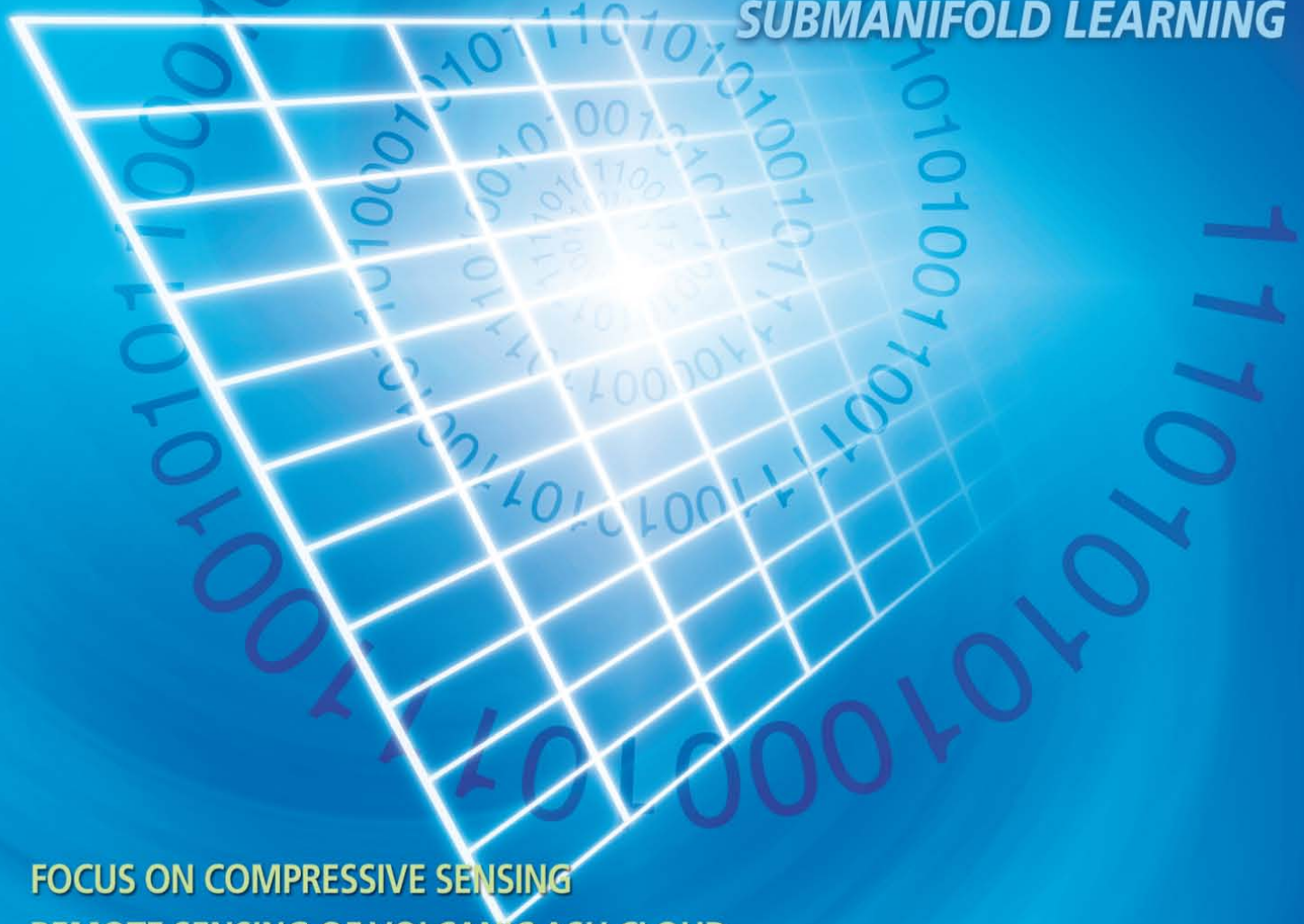
IEEE Signal Processing MAGAZINE

[VOLUME 28 NUMBER 2 MARCH 2011]

DIMENSIONALITY REDUCTION

VIA SUBSPACE AND SUBMANIFOLD LEARNING



FOCUS ON COMPRESSIVE SENSING

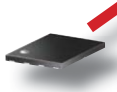
REMOTE SENSING OF VOLCANIC ASH CLOUD

MODELING SOCIAL PERCEPTION OF FACES

IEEE Signal Processing Society

IEEE

FLAT GAIN AMPLIFIERS



NF *as low as* **2.5 dB, Pout** *up to* **+20.5 dBm, 800 MHz** *to* **3.8 GHz** *from* **$1.79** *ea. (qty. 1000)*

Ultra flat gain, as low as ±0.2 dB across the entire frequency range, paves the way to all kinds of applications for our new YSF amplifiers. Together, these 7 models cover the 800-3200 MHz spectrum, from cellular and satellite L bands to GPS, PCS, UMTS, and WiMAX. Whenever gain flatness and repeatability are critical, and high dynamic range (low NF and high IP3) are required, Mini-Circuits YSF amplifiers are an ideal solution.

Model No.	Freq. (MHz)	Gain (dB)	Gain Flatness (±dB)	Pout (dBm) @ Comp		Dynamic Range NF	Dynamic Range IP3	Price $ ea.
				1dB	3dB	dB	dBm	Qty. 10
	fL-fU	Typ.		Typ.	Typ.	Typ.	Typ.	
YSF-122+	800-1200	20.4	0.2	20.5	21.3	3.4	36	2.69
YSF-2151+	900-2150	20.0	0.4	20.0	21.0	3.1	35	2.95
YSF-162+	1200-1600	20.1	0.2	20.0	21.0	3.2	35	2.69
YSF-232+	1700-2300	20.0	0.2	20.0	21.0	2.8	35	2.69
YSF-272+	2300-2700	19.0	0.7	20.0	21.0	2.5	35	2.59
YSF-382+	3300-3800	14.5	0.9	20.0	21.0	2.5	36	2.59
YSF-322+	900-3200	17.0	2.2	20.0	21.0	2.5	35	2.85

DC PWR. Voltage (nom.) 5v Current (max.) 145 mA RoHS compliant

Excellent combination of gain, noise, and distortion parameters. These amplifiers meet or exceed other key performance criteria with 20 dB gains, noise factors as low as 2.5, a 20 dBm P1dB, and a 35 dBm IP3.

They even simplify PCB configuration, with a small footprint (5 x 6 mm) and no external matching requirements. Our MSiP™ design provides the internal feedback, matching, bias, and DC blocking that make it all possible. So why wait? Place your order today, and we'll have them in your hands as early as tomorrow.

MSiP™ Mini-Circuits System In Package

Mini-Circuits...we're redefining what VALUE is all about!

Mini-Circuits®
ISO 9001 ISO 14001 AS9100 CERTIFIED

P.O. Box 350166, Brooklyn, New York 11235-0003 (718) 934-4500 Fax (718) 332-4661

The Design Engineers Search Engine finds the model you need, Instantly • For detailed performance specs & shopping online see minicircuits.com

U.S. patent 7739260

IF/RF MICROWAVE COMPONENTS

486 rev org

VOLUME 28 NUMBER 2

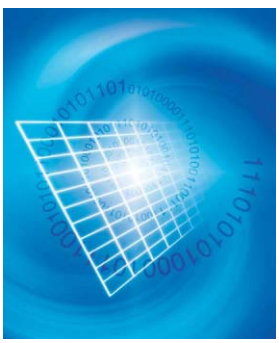
CONTENTS

SPECIAL SECTION—DIMENSIONALITY REDUCTION METHODS

COLUMNS

DEPARTMENTS

COVER © GETTY IMAGES



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.

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 $20.00 (first copy only), nonmembers $141.00 per copy. 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.

Digital Object Identifier 10.1109/MSP.2010.940001

[from the EDITORS]

Z. Jane Wang
Area Editor for eNews
zjanew@ece.ubc.ca

Li Deng
Editor-in-Chief
deng@microsoft.com
http://signalprocessingsociety.org/publications/periodicals/spm

Democratizing Signal Processing

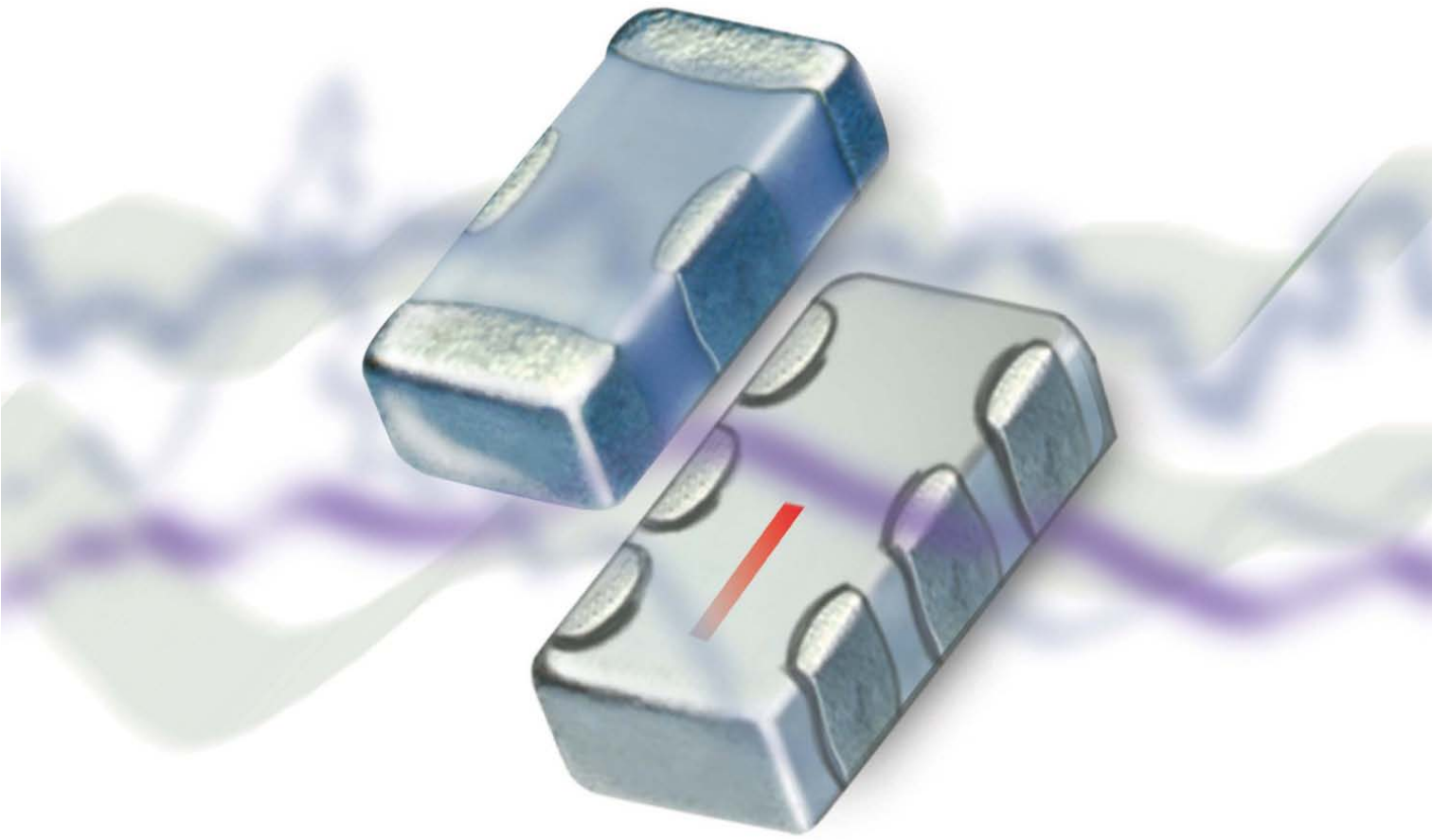
It will be March when you receive this issue of *IEEE Signal Processing Magazine* (*SPM*), and it has been almost four years since the IEEE Signal Processing Society (SPS)'s *Inside Signal Processing eNewsletter* (eNews) was first launched in April 2007, as the latest addition to *SPM*. When spring begins, flowers blossom and a new start brings new hopes. As the old Chinese saying goes, "A year's plan starts with spring." This is precisely what we are doing for the bimonthly *SPM* and the monthly eNews that complements the magazine.

While planning, we are guided by the motto of "democratizing signal processing." In a recent presentation in Shanghai that President Mos Kaveh gave on the history of signal processing and SPS, he used Matlab as an example of democratization of signal processing during the 1980s. Today, this trend is ever more prevalent than any time in the history. Signal processing techniques are no longer confined within the province of the privileged group of elite professionals. Rather, scientists, engineers, mathematicians, and even financial analysts have been using signal processing as a fundamental tool in problem solving, as evidenced by a series of recent and upcoming special issues in *SPM*. Riding on the tidal wave of the ubiquity of signal processing, our eNews has special roles to play, and we aim to fulfill these roles and maximize the societal impact of signal processing as well as the benefits of our *SPM* readers.

The IEEE has developed a global strategy to encourage active participation among all electrical engineers worldwide. Embracing this strategy, our SPS vision statement reads as follows: "The Signal Processing Society is a dynamic organization that is the preeminent source of signal processing information and resources for a *global* community. We do this by: being a one-stop source of signal processing resources; providing a variety of high quality resources to a variety of users in formats customized to their interests; adapting to a rapidly changing technical community; and being intimately involved in the education of signal processing professionals at all levels." Recently, a long-range strategic retreat was held by the SPS during IEEE ICIP 2010 in Hong Kong, where better

Digital Object Identifier 10.1109/MSP.2011.940297
Date of publication: 17 February 2011

Digital Object Identifier 10.1109/MSP.2011.940298



CERAMIC FILTERS

LOW PASS BANDPASS HIGH PASS

Value Packed Recession Busters! from 99¢ ea. qty. 1000

In today's tough economic situation there is no choice: Reducing cost while improving value is a must. Mini-Circuits has the solution…***pay less and get more*** for your purchases with our industry leading ultra small ceramic filters.

Over ~~141~~ 156 models…45 MHz to 13 GHz

Measuring only 0.12" x 0.06", these tiny hermetically sealed filters utilize our advanced Low Temperature Co-fired Ceramic *(LTCC)* technology to offer superior thermal stability, high reliability, and very low cost, making them a must for your system requirements. Visit our website to choose and view comprehensive performance curves, data sheets, pcb layouts, and environmental specifications. And you can even order direct from our web store and have a unit in your hands as early as tomorrow!

Mini-Circuits…we're redefining what VALUE is all about!

***Wild Card** KWC-LHP LTCC Filter Kits only $98*

Choose any 8, LFCN, HFCN models.
Receive 5 of ea. model, for a total of 40 filters.
Order your KWC-LHP FILTER KIT *TODAY!*

RoHS compliant U.S. Patent 6,943,646

Mini-Circuits®
ISO 9001 ISO 14001 AS9100 CERTIFIED

P.O. Box 350166, Brooklyn, New York 11235-0003 (718) 934-4500 Fax (718) 332-4661

U.S. patent 7739260

***The Design Engineers Search Engine** finds the model you need, Instantly* • For detailed performance specs & shopping online see minicircuits.com

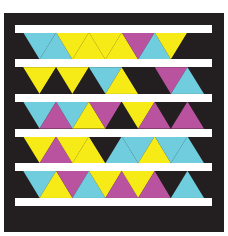
IF/RF MICROWAVE COMPONENTS

432 rev L

[FIG1]

serving our SPS members was the central theme of the retreat discussion. Using the eNews as the light-speed vehicle to inform our worldwide SPS members of how signal processing is democratized in all walks of life, *SPM* is equipped with a special tool to support IEEE's global strategy and to effectively serve our members.

The contents of the eNews are available at the Web site: http://enews.ieee-spm.org, and highlights of each issue are e-mailed monthly to SPS members. There were many reasons for the launch of eNews, with a critical one being its role of global reach. In her November 2007 *SPM* article "Signal Processing Magazine E-Newsletter: Inside Out," then Area Editor Min Wu stated that "As suggested by its name, *Inside Signal Processing*, the overall goal is to make the e-newsletter a gateway for different units of the SPS organizations to reach out, a cyber stop where students and professionals who are interested in signal processing can find out the latest happenings in this vibrant field." Since then, a number of new features have been introduced, and redesign efforts have been undertaken. Since its inception, it has been the eNews team's belief that eNews must include content that is timely, member-centric, relevant, targeted, and compelling. In the initial design of eNews, there were nine sections of news and updates: "Society News," "Conference News," "Publication News," "TC News," "Chapter/DL News," "Initiatives and Trends," "Ph.D. Theses," "New Books," and "Research Opportunities." Last year, as part of the vision and efforts of SPS Publications Board to fully leverage electronic medium in SPS, our eNews team redesigned and upgraded the newsletter by introducing a modern layout with enhanced features and improved navigations, the design you are seeing today.



