HOS.

FILTERING

LINEAR AND WIDELY LINEAR FILTERING USING MSE AND GAUSSIAN ENTROPY

In the previous section, we discussed WLMMSE filtering. An obvious disadvantage of WLMMSE filtering is that it doubles the dimension of the filter. A model with more parameters is more prone to overfitting and also leads to slower convergence when gradient-type algorithms are employed [40], [68]. An alternative way to incorporate full second-order statistics is to use a strictly linear filter with Gaussian entropy as the cost [35], which is equivalent to minimizing the determinant of the augmented error covariance matrix, $\det(\underline{\mathbf{Q}})$. The determinant of $\underline{\mathbf{Q}}$ depends on both the covariance and the complementary error covariance, as opposed to the MSE, $\mathrm{Trace}\{\mathbf{Q}\} = (1/2)\mathrm{Trace}(\underline{\mathbf{Q}})$, which only accounts for the covariance matrix.

Let us consider the linear filtering example given in the section "Optimization: Vector Case," where we estimate $x(k)$ from observations taken at M time instants $\mathbf{y}(k) = [y(k), y(k-1), \ldots, y(k-M+1)]^{T}$ as $\hat{x}(k) = \mathbf{w}^{H}(k)\mathbf{y}(k)$. We can either use the MSE

$$J_{\mathrm{mse}}(\mathbf{w}) = E\{|e(k)|^{2}\} \qquad (43)$$

or the Gaussian entropy cost defined as

$$J_{\mathrm{ent}}(\mathbf{w}) = [E\{|e(k)|^{2}\}]^{2} - |E\{e^{2}(k)\}|^{2}, \qquad (44)$$

where $e(k) = x(k) - \hat{x}(k)$ and $\hat{x}(k)$ can be estimated using either a linear filter

$$\hat{x}(k) = \mathbf{w}^{H}\mathbf{y}(k)$$

or a widely linear filter

$$\hat{x}_{wl}(k) = \mathbf{v}^{H}\underline{\mathbf{y}}(k),$$

where the weight vector $\mathbf{v} = [v_0, v_1, \cdots, v_{2M-1}]^{T}$ has twice the dimension of the linear filter.

As shown in (18), we can directly calculate the optimal linear weight vector using Wirtinger derivatives as

$$\frac{\partial J_{\mathrm{mse}}(\mathbf{w})}{\partial \mathbf{w}^{*}} = 0 \Rightarrow \mathbf{w}_{\mathrm{opt}} = \mathbf{C}_{yy}^{-1}\mathbf{c}_{yx},$$

where $\mathbf{c}_{yx}(k) = E\{x^{*}(k)\mathbf{y}(k)\}$. If $x(k)$ and $y(k)$ are jointly WSS, this equation is independent of k.

Similarly, we can calculate the Wirtinger derivative of $J_{\mathrm{ent}}(\mathbf{w})$ and write

$$\begin{aligned} \frac{\partial J_{\mathrm{ent}}(\mathbf{w})}{\partial \mathbf{w}^{*}} &= -2E\{|e(k)|^{2}\}E\{e^{*}(k)\mathbf{y}(k)\} + 2E\{[e^{*}(k)]^{2}\}E\{e(k)\mathbf{y}(k)\} \\ &= 0 \Rightarrow E\{(e^{*}(k) - \rho_{e}^{*})\mathbf{y}(k)\} = 0, \end{aligned}$$

by defining the correlation coefficient $\rho_e = E\{e^{2}(k)\}/E\{|e(k)|^{2}\}$. The entropy cost does not admit a closed-form solution for $\mathbf{w}_{\mathrm{opt}}$, but we can use either Newton variant updates [35] or stochastic gradient updates

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \mu[e^{*}(k) - \hat{\rho}_{e}^{*}(k)e(k)]\mathbf{y}(k). \qquad (45)$$

The latter leads to the least stochastic entropy (LSE) algorithm [35]. Here, the correlation coefficient $\rho_e(k)$ is estimated using the sample average up to time k. Note that the term $\hat{\rho}_{e}^{*}(k)e(k)\mathbf{y}(k)$ takes impropriety of the error signal into account. When the error is proper, i.e., $\hat{\rho}_{e}^{*}(k) = 0$, the update in (45) reduces to the well-known least mean squares (LMS) update rule.

Similarly, we can also derive a widely linear LSE update rule as

$$\mathbf{v}(k+1) = \mathbf{v}(k) + \mu[e^{*}(k) - \hat{\rho}_{e}^{*}(k)e(k)]\underline{\mathbf{y}}(k), \qquad (46)$$

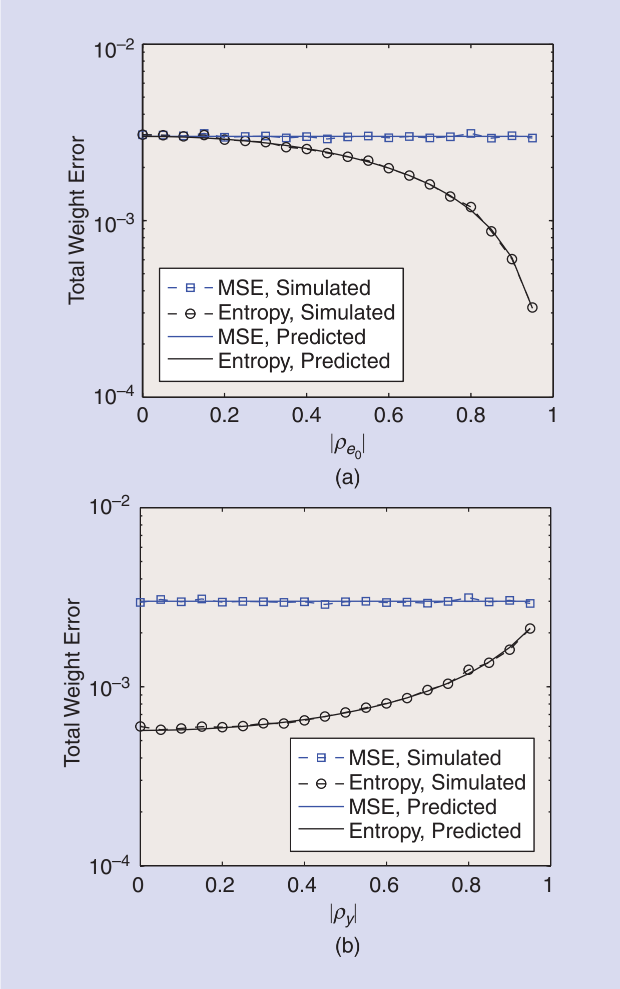
where the error is computed as $e(k) = x(k) - \mathbf{v}^H \underline{\mathbf{y}}(k)$. This update rule coincides with the LMS update for $\hat{\rho}_e(k) = 0$.

The performance of these filters depends on whether they are implemented adaptively or batch-wise. Let us first consider a batch implementation, where K samples are processed at a time. To compare the performance of the linear filters designed using the MSE and the Gaussian entropy criteria, we assume that $x(k)$ is generated using the linear regression model

$$x(k) = \mathbf{w}_0^H \mathbf{y}(k) + e_0(k), \tag{47}$$

where $e_0(k)$ is a white noise process, uncorrelated with the input $y(k)$, which is an uncorrelated process, $E\{\mathbf{y}(k)\mathbf{y}^H(l)\} = E\{\mathbf{y}(k)\mathbf{y}^T(l)\} = \mathbf{0}$ for $k \neq l$, i.e., both covariance and complementary covariance functions are zero for $k \neq l$.

We evaluate performance in terms of the weight error vector $\boldsymbol{\epsilon} = \mathbf{w} - \mathbf{w}_0$. The total weight error using the entropy criterion, for large K, is given by [35]



[FIG4] The total weight error for a linear filter with the Gaussian entropy and MSE criteria. Each simulation point is averaged over 1,000 independent runs. The predicted curves are calculated based on (48) and (49).

$$\varepsilon_{\text{ent}} \triangleq E\{\|\boldsymbol{\epsilon}\|^2\} = (1/2K)(1-|\rho_{e_0}|^2)\sigma_{e_0}^2 \operatorname{Trace}(\mathbf{K}^{-1}) + O(1/K^2), \tag{48}$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{C}_{yy} & -\rho_{e_0}^* \widetilde{\mathbf{C}}_{yy} \\ -\rho_{e_0} \widetilde{\mathbf{C}}_{yy}^* & \mathbf{C}_{yy}^* \end{bmatrix},$$

and $\mathbf{C}_{yy}$ and $\widetilde{\mathbf{C}}_{yy}$ are the covariance and the complementary covariance matrices of the input. For $\rho_{e_0} = 0$, the expression reduces to the total weight error for the MSE criterion

$$\varepsilon_{\text{mse}} = (1/K)\sigma_{e_0}^2 \operatorname{Trace}(\mathbf{C}_{yy}^{-1}) + O(1/K^2). \tag{49}$$

When the error is maximally improper, $|\rho_{e_0}| = 1$, we have $\varepsilon_{\text{ent}} = 0$. One can show that $\varepsilon_{\text{ent}} \leq \varepsilon_{\text{mse}}$, with equality only when $\rho_{e_0} = 0$ or $y(k)$ is maximally improper. In addition, it is the entropy criterion that leads to a best linear unbiased estimator (BLUE) for $\mathbf{w}_0$, where BLUE is to be understood with respect to a given vector, which in the case of widely linear filter is a filter in double dimension, the input vector augmented with its complex conjugate.

A similar study can be performed for the adaptive implementations, LMS and LSE, which minimize MSE and entropy, respectively, using stochastic gradient updates. In general, the LSE algorithm leads to better performance in terms of the total weight error. For a proper input, the LSE yields a smaller weight error if the additive noise is improper. However, as the degree of impropriety of the input increases, the LSE algorithm suffers from slower convergence rate due to an increase in the eigenvalue spread, the ratio of maximum to minimum eigenvalues. This is analogous to the widely linear LMS algorithm.

The situation is quite different for widely linear filters. If implemented batch-wise, the widely linear filters using the entropy and the MSE criteria lead to the same solution. This is due to the fact that the WLMMSE filter minimizes $\underline{\mathbf{Q}}$, and therefore all increasing functions of $\underline{\mathbf{Q}}$, in particular, both $\operatorname{Trace}(\underline{\mathbf{Q}})$ and $\det(\underline{\mathbf{Q}})$, are minimized. The total weight error (both for MSE and entropy cost function) is given by (49), except that $\mathbf{C}$ is replaced by $\underline{\mathbf{C}}$. Hence, both error criteria provide the same asymptotically BLUE for a widely linear optimal filter $\mathbf{v}_{\text{opt}}$.

With a stochastic gradient approach, however, the entropy criterion yields a smaller steady-state error than the MSE criterion if the residual error $e_{\text{opt}}(k)$ is noncircular. However, this gain comes at a price: An improper error also increases the overall eigenvalue spread of the augmented joint input-error covariance matrix that defines the modes of the algorithm [35]. This slows down its convergence rate. Hence, for the widely linear case, the widely linear LSE algorithm has no clear advantage over the widely linear LMS algorithm.

EXAMPLE

Here is an example to compare the performance of batch estimation of filter weights using either the MSE or the entropy criterion. Consider (47), where the input is a white sequence, and its covariance and complementary covariance matrices are given by $\mathbf{C}_{yy} = \mathbf{I}$ and $\widetilde{\mathbf{C}}_{yy} = \rho_y \mathbf{I}$. The real and imaginary parts of the filter coefficients of w are randomly drawn from the standard Gaussian distribution. The additive noise e_0 is a white noise process with unit variance. The order of the linear filter is $N = 3,$ and the sample size is 1,000. We vary the circularity coefficients of the input and the additive noise to observe their impact on the estimation of w.

Figure 4 shows the average total weight error of a linear filter with the two criteria, for input circularity coefficient $|\rho_y| = 0.3$ and noise circularity coefficient $|\rho_{e_0}| = 0.9$. For both cases, the Gaussian entropy criterion yields better performance in terms of total weight error. In the first case, the gain by using the entropy criterion increases with increasing impropriety of the additive noise. In the second case, the gain decreases with increasing input impropriety. It is also shown that the estimated performance gain closely matches the gain predicted by (48) and (49). ■

NONLINEAR FILTERING

As discussed above, full second-order statistics can be taken into account by a linear filter with Gaussian entropy cost or by a widely linear filter with either MSE or Gaussian entropy cost. When the assumption of Gaussianity is not realistic and prior information on the distribution of the residual error is available, the cost can be modified as

$$J_p(\mathrm{w}) = E\{|e(k)|^p\} = E\{|x(k) - \hat{x}(k)|^p\},$$

with $p \geq 1$. Alternatively, the complete statistics of the error can be taken into account by choosing as the cost the entropy of the error, $J_{\text{ent}}(\mathrm{w}) = -E\{\log p_e(e)\}$. In either case, Wirtinger calculus enables optimization of the selected cost function once it is expressed as a function of w and w^*, a straightforward task in most cases. For $J_{\text{ent}}(\mathrm{w})$, we can write $p_e(e_r, e_i)$ as a function of e and e^* to enable the use of Wirtinger derivates. The entropy can be approximated using either parametric or nonparametric methods. In [75], a semiparametric approach is used to propose a complex-valued filter based on the minimization of error entropy, where the entropy is estimated by choosing the tightest bound among a number of candidates using entropy bound minimization (EBM) [32], [76].

Given the expected dynamics of the underlying problem, the desired sequence $\hat{x}(k)$ can be approximated using either a linear, widely linear, or a nonlinear filter such that $\hat{x}(k) = g(\mathrm{w}, \mathrm{y}(k))$. Most of the real-valued adaptive filters have been extended to the complex domain, such as kernel filters [77]—which are closely related to the radial basis function networks—tapped-delay MLP structures [78], [79], and Volterra filters [80]. For most of these nonlinear structures, the popular MSE has been the cost function of choice.

When designing tapped-delay line MLP filters, the typical structure is a hidden layer with nonlinear activation functions $f(\cdot)$—typically of the squashing type (the hyperbolic tangent function)—followed by a linear output layer. When choosing the activation function $f(\cdot)$ for a complex MLP, a number of solutions [81]–[83] have emphasized boundedness and advocated the use of functions that process either the real and imaginary parts or the magnitude and phase of the complex variable separately, by defining bounded complex functions that are not complex-differentiable. A second approach adopts the use of nonlinearities that are complex-differentiable functions $f : \mathbb{C} \mapsto \mathbb{C}$ and hence have to possess singular points as stated by Liouville's theorem [42]. As an example, the commonly used nonlinearity $\tanh(z)$ has periodic singular points. MLP filters employing such nonlinearities are called *fully complex*. These filters are shown to be more efficient in approximating nonlinear functions, and lead to better performance in challenging problems such as equalization of highly nonlinear channels [50], and when using gradient-adaptive step-size algorithms [51]. More importantly, as in the real-valued case, it can be shown that an MLP that uses fully complex nonlinearities is a universal approximator of any smooth nonlinear mapping [79] despite the presence of singular points. For MLPs using either the split or the fully complex type nonlinearities, the backpropagation update rule as well as other second-order algorithms can be easily derived using Wirtinger calculus, making many efficient learning procedures developed for the real-valued case readily accessible in the complex domain [40].

Another important class of nonlinear adaptive filters is the class of kernel filters. Kernel methods make use of the theory of reproducing kernel Hilbert spaces (RKHS) to transform the typically low-dimensional input space to a high-dimensional, possibly infinite-dimensional, feature space. Hence, the nonlinear problem in the input space is transformed into a linear one in the feature space, which can now be solved by linear processing, i.e., simple quadratic optimization. Since most kernel methods are variations on the well-known support vector machine framework, they are best suited to batch processing. However, they have been extended to allow online processing [84], and a number of kernel adaptive filters including the kernel version of the LMS algorithm [85] have been proposed. A key result that enables the development of complex-valued kernel filters is the extension of Wirtinger calculus to functional spaces using Fréchet derivatives, which generalizes differentiability to Hilbert spaces. In [77], after extending Wirtinger calculus to include complex RKHS, a number of kernel LMS algorithms are derived by using real-valued reproducing kernels through a process called *complexification* as well as using complex Gaussian kernels. A number of practical examples along with a framework for adaptive learning in complex RKHS are given in [86]. The choice and design of effective complex kernels as well as ways to cope with the issue of increasing memory length in the implementation of kernel adaptive filters is an important research problem.

Finally, it is worth pointing out that widely linear filters can be extended to account for HOS. The next logical step is the extension to widely linear-quadratic processing [46], [80], [87], which requires statistical information up to fourth order. We should, however, add the cautionary note here that there is no difference between the optimum, generally nonlinear, conditional mean estimator $E\{x|y\}$, and $E\{x|y, y^*\}$. Conditioning on y and y^* changes nothing, since $E\{x|y\}$ already extracts all the information there is about x from y.

So the "widely nonlinear" estimator $E\{x\,|\,y,y^*\}$ is simply $E\{x\,|\,y\}$. A number of examples in nonlinear filtering can be found in [49].

INDEPENDENT COMPONENT ANALYSIS

Data-driven approaches such as BSS minimize the assumptions on the data and thus have become attractive alternatives to traditional model-based techniques, especially for problems where the underlying dynamics are difficult to characterize. BSS methods are typically based on a linear mixing model where the goal is to identify the underlying components, which may or may not correspond to physical quantities. Since independence of the underlying components in a given data set is a natural assumption, ICA has been the most popular way to achieve source separation. In addition, independence allows for easy interpretation of the results, and because it is a strong condition, admits a solution subject to only a scaling and permutation ambiguity.

The standard linear mixing model for ICA is given by

$$\mathbf{x}(v) = \mathbf{A}\mathbf{s}(v), \quad v = 1,\dots V, \tag{50}$$

where $\mathbf{x}(v) \in \mathbb{C}^N$ denotes the observation vector, $\mathbf{s}(v) \in \mathbb{C}^N$ the sources (or components), and $\mathbf{A} \in \mathbb{C}^{N\times N}$ is the nonsingular mixing matrix. The index v can be time, a spatial, or a volume index, e.g., a voxel in the case of fMRI analysis. The sources $s_n, n = 1,\dots,N$, are identifiable up to a scaling and permutation ambiguity because independence is invariant to those. The scaling ambiguity includes the phase as well as the magnitude since both $\mathbf{s}$ and $\mathbf{A}$ are assumed to be complex.

Given this simple linear model, ICA decomposition is achieved by determining a demixing matrix $\mathbf{W}$ such that $\mathbf{u}(v) = \mathbf{W}\mathbf{x}(v)$ are the source estimates. For this task, one has to make use of the statistical properties of the signals, i.e., some form of diversity. Non-Gaussianity of individual sources—exploited through HOS—has been the most commonly used type of diversity followed by the sample-to-sample correlation within each source. If there is sample correlation in the real-valued case, it allows separation of Gaussian sources as well. An important result for complex-valued ICA is that the circularity coefficients—discussed in the section "Circularity Coefficients and Entropy"—provide yet another source of diversity one can use. They enable separation of Gaussian sources even without sample correlation using the strong uncorrelating transform (SUT) [26], [27], provided that all the sources in the mixture are improper with distinct circularity coefficients.

Assuming that the sources are stationary, one can bring the two approaches for achieving ICA—use of HOS and sample dependence—under one umbrella by using mutual information rate as the cost function [88]

$$\mathrm{I}_r(\mathbf{W}) = \sum_{n=1}^{N} H_r(u_n) - \log|\det(\mathbf{W}_{\mathbb{R}})| - H_r(\mathbf{x}), \tag{51}$$

where $H_r(u_n) = \lim_{v\to\infty} H[u_n(1),\dots,u_n(v)]/v$ is the entropy rate of the nth source estimate u_n with pdf $p_{s_n}(u_n) = p_{s_{nr},s_{ni}}(u_{nr},u_{ni})$, and

$$\mathbf{W}_{\mathbb{R}} = \begin{bmatrix} \mathbf{W}_r & -\mathbf{W}_i \\ \mathbf{W}_i & \mathbf{W}_r \end{bmatrix}$$

is the real representation of complex $\mathbf{W} = \mathbf{W}_r + j\mathbf{W}_i$. It results from the use of Jacobian transformation $p_x(\mathbf{x}) = |\det \mathbf{W}_{\mathbb{R}}|\, p_u(\mathbf{W}\mathbf{x})$, where $p_u(\mathbf{W}\mathbf{x}) = p_u(\mathbf{u})$. Since $|\det \mathbf{W}_{\mathbb{R}}| = \det(\mathbf{W}\mathbf{W}^H) = |\det(\mathbf{W})|^2$, we can rewrite (51) as

$$\mathrm{I}_r(\mathbf{W}) = \sum_{n=1}^{N} H_r(u_n) - 2\log|\det(\mathbf{W})| - H_r(\mathbf{x}) \tag{52}$$

and use (19) along with the definition of Wirtinger derivatives (3) to develop gradient update rules for optimizing the cost.

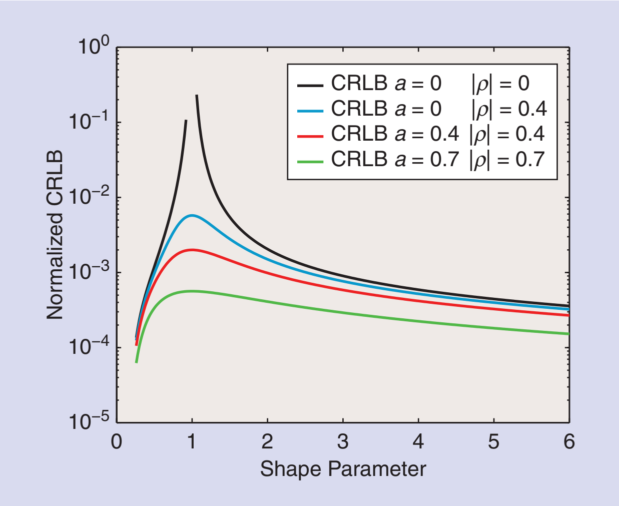
The entropy rate $H_r(\mathbf{x}) = \lim_{v\to\infty} H[\mathbf{x}(1),\dots,\mathbf{x}(v)]/v$ of the observations is constant with respect to $\mathbf{W}$, and thus the first term is sufficient to minimize statistical dependence among the sources. The second term $\log|\det(\mathbf{W})|$ helps with the regularization by penalizing ill-conditioned matrices. Minimization of the mutual information rate is equivalent to ML estimation when we write the expectations in (52) using the given set of observations $\{\mathbf{x}(v)\}_{v=1}^{V}$. When $\mathbf{W}$ is constrained to be unitary $(\mathbf{W}\mathbf{W}^H = \mathbf{I})$, then $\log|\det(\mathbf{W})| = 0$ and (52) becomes equivalent to maximization of negentropy rate as cost. Maximization of negentropy is achieved by minimizing the entropy of the source estimates under a variance constraint, another effective way to make use of HOS for separation. While this constraint provides a natural decoupling among the source estimates, it limits the search space for the optimal demixing matrix and results in suboptimal performance [28]. An effective decoupling approach [89], [90] allows one to retain the bigger optimization space of nonunitary matrices by decomposing the determinant term in (52) into a series of vector optimization problems. It is also important to remember that the whitening step typically employed in ICA algorithms implies constraining the demixing matrix to be unitary only when the number of samples tend to infinity.

A number of algorithms based on ML or maximization of negentropy have been derived for complex ICA, which has been an active field of study. Wirtinger calculus has played an important role in the development and analysis of algorithms especially when considering the general case that avoids the circularity assumption. Next, we summarize some of the work in the area, in terms of identifiability, performance, and algorithms.

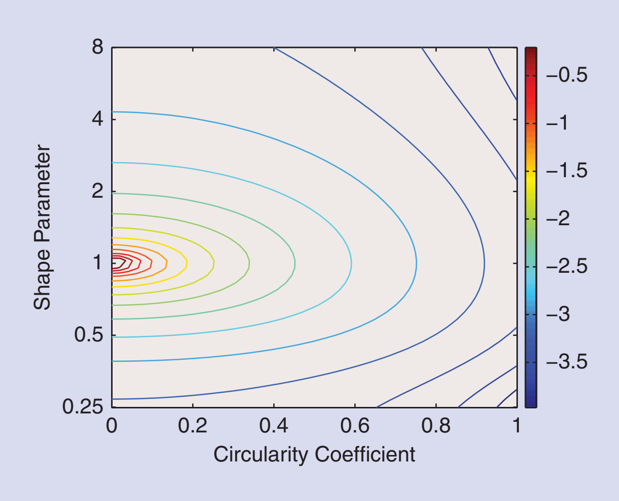
For the real-valued case, identification—up to the two ambiguities of ICA, permutation and scaling—is possible as long as there are no two sources in the mixture that are both Gaussian and have proportional covariance matrices. If sample correlation is not taken into account—or is absent because the samples are independent and identically distributed (i.i.d.)—and only HOS are used, then we can identify only a single Gaussian source [91]–[95]. For the complex case, with the additional diversity offered through the complementary covariance, i.i.d. Gaussian sources can be identified as well, as long as no two sources have the same circularity coefficient. When sample correlation is taken into account, the availability of complementary covariance as an additional source of information makes identifiability of the ICA model even easier. In this case, the ICA problem becomes nonidentifiable only when there are two Gaussian sources that have both their covariance and complementary covariance matrices proportional to each other, and proportional through a complex constant for the latter, as implied by the analyses in [56] and [96]. Hence, it is again the second-order properties that determine the identifiability. In [97], the performance of the SUT is analyzed using interference-to-source ratio as the metric. It is also shown that a maximally improper source can be perfectly separated from all other sources as long as these are not maximally improper as well.

In Figure 5, we demonstrate the role of the three types of diversity available for the complex ICA problem on performance and in terms of identifiability. We plot the Cramér–Rao lower bound (CRLB) for two sources: an i.i.d. source that comes from a generalized Gaussian distribution [61]—a unimodal symmetric density that is Gaussian when the shape parameter $\beta = 1$, super-Gaussian for $0 < \beta < 1$, and sub-Gaussian when $\beta > 1$. We consider three levels of noncircularity for the source by changing its circularity coefficient such that $|\rho| = 0, 0.4$ and 0.7. The second source is a first-order autoregressive (AR) process generated by a circular generalized Gaussian distributed (GGD) innovation process. By changing the value of AR coefficient a, we consider three cases with increasing sample dependence, $a = 0$ for which samples are i.i.d., as well as $a = 0.4$ and $a = 0.7$. We see that the only case that is not identifiable is when both sources are i.i.d. and circular Gaussians (black curve, $\beta = 1$, $a = 0$, and $\rho = 0$). With the addition of noncircularity, the ICA problem becomes identifiable (green curve), and as demonstrated by the trends of all four curves, performance improves when the degree of non-Gaussianity increases (when we move away from 1), and when sample dependence and noncircularity of the sources increase. Next, we consider the i.i.d. case and plot the CRLB for two sources, a GGD and a Gaussian, to show the role of diversity in a continuous scale. We change the degree of noncircularity of the Gaussian source and the shape parameter β of the second GGD source. As shown in Figure 6, performance improves with increasing noncircularity and non-Gaussianity. As these simple examples also demonstrate, the three types of diversity—noncircularity, non-Gaussianity, and sample correlation—all help improve the performance, and help with the identifiability of the ICA model.