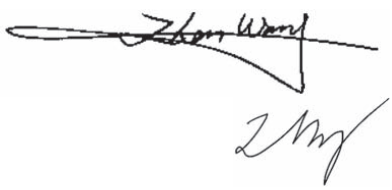
[FIG2]

The cultural and geographical diversity of SPS members is widespread. As the first step to enhance the team functionality and to strengthen the geographic and technical coverage of eNews content, two associate editors (AE) at-large were added in late 2009 to the eNews team to complement the original section AEs, each of whom is responsible for editing particular sections. Marcelo Bruno is AE at-large representing the Latin America region, and Gwenael Doerr is AE at-large representing the Europe region. Also, to strengthen its content coverage and improve its ability as the SPS gateway, many new content links have been integrated into eNews from different SPS sections. For instance, "Recent Patents in Signal Processing Areas" was introduced to the "Initiatives & Trends" section in February 2010; links to individual TC newsletters (e.g., the SLTC newsletter) was added to "TC News"; the "Conference Organizer Newsletter" link was added to "Conference News"; the "Top Ten Viewed Articles" was added to "Publication News"; and the newsletter "IEEE Chapter Briefs" and featured SPM articles became valuable new resources for the eNews contents.

The December 2010 eNews featured the article "Think Long-Range for Signal Processing Society" by President-Elect Ray Liu. The 2010 SPS long-range strategic retreat identified several action items as key new initiatives that SPS can consider taking on to offer greater benefit and value to SPS members, including bringing the visibility of signal processing to the general public, meeting the needs of continuing SP education and SP industry, attracting more student members, and so on. Recently there have been discussions between the SPS leadership team and the eNews team on how to incorporate the ideas from the 2010 Long-Range Planning Meeting into eNews to better serve the needs of SPS members, especially our student members and industry members. This will be a new focus of eNews. For instance, the eNews team, with the help from other SPM teams and SPS technical committees, will focus on enhancing or expending "Industry News" and "Job Opportunity News." We invite you, our readers, to share your insights with the eNews team and write to us. We need your feedbacks on how better design eNews and use this tool to serve your information needs. We seek your help to further democratize signal processing and to increase its societal impact in the most effective manner.

Aiming to democratize signal processing on the global scale, our *SPM* recently pioneered the innovation of translating our articles to Chinese, receiving highly positive feedback. More related details can be found in the article "Introducing a Translated Reprint Edition of *IEEE Signal Processing Magazine*" in the October 2010 issue of eNews. Following the success of the Chinese translation experiment, more recently our SPS Publication Board and ExCom unanimously approved the extension of our translation efforts to Portuguese as well as the continuation of Chinese translation. Another innovation we are planning to experiment with is to use the tag technology (via the two-dimensional bar codes) to enable *SPM* readers to conveniently connect from the printed material to the most relevant online content. Tags allow readers to quickly scan them with the cell phone camera, immediately launching a Web site with supplementary video/audio or text information such as video lectures, slide presentations, detailed technical reports, or animation of figures. Two examples of the tags (for the QR reader in Figure 1 and the MS reader in Figure 2) that link to an IEEE-TV video and a data file are provided here, ready for your scan after free downloading the respective readers to your smartphones (e.g., iPhones).

We hope you enjoy reading *SPM* and in particular the eNews as *SPM*'s key section. We are eager to receive your valuable feedback on any topic discussed above and sincerely invite you to contribute news and articles to *SPM* and the eNews.



[SP]

CAMSAP 2011

International Workshop on Computational Advances in Multi-Sensor Adaptive Processing

IEEE Signal Processing Society

The Fourth International Workshop on Computational Advances in Multi-Sensor Adaptive Processing

December 13-16 2011; San Juan Marriott Resort - San Juan, Puerto Rico

CALL FOR PAPERS

Following the success of the first three editions, we are pleased to announce the fourth workshop in the CAMSAP series, sponsored by the Sensor Array and Multi-channel Technical Committee of the IEEE Signal Processing Society. CAMSAP 2011 will be held at the San Juan Marriott Resort in San Juan, Puerto Rico, and will feature plenary talks from the world's leading researchers in the area, special focus sessions, and contributed papers. To provide feedback to the authors and ensure a high-quality program, all papers will undergo peer review.

IMPORTANT DATES

Special session proposals:
March 18th 2011
Full four-page paper submission:
July 8th 2011
Notification of acceptance:
September 16th 2011
Final camera-ready papers:
October 14th 2011
Early registration:
November 4th 2011

TOPICS OF INTEREST:

- Convex optimization and relaxation
- Computational linear algebra
- Computer-intensive methods in statistical SP (bootstrap, MCM, EM, particle filtering)
- Distributed computing, estimation, and detection algorithms
- Sampling and signal processing methods that exploit sparsity
- Emerging techniques

APPLICATIONS:

- Array processing: beamforming, space-time processing, and waveform design
- Communication systems
- Sensor networks
- Biomedicine
- Computational imaging
- Emerging topics

COMMITTEE

General Co-Chairs

Aleksandar Dogandzic, *Iowa State University, USA*
ald@iastate.edu
Maria Sabrina Greco, *University of Pisa, Italy*
m.greco@iet.unipi.it

Technical Program Co-Chairs

Sergiy A. Vorobyov, *University of Alberta, Canada*
vorobyov@ece.ualberta.ca
Lee Swindlehurst, *University of California – Irvine, USA*
swindle@uci.edu

Finance Chair

Dominic K. C. Ho, *University of Missouri – Columbia, USA*
hod@missouri.edu

Local arrangement Co-Chairs

Juan F. Arratia, Ana G. Méndez *University System, San Juan, Puerto Rico*
Luis M. Vicente, *Polytechnic University of Puerto Rico*

Technical Program Committee

Sofiene Affes (INRS-EMT, Canada)
Kristine Bell (Metron Inc., USA)
Rick Blum (Lehigh Univ., USA)
Volkan Cevher (EPFL, Switzerland)
Alex Dimakis (USC, USA)
Petar Djuric (State Univ. of New York, USA)
Alex Gershman (Darmstadt Un. of Tech., Germany)
Fulvio Gini (University of Pisa, Italy)
Simon Godsill (University of Cambridge, UK)
Zhu Han (University of Houston, USA)
Yingbo Hua (Univ. California Riverside, USA)
Andreas Jakobsson (Lund University, Sweden)
Mos Kaveh (University of Minnesota, USA)
Visa Koivunen (Aalto University, Finland)
Amir Leshem (Bar-Ilan University, Israel)
Hongbin Li (Stevens Inst. of Tech., USA)
Jian Li (University of Florida, USA)
Tom Luo (University of Minnesota, USA)
Arye Nehorai (Washington Univ. in S.Louis, USA)
Daniel Palomar (HKUST, Hong Kong)
Alejandro Ribeiro (Univ. of Pennsylvania, USA)
Brian Sadler (ARL, USA)
Shahram Shahbazpanahi (UOIT, Canada)
Nikos Sidiropoulos (TUC, Greece)
Ananthram Swami (ARL, USA)
Joseph Tabrikian (Ben-Gurion Univ., Israel)
Zhi Tian (Michigan Technological University, USA)
Jean-Yves Tourneret (IRIT/ENSEEIHT/TéSA, France)
Mats Viberg (Chalmers Univ. of Tech., Sweden)
Aylin Yener (Pennsylvania State University, USA)
Abdelhak Zoubir (Univ. of Darmstadt, Germany)

For more information visit the website at:
www.conference.iet.unipi.it/camsap11/

Technically co-sponsored by:

EURASIP

president's **MESSAGE**

Mostafa (Mos) Kaveh
2010–2011 SPS President
mos@umn.edu

Signal Processing Everywhere

During the past few months, I have had a chance to examine and reflect on the history of our field and the growth and diversification of the Society's technical and professional activities. I have also had several opportunities to give presentations on these topics to colleagues and students in meetings. The core phrase for the title of such a presentation is obvious and compelling: "Signal Processing Everywhere." Indeed, in the process of developing the IEEE Signal Processing Society's (SPS's) field of interest, the SPS Board of Governors boldly declared that "Signal processing is essential to integrating the contributions of other engineering and scientific disciplines in the design of complex systems that interact with humans and the environment, both as a fundamental tool, due to the signals involved, and as a driver of new design methodologies. As such, signal processing is a core technology for addressing critical societal challenges that include: healthcare, energy systems, sustainability, transportation, entertainment, education, communication, collaboration, defense, and security," (see http://www.signalprocessingsociety.org/about-sps/scope-mission/).

One need not look back very far to appreciate how far the field has come, and how its fundamental advances and transformative contributions to the way we live, work, and play are taken for granted. A major step in the broadening of the field happened in March 1966 with the renaming of *IEEE Transactions on Audio* as the quarterly *IEEE Transactions on Audio and Electroacoustics*. A bit over a year later, in June 1967, a special issue of the transactions was published focusing on the fast Fourier transform (FFT) and its applications to digital filtering and spectral analysis. The guest editorial for that issue by Bruce Bogert ended with these prophetic observations: "The audio engineer who naturally thinks only in terms of analog processing might well become familiar with what the digital approach is now able to offer. He may be surprised. What lies over the horizon in digital processing is anyone's guess, but I think it will surprise us all." The first article of that same issue was titled "What Is the Fast Fourier Transform?," written by a group of pioneers representing the IEEE Group on Audio and Electroacoustics Subcommittee on Measurement Concepts.

A MORE RECENT FORMALIZED AREA OF ACTIVITY WITHIN THE SOCIETY IS FOCUSED ON BIOLOGY AND MEDICINE.

Fast forward to January 2011 with *IEEE Signal Processing Magazine* devoting its special section on immersive communication, including several fascinating articles on a range of technique possibilities for immersive audio. SPS owes its genesis to audio, and it is appropriate to recognize the continuing signal processing challenges and opportunities this area provides. What is presented in the magazine is certainly not the audio many of us grew up with, or dreamt of!

A more recent formalized area of activity within the Society is focused on biology and medicine. Following the launch of its first major cross-organizational unit initiative on the smart grid, the IEEE has introduced a new one on life sciences. A meeting last November on this initiative was attended by representatives from many IEEE Societies and organizational units. The plan is, again, to provide a portal on the subject that integrates the activities across many organizational units and helps develop a "coherent IEEE life sciences strategy" for assisting volunteers and staff in expanding contributions and product development in this sphere. Not surprisingly, signal and image processing occupy central roles in the initiative, as indicated in a diagram on the intersection of life sciences and the IEEE's more traditional boundaries, presented in June 2010 to the IEEE Board of Directors. The chair of the SPS Bio Imaging and Signal Processing Technical Committee, Jean-Christophe Olivo-Marin, is serving as the Society's representative for this initiative.

ICASSP 2011 in beautiful Prague is just around the corner, and advance registration for the conference is almost upon us, along with the arrival of spring in the northern hemisphere. My best wishes to all of you for another season of renewal.

[SP]

Digital Object Identifier 10.1109/MSP.2010.940131
Date of publication: 17 February 2011

Member Awards, Fellows, and Call for Nominations

In this column, IEEE Signal Processing Society (SPS) award recipients are announced, 2011 SPS Fellows are introduced, and nominations are sought for Board of Governors members-at-large.

SPS MEMBERS RECEIVE IEEE AWARDS

The *IEEE Medal for Innovations in Healthcare and Technology* is being presented to Harrison H. Barrett, from the University of Arizona, for outstanding contributions and/or innovations in engineering within the fields of medicine, biology, and healthcare technology.

The *IEEE/Royal Society of Edinburgh Wolfson James Clark Maxwell Award* will be presented to Marcian Edward Hoff, of Teklicon, Inc. This award was established to acknowledge groundbreaking contributions that have had an exceptional impact on the development of electronics and electrical engineering or related fields.

The *IEEE Donald G. Fink Prize Paper* will be given to Andreas F. Mourlisch from the University of Southern California, Larry J. Greenstein of Rutgers University, and Mansoor Shafi of Telecom New Zealand. The award is presented to the most outstanding survey, review, or tutorial paper published in IEEE transactions, journals, magazines, or the *Proceedings of the IEEE* between 1 January and 31 December of the preceding year.

Ingrid Daubechies of Princeton University has been honored with the IEEE Jack S. Kilby Sign Processing Medal. Established in 1995, the medal is given for outstanding achievements in signal processing.

Digital Object Identifier 10.1109/MSP.2011.940238
Date of publication: 17 February 2011

49 SPS MEMBERS ELEVATED TO FELLOW

Each year, the IEEE Board of Directors confers the grade of Fellow on up to one-tenth percent of the members. To qualify for consideration, an individual must have been a Member, normally for five years or more, and a Senior Member at the time for nomination to Fellow. The grade of Fellow recognizes unusual distinction in the IEEE's designated fields.

The SPS congratulates the following 49 SPS members who were recognized with the grade of Fellow as of 1 January 2011.

Mohamed Abdel-Mottaleb, Coral Gables, Florida: For contributions to biometrics, content-based image and video retrieval, and digital mammography.

Mark Bell, West Lafayette, Indiana: For contributions to signal design and processing in radar and communication systems.

Shuvra Bhattacharyya, College Park, Maryland: For contributions to design optimization for signal processing.

Holger Boche, Berlin, Germany: For contributions to signal processing and multiuser wireless communications.

Wayne Burleson, Amherst, Massachusettes: For contributions to integrated circuit design and signal processing.

Jonathon Chambers, Loughborough, Leicestershire, United Kingdom: For contributions to adaptive signal processing and its applications.

Marco Chiani, Bologna, Italy: For contributions to wireless communication systems.

Pak Chung Ching, Shatin, Hong Kong, China: For leadership in engineering education and accreditation.

Ajay Divakaran, Princeton, New Jersey: For contributions to multimedia content analysis.

Emad Ebbini, Minneapolis, Minnesota: For contributions to ultrasound temperature imaging and dual-mode ultrasound.

Elza Erkip, Brooklyn, New York: For contributions to multiuser and cooperative communications.

Moncef Gabbouj, Tampere, Finland: For contributions to nonlinear signal processing and video communication.

Mark Gales, Cambridge, United Kingdom: For contributions to acoustic modeling for speech recognition.

David Gesbert, Sophia-Antipolis, France: For contributions to multi-antenna and multiuser communication theory and their applications.

Maria Greco, Pisa, Italy: For contributions to non-Gaussian radar clutter modeling and signal processing algorithms.

Arun Hampapur, Yorktown Heights, New York: For contributions to video indexing, video search, and surveillance systems.

Robert Heath, Austin, Texas: For contributions to multiple antenna wireless communications.

Visa Koivunen, Aalto, Finland: For contributions to statistical signal processing for multichannel signals and sensor arrays.

Ying-Chang Liang, Singapore: For contributions to cognitive radio-communications.

Johan Paul Linnartz, Eindhoven, The Netherlands: For leadership in security with noisy data.

Te-Won Lee, San Diego, California: For contributions to independent component algorithm analysis.

Shipeng Li, Beijing, China: For contributions to the advancement of image and video coding.

Patrick Loughlin, Pittsburgh, Pennsylvania: For contributions to

CALL FOR PAPERS
45th Annual Asilomar Conference on Signals, Systems, and Computers

IEEE Signal Processing Society

Asilomar Hotel and Conference Grounds
Pacific Grove, California
November 6-9, 2011
www.asilomarssc.org

Authors are invited to submit papers before **May 1st, 2011**, in the following areas:

A. Communications Systems: 1. Modulation and Detection, 2. Error Control Coding, 3. OFDM / Multicarrier, 4. Cognitive Radio, 5. Adaptive Waveforms, 6. Wireless Security/Privacy, 7. Power line communication, 8. DSL and Wireline Technologies, 9. 60GHz, 10. Optical Communications, 11. 4G Applications

B. MIMO Communications and Signal Processing: 1. Space-Time Coding and Decoding, 2. Channel Estimation and Equalization, 3. Multi-user MIMO, 4. Base Station Cooperation, 5. Limited Feedback Techniques, 6. Interference Management

C. Networks: 1. Transmission Techniques for Ad Hoc Networks, 2. Wireless Sensor Networks, 3. Network Information Theory, 4. Cooperative Diversity, 5. Relays, 6. Heterogeneous Networks, 7. Cognitive/Adaptive Sensor Networks

D. Signal Processing and Adaptive Systems: 1. Compressive Sensing, 2. Machine Learning Based Statistical Signal Processing, 3.Information Theoretic Signal Processing, 4. Cognitive Information Processing, 5. Adaptive Filtering, 6. Fast Algorithms

E. Array Signal Processing: 1. Source Localization, 2. Source Separation, 3. Adaptive Beamforming 4. Robust Methods, 5. Computational Aspects, 6. Applications (Sonar, Radar, Microphone arrays, etc.)

F. Biomedical Signal and Image Processing: 1. Medical Image Analysis, 2. Imaging Modalities, 3. Advances in Medical Imaging, 4. Biomedical Signal Processing, 5. Biomedical Applications, 6. Bioinformatics, 7. Image Registration and Multi-modal Imaging, 8. Image Reconstruction, 9. Computer Aided Diagnosis, 10. Functional Imaging, 11. Visualization

G. Architecture and Implementation: 1. Energy efficient design, 2. High-speed computer arithmetic, 3. Reconfigurable signal processing, 4. Multicore, manycore and distributed systems, 5. Algorithm and architecture co-optimization, 6. System-level representation and synthesis, 7. Cyber-physical system prototypes/testbeds

H. Speech, Image and Video Processing: 1. Speech Processing, 2. Speech Coding, 3. Speech Recognition, 4. Narrowband / Wideband Speech and Audio Coding, 5. Document Processing, 6. Models for Signal and Image Processing, 7. Image and Video Coding, 8. Image and Video Segmentation, 9. Image and Video Analysis, 10. Image / Video Security, Retrieval and Watermarking, 11. Image and Video Enhancement / Filtering, 12. Biometrics and Security, 13. Wavelets

Submissions should include a 50 to 100 word abstract and an extended summary (500 to 1000 words, plus figures). Submissions must include the title of the paper, each author's name and affiliation, and the technical area(s) in which the paper falls with number(s) from the above list. Check the conference website (www.asilomarssc.org) for specific information on the electronic submission process. Submissions will be accepted starting February 1, 2011. **No more than FOUR submissions are allowed** per contributor, as author or co-author. **All submissions must be received by May 1st, 2011**. Notifications of acceptance will be mailed by mid July 2011, and author information will be available on the conference website by late July 2011. Full papers will be due shortly after the conference and published in early 2012. All technical questions should be directed to the Technical Program Chair, **Dr. Robert W. Heath Jr.,** e-mail **rheath@ece.utexas.edu** or the General Chair, **Dr. Jim Schroeder**, e-mail **jim.schroeder@harris.com**.

CONFERENCE COMMITTEE

General Chair:	Jim Schroeder, *Harris Corporation*
Technical Program Chair:	Robert W. Heath Jr., *The University of Texas at Austin*
Conference Coordinator:	Monique P. Fargues, *Naval Postgraduate School*
Publication Chair:	Michael Matthews, *ATK Space Systems*
Publicity Chair:	Linda DeBrunner, *Florida State University*
Finance Chair:	Frank Kragh, *Naval Postgraduate School*

The site for the 2011 Conference is at the Asilomar Conference Grounds, in Pacific Grove, CA. The grounds border the Pacific Ocean and are close to Monterey, Carmel, and the scenic Seventeen Mile Drive in Pebble Beach.

The Conference is organized in cooperation with the Naval Postgraduate School, Monterey, CA, and ATK Space Systems, Monterey, CA. The IEEE Signal Processing Society is a technical co-sponsor of the conference.

time-frequency analysis and nonstationary signal processing.

Wei-Ying Ma, Beijing, China: For contributions to multimedia information retrieval.

Rainer Martin, Bochum, Germany: For contributions to speech enhancement for mobile communications and hearing aids.

Stephen McLaughlin, Edinburgh, Lothian, Scotland, United Kingdom: For contributions to statistical and nonlinear signal processing techniques in communication systems.

Nasir Memon, Brooklyn, New York: For contributions to media security and compression.

Asoke Nandi, Liverpool, United Kingdom: For contributions to signal processing and its applications.

Hermann Ney, Aachen, Germany: For contributions to statistical language modeling, statistical machine translation, and large vocabulary speech recognition.

Christof Paar, Bochum, Germany: For contributions to cryptographic engineering.

Eric Pottier, Rennes, Bretagne, France: For contributions to polarimetric specific absorption rate.

Susanto Rahardja, Singapore: For leadership in digital audio and signal processing.

Philippe Salembier, Barcelona, Spain: For contributions to region-based image analysis and mathematical morphology for compression and indexing.

Anna Scaglione, Davis, California: For contributions to filterbank precoding for wireless transmission and signal processing for cooperative sensor networks.

Laurence Simar, Richmond, Texas: For leadership in digital signal processor architecture development.

Andreas Stolcke, Menlo Park, California: For contributions to statistical language modeling, automatic speech recognition and understanding, and automatic speaker recognition.

Akihiko Sugiyama, Kawasaki, Kanazawa, Japan: For contributions to speech and audio signal processing.

Qibin Sun, Singapore: For contributions to multimedia security.

Myung Sunwoo, Suwon, Gyeonggi-Do, Korea: For contributions to multimedia and communications.

Isabel Trancoso, Lisbon, Portugal: For sustained contributions to speech technology, especially in the provision of research in and resources for the Portuguese language.

Mitchell Trott, Palo Alto, California: For contributions to wireless communication.

Vinay Vaishampayan, Florham Park, New Jersey: For contributions to error-resilient compression systems.

Anthony Vetro, Cambridge, Massachusetts: For contributions to video coding, three-dimensional television, and multimedia adaptation.

Narayanan Vijaykrishnan, University Park, Pennsylvania: For contributions to power-aware systems and estimation tools.

Emanuele Viterbo, Rende, Italy: For contributions to coding and decoding for wireless digital communications.

Li-Chun Wang, Hsinchu, Taiwan: For contributions to cellular architectures and radio resource management in wireless networks.

Min Wu, College Park, Maryland: For contributions to multimedia security and forensics.

Xiaolin Wu, Hamilton, Ontario, Canada: For contributions to image coding, communication, and processing.

Fan-Gang Zeng, Irvine, California: For contributions to metrology techniques for electromagnetic compatibility.

CALL FOR NOMINATIONS: BOARD OF GOVERNORS MEMBERS-AT-LARGE

In accordance with the SPS Bylaws, the membership will elect, by direct ballot, three members-at-large to the Board of Governors (BoG) for three-year terms commencing 1 January 2012 and ending 21 December 2014.

BoG members-at-large are directly elected by the Society's membership to represent the member viewpoint in Board decision making. Candidates must demonstrate that they are active in one or more signal processing disciplines and must have been a member of the Society for five years or more.

José M.F. Moura, SPS past president and chair of the Nominations and Appointments (N&A) Committee, has provided the following formal procedures for the SPS's 2011 BoG members-at-large elections.

- Publication of a call for nominations for positions of BoG members-at-large. Nominees must hold SPS member grade to hold elective office (March).
- From the responses received, a list of candidates will be assembled by the past president for presentation to the N&A Committee (April).
- The N&A ballots to create a short list of at least six candidates (by bylaw, at least two candidates must be submitted for each BoG member-at-large position becoming vacant) (April–May).
- After the N&A ranking ballot, the top candidates who are willing and able to serve for members-at-large are advanced for ballot to the SPS's voting members (July).
- Collection and tabulation of returned ballots will again be handled by the IEEE Technical Activities Society Services Department on behalf of the SPS (July–September).
- The three candidates receiving the highest number of votes who confirm their ability to serve will be declared elected members-at-large to the Board of Governors with three-year terms commencing 1 January 2012 (September).

Please provide nominations for members-at-large to Past President José M.F. Moura via e-mail to t.argiropoulos@ieee.org or via fax to +1 732 235 1627. Please provide the name, address, phone, fax, e-mail, or other contact information of the nominee, along with a brief background on the individual (no more than 100 words, please) and any information about the individual's current activities in the SPS, IEEE, or other professional societies.

[SP]

John Edwards

Focus on Compressive Sensing

"A unique single-pixel camera inspires a new generation of faster, cheaper imaging technologies."
—John Edwards

Back in 2004, Rice University researchers Richard Baraniuk and Kevin F. Kelly shook the imaging field, and more than a few accepted notions, by developing the first single-pixel camera. The prototype device demonstrated that a new technique, compressive sensing, provided a practical and relatively inexpensive way of creating high-resolution images in less time and with fewer sensors.

In the years since the single-pixel camera's introduction, researchers in several technology areas, including consumer and commercial photography, magnetic resonance imaging (MRI), and radar, have begun using compressive sensing to speed imaging, create high-quality images using only a limited number of sensors, or a combination of both attributes. Today, compressive sensing (also sometimes called compressed sensing) is increasingly seen as a way of helping developers enhance a wide range of existing imaging technologies without driving up costs.

The technique is, in essence, a shortcut. "Compressive sensing is really a mathematical basis for estimating [a] signal when you haven't made enough measurements," says David J. Brady, a Duke University electrical and computer engineering professor and compressive sensing researcher. "Basically, you're trying to estimate more signal values than you measured and it applies, very broadly, to many different measurement systems" (Figure 1).

Digital Object Identifier 10.1109/MSP.2010.939750
Date of publication: 17 February 2011

Brady observes that compressive sensing makes sense out of incomplete data. "In the 1990s, as people applied more and more computation to imaging, they began to realize that they could use algorithms to take data that was measured badly and fix it," he says. "Well, if I can take bad measurements and fix them with algorithms, maybe I could get some advantages by making deliberately bad measurements with reduced physical structures and still get good measurements."

A NEW APPROACH

Baraniuk, Rice's Victor E. Cameron Professor of Electrical and Computer Engineering, says that the single-pixel camera was the result of a series of extraordinary advances over the previous several years in computational science, particularly signal processing. "[It was] such that we could envision building entirely new kinds of sensors and, in particular, things like cameras that performed far beyond what you would expect using standard theory."

[FIG1] David J. Brady, electrical and computer engineering professor at Duke University, sees compressive sensing as a practical way of making sense out of incomplete data. (Photo used with permission from David J. Brady.)

Baraniuk notes that until the early 21st century, sensor research was largely driven by mathematics developed in the 1940s, primarily by the Shannon-Nyquist sampling theorem. "That theory told us that if you wanted a certain resolution, say in a camera or in an MRI scanner or an analog to digital converter, you had to sample at least twice as fast as the highest frequency in the signal." The results, he says, were highly predictable and fundamentally inflexible. "If we want a resolution in, say a digital camera of 10 megapixels, the theorem basically tells us that we need 10 million little sensors," Baraniuk says.

By 2004, however, new theoretical research had arrived to allow developers to push beyond the compression limits dictated by Shannon-Nyquist. "[It] made people realize that if we know just a little bit more about the signals, namely that they're compressible by an algorithm like JPEG, then you can actually sample the signals at a much, much lower rate," Baraniuk says. "Our realization was that we could build a camera that had 10 million effective pixels, but had a far fewer number of sensors."

Baraniuk and Kelly decided to push the new technique to its logical limit. "We took it to the extreme end of things, which was one single pixel, one single sensor," Baraniuk says. According to Kelly, a Rice associate professor of electrical and computer engineering, a single-pixel camera prototype seemed like an ideal way to both develop the concept

and to demonstrate compressive sensing's potential. "Rich and I realized that the best way to realize the mathematics was its direct implementation in hardware platforms," he says.

Baraniuk notes that while the camera represented a practical use of compressive sensing, the device itself wasn't the most important thing. "The real discovery was this new signal processing mathematics, which tells us that we can build sensors that take far fewer measurements than the classical theory would tell us," he says.

The realization led Baraniuk and Kelly to a new understanding of data compression. "It's commonly thought that if we want a 10-megapixel image, that we need to tap a whole bunch of little sensors on our camera chip, and that's actually not true," Baraniuk says. "Likewise, if we think we want a certain resolution with an MRI scan of our brain, that we need to sit in the scanner for 20 minutes. Well, that's not true, either."

COMPRESSED MEDICAL IMAGING

Michael Lustig, an assistant professor of electrical engineering and computer sciences at the University of California, Berkeley, feels that compressive sensing has the potential to revolutionize MRI technology, opening the door to faster medical imaging and, potentially, even video-style. Lustig is currently engaged in research that aims to make MRI scans better, faster, and more comfortable for patients.

"One of the shortcomings of MRI is that the scan time is relatively long," Lustig explains. He notes that until compressive sensing came along, MRI researchers were facing a brick wall when it came to reducing the amount of time required to conduct an MRI without seriously impairing resolution quality. "We were really at the limit of being able to collect data as fast as possible, just because of the physical and physiological constraints of the system," Lustig says.

Lustig feels that compressive sensing provides an almost ideal way of speeding up uncomfortable and claustrophobia-inducing MRI scans, which currently take up to 30 minutes or more. "One of the only ways to accelerate the scan time is to reduce the amount of data that's needed to reconstruct the image, so when the idea of compressed sensing came about it immediately occurred to us that MRI would be a great application to apply it to," he says.

Compressive sensing provides an entirely new approach to MRI image reconstruction. "Compressed sensing uses the fact that images are compressible, or that they can be represented expressly after applying some mathematical transformation," he says. "Up until now, none of the reconstruction techniques took advantage of that fact." Lustig adds that compressive sensing can also ensure a higher degree of image integrity than other data compression techniques. "If you consider the fact that the image you're expecting to reconstruct is compressible, you can reduce the amount of data but still be able to reconstruct the original image almost exactly," Lustig observes.

Lustig notes that image loss is a much larger concern in medicine than in conventional photography. In medical imaging, patients' lives often hinge on the inclusion—or omission—of just a few pixels. "That is one of the main issues with compressed sensing: to actually show in the clinical setting that you can robustly collect less data but still be able to get what you're interested in," Lustig says.

Compressive sensing can be a relatively low-cost way of improving MRI technology. "You really don't need to change the hardware," Lustig says. "It turns out that just by changing the software, and the way we acquire data in an MRI, we can do the kind of random sampling that is needed for compressed sensing."

Lustig says that most of the cost of using compressive sensing in MRI systems lies in creating the software. "Because, if you have an MRI system, you'll be able to do compressive sensing just by changing the pulse sequences … basically just by changing the software."

Most of the major MRI vendors are now working to add compressive sensing software to their systems. "I know GE has a team working on it, so do Siemens and Philips," Lustig says.

Lustig's University of California, Berkeley, team is also striving to bring compressive sensing MRI machines into the real world. "We have a project at Lucile Packard Children's Hospital [in Palo Alto, California]," he says. "We're trying to use this technique to accelerate the [scanning] of pediatric patients."

Infants are among the most problematic MRI patients. "This is a very vulnerable population, and they're very hard to image with MRI," Lustig says. "Because of the long scan times, they have to be put under general anesthesia." In an effort to reduce or even eliminate the need for anesthesia, the researchers are working with the hospital's pediatric radiologists to cut scanning times to the absolute minimum. "So if something would have taken two or ten minutes, then it might be reduced to maybe 20 seconds or just a minute instead of the full exam time, Lustig says.

After the data is acquired, it's sent to a "reconstruction machine," a computer equipped with high-speed general-purpose graphics processors. "Basically, it's a lot of very powerful processors to process the data very quickly," Lustig says. "Within less than a minute, you get images showing on your screen."

Stretching compressing sensing's limits requires a great deal of caution, however. "You have to be careful not to push the technology too much, because then you'll start degrading the image quality," Lustig says. Fortunately, errors are relatively easy to detect. "A trained radiologist, when he looks at images, he kind of knows if he's seeing something that looks like an artifact [or] if it's too low resolution," Lustig says.

Despite ongoing improvements, image reconstruction time continues to be an important issue. "It used to take several hours to reconstruct a simple three-dimensional (3-D) volume; we've been trying to address that by using parallel computing and fast algorithms," Lustig says. "We're now able to go just below a minute for reconstruction, but that's just for static images."

Lustig now wants to use faster MRI scanning rates to create 3-D MRI movies. "There are a lot of exams where you

inject a contrast solution into the body and you want to follow that contrast and to see it work dynamically," he says. "You mostly think of [MRI] as a still camera, but what we'd like to do is make it more like a video camera."

Creating full-motion MRIs will be a challenge however. "The problems are just huge … incredible," Lustig says. "We're talking about billions of variables to solve in order to get the images and huge amounts of memory."

A typical single, static MRI exam currently generates about five or six gigabytes of data, Lustig says. But an entire sequence of MRI images would boost storage requirements by several magnitudes. "We're talking about dynamics that can go up to hundreds of gigabytes," Lustig says. While that's not very much data in a world where 1 TB hard drives sell for under $100, reliably processing all of that information within a reasonable amount of time, even with the help of the most sophisticated compressive sensing techniques running on the most powerful processors, could result in very long processing times for both MRI systems and patients.