Thus, the conditions for identifiability of the complex ICA model are quite relaxed. However, to achieve the desirable large sample properties of the ML estimator, one needs to estimate the density of the sources along with the demixing matrix. Modeling of the density for a complex random process has more degrees of freedom than a real-valued one, making the problem more difficult. Most of the complex ICA solutions to date have a focus on the use of HOS, ignoring sample correlation where the entropy rate in (52) is simply replaced by the entropy. Among those, early solutions assumed that the sources are circular [98], [99] which simplifies the derivation but fails to take the additional source of diversity into account. Thus, it limits the type of sources that can be successfully separated. Another approach used analytic nonlinearities within a nonlinear correlations framework, hence bypassing the need to directly optimize (52) [100]. On the other hand, by writing $p_s(u) = p_s(u_r, u_i)$ as a function of u and u^*, one can make use of Wirtinger derivatives and derive algorithms by directly minimizing (52) as shown in [28], thus eliminating the need to assume that the sources are circular. Among the solutions based on ML—or maximization of negentropy, which constrains $\mathbf{W}$ to be unitary in the ML cost—there are those that use complex nonlinear functions [30], [31], generalized Gaussian density model



[FIG5] **The CRLB for two sources, an i.i.d. GGD source with circularity coefficient of 0, 0.4, and 0.7 and a first-order AR source driven by a circular GGD innovation process with AR coefficient $a = 0$ (i.i.d.), $a = 0.4$, and $a = 0.7$. Note the improved performance as diversity increases in terms of noncircularity, sample dependence, and non-Gaussianity. The problem is not identifiable when both sources are i.i.d. circular Gaussians.**



[FIG6] **The CRLB contour plot for the i.i.d. case for separation of two sources: a Gaussian source with increasing noncircularity and a GGD source with changing shape parameter, degree of non-Gaussianity.**

for the sources, a good model for symmetric unimodal distributions [55], [61], and a semiparametric approach based on EBM [32]. All these solutions take potential noncircularity of the sources into account and provide better separation performance in terms of the minimum achievable interference-to-source ratio when the sources are noncircular. In [56], the CRLB is derived for ML ICA. It is shown that ML ICA, with the nonlinearity exactly matched to the source density, approaches this bound. Moreover, the complex EBM algorithm [32], which is not matched to the specific density but uses a flexible set of nonlinearities, also approaches the CRLB for large enough sample size. Many of the complex-valued ICA algorithms have been made available in a MATLAB-based [101] toolbox, LYCIA, which stands for Library of Complex Independent component analysis Algorithms, and is available at http://mlsp.umbc.edu/lycia/lycia.html. The toolbox allows for comparison of multiple algorithms, including algorithms supplied by the user through a number of metrics and visualization tools.

In [96], a second-order algorithm, entropy rate minimization (ERM), is derived by minimizing (52) directly and making use of both full second-order statistics and sample correlation. Information in terms of sample dependence is exploited by whitening the source estimates u_n through a widely linear filter $v = \mathbf{a}^H \underline{\mathbf{u}}$, such that $\mathbf{a} \in \mathbb{C}^{2p}$ is estimated by minimizing the Gaussian entropy subject to $||a(0)|^2 - |a(p)|^2| = 1$. Then the demixing matrix is estimated by minimizing $\mathcal{J}(\mathbf{W}, \mathbf{a}_1, \ldots, \mathbf{a}_N) = \sum_{n=1}^{N} \log E\{|v_n|^2\} - 2\log|\det(\mathbf{W})|$. The SUT [26], [27] becomes a special case of the ERM when the correlation lag is set to 0, i.e., sample correlation is not taken into account. However, the SUT can be computed directly in a straightforward manner by first estimating the coherence matrix in (27). Then, after computing a Takagi factorization of the coherence matrix as in (28), the SUT is $\mathbf{W} = \mathbf{F}^H \mathbf{C}_{xx}^{-1/2}$. Other complex ICA solutions that have been proposed include approaches that explicitly compute and maximize HOS such as kurtosis, either by taking the full statistics into account [33], [102] or by assuming circularity [103]; a noncircular version of the second-order blind identification algorithm [96] that diagonalizes time-lagged covariance and complementary covariances; and a quaternion ICA algorithm that uses full second-order statistics [104]. A recent review of complex ICA algorithms, both those based on joint diagonalization and ML estimation can be found in the book [88].

Development and performance analysis of complex ICA algorithms have been an active field of study, and noncircularity has played an important role in the development. On the one hand, it has provided an additional type of diversity relaxing the conditions for identifiability of the ICA model. On the other hand, it has made the problem more complicated due to two factors: density models that account for the potential noncircularity of the sources are more complicated making the development more challenging, and the local stability of the algorithms is impacted by the circularity coefficients where the algorithms become more prone to instability when sources are highly noncircular [31], [55]. Hence, the complex ICA problem is indeed a rich one, and there is still a need for algorithms that can directly minimize (52) by powerful complex density models that account for both sample dependence and HOS.

DISCUSSION

We have discussed the two key components for optimization and estimation in the complex domain: 1) the optimization of real-valued cost functions with respect to complex parameters using Wirtinger calculus and 2) utilizing the complete statistical description of complex random signals. With respect to the latter, we provided details on estimation using the full second-order statistics in terms of covariance and complementary covariances. Widely linear transformations provide an easy and straightforward way to incorporate such information. Most of the development for widely linear filters follows that of linear filters, as in the case of MSE estimation. This is a direct consequence of the homomorphism between the double-dimensional real and augmented complex representations. Doubling the dimensionality, however, increases the number of parameters to be estimated. This leads to potential overfitting as well as other undesirable effects such as slow convergence when gradient-type algorithms are used. In general, circular models are to be preferred when the signal-to-noise ratio is low, the number of samples is small, or the degree of noncircularity is low [68], [69]. As an alternative, accounting for the full second-order statistics with a linear filter, as discussed in the section "Filtering" might offer advantages.

This article reviewed some of the fundamental results and selected more recent developments in the field. There is a lot that we have not included, such as Cramér–Rao type performance bounds and practical applications in areas such as communications, medical image analysis, and array processing. Moreover, the only noncircular distribution we formally introduced is the Gaussian distribution. Of course, we do not live in a perfectly Gaussian world, so non-Gaussian noncircular distributions are important as well. For an account of these, we refer the reader to [61], [70], [73], and [74].

There are a number of recent references that provide a more comprehensive overview of complex-valued signal processing. Among those, [46] presents the theory for statistical signal processing, [40] discusses complex-valued optimization, robust estimation, and ICA, and [88] concentrates on complex-valued ICA. Another book on the topic is [49], which has a focus on neural networks, and two recent, more technical, overview papers are [68] and [74]. We hope that this review article will help further increase the activity in this growing and important area of signal processing.

ACKNOWLEDGMENTS

Tülay Adalı was supported by the National Science Foundation (NSF) grants NSF-IIS 1017718 and NSF-CCF 1117056. Peter J. Schreier was supported by the Alfried Krupp von Bohlen und Halbach foundation, under its program "Return of German Scientists from Abroad." He gratefully acknowledges the kind permission from Cambridge University Press to use some material from the book *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*.

AUTHORS

Tülay Adalı (adali@umbc.edu) is a professor in the Department of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County. She has actively assisted the IEEE in organizing numerous international conferences and workshops and chaired or served on various technical committees and editorial boards, including *Proceedings of the IEEE*. She is a Fellow of the IEEE and the American Institute for Medical and Biological Engineering, an IEEE Distinguished Lecturer, and a recipient of an NSF CAREER Award, 2010 IEEE Signal Processing Society Best Paper Award, and 2013 University System of Maryland Regents' Award for Research. Her research interests include statistical signal processing, machine learning for signal processing, and medical data analysis.

Peter J. Schreier (peter.schreier@sst.upb.de) is a professor and the head of the Signal and System Theory Group in the Department of Electrical Engineering and Information Technology at the Universität Paderborn, Germany. He received his master of science degree from the University of Notre Dame, Indiana, in 1999 and his Ph.D. degree from the University of Colorado at Boulder in 2003, both in electrical engineering. From 2004 until 2011, he was on the faculty of the University of Newcastle, Australia. He is currently a senior area editor of *IEEE Transactions on Signal Processing*.

REFERENCES

[1] C. N. K. Mooers, "A technique for the cross spectrum analysis of pairs of complex-valued time series, components and rotational invariants," *Deep-Sea Research*, vol. 20, pp. 1129–1141, 1973.

[2] W. A. Gardner, "Cyclic Wiener filtering: theory and method," *IEEE Trans. Commun.*, vol. 41, pp. 151–163, 1993.

[3] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1293–1302, July 1993.

[4] B. Picinbono, "On circularity," *IEEE Trans. Signal Processing*, vol. 42, pp. 3473–3482, Dec. 1994.

[5] B. Picinbono and P. Bondon, "Second-order statistics of random signals," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 411–419, Feb. 1997.

[6] W. Wirtinger, "Zur formalen theorie der funktionen von mehr komplexen veränderlichen," *Math. Ann.*, vol. 97, no. 1, pp. 357–375, 1927.

[7] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proceedings*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[8] A. van den Bos, "Estimation of complex parameters," in *Proc. 10th IFAC Symp.*, July 1994, vol. 3, pp. 495–499.

[9] A. van den Bos, "Complex gradient and Hessian," *IEE Proc. Vision, Image, Signal Processing*, vol. 141, no. 6, pp. 380–382, Dec. 1994.

[10] Y. C. Yoon and H. Leib, "Maximizing SNR in improper complex noise and applications to CDMA," *IEEE Commun. Letters*, vol. 1, pp. 5–8, 1997.

[11] S. Buzzi, M. Lops, and A. M. Tulino, "A new family of MMSE multiuser receivers for interference suppression in DS/CDMA systems employing BPSK modulation," *IEEE Trans. Commun.*, vol. 49, pp. 154–167, 2001.

[12] R. Nilsson, F. Sjoberg, and J. P. LeBlanc, "A rank-reduced LMMSE canceller for narrowband interference suppression in OFDM-based systems," *IEEE Trans. Commun.*, vol. 51, pp. 2126–2140, 2003.

[13] A. Napolitano and M. Tanda, "Doppler-channel blind identification for noncircular transmissions in multiple-access systems," *IEEE Trans. Commun.*, vol. 52, pp. 2073–2078, 2004.

[14] J. J. Jeon, J. G. Andrews, and K. M. Sung, "The blind widely linear minimum output energy algorithm for DS-CDMA systems," *IEEE Trans. Signal Processing*, vol. 54, pp. 1926–1931, 2006.

[15] A. S. Cacciapuoti, G. Gelli, and F. Verde, "FIR zero-forcing multiuser detection and code designs for downlink MC-CDMA," *IEEE Trans. Signal Processing*, vol. 55, pp. 4737–4751, 2007.

[16] M. Valkama, M. Renfors, and V. Koivunen, "Advanced methods for I/Q imbalance compensation in communication receivers," *IEEE Trans. Signal Process.*, vol. 49, pp. 2335–2344, 2001.

[17] L. Anttila, M. Valkama, and M. Renfors, "Circularity-based I/Q imbalance compensation in wideband direct-conversion receivers," *IEEE Trans. Vehicular Techn.*, vol. 57, pp. 2099–2113, 2008.

[18] P. Rykaczewski, M. Valkama, and M. Renfors, "On the connection of I/Q imbalance and channel equalization in direct-conversion transceivers," *IEEE Trans. Vehicular Techn.*, vol. 57, pp. 1630–1636, 2008.

[19] Y. Zou, M. Valkama, and M. Renfors, "Digital compensation of I/Q imbalance effects in space-time coded transmit diversity systems," *IEEE Trans. Signal Processing*, vol. 56, pp. 2496–2508, 2008.

[20] D. R. Morgan, "Variance and correlation of square-law detected allpass channels with bandpass harmonic signals in Gaussian noise," *IEEE Trans. Signal Processing*, vol. 54, pp. 2964–2975, 2006.

[21] D. R. Morgan and C. K. Madsen, "Wideband system identification using multiple tones with allpass filters and square-law detectors," *IEEE Trans. Circuits Syst. I, Reg. Papers,* vol. 53, pp. 1151–1165, 2006.

[22] T. McWhorter and P. Schreier, "Widely-linear beamforming," in *Proc. 37th Asilomar Conf. Signals, Systems, Comput.*, 2003, pp. 753–759.

[23] P. Chevalier and A. Blin, "Widely linear MVDR beamformers for the reception of an unknown signal corrupted by noncircular interferences," *IEEE Trans. Signal Processing*, vol. 55, pp. 5323–5336, 2007.

[24] P. Chevalier, J. P. Delmas, and A. Oukaci, "Optimal widely linear MVDR beamforming for noncircular signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 3573–3576, Apr. 2009.

[25] S. C. Douglas, "Widely-linear recursive least-squares algorithm for adaptive beamforming," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2009.

[26] L. De Lathauwer and B. De Moor, "On the blind separation of non-circular sources," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Toulouse, France, 2002.

[27] J. Eriksson and V. Koivunen, "Complex random vectors and ICA models: Identifiability, uniqueness and separability," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1017–1029, 2006.

[28] T. Adali, H. Li, M. Novey, and J.-F. Cardoso, "Complex ICA using nonlinear functions," *IEEE Trans. Signal Processing*, vol. 56, no. 9, pp. 4356–4544, Sept. 2008.

[29] E. Ollila and V. Koivunen, "Complex ICA using generalized uncorrelating transform," *Signal Processing*, vol. 89, pp. 365–377, Apr. 2009.

[30] M. Novey and T. Adali, "Complex ICA by negentropy maximization," *IEEE Trans. Neural Networks*, vol. 19, no. 4, pp. 596–609, Apr. 2008.

[31] M. Novey and T. Adali, "On extending the complex FastICA algorithm to noncircular sources," *IEEE Trans. Signal Processing*, vol. 56, no. 5, pp. 2148–2154, Apr. 2008.

[32] X.-L. Li and T. Adali, "Complex independent component analysis by entropy bound minimization," *IEEE Trans. Circuits Syst. I, Reg. Papers,* vol. 57, no. 7, pp. 1417–1430, July 2010.

[33] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Trans. Neural Networks*, vol. 21, no. 2, pp. 248–261, Feb. 2010.

[34] H. Li, N. Correa, P. A. Rodriguez, V. D. Calhoun, and T. Adali, "Application of independent component analysis with adaptive density model to complex-valued fMRI data," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 10, pp. 2794–2803, Oct. 2011.

[35] X.-L. Li and T. Adali, "Complex-valued linear and widely linear filtering using MSE and Gaussian entropy," *IEEE Trans. Signal Processing*, vol. 60, no. 11, pp. 5672–5684, Nov. 2012.

[36] W. M. Brown and R. B. Crane, "Conjugate linear filtering," *IEEE Trans. Inform. Theory*, vol. 15, pp. 462–465, 1969.

[37] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Processing*, vol. 43, pp. 2030–2033, Aug. 1995.

[38] P. O. Amblard, M. Gaeta, and J. L. Lacoume, "Statistics for complex variables and signals–Part 1: Variables," *Signal Processing*, vol. 53, no. 1, pp. 1–13, 1996.

[39] P. O. Amblard, M. Gaeta, and J. L. Lacoume, "Statistics for complex variables and signals–Part 2: Signals," *Signal Processing*, vol. 53, no. 1, pp. 15–25, 1996.

[40] T. Adali and S. Haykin, *Adaptive Signal Processing: Next Generation Solutions*. Hoboken, NJ: Wiley Interscience, 2010.

[41] K. Kreutz-Delgado, "The complex gradient operator and the CR-calculus," Univ. California, San Diego, CA, Tech. Rep. UCSD-ECE275CG-S2009v1.0, 2009.

[42] M. J. Ablowitz and A. S. Fokas, *Complex Variables*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[43] R. Remmert, *Theory of Complex Functions*. Harrisonburg, VA: Springer-Verlag, 1991.

[44] J. Eriksson, E. Ollila, and V. Koivunen, "Essential statistics and tools for complex random variables," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5400–5408, Oct. 2010.

[45] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: SIAM, 2000.

[46] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge, U.K.: Cambridge, 2010.

[47] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1985.

[48] D. R. Morgan, "Adaptive algorithms for a two-channel structure employing allpass filters with applications to polarization mode dispersion compensation," *IEEE Trans. Circuits Syst. I, Reg. Papers,* vol. 51, no. 9, pp. 1837–1847, 2004.

[49] D. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters*. Chippenham, Wiltshire, U.K.: Wiley, 2009.

[50] T. Kim and T. Adali, "Fully complex multi-layer perceptron network for nonlinear signal processing," *J. VLSI Signal Processing Syst. Signal, Image, Video Technol.,* vol. 32, pp. 29–43, Aug.–Sept. 2002.

[51] S. L. Goh and D. P. Mandic, "Stochastic gradient-adaptive complex-valued nonlinear neural adaptive filters with a gradient-adaptive step size," *IEEE Trans. Neural Networks*, vol. 18, pp. 1511–1516, 2007.

[52] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Oct. 2008, Version 20081110.

[53] A. Hjōrungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Processing*, vol. 55, no. 6, pp. 2740–2746, 2007.

[54] A. Hjørungnes, *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[55] H. Li and T. Adali, "Algorithms for complex ML ICA and their stability analysis using Wirtinger calculus," *IEEE Trans. Signal Processing*, vol. 58, no. 12, pp. 6156–6167, Dec. 2010.

[56] B. Loesch and B. Yang, "Cramér–Rao bound for circular and noncircular complex independent component analysis," *IEEE Trans. Signal Processing*, vol. 61, no. 2, pp. 365–379, 2013.

[57] A. van den Bos, "The multivariate complex normal distribution—A generalization," *IEEE Trans. Inform. Theory*, vol. 41, no. 2, pp. 537–539, Mar. 1995.

[58] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Trans. Signal Processing*, vol. 44, no. 10, pp. 2637–2640, Oct.1996.

[59] P. Schreier and L. Scharf, "Second-order analysis of improper complex random vectors and processes," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 714–725, Mar. 2003.

[60] F.D. Neeser and J.L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1293–1302, July 1993.

[61] M. Novey, T. Adali, and A. Roy, "A complex generalized Gaussian distribution—characterization, generation, and estimation," *IEEE Trans. Signal Processing*, vol. 58, no. 3, pp. 1427–1433, Mar. 2010.

[62] R. A. Wooding, "The multivariate distribution of complex normal variables," *Biometrika*, vol. 43, no. 1/2, pp. 212–215, 1956.

[63] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution," *Annals Math. Stats.*, vol. 34, pp. 152–176, 1963.

[64] T. L. Grettenberg, "A representation theorem for complex normal processes," *IEEE Trans. Inform. Theory*, vol. 11, no. 2, pp. 305–306, Apr. 1965.

[65] P. Schreier, L. Scharf, and A. Hanssen, "A generalized likelihood ratio test for impropriety of complex signals," *IEEE Signal Processing Lett.*, vol. 13, no. 7, pp. 433–436, July 2006.

[66] J. P. Delmas and H. Abeida, "Asymptotic distribution of circularity coefficients estimate of complex random variables," *Signal Processing*, vol. 89, no. 12, pp. 2670–2675, 2009.

[67] P. Schreier, "Bounds on the degree of improperiety of complex random vectors," *IEEE Signal Processing Lett.*, vol. 15, pp. 190–193, 2008.

[68] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *IEEE Trans. Signal Processing*, vol. 59, no. 11, pp. 5101–5123, Nov. 2011.

[69] X.-L. Li, T. Adali, and M. Anderson, "Noncircular principal component analysis and its application to model selection," *IEEE Trans. Signal Processing*, vol. 59, no. 10, pp. 4516–4528, Oct. 2011.

[70] E. Ollila and V. Koivunen, "Generalized complex elliptical distributions," in *Proc. 3rd Sensor Array Multichannel Signal Processing Workshop*, Sitges, Spain, July 2004, pp. 460–464.

[71] A. T. Walden and P. Rubin-Delanchy, "On testing for impropriety of complex-valued Gaussian vectors," *IEEE Trans. Signal Processing*, vol. 57, pp. 825–834, Mar. 2009.

[72] E. Ollila, J. Eriksson, and V. Koivunen, "Complex elliptically symmetric random variables—Generation, characterization and circularity tests," *IEEE Trans. Signal Processing*, vol. 59, no. 1, pp. 58–69, 2011.

[73] E. Ollila, V. Koivunen, and H. V. Poor, "A robust estimator and detector of circularity of complex signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* 2011.

[74] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Processing*, vol. 60, pp. 5597–5625, 2012.

[75] S. Huang and C. Li, "Complex-valued adaptive filtering based on the minimization of complex-error entropy," *IEEE Trans. Neural Networks Learning Syst.,* vol. 24, no. 5, pp. 695–708, May 2013.

[76] X.-L. Li and T. Adali, "Independent component analysis by entropy bound minimization," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5151–5164, Oct. 2010.

[77] P. Bouboulis and S. Theodoridis, "Extension of Wirtinger's calculus to reproducing kernel Hilbert spaces and the complex kernel LMS," *IEEE Trans. Signal Processing*, vol. 59, no. 3, pp. 964–978, Mar. 2011.

[78] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[79] T. Kim and T. Adali, "Approximation by fully complex multilayer perceptrons," *Neural Computation*, vol. 15, no. 7, pp. 1641–1666, July 2003.

[80] P. Chevalier, P. Duvaut, and B. Picinbono, "Complex transversal volterra filters optimal for detection and estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Toronto, ON, Canada, 1991, pp. 3537–3540.

[81] A. Hirose, "Continuous complex-valued back-propagation learning," *Electronics Lett.*, vol. 28, no. 20, pp. 1854–1855, 1992.

[82] H. Leung and S. Haykin, "The complex backpropagation algorithm," *IEEE Trans. Signal Processing*, vol. 39, no. 9, pp. 2101–2104, Sept. 1991.

[83] A. Uncini and F. Piazza, "Blind signal processing by complex domain adaptive spline neural networks," *IEEE Trans. Neural Networks*, vol. 14, no. 2, pp. 399–412, 2003.

[84] J. Kivinen, A. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[85] W. Liu, P Pokharel, and J. Principe, "The kernel least-mean-square algorithm," *IEEE Trans. Signal Processing*, vol. 56, no. 2, pp. 543–554, 2008.

[86] P. Bouboulis and S. Theodoridis, "Adaptive learning in complex reproducing kernel Hilbert spaces employing Wirtinger's subgradients," *IEEE Trans. Neural Networks*, vol. 23, no. 3, pp. 425–438, Mar. 2012.

[87] P. Chevalier and B. Picinbono, "Complex linear-quadratic systems for detection and array processing," *IEEE Trans. Signal Processing*, vol. 44, no. 10, pp. 2631–2634, Oct. 1996.

[88] E. Moreau and T. Adali, *Blind Identification and Separation of Complex-valued Signals*. Hoboken, NJ: Wiley, 2013.

[89] X.-L. Li and X.-D. Zhang, "Nonorthogonal joint diagonalization free of degenerate solution," *IEEE Trans. Signal Processing*, vol. 55, no. 5, pp. 1803–1814, May 2007.

[90] M. Anderson, X.-L. Li, P. A. Rodriguez, and T. Adali, "An effective decoupling method for matrix optimization and its application to the ICA problem," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 2012, pp. 1885–1888.

[91] P. Comon, "Independent component analysis, A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[92] E. Ollila, K. Hyon-Jung, and V. Koivunen, "Compact Cramér–Rao bound expression for independent component analysis," *IEEE Trans. Signal Processing*, vol. 56, no. 4, pp. 1421–1428, Apr. 2008.

[93] Z. Koldovský, P. Tichavský, and E. Oja, "Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér–Rao lower bound," *IEEE Trans. Neural Networks*, vol. 17, no. 5, pp. 1265–1277, 2006.

[94] A. Yeredor, "Blind separation of Gaussian sources with general covariance structures: Bounds and optimal estimation," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5057–5068, Oct. 2010.

[95] J.-F. Cardoso, "Likelihood," in *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, P. Comon and C. Jutten, Eds. New York: Academic, 2010, ch. 4, pp. 107–154.

[96] X.-L. Li and T. Adali, "Blind separation of noncircular correlated sources using Gaussian entropy rate," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2969–2975, June 2011.

[97] A. Yeredor, "Performance analysis of the strong uncorrelating transformation in blind separation of complex-valued sources," *IEEE Trans. Signal Processing*, vol. 60, no. 1, pp. 478–483, Jan. 2012.

[98] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Systems*, vol. 10, no. 1, pp. 1–8, Feb. 2000.

[99] J. Anemüller, T. J. Sejnowski, and S. Makeig, "Complex independent component analysis of frequency-domain electroencephalographic data," *Neural Networks*, vol. 16, no. 9, pp. 1311–1323, 2003.

[100] T. Adali, T. Kim, and V. D. Calhoun, "Independent component analysis by complex nonlinearities," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Montreal, QC, Canada, May 2004, vol. V, pp. 525–528.

[101] *MATLAB*, The MathWorks Inc., Natick, Massachusetts, version 7.10.0 (r2010a) edition, 2010.

[102] H. Li and T. Adali, "A class of complex ICA algorithms based on the kurtosis cost function," *IEEE Trans. Neural Networks*, vol. 19, no. 3, pp. 408–420, Apr. 2008.

[103] B. Sällberg, N. Grbić, and I. Claesson, "Complex-valued independent component analysis for online blind speech extraction," *IEEE Trans. Signal Processing*, vol. 16, no. 8, pp. 1624–1632, 2008.

[104] J. Via, D. Palomar, L. Vielva, and I. Santamaria, "Quaternion ICA from second-order statistics," *IEEE Trans. Signal Processing*, vol. 59, no. 4, pp. 1586–1600, 2011.

[SP]

Rishee K. Jain, José M.F. Moura,
and Constantine E. Kontokosta

Big Data + Big Cities: Graph Signals of Urban Air Pollution

For the first time in human history, the majority of the world's inhabitants now reside in cities. Urban inhabitants are expected to account for a staggering 67% of the world's population (6.3 billion people) by 2050 [1]. This enormous migration toward urban environments has brought with it a host of challenges related to sustainability, health, and development. Engineers, scientists, and policy makers must grapple with the daunting task of providing the next generation of urban citizens with such core necessities as clean water, energy, and air.

In parallel, the proliferation of low-cost sensing has led to an explosion of data from the urban-built environment. Large amounts of data at a high degree of spatial granularity and temporal frequency (i.e., big data)—such as energy and water usage, environmental emissions, and human activity—are rapidly becoming available in cities around the world. Urban informatics—applying "big data" analytics to the context of "big cities"—offers an unprecedented opportunity to understand, analyze, and improve how our cities develop and operate. Processing unstructured and high-dimensional data from urban systems will require combining expertise from the fields of signal processing, graph theory, and data science with the application domains of civil engineering, environmental science, and urban planning, among others. In this column, we consider unstructured data sets from the urban-built environment and propose how to represent them as a high-dimensional and geometrically structured graph signal. We illustrate the impact and merits of this approach by applying it to a pertinent sustainability and health issue in New York City—air pollution from the burning of heavy fuel oils for heating and hot water in buildings.