While MRI video remains on the drawing board, Lustig expects that basic, static compressive sensing MRI technology will become widely available over the next several years. "I think you'll see this kind of technology appearing in clinical practice in some form," he says. "It may not be the exact, true compressed sensing that was described in the theoretical papers, but a lot of these ideas will definitely penetrate [and] we'll be able to get much faster scans that produce much higher quality just by using these ideas."

ON THE RADAR SCREEN

Radar is another technology that could potentially benefit from compressed sensing. "One of the more high profile compressed sensing projects going on right now is building … a new radar receiver that can operate at very high frequencies," says Justin Romberg, an assistant professor of electrical and computer engineering at Georgia Tech. Romberg is investigating how the technique can be used to improve radar performance while cutting system costs. Teams at Cal Tech and Rice are also participating in the research, which is being funded by DARPA.

"The main thing all the different compressive sensing projects have in common is that you encode or scramble the data before you sample it," Romberg explains. "Rich [Baraniuk] does that with the single-pixel camera through using a mirror array; we do that for our radar receivers using high-frequency modulators," Romberg says. "They're just very different physical instantiations of the same principle."

The new receiver is being designed to bring high-frequency radar technology into the digital age. "It's impossible to build traditional hardware that acquires [radar pulses] digitally," Romberg says. "You have to basically build expensive analog circuits to see what's out there," he states.

Scrambling data with compressive sensing paves the way for cutting costs. "If you put this kind of scrambling on the front end, you can use more traditional acquisitions and still work your way into these high frequencies," Romberg says. "So the idea is to build cheaper hardware that lets you access more of the spectrum."

As the receiver researchers experiment with and refine various compressive sensing approaches, they're looking to achieve a technology balance. "It requires … effort to tease the information that you want out of the data that you've taken, so you might have to have a little more advanced processing algorithms on the back end," Romberg says. "It's like we're sort of trading off front-end sensor complexity versus back-end computing."

Romberg says he's excited by compressive sensing's potential to provide low cost, high speed analog to digital conversion. "It can allow you to reach parts of the spectrum, high frequencies, that you just can't get with any kind of traditional hardware," he says. "I mean, they're just totally inaccessible right now, except with some very expensive [hardware]."

A possible application Romberg and other researchers are looking at is ground penetrating radar. "When you're sweeping an area for land mines, for example, it takes a while to create a scan of the entire area," he says. "What this [compressive sensing technology] will do is reduce the amount of time that the scanning will take."

Besides radar, compressive sensing has a wide range of other potential communication applications, Romberg says. "There are people who are very interested in using these same types of ideas for very low power communications," he says. "You can have sensor networks that operate over a very long time period without having to have their batteries recharged." Romberg says there's also a chance that the technology might be integrated into next-generation analog-to-digital converters. "It could actually appear in, say, your cell phone," he says.

DOWN THE ROAD

Widespread commercialization of compressive sensing technologies is only a matter of time, predicts Baraniuk, who serves as a director of InView Technology, a company that he cofounded with Kelly to develop and market compressive sensing camera products. The Austin, Texas-based firm is currently developing a series of infrared cameras that it promises will be anywhere from five to ten times cheaper than currently available counterparts.

Brady is also bullish on compressive sensing's commercial prospects. "I think it's certainly going to come very soon," he predicts. Brady sees compressive sensing technology popping up in multiple areas. "The main markets are security markets, consumer imaging markets, and machine vision things," he says.

Baraniuk, meanwhile, thinks that space could be compressive sensing's next frontier. He sees cameras based on the technology being used on space probes to analyze alien environments. "That's something that's available today through hyperspectral cameras—but they cost hundreds of thousands of dollars," Baraniuk says. "We hope, and plan, to be able to build one of these [cameras] for thousands of dollars." [SP]

from the **GUEST EDITORS**

Yi Ma, Partha Niyogi,
Guillermo Sapiro, and René Vidal

Dimensionality Reduction via Subspace and Submanifold Learning

The problem of finding and exploiting low-dimensional structures in high-dimensional data is taking on increasing importance in image, video, or audio processing; Web data analysis/search; and bioinformatics, where data sets now routinely lie in observational spaces of thousands, millions, or even billions of dimensions. The curse of dimensionality is in full play here: We often need to conduct meaningful inference with a limited number of samples in a very high-dimensional space. Conventional statistical and computational tools have become severely inadequate for processing and analyzing such high-dimensional data.

Although the data might be presented in a very high-dimensional space, their intrinsic complexity and dimensions are typically much lower. One of the earliest and most popular models assumes that the data lie on a low-dimensional linear subspace. The subspace can be effectively computed via principal component analysis (PCA), whose origin can be traced back many decades ago [1]. In recent years, many new mathematical models have been proposed to capture more complex low-dimensional structures than a single subspace. For instance, the low-dimensional structure can be assumed to be nonlinear and the data hence lie on a low-dimensional submanifold [2], [3]. Or the structure can be assumed to be hybrid, i.e., different pieces might have different dimensions, and the data lie on an arrangement of multiple linear subspaces [4]–[6]. Similarly, one may assume that the data have a sufficiently sparse representation with respect to some basis or dictionary [7], [8], in which case one is dealing with a very special class of subspace arrangements, known as perfect arrangements. Most generally, one can even assume that the data lie on an arrangement of submanifolds, each with a possibly different dimension. Such a structure is also known as a stratification [9].

In the past few years, numerous new methods have been developed for learning the above low-dimensional structures from the data. These methods have emerged from many different but related research areas such as signal/image processing, machine learning, computer vision, and pattern recognition. Although they tackle the same kind of problems, these methods have employed drastically different mathematical tools, ranging from statistical, geometrical, algebraic, to graphical tools, and from parametric to nonparametric techniques. These complementary perspectives have significantly improved our understanding of many new statistical and geometric phenomena in high-dimensional data, which we normally do not see in low-dimensional spaces, such as concentration of measure. As a result, people have developed a variety of effective and efficient algorithms that can learn intricate low-dimensional structures for high-dimensional data, even with very limited amount of samples, and even when the samples are incomplete or corrupted. These new methods have seen great success in many important practical applications in image, video, and audio processing, often significantly advancing state of the art.

A timely special section can help consolidate results and efforts from all the related research areas and foster closer collaboration among them. Our goal with this special section is to bring together leading experts in the areas of subspace analysis and manifold learning to jointly explore the impact of these new dimensionality-reduction methods on signal processing, image processing, and many other applications. This special section provides a comprehensive overview of recent developments in these areas, features some new exciting results, and lays out new research problems and directions. We hope that this special issue will help the readers quickly grasp the essence and scope of these research areas, learn state-of-the-art dimensionality-reduction tools, and also apply them to problems in their own research domain.

AN OVERVIEW OF THIS SPECIAL SECTION

More than 60 teams have answered the original call for papers with white papers. Fifteen were invited to submit their full manuscripts, all of which were carefully scrutinized by reviewers, and eventually seven were selected for final publication. The articles range from technical articles that have significant original contributions to survey papers that give timely and comprehensive reviews of certain emerging new topics.

The featured articles cover very representative models and techniques that people have developed in recent years for modeling and extracting low-dimensional structures of high-dimensional data. They include:

- low-dimensional linear subspaces (see "Linear Subspace Learning-Based Dimensionality Reduction" by Xudong Jiang)
- sparse representations and sparsity-promoting dictionary learning, which is mathematically equivalent

Digital Object Identifier 10.1109/MSP.2010.940005
Date of publication: 17 February 2011

to an arrangement of subspaces with equal dimension (see "Dictionary Learning" by Ivana Tošić and Pascal Frossard, and "Learning Low-Dimensional Signal Models" by Lawrence Carin et al.)

- unions of general linear subspaces (see "Subspace Clustering" by René Vidal)
- nonlinear manifolds and mappings (see "Geometric Manifold Learning" by Arta A. Jamshidi, Michael Kirby, and Dave S. Broomhead and "Preimage Problem in Kernel-Based Machine Learning" by Paul Honeine et al.)
- probabilistic distributions (see "Information-Geometric Dimensionality Reduction" by Kevin M. Carter et al.).

We believe that this special section can serve as a good introduction to various topics in this quickly evolving new research area. Therefore, although many of the articles contain original technical contributions, the authors have prepared them in such a way that their pedagogical value is maximized. Indeed, each article has provided a solid review of its topic and ample references to related results so that any interested researchers, especially beginners, can find the articles more useful in terms of getting them familiar with the topic.

For very much the same purpose, the guest editors have compiled a list of related online resources, including tutorials, codes, demos, and data sets of popular methods and algorithms for dimensionality reduction. The list is not meant to be complete but will give the readers, especially beginners, a very good starting point to play with many of the techniques introduced in this special issue.

SOFTWARE AND WEB RESOURCES

DIMENSIONALITY REDUCTION

CODE

Code for various methods (mostly implemented in MATLAB, some in C)

- Matlab Toolbox for Dimensionality Reduction: http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
- ISOMAP: http://isomap.stanford.edu/
- MDS: http://www.newmdsx.com/permap/permap.html
- Robust PCA: http://perception.csl.uiuc.edu/matrix-rank/home.html

TUTORIALS AND DEMOS

- A tutorial that uses the MATLAB Toolbox to compare 14 different methods: http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction_files/TR_Dimensiereductie.pdf
- Slides and a MATLAB demo comparing eight different methods: http://www.math.ucla.edu/~wittman/mani/index.html#downloads
- Java applet demos for PCA: http://www.cs.mcgill.ca/~sqrt/dimr/dimreduction.html
- Tutorial and references on low-rank matrix recovery and completion: http://perception.csl.uiuc.edu/matrix-rank/references.html

COMMON DATA SETS

- Data sets used in the Isomap paper: swiss roll, faces, twos, and hands http://isomap.stanford.edu/datasets.html
- A collection of data sets for digits (USPS, MNIST) and multiview object recognition COIL20: http://www.zjucadcg.cn/dengcai/Data/MLData.html
- Official Web site of the COIL20 data set: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
- Official Web site of the MNIST data set: http://yann.lecun.com/exdb/mnist/

ONLINE ARTICLES

- Wikipedia: http://en.wikipedia.org/wiki/Dimension_reduction
- http://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction

SUBSPACE CLUSTERING

CODE

Code for various methods (mostly implemented in Matlab, some in C)

- GPCA, LSA, RANSAC, and SSC: http://www.vision.jhu.edu/code/
- GPCA and extensions of GPCA: http://perception.csl.uiuc.edu/gpca/home.htm
- ALC: http://perception.csl.uiuc.edu/coding/home.htm
- SCC, k-flats, MPPCA: http://www.math.duke.edu/~glchen/scc.html
- MSL and Improved MSL (motion segmentation): http://www.iim.cs.tut.ac.jp/~sugaya/public-e.html
- Median k-flats: http://math.umn.edu/~zhang620/docs/
- k-means projective clustering: http://www.mathworks.com/matlabcentral/fileexchange/13443-k-means-projective-clustering

COMMON DATA SETS

- Motion segmentation: Hopkins 155 data set and other sequences: http://www.vision.jhu.edu/data/
- Face clustering data sets: http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html http://www.multipie.org/
- Applications of ALC to image segmentation and motion segmentation: http://perception.csl.uiuc.edu/coding/home.htm

SPARSE REPRESENTATION

CODE

- Sparsity Toolbox: This toolbox contains functions related to sparsity optimization in signal processing, including general purpose sparse solvers (such as MP, OMP, BP, IT, KSVD, and MOD): http://www.mathworks.com/matlabcentral/fileexchange/16204
- SPARSELAB: SparseLab is a Matlab software package designed to find sparse solutions to systems of linear equations, particularly underdetermined systems. http://sparselab.stanford.edu/
- SPAMS (SPArse Modeling Software): This is an optimization toolbox that is composed of a set of binaries implementing algorithms to address

(continued on page 126)

Xudong Jiang

Linear Subspace Learning-Based Dimensionality Reduction

A feature extraction module in the pattern recognition system

The ultimate goal of pattern recognition is to discriminate the class membership of the observed novel objects with the minimum misclassification rate. An observed object is often represented by a high-dimensional real-valued vector after some preprocessing while its class membership can be represented by a much lower dimensional binary vector. Thus, in the discriminating process, a pattern recognition system intrinsically reduces the dimensionality of the input data into the number of classes. In fact, dimensionality reduction often occurs implicitly in all modules of a recognition system: preprocessing, feature extraction, and classification. In some applications such as visual object detection and recognition, bioinformatics, and data mining, high data dimensionality imposes great burdens on the robust and accurate recognition due to insufficient knowledge about the data population and limited number of training samples. Dimensionality reduction thus becomes a separate and maybe the most critical module of such recognition systems. Linear subspace analysis is a powerful tool for dimensionality reduction. It also provides a solid foundation for various nonlinear approaches. This is evidenced by numerous techniques published in the past two decades. While some of them, such as sparse representation [1], [2] and subspace arrangements [3], directly solve the classification and clustering problems, most approaches such as the principal component analysis (PCA) [4], linear discriminant analysis (LDA) [4], null-space LDA (NLDA) [5], locality preserving projections (LPP) [6], [7], marginal Fisher analysis (MFA) [8], and their numerous variants serve as a means of feature extraction.

Dimensionality reduction functioning as a feature extraction has two objectives. One objective is to reduce the computational complexity of the subsequent classification with the minimum loss of

Digital Object Identifier 10.1109/MSP.2010.939041
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

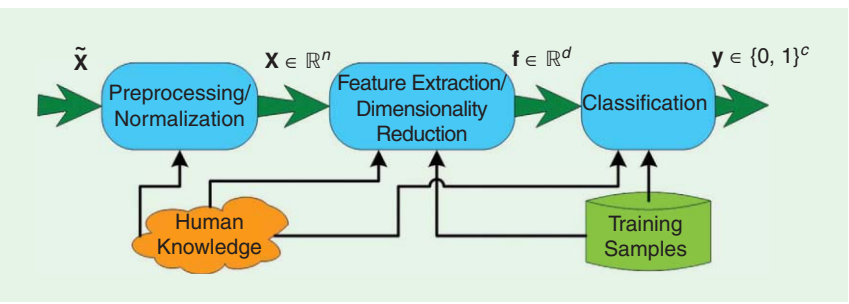
information needed for classification. The second objective is to circumvent the generalization problem of the subsequent classification and hence enhance its accuracy and robustness. To achieve the first objective, it is straightforward that we should maximize the information carried by the data in the extracted low-dimensional subspace. Although PCA does maximize the data structure information in the principal space and hence is optimal for data reconstruction, it is the discriminative information that plays roles in pattern recognition. Thus, most researchers prefer discriminant analysis to the principal component analysis, as evidenced by the fact that the vast majority of the published approaches are based on some kind of the "most discriminative" criteria. There is no doubt that various discriminant analyses can effectively achieve the first objective. The second objective of the dimensionality reduction is, however, far from straightforward. The most discriminative subspace may not be an effective criterion for it because any dimensionality reduction causes a loss of information, including the discriminative information. Any subspace cannot contain more discriminative information than any larger one that includes the former. Why can the dimensionality reduction boost the classification accuracy if the discriminative information is the most critical for classification? Although some general phenomena, such as the curse of dimensionality, small sample size problem, noise removal effect of dimensionality reduction, and better generalization in a lower dimensional space, are well known in the pattern recognition community, they have not indicated what dimensions should be extracted or what else should be removed for a more robust classification. We cannot develop an effective dimensionality reduction technique to maximize the classification accuracy just based on these general phenomena.

ALTHOUGH THE ULTIMATE GOAL OF ALL MODULES OF A PATTERN RECOGNITION SYSTEM IS TO EXTRACT THE MOST DISCRIMINATIVE INFORMATION, IT IS THE MOST DISCRIMINATIVE INFORMATION ABOUT THE WHOLE DATA POPULATION, NOT ON A SPECIFIC TRAINING SET.

It is thus necessary to study the underlying principles and insights of why and how the dimensionality reduction can enhance the generalization accuracy and robustness of the subsequent classification. This is critical because the second objective of the dimensionality reduction is more important than the first one in most applications with the rapid growth of computation power. The study will also help us find the commonalities and differences of various dimensionality reduction techniques and their pros and cons. Without a thorough analysis and gaining an in-depth understanding of the underlying principles, it is difficult to bring the research in this area to a significantly higher level. This article studies the linear subspace learning-based dimensionality reduction as a feature extraction module in the pattern recognition system. Hopefully, some doubts, misunderstandings, ambiguities, and paradoxes in this area can be resolved by this study. For an in-depth analysis, we need to start from some fundamental yet critical issues in pattern recognition and then explore some problems of the statistical classification.