Air pollution from the burning of heavy fuel oil has been shown to have deleterious effects on human health. The combustion of heavy fuel oils results in the release of particulate matter smaller than 2.5 μm ($PM_{2.5}$) and mono-nitrogen oxides (NO_x). Both pollutants have been linked to increased airway inflammation, decreased lung function, and the worsening of asthma leading to a rise in hospital emergency room visits, hospital admissions, and deaths from cardiovascular and respiratory diseases [2], [3]. Recent work [4] has successfully applied methods from data science and machine learning to improve models of urban air quality in Beijing and Shanghai. As a result, policy makers and citizens around the world are increasingly aware of the potential to utilize data-driven methods to understand trends on $PM_{2.5}$ and NO_x emissions. For example, identifying neighborhoods or clusters of buildings with high emissions could have widespread implications on the deployment of air quality sensor infrastructure, the design of targeted programs aimed at accelerating the transition of buildings to cleaner fuels, the formulation of public health initiatives regarding respiratory diseases, and even real estate prices in neighborhoods identified as high or low emitters.

> ENGINEERS, SCIENTISTS, AND POLICY MAKERS MUST GRAPPLE WITH THE DAUNTING TASK OF PROVIDING THE NEXT GENERATION OF URBAN CITIZENS WITH SUCH CORE NECESSITIES AS CLEAN WATER, ENERGY, AND AIR.

In this article, we apply signal processing and data science methodologies to study the environmental impact of burning different types of heating oil in New York City, where currently the burning of heavy fuel oil in buildings produces more annual black carbon, a key component of PM2.5, emissions, than all cars and trucks combined [5]. The data utilized in this article are collected through New York City's Local Law 84 (LL84) energy disclosure mandate [6]. The mandate requires annual energy consumption reporting for large buildings (i.e., approximately greater than 50,000 gross feet) of all use types. This analysis utilizes actual heating oil consumption data for calendar year 2012. The LL84 data set was merged with land use and geographic data at the tax lot level from the Primary Land Use Tax Lot Output (PLUTO) data set from the New York City Department of City Planning. The PLUTO data set provides building and tax lot characteristics, as well as their geographic location.

THE URBAN-BUILT ENVIRONMENT AS GRAPH SIGNALS

Consider a data set where each data element is represented by a building. For each of the N data elements or buildings, we have the corresponding geographic location and subsequently can infer some relational information about the data elements. This information can

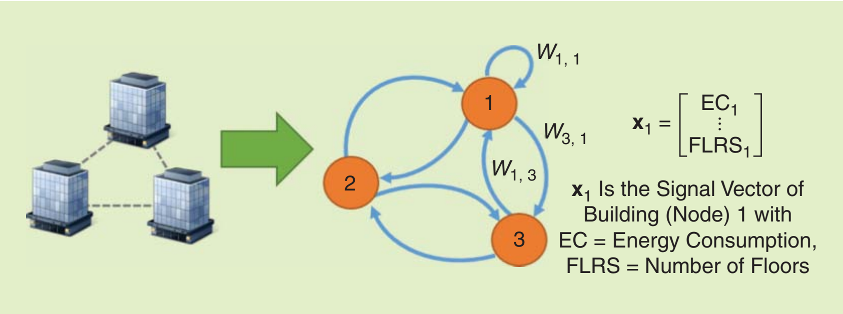
Digital Object Identifier 10.1109/MSP.2014.2330357
Date of publication: 19 August 2014

be represented by a graph $G = (V, \mathrm{W})$, where $V = \{v_0, \dots, v_{N-1}\}$ is the set of nodes and W is the weighted adjacency matrix of the graph. Each data element or, in this case, building corresponds to node v_n. The entry $\mathrm{W}_{\mathrm{i,j}}$ is the weight of a directed edge that reflects the degree of relation of the jth building to the ith building. We define G generally so that it can take the form of an undirected or directed weighted graph. For example, if we define the relation between two buildings as the physical distance between them, then $\mathrm{W}_{\mathrm{i,j}} = \mathrm{W}_{\mathrm{j,i}}$ and G would be an undirected weighted graph. However, the relation between two buildings can also be defined as a function of each building's properties (e.g., building size) causing $\mathrm{W}_{\mathrm{i,j}} \neq \mathrm{W}_{\mathrm{j,i}}$ and G to take the form of a directed weighted graph. In this general model for representing buildings data in a graph, we do not restrict the edge weights of $\mathrm{W}_{\mathrm{i,j}}$ to nonnegative reals, but physical constraints related to the application will likely govern such values. We also define our model to allow self-loops, an edge that connects a node to itself, to account for specific applications related to the built environment (e.g., a building's air pollution would impact itself in addition to neighboring buildings). As a result, $\mathrm{W}_{\mathrm{i,i}} \neq 0$ is a valid entry in our weighted adjacency matrix.

In the context of the built environment, the edge weights will most likely be naturally defined by the application. However, this natural definition may not be apparent during the initial construction of a graph and thus one can define an edge connecting node j to i by using the common thresholded Gaussian kernel weighting function:

$$W_{i,j} = \begin{cases} \exp\left(\dfrac{[\mathrm{dist(i,j)}]^2}{2\theta^2}\right) & \text{if } \mathrm{dist(i,j)} \leq \gamma \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where γ and θ are parameters defined by the user and dist(i,j) represents the physical distance between buildings (nodes) j and i or the Euclidean distance between two feature vectors describing j and i.


$W_{1,1}$
1
$W_{3,1}$
$W_{1,3}$
2
3
$\mathbf{x}_1 = \begin{bmatrix} \mathrm{EC}_1 \\ \vdots \\ \mathrm{FLRS}_1 \end{bmatrix}$
$\mathbf{x}_1$ Is the Signal Vector of Building (Node) 1 with
EC = Energy Consumption,
FLRS = Number of Floors

[FIG1] A diagram of the built environment represented as a graph signal.

We define a graph signal of buildings as $f: V \rightarrow \mathbb{R}$ defined on the nodes of the graph. The signal can be represented as a vector $\mathrm{f} \in \mathbb{R}^N$ where the ith component of the vector f is the signal at the ith building (node) in V. Alternatively, if a building has multiple signals, we can define a vector $\mathrm{x_i}$ comprising all signals related to the ith building where $\mathrm{X} = \{\mathrm{x}_0, \dots, \mathrm{x}_{\mathrm{N}-1}\}$ is the set of signal vectors. For example, $\mathrm{x_i}$ could consist of discrete data on energy, water, and natural gas consumption or data on the physical properties of building i. A diagram of the built environment represented as a graph is provided in Figure 1.

APPLICATION TO AIR POLLUTION IN NEW YORK CITY

ADAPTING THE GENERAL MODEL TO URBAN AIR POLLUTION

To adapt the general model described above to air pollution in New York City, we define the edges between nodes using a modified Gaussian dispersion plume model. The Gaussian dispersion plume model has been utilized in previous work [7] to gain a high-level understanding of urban air pollution dynamics. Assuming the most conservative boundary condition (i.e., perfectly reflective surface), the Gaussian dispersion equation takes the following form:

$$C(x, y, z; H) = \frac{Q}{2\pi \cdot u \cdot \sigma_y \cdot \sigma_z} \cdot \exp\left[-\frac{y^2}{2\sigma_y^2}\right] \cdot \left\{\exp\left[-\frac{(z-H)^2}{2\sigma_z^2}\right] + \exp\left[-\frac{(z+H)^2}{2\sigma_z^2}\right]\right\}, \tag{2}$$

where C is the mean concentration (in g/m^3), x is the distance from the sources in the direction of the wind (in m), y is the cross wind distance from the source (in m), z is the vertical height from the ground (in m), H is the effective height above ground of where the pollutant is being released (in m), u is the wind speed (in m/s), Q is the strength of the emission source (in g/s), σ_y is the urban dispersion parameter (i.e., the standard deviation of the emission distribution, a function of stability class and x) in the horizontal direction, and σ_z is the urban dispersion parameter in the vertical direction.

We aim to use the Gaussian dispersion model as a basis for gaining a simple understanding of differences in air pollution across an urban area by analyzing data from several thousand buildings. It should be noted that our goal is not to specifically quantify the concentration of a pollutant at a given location but to derive data-driven estimates of where air pollution is highest in an urban area. Obtaining an accurate quantification of concentration would require the development of a much more complex and input intensive computational fluid dynamic (CFD) model for the area surrounding each building. We make several simplifying assumptions in our analysis: the wind speed u is assumed to be constant across the study area in a single direction (west to east), $z = H = 0$ meaning that we assume emissions and exposure to occur at the ground level with no dispersion in the z direction (i.e., σ_z is a constant), and Q is a static value of the total emissions released (in our case a function of kBTU) and not a rate. As a result, C is not the

[TABLE 1] THE WEIGHTING VALUES FOR EACH POLLUTANT RELATIVE TO HEATING OIL #6 (SOURCE: [10]).

HEATING OIL TYPE	$PM_{2.5}$	NO_X
#6	1	1
#4	0.53	0.675
#2	0.06	0.32

concentration but a representative value of emissions (a function of kBTU). Removing the constants $(u, 2\pi, \sigma_z)$ and applying our assumptions to (2), the Gaussian plume equation takes the following reduced form:

$$C = \frac{Q}{\sigma_y} \cdot \exp\left[-\frac{y^2}{2\sigma_y^2}\right], \qquad (3)$$

where the urban dispersion parameter σ_y takes the following form based on [8] for New York City (classified as Pasquill atmospheric stability urban class C; see [9] for details on Pasquill stability classes):

$$\sigma_y = 0.22x(1 + 0.0004x)^{-1/2}. \qquad (4)$$

We translate (3) into a graph and signal form by breaking up the Gaussian dispersion component and the strength of the emission source Q into a matrix A and signal vectors $\mathbf{q}$ in the set $Q = \{\mathbf{q}_0, \ldots, \mathbf{q}_{N-1}\}$. Matrix A is defined as:

$$A_{i,j} = \begin{cases} \dfrac{1}{\sigma_y}\exp\left(-\dfrac{[y_{i,j}]^2}{2\sigma_y^2}\right) & \text{if } y_{i,j} \geq \gamma \\ 0 & \text{otherwise,} \end{cases} \qquad (5)$$

where $y_{i,j}$ is the Euclidean physical distance between building j and building i. The threshold γ is set as 2.54E-10 [derived from setting x and y to 1,500 m in (3) and (4)], and we assign one to the diagonals of matrix A to account for the impact of a building's own emissions on itself (self-loop). The signal vector $\mathbf{q}$ is defined as

$$q_i = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}_i, \qquad (6)$$

where $\mathbf{h}_i$ is a 3×1 vector corresponding to building i's consumption of heating oil #2, heating oil #4, and heating oil #6 where $\mathrm{H} = \{\mathbf{h}_0, \ldots, \mathbf{h}_{N-1}\}$ is the set of consumption vectors. Each type of heating oil contributes to air pollution by a different magnitude based on the pollutant being examined (i.e., $PM_{2.5}$ or NO_X). Therefore, $\mathbf{h}_i$ is multiplied by the transpose of a 3×1 weighting vector $\boldsymbol{\lambda}$. Table 1 provides the weighting values for each pollutant relative to heating oil #6. Intuitively, this means that burning one kBTU equivalent of heating oil #4 emits 47% less $PM_{2.5}$ and 32.5% less NO_x than burning one kBTU of heating oil #6.

Finally, we obtain the weighted directed adjacency matrix as follows from A and $\mathbf{q}$:

$$W_{i,j} = A_{i,j}q_j. \qquad (7)$$

GRAPH METRICS AND COMMUNITY DETECTION

A large amount of information can be abstracted from a graph by examining the properties of its nodes and edges. We define several common graph metrics as they relate to the study of urban air pollution in Table 2.

Additionally, we are interested in identifying clusters or communities of buildings with large amounts of emissions. Identifying such "hot spots" of emissions are valuable for data-driven deployment of air quality sensors and the development of policy measures to reduce building emissions. For our analysis, we employ a commonly used community detection algorithm, the Louvain method [12]. Community detection differs from other methods of graph partitioning in that the experimenter does not make any a priori assumptions on the number and size of the partitions [11]. Thus, community detection algorithms are able to infer the natural substructure of the graph.

The Louvain method is a heuristic method based on modularity maximization. Modularity is a measure of the density of links inside communities as compared to the links between communities [11] and can be derived as follows. Let us first formulate in mathematical terms the value of edges that run between nodes in the same community:

$$\frac{1}{2}\sum\nolimits_{ij} W_{i,j}\delta(c_i, c_j), \qquad (8)$$

where W is the weighted adjacency matrix of an undirected graph, c_i represents the community to which node i is assigned, and function δ is 1 if $c_i = c_j$ and 0 otherwise.

Next, we formulate a mathematical expression for the expected value of edges between all pairs of nodes in the same community if connections are made purely at random:

$$\frac{1}{2}\sum\nolimits_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j), \qquad (9)$$

where $k_i = \sum_j W_{i,j}$ is the sum of all the weights of the edges attached to node i (i.e., degree of i) and m is the total value of edges in the graph.

Taking the difference between (8) and (9), we obtain the mathematical expression [given in (10)] for the values of edges that are present in a community beyond what would be expected at random, a quantity known as modularity (M) [11]:

$$M = \frac{1}{2m}\sum\nolimits_{ij}\left(W_{i,j} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j). \qquad (10)$$

The Louvain algorithm optimizes modularity across a network in two phases. First, the algorithm optimizes modularity by allowing only local changes of communities. Second, the algorithm aggregates communities to build a new network of supercommunities and optimizes modularity again. These two phases are repeated iteratively until an increase in modularity is no longer observed. The method returns subgraphs of G such that a subgraph $S \subseteq G$. More details regarding the Louvain method can be found in [12].

To apply the Louvain method for community detection, we translate our graph from directed to undirected by mapping the signal vector $\mathbf{q}$ to an $n \times n$ symmetric matrix Q as:

$$q_i \rightarrow \mathrm{Q}_{i,j} = \mathrm{Q}_{j,i} = \boldsymbol{\lambda}^{\mathrm{T}}\mathrm{Z}(\mathbf{h}_i, \mathbf{h}_j), \qquad (11)$$

where $\boldsymbol{\lambda}$ is the 3×1 weighting matrix [as defined in (6)] and Z is a proportional weighting function defined as:

$$\mathrm{Z}(\mathbf{h}_i, \mathbf{h}_j) = \left[\frac{h_{i1}^2 + h_{j1}^2}{h_{i1} + h_{j1}} \quad \frac{h_{i2}^2 + h_{j2}^2}{h_{i2} + h_{j2}} \quad \frac{h_{i3}^2 + h_{j3}^2}{h_{i3} + h_{j3}}\right], \qquad (12)$$

[TABLE 2] THE COMMON GRAPH METRICS APPLIED TO URBAN AIR POLLUTION.

GRAPH METRIC	FORMAL DEFINITION (DEFINITIONS ADAPTED FROM [11])	MATHEMATICAL FORMULATION	APPLICATION TO URBAN AIR POLLUTION
SELF-LOOP OF A VERTEX	THE VALUE OF AN EDGE CONNECTING A VERTEX TO ITSELF.	$W_{i,j}$ WHERE $i = j$	A MEASURE OF THE EMISSIONS OF BUILDING i.
IN-DEGREE OF A VERTEX	THE VALUE OF ALL IN-GOING EDGES CONNECTED TO A VERTEX ON A DIRECTED GRAPH.	$k_i^{\text{in}} = \sum_{j=1}^{N} W_{i,j}$	A MEASURE OF HOW MUCH NEIGHBORING BUILDINGS ARE CONTRIBUTING TO BUILDING i'S AIR QUALITY.
OUT-DEGREE OF A VERTEX	THE VALUE OF ALL OUT-GOING EDGES CONNECTED TO A VERTEX ON A DIRECTED GRAPH.	$k_j^{\text{out}} = \sum_{i=1}^{N} W_{i,j}$	A MEASURE OF HOW MUCH BUILDING j IS CONTRIBUTING TO THE AIR QUALITY OF ITS NEIGHBORS.

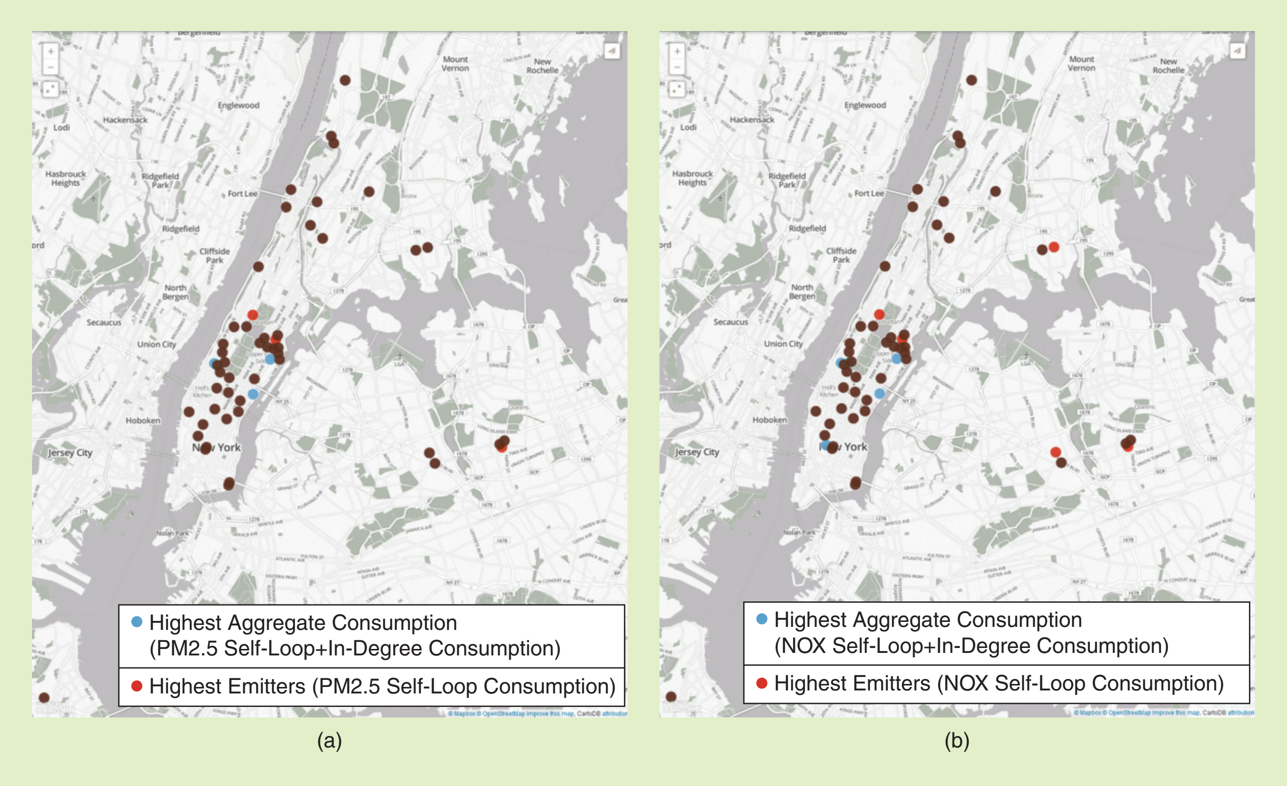
Note: For the specific application of air pollution, we deviate from the standard definition of in-degree and out-degree by excluding the contribution of a self-loop $(W_{i,j})$, where $i = j$ since we are trying to ascertain the impact neighboring buildings have on building i's air quality or building j has on the air quality of neighboring buildings.

where h_{i1}, h_{i2}, h_{i3} correspond to building i's consumption of heating oil #2, heating oil #4, and heating oil #6 and h_{j1}, h_{j2}, h_{j3} correspond to building j's consumption of heating oil #2, heating oil #4, and heating oil #6, respectively.

Additionally, to translate matrix A into an $n \times n$ symmetric matrix, we superimpose two symmetrical wind directions (i.e., west to east, east to west).

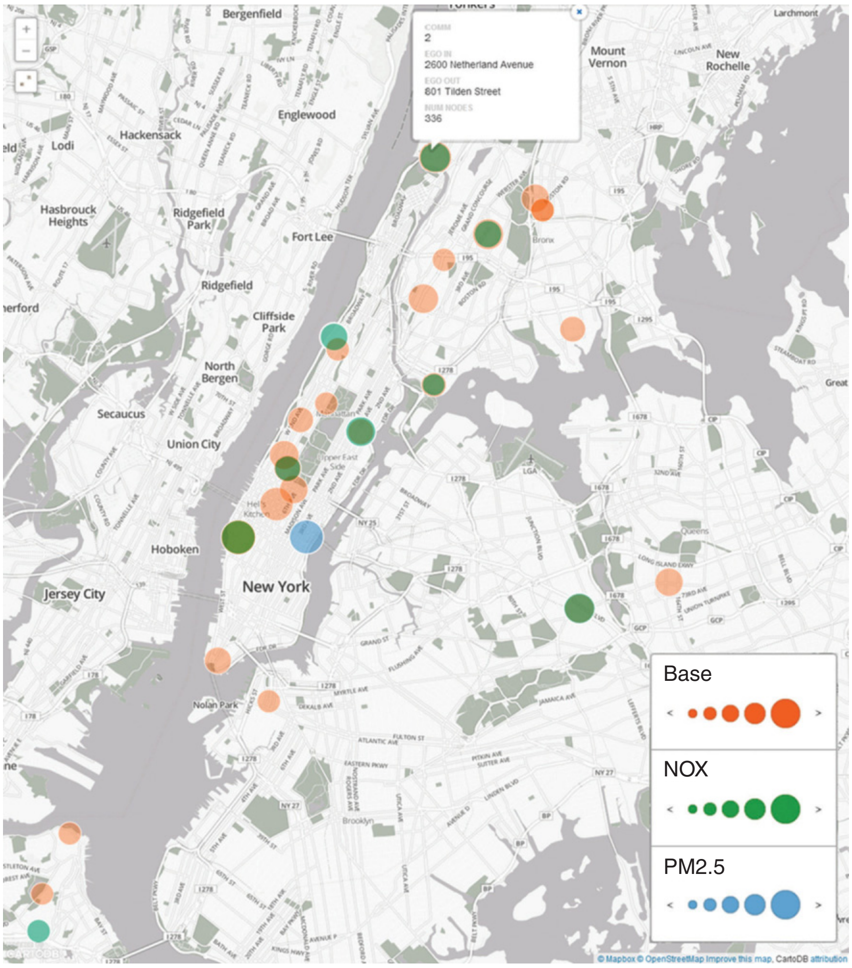
DESCRIPTION OF NEW YORK CITY DATA SET

The NYC LL84 and PLUTO data sets were merged on Borough Block Lot (BBL) numbers, unique identifiers used by the City of New York to track tax lot parcels, to form a composite data set of heating oil consumption and associated geographic location of buildings covered by the energy disclosure mandate. A conversion process was undertaken using CORPSCON [13], an open-source coordinate conversion program from the U.S. Army Corps of Engineers that can batch convert coordinates between map projections, to convert data in the New York–Long Island State Plane Coordinate System to corresponding latitude and longitude values. An initial preprocessing step was conducted on the composite data set


Englewood
Hackensack
Lodi
Hasbrouck Heights
Ridgefield Park
Fort Lee
Ridgefield
Cliffside Park
North Bergen
Secaucus
Union City
Hoboken
Jersey City
New York
Mount Vernon
New Rochelle
● Highest Aggregate Consumption (PM2.5 Self-Loop+In-Degree Consumption)
● Highest Emitters (PM2.5 Self-Loop Consumption)
(a)
● Highest Aggregate Consumption (NOX Self-Loop+In-Degree Consumption)
● Highest Emitters (NOX Self-Loop Consumption)
(b)

[FIG2] Visualization of the buildings estimated to be exposed to high levels of (a) $PM_{2.5}$ and (b) NO_x. The top 50 emitting buildings are indicated by red markers and the top 50 buildings exposed to the highest aggregate consumption by blue markers. Red markers are determined by taking the nodes with highest weighted heating oil consumption. Blue markers are determined by taking the nodes with the highest combined self-loop and in-degree consumption. Dark red markers indicate where a red and blue markers overlap.

to remove duplicate data points and data points that were incomplete or contained missing information (i.e., energy usage, square footage, geographic information). A secondary preprocessing step was conducted to identify and remove erroneous (i.e., energy usage exorbitantly too high or too low) and outlier data points (i.e., top/bottom 1% of energy usage). Both preprocessing steps are consistent with the data cleaning methodology established by the City of New York in their annual report regarding the LL84 energy disclosure data [14]. The postprocessed data set consisted of 11,196 valid data points and represented nearly 2 billion gross ft^2 with an average building size of 173,707 ft^2. Seventy-six percent of the data points correspond to multifamily residential buildings, and 11% correspond to commercial office buildings. The remaining percent of buildings have a multitude of uses (e.g., retail, hotel education). The geographic distribution of the data points across the five New York City boroughs are as follows: 44% in Manhattan, 17% in the Bronx, 18% in Brooklyn, 19% in Queens, and 2% in Staten Island. The geographic bias toward Manhattan is expected as Manhattan contains the bulk of large buildings subject to the reporting requirements of the disclosure mandate. We acknowledge this geographic bias as a limitation of our analysis and aim to mitigate this issue in future work by incorporating other disparate data sets on smaller buildings in New York City. A subset of the overall data set (4,702 data points, 42% of the total) accounted for over 27.5 billion kBTU of heating oil consumption in the 2012 calendar year with an average consumption of 5.8 million kBTU per building.



[FIG3] **Visualization of building clusters that form "hot spots" of pollution. The base case does not weigh consumption by pollutant and is in orange. PM$_{2.5}$ and NO$_x$ are represented by blue and green markers, respectively. The size of each circle marker is indicative of the number of buildings in each community with the center located at the geographic coordinates of the "Ego In" node (i.e., the building with the highest in-degree plus self-loop value). Each cluster's information box provides: the "Ego In" node's address, the "Ego Out" (i.e., the building with the highest out-degree value indicating that it significantly contributes to the poor air quality of its neighboring buildings) node's address and the total number of nodes in the cluster that burn heating oil.**

ANALYSIS, RESULTS, AND IMPLICATIONS

All analysis was conducted using NetworkX [15], an open-source Python language software package for the creation, manipulation, and analysis of complex graphs. Results were visualized using CartoDB [16], an online visualization tool for geotagged data.

IDENTIFYING BUILDINGS MOST EXPOSED TO POLLUTION

We aim to illustrate the benefits of our graph-based approach by identifying specific buildings in New York City that are susceptible to high levels of PM$_{2.5}$ and NO$_x$ pollution. For comparison, we employed both a conventional analysis method and a method derived from representing the data as a graph signal. The conventional method consisted of ranking the buildings by their weighted heating oil consumption (q_i) to determine the top emitters for each pollutant. The second method utilizes the graph structure of our model to quantify and rank the combined impact a building's own heating oil consumption and the consumption of its neighbors has on surrounding air quality. In graph terms, this quantity is calculated by summing the in-degree and the self-loop for each vertex (as defined in Table 2). A visualization of the results for both methods is presented in Figure 2 for PM$_{2.5}$ and NO$_x$.

Significant overlap exists between the two analysis methods as indicated by the dark red markers in Figure 2. As expected, several buildings that are the highest emitters for both PM$_{2.5}$ and NO$_x$ also have the highest aggregated consumption, indicative of buildings most exposed to air pollutants. However, a discrepancy is also apparent between the two methods. Several buildings are identified to be high emitters (red markers), but the air pollution around the building may not be an issue given that surrounding buildings are not contributing significant amounts of pollution. Conversely, the conventional method fails to identity several buildings in Manhattan (blue markers) where the combination of a building's own emissions and those of its neighbors together are indicative of locations where the surrounding air quality maybe poor. By not taking into account the geometric structure of the data, the conventional method fails to utilize all of the available information and as a result may not provide a complete picture of what buildings may be the most susceptible to high levels of $PM_{2.5}$ and NO_x. Identifying buildings that are exposed to high levels of pollutants can allow policy makers to develop targeted measures aimed at reducing heating oil consumption in specific properties. Additionally, such information could also be valuable to public health workers aiming to understand and reduce respiratory diseases. Future analysis could be undertaken to explore health issues and hospitalization rates surrounding the identified buildings and utilized to inform public health policy.

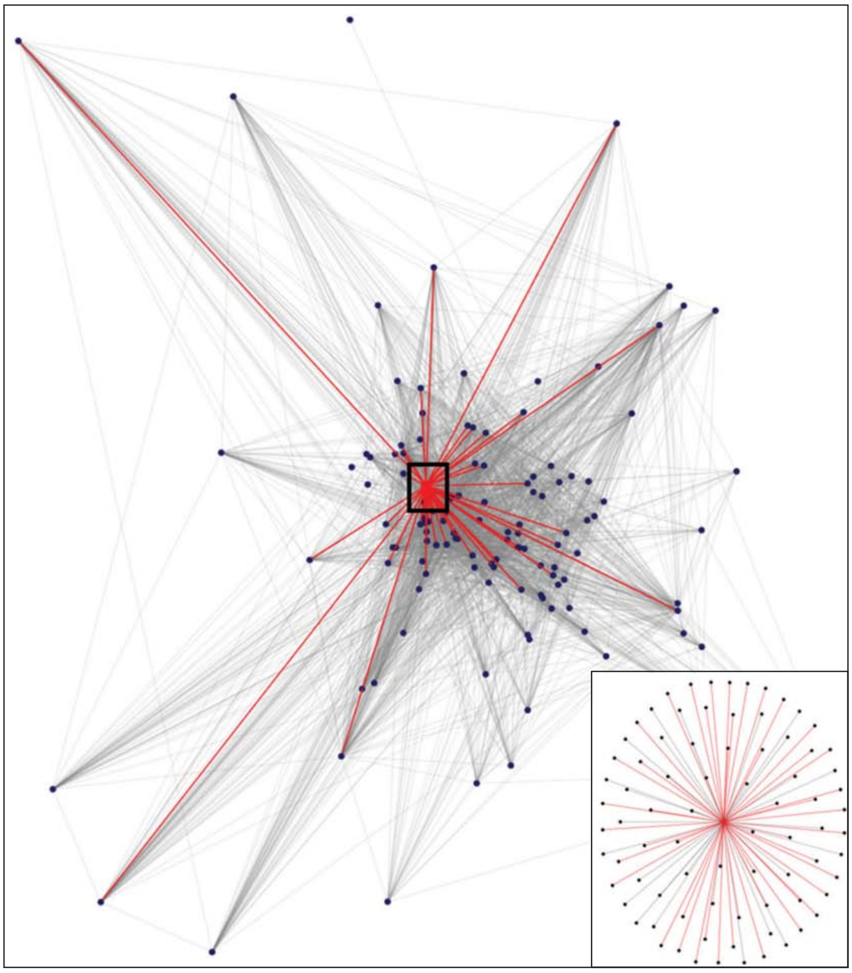
IDENTIFYING "HOT SPOTS" OF AIR POLLUTION

We extend the previous analysis to demonstrate how our graph-based approach can be used to identify "hot spots" of air pollution. As described in a previous section, the graph partitioning method employed makes no a priori assumptions on the structure of the graph and therefore allows the natural structure of clusters to emerge from the data. A visualization of the results for a base case (all pollutant weights set equal to one) for $PM_{2.5}$ and NO_x levels is provided in Figure 3 and a representative network diagram for a sample cluster is provided in Figure 4.

Clusters of $PM_{2.5}$ and NO_x pollution are consistent and the largest clusters are present in midtown Manhattan, the Upper East Side, and northern Manhattan/Bronx. The base case clusters follow a similar pattern, but not weighting the consumption based on fuel oil type and pollutant is seen to shift the center of clusters in many areas, such as northern Manhattan/Bronx. While the results are not surprising given the geographic distribution of building types and building age across New York City neighborhoods, this approach provides an alternative method based on point source consumption and emissions data to corroborate traditional air quality monitoring and modeling studies [18].

WE APPLY SIGNAL PROCESSING AND DATA SCIENCE METHODOLOGIES TO STUDY THE ENVIRONMENTAL IMPACT OF BURNING DIFFERENT TYPES OF HEATING OIL IN NEW YORK CITY.

The graph-based approach also allows us to deepen our analysis and abstract additional information on each cluster of buildings including the location of the "Ego In" and "Ego Out" nodes. The "Ego In" node is the building with the highest in-degree plus self-loop value and therefore estimated to be where the concentration of pollutants is the highest. Identifying where air pollution is expected to be the worst in each cluster of buildings could be utilized for intelligent and data-driven positioning of air quality monitoring equipment. Previous research [19] has found that pollution hot spots in urban areas maybe inaccurately characterized



[FIG4] Visualization of a sample cluster located in the Manhattan borough of New York City consisting of 127 nodes and 5,702 edges (2,942 visible). The large red marker and edges indicate the "Ego In" node (i.e., the building with the highest in-degree plus self-loop value within the cluster) and its edges. The inset visualizes the "Ego In" node and all 80 of its edges with gray edges representing a connection not visible in the main visualization. The visualization was created using NodeXL [17].

due to inappropriate positioning of air quality monitoring equipment. Given that air quality monitoring equipment is often deployed based on site availability rather than the need for measurement at that particular location, a data-driven approach could drastically improve the reliability of air quality models and enhance our understanding of urban air pollution dynamics. The "Ego Out" node is the building with the highest out-degree value indicating that it significantly contributes to the poor air quality of its neighboring buildings. Identifying and disseminating the "worst neighbor" buildings in communities could impact the real-estate market (e.g., such information could alter a buyer's decision to purchase in a particular building) and even the social dynamics driving the adoption of cleaner heating technologies (e.g., a property owner could be incentivized to adopt a cleaner heating technology to avoid social scrutiny from occupants of surrounding buildings). In particular, social norms have been observed to have an impact on other environmental behavior, such as energy consumption [20], [21]. Thus, similar social dynamics could be utilized to accelerate the adoption and penetration of clean heating systems in New York City.

TOWARD A DATA-DRIVEN URBAN ENVIRONMENT

This article represents data from the urban built environment as a high-dimensional and geometrically structured graph signal. We demonstrate the merits of this approach by applying it to the issue of air quality in New York City and illustrate how the geometric structure of the data can be utilized to abstract valuable information for both urban citizens and policy makers. This work represents an important first step in rethinking how we structure and analyze data from the urban built environment and could be expanded in numerous ways, including: relaxing the assumption that an urban environment is flat by incorporating dispersion along the z-axis, supplementing the current data set with information on smaller buildings not captured by the current energy disclosure mandate, incorporating data from additional sources (e.g., social media, health-care informatics) to observe the interdynamics between air pollution and human health or behavior, and applying new methods from the emerging field of signal processing on graphs (e.g., [22] and [23]) to further deepen our analysis.

More importantly, this article contributes to the literature at the intersection of big data and urban environments (i.e., "urban informatics") and aims to catalyze future research on how urban data can be collected, processed, represented, and analyzed to make our cities more sustainable. Moving toward a more data-driven urban environment will provide an enormous opportunity to not just accommodate but enhance the lives of the world's urban inhabitants.

ACKNOWLEDGMENTS

We would like to thank the New York City Mayor's Office of Long-Term Planning and Sustainability for providing the data for this study. José M.F. Moura was partially supported by the Air Force Office of Scientific Research through grant number FA95501210087.

AUTHORS

Rishee K. Jain (rishee.jain@nyu.edu) is the director's postdoctoral fellow at the Center for Urban Science and Progress, New York University. His research interests encompass urban infrastructure systems, smart buildings, and data science.

José M.F. Moura (moura@ece.cmu.edu) is the Philip and Marsha Dowd University Professor in the Department of Electrical and Computer Engineering at Carnegie Mellon University and is currently a visiting professor, on sabbatical leave, at the Center for Urban Science and Progress, New York University. He is a member of the National Academy of Engineering, a corresponding member of the Portugal Academy of Science, an IEEE Fellow, and a fellow of the AAAS.

Constantine E. Kontokosta (ckontokosta@nyu.edu) is the deputy director of the Center for Urban Science and Progress, the founding director of the Center for the Sustainable Built Environment, and an associate research professor in the Department of Civil and Urban Engineering, New York University.

REFERENCES

[1] United Nations Department of Economic and Social Affairs, "World urbanization prospects, the 2011 revision," NY, 2011.

[2] U.S. Environmental Protection Agency, "Integrated science assessment for particulate matter," Research Triangle Park, NC, Rep. EPA/600/R-08/139F, 2009.

[3] A. G. Cornell, S. N. Chillrud, R. B. Mellins, L. M. Acosta, R. L. Miller, J. W. Quinn, B. Yan, A. Divjan, O. E. Olmedo, S. Lopez-Pintado, P. L. Kinney, F. P. Perera, J. S. Jacobson, I. F. Goldstein, A. G. Rundle, and M. S. Perzanowski, "Domestic airborne black carbon and exhaled nitric oxide in children in NYC," *J. Expo. Sci. Environ. Epidemiol.*, vol. 22, no. 3, pp. 258–266, 2012.

[4] Y. Zheng, F. Liu, and H. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2013, pp. 1436–1444.

[5] The City of New York, "City of New York press release," PR 309-12, 2012.

[6] C. Kontokosta, "Energy disclosure, market behavior, and the building data ecosystem," *Ann. N. Y. Acad. Sci.*, vol. 1295, pp. 34–43, Aug. 2013.

[7] E. Gilmore, L. Lave, and P. Adams, "The costs, air quality, and human health effects of meeting peak electricity demand with installed backup generators," *Environ. Sci. Technol.*, vol. 40, no. 22, pp. 6887–6893, 2006.

[8] G. Briggs, "Diffusion estimation for small emissions," NOAA Rep. ATDL-106, 1973.

[9] D. Turner, *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling*. Boca Raton, FL: CRC, 1994.

[10] Environmental Defense Fund, "The bottom of the barrel: How the dirtiest heating oil pollutes our air and harms our health," NY, 2009.

[11] M. Newman, *Networks: An Introduction*, 1st ed. London, U.K.: Oxford Univ. Press, 2010, 720 pp.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, no. 10, pp. 1–12, Oct. 2008.

[13] U.S. Army Corp of Engineers, "CORPSCON."

[14] The City of New York, "New York City local law 84 benchmarking report—August 2012," NY, 2013.

[15] A. Hagberg, P. Swart, and D. Chult, "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python in Science Conf. (SciPy2008)*, pp. 11–15.

[16] Vizzuality. (2014, May 5). CartoDB. [Online]. Available: http://cartodb.com/

[17] M. Smith, N. Milic-Fraying, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec, and C. Dunne, *NodeXL: A Free and Open Network Overview, Discovery and Exploration Add-in for Excel 2007/2010*. Social Media Research Foundation, 2010.

[18] New York City Department of Health and Mental Hygiene, "The New York City Community Air Survey. Results from years one and two: December 2008-December 2010," NY, 2012.

[19] S. Vardoulakis, E. Solazzo, and J. Lumbreras, "Intra-urban and street scale variability of BTEX, NO2 and O3 in Birmingham, UK: Implications for exposure assessment," *Atmos. Environ.*, vol. 45, no. 29, pp. 5069–5078, Sept. 2011.

[20] H. Allcott, "Social norms and energy conservation," *J. Public Econ.*, vol. 95, no. 9–10, pp. 1082–1095, Oct. 2011.

[21] R. K. Jain, R. Gulbinas, J. E. Taylor, and P. J. Culligan, "Can social influence drive energy savings? Detecting the impact of social influence on the energy consumption behavior of networked users exposed to normative eco-feedback," *Energy Build.*, vol. 66, pp. 119–127, Nov. 2013.

[22] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[23] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Processing*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

Anthony G. Christodoulou,
Peter Kellman, and Zhi-Pei Liang

Accelerating Cardiovascular Magnetic Resonance Imaging: Signal Processing Meets Nuclear Spins

Cardiovascular diseases are still the leading cause of death worldwide, accounting for an estimated 30% of all deaths across the globe, more than cancer, injury, and HIV/AIDS combined (Figure 1). Efforts to address cardiovascular disease with technology can be traced back nearly 200 years to the invention of the stethoscope in 1816. The many successful technological advances since then have significantly transformed the detection, diagnosis, and treatment of cardiovascular diseases over the last two centuries.

Cardiovascular imaging technology has enabled measurement and visualization of the structure and function of the beating heart and has become an indispensable part of cardiac health care. A number of cardiac imaging modalities are available to the new generation of cardiologists: cardiac ultrasound, known as echocardiography (ECHO), and X-ray computed tomography (CT) are typically used to image cardiac structure; positron emission tomography (PET) and single photon emission computed tomography (SPECT) are typically used to image cardiac function. Magnetic resonance imaging (MRI), the topic of this column, is suitable for both structural and functional imaging (Figure 2).

CARDIOVASCULAR MRI

Because of the particular properties of the magnetic resonance (MR) phenomenon, MRI has unique potential for cardiovascular imaging. MRI is already the gold standard modality for cardiac chamber anatomy and function, detection and assessment of myocardial infarction, evaluation of congenital heart disease, and more [1]. Further advances in myocardial perfusion imaging, blood flow velocity imaging, spectroscopic imaging, and other applications have brought MRI closer to achieving its potential as the premier all-around imaging modality for cardiologists, although technological challenges still exist for each of these applications.

The primary challenge facing cardiovascular MRI is imaging speed. There is a fundamental tradeoff between the temporal resolution and spatial resolution of MRI, both of which are major concerns when imaging the beating heart. Conventionally, the sampling requirements for MRI are governed by the Nyquist–Shannon theorem. The earliest methods to accelerate MRI were fast-scanning methods focused on manipulating nuclear spins for fast data acquisition, all within the Nyquist–Shannon framework. Fast-scanning technology is now a relatively mature area of research, giving way to solutions which leverage different signal processing frameworks for sub-Nyquist imaging within the sampling constraints of nuclear spin physics.

OVERCOMING THE NYQUIST BARRIER

In cardiovascular MRI, the desired spatiotemporal image $\rho(\mathbf{r},t)$ is related to the signal $\{d_q(\mathbf{k},t)\}_{q=1}^{Q}$ from an array of Q receive coils as

$$d_q(\mathbf{k},t) = \int_{-\infty}^{\infty} S_q(\mathbf{r})\rho(\mathbf{r},t)e^{-i2\pi\mathbf{k}\cdot\mathbf{r}}d\mathbf{r}, \quad (1)$$

where $S_q(\mathbf{r})$ represents the spatial sensitivity of the qth receive coil. For conventional image reconstruction, the coil sensitivities are absorbed into the desired image function for each coil, $\rho_q(\mathbf{r},t) = S_q(\mathbf{r})\rho(\mathbf{r},t)$, which are then independently reconstructed from the Nyquist-sampled $(\mathbf{k},t)$-space data via the inverse Fourier transform. The resulting coil images are combined to form the final reconstructed image.

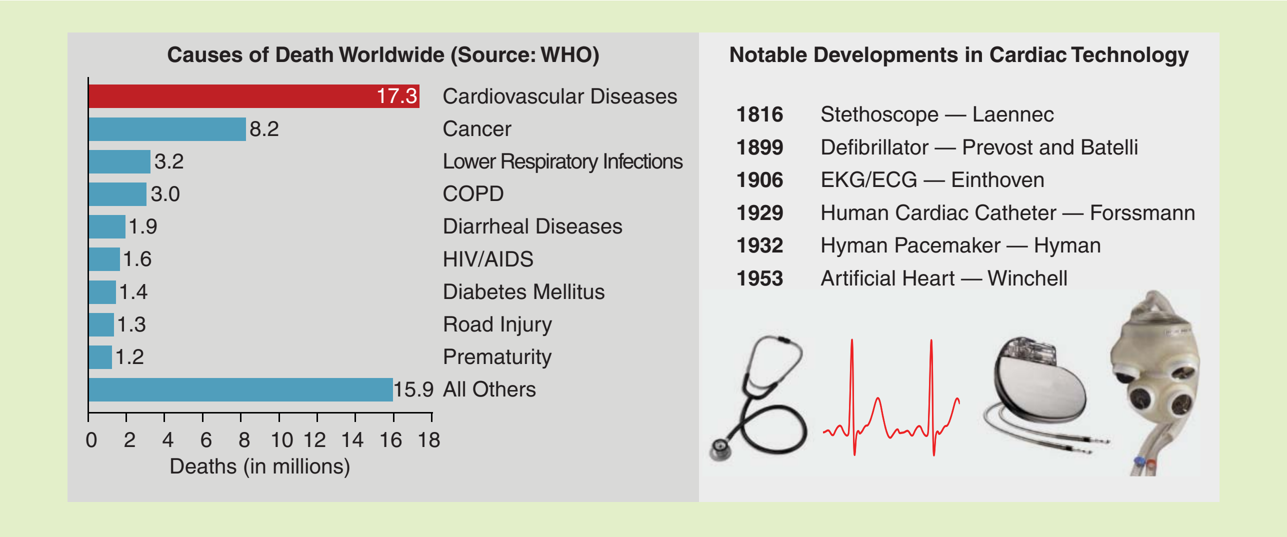
PARALLEL IMAGING

Parallel imaging utilizes the additional encoding power of the receive coil sensitivities $\{S_q(\mathbf{r})\}_{q=1}^{Q}$ to solve the reconstruction problem from the sub-Nyquist data. This approach is an ingenious application of Papoulis' multichannel sampling theorem to MRI [2]. It is well known that, under certain conditions, a signal that is bandlimited in either time or space can be exactly recovered from sub-Nyquist measurements of the signal from multiple sensors. More specifically, consider that $\rho(\mathbf{r},t)$ is spatially bandlimited to the region $[-B/2, B/2]$ and $\{d_q(\mathbf{k},t)\}_{q=1}^{Q}$ are the outputs from a bank of Q linear and $\mathbf{k}$-shift-invariant filters $\{S_q(\mathbf{r})\}_{q=1}^{Q}$. Papoulis' multichannel sampling theorem states that $d(\mathbf{k},t) = \int_{-\infty}^{\infty} \rho(\mathbf{r},t)e^{-i2\pi\mathbf{k}\cdot\mathbf{r}}d\mathbf{r}$, and therefore $\rho(\mathbf{r},t)$, can be recovered from complex samples of $\{d_q(\mathbf{k},t)\}_{q=1}^{Q}$ taken at rate $\Delta\hat{k} = Q/B$ (i.e., a factor of Q above the Nyquist rate $\Delta k = 1/B$) using interpolation kernels $\{g_q(\mathbf{k})\}_{q=1}^{Q}$ derived from $\{S_q(\mathbf{r})\}_{q=1}^{Q}$.

In the situation with known $\{S_q(\mathbf{r})\}_{q=1}^{Q}$, $\rho(\mathbf{r},t)$ can be recovered in image space by inverting (1) (e.g., sensitivity encoding for fast MRI (SENSE) [3]). When $\{S_q(\mathbf{r})\}_{q=1}^{Q}$ are unknown, $\rho(\mathbf{r},t)$ can be recovered using $\mathbf{k}$-space interpolation kernels (analogous to the $\{g_q(\mathbf{k})\}_{q=1}^{Q}$ in Papoulis' sampling theorem), which are learned from auxiliary data (e.g., generalized autocalibrating partially parallel acquisitions (GRAPPA) [4]). Although Papoulis' multichannel sampling framework permits acceleration factors up to Q,

Digital Object Identifier 10.1109/MSP.2014.2328503
Date of publication: 19 August 2014

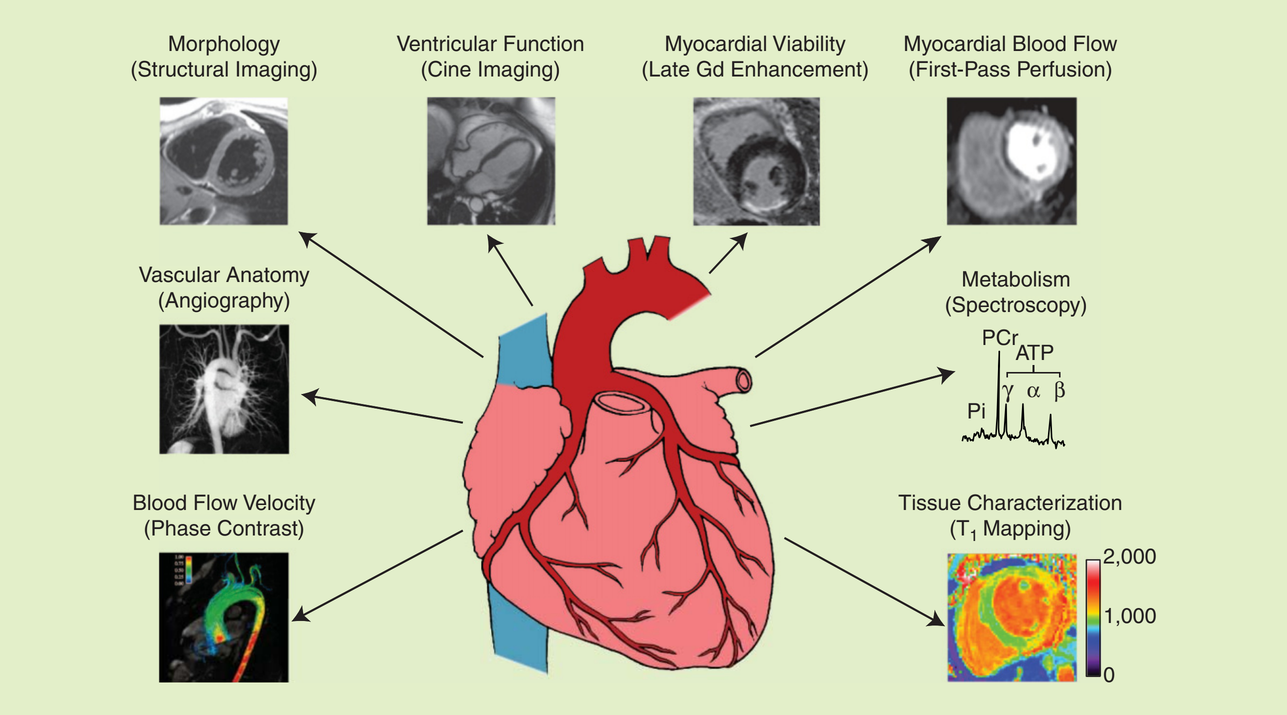
1053-5888/14©2014IEEE



Notable Developments in Cardiac Technology

1816	Stethoscope — Laennec
1899	Defibrillator — Prevost and Batelli
1906	EKG/ECG — Einthoven
1929	Human Cardiac Catheter — Forssmann
1932	Hyman Pacemaker — Hyman
1953	Artificial Heart — Winchell

[FIG1] Notable facts about cardiovascular diseases and cardiac technologies.


Morphology (Structural Imaging)
Ventricular Function (Cine Imaging)
Myocardial Viability (Late Gd Enhancement)
Myocardial Blood Flow (First-Pass Perfusion)
Vascular Anatomy (Angiography)
Metabolism (Spectroscopy)
PCr
ATP
γ α β
Pi
Blood Flow Velocity (Phase Contrast)
Tissue Characterization (T_1 Mapping)
2,000
1,000
0

[FIG2] Applications of cardiovascular MRI.

measurement noise, auxiliary data acquisition, and ill-conditioning of the reconstruction problem limit the practically achievable acceleration factor. As a result, acceleration factors well below Q are applied in practice. To achieve greater acceleration, parallel imaging is often applied jointly with complementary acceleration approaches such as compressed sensing (CS) and/or subspace imaging.

COMPRESSED SENSING

CS theory enables recovery of sparse signals from sub-Nyquist measurements and has found important application in MRI, especially cardiac MRI. After discretizing and vectorizing $\rho(\mathbf{r}, t)$ and $d(\mathbf{k}, t)$ as $\boldsymbol{\rho}$ and $\mathbf{d}$, respectively, the data acquisition equation (1) can be formulated as $\mathbf{A}\boldsymbol{\rho} = \mathbf{d}$. Under the CS theory, $\boldsymbol{\rho}$ can be recovered from $\mathbf{d}$ by minimizing $\|\mathbf{T}\boldsymbol{\rho}\|_1$ subject to $\mathbf{A}\boldsymbol{\rho} = \mathbf{d}$ or $\|\mathbf{d} - \mathbf{A}\boldsymbol{\rho}\|_2^2 < \epsilon$ (in the case with noise), where $\mathbf{T}$ is a sparsifying transform.

Cardiovascular images are sparse in a number of transform domains [5], including the $(\mathbf{r}, f)$-space (spatial-spectral), wavelet-spectral, or spatiotemporal finite-difference domains. $\mathbf{T}$ is commonly chosen to transform the image vector $\boldsymbol{\rho}$ into one of these domains. Randomly ordered $(\mathbf{k}, t)$-space sampling is also used for CS MRI, as it generally results in a sampling basis A which is incoherent with any of the previously mentioned sparse bases. Image reconstruction can be performed by solving an unconstrained optimization problem:

$$\arg\min_{\boldsymbol{\rho}} \| \mathbf{d} - \mathbf{A}\boldsymbol{\rho} \|_2^2 + \lambda \| \mathbf{T}\boldsymbol{\rho} \|_1,$$

where λ is a regularization parameter.

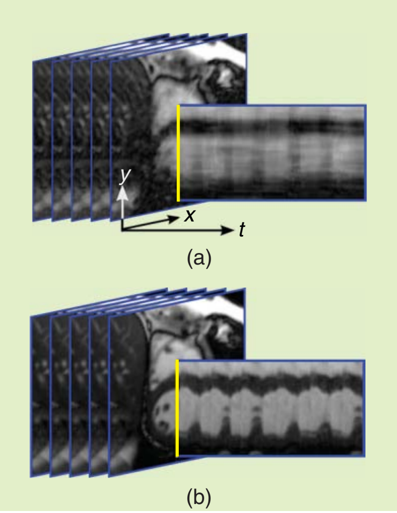
Like parallel imaging, CS is an effective strategy to accelerate cardiovascular MRI and is most effective when jointly used with parallel imaging and/or subspace imaging.

SUBSPACE IMAGING

Subspace imaging exploits the fact that cardiovascular signals have a high degree of spatiotemporal correlation (or reside in a low-dimensional subspace). More specifically, the spatiotemporal changes of cardiac MR data can be expressed as $d(\mathbf{k},t) = \sum_{\ell=1}^{L} u_\ell(\mathbf{k}) v_\ell(t)$ [6]. In other words, $d(\mathbf{k},t)$ is Lth-order partially separable. It can be shown that the Casorati matrix $\mathbf{C}$ formed with elements $C_{ij} = d(\mathbf{k}_i, t_j)$ has a rank no more than L. This property enables recovery of $d(\mathbf{k},t)$ (or the missing entries of $\mathbf{C}$) from highly undersampled measurements by imposing a rank or subspace constraint.

MR cardiac signals are highly correlated and as a result, the separation rank L is rather low (around 32). In addition, special data acquisition schemes can be implemented, which acquire at least L rows of $\mathbf{C}$ in full and sparsely sample the remaining rows of $\mathbf{C}$. This allows the temporal subspace to be predetermined and the rank-constrained matrix completion problem to be solved with this known subspace, which significantly simplifies the subspace imaging reconstruction problem.