FUNCTIONALITIES OF PATTERN RECOGNITION MODULES

To study how the dimensionality reduction enhances the recognition accuracy, we need to explore the roles of different modules of the recognition system. A statistical pattern recognition system can be partitioned into three modules as shown in Figure 1. The preprocessing/normalization module segments the object of interest from the background, removes noise, and normalizes its representation. This module is usually designed based on some human knowledge to reduce the intraclass variation of patterns with minimum loss of their interclass distinction, i.e., to extract the most discriminative information from the pattern. Although its input $\tilde{\mathbf{x}}$ and output $\mathbf{x}$ may lie in the same domain, e.g., both are images, dimensionality reduction implicitly occurs at this early stage. Among various pattern representations after the first module, we consider the most widely applied vector format $\mathbf{x} \in \mathbb{R}^n$ in an n-dimensional Euclidian space, called data space. The feature extraction/dimensionality-reduction module transfers the pattern from the data space $\mathbf{x} \in \mathbb{R}^n$ to a feature space $\mathbf{f} \in \mathbb{R}^d$. Some approaches are based on the human knowledge about the pattern, e.g., extracting image local structures such as corner, blob, and local orientation [9], [10], and global structures such as Fourier transform and various moments [11]. For many difficult recognition tasks, human beings lack sufficient knowledge about the discriminative features hidden in the data, and hence machine learning from training samples becomes more prevalent. Obviously, the objective of this module is the same as the first one: extracting the most discriminative information. Dimensionality reduction ($d < n$) often explicitly occurs at this intermediate stage. The last module, classification, establishes decision boundaries in the feature space that separate patterns of different classes. As the extracted features are often abstract with little physical interpretation, this module is mainly designed based on the



[FIG1] A general model of the statistical pattern recognition system.

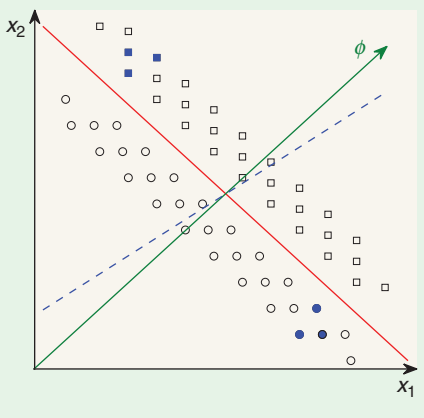
machine learning with limited human interference such as some assumptions of the data distribution model, class prior probability, and loss function. The class label can be represented by a c-dimensional binary vector for a c-class problem. Thus, classification transforms the feature vector, $\mathbf{f} \in \mathbb{R}^d$, into the class label vector, $\mathbf{y} \in \{0, 1\}^c$, which again extracts the most discriminative information and, in most applications, implicitly reduces the dimensionality ($c < d$).

We see from above that all three modules in fact have a common objective but are realized in different ways based on different rules because one way or one rule cannot fully achieve the challenging objective. This common objective in all modules is to extract the most discriminative pattern representations or equivalently, to discard the redundant representations. This is some kind of dimensionality reduction based on some rules generated by human knowledge or machine learning (or both). To understand how the dimensionality reduction in the first two modules helps the final classification, let's explore a simple classification example graphically illustrated in Figure 2.

Suppose the circles and squares in Figure 2 represent the whole data population of two classes, respectively. A classifier can be easily trained by them to form a decision boundary shown by the red solid line, which perfectly classifies all data. Obviously, the dimension spanned by its normal vector ϕ (the green arrow) contains the most discriminative information and the one orthogonal to ϕ has hardly discriminative information. Nevertheless, this redundant dimension causes no harm to the classification because it is ignored by the classifier trained to extract the most discriminative information. Why do we need the first two modules to reduce the dimensionality or to extract the most discriminative pattern representations? It is well known that the probability of misclassification decreases or at least does not increase as the data dimensionality increases, as long as the decision is based on the knowledge about the whole data population. This was theoretically proven in [4], [12], and [13]. However, it is also well known that high dimensionality often degrades the classification performance in practice (curse of dimensionality) [4], [13]. This paradox can be resolved by distinguishing the discriminative information about the data population from that on the training set. The trained classifier can only capture the most discriminative information on the training data. If some statistics estimated on the training data deviate from those of the data population, the misclassification rate on the novel data increases. This is always the case in the practice. The question is only how severe it is. For example, if the available training data are only the blue solid points as shown in Figure 1, the decision boundary of the trained classifier will be the blue dashed line. The misclassification rate on the data population or on the novel data can approach the maximum 50%. The increasing probability of misclassification along with the increase of the data dimensionality for a fixed number of training samples was theoretically proven in [12] with a simple example.

TO BOOST THE CLASSIFICATION ACCURACY, THE DIMENSIONALITY REDUCTION SHOULD BE TARGETED AT REMOVING THE DIMENSIONS UNRELIABLE FOR THE CLASSIFICATION.

If the first two modules can extract only the dimension ϕ based on some human knowledge about the whole data population, the classifier can easily perform a perfect classification in this one-dimensional subspace even if the solid points are the only available training data. This dimensionality reduction is quite possible if proper human knowledge such as some physical characteristics of the pattern is applied in the segmentation and feature extraction. However, if the dimensionality reduction is based on the machine learning from the training samples (the solid points), it cannot extract the right dimension ϕ based on any kind of the "most discriminative" criterion because it in principle just duplicates the classification process. Therefore, some criterion other than the most discriminative should be developed for the dimensionality reduction via machine learning. As the classifier is trained to capture some statistics on the training samples, a problem occurs if they are unreliable in some dimensions (largely deviating from those on the data population). To boost the subsequent classification accuracy or robustness, the dimensionality reduction should be targeted at circumventing this problem. Although the ultimate objective of all modules of a pattern recognition system is to extract the most discriminative information, it is the most discriminative information about the whole data population, not on a specific training set. A classifier is trained to capture the most discriminative information on the training samples. Therefore, to boost the classification accuracy, the dimensionality reduction should be targeted at removing the dimensions unreliable for the classification. Hence, to develop effective techniques of dimensionality reduction via machine learning, we need to study where the possible problem of a statistical classification lies.



[FIG2] **A simple example showing the problem of classification with unrepresentative training samples. The decision boundary (the red solid line) trained by the circles and squares largely deviates from that (the blue dashed line) trained by the solid points.**

PROBLEMS OF CLASSIFICATION, REGULARIZATION, AND SEMIDIMENSIONALITY REDUCTION

Classification is to assign a given novel pattern, here represented by a column vector $\mathbf{x} \in \mathbb{R}^n$ if no feature extraction is imposed, to one of the c categories, ω_i. The minimum probability of misclassification is achieved by assigning the pattern to the class that has the maximum probability after the pattern $\mathbf{x}$ has been observed, called a posteriori probability $P(\omega_i|\mathbf{x})$. This maximum a posterior (MAP) rule is a Bayes decision rule with the 0/1 loss function. It leads to the optimal classification called Bayes classification. As $P(\omega_i|\mathbf{x}) = P(\omega_i)p(\mathbf{x}|\omega_i)p^{-1}(\mathbf{x})$ and $p(\mathbf{x})$ is not a function of ω_i, the Bayes classification is to evaluate the discriminant functions that can be defined as

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \tag{1}$$

and find the class ω_i that has the maximum value of the discriminant function for a given pattern $\mathbf{x}$. Here, a natural logarithm ln is applied as it is a monotonically increasing function that does not affect the decision result but will simplify its evaluation if $p(\mathbf{x}|\omega_i)$ is an exponential function.

Further quantitative analysis needs an analytical form of the class-conditional probability function $p(\mathbf{x}|\omega_i)$. We take the multivariate Gaussian distribution as an example due to several reasons. First, it is the most natural distribution and the sum of a large number of independent random distributions obeys Gaussian distribution. It has the maximum uncertainty of all distributions having a given mean and variance. Moreover, it is an appropriate model for many situations, from handwritten characters to some speech sounds, where the data can be viewed as some prototype corrupted by a large number of random processes [4]. Multiprototype distribution can be well approximated by Gaussian mixture, the weighted sum of a number of Gaussian distributions. Last, dimensionality reduction techniques such as PCA and LDA and many classifiers are only specified by the second-order statistics, and so is the Gaussian distribution. Although LDA, Mahalanobis distance, and many classifiers are proven optimal only under Gaussian assumption, they are successfully employed in many applications. Under the Gaussian assumption

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)\right], \tag{2}$$

the discriminant function (1) becomes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \mathrm{b}_i. \tag{3}$$

In practice, b_i is often not strictly determined by (1) and (2) but used as a threshold for users to control the error rate of class ω_i at a price of the other classes, e.g., to compromise between the false acceptance and false rejection rates in a biometric verification or object detection application.

> THE LARGE DEVIATIONS OF THE SMALL EIGENVALUES FROM THE POPULATION VARIANCES RESULT IN A SEVERE OVER-FITTING PROBLEM OF THE CLASSIFIER THAT GREATLY AFFECTS THE CLASSIFICATION ACCURACY ADVERSELY.

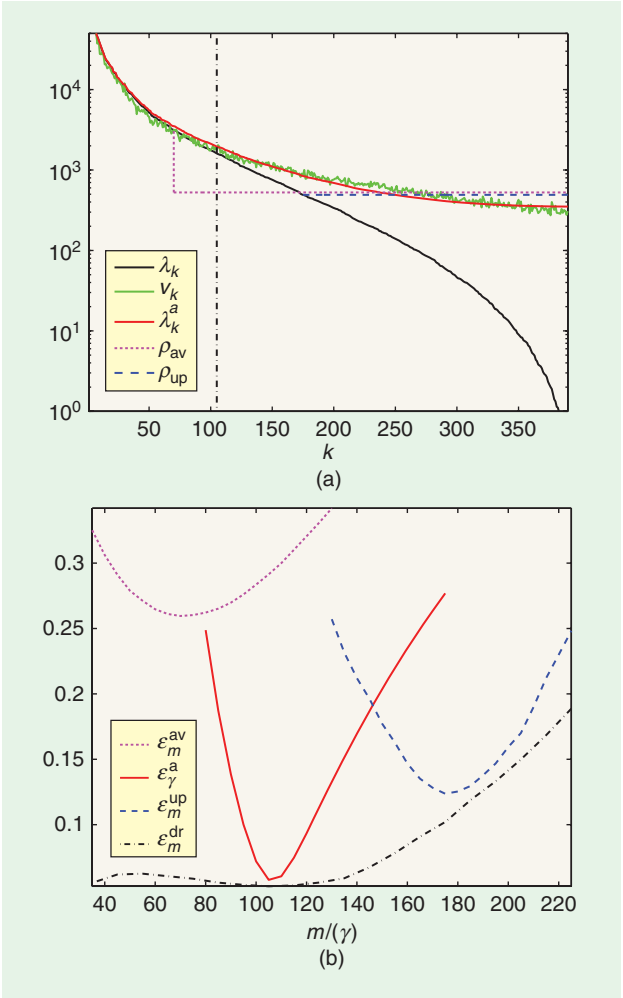
The problem is that human knowledge cannot provide the class mean $\bar{\mathbf{x}}_i$ and covariance matrix $\mathbf{\Sigma}_i$ of the data population, which can only be estimated or learned by machine from the available training samples. If some estimates largely deviate from those of the data population, we will face a large misclassification rate. From (3) we see that the discriminant function is very sensitive to the covariance matrix $\mathbf{\Sigma}_i$ because the data vector is multiplied by its inverse. However, it is very difficult to study the problems of $\mathbf{\Sigma}_i$ directly as it carries two different kinds of information by n^2 estimates: data variations and correlations. Eigen-decomposition provides an effective tool to simplify the problem. As the covariance matrix is symmetric, its eigenvectors provide an orthogonal basis for n-space. After applying eigen-decomposition, $\mathbf{\Phi}_i^T\mathbf{\Sigma}_i\mathbf{\Phi}_i = \mathbf{\Lambda}_i = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, the discriminant function (3) is simplified as

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\mathbf{\Phi}_i\mathbf{\Lambda}_i^{-1}\mathbf{\Phi}_i^T(\mathbf{x}-\bar{\mathbf{x}}_i) + b_i \\ &= -\frac{1}{2}\sum_{k=1}^{n}\frac{(z_k-\bar{z}_k)^2}{\lambda_k} + b_i, \end{aligned} \tag{4}$$

where z_k and $\bar{z}_k$ are respectively the projections of $\mathbf{x}$ and $\bar{\mathbf{x}}_i$ on the orthonormal eigenvector Φ_k corresponding to the eigenvalue λ_k of $\mathbf{\Sigma}_i$. For symbolic simplicity, the class index i is omitted where the index k is necessary. As an eigenvalue λ_k is the variance of the training samples of a class projected on the eigenvector Φ_k, it is an estimate of the class population variance based on the available training data. If it deviates from the population variance, the decision rule (3) or (4) overfits the training samples and hence leads to a poor generalization or prediction on the novel testing data. This problem will become very severe if some eigenvalues largely deviate from the population variances.

The black curve of Figure 3(a) shows an eigen-spectrum (λ_k sorted in descending order) obtained from 400 face images of size 20×20 and the green curve shows the variances v_k of other 8,500 face images (representing the face population) projected on the eigenvectors Φ_k. They are plotted in logarithm scale for comparison because we see from (4) that it is not the amount of the difference but the amount of the ratio between λ_k and v_k that affects the accuracy of the discriminant function (4). All images are taken from a face detection database used in [14]. Other sets of training images produce results very similar to Figure 3. It shows deviations between the eigenvalues and the population variances. One way to quantify this disparity over the range space is to compute $e(\lambda) = \mu\{(\ln v_k - \ln\lambda_k)^2\}_{1\le k\le r}$, where $\mu\{\cdot\}_{1\le k\le r}$ is a mean operator over $1 \le k \le r$ and r is the rank of $\mathbf{\Sigma}_i$.

Figure 3 shows significantly larger deviations of the smallest eigenvalues. This phenomenon was elucidated in [14], [15], and [16], where more examples on several other real data sets can be



[FIG3] Problems of eigenvalues and their regularization. Part (a) shows the eigen-spectrum λ_k and its regularized versions computed from 400 face images, and variances v_k of other 8,500 face images projected on the eigenvectors Φ_k. Part (b) shows the normalized disparity between the regularized eigen-spectrum and the variances v_k.

found. It seems to be a general problem verified in [14] by synthetic data with known true population variances. Although, in general, the largest sample-based eigenvalues are biased upwards and the smallest ones are biased downwards, the bias is more pronounced when the population variances tend toward equality, and it is correspondingly less severe when their values are highly disparate [15]. In most applications, population variances often first decay very rapidly and then stabilize so that the smallest eigenvalues are biased much more severely than the largest ones [14], [16]. This is evidenced by Figure 3. The large deviations of the small eigenvalues from the population variances result in a severe overfitting problem of the classifier that greatly affects the classification accuracy adversely.

One solution is to regularize the covariance matrix $\mathbf{\Sigma}_i$. A common practice in classification and data regression is to add a constant to its diagonal elements, $\mathbf{\Sigma}_i^a = \mathbf{\Sigma}_i + a\mathbf{I}$. We can let $a = \gamma\,\mathrm{trace}(\mathbf{\Sigma}_i)/r$ so as to select γ invariably to the data scale. The normalized disparity of the regularized eigen-spectrum $\varepsilon_\gamma^a = e(\lambda^a)/e(\lambda)$ against γ is shown by the red curve of Figure 3(b). Its minimum is $\varepsilon_{0.08}^a = 0.06$. The regularized eigen-spectrum λ_k^a with $\gamma = 0.08$ is shown by the red curve of Figure 3(a). Although this method was originally proposed to circumvent the singularity of $\mathbf{\Sigma}_i$ and the numerical instability of its inverse, we see from Figure 3 that the regularized eigen-spectrum can be very close to the population variances. It is thus not a surprise that numerous algorithms for classification, data regression, dimensionality reduction, and manifold learning adopt this classical technique [15], [17]–[19]. The underlying principle of $\mathbf{\Sigma}_i^a = \mathbf{\Sigma}_i + a\mathbf{I}$ can been seen by its equivalence to adding the constant to all eigenvalues $\lambda_k^a = \lambda_k + a$. From $(\lambda_k + a)/v_k = (1 + a/\lambda_k)\lambda_k/v_k$, we see that the factor $(1 + a/\lambda_k)$ is larger for smaller λ_k and smaller for larger λ_k. Therefore, the regularized eigen-spectrum can be very close to the population variances as shown in Figure 3. Problems of this method are the increased disparity of large eigenvalues and no dimensionality reduction effect. Either the $n \times n$ covariance matrix or the $n \times n$ eigenvector matrix is needed to compute the discriminant function (3) or (4).

Another solution, called probabilistic subspace learning [20], [21], decomposes the discriminant function (4) into two parts and replaces the small eigenvalues by a constant as

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[\sum_{k=1}^{m}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} + \sum_{k=m+1}^{n}\frac{(z_k - \bar{z}_k)^2}{\rho}\right] + b_i. \quad (5)$$

The constant is computed by $\rho_{av} = \mu\{\lambda_k\}_{m<k\le r}$ in [20] and [21] as it is the optimal approximation to λ_k for $m < k \le r$. This method leads to one of the best performers, called the Bayesian algorithm [22], in the face recognition community and is adopted in many other approaches of visual object recognition [23]–[25]. In fact, this method regularizes the eigen-spectrum by setting $\lambda_k^{\rho_{av}} = \rho_{av}$ for $m < k \le n$. The normalized disparity $\varepsilon_m^{av} = e(\lambda^{\rho_{av}})/e(\lambda)$ against m is shown by the magenta dotted curve of Figure 3(b). Its minimum is $\varepsilon_{70}^{av} = 0.26$. The regularized eigen-spectrum $\lambda_k^{\rho_{av}}$ for $70 < k \le n$ is shown by the magenta dotted line in Figure 3(a). We see a much greater disparity than λ_k^a. The problem is the computation of the constant ρ. The purpose of the regularization is not best approximating to the eigenspectrum but to the population variances. Eigenvalues in the subspace $m < k \le n$ are replaced by a constant ρ because they are unreliable, and so is their arithmetic average ρ_{av}. As they are biased downwards, it is proposed in [26] to use their upper bound as the constant $\rho_{up} = \max\{\lambda_k\}_{k>m}$, which is also adopted in [27]. The normalized disparity $\varepsilon_m^{up} = e(\lambda^{\rho_{up}})/e(\lambda)$ against m is shown by the blue dashed curve of Figure 3(b). Its minimum is $\varepsilon_{175}^{up} = 0.12$. The regularized eigen-spectrum $\lambda_k^{\rho_{up}}$ for $175 < k \le n$ is shown by the blue dashed line in Figure 3(a). We see a much smaller disparity than λ_k^{av}, which is greater than ε^a in this example but smaller than it in another (Figure 4). The upper bound ρ_{up} leads to significantly higher face recognition accuracy than the average ρ_{av} [26].

In fact, this regularization has some role of dimensionality reduction as it is not necessary to project the data to the

eigenvectors Φ_k for $k > m$. As we choose orthonormal eigenvectors, Euclidian distance between two vectors in the eigen-space is identical to that in the data space and hence,

$$\sum_{k=m+1}^{n} \frac{(z_k - \bar{z}_k)^2}{\rho} = \frac{1}{\rho}\left[(\mathbf{x} - \bar{\mathbf{x}}_i)^T(\mathbf{x} - \bar{\mathbf{x}}_i) - \sum_{k=1}^{m}(z_k - \bar{z}_k)^2\right]. \quad (6)$$

Thus, we only need $n \times m$ eigenvector matrix to compute (5) for classification. However, the n-dimensional class mean vectors $\bar{\mathbf{x}}_i$ are still required. We call it semidimensionality reduction.

Besides adding a constant to all eigenvalues or replacing the unreliable eigenvalues by a constant as discussed above, another regularization technique [16] replaces the unreliable eigenvalues λ_k, $m < k \le r$ by a model $\alpha(k + \beta)^{-1}$, where α and β are two constants determined by the reliable eigenvalues. The rationale behind it is that the population variance is not constant in the unreliable subspace but decays much slower than the eigenvalue does. This decaying nature can be modeled by $\alpha(k + \beta)^{-1}$, which will be certainly closer to the population variances than the constant ρ_{av} or ρ_{up} if proper values of α and β are chosen.

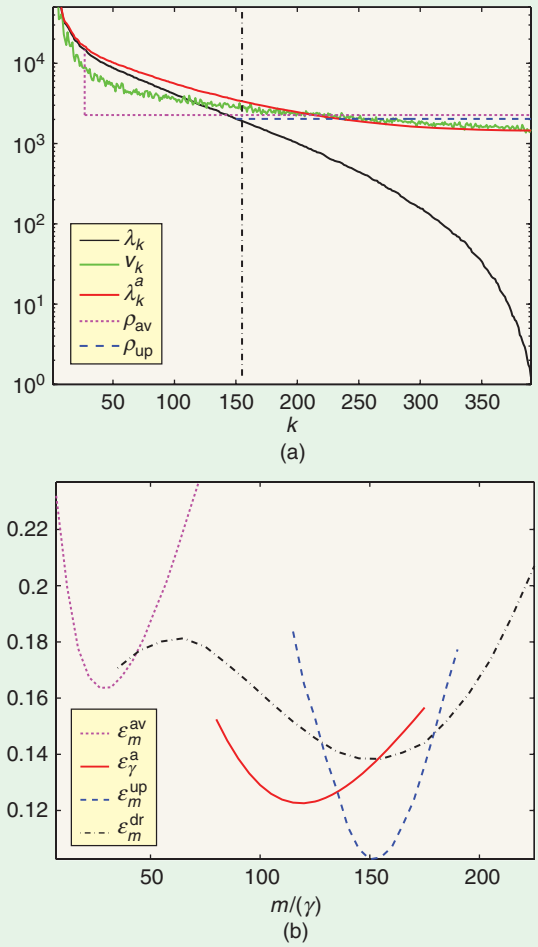
DIMENSIONALITY REDUCTION FOR REMOVING UNRELIABLE DIMENSIONS

Various regularization techniques that greatly improve the classification accuracy are evidenced by a large amount of publications. As analyzed in the last section, the underlying principle behind the regularization is that it reduces the disparity between the eigenvalues and the population variances and hence attenuates the overfitting problem. Obviously, we can also remove the unreliable dimensions to reduce the disparity in the remaining subspace. The normalized disparity of the eigen-spectrum in the subspace $(1 \le k \le m)$ against m, $\varepsilon_m^{dr} = e(\lambda^{dr})/e(\lambda)$, is shown by the black dot-dashed curve in Figure 3(b). The minimum is $\varepsilon_{105}^{dr} = 0.05$. The extracted and removed subspaces resulting in the minimum ε^{dr} are separated by a vertical black dot-dashed line in Figure 3(a). It shows that the dimensionality reduction effectively reduces the disparity because large disparity occurs at small eigenvalues. Therefore, similar to various regularization techniques that modify the smallest eigenvalues, removing the subspace spanned by the eigenvectors corresponding to the smallest eigenvalues improves the inference of the classifier, i.e., reduces the misclassification rate on the novel testing data.

However, this dimensionality reduction may also reduce the interclass distinction and the discriminant functions (3) or (4) of different classes in general should be evaluated in a common feature subspace for comparison. To extract a common subspace reliable for all classes and to prevent possible significant loss of the interclass distinction, we combine all class-conditional covariance matrices plus the covariance matrix of class mean to create a covariance mixture as

$$\mathbf{\Sigma}_\alpha = \sum_{i=1}^{c} \alpha_i \mathbf{\Sigma}_i + \eta \mathbf{\Sigma}_\mu, \quad (7)$$

where α_i and η are weights and



[FIG4] Parts (a) and (b) show results of the same program as of Figure 3 but using 400 nonface training images and 8,500 nonface test images.

$$\mathbf{\Sigma}_\mu = \sum_{i=1}^{c} \frac{q_i}{q}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^2. \quad (8)$$

The covariance matrix of class mean $\mathbf{\Sigma}_\mu$ is also called interclass scatter matrix, where q_i is the sample size of class ω_i, and $\bar{\mathbf{x}}$ and q are respectively the mean and the sample size of the whole training set. Eigen-decomposition is then applied to the constructed covariance mixture

$$\mathbf{\Phi}^T \mathbf{\Sigma}_\alpha \mathbf{\Phi} = \mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}. \quad (9)$$

If we remove a subspace spanned by eigenvectors corresponding to the smallest eigenvalues of $\mathbf{\Sigma}_\alpha$, it tends to remove unreliable dimensions of all class-conditional covariance matrices $\mathbf{\Sigma}_i$ and retain large interclass distinction residing in a subspace that has large eigenvalues of $\mathbf{\Sigma}_\mu$. Therefore, classification on the m-dimensional feature vector

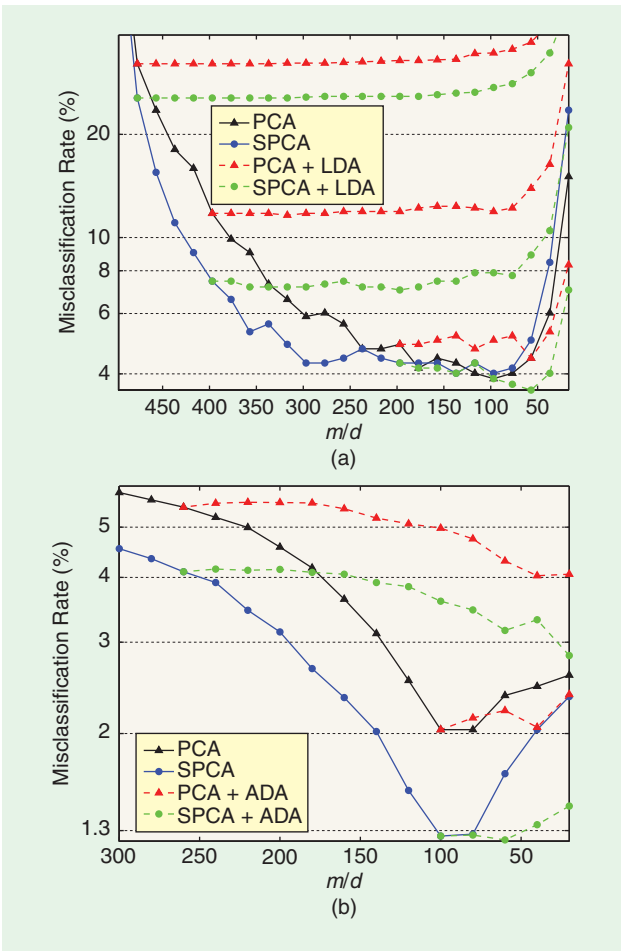
$$\mathbf{f} = \mathbf{\Phi}_m^T \mathbf{x} \quad (10)$$

is most likely to perform better than on the n-dimensional data vector $\mathbf{x}$, where $\mathbf{\Phi}_m$ consists of m eigenvectors corresponding to the m largest eigenvalues of the covariance mixture $\mathbf{\Sigma}_\alpha$. If we regard (7), (9), and (10) as a separate module called dimensionality reduction, the subsequent classification (3) is simplified in the m-dimensional subspace

$$g_i(\mathbf{f}) = -\frac{1}{2}(\mathbf{f}-\bar{\mathbf{f}}_i)^T\mathbf{\Sigma}_{fi}^{-1}(\mathbf{f}-\bar{\mathbf{f}}_i) + b_i. \tag{11}$$

It is not necessary to project training samples into the subspace as $\bar{\mathbf{f}}_i = \mathbf{\Phi}_m^T\bar{\mathbf{x}}_i$ and $\mathbf{\Sigma}_{fi} = \mathbf{\Phi}_m^T\mathbf{\Sigma}_i\mathbf{\Phi}_m$. The objective of the dimensionality reduction by (7), (9), and (10) is to facilitate an effective removal of the unreliable dimensions and hence boost the classification accuracy (11). Thus, larger values of α_i should be assigned to less reliable covariance matrices so that more dimensions characterized by the smallest eigenvalues of less reliable classes can be removed by the eigen-decomposition of $\mathbf{\Sigma}_\alpha$.

The weights α_i are critical in some applications because different classes may have different characteristics and hence the reliability of the estimated covariance matrices can be significantly different. Figure 4 is generated by the same program as Figure 3 but uses 400 and 8,500 nonface training and testing images, respectively, from a face detection database used in [14]. Other partitions of the training and testing sets produce very similar results to Figure 4. It shows much larger disparity between the eigenvalue and the population variance than that of Figure 3. More examples can be found in [14]. This is a general problem caused by the different characteristics of different classes. In the applications of biometric verification and object detection, for example, the positive and negative classes are highly asymmetric because the former represents only one particular person or object while the latter represents the whole "rest of the world" that contains all other people or objects. Thus, it is much more difficult to collect a representative training set for the negative class than for the positive one. This often results in a larger eigenvalue bias of the negative class. Furthermore, as pointed out in [15] and further evidenced by Figures 3 and 4, the bias is more pronounced when population variances tend toward equality, and less severe when their values are highly disparate. This is also verified in [14] by synthetic data with known true population variances. As the negative class occupies a much larger subspace and hence has flatter eigen-spectrum, in general, we need to assign a larger weight to the negative class than to the positive one.



[FIG5] Misclassification rate against the reduced dimensionality by PCA/SPCA and PCA/SPCA+LDA/ADA of (a) face identification and (b) face detection problems. The left-most point of each dashed curve indicates the dimensionality *m* of the PCA/SPCA subspace in which LDA/ADA further reduces it to *d* indicated by the other points on the same dashed curve.

It is very interesting to see that if we set $\eta = 1$ and $\alpha_i = q_i/q$, the constructed covariance mixture $\mathbf{\Sigma}_\alpha$ will be identical to the covariance matrix $\mathbf{\Sigma}_t$ of all training data without considering their class labels. It is also called a total scatter matrix. This shows that the well-known PCA is a specific case of the aforementioned dimensionality reduction method. Therefore, this study also reveals the underlying principle of why PCA, though an unsupervised method that minimizes the data reconstruction error rather than maximizes the class discrimination, can improve the classification accuracy. Although many approaches apply PCA only aimed at circumventing the singularity problem of the intraclass scatter matrix for the subsequent discriminant analysis, as analyzed above, the role of PCA for classification is in fact far beyond that. Figure 5 (refer to the experimental section) demonstrates the significant gains in classification accuracy by using PCA to reduce the dimensionality much lower than the rank of the intraclass scatter matrix. More evidence can be found in the experimental results of [8], [14], [19], [25], and [28].

Nevertheless, PCA is not optimized for classification. The weights $\eta = 1$ and $\alpha_i = P(\omega_i)$ or $\alpha_i = q_i/q$ are required for PCA to achieve the least-mean-square data reconstruction error, which is irrelevant to classification. Our objective for classification is to remove the unreliable dimensions in which the sample-based class-conditional variances are largely deviate from the population variances. The reliability of a covariance matrix does not depend on the class prior probability. More training samples of a class may result in a more reliable covariance matrix if they are properly collected. However, it is the less reliable covariance matrix that should be heavier weighted in the covariance mixture so that more dimensions characterized by the small variances of this class can be removed. From the analysis, we see that PCA helps improve the

classification accuracy, not because it minimizes the data reconstruction error, but because it has some roles in removing the unreliable dimensions. As its objective is not from the classification point of view, PCA may not effectively remove the unreliable dimensions. In sharp contrast to PCA that weights $\mathbf{\Sigma}_i$ proportionally to q_i, it is suggested in [14] to pool $\mathbf{\Sigma}_i$ with weights inversely proportional to q_i if there is no prior knowledge about the class characteristics and the data collection procedure. Even $\eta = 1$ in PCA may not be optimal for classification. Although a larger value of η ensures less loss of the interclass distinction, it leads to less effective removal of the unreliable dimensions. Hence, more dimensions have to be removed, which in turn results in more loss of the interclass distinction. The aforementioned limitations of PCA are verified by the experimental results shown in Figure 5.

PCA is an unsupervised technique, as no class label is needed. For a two-class problem, dimensionality reduction (7), (9), (10) is called asymmetric principal component analysis (APCA) [14] due to the asymmetric treatment of the two covariance matrices. For a multiple-class problem, more generally, we call it supervised principal component analysis (SPCA), as it utilizes the class label and other class-specific information by imposing different weights on the covariance matrices. The optimal values of weights are application dependent. The objective of the SPCA (7), (9), (10) is to effectively remove the unreliable dimensions and hence boost the classification accuracy (11). Thus, it may not greatly reduce the dimensionality for a fast classification in some applications. In addition, APCA or SPCA may not work well for a classifier that neither explicitly nor implicitly weights the feature by the inverse of its variance, such as the classical nearest-neighbor classifier (NNC) with Euclidian distance and the sparse representation-based classifier (SRC) where the ℓ^1-minimization is applied [1], [2].