Subspace imaging provides a powerful tool to accelerate cardiovascular MRI. It produces best results when jointly imposed alongside parallel imaging and/or CS, as demonstrated in Figure 3.



[FIG3] An illustration of sub-Nyquist cardiac MRI. (a) Nearest-neighbor temporal interpolation of the (*k,t*)-space data demonstrates the low temporal sampling rate of MRI as well as the resulting spatiotemporal artifacts and blurring. (b) Reconstruction of the same data using parallel imaging, CS, and subspace imaging shows the power of accelerated imaging. Images are shown stacked along the time dimension, and spatiotemporal slices show the temporal profiles over the yellow lines.

APPLICATIONS

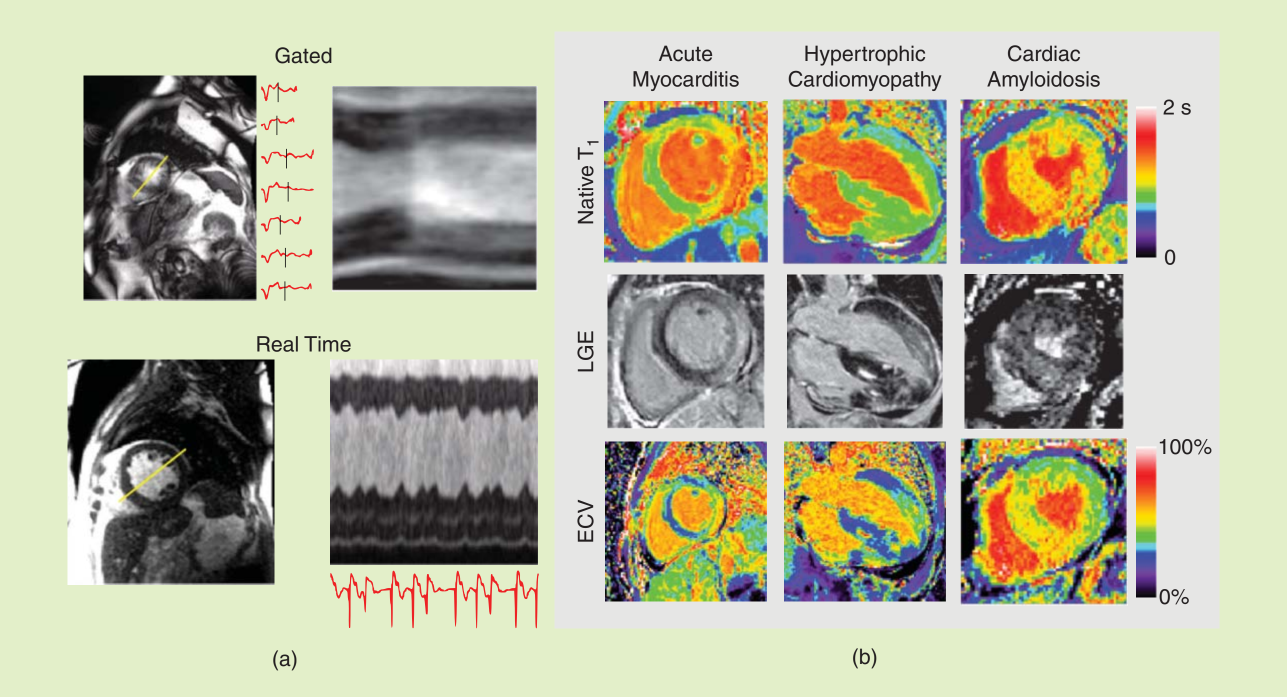
CINE IMAGING

Dynamic cine image sequences depict the structure and function of the heart, including the mechanical contraction, timing, and extent of wall motion and thickening, as well as the function of valves. From these images, it is possible to perform a multitude of cardiac assessments. Global measures such as cardiac mass, blood volume, and ejection fraction can be measured from time-resolved images at different cardiac phases. Regional wall motion may be used to determine and localize abnormal tissue function: akinetic regions of the myocardium (i.e., the cardiac wall) can be well visualized, helping to determine the extent of injury to the myocardium. Functional cine imaging may augment morphological imaging to better assess complex structural abnormalities and congenital heart defects by visualizing the motion of the blood and valves. Cine imaging may also be used to assess the mechanical activation of the heart, which may be important in understanding arrhythmias and in guiding treatment. When used in conjunction with contrast enhanced viability imaging, it may further be used to distinguish irreversibly damaged myocardium from stunned myocardium after ischemia.

The cornerstone of cine imaging is cardiac motion; however, it is challenging to acquire high spatial resolution images quickly enough to resolve the motion of the heart. For this reason, "gated" methods are commonly employed to utilize data acquired across multiple heartbeats, with the underlying assumption that each heartbeat is the same, i.e., that $\rho(\mathbf{r},t)$ is periodic. This is achieved by using the electrocardiogram (ECG) as a reference signal and instructing the subject to hold his or her breath; the data from multiple heartbeats are then combined to reconstruct a single representative heartbeat. However, many patients are unable to hold their breath adequately or may have variations in their heart rhythms that violate the assumption of a stationary (i.e., periodic) heart, leading to poor image quality using gated methods. For this reason, it is often preferable to use accelerated methods that can produce high spatial-resolution images quickly enough to resolve cardiac and respiratory motions without resorting to ECG triggering or breath-holding. These accelerated methods are referred to as *real-time* imaging methods. Figure 4(a) shows a comparison between gated and real-time imaging on patients with atrial fibrillation.

Advanced image reconstruction methods that use signal processing to permit rapid imaging and fill in the missing data from undersampled acquisitions are routinely applied to cardiac functional cine. These methods are used to reduce the breath-hold duration for gated, segmented scans to several heartbeats, as well as for real-time imaging. Indeed, cine imaging has advanced to the point where it is now possible to image two-dimensional (2-D) slices of the heart at 1.0 mm in-plane spatial resolution and 20 frames per second (fps), using hybrid fast-scanning, parallel imaging, CS, and subspace imaging methods [7].

Methods that can acquire time-resolved three-dimensional (3-D) volumes have potential to greatly simplify the workflow and improve the analysis of cardiac function. Three-dimensional methods require an even higher degree of acceleration and



[FIG4] **Example images from different applications of cardiac MRI. (a) Examples of gated and real-time imaging of patients with irregular heartbeats. Individual heartbeats do not match up when "stacked" as in gated imaging, producing artifacts. Real-time imaging is fast enough to image each heartbeat individually, avoiding these artifacts. (b) Examples of both qualitative (LGE) and quantitative [Native T_1 and extracellular volume (ECV) fraction] cardiovascular MR images. Quantitative imaging has advantages over qualitative imaging when the disease is globally diffuse, which is more difficult to discern using qualitative imaging (as in the cardiac amyloidosis example).**

are a subject of active investigation. There is also demand for even higher spatiotemporal resolution: submillimeter resolution to capture detail of small structures such as atria, coronary arteries, or thin walls such as the right ventricle, and temporal resolution on the order of 10–20 ms for the assessment of diastolic function [8]. Validating the fidelity of advanced image reconstruction methods is an important area of research.

VIABILITY IMAGING

Late gadolinium enhancement (LGE) imaging (also known as *delayed enhancement imaging*) is used to assess the viability of myocardial tissue (i.e., whether the tissue is dead or alive). The heart is typically imaged 10–20 min after the administration of gadolinium-based contrast agent into the blood stream. Gadolinium contrast agents shorten the spin-lattice relaxation time constant T_1 (a key mechanism in the contrast of ρ), boosting the signal when using T_1-weighted imaging and therefore brightening voxels in which the contrast agent is concentrated. After a period of time following the administration of the gadolinium based contrast agent, concentration is higher in fibrous scar tissue than in normal myocardium, since the contrast agent in that tissue washes out at a slower rate. With T_1-weighted sequences such as inversion recovery, the normal myocardium appears dark and scar tissue appears bright, leading to positive contrast. LGE has become the gold standard for viability imaging.

MRI IS ALREADY THE GOLD STANDARD MODALITY FOR CARDIAC CHAMBER ANATOMY AND FUNCTION, DETECTION AND ASSESSMENT OF MYOCARDIAL INFARCTION, EVALUATION OF CONGENITAL HEART DISEASE, AND MORE.

To measure enough data (lines of k-space) to achieve the desired spatial resolution, it is customary to acquire data over multiple heartbeats in a gated, segmented fashion. This approach presumes that the subject has a stable heart period and is able to reliably hold their breath, but it is difficult (or for sicker subjects, impossible) to fulfill this requirement. Accelerated parallel imaging may be used to acquire LGE images in a single heartbeat [9]. Using this approach, the patients may breathe freely, and imaging is not sensitive to arrhythmias.

TISSUE CHARACTERIZATION

The physics of MR provides a rich set of contrasts useful in characterizing tissue and answering a number of clinical questions. Various contrasts such as T_1, T_2, or T_2* may be achieved by varying the pulse sequence used in data acquisition, revealing characteristics of the local chemical environment. For instance, it is possible to image and quantify water, fat, and iron content, which can be used to diagnose and differentiate disease. While qualitative T_1-weighted imaging may reveal regional differences in the T_1 of tissue (such as the elevated T_1 due to edema in acute myocarditis), it is more challenging to detect when there is a global shift in T_1. In this instance, there will be no regional differences or spatial contrast observed. To detect diseases that result in a global abnormality (i.e., a uniform contrast change), it is required to quantify the actual value of T_1 or other parameters (e.g., T_2 or T_2*).

It is possible to quantify these parameters by collecting multiple images with different parameter weightings to generate parametric maps. These maps have proven useful in cases of globally diffuse disease processes such as fibrosis and edema [10]. Figure 4(b) shows some of the advantages of quantitative imaging over qualitative imaging, particularly for the cardiac amyloidosis example, wherein amyloids have globally infiltrated the myocardium and therefore do not exhibit the local enhancements required to allow detection via qualitative imaging. Quantitative MRI is more objective and provides a means to perform serial measurements which may be used to evaluate the effectiveness of therapies in the long term.

Parametric mapping places additional demands on accelerated imaging to achieve the desired image quality and spatiotemporal resolution in the presence of motion. For example, in creating a pixelwise map of the time constant T_1, images are acquired at varying delays following an inversion or saturation of the magnetization, and measurements of the signal at each pixel are fit to an exponential recovery curve. Quantitative measurements have great potential for disease detection but increase the demand for reliable and validated image reconstruction methods to achieve the desired accuracy and precision.

MYOCARDIAL PERFUSION IMAGING

Myocardial perfusion imaging measures blood flow through the myocardium to detect coronary artery disease. Imaging is performed to measure the wash-in and wash-out during the first passage of a bolus of gadolinium-based contrast agent. Regions with normal flow will appear brighter than regions with reduced flow (signal intensity is proportional to the concentration of contrast agent on a T_1-weighted image). Perfusion measurements can then be extracted from the signal intensity curve $\rho(\mathbf{r}_0, t)$ for any voxel $\mathbf{r}_0$ inside the myocardium. Myocardial perfusion contrast dynamics are transient, so real-time imaging must be performed with adequate temporal resolution to freeze cardiac motion. Spatial coverage of the heart is achieved by imaging several 2-D slices in rapid succession or by a highly accelerated 3-D volumetric acquisition. Myocardial perfusion imaging is commonly performed under both stress (wherein the patient is generally administered a pharmacological stress agent) and at rest. Imaging during stress presents additional challenges due to increased heart rates and the subsequent requirement for even faster imaging. Myocardial perfusion imaging after exercise-induced stress (such as after a treadmill session) is even more challenging due to the need to image within seconds of reaching peak stress.

[METHODS THAT CAN ACQUIRE TIME-RESOLVED THREE-DIMENSIONAL VOLUMES HAVE POTENTIAL TO GREATLY SIMPLIFY THE WORKFLOW AND IMPROVE THE ANALYSIS OF CARDIAC FUNCTION.]

Research in myocardial perfusion imaging is largely focused on increased speed and spatial coverage using highly undersampled acquisitions and advanced reconstruction. Although MR images are conventionally acquired by sampling **k**-space on a Cartesian grid, new approaches to highly accelerated myocardial perfusion imaging have explored non-Cartesian sampling patterns such as radial or spiral **k**-space trajectories [11]. Non-Cartesian trajectories also have the potential to achieve full 3-D coverage using CS reconstruction [12]. Respiratory motion correction via image registration allows for free breathing during the acquisition of first-pass contrast-enhanced images, and advanced image reconstruction methods that incorporate motion correction directly into the image reconstruction problem have been demonstrated. Myocardial perfusion is an active area of research with goals of achieving reliable, artifact free, high spatial resolution imaging and providing fully quantitative measurement of myocardial blood flow and flow reserve.

PHASE CONTRAST VELOCITY MAPPING

Magnetic field gradients can also manipulate nuclear spins to encode the blood flow velocity in the phase of the complex image $\rho(\mathbf{r}, t)$. Velocity encoding in a single carefully chosen direction can provide flow measurements for targeted areas; velocity encoding in three directions can provide a vector field showing the path and speed of blood flow through all of the imaged chambers and vessels. This vector field can then be used to identify forward, regurgitant, and shunt flows and potentially be used to measure flow pressure as well as shear stress on the vessel wall.

Velocity encoding in multiple directions cannot be performed simultaneously; the data from each velocity encoding direction are collected separately, resulting in a threefold loss of temporal resolution. Phase contrast (PC) velocity mapping has been performed in 2-D without the use of ECG gating through a combination of fast-scanning, parallel imaging, and sparse sampling, achieving 1.8 mm in-plane spatial resolution at 23 fps [13], but the additional acceleration requirements of PC MRI have ensured that 3-D PC MRI is still solidly dominated by gated techniques [14]. Opportunities exist to accelerate 3-D PC MRI to the point where ECG gating is no longer required.

SPECTROSCOPIC IMAGING

MR spectroscopic imaging (MRSI) collects a nuclear MR (NMR) chemical spectrum for each voxel, adding yet another image dimension. For example, ^{1}H spectroscopic imaging separates the hydrogen signals in water, fat, creatine, lactate, etc., from each other, generating images for each molecule. Phosphorus (^{31}P) imaging is even more useful, allowing monitoring of cardiac metabolism by isolating phosphocreatine (PCr), inorganic phosphate (Pi), and the phosphate groups in adenosine diphosphate (ADP) and adenosine triphosphate (ATP). Sodium (^{23}Na) imaging also has potential to assess extracellular volume (ECV) without the need for contrast agents due to the elevated sodium levels in myocardial scars.

The low abundance of spins in metabolites ensures that MRSI is extremely signal starved. Coupled with the sampling difficulties brought on by the additional imaging dimension, real-time dynamic cardiac MRSI has yet to be demonstrated. Static cardiac MRSI methods have addressed the signal-to-noise ratio problem by using large voxel sizes on the order of 20 mm and by averaging the signal from many different acquisitions over the course of 30 min or more [15]. The ability to perform high-resolution cine MRSI would represent a major step forward in cardiac imaging.

OUTLOOK

Cardiovascular MRI has come a long way since its inception and has had important clinical impact. Technological advancements based on signal processing for reconstruction of sparsely sampled data are making important impacts in many areas, including cardiovascular stress imaging, pediatric imaging of congenital heart disease, quantitative tissue characterization, vessel wall imaging, and image-guided therapeutic procedures. Quantitative imaging has the potential for earlier and more reliable detection of diseases. Advances in spectroscopic imaging could allow myocardial viability and ECV assessments without the use of contrast agents. Improvements in rapid imaging may be used to streamline the imaging workflow and reduce the cost of studies. Rapid 3-D whole-heart imaging will streamline the workflow even further by eliminating the need for scout scans and scan plane localization. These advances are quickly moving us toward an all-free-breathing paradigm for whole-heart cardiac MRI, allowing shorter scans with increased patient comfort.

Comprehensive physiological imaging is also an exciting possibility, which would provide metabolic and biochemical information about cardiac tissues in a wide number of conditions: hypertensive, valvular, and ischemic heart diseases, heart failure, cardiac transplantation, and cardiomyopathies. The fusion of MRI with other imaging modalities such as PET can leverage the various strengths of different modalities to provide even more physiological information.

Tracking and fully utilizing this physiological information will require advances in cardiovascular health informatics, a hugely important topic related to cardiac MRI as well as the broader signal processing community. Leveraging MRI's capability to acquire structural, functional, and physiological information of the heart may allow development of personalized computational models of the cardiovascular system. When combined with advances in health informatics, longitudinal personalized models could be retrieved and updated from any properly equipped clinical facility. These models, coupled with other real-time sensory data such as ECG signals, would not only quantitatively assess the current state of the heart, but could also be used to design personalized treatment (e.g., cardiac implants, pharmacotherapy) or even to predict future cardiac events.

AUTHORS

Anthony G. Christodoulou (christo8@illinois.edu) is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign.

Peter Kellman (kellmanp@nhlbi.nih.gov) is a staff scientist with the National Heart Lung and Blood Institute, National Institutes of Health.

Zhi-Pei Liang (z-liang@illinois.edu) is Franklin W. Woeltge Professor of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign.

REFERENCES

[1] D. J. Pennell, U. P. Sechtem, C. B. Higgins, W. J. Manning, G. M. Pohost, F. E. Rademakers, A. C. van Rossum, L. J. Shaw, and E. K. Yucel, "Clinical indications for cardiovascular magnetic resonance (CMR): Consensus Panel report," *Eur. Heart J.*, vol. 25, pp. 1940–1965, Nov. 2004.

[2] L. Ying and Z.-P. Liang, "Parallel MRI using phased array coils," *IEEE Signal Processing Mag.*, vol. 27, pp. 90–98, July 2010.

[3] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, pp. 952–962, Nov. 1999.

[4] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Magn. Reson. Med.*, vol. 47, pp. 1202–1210, June 2002.

[5] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Mag.*, vol. 25, pp. 72–82, Mar. 2008.

[6] Z.-P. Liang, "Spatiotemporal imaging with partially separable functions," in *Proc. IEEE Int. Symp. Biomedical Imaging*, 2007, pp. 988–991.

[7] A. G. Christodoulou, H. Zhang, B. Zhao, T. K. Hitchens, C. Ho, and Z.-P. Liang, "High-resolution cardiovascular MRI by integrating parallel imaging with low-rank and sparse modeling," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 11, pp. 3083–3092, Nov. 2013.

[8] R. Krishnamurthy, A. Pednekar, B. Cheong, and R. Muthupillai, "High temporal resolution SSFP cine MRI for estimation of left ventricular diastolic parameters," *J. Magn. Reson. Imaging*, vol. 31, no. 4, pp. 872–880, 2010.

[9] P. Kellman and A. E. Arai, "Cardiac imaging techniques for physicians: Late enhancement," *J. Magn. Reson. Imaging*, vol. 36, pp. 529–42, Sept. 2012.

[10] P. Kellman, J. R. Wilson, H. Xue, M. Ugander, and A. E. Arai, "Extracellular volume fraction mapping in the myocardium, part 1: Evaluation of an automated method," *J. Cardiovasc. Magn. Reson.*, vol. 14, no. 63, Sept. 2012.

[11] M. Salerno, C. Sica, C. M. Kramer, and C. H. Meyer, "Improved first-pass spiral myocardial perfusion imaging with variable density trajectories," *Magn. Reson. Med.*, vol. 20, pp. 976–989, Dec. 2013.

[12] E. V. R. DiBella, L. Chen, M. C. Schabel, G. Adluru, and C. J. McGann, "Myocardial perfusion acquisition without magnetization preparation or gating," *Magn. Reson. Med.*, vol. 67, no. 3, pp. 609–613, Mar. 2012.

[13] A. A. Joseph, K.-D. Merboldt, D. Voit, S. Zhang, M. Uecker, J. Lotz, and J. Frahm, "Real-time phase-contrast MRI of cardiovascular blood flow using undersampled radial fast low-angle shot and nonlinear inverse reconstruction," *NMR Biomed.*, vol. 25, no. 7, pp. 917–924, July 2012.

[14] M. Markl, P. J. Kilner, and T. Ebbers, "Comprehensive 4D velocity mapping of the heart and great vessels by cardiovascular magnetic resonance," *J. Cardiovasc. Magn. Reson.*, vol. 13, no. 7, Jan. 2011.

[15] C. T. Rodgers, W. T. Clarke, C. Snyder, J. T. Vaughan, S. Neubauer, and M. D. Robson, "Human cardiac ^{31}P magnetic resonance spectroscopy at 7 tesla," *Magn. Reson. Med.* [Online]. Available: http://dx.doi.org/10.1002/mrm.24922

[SP]

Guangtao Zhai and Xiaolin Wu

Multiuser Collaborative Viewport via Temporal Psychovisual Modulation

Consider a multiuser visualization scenario. When making or evaluating emergency response plans for a large city, professionals from police, fire, civil defense, health, environment, transportation, and social services departments meet and discuss. As emergency management involves coordinated activities of various stakeholders, all participants desire to have visual representation of location-sensitive data on a common, integrated display. Separate displays for different types of data cause semantic fragmentation; as one's eyes switch between displays to associate related information, mental transformation in cognitive psychology has to take place, reducing an individual's performance on the task. Moreover, separate personal displays create a feeling of isolation from others and hinders face-to-face communication.

While sharing the same physical display, different experts may need to independently consult specialty maps of their own disciplines (e.g., political, topographic, hydrological, geological, atmospheric, seismic, underground utility, satellite images, etc.) without distracting others. The underground maze of water and sewage pipes, electricity and telecommunication lines, etc., may appear perfectly clear and legible to someone in charge of public utilities but bewildering to an ambulance dispatcher. In other words, the optimal level of details varies from user to user and task to task. Therefore, clutter-free presentation of complex geodata in the above case—or any type of big data in general—to a team of collaborating users requires a single physical display to generate concurrent multiple visuals tailored to different viewers. The previously mentioned collaborative multiuser visualization can be facilitated by an information display technology called *temporal psychovisual modulation* (*TPVM*) [1], which can generate a number of interference-free visuals on a common exhibition medium.

BACKGROUND

Multiple exhibitions on a lone display (MELD) refers to the display capability of concurrently generating multiple individual-tailored interference-free views on a common physical medium. Conceptually, displaying multiple images on a common optoelectronic medium surface can be considered as a problem of two-dimensional (2-D) optical communication.

THE TWO STRAIGHTFORWARD APPROACHES FOR MELD ARE SPACE MULTIPLEXING AND TIME MULTIPLEXING OF LIGHT SIGNALS EMITTED FROM THE DISPLAY SURFACE.

The two straightforward approaches for MELD are space multiplexing and time multiplexing of light signals emitted from the display surface (a rectangular array of light transmitters). The classic multiview display techniques of lenticular lens [2] and parallel barriers [3] are the approach of space multiplexing. A more sophisticated space multiplexing-based MELD technique is random hole display [4]. The weakness of space multiplexing methods is that the spatial resolution of displayed image for each viewer is reduced and the image quality rapidly deteriorates as more participants join the session. The user experience is further compromised by the fact that the image quality of these displays varies in viewers positions, with only a few so-called sweet spots. The SecondLight system [5], the ThirdEye system [6], and the dual-view display of Sony Corporation [7] are methods of simple time multiplexing. Time multiplexing-based MELD can only be achieved on displays of high refresh rate, and it is a terrible waste of available optical bandwidth; a time multiplexing display has to run at 60 KHz to support K users. Moreover, the light influx for each viewer drops as the number of concurrent views increases, further limiting the number of users served. In contrast, the new TPVM information display paradigm published in the January 2013 issue of *IEEE Signal Processing Magazine* [1] is ideally suited to achieve the MELD functionality; it can exhibit a much larger number of concurrent views than space and time multiplexing without suffering from loss of spatial resolution or depletion of light influx.

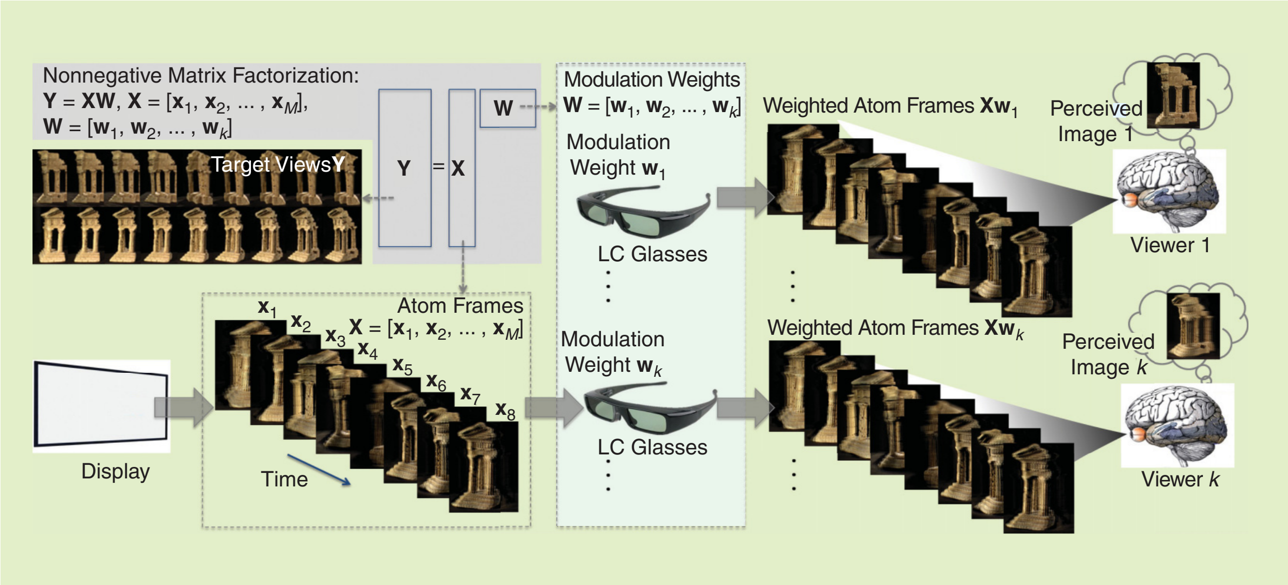
MELD SYSTEM BASED ON TEMPORAL PSYCHOVISUAL MODULATION

Unlike conventional displays, in TPVM, each output frame alone is generally not a complete image but rather a so-called atom frame, which is meant to be linearly combined with other atom frames to form different concurrent images all on the same display medium. As long as these atom frames are refreshed at a speed higher than 60 Hz, the critical

Digital Object Identifier 10.1109/MSP.2014.2328506
Date of publication: 19 August 2014

1053-5888/14©2014IEEE

From Imagination to Market



[FIG1] Image formation by temporal psychovisual modulation (TPVM). The basis images and modulation vectors are computed from non-negative matrix factorization. $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)$ are K target images to be concurrently displayed to different viewers. The $N \times K$ matrix $\mathbf{Y}$, where N is the number of pixels in each target image, is decomposed into $\mathbf{Y} = \mathbf{XW}$, with the $N \times M$ matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$ being the set of atom frames and the $M \times K$ matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K)$ being the K modulation coefficient vectors corresponding to the K target images. The atom frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ are cyclically displayed and temporally modulated by active LC glasses according to weights $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$. The optoelectronic display-glass coupling and the psychovisual temporal fusion mechanism of HVS jointly render the K concurrent target images as different linear combinations of the atom frames.

flicker frequency, the human visual system (HVS) cannot distinguish individual basis frames but rather psychovisually fuse them into an image. Therefore, if the temporal psychovisual fusion process can be manipulated, then different images can be formed out of the same set of atom frames for different viewers. The control of temporal psychovisual fusion can be obtained by placing a display-synchronized light amplitude modulator between a viewer's eyes and the display medium.