DIMENSIONALITY REDUCTION BY RETAINING DISCRIMINATIVE DIMENSIONS

As discussed in the last two sections, if the dimensionality reduction is aimed at enhancing the inference accuracy of the subsequent classification, it should be targeted at removing the unreliable dimensions. In some applications, we need to reduce a very high dimensional data vector to a very low dimensional feature vector to facilitate a simple and fast classification. This can be effectively achieved by extracting the most discriminative dimensions, which ensures the minimum loss of the discriminative information in the extracted subspace among all other subspaces of the same dimensionality. Linear discriminant analysis and its various variants are the most widely studied approaches.

In the identification applications, we often have a large number of classes with only a few samples per class for training so that each individual $\mathbf{\Sigma}_i$ is extremely unreliable. One solution to regularize them is to pool them together to form a common covariance matrix, $\mathbf{\Sigma}_w = \sum_{i=1}^{c} q_i \mathbf{\Sigma}_i / q$, which is also called intraclass scatter matrix. The discriminant function (3) is thus simplified as

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{\Sigma}_w^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) + \mathrm{b}_i = \mathbf{x}^T \mathbf{\Sigma}_w^{-1} \bar{\mathbf{x}}_i + t_i, \quad (12)$$

where t_i absorbs all terms that is either constant to $\mathbf{x}$ or constant to i. We see that it is a linear function of $\mathbf{x}$ and hence the decision boundary $g_i(\mathbf{x}) = g_j(\mathbf{x})$ between any two classes ω_i and ω_j is a hyperplane specified by its normal vector $\boldsymbol{\psi}_{ij} = \mathbf{\Sigma}_w^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$ and the threshold $t_i - t_j$. This means that for the optimal classification between two classes ω_i and ω_j, only one dimension spanned by $\boldsymbol{\psi}_{ij}$ is necessary. Thus, under the constraint of the linear classification, this dimension contains the most (in fact, all) discriminative information to differentiate class ω_i and ω_j. It is easy to see that the training data in this dimension have the maximum ratio κ between the interclass and intraclass variances. Therefore, we can define this ratio κ as a discriminant value to assess the discriminating power of a dimension. Although we need $(c-1)!$ hyperplanes to classify c classes, their normal vectors $\boldsymbol{\psi}_{ij} = \mathbf{\Sigma}_w^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$ only span a $(c-1)$-dimensional subspace as only $c-1$ of them are linear independent. Therefore, we can reduce the n-dimensional data space to this $(c-1)$-dimensional subspace without losing any discriminative information as the linear classification (12) produces exactly the same results in the two spaces. However, if the dimensionality is reduced to d, $d < c-1$, some discriminative information will be lost. The subspace spanned by the eigenvectors corresponding to the d largest eigenvalues of the matrix $\mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_\mu$ contains the most discriminative information among all possible d-dimensional subspaces for the linear classification (12) because an eigenvalue of $\mathbf{\Sigma}_w^{-1}\mathbf{\Sigma}_\mu$ is the ratio κ between the interclass and intraclass variances in the dimension spanned by the corresponding eigenvector. This is the well-known LDA that performs the eigen-decomposition

$$\mathbf{\Psi}^T \mathbf{\Sigma}_w^{-1} \mathbf{\Sigma}_\mu \mathbf{\Psi} = \mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ \ldots, \lambda_n\}. \quad (13)$$

We see from the above analysis that the objective of LDA is to find the one among all possible d-dimensional subspaces in which the linear classification (12) achieves the closest result to that in the original n-space. It is undoubtedly an effective method to largely reduce the data dimensionality with the minimum loss of the classification capability in a linear sense.

For a two-category classification problem, LDA can only extract one dimension. It is insufficient for a reasonable classification for some problems such as various tasks of verification and object detection because the two class-conditional covariance matrices are significantly different and hence the optimal classification is obviously not linear. To apply the discriminant analysis in such problems, an asymmetric discriminant analysis (ADA) is proposed in [14] to extract a rich number of features. It solves the following eigen-decomposition problem:

$$\mathbf{\Psi}^T(\mathbf{\Sigma}_1 + \beta\mathbf{\Sigma}_2)^{-1}(\mathbf{\Sigma}_1 + \gamma\mathbf{\Sigma}_\mu)\mathbf{\Psi} = \mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \quad (14)$$

in the APCA subspace. The underlying principle is that the discriminative information is not only carried by the distinction of the two class means but also by the distinction of the two class variances. The constant γ weights the discriminative information about the class mean against that about the variance. The asymmetry of the two classes is balanced by the constant β. It is proven

in [14] that the ADA with $\gamma = \beta = 1$ maximizes the Bhattacharyya distance [29] between two classes in the subspace spanned by the eigenvectors corresponding to the largest $\max(\lambda_k, 1 - \lambda_k)$. Note that, different from those in the last two sections, eigenvectors of LDA or ADA are not orthogonal. The Euclidian distance in a space using LDA/ADA eigenvectors as a base will be different from that using an orthogonal base. However, it is easy to show that the Mahalanobis distance is not affected by the orthogonality of the base.

As the rank of $\mathbf{\Sigma}_w$ is at most $\min(n, q - c)$, $\mathbf{\Sigma}_w$ is often singular in some applications so that the discriminant value of LDA (13) and ADA (14) cannot be evaluated. Numerous variants or generalizations of LDA have been proposed to circumvent this problem, which are summarized under a common framework graphically [8] and algebraically [19]. A popular approach called Fisherface or Fisher LDA (FLDA) [30] applies PCA so as to make $\mathbf{\Sigma}_w$ nonsingular before LDA. Another approach called direct LDA (DLDA) [17] removes null space of $\mathbf{\Sigma}_\mu$ and extracts the eigenvectors corresponding to the smallest eigenvalues of $\mathbf{\Sigma}_w$. This is under the assumption that the most discriminative information resides in the range space of $\mathbf{\Sigma}_\mu$. NLDA [5] extracts features from the null space of $\mathbf{\Sigma}_w$. Interestingly, this appears to contradict the popular FLDA that only uses the range space and discards the null space of $\mathbf{\Sigma}_w$. A common aspect of all these methods is that they all remove some dimensions, either in the principal or the null space, before the LDA process. It is difficult to compare the effectiveness of the aforementioned LDA variants because we see from (13) that both $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_\mu$ contribute to the discriminant value κ in a dimension. NLDA and DLDA appear to retain more discriminative information as any dimension in the intersection of the null space of $\mathbf{\Sigma}_w$ and the range space of $\mathbf{\Sigma}_\mu$ has infinite discriminant value κ according to (13). DLDA ensures the class mean distinction $\mathbf{\Sigma}_\mu$ untouched in the first stage. However, small and zero eigenvalues of $\mathbf{\Sigma}_w$ are unreliable that may cause severe problem as we analyzed in the last two sections. Just a small decrease or increase in the number of training samples may greatly change them. Furthermore, the most discriminative dimensions are not restricted within the range space of $\mathbf{\Sigma}_\mu$ or the null space of $\mathbf{\Sigma}_w$. Therefore, the above LDA variants are criticized in the literature [27], [31], [32] as a significant amount of discriminative information could be lost before the LDA process.

To avoid losing discriminative information before the LDA process, the dual-space LDA approach (DSL) [31], [32] performs LDA on the principal space of $\mathbf{\Sigma}_w$ and its complementary space separately and combines the two sets of the extracted LDA features. Obviously, it is suboptimal to extract features separately from the two subspaces. Furthermore, how to fuse the two feature sets properly is an open problem as they do not share the same metric measurement. Features from the principal space, $k \le m$, are weighted by the inverse of their intraclass variance and those from the complementary space, $k > m$, are equally weighted by some constant. From the last two sections, we see that this feature weighting is problematic in the principal space for a large value of m and is problematic in the complementary subspace for a small value of m. One solution to these problems is first to partition the data space into three subspaces: reliable, unreliable, and null space of $\mathbf{\Sigma}_w$, then to regularize the eigenvalues differently in these three subspaces and finally to apply LDA in the whole space [16]. Consistent gains in face recognition accuracy of this approach were reported in [16]. Another way [33] to avoid losing information of $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_\mu$ before the discriminant evaluation is to modify the LDA (13) to

$$\mathbf{\Psi}^T \mathbf{\Sigma}_t^{-1} \mathbf{\Sigma}_\mu \mathbf{\Psi} = \mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ \ldots, \lambda_n\}. \tag{15}$$

As $\mathbf{\Sigma}_t = \mathbf{\Sigma}_w + \mathbf{\Sigma}_\mu$ and hence the null space of $\mathbf{\Sigma}_t$ is the intersection of the null spaces of $\mathbf{\Sigma}_\mu$ and $\mathbf{\Sigma}_w$, no discriminative information is lost by evaluating (15) in the range space of $\mathbf{\Sigma}_t$. However, (15) deviates from the LDA (13) and hence the extracted subspace may not be the most discriminative in a sense of LDA or of the classification (12). Moreover, it puts an undue emphasis on the null space of $\mathbf{\Sigma}_w$ as the discriminant value from $\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_\mu$ in the range space of $\mathbf{\Sigma}_w$ ($\kappa < 1$) is always smaller than that in its null space ($\kappa = 1$). In addition, there is also a problem of how to properly scale the features from the principal and null spaces of $\mathbf{\Sigma}_w$, which may not be of full rank even in the reduced subspace.

Most aforementioned approaches focus on the singularity problem of $\mathbf{\Sigma}_w$. In fact, as analyzed in the last two sections, the unreliability, bias, and instability of the small eigenvalues of $\mathbf{\Sigma}_w$ cause great problems wherever its inverse is applied in the discriminant evaluation (13), (14) or the classification (3), (12). Any regularization or dimensionality reduction technique discussed in the last two sections can be applied to attenuate this problem before applying discriminant analysis to further reduce the dimensionality for a fast classification. Significant gains in classification accuracy were reported by applying various regularization techniques in the LDA approaches [15], [16], [19], [27], [34]. Also, great gains in classification accuracy were reported by applying PCA, APCA, or SPCA to reduce the data dimensionality much smaller than the rank of the $\mathbf{S}_w$ before applying LDA or other discriminative methods [8], [14], [19], [25], [28].

EXPERIMENTAL STUDIES

The different roles of dimensionality reduction by PCA, SPCA, and LDA/ADA for pattern recognition are further explored in two experiments. One is a face identification problem on a data set [16] extracted from the facial recognition technology (FERET) database with many classes (1,194 people) and only two samples per class, and the other is a face detection problem in the database used in [14] with only two classes (face and nonface) and many (9,000) samples per class. Images are cropped into the size of 33×38 for the identification problems and 20×20 for the detection problem. In the identification experiment, 497 people are randomly selected for training, the remaining 697 people are used for testing, and the linear classifier (12) is applied in the feature space. In the detection experiment, four experiments, each with a distinct 25% images as testing set and the remaining images as training set, are conducted and the average misclassification rate over the four distinct testing sets is computed. The detection

applies the quadratic classifier (11) for PCA-related approaches and its asymmetric version with $\beta = 0.75$ [14] for SPCA-related approaches, where b_i is set so that the two classes have the same misclassification rate. For the identification problem, as there is no ground for significantly different distributions of different persons, the same parameter $\alpha_i = 1/497$ is chosen for $\mathbf{\Sigma}_\alpha$. We choose $\eta = 1/4$ to differentiate the covariance mixture $\mathbf{\Sigma}_\alpha$ significantly from the total scatter matrix $\mathbf{\Sigma}_t$ where $\eta = 1$. For the detection problem, we choose $\eta = 1$ (same as in PCA) but $\alpha_1 = 1/5$ (for face class) and $\alpha_2 = 4/5$ (for nonface class) to remove significantly more unreliable dimensions of the nonface class in the SPCA stage as discussed in the last section. For ADA, $\gamma = 10$ and $\beta = 0.75$ is chosen [14]. Figure 5 shows the misclassification rates against the dimensionality m reduced by PCA and SPCA and d reduced by PCA+LDA/ADA and SPCA+LDA/ADA. The most left point of each dashed curve indicates the dimensionality m of the PCA/SPCA subspace, in which the LDA/ADA further reduces it to d indicated by the other points on the same dashed curve.

The experimental results shown in Figure 5 further verify the analysis of this article. It is the regularization technique or the dimensionality reduction by the supervised principal component analysis (including PCA) that plays the most vital role in boosting the classification accuracy while the discriminative method can greatly reduce the dimensionality with the minimum loss of the discriminative information. The question may arise as to why NLDA can work well in some applications if the smallest and zero eigenvalues are the most unreliable. The reason behind it is that the classification of the NLDA features does not use the variance due to zero eigenvalues in all dimensions of the null space. Thus, it implicitly circumvents the problem of the unreliable small eigenvalues to a certain extent by evenly weighting all features. Another question is why some approaches using LDA alone can also work well on some data sets. The underlying causes include the avoidance of feature scaling in the classification and the linearity of LDA but the nonlinearity of the classifier. These approaches, though applying LDA (13) for feature extraction, do not apply its origin (12) as classifier. Most of them apply the NNC with Euclidian distance. While the simple Euclidian distance ignores the data variance and hence circumvents the problem of the unreliable small eigenvalues to a certain extent, the complex data distribution is captured by the NNC that computes all distances from a novel pattern to all training samples. The NNC, though very simple, is highly nonlinear, can form arbitrary complex, nonlinear decision boundary and classifies all training samples without error. LDA restricts such highly nonlinear classifier to a subspace, which is, though the most discriminative, only in a linear sense. This restriction has similar role to the regularization. Therefore, the improvement of the classification accuracy by LDA is most likely contributed by its linearity constraint rather than its most discriminative nature. However, LDA that represents the class distinction by using the difference of class mean only may impose too strict constraint on some complex data structure. Therefore, some approaches that utilize the locality and neighborhood of the training samples such as LPP [6], [7] and MFA [8] extract more discriminative features than LDA. Nevertheless, experiments in [7] and [8] still show that a PCA stage either is necessary to "remove the noise" [7] or significantly improves the performance [8] of these discriminative approaches.

CONCLUSIONS

To recognize unknown data, a pattern recognition system is designed based on the human knowledge about the data population and the machine learning from the known training samples. The difficult recognition task is performed in several stages. Classification as the last stage is mainly trained by the available training samples. Thus, it extracts the most discriminative information on the training data, which in general deviates from that about the whole data population as only a finite set of training samples is applicable. This deviation increases the misclassification rate on the novel data. The problem becomes very severe if the data lie in a high-dimensional space. Moreover, high dimensionality also makes it difficult to apply sophisticated classifiers. Linear subspace learning-based dimensionality reduction provides a powerful tool to circumvent these problems. It also serves as a solid foundation for various nonlinear approaches. Dimensionality reduction as an intermediate stage of a pattern recognition process has two objectives. One is to reduce the computational complexity of the subsequent classification with the minimum loss of the discriminative information, and the other is to circumvent the over-fitting problems of the classification and hence enhance its inference accuracy and robustness.

To achieve the first objective, we need to maximize the discriminative information in the reduced low-dimensional space. Discriminative approaches such as LDA, NLDA, DLDA, ADA, LPP, MFA, and their various variants can undoubtedly reduce the data dimensionality in large scale with the minimum loss of the discriminative information. Since these approaches in general have similar objective to that of classification, i.e., extracting the most discriminative information on the training samples, problems of misclassification on novel data or poor generalization/inference capability caused by the high dimensionality of the data may not be effectively circumvented. However, some constraints on these discriminative approaches such as the linearity and the limitation to the zero intra-class variation, which are not imposed on the subsequent classification, play some roles in improving the classification accuracy.

The second objective cannot be effectively achieved only based on the consideration of some general phenomena, such as the curse of dimensionality, small sample size problem, noise removal effect of the dimensionality reduction and better generalization in a lower dimensional space. For an effective dimensionality reduction, we have to find out which dimensions are more problematic or harmful than others for a robust classification and hence should be removed. It is shown that the smallest eigenvalues of the class-conditional covariance matrix have the largest deviation from the population variances and hence cause the most severe problem in classification and LDA/ADA evaluation.

Therefore, regularization of these unreliable statistics or removal of the corresponding dimensions by SPCA greatly enhances the classification accuracy. They also help the discriminant evaluation of LDA, ADA, LPP, and MFA to find a portable set of reliable and most discriminative dimensions. However, they may not be effective for a classifier that neither explicitly nor implicitly weights the feature by the inverse of its variance, such as the classical NNC with Euclidian distance and the sparse representation-based classifier SRC.