Figure 1 is a schematic depiction of a TPVM-based MELD display system, in which active liquid crystal (LC) glasses play the role of the display-synchronized light amplitude modulator. The LC glasses, if synchronized with the high-speed display, can regulate how much of the light energy of each atom frame to pass through and reach retina, particularly, perform amplitude modulation of rapidly fired atom frames. At the heart of the multiuser display system is a problem of nonnegative matrix factorization (NMF) [8]. Let $Y = (y_1, y_2, \ldots, y_K)$ be the K target images to be concurrently displayed to different viewers. The $N \times K$ matrix Y, where N is the number of pixels in each target image, needs to be decomposed into $Y = XW$, with the $N \times M$ matrix $X = (x_1, x_2, \ldots, x_M)$ being the set of atom frames and the $M \times K$ matrix $W = (w_1, w_2, \ldots, w_K)$ being the K modulation coefficient vectors corresponding to the K target images. The resulting atom frames $x_1, x_2, \ldots, x_M$ are cyclically displayed at a refresh rate above $60M$ Hz, and the corresponding 2-D optical signals are temporally modulated by active LC glasses according to weights $w_1, w_2, \ldots, w_K$. This optoelectronic display-glass coupling and the psychovisual temporal fusion mechanism of HVS jointly render the K concurrent target images $y_1, y_2, \ldots, y_K$ as different linear combinations of the $x_1, x_2, \ldots, x_M$ atom frames.

In practice, the image decomposition underlying TPVM has to respect a condition of nonnegativity because the light energy emitted by the display cannot be negative, and active LC glasses can only implement modulation weights between 0 and 1. Therefore, the introduced MELD display system needs to solve the following problem of NMF:

$$\min_{X,W} \| Y - XW \|_F^2, \qquad \text{subject to} \quad 0 \le X, W \le 1, \qquad (1)$$

with $Y \in \Re^{N \times K}, X \in \Re^{N \times M}, W \in \Re^{M \times K}$, and where $\| \cdot \|_F$ is the Frobenius norm; $\le$ operates on each element of the matrices.

In the MELD display system, all K target views y_k, $1 \le k \le K$, which are generated through synchronized active LC glasses performing light amplitude modulation, are called *personal views* in a multiuser environment, as they are intended to provide individualized visual experience. However, in concurrence with these personal views, yet another image will also result for those viewers who use no light modulation devices. This is the view that the HVS forms by fusing all unattenuated atom frames displayed in rapid succession, i.e., $y_0 = x_1 + x_2 + \ldots + x_M$; y_0 is the default image seen by all viewers without using personalized viewing devices. In most multiuser applications, the shared view y_0 should be semantically meaningful and visually pleasing in coexistence of intended personal views on the same display medium. Once given in conjunction of personal views in a multiuser setting, $\mathbf{y}_0$ needs to be factored into the MELD design criteria. This expands the objective function (1) to

$$\min_{\mathbf{X},\mathbf{W}}\left(\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2+\lambda_1\|\mathbf{y}_0-\mathbf{X}\mathbf{1}\|_F^2\right),$$
$$\text{subject to}\quad 0\le \mathbf{X},\mathbf{W}\le 1, \tag{2}$$

where $\mathbf{1}$ stands for a column vector of all 1s, and multiplier λ_1 determines quality tradeoff between the shared view $\mathbf{y}_0$ and personal views $\mathbf{y}_k,\ 1\le k\le K$.

How many personal views can be sustained by the system in addition to a shared view and at what quality depend on the number of atom frames M. In TPVM, supporting M atom frames without flicker artifacts requires the display refresh rate and light modulator speed to reach 60M Hz. There are high-speed optoelectronic displays that can operate at 240 Hz or much higher. For example, the new DLP9500 DMD (which stands for "Digital Micromirror Device") from Texas Instruments can operate at 1,700 Hz for full high-definition resolution [9]. The speed active LC glasses is significantly lower. Although there are advanced LC glasses whose response speeds reach 1,000 Hz or above in laboratories [10], the speed of the off-the-shelf LC glasses cannot exceed 240 Hz. One way to accommodate this device limitation is to impose suitable constraints on the modulation vectors in $\mathbf{W}$.

Due to the material property of LC, the expected response time of active LC shutters is shorter if the controlling electric signal is sparse in time (i.e., fewer "on" states), meaning in our case a fewer number of large elements in the modulation weighting vector $\mathbf{w}_k,\ 1\le k\le K$. This sparsity constraint is added to the objective function (2) by making $\|\mathbf{w}_k\|_{\ell_1}$ as small as possible, namely,

$$\min_{\mathbf{X},\mathbf{W}}\left(\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2+\lambda_1\|\mathbf{y}_0-\mathbf{X}\mathbf{1}\|_F^2\right.$$
$$\left.+\lambda_2\sum_{k=1}^{K}\|\mathbf{w}_k\|_{\ell_1}\right),$$
$$\text{subject to}\quad 0\le \mathbf{X},\mathbf{W}\le 1, \tag{3}$$

where the second multiplier λ_2 governs the desired level of sparsity in the modulation weighting vectors. The sparsity-based NMF problem (3) can be simplified and solved by the two-block coordinate descent type of algorithms [11], e.g., the active set method [12].

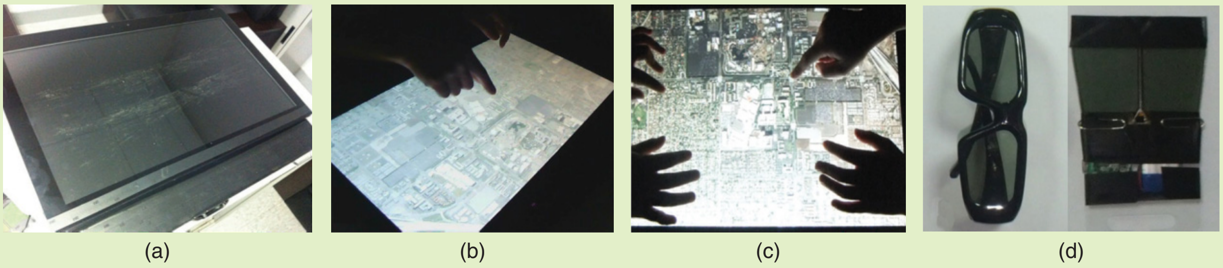
APPLICATION SHOWCASES AND DISCUSSIONS

A prototyped MELD viewport in action is illustrated in Figure 2. We showcase MELD functionalities and visual effects in two mock-up application scenarios: 1) collaborative visualization of a large and complex data set and 2) MELD display for multiuser virtual reality (VR) in surgical planning.

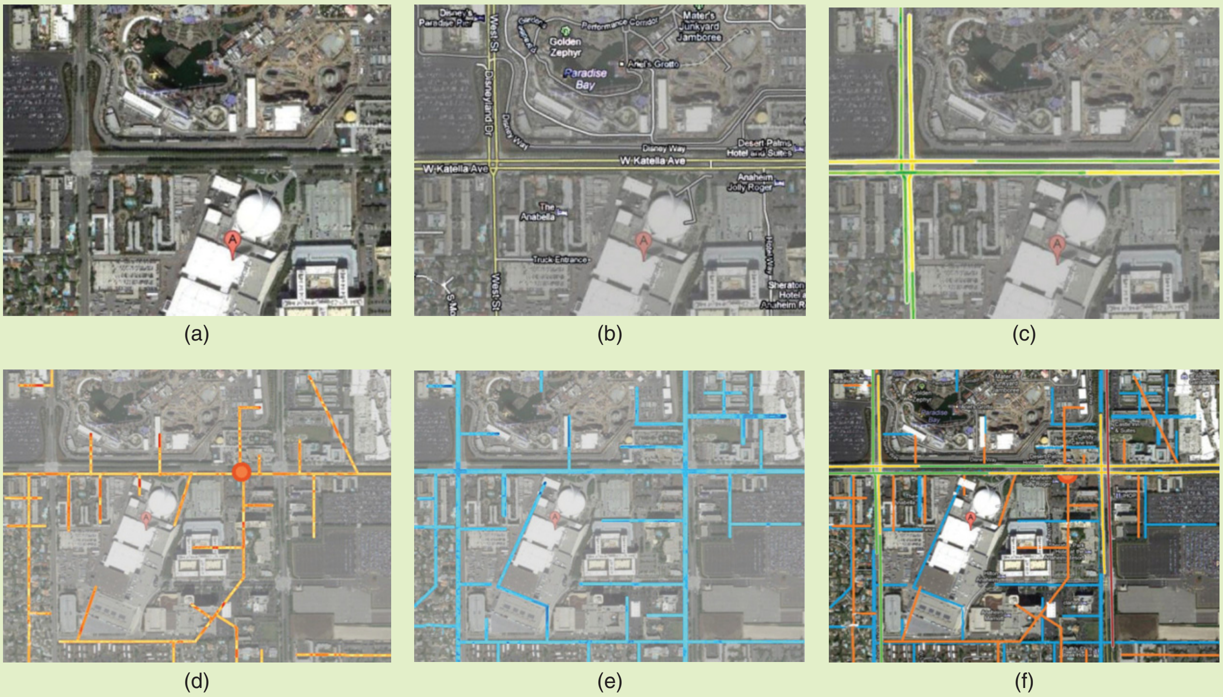
Figure 3 shows the screen captures of a multiuser collaborative visualization session as described in the introduction, where a group of interdisciplinary experts and municipal administrators congregate to discuss a city's emergency response plans. The available geographical data are highly complex with many specialty layers. Displaying all map layers together on a conventional screen generates severe visual clutters as shown in Figure 3(f). This problem is alleviated by the MELD system that concurrently presents multiple interference-free views tailored to individual participants: the view of the satellite image with buildings and roads annotated [Figure 3(b)], the view of the same satellite image but with color coded traffic patterns [Figure 3(c)], and the views of the same base image but coupled with different layers of underground structures [gas pipelines in Figure 3(d) and sewage system in Figure 3(e)]. In addition, the MELD system presents a shared (or default) view as a common reference for those who do not use modulation glasses [Figure 3(a)]. Unlike in other multiview display systems that restrict personal views in locations, viewpoints, and spatial resolution, MELD personal views are visible from any angle, presented at the full spatial resolution of the display, and can completely overlap each other without interference.

The second application scenario is surgical planning in a setting of multiuser mixed reality (combined virtual and physical realities), where collaborating surgeons and nurses "operate" on a virtual patient in physical copresence. If a conventional display is used to render the virtual patient, then participants are forced to have an identical view despite their different eye positions and viewing angles. Their visual experience will be distorted and disconnected from the VR, causing disorientation and cognitive impairment. In contrast, a MELD display concurrently generates perspective correct views for different participants, even as they move and physically interact with each other.

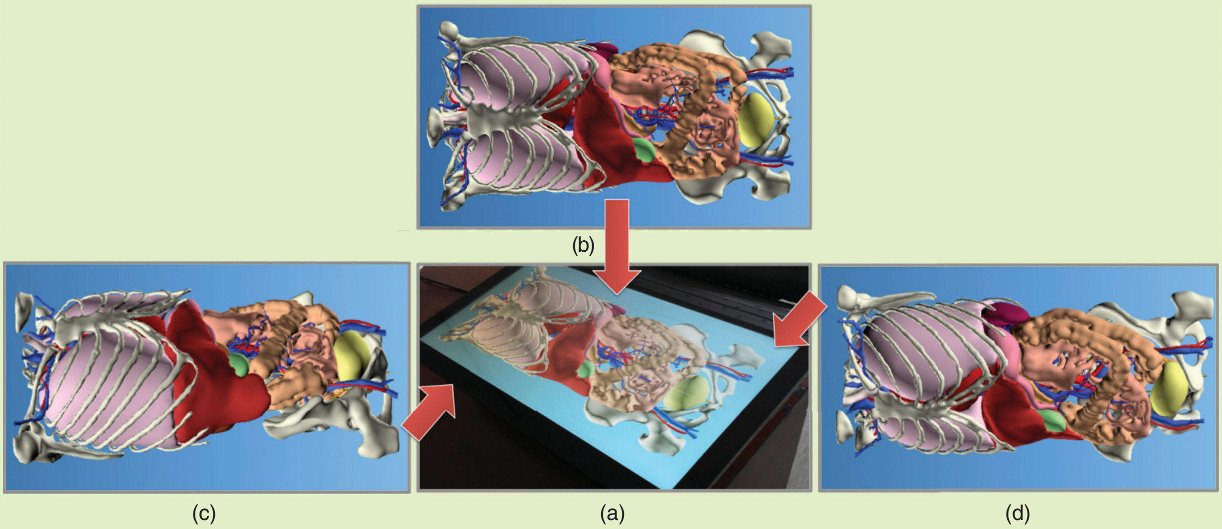
Figure 4 illustrates actions and effects of the MELD system in surgical planning. By feeding each participant her/his own eye-tracked perspective-correct image of the virtual patient's anatomy, the MELD


(a) (b) (c) (d)

[FIG2] (a) A prototyped MELD viewport. (b) and (c) Different personal views in collaborative visualization rendered on a common desktop display. (d) The LC viewing devices, glasses, and viewport.



[FIG3] Snapshots of the MELD prototype system used in multiuser collaborative visualization of a multilayer geographic information system data set. (a) Shared view: satellite image of a target area. (b) Personal view 1: with annotations of building and roads. (c) Personal view 2: with live traffic conditions shown. (d) Personal view 3: with mock underground gas pipelines and storage facility shown. (e) Personal view 4: with mock underground sewage system shown. (f) Visual clutter when all data layers are displayed superimposed to each other. Parts of data are from Google Maps and parts of data are imaginative for demonstration purpose only.



[FIG4] Snapshots of the MELD system when used in multiuser VR for medical applications. (a) The prototyped MELD desktop viewport. (b) A shared top-down view of a patient's anatomy. (c) A personal view from the top left of the virtual patient. (d) A personal view from the top right of the virtual patient.

system can create a fairly realistic collaborative team experience with respect to a common virtual environment rendered on a single desktop viewport. The participants can stand around the viewport as they would around an operation table, interact with each other face to face unobstructed, and yet have their own perspective-correct views of the virtual patient from the MELD viewport.

SUMMARY

We discussed a novel multiview display system suitable for multiuser VR/augmented reality and collaborative visualization that utilizes the TPVM principle for concurrent images formation. We also discussed an algorithmic approach of sparsity-based NMF for the generation of concurrent images. Finally we demonstrated the functionalities of the new display technology with multiple concurrent interference-free user-manipulable views at high quality on a common physical medium. The future directions of this research include the design of fast TPVM algorithms and the development of hardware solutions for real-time applications.

ACKNOWLEDGMENT

This work is supported by National Science Foundation of China grant 61331014.

AUTHORS

Guangtao Zhai (zhaiguangtao@sjtu.edu.cn) is a research professor at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, China. He is a Member of the IEEE.

Xiaolin Wu (xwu@ece.mcmaster.ca) is a professor of electrical and computer engineering at McMaster University, Hamilton, Ontario, Canada, and a guest professor at Shanghai Jiao Tong University, China. He is a Fellow of the IEEE.

REFERENCES

[1] X. Wu and G. Zhai, "Temporal psychovisual modulation: A new paradigm of information display," *IEEE Signal Processing Mag.*, vol. 30, no. 1, pp. 136–141, 2013.

[2] W. Matusik and H. Pfister, "3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," in *Proc. ACM SIGGRAPH*, 2004, pp. 814–824.

[3] K. Perlin, S. Paxia, and J. S. Kollin, "An autostereoscopic display," in *Proc. ACM SIGGRAPH*, 2000, pp. 319–326.

[4] A. Nashel and H. Fuchs, "Random hole display: A non-uniform barrier autostereoscopic display," in *Proc. IEEE 3DTV Conf. The True Vision–Capture, Transmission, and Display of 3D Video*, May 2009, pp. 1–4.

[5] S. Izadi, S. Hodges, S. Taylor, D. Rosenfeld, N. Villar, A. Butler, and J. Westhues, "Going beyond the display: A surface technology with an electronically switchable diffuser," in *Proc. ACM Symp. User Interface Software Technology*, 2008, pp. 269–278.

[6] P. Mistry, "ThirdEye: A technique that enables multiple viewers to see different content on a single display screen," in *Proc. ACM SIGGRAPH ASIA 2009 Posters*, p. 29:1.

[7] Sony, "Stereoscopic screen sharing method and apparatus," U.S. Patent Application 2010/0177172 A1, July 2010.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[9] Texas Instruments. (2012). PMU for active shutter 3D glasses. [Online]. Available: http://www.ti.com/product/dlp9500

[10] L. Komitov, G. Hegde, and D. Kolev, "Fast liquid crystal light shutter," *J. Phys. D*, vol. 44, no. 44, pp. 1–5, 2011.

[11] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computat. Stat. Data Anal.*, vol. 52, no. 1, pp. 155–173, Sept. 2007.

[12] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM J. Sci. Comput.*, vol. 33, no. 6, pp. 3261–3281, 2011.

[SP]

[dates AHEAD]

Please send calendar submissions to:
Dates Ahead, c/o Jessica Barragué
IEEE Signal Processing Magazine
445 Hoes Lane
Piscataway, NJ 08855 USA
e-mail: j.barrague@ieee.org
(Colored conference title indicates SP-sponsored conference.)

2014

[AUGUST]

11th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)
26–29 August, Seoul, South Korea.
General Chair: Hanseok Ko
General Cochair: Jin Young Choi
URL: http://www.avss2014.org/

IEEE Signal Processing Society Summer School on "Internet of Things and Machine-to-Machine Systems" (M2M)
26–29 August, Taipei, Taiwan.
General Chair: Sy-Yen Kuo
URL: http://wmnlab.ee.ntu.edu.tw/IEEESummerSchool/index.html

[SEPTEMBER]

22nd European Signal Processing Conference (EUSIPCO)
1–5 September, Lisbon, Portugal.
Honorary Chair: Carlos Salema
General Chair: Leonel Sousa
URL: http://www.eusipco2014.org/

2014 Sensor Signal Processing for Defence (SSPD)
8–9 September, Edinburgh, United Kingdom.
General Chairs: Mike Davies, Paul Thomas, and Jonathon Chambers
URL: http://www.see.ed.ac.uk/drupal/udrc/sspd/

24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)
21–24 September, Reims, France.
General Chair: Mamadou Mboup
URL: http://mlsp2014.conwiz.dk/home.htm

16th IEEE International Workshop on Multimedia Signal Processing (MMSP)
22–24 September, Jakarta, Indonesia.
General Chairs: Susanto Rahardja and Zhengyou Zhang
URL: http://mmsp2014.ilearning.me/call-for-paper/

[OCTOBER]

IEEE Workshop on Signal Processing Systems (SIPS)
20–23 October, Belfast, Ireland.

[NOVEMBER]

48th Asilomar Conference on Signals, Systems, and Computers
2–5 November, Pacific Grove, California.
General Chair: Roger Woods
Technical Program Chair: Geert Leus
URL: http://www.asilomarsscconf.org/

[DECEMBER]

IEEE Global Conference on Signal and Information Processing (GlobalSIP)
3–5 December, Atlanta, Georgia.
General Chairs: Geoffrey Li and Fred Juang
URL: http://renyi.ece.iastate.edu/globalsip2014/

IEEE International Workshop on Information Forensics and Security (WIFS)
3–5 December, Atlanta, Georgia.
General Chairs: Yan (Lindsay) Sun and Vicky H. Zhao
URL: http://ieeewifs.org/

IEEE Spoken Language Technology Workshop (SLT)
6–9 December, South Lake Tahoe, California.
General Chairs: Murat Akbacak and John Hansen

2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)
9–12 December, Chiang Mai, Thailand.
Honorary Cochairs: Sadaoki Furui, K.J. Ray Liu, and Prayoot Akkaraekthalin
General Cochairs: Kosin Chamnongthai, C.-C. Jay Kuo, and Hitoshi Kiya
URL: http://www.apsipa2014.org/home/

2015

[APRIL]

Data Compression Conference (DCC)
7–9 April, Snowbird, Utah.
URL: http://www.cs.brandeis.edu/~dcc/index.html

IEEE 12th International Symposium on Biomedical Imaging (ISBI)
16–19 April, Brookyln, New York.
General Chairs: Elsa Angelini and Jelena Kovacevic
URL: http://biomedicalimaging.org/2015/

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
19–24 April, Brisbane, Australia.
General Cochairs: Vaughan Clarkson and Jonathan Manton
URL: http://icassp2015.org/

[JUNE]

IEEE International Conference on Multimedia and Expo (ICME)
29 June–3 July, Turin, Italy.
General Chairs: Enrico Magli, Stefano Tubaro, and Anthony Vetro
URL: http://www.icme2015.ieee-icme.org/index.php

[SEPTEMBER]

IEEE International Conference on Image Processing (ICIP)
28 September–1 October, Quebec City, Quebec, Canada.

Digital Object Identifier 10.1109/MSP.2014.2332105
Date of publication: 19 August 2014

[advertisers INDEX]

The Advertisers Index contained in this issue is compiled as a service to our readers and advertisers: the publisher is not liable for errors or omissions although every effort is made to ensure its accuracy. Be sure to let our advertisers know you found them through *IEEE Signal Processing Magazine*.

ADVERTISER	PAGE	URL	PHONE
ICASSP 2015	3	www.ICASSP2015.org	+61 2 9265 0700
IEEE Marketing Department	7	www.ieee.org/tryieeexplore	
IEEE MDL/Marketing	11	www.ieee.org/go/trymdl	
Mathworks	CVR 4	www.mathworks.com/accelerate	+1 508 647 7040
Mini-Circuits	CVR 2, 5, CVR 3	www.minicircuits.com	+1 718 934 4500
Norwegian University of Science & Technology	13	www.jobbnorg.no	+47 735 92 023

[advertising SALES OFFICES]

James A. Vick
Sr. Director, Advertising
Phone: +1 212 419 7767;
Fax: +1 212 419 7589
jv.ieeemedia@ieee.org

Marion Delaney
Advertising Sales Director
Phone: +1 415 863 4717;
Fax: +1 415 863 4717
md.ieeemedia@ieee.org

Susan E. Schneiderman
Business Development Manager
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org

Product Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, PA, DE, MD, DC, KY, WV

NEW ENGLAND/SOUTH CENTRAL/ EASTERN CANADA
Jody Estabrook
Phone: +1 774 283 4528;
Fax: +1 774 283 4527
je.ieeemedia@ieee.org
ME, VT, NH, MA, RI, CT, AR, LA, OK, TX
Canada: Quebec, Nova Scotia,
Newfoundland, Prince Edward Island,
New Brunswick

SOUTHEAST
Thomas Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
tf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

Digital Object Identifier 10.1109/MSP.2013.2290964

MIDWEST/CENTRAL CANADA
Dave Jones
Phone: +1 708 442 5633;
Fax: +1 708 442 7620
dj.ieeemedia@ieee.org
IL, IA, KS, MN, MO, NE, ND,
SD, WI, OH
Canada: Manitoba,
Saskatchewan, Alberta

MIDWEST/ ONTARIO, CANADA
Will Hamilton
Phone: +1 269 381 2156;
Fax: +1 269 381 2556
wh.ieeemedia@ieee.org
IN, MI. Canada: Ontario

WEST COAST/MOUNTAIN STATES/ WESTERN CANADA
Marshall Rubin
Phone: +1 818 888 2407;
Fax: +1 818 888 4907
mr.ieeemedia@ieee.org
AZ, CO, HI, NM, NV, UT, AK, ID, MT,
WY, OR, WA, CA. Canada: British
Columbia

EUROPE/AFRICA/MIDDLE EAST ASIA/FAR EAST/PACIFIC RIM
Louise Smith
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
les.ieeemedia@ieee.org
Europe, Africa, Middle East
Asia, Far East, Pacific Rim, Australia,
New Zealand

Recruitment Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, CT, PA, DE, MD, DC, KY, WV

NEW ENGLAND/EASTERN CANADA
Liza Reich
Phone: +1 212 419 7578;
Fax: +1 212 419 7589
e.reich@ieee.org
ME, VT, NH, MA, RI. Canada: Quebec,
Nova Scotia, Prince Edward Island,
Newfoundland, New Brunswick

SOUTHEAST
Cathy Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
cf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

MIDWEST/SOUTH CENTRAL/ CENTRAL CANADA
Darcy Giovingo
Phone: +224 616 3034;
Fax: +1 847 729 4269
dg.ieeemedia@ieee.org;
AR, IL, IN, IA, KS, LA, MI, MN, MO, NE,
ND, SD, OH, OK, TX, WI. Canada:
Ontario, Manitoba, Saskatchewan, Alberta

WEST COAST/SOUTHWEST/ MOUNTAIN STATES/ASIA
Tim Matteson
Phone: +1 310 836 4064;
Fax: +1 310 836 4067
tm.ieeemedia@ieee.org
AZ, CO, HI, NV, NM, UT, CA, AK, ID, MT,
WY, OR, WA. Canada: British Columbia

EUROPE/AFRICA/MIDDLE EAST
Louise Smith
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
les.ieeemedia@ieee.org
Europe, Africa, Middle East

Find it at
mathworks.com/accelerate
datasheet
video example
trial request

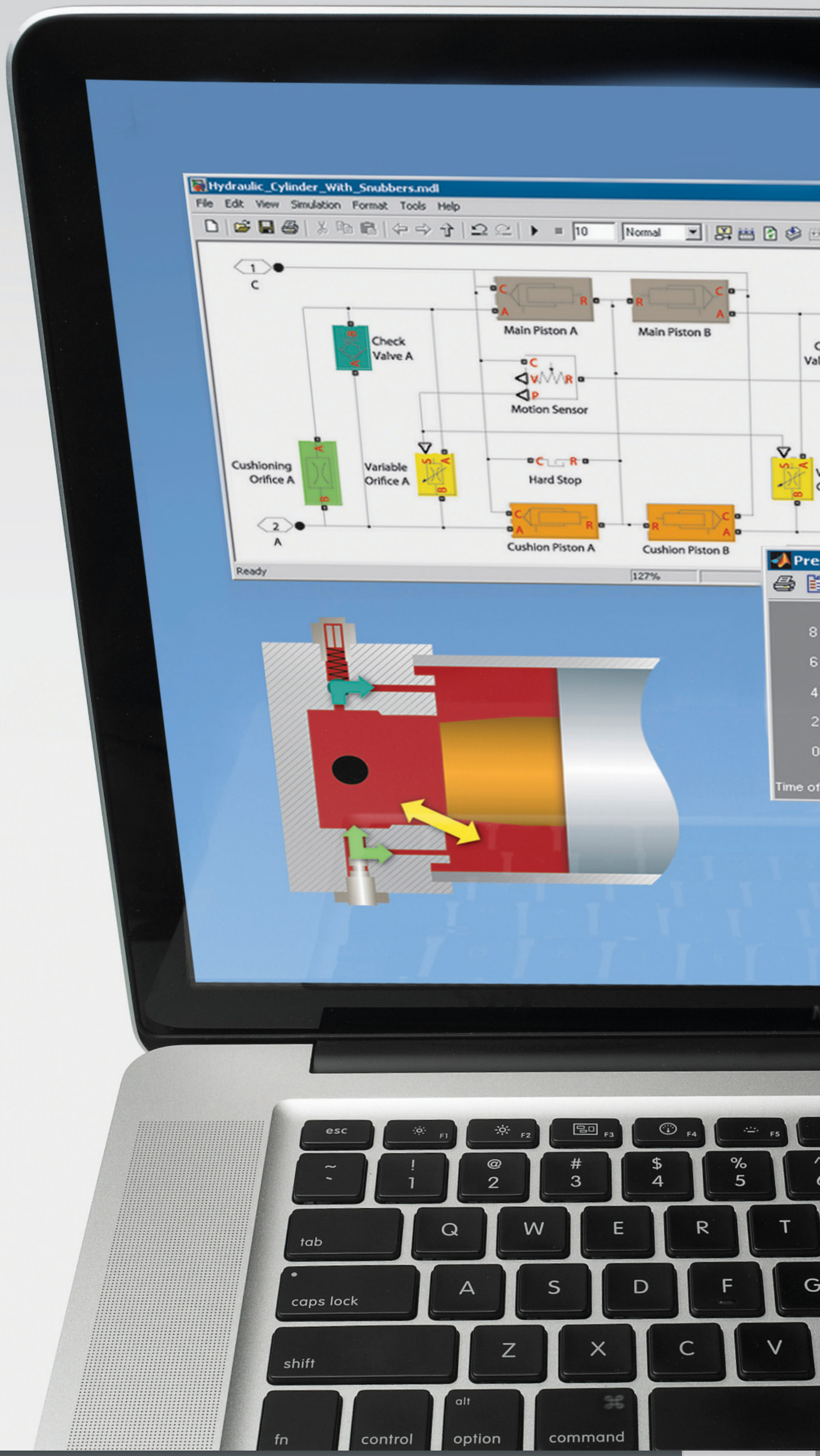
MODEL PHYSICAL SYSTEMS

in

Simulink

with **Simscape™**

- Electrical
- Mechanical
- Hydraulic
  and more



Use SIMSCAPE with SIMULINK to model and simulate the plant and controller of an embedded system. Assemble your model with a graphical interface, or import physical models from CAD systems. Use built-in components or create your own with the Simscape language.

MATLAB® & SIMULINK®

MathWorks®
Accelerating the pace of engineering and science

IEEE SIGNAL PROCESSING SOCIETY

CONTENT GAZETTE

[ISSN 2167-5023]

SEPTEMBER 2014

IEEE Signal Processing Society®

IEEE®

2015

EMB

IEEE Signal Processing Society

IEEE International Symposium on Biomedical Imaging

April 16th — 19th 2015, Brooklyn, NY USA

Conference Chairs

Elsa Angelini
Telecom ParisTech, France
Columbia University, USA

Jelena Kovačević
Carnegie Mellon University, USA

Program Chairs

Sebastien Ourselin
University College London, UK

Jens Rittcher
Oxford University, UK

Organizing Committee

Stephen Aylward, Kitware
Dana Brooks, Northeastern U.
Qi Duan, NIH
Elisa Konofagou, Columbia U.
Jan Kybic, Czech Tech. University
Erik Meijering, Erasmus MC
Wiro Niessen, Erasmus MC
Ricardo Otazo, NYU
Dirk Padfield, GE Healthcare
Gustavo Rohde, Carnegie Mellon
Badri Roysam, U. of Houston
Ivan Selesnick. Polytech NYU
Dimitri Van De Ville, EPFL
Simon Warfield, Harvard
Ge Yang, Carnegie Mellon

Contact

d.bernstein@ieee.org

The IEEE International Symposium on Biomedical Imaging (ISBI) is a premier interdisciplinary conference encompassing all scales of imaging in medicine and the life sciences. The 2015 meeting will continue its tradition of fostering knowledge transfer among different imaging communities and contributing to an integrative approach to biomedical imaging across all scales of observation.

ISBI is a joint initiative from the IEEE Signal Processing Society (SPS) and the IEEE Engineering in Medicine and Biology Society (EMBS). The 2015 meeting will open with a morning of tutorials, followed by a scientific program of plenary talks, invited special sessions, challenges, as well as oral and poster presentations of peer-reviewed papers.

High-quality papers are requested containing original contributions to mathematical, algorithmic, and computational aspects of biomedical imaging, from nano- to macro-scale. Topics of interest include image formation and reconstruction, computational and statistical image processing and analysis, dynamic imaging, visualization, image quality assessment, and physical, biological, and statistical modeling. We also encourage papers that elucidate biological processes (including molecular mechanisms) or translational ramification through integration of image-based data. Accepted 4-page regular papers will be published in the symposium proceedings and included in IEEE Xplore.

To encourage attendance by a broader audience of imaging scientists (in particular from the biology, radiology, and physics community) and offer additional opportunities for cross-fertilization, ISBI will again propose a second track featuring posters selected from abstract submissions without subsequent archival publication.

Important Dates

Tutorials, Special, Sessions & Challenges
Proposal Submission
June — Sept. 2014

4-Page Paper Submission

*Aug. 1st — **Nov. 10th**, 2014*
Notification
Dec. 20th, 2014
Upload & Registration
Jan. 10th, 2015

1-Page Paper Submission

*Nov. 20th , 2014 — **Dec. 20th** , 2014*
Notification
Feb. 1st , 2015
Upload & Registration
Feb. 15th , 2015

IEEE

Venue: ISBI 2015 will be held at the ***Marriott hotel at the Brooklyn bridge***, located on Adams street, next to the historical Court House building, with premier shopping, dining, and attractions in the heart of the Dumbo district. A short walk will take you to eight subway lines, a city bike station or a yellow cab to explore Brooklyn or to reach Manhattan just 1.5 miles (2 subway stations) across the East river for memorable nights in the Big Apple.

http://biomedicalimaging.org/2015

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed

MEDLINE U.S. National Library of Medicine

JULY 15, 2014 VOLUME 62 NUMBER 14 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

Adaptive Signal Processing

Digital and Multirate Signal Processing

IEEE TRANSACTIONS ON SIGNAL PROCESSING (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $569.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

Machine Learning

Signal Processing for Networks

Radar and Sonar Signal Processing

MIMO Communications & Signal Processing

Signal Processing for Communications

Statistical Signal Processing

COMMENTS AND REPLIES

Signal Processing for Networks

IEEE TRANSACTIONS ON

SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

AUGUST 1, 2014 VOLUME 62 NUMBER 15 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

Adaptive Signal Processing

Biomedical Signal Processing

Design and Implementation of Signal Processing Systems

IEEE TRANSACTIONS ON SIGNAL PROCESSING (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $569.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

Digital and Multirate Signal Processing

Multidimensional Signal Processing

Machine Learning

Signal Processing for Networks

Optimization Methods for Signal Processing

Sensor Array and Multichannel Processing

Signal Processing for Communications

Statistical Signal Processing

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

AUGUST 15, 2014 VOLUME 62 NUMBER 16 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

Adaptive Signal Processing

Design and Implementation of Signal Processing Systems

Digital and Multirate Signal Processing

Multidimensional Signal Processing

IEEE Transactions on Signal Processing (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $569.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE Transactions on Signal Processing, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

Machine Learning

MIMO Communications & Signal Processing

Signal Processing for Networks

Sensor Array and Multichannel Processing

Signal Processing for Communications

Signal Processing for Energy and Power Systems

Statistical Signal Processing

IEEE/ACM TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society ® Association for Computing Machinery www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

AUGUST 2014 VOLUME 22 NUMBER 8 ITASFA (ISSN 2329-9290)

REGULAR PAPERS

Modeling, Analysis and Synthesis of Acoustic Environments

Auditory Modeling and Hearing Aids

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (ISSN 2329-9290) is published bimonthly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $307.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

Acoustic Sensor Array Processing

Music Signal Analysis, Processing and Synthesis

Speech Enhancement

Robust Speech Recognition

IEEE TRANSACTIONS ON IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

JULY 2014 VOLUME 23 NUMBER 7 IIPRE4 (ISSN 1057-7149)

PAPERS

IEEE

IEEE TRANSACTIONS ON IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society® www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

AUGUST 2014 VOLUME 23 NUMBER 8 IIPRE4 (ISSN 1057-7149)

PAPERS

Image and Video Sensing and Acquisition

Structural-Model Based Methods

Image and Video Representation

Perception and Quality Models for Images and Video

Linear and Nonlinear Filtering of Images and Video

IEEE

Partial Differential Equation Based Processing of Images and Video

Multiresolution Processing of Images and Video

Restoration and Enhancement

Interpolation, Super-Resolution, and Mosaicing

Formation and Reconstruction

Lossy Coding of Images and Video

Color and Multispectral Imaging

Printing and Halftoning

Stereoscopic and Multiview Processing and Display

Radar Imaging, Remote Sensing, and Geophysical Imaging

Image and Video Mid Level Analysis

Image and Video Interpretation and Understanding

Image and Video Biometric Analysis

Image and Video Storage and Retrieval

IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.signalprocessingsociety.org

AUGUST 2014 VOLUME 9 NUMBER 8 ITIFA6 (ISSN 1556-6013)

PAPERS

Applied Cryptography

Biometrics

Communication and Physical-Layer Security

Forensics Analysis

Network Security

Security of Large Networked Systems

Steganography and Covert Communications

ANNOUNCEMENTS

General Chairs
Vaughan Clarkson
University of Queensland
Jonathan Manton
University of Melbourne

Technical Program Chairs
Doug Cochran
Arizona State University
Doug Gray
University of Adelaide

Finance Chair
Lang White
University of Adelaide

Special Session Chairs
Robert Calderbank
Duke University
Stephen Howard
DSTO
Songsri Sirianunpiboon
DSTO

Tutorials Chair
Daniel Palomar
Hong Kong University of S&T

Local Arrangements Chair
Andrew Bradley
University of Queensland

Registration Chair
Paul Teal
Victoria University of Wellington

Publicity Chair
Matt McKay
Hong Kong University of S&T

Publication Chair
Leif Hanlen
NICTA

Exhibits Chair
Iain Collings
CSIRO

Student Paper Contest Chair
Nikos Sidiropoulos
University of Minnesota

Conference Managers

Registration & Program Enquiries:
Conference Management Services, Inc
3833 S Texas Ave, Ste 221,
Bryan TX 77802, USA

General Enquiries:
arinex pty limited
S3, The Precinct, 12 Browning St
Brisbane QLD 4101, Australia
Ph: +61 2 9265 0700
Email: icassp2015@arinex.com.au

ICASSP
2015
Brisbane Australia

Second Call for Papers

ICASSP 2015

2015 IEEE International Conference on Acoustics,
Speech, and Signal Processing (ICASSP)
Brisbane Convention & Exhibition Centre
April 19 – 24, 2015 • Brisbane, Australia
www.ICASSP2015.org

The 40th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) will be held in the Brisbane Convention & Exhibition Centre, Brisbane, Australia, between April 19th and 24th, 2015. ICASSP is the world's largest and most comprehensive technical conference focused on signal processing and its applications. The conference will feature world-class speakers, tutorials, exhibits, and over 120 lecture and poster sessions. Topics include but are not limited to:

- Audio and acoustic signal processing
- Bio- imaging and biomedical signal processing
- Signal processing education
- Speech processing
- Industry technology tracks
- Information forensics and security
- Machine learning for signal processing
- Localisation and tracking
- Multimedia signal processing
- Sensor array & multichannel signal processing
- Design /implementation of signal processing systems
- Signal processing for communications & networking
- Image, video & multidimensional signal processing
- Signal processing theory & methods
- Spoken language processing
- Remote sensing signal processing

Submission of Papers: Prospective authors are invited to submit full-length papers, with up to four pages for technical content including figures and possible references, and with one additional optional 5th page containing only references. A selection of best papers will be made by the ICASSP 2015 committee upon recommendations from the Technical Committees.

Signal Processing Letters: Authors of IEEE Signal Processing Letters (SPL) papers will be given the opportunity to present their work at ICASSP 2015, subject to space availability and approval by the ICASSP Technical Program Chairs. SPL papers published on or after January 1, 2014 and SPL manuscripts accepted on or before November 15, 2014 are eligible for presentation at ICASSP 2015. Because they are already peer-reviewed and published, SPL papers presented at ICASSP 2015 will neither be reviewed nor included in the ICASSP proceedings. Requests for presentation of SPL papers should be made through the ICASSP 2015 website on or before 16 December 2014. Approved requests for presentation must have one author/presenter register for the conference according to the ICASSP 2015 registration instructions.

Important Deadlines:

Submission of regular papers	Sunday, October 5th 2014
Early registration opens	Monday, January 12th 2015
Notification of paper acceptance	Wednesday, January 14th 2015
Revised paper upload	Friday, February 13th 2015
Author registration	Friday, February 13th 2015

IEEE

IEEE Signal Processing Society ®

IEEE TRANSACTIONS ON MULTIMEDIA

A PUBLICATION OF
THE IEEE CIRCUITS AND SYSTEMS SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY

http://www.signalprocessingsociety.org/tmm/

AUGUST 2014 VOLUME 16 NUMBER 5 ITMUF8 (ISSN 1520-9210)

SPECIAL SECTION ON MUSIC DATA MINING

GUEST EDITORIAL

SPECIAL SECTION PAPERS

REGULAR PAPERS

Audio/Video Analysis and Synthesis

Compression and Coding

Watermarking, Encryption, and Data Hiding

3–D Audio/Video Processing

3–D Presentation

Multimodal Telepresence and Collaboration

Multimedia Search and Retrieval

Media Cloud Computing and Communication

Multimedia Streaming and Transport

Multimedia Algorithms, Systems, and Interfaces

CORRESPONDENCE

Watermarking, Encryption, and Data Hiding

IEEE JOURNAL OF
SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society ®

www.ieee.org/sp/index.html

AUGUST 2014 VOLUME 8 NUMBER 4 IJSTGY (ISSN 1932-4553)

EDITORIAL

PAPERS

IEEE

ANNOUNCEMENTS

CALL FOR PAPERS

IEEE Signal Processing Society

IEEE Journal of Selected Topics in Signal Processing

Special Issue on Signal and Information Processing for Privacy

Aims and Scope

There has been a remarkable increase in the usage of communications and information technology over the past decade. Currently, in the backend and in the cloud, reside electronic repositories that contain an enormous amount of information and data associated with the world around us. These repositories include databases for data-mining, census, social networking, medical records, etc. It is easy to forecast that our society will become increasingly reliant on applications built upon these data repositories. Unfortunately, the rate of technological advancement associated with building applications that produce and use such data has significantly outpaced the development of mechanisms that ensure the privacy of such data and the systems that process it. As a society we are currently witnessing many privacy-related concerns that have resulted from these technologies—there are now grave concerns about our communications being wiretapped, about our SSL/TLS connections being compromised, about our personal data being shared with entities we have no relationship with, etc. The problems of information exchange, interaction, and access lend themselves to fundamental information processing abstractions and theoretical analysis. The tools of rate-distortion theory, distributed compression algorithms, distributed storage codes, machine learning for feature identification and suppression, and compressive sensing and sampling theory are fundamental and can be applied to precisely formulate and quantify the tradeoff between utility and privacy in a variety of domains. Thus, while rate-distortion theory and information-theoretic privacy can provide fundamental bounds on privacy leakage of distributed data systems, the information and signal processing techniques of compressive sensing, machine learning, and graphical models are the key ingredients necessary to achieve these performance limits in a variety of applications involving streaming data, distributed data storage (cloud), and interactive data applications across a number of platforms. This special issue seeks to provide a venue for ongoing research in information and signal processing for applications where privacy concerns are paramount.

Topics of Interest include (but are not limited to):

- Signal processing for information-theoretic privacy
- Signal processing techniques for access control with privacy guarantees in distributed storage systems
- Distributed inference and estimation with privacy guarantees
- Location privacy and obfuscation of mobile device positioning
- Interplay of privacy and other information processing tasks
- Formalized models for adversaries and threats in applications where consumer and producer privacy is a major concern
- Techniques to achieve covert or stealthy communication in support of private communications
- Competitive privacy and game theoretic formulations of privacy and obfuscation

Important Dates:

Manuscript submission due: October 1, 2014
First review completed: December 15, 2014
Revised manuscript due: February 1, 2015
Second review completed: March 15, 2015
Final manuscript due: May 1, 2015
Publication date: October 2015

Prospective authors should visit http://www.signalprocessingsociety.org/publications/periodicals/jstsp/ for information on paper submission. Manuscripts should be submitted using Manuscript Central at http://mc.manuscriptcentral.com/jstsp-ieee.

Wade Trappe Rutgers University, USA trappe@winlab.rutgers.edu	Lalitha Sankar Arizona State University, USA lalithasankar@asu.edu	Radha Poovendran University of Washington, USA rp3@u.washington.edu
Heejo Lee Korea University, Korea heejo@korea.ac.kr	Srdjan Capkun ETH-Zurich srdjan.capkun@inf.ethz.ch	

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society ®

www.ieee.org/sp/index.html

SEPTEMBER 2014 VOLUME 21 NUMBER 9 ISPLEM (ISSN 1070-9908)

LETTERS

Audio and Acoustic Signal Processing

Digital and Multirate Signal Processing

Image and Multidimensional Signal Processing

IEEE SIGNAL PROCESSING LETTERS (ISSN 1070–9908) is published quarterly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $304.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE SIGNAL PROCESSING LETTERS, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

Machine Learning and Statistical Signal Processing

Sensor Array and Multichannel Signal Processing

Signal Processing for Communications

Speech Processing

Statistical and Adaptive Signal Processing

ANNOUNCEMENTS

IEEE SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.ieee.org/sp/index.html

OCTOBER 2014 VOLUME 21 NUMBER 10 ISPLEM (ISSN 1070-9908)

LETTERS

Audio and Acoustic Signal Processing

Biomedical Signal Processing

Digital and Multirate Signal Processing

IEEE SIGNAL PROCESSING LETTERS (ISSN 1070–9908) is published quarterly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $304.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE SIGNAL PROCESSING LETTERS, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

Image and Multidimensional Signal Processing

Machine Learning and Statistical Signal Processing

Signal Processing for Communications

Speech Processing

Statistical and Adaptive Signal Processing

IEEE SignalProcessing MAGAZINE

VOLUME 31 NUMBER 5 SEPTEMBER 2014

BIG DATA

THEORETICAL AND ALGORITHMIC FOUNDATIONS

OPTIMIZATION AND ESTIMATION OF COMPLEX-VALUED SIGNALS

MULTIUSER COLLABORATIVE VIEWPORT VIA TEMPORAL PSYCHOVISUAL MODULATION

NEW WIRELESS TECHNOLOGIES

IEEE Signal Processing Society®

IEEE

CONTENTS

[VOLUME 31 NUMBER 5]

SPECIAL SECTION—BIG DATA

FEATURE

COLUMNS

DEPARTMENT

Digital Object Identifier 10.1109/MSP.2014.2332376

The 15th ACM/IEEE International Conference on Information Networks IPSN Call for Papers

The International Conference on Information Processing in Sensor Networks (IPSN) is a leading, single-track, annual forum on research in networked sensing and control, broadly defined. IPSN brings together researchers from academia, industry, and government to present and discuss recent advances in both theoretical and experimental research. Its scope includes signal and image processing, information and coding theory, databases and information management, distributed algorithms, networks and protocols, wireless communications, collaborative objects and the Internet of Things, machine learning, mobile and social sensing, and embedded systems design. Of special interest are contributions at the confluence of a multiple of these areas.

In addition to regular research papers, in IPSN 2015, we also encourage submissions of Challenge Papers that lay out visions and future challenges in the field of information processing in sensor networks. Challenge papers are up to 6 page long and the title should start with “Challenge: … ” These submissions are reviewed based on the novelty of the concepts and potential of impacting the field.

The conference features two submission focus areas: one on Information Processing (IP), and one on Sensor Platforms, Tools and Design Methods (SPOTS). The entire program committee is eligible to review both focus areas, but authors are encouraged to make indications in the submission site accordingly to aid in reviewer selection.

The **IP area** focuses on algorithms, theory, and systems for information processing using networks of embedded, human-in-the-loop, or social sensors. Topics covered in the IP area include, but are not limited to:

- Sensor data processing, mining, and machine learning
- Data storage, management, and retrieval
- Coding, compression and information theory
- Detection, classification, tracking, reasoning, and decision making
- Sensor tasking, control, and actuation
- Theoretical foundation and fundamental bounds
- Network and system architectures and protocols
- Location, time, and other network services
- Programming models and languages
- Mobile, participatory, and social sensing
- Innovative applications and deployment experiences

Key Dates

Abstract registration:	October 3, 2014
Submission deadline:	October 10, 2014
	17, 2015

Submission

Formatting guidelines for regular and challenge papers are available here.

The **SPOTS area** focuses on new hardware and software architectures, modeling, evaluation, deployment experiences, design methods, implementations, and tools for networked embedded sensor systems. Submissions are expected to refer to specific hardware, software, and implementations. Topics covered in SPOTS include, but are not limited to:

- Novel components, devices and architectures for networked sensing
- Innovative sensing and processing platforms including cloud,crowd, and Internet-of-Things
- Embedded software for sensor networks
- System modeling, simulation, measurements, and analysis
- Design tools and methodologies for sensor networks
- Network health monitoring and management
- Operating systems and runtime environments
- User interfaces for sensing applications and systems
- Case studies highlighting experiences, challenges, and comparisons of platforms and tools

Organizers

General Chair: Suman Nath, MSR
TPC Co-Chair (IP): Bhaskar Krishnamachari, USC
TPC Co-Chair (SPOTS): Anthony Rowe, CMU
Steering Committee Chair: Feng Zhao, MSR Asia

Subject: IEEE Signal Processing Cup 2015 at ICASSP2015

Call for Participation: IEEE Signal Processing Cup 2015
http://icassp2015.org/signal-processing-cup-2015/

Challenge: **Heart Rate Monitoring During Physical Exercise Using Wrist-Type Photoplethysmographic (PPG) Signals**
For details of the competition project, please visit: www.zhilinzhang.com/spcup2015/

The IEEE Signal Processing Society organizes the SP Cup competition for undergraduates at ICASSP2015. This competition aims to provide undergraduate students with the opportunity to form teams and work together to solve a challenging and interesting real-world problem using signal-processing techniques. Three teams will be selected to present their work, and the prizes will be awarded at ICASSP 2015.

You are very welcome to participating in the competition. Please also help us to circulate this email to other colleagues or students you know who may be interested in this competition.

Participation in the Competition:

Each team participating in the competition is to be composed of one faculty member (whose role is the supervisor of the team members), at most one graduate student (who will assist the supervisor in supervising the undergraduate team members), and at least 3 but no more than 10 undergraduates. At least three of the undergraduate team members must be either IEEE SP members or student members.

Participating teams must submit their project by February 6, 2015. Each submission should include a report, in the form of an IEEE conference paper, on the technical details of the methods used and the results, as well as the programs developed (MATLAB is preferred). Participating teams must register to join the competition by January 16, 2015. The online registration system will be open in September 2014. By February 27, 2015, the best 3 teams will be identified to participate in the final competition at ICASSP2015.

Important Dates:

January 16, 2015 (Friday): Team registration to join the SP Cup competition
February 6, 2015 (Friday): Submission deadline for participating teams
February 27, 2015 (Friday): Announcement of the best 3 teams
April 20, 2015: Final competition at ICASSP 2015

Team Prizes:

The champion:	$5,000
The first runner-up:	$2,500
The second runner-up:	$1,500

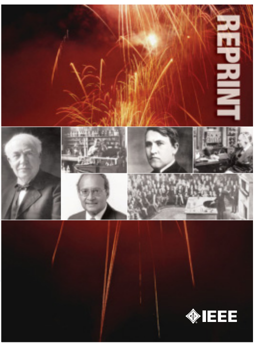
Each team invited to ICASSP2015 will have their travel expenses supported by the SP Society. Each team member is offered up to $1200 for continental travel, or $1,700 for intercontinental travel, and at most 3 people from each team will be supported.

Enquiries:

Technical problems: zhilinzhang@ieee.org
General enquiry: sp-enq-spcup@ieee.org

Organizers:

Bio Imaging and Signal Processing Technical Committee (BISP TC)
IEEE SPS Student Services Committee



IEEE ORDER FORM FOR REPRINTS

Purchasing IEEE Papers in Print is easy, cost-effective and quick.

Complete this form, tear it out, and either fax it (24 hours a day) to 732-981-8062 or mail it back to us.

PLEASE FILL OUT THE FOLLOWING

Author: ____________________________________

Publication Title: ____________________________________

Paper Title: ____________________________________

____

RETURN THIS FORM TO:
IEEE Publishing Services
445 Hoes Lane
Box 1331
Piscataway, NJ 08855-1331
Call Reprint Department at (732) 562-3941
for questions regarding this form
(732) 981-8062 - FAX

PLEASE SEND ME

□ 50 □ 100 □ 200 □ 300 □ 400 □ 500 or _________ (in multiples of 50) reprints.

□ YES □ NO Self-covering/title page required. COVER PRICE: $74 per 100, $39 per 50.

□ $58.00 Air Freight must be added for all orders being shipped outside the U.S.

□ $18.50 must be added for all USA shipments to cover the cost of UPS shipping and handling.

PAYMENT

□ Check enclosed. Payable on a bank in the USA.

□ Charge my: □ Visa □ Mastercard □ Amex □ Diners Club

Account # ____________________________________ Exp. date ______________________________

Cardholder's Name (please print): ____________________________________

____

□ Bill me (you must attach a purchase order) Purchase Order Number ______________________________

Send Reprints to:

____
____
____
____

Bill to address, if different:

____
____
____
____

Because information and papers are gathered from various sources, there may be a delay in receiving your reprint request. This is especially true with postconference publications. Please provide us with contact information if you would like notification of a delay of more than 12 weeks.

Telephone: ____________________ Fax: ____________________ Email Address: ____________________

2011 REPRINT PRICES (without covers)

Number of Text Pages

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40	41-44	45-48
50	$126	$211	$243	$245	$285	$340	$371	$408	$440	$477	$510	$543
100	$242	$423	$476	$492	$570	$680	$742	$817	$885	$953	$1021	$1088

Larger quantities can be ordered. Email reprints@ieee.org with specific details.

Tax Applies on shipments of regular reprints to CA, DC, FL, MI, NJ, NY, OH and Canada (GST Registration no. 12534188).
Prices are based on black & white printing. Please call us for full color price quote, if applicable.

Authorized Signature: ____________________________________ Date:________________

2014 IEEE MEMBERSHIP APPLICATION

(students and graduate students must apply online)

Start your membership immediately: Join online www.ieee.org/join

Please complete both sides of this form, typing or **printing in capital letters**. Use only English characters and abbreviate only if more than 40 characters and spaces per line. We regret that incomplete applications cannot be processed.

1 Name & Contact Information

Please PRINT your name as you want it to appear on your membership card and IEEE correspondence. As a key identifier for the IEEE database, circle your last/surname.

□ Male □ Female Date of birth (Day/Month/Year) ____/____/____

Title First/Given Name Middle Last/Family Surname

▼ Primary Address □ Home □ Business (All IEEE mail sent here)

Street Address

City State/Province

Postal Code Country

Primary Phone

Primary E-mail

▼ Secondary Address □ Home □ Business

Company Name Department/Division

Street Address City State/Province

Postal Code Country

Secondary Phone

Secondary E-mail

To better serve our members and supplement member dues, your postal mailing address is made available to carefully selected organizations to provide you with information on technical services, continuing education, and conferences. Your e-mail address is not rented by IEEE. Please check box only if you do not want to receive these postal mailings to the selected address. □

2 Attestation

I have graduated from a three- to five-year academic program with a university-level degree.
□ Yes □ No

This program is in one of the following fields of study:

□ Engineering
□ Computer Sciences and Information Technologies
□ Physical Sciences
□ Biological and Medical Sciences
□ Mathematics
□ Technical Communications, Education, Management, Law and Policy
□ Other (please specify): ________________

This academic institution or program is accredited in the country where the institution is located. □ Yes □ No □ Do not know

I have ______ years of professional experience in teaching, creating, developing, practicing, or managing within the following field:

□ Engineering
□ Computer Sciences and Information Technologies
□ Physical Sciences
□ Biological and Medical Sciences
□ Mathematics
□ Technical Communications, Education, Management, Law and Policy
□ Other (please specify): ________________

IEEE Advancing Technology for Humanity

3 Please Tell Us About Yourself

Select the numbered option that best describes yourself. This information is used by IEEE magazines to verify their annual circulation. Please enter numbered selections in the boxes provided.

A. Primary line of business → □

1. Computers
2. Computer peripheral equipment
3. Software
4. Office and business machines
5. Test, measurement and instrumentation equipment
6. Communications systems and equipment
7. Navigation and guidance systems and equipment
8. Consumer electronics/appliances
9. Industrial equipment, controls and systems
10. ICs and microprocessors
11. Semiconductors, components, sub-assemblies, materials and supplies
12. Aircraft, missiles, space and ground support equipment
13. Oceanography and support equipment
14. Medical electronic equipment
15. OEM incorporating electronics in their end product (not elsewhere classified)
16. Independent and university research, test and design laboratories and consultants (not connected with a mfg. co.)
17. Government agencies and armed forces
18. Companies using and/or incorporating any electronic products in their manufacturing, processing, research or development activities
19. Telecommunications services, telephone (including cellular)
20. Broadcast services (TV, cable, radio)
21. Transportation services (airline, railroad, etc.)
22. Computer and communications and data processing services
23. Power production, generation, transmission and distribution
24. Other commercial users of electrical, electronic equipment and services (not elsewhere classified)
25. Distributor (reseller, wholesaler, retailer)
26. University, college/other educational institutions, libraries
27. Retired
28. Other______________________

B. Principal job function → □

1. General and corporate management
2. Engineering management
3. Project engineering management
4. Research and development management
5. Design engineering management —analog
6. Design engineering management —digital
7. Research and development engineering
8. Design/development engineering —analog
9. Design/development engineering—digital
10. Hardware engineering
11. Software design/development
12. Computer science
13. Science/physics/mathematics
14. Engineering (not elsewhere specified)
15. Marketing/sales/purchasing
16. Consulting
17. Education/teaching
18. Retired
19. Other ________________

C. Principal responsibility → □

1. Engineering and scientific management
2. Management other than engineering
3. Engineering design
4. Engineering
5. Software: science/mngmnt/engineering
6. Education/teaching
7. Consulting
8. Retired
9. Other ________________

D. Title → □

1. Chairman of the Board/President/CEO
2. Owner/Partner
3. General Manager
4. VP Operations
5. VP Engineering/Dir. Engineering
6. Chief Engineer/Chief Scientist
7. Engineering Management
8. Scientific Management
9. Member of Technical Staff
10. Design Engineering Manager
11. Design Engineer
12. Hardware Engineer
13. Software Engineer
14. Computer Scientist
15. Dean/Professor/Instructor
16. Consultant
17. Retired
18. Other ________________

Are you now or were you ever a member of IEEE?
□ Yes □ No If yes, provide, if known:

Membership Number Grade Year Expired

4 Please Sign Your Application

I hereby apply for IEEE membership and agree to be governed by the IEEE Constitution, Bylaws, and Code of Ethics. I understand that IEEE will communicate with me regarding my individual membership and all related benefits. **Application must be signed.**

Signature Date

Over Please

Information for Authors

(Updated March 2012)

The IEEE Transactions are published monthly covering advances in the theory and application of signal processing. The scope is reflected in the EDICS: the Editor's Information and Classification Scheme. Please consider the journal with the most appropriate scope for your submission.

Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Correspondences (brief items that describe a use for or magnify the meaning of a single technical point, or provide comment on a paper previously published in the Transactions). Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.

Every manuscript must (a) provide a clearly defined statement of the problem being addressed, (b) state why it is important to solve the problem, and (c) give an indication as to how the current solution fits into the history of the problem.

By submission/resubmission of your manuscript to this Transactions, you are acknowledging that you accept the rules established for publication of manuscripts, including agreement to pay all overlength page charges, color charges, and any other charges and fees associated with publication of the manuscript. Such charges are not negotiable and cannot be suspended.

New and revised manuscripts should be prepared following the "New Manuscript Submission" guidelines below, and submitted to the online manuscript system ScholarOne Manuscripts. After acceptance, finalized manuscripts should be prepared following the "Final Manuscript Submission Guidelines" below. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will access your manuscript electronically via the ScholarOne Manuscripts system.

New Manuscript Submission. Please follow the next steps.

1. *Account in ScholarOne Manuscripts*. If necessary, create an account in the on-line submission system ScholarOne Manuscripts. Please check first if you already have an existing account which is based on your e-mail address and may have been created for you when you reviewed or authored a previous paper.
2. *Electronic Manuscript*. Prepare a PDF file containing your manuscript in double-spaced format (one full blank line between lines of type) using a font size of 11 points or larger, having a margin of at least 1 inch on all sides. For a regular paper, the manuscript may not exceed 30 double-spaced pages, including title; names of authors and their complete contact information; abstract; text; all images, figures and tables; and all references.

   Upload your manuscript as a PDF file "manuscript.pdf" to the ScholarOne Manuscripts site, then proofread your submission, confirming that all figures and equations are visible in your document before you "SUBMIT" your manuscript. Proofreading is critical; once you submit your manuscript, the manuscript cannot be changed in any way. You may also submit your manuscript as a PostScript or MS Word file. The system has the capability of converting your files to PDF, however it is your responsibility to confirm that the conversion is correct and there are no font or graphics issues prior to completing the submission process.
3. *Double-Column Version of Manuscript*. You are required to also submit a roughly formatted version of the manuscript in single-spaced, double column IEEE format (10 points for a regular submission or 9 points for a Correspondence) using the IEEE style files (it is allowed to let long equations stick out). *If accepted for publication,* over length page charges are levied beginning with the 11th published page of your manuscript. You are, therefore, advised to be conservative in your submission. This double-column version submitted will serve as a confirmation of the approximate publication length of the manuscript and gives an additional confirmation of your understanding that over length page charges will be paid when billed upon publication.

   Upload this version of the manuscript as a PDF file "double.pdf" to the ScholarOneManuscripts site.
4. *Additional Material for Review*. Please upload pdf versions of all items in the reference list which are not publicly available, such as unpublished (submitted) papers. Other materials for review such as supplementary tables and figures, audio fragments and QuickTime movies may be uploaded as well. Reviewers will be able to view these files only if they have the appropriate software on their computers. Use short filenames without spaces or special characters. When the upload of each file is completed, you will be asked to provide a description of that file.
5. *Submission*. After uploading all files and proofreading them, submit your manuscript by clicking "Submit." A confirmation of the successful submission will open on screen containing the manuscript tracking number and will be followed with an e-mail confirmation to the corresponding and all contributing authors. Once you click "Submit," your manuscript cannot be changed in any way.
6. *Copyright Form and Consent Form*. By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with the IEEE copyright policies, authors are required to sign and submit a completed "IEEE Copyright and Consent Form" prior to publication by the IEEE.

   The IEEE recommends authors to use an effective electronic copyright form (eCF) tool within the ScholarOne Manuscripts system. You will be redirected to the "IEEE Electronic Copyright Form" wizard at the end of your original submission; please simply sign the eCF by typing your name at the proper location and click on the "Submit" button.

Correspondence Items. Correspondence items are short disclosures with a reduced scope or significance that typically describe a use for or magnify the meaning of a single technical point, or provide brief comments on material previously published in the Transactions. These items may not exceed 12 pages in double-spaced format (3 pages for Comments), using 11 point type, with margins of 1 inch minimum on all sides, and including: title, names and contact information for authors, abstract, text, references, and an appropriate number of illustrations and/or tables. Correspondence items are submitted in the same way as regular manuscripts (see "New Manuscript Submission" above for instructions).

Manuscript Length. Papers published on or after 1 January 2007 can now be up to 10 pages, and any paper in excess of 10 pages will be subject to over length page charges. The IEEE Signal Processing Society has determined that the standard manuscript length shall be no more than 10 published pages (double-column format, 10 point type) for a regular submission, or 6 published pages (9 point type) for a Correspondence item, respectively. Manuscripts that exceed these limits will incur mandatory over length page charges, as discussed below. Since changes recommended as a result of peer review may require additions to the manuscript, it is strongly recommended that you practice economy in preparing original submissions.

Exceptions to the 30-page (regular paper) or 12-page (Correspondences) manuscript length may, under extraordinary circumstances, be granted by the Editor-in-Chief. However, such exception does not obviate your requirement to pay any and all over length or additional charges that attach to the manuscript.

Resubmission of Previously Rejected Manuscripts. Authors of rejected manuscripts are allowed to resubmit their manuscripts only once. The Signal Processing Society strongly discourages resubmission of rejected manuscripts more than once. At the time of submission, you will be asked whether you consider your manuscript to be a new submission or a resubmission of an earlier rejected manuscript. If you choose to submit a new version of your manuscript, you will be asked to submit supporting documents detailing how your new version addresses all of the reviewers' comments.

Full details of the resubmission process can be found in the Signal Processing Society "Policy and Procedures Manual" at http://www.signalprocessingsociety.org/about/governance/policy-procedure/. Also, please refer to the decision letter and your Author Center on the on-line submission system.

Author Misconduct.

Author Misconduct Policy: Plagiarism includes copying someone else's work without appropriate credit, using someone else's work without clear delineation of citation, and the uncited reuse of an authors previously published work that also involves other authors. Plagiarism is unacceptable.

Self-plagiarism involves the verbatim copying or reuse of an authors own prior work without appropriate citation; it is also unacceptable. Self-plagiarism includes duplicate submission of a single journal manuscript to two different journals, and submission of two different journal manuscripts which overlap substantially in language or technical contribution.

Authors may only submit original work that has not appeared elsewhere in a journal publication, nor is under review for another journal publication. Limited overlap with prior journal publications with a common author is allowed only if it is necessary for the readability of the paper. If authors have used their own previously published work as a basis for a new submission, they are required to cite the previous work and very briefly indicate how the new submission offers substantively novel contributions beyond those of the previously published work.

It is acceptable for conference papers to be used as the basis for a more fully developed journal submission. Still, authors are required to cite related prior work; the papers cannot be identical; and the journal publication must include novel aspects.

Author Misconduct Procedures: The procedures that will be used by the Signal Processing Society in the investigation of author misconduct allegations are described in the IEEE SPS Policies and Procedures Manual.

Author Misconduct Sanctions: The IEEE Signal Processing Society will apply the following sanctions in any case of plagiarism, or in cases of self-plagiarism that involve an overlap of more than 25% with another journal manuscript:

1) immediate rejection of the manuscript in question;
2) immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors; immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors;
3) prohibition against each of the authors for any new submissions, either individually, in combination with the authors of the plagiarizing manuscript, or in combination with new co-authors, to all of the Society's publications (journals, conferences, workshops). The prohibition shall continue for one year from notice of suspension.

Further, plagiarism and self-plagiarism may also be actionable by the IEEE under the rules of Member Conduct.

Submission Format.

Authors are encouraged to prepare manuscripts employing the on-line style files developed by IEEE. All manuscripts accepted for publication will require the authors to make final submission employing these style files. The style files are available on the web at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2 under "Template for all Transactions." (LaTeX and MS Word).

Authors using LaTeX: the two PDF versions of the manuscript needed for submission can both be produced by the IEEEtran.cls style file. A double-spaced document is generated by including \documentclass[11pt,draftcls,onecolumn]{IEEEtran} as the first line of the manuscript source file, and a single-spaced double-column document for estimating the publication page charges via \documentclass[10pt,twocolumn,twoside]{IEEEtran} for a regular submission, or \documentclass[9pt,twocolumn,twoside]{IEEEtran} for a Correspondence item.

- *Title page and abstract:* The first page of the manuscript shall contain the title, names and contact information for all authors (full mailing address, institutional affiliations, phone, fax, and e-mail), the abstract, and the EDICS. An asterisk * should be placed next to the name of the Corresponding Author who will serve as the main point of contact for the manuscript during the review and publication processes.

  An abstract should have not more than 200 words for a regular paper, or 50 words for a Correspondence item. The abstract should indicate the scope of the paper or Correspondence, and summarize the author's conclusions. This will make the abstract, by itself, a useful tool for information retrieval.
- *EDICS:* All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at http://www.signalprocessingsociety.org/publications/periodicals/tsp/TSP-EDICS/
- NOTE: EDICS are necessary to begin the peer review process. Upon submission of a new manuscript, please choose the EDICS categories that best suit your manuscript. Failure to do so will likely result in a delay of the peer review process.
- The EDICS category should appear on the first page—i.e., the title and abstract page—of the manuscript.
- *Illustrations and tables:* Each figure and table should have a caption that is intelligible without requiring reference to the text. Illustrations/tables may be worked into the text of a newly-submitted manuscript, or placed at the end of the manuscript. (However, for the final submission, illustrations/tables must be submitted separately and not interwoven with the text.)

  Illustrations in color may be used but, unless the final publishing will be in color, the author is responsible that the corresponding grayscale figure is understandable.

  In preparing your illustrations, note that in the printing process, most illustrations are reduced to single-column width to conserve space. This may result in as much as a 4:1 reduction from the original. Therefore, make sure that all words are in a type size that will reduce to a minimum of 9 points or 3/16 inch high in the printed version. Only the major grid lines on graphs should be indicated.
- *Abbreviations:* This TRANSACTIONS follows the practices of the IEEE on units and abbreviations, as outlined in the Institute's published standards. See http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/authors/transjnl/auinfo07.pdf for details.
- *Mathematics:* All mathematical expressions must be legible. Do not give derivations that are easily found in the literature; merely cite the reference.

Final Manuscript Submission Guidelines.

Upon formal acceptance of a manuscript for publication, instructions for providing the final materials required for publication will be sent to the Corresponding Author. Finalized manuscripts should be prepared in LaTeX or MS Word, and are required to use the style files established by IEEE, available at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2.

Instructions for preparing files for electronic submission are as follows:

- Files must be self-contained; that is, there can be no pointers to your system setup.
- Include a header to identify the name of the TRANSACTIONS, the name of the author, and the software used to format the manuscript.
- Do not import graphics files into the text file of your finalized manuscript (although this is acceptable for your initial submission). If submitting on disk, use a separate disk for graphics files.
- Do not create special macros.
- Do not send PostScript files of the text.
- File names should be lower case.
- Graphics files should be separate from the text, and not contain the caption text, but include callouts like "(a)," "(b)."
- Graphics file names should be lower case and named fig1.eps, fig2.tif, etc.
- Supported graphics types are EPS, PS, TIFF, or graphics created using Word, Powerpoint, Excel or PDF. Not acceptable is GIF, JPEG, WMF, PNG, BMP or any other format (JPEG is accepted for author photographs only). The provided resolution needs to be at least 600 dpi (400 dpi for color).
- Please indicate explicitly if certain illustrations should be printed in color; note that this will be at the expense of the author. Without other indications, color graphics will appear in color in the online version, but will be converted to grayscale in the print version.

Additional instructions for preparing, verifying the quality, and submitting graphics are available via http://www.ieee.org/publications_standards/publications/ authors/authors_journals.html.

Multimedia Materials.

IEEE Xplore can publish multimedia files and Matlab code along with your paper. Alternatively, you can provide the links to such files in a README file that appears on Xplore along with your paper. For details, please see http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect6 under "Multimedia." To make your work reproducible by others, the TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper.

Page Charges.

Voluntary Page Charges. Upon acceptance of a manuscript for publication, the author(s) or his/her/their company or institution will be asked to pay a charge of $110 per page to cover part of the cost of publication of the first ten pages that comprise the standard length (six pages, in the case of Correspondences).

Mandatory Page Charges. The author(s) or his/her/their company or institution will be billed $220 per each page in excess of the first ten published pages for regular papers and six published pages for correspondence items. These are mandatory page charges and the author(s) will be held responsible for them. They are not negotiable or voluntary. The author(s) signifies his willingness to pay these charges simply by submitting his/her/their manuscript to the TRANSACTIONS. The Publisher holds the right to withhold publication under any circumstance, as well as publication of the current or future submissions of authors who have outstanding mandatory page charge debt.

Color Charges. Color figures which appear in color only in the electronic (Xplore) version can be used free of charge. In this case, the figure will be printed in the hardcopy version in grayscale, and the author is responsible that the corresponding grayscale figure is intelligible. Color reproduction in print is expensive, and all charges for color are the responsibility of the author. The estimated costs are as follows. There will be a charge of $62.50 for each figure; this charge may be subject to change without notification. In addition, there are printing preparation charges which may be estimated as follows: color reproductions on four or fewer pages of the manuscript: a total of approximately $1045; color reproductions on five pages through eight pages: a total of approximately $2090; color reproductions on nine through 12 pages: a total of approximately $3135, and so on. Payment of fees on color reproduction is not negotiable or voluntary, and the author's agreement to publish the manuscript in the TRANSACTIONS is considered acceptance of this requirement.

To find the Information for Authors for IEEE Signal Processing Letters, the IEEE Journal of Selected Topics in Signal Processing or the IEEE Signal Processing Magazine, please refer to the IEEE Signal Processing website at www.signalprocessingsociety.org.

2014 IEEE SIGNAL PROCESSING SOCIETY MEMBERSHIP APPLICATION

(Current and reinstating IEEE members joining SPS complete areas 1, 2, 8, 9.)
Mail to: IEEE OPERATIONS CENTER, ATTN: Louis Curcio, Member and Geographic Activities, 445 Hoes Lane, Piscataway, New Jersey 08854 USA
or Fax to (732) 981-0225 (credit card payments only.)
For info call (732) 981-0060 or 1 (800) 678-IEEE or E-mail: new.membership@ieee.org

IEEE

1. PERSONAL INFORMATION

NAME AS IT SHOULD APPEAR ON IEEE MAILINGS: **SEND MAIL TO**: □ **Home Address** OR □ **Business/School Address**
If not indicated, mail will be sent to home address. Note: Enter your name as you wish it to appear on membership card and all correspondence.
PLEASE PRINT Do not exceed 40 characters or spaces per line. Abbreviate as needed. Please circle your last/surname as a key identifier for the IEEE database.

____
TITLE FIRST OR GIVEN NAME MIDDLE NAME SURNAME/LAST NAME

____
HOME ADDRESS

____
CITY STATE/PROVINCE POSTAL CODE COUNTRY

2. Are you now or were you ever a member of IEEE? □ Yes □ No

If yes, please provide, if known:

MEMBERSHIP NUMBER______|__|__|__|__|__|__|__|__|__|

Grade_____________**Year Membership Expired:**_______________

3. BUSINESS/PROFESSIONAL INFORMATION

____
Company Name

____
Department/Division

____
Title/Position Years in Current Position

____
Years in the Profession Since Graduation □ PE State/Province

____
Street Address

____
City State/Province Postal Code Country

4. EDUCATION

A baccalaureate degree from an IEEE recognized educational program assures assignment of "Member" grade. For others, additional information and references may be necessary for grade assignment.

A. ____
Baccalaureate Degree Received Program/Course of Study

____
College/University Campus

____
State/Province Country Mo./Yr. Degree Received

B. ____
Highest Technical Degree Received Program/Course of Study

____
College/University Campus

____
State/Province Country Mo./Yr. Degree Received

5. ____

Full signature of applicant

6. DEMOGRAPHIC INFORMATION – ALL APPLICANTS -

Date Of Birth ____________________ □ Male □ Female
Day Month Year

7. CONTACT INFORMATION

____________________ ____________________
Office Phone/Office Fax Home Phone/Home Fax

____________________ ____________________
Office E-Mail Home E-Mail

8. 2014 IEEE MEMBER RATES

IEEE DUES	16 Aug 12-28 Feb 14	1 Mar 13-15 Aug 14
Residence	Pay Full Year	Pay Half Year**
United States	$187.00 □	$93.50 □
Canada (incl. GST)	$168.10 □	$84.05 □
Canada (incl. HST for NB, NF and ON)	$179.46 □	$89.73 □
Canada (incl. HST for Nova Scotia)	$182.30 □	$91.15 □
Canada (incl. HST for PEI)	$180.88 □	$90.44 □
Africa, Europe, Middle East	$155.00 □	$77.50 □
Latin America	$146.00 □	$73.00 □
Asia, Pacific	$147.00 □	$73.50 □

Canadian Taxes (GST/HST): All supplies, which include dues, Society membership fees, online products and publications (except CD-ROM and DVD media), shipped to locations within Canada are subject to the GST of 5% or the HST of 13%, 14% or 15%, depending on the Province to which the materials are shipped. GST and HST do not apply to Regional Assessments. (IEEE Canadian Business Number 12563 4188 RT0001)
Value Added Tax (VAT) in the European Union: In accordance with the European Union Council Directives 2002/38/EC and 77/388/EEC amended by Council Regulation (EC)792/2002, IEEE is required to charge and collect VAT on electronic/digitized products sold to private consumers that reside in the European Union. The VAT rate applied is the EU member country standard rate where the consumer is resident. (IEEE's VAT registration number is EU826000081)
U.S. Sales Taxes: Please add applicable state and local sales and use tax on orders shipped to **Alabama, Arizona, California, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Massachusetts, Maryland, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, West Virginia, Wisconsin.** Customers claiming a tax exemption must include an appropriate and properly completed tax-exemption certificate with their first order.

IEEE Signal Processing Society

2014 SPS MEMBER RATES

		16 Aug-28 Feb Pay Full Year	1 Mar-15 Aug Pay Half Year
Signal Processing Society Membership Fee*		$ 20.00 □	$ 10.00 □

Fee includes: **IEEE Signal Processing Magazine** (electronic and digital), **Inside Signal Processing eNewsletter** (electronic) and **IEEE Signal Processing Society Content Gazette** (electronic).

Add $15 to enhance SPS Membership and also receive:		$ 15.00 □	$ 7.50 □

IEEE Signal Processing Magazine (print) and **SPS Digital Library**: online access to Signal Processing Magazine, Signal Processing Letters, Journal of Selected Topics in Signal Processing, Trans. on Audio, Speech, and Language Processing, Trans. on Image Processing, Trans. on Information Forensics and Security and Trans. on Signal Processing.

Publications available only with SPS membership:

Signal Processing, IEEE Transactions on:	Print	$190.00 □	$ 95.00 □
Audio, Speech, and Language Proc., IEEE/ACM Trans. on:	Print	$145.00 □	$ 72.50 □
Image Processing, IEEE Transactions on:	Print	$188.00 □	$ 94.00 □
Information Forensics and Security, IEEE Trans. on:	Print	$163.00 □	$ 81.50 □
IEEE Journal of Selected Topics in Signal Processing:	Print	$160.00 □	$ 80.00 □
Affective Computing, IEEE Transactions on:	Electronic	$ 33.00 □	$ 16.50 □
Biomedical and Health Informatics, IEEE Journal of:	Print	$ 55.00 □	$ 27.50 □
	Electronic	$ 40.00 □	$ 20.00 □
	Print & Electronic	$ 65.00 □	$ 32.50 □
IEEE Cloud Computing	Electronic and Digital	$ 39.00 □	$ 19.50 □
Computing in Science & Engineering:	Electonic and Digital	$ 45.00 □	$ 22.50 □
	Print & Electronic	$ 49.00 □	$ 24.50 □
Medical Imaging, IEEE Transactions on:	Print	$ 74.00 □	$ 37.00 □
	Electronic	$ 53.00 □	$ 26.50 □
	Print & Electronic	$ 89.00 □	$ 44.50 □
Mobile Computing, IEEE Trans. on:	ELE/Print Abstract/CD-ROM	$ 36.00 □	$ 18.00 □
Multimedia, IEEE Transactions on:	Electronic	$ 42.00 □	$ 21.00 □
IEEE MultiMedia Magazine:	Electronic and Digital	$ 45.00 □	$ 22.50 □
	Print & Electronic	$ 49.00 □	$ 24.50 □
Network Science and Engrg, IEEE Trans. on:	Electronic	$ 32.00 □	$ 16.00 □
IEEE Reviews in Biomedical Engineering:	Print	$ 25.00 □	$ 12.50 □
	Electronic	$ 25.00 □	$ 12.50 □
	Print & Electronic	$ 40.00 □	$ 20.00 □
IEEE Security and Privacy Magazine:	Electronic and Digital	$ 45.00 □	$ 22.50 □
	Print and Electronic	$ 49.00 □	$ 24.50 □
IEEE Biometrics Compendium	Online	$ 30.00 □	$ 15.00 □
IEEE Sensors Journal:	Print	$150.00 □	$ 75.00 □
	Electronic	$ 50.00 □	$ 25.00 □
	Print & Electronic	$190.00 □	$ 95.00 □
Smart Grid, IEEE Transactions on:	Print	$100.00 □	$ 50.00 □
	Electronic	$ 40.00 □	$ 20.00 □
	Print & Electronic	$120.00 □	$ 60.00 □
IEEE Transactions on Engineering Management:	Print & Electronic	$ 39.00 □	$ 19.50 □
IEEE Engineering Management Review:	Print & Electronic	$ 29.00 □	$ 14.50 □
IEEE Technology Management Package:	Print & Electronic	$ 60.00 □	$ 30.00 □
Wireless Communications, IEEE Transactions on:	Print	$ 87.00 □	$ 43.50 □
	Electronic	$ 42.00 □	$ 21.00 □
	Print & Electronic	$ 87.00 □	$ 43.50 □
IEEE Wireless Communications Letters:	Print	$ 80.00 □	$ 40.00 □
	Electronic	$ 18.00 □	$ 9.00 □
	Print & Electronic	$ 95.00 □	$ 47.50 □

*IEEE membership required or requested
Affiliate application to join SP Society only. Amount Paid **$**_______________

9. IEEE Membership Dues (See pricing in Section 8) $_______________

Signal Processing Society Fees $_______________
Canadian residents pay 5% GST or 13% HST
Reg. No. 125634188 on Society payment(s) & pubs only Tax $_______________
AMOUNT PAID WITH APPLICATION TOTAL $_______________
Prices subject to change without notice.
□ **Check or money order enclosed Payable to IEEE on a U.S. Bank**
□ **American Express** □ **VISA** □ **MasterCard**
□ **Diners Club**

|__|__|__|__|__|__|__|__|__|__|__|__|__|__|__|__|

Exp. Date Mo./Yr. |__|__|__|__|

Cardholder 5 Digit Zip Code Billing Statement Address/USA Only |__|__|__|__|__|

____
Full signature of applicant using credit card Date

10. WERE YOU REFERRED?

□ Yes □ No If yes, please provide the follow information:
Member Recruiter Name:______________________________________
IEEE Recruiter's Member Number (Required): ____________________

IEEE Signal Processing Society®

IEEE®