As regularization does not reduce or fully reduce the data dimensionality and the removal of the unreliable dimensions by SPCA may not lead to a portable feature vector, discriminative approaches such as LDA, ADA, LPP, MFA, and their variants can be followed to greatly reduce the dimensionality for a simple and fast classification. Although various regularization techniques are also applied in many classifiers, they should be applied before the dimensionality reduction because the regularization in the classification stage cannot recover the improperly removed dimensions in the dimensionality reduction stage. With the in-depth understanding of the roles of dimensionality reduction for pattern recognition and the underlying principles revealed in this article, it is not a surprise that most top performers of the state-of-the-art techniques either apply various regularized discriminative analyses or apply two-stage approaches, such as PCA+LDA, PCA+LPP, SPCA+ADA, and PCA+MFA to accomplish the both objectives of the dimensionality reduction.

ACKNOWLEDGMENT

This work was supported in part by Singapore National Science and Engineering Research Council Thematic Strategic Research Programme Grant 062 130 0056.

AUTHOR

Xudong Jiang (exdjiang@ntu.edu.sg) received the B.Sc. and M.Sc. degrees from the University of Electronic Science and Technology of China in 1983 and 1986, respectively, and the Ph.D. degree from Helmut Schmidt University, Hamburg, Germany, in 1997, all in electrical and electronic engineering. From 1998 to 2004, he was a lead scientist and head of the Biometrics Laboratory at the Institute for Infocomm Research, A*Star, Singapore. He has been a faculty member since 2003 and is currently a tenured associate professor and director of the Centre for Information Security at Nanyang Technological University, Singapore. He has published over 90 papers in journals and conferences. His research interests include pattern recognition, computer vision, signal and image processing, and biometrics. He is a Senior Member of the IEEE.

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[2] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.

[3] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[5] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 1, pp. 4–13, Jan. 2005.

[6] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, 2008.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using faces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[8] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[9] X. D. Jiang, "On orientation and anisotropy estimation for online fingerprint authentication," *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 4038–4049, Oct. 2005.

[10] X. D. Jiang, "Extracting image orientation feature by using integration operator," *Pattern Recogni.*, vol. 40, no. 2, pp. 705–717, Feb. 2007.

[11] P. Yap, X. D. Jiang, and A. Kot, "Two dimensional polar harmonic transforms for invariant image representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 7, pp. 1259–1270, July 2010.

[12] G. V. Trunk, "Problem of dimensionality: A simple example," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, no. 3, pp. 306–307, July 1979.

[13] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[14] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 5, pp. 931–937, May 2009.

[15] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.

[16] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

[17] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recogn. Lett.*, vol. 26, no. 2, pp. 181–191, 2005.

[18] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Machine Learn. Res.*, vol. 7, no.11, pp. 2399–2434, 2006.

[19] S. Ji and J. Ye, "Generalized linear discriminant analysis: A unified framework and efficient model selection," *IEEE Trans. Neural Netw.*, vol. 19, no. 10, pp. 1768–1782, Oct. 2008.

[20] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 696–710, July 1997.

[21] B. Moghaddam, "Principal manifolds and probabilistic subspace for visual recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 6, pp. 780–788, June 2002.

[22] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, Nov. 2000.

[23] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[24] C. Liu, "A Bayesian discriminating features method for face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 6, pp. 725–740, 2003.

[25] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 9, pp. 1222–1228, Sept. 2004.

[26] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *Electron. Lett.*, vol. 42, no. 19, pp. 1089–1090, Sept. 2006.

[27] X. D. Jiang, B. Mandal, and A. C. Kot, "Complete discriminant evaluation and feature extraction in kernel space for face recognition," *Mach. Vis. Appl.*, vol. 20, no. 1, pp. 35–46, Jan. 2009.

[28] A. Martinez and A. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 2, pp. 228–233, 2001.

[29] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[30] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711–720, July 1997.

[31] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[32] W. Zheng and X. Tang, "Fast algorithm for updating the discriminant vectors of dual-space LDA," *IEEE Trans. Inform. Forensics Security*, vol. 4, no. 3, pp. 418–427, Sept. 2009.

[33] J. Ye, R. Janardan, C. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 8, pp. 982–994, Aug. 2004.

[34] B. Mandal, X. D. Jiang, H. Eng, and A. Kot, "Prediction of eigenvalues and regularization of eigenfeatures for human face verification," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 717–724, June 2010.

[SP]

Ivana Tošić and Pascal Frossard

Dictionary Learning

What is the right representation for my signal?

Huge amounts of high-dimensional information are captured every second by diverse natural sensors such as the eyes or ears, as well as artificial sensors like cameras or microphones. This information is largely redundant in two main aspects: it often contains multiple correlated versions of the same physical world and each version is usually densely sampled by generic sensors. The relevant information about the underlying processes that cause our observations is generally of much reduced dimensionality compared to such recorded data sets. The extraction of this relevant information by identifying the generating causes within classes of signals is the central topic of this article. We present methods for determining the proper representation of data sets by means of reduced dimensionality subspaces, which are adaptive to both the characteristics of the signals and the processing task at hand. These representations are based on the principle that our observations can be described by a sparse subset of atoms taken from a redundant dictionary, which represents the causes of our observations of the world. We describe methods for learning dictionaries that are appropriate for the representation of given classes of signals and multisensor data. We further show that dimensionality reduction based on dictionary representation can be extended to address specific tasks such as data analysis or classification when the learning includes a class separability criteria in the objective function. The benefits of dictionary learning clearly show that a proper understanding of causes underlying the sensed world is key to task-specific representation of relevant information in high-dimensional data sets.

WHAT IS THE GOAL OF DIMENSIONALITY REDUCTION?

Natural and artificial sensors are the only tools we have for sensing the world and gathering information about physical processes and their causes. These sensors are usually not aware of the physical process underlying the phenomena they "see," hence they often sample the information with a higher rate than the effective dimension of the process. However, to store, transmit or analyze the processes we observe, we do not need such abundant data: we only need the information that is relevant to understand the causes, to reproduce the physical processes, or to make decisions. In other words, we can reduce the

Digital Object Identifier 10.1109/MSP.2010.939537
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

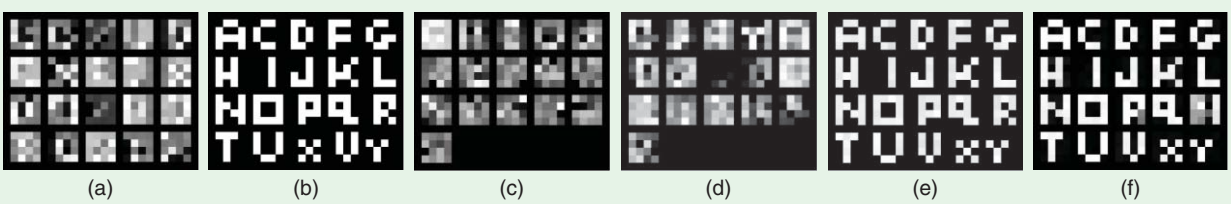
dimension of the sampled data to the effective dimension of the underlying process without sensible penalty in the subsequent data analysis procedure.

An intuitive way to approach this dimensionality reduction problem is first to look at what generates the dimensionality gap between the physical processes and the observations. The most common reason for this gap is the difference between the representation of data defined by the sensor and the representation in the physical space. In some cases, this discrepancy is, for example, a simple linear transform of the representation space, which can be determined by the well-known principal component analysis (PCA) [1] method. It may however happen that the sensors observe simultaneously two or more processes with causes lying within different subspaces. Other methods such as independent component analysis (ICA) [2] are required to understand the different processes behind the observed data. ICA is able to separate the different causes or sources by analyzing the statistical characteristics of the data set and minimizing the mutual information between the observed samples. However, ICA techniques respect some orthogonality conditions such that the maximal number of causes is often limited to the signal dimension. In Figure 1(a), we show some examples of noisy images whose underlying causes are linear combinations of two English letters chosen from a dictionary in Figure 1(b). These images are 4×4 pixels, hence their dimensionality in the pixel space is 16, while the number of causes is 20 (total number of letters). When applied to 5,000 randomly chosen noisy samples of these letters, PCA finds a linear transform of the pixels space into another 16-dimensional space represented by vectors in Figure 1(c). This is done by finding the directions in the original space with the largest variance. However, this representation does not identify the processes that generate the data, i.e., it does not find our 20 letters. ICA [2] differs from PCA because it is able to separate sources not only with respect to the second order correlations in a data set, but also with respect to higher order statistics. However, since the maximal number of causes is equivalent to the signal dimension in the standard ICA, the subspace vectors found by ICA in the example of Figure 1(d) do not explain the underlying letters.

The obvious question is: Why should we constrain our sensors to observe only a limited number of processes? Why do we need to respect orthogonality constraints in the data representation subspace? There is no reason to believe that the number of all observable processes in nature is smaller than the maximal dimension in existing sensors. If we look for an example in a 128×128 dimensional space of face images for all the people in the world, we can imagine that all the images of a single person belong to the same subspace within our 16,384-dimensional space, but we cannot reasonably accept that the total number of people in the world is smaller than our space dimension. We conclude that the representation of data could be overcomplete, i.e., that the number of causes or the number of subspaces used for data description can be greater than the signal dimension.

Where does the dimensionality reduction occur in this case? The answer to this question lies in one of the most important principles in sensory coding—efficiency, as first outlined by Barlow [3]. Although the number of possible processes in the world is huge, the number of causes that our sensors observe at a single moment is much smaller: the observed processes are sparse in the set of all possible causes. In other words, although the number of representation subspaces is large, only few ones will contain data samples from sensor measurements. By identifying these few subspaces, we find the representation in the reduced space.

An important question arises here: given the observed data, how to determine the subspaces where the data lie? The choice of these subspaces is crucial for efficient dimensionality reduction, but it is not trivial. This question has triggered the emergence of a new and promising research field called dictionary learning. It focuses on the development of novel algorithms for building dictionaries of atoms or subspaces that provide efficient representations of classes of signals. Sparsity constraints are keys to most of the algorithms that solve the dictionary learning problems; they enforce the identification of the most important causes of the observed data and favor the accurate representation of the relevant information. Figure 1(e) shows that one of the first dictionary learning methods called sparse coding [4] succeeds in learning all 20 letters that generate 5,000 observations



[FIG1] Learning underlying causes from a set of noisy observations of English letters. A subset of 20 noisy 4×4 images is shown in (a). These samples have been generated as linear combinations of two letters randomly chosen from the alphabet in (b), and they have been corrupted by additive Gaussian noise. When run of 5,000 such samples, PCA and ICA find the same number of components as the dimension of the signal. Therefore, they cannot find the underlying 20 letters. Sparse coding [4] learns an overcomplete dictionary of 20 components, thus it can separate these causes and find all 20 letters from the original alphabet. K-SVD [5] performs similarly, i.e., it finds almost all of the letters. However, since the implementation of K-SVD [5] uses MP for the sparse approximation step, it converges to a local minimum resulting in some repeated letters in the learned dictionary. (a) Noisy samples; (b) original causes; (c) PCA; (d) ICA; (e) sparse coding; and (f) KSVD.

in our simple example. In the course of the last decade, dictionary optimization has led to significant performance improvements in high-dimensional signal processing tasks such as audio, image, multiview, and multimodal data analysis.

This article presents the main challenges in the field of dictionary learning for dimensionality reduction. We first present a brief description of sparse approximations. Next, we give a tutorial overview of the main algorithms that permit the construction of dictionaries for the sparse representation of given classes of signals, possibly with properties such as large incoherence or model-based structures. In the section "Applications of Dictionary Learning," we present a few signal processing applications where the objectives of the learning algorithms is adapted to specific problems such as the joint analysis of correlated signals like audio-visual signals and stereo images. We later show in the section "Learning for Classification" that the construction of dictionaries can also be constrained in order to satisfy discriminative objectives; the dimensionality reduction steps not only lead to good approximation but also efficient classifications of signals.

SPARSE APPROXIMATIONS

The goal of sparse representation is to express a given signal $\mathbf{y}$ of dimension n as a linear combination of a small number of signals taken from a "resource" database, which is called the dictionary. Elements of the dictionary are typically unit norm functions called atoms. Let us denote the dictionary as $\mathcal{D}$ and the atoms as ϕ_k, $k = 1, \ldots, N$, where N is the size of the dictionary. The dictionary is overcomplete $(N > n)$ when it spans the signal space and its atoms are linearly dependent. In that case, every signal can be represented as a linear combination of atoms in the dictionary

$$\mathbf{y} = \mathbf{\Phi a} = \sum_{k=1}^{N} a_k \phi_k. \tag{1}$$

Because the dictionary is overcomplete, $\mathbf{a}$ is not unique. This is where the sparsity constraint comes into play. To achieve efficient and sparse representations, we generally relax the requirement for finding the exact representation. We look for a sparse linear expansion with an approximation error η of bounded energy ϵ. The objective is now to find a sparse vector $\mathbf{a}$ that contains a small number of significant coefficients, while the rest of the coefficients are close or equal to zero. In other words, we want to minimize the resources (atoms) that we use to accomplish the task of signal representation. This optimization problem can be formulated as follows:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_0 \text{ subject to } \mathbf{y} = \mathbf{\Phi a} + \eta \text{ and } \|\eta\|_2^2 < \epsilon, \tag{2}$$

where $\|\cdot\|_p$ denotes the l_p norm. Unfortunately, this problem is NP-hard. However, there exist polynomial time approximation algorithms that find a suboptimal solution for the sparse vector $\mathbf{a}$. These algorithms can be classified in two main groups. The first group includes greedy algorithms such as the matching pursuit (MP) [6] and the orthogonal MP (OMP) [7], which iteratively select locally optimal basis vectors. In the second group, we find algorithms based on convex relaxation methods such as the basis pursuit denoising [8] or least absolute shrinkage and selection operator (LASSO) [9], which solve the following problem:

$$\min_{\mathbf{a}} \left(\|\mathbf{y} - \mathbf{\Phi a}\|_2^2 + \lambda \|\mathbf{a}\|_1\right). \tag{3}$$

The convex relaxation permits to replace the nonconvex l_0 norm in the original problem by the convex l_1 norm. The l_0 norm of a vector is equal to the number of nonzero elements in that vector. It is called a "norm" because it is the limit of p-norms as p approaches zero. However, note that it is not a true norm, unlike the l_1 norm that has all properties of a norm. Besides pursuit algorithms, there exist other sparse approximation algorithms such as the focal underdetermined system solver (FOCUSS) [10] and sparse Bayesian learning [11], for example. A recent review of the sparse recovery algorithms can be found in [12]. The performance of these algorithms in terms of the approximation quality and the sparsity of the coefficient vector $\mathbf{a}$ depends not only on the signal itself, but also on the overcomplete dictionary $\mathcal{D}$. Once the algorithms are used on a specific class of signals $\mathbf{y}$, we easily understand that not all dictionaries provide the same approximation performance. There exist dictionaries that are more likely to lead to sparse solutions than others. These are the dictionaries that include atoms explaining best the causes of the target data set. It is exactly the goal of dictionary learning methods to find such optimized dictionaries.

DICTIONARY LEARNING METHODS

The research in dictionary learning has followed three main directions that correspond to three categories of algorithms: i) the probabilistic learning methods; ii) the learning methods based on clustering or vector quantization; and iii) the methods for learning dictionaries with a particular construction. This construction is typically driven by priors on the structure of the data or to the target usage of the learned dictionary. This section presents the main principles of representative algorithms in each of these three dictionary learning categories.

PROBABILISTIC METHODS

Representation and coding of images have always been a great challenge for researchers because of the high dimensionality and complex statistics of such signals. Thus, it is not surprising that one of the earliest works addressing the problem of learning overcomplete dictionaries appeared exactly for image representation. In 1997, Olshausen and Field [4] developed a maximum likelihood (ML) dictionary learning method for natural images under the sparse approximation assumption. Their method is called sparse coding. The goal of the work was to give evidence that the coding in the primary visual area V1 in the human cortex probably follows a sparse coding model. In other words, their hypothesis was that the visual cortex reduces the high-dimensional representation of each retinal image into a reduced space defined by the receptive fields of a small number of active neurons. Given the linear generative image model in (1), the objective of the ML

learning method is to maximize the likelihood that natural images have efficient, sparse representations in a redundant dictionary given by the matrix $\mathbf{\Phi}$. Formally, the goal of learning is to find the overcomplete dictionary $\mathbf{\Phi}^*$ such that

$$\begin{aligned}\mathbf{\Phi}^* &= \arg\max_{\mathbf{\Phi}}[\log P(\mathbf{y}|\mathbf{\Phi})] \\ &= \arg\max_{\Phi}\left[\log\int_{\mathbf{a}} P(\mathbf{y}|\mathbf{a},\mathbf{\Phi})P(\mathbf{a})\mathbf{da}\right].\end{aligned} \tag{4}$$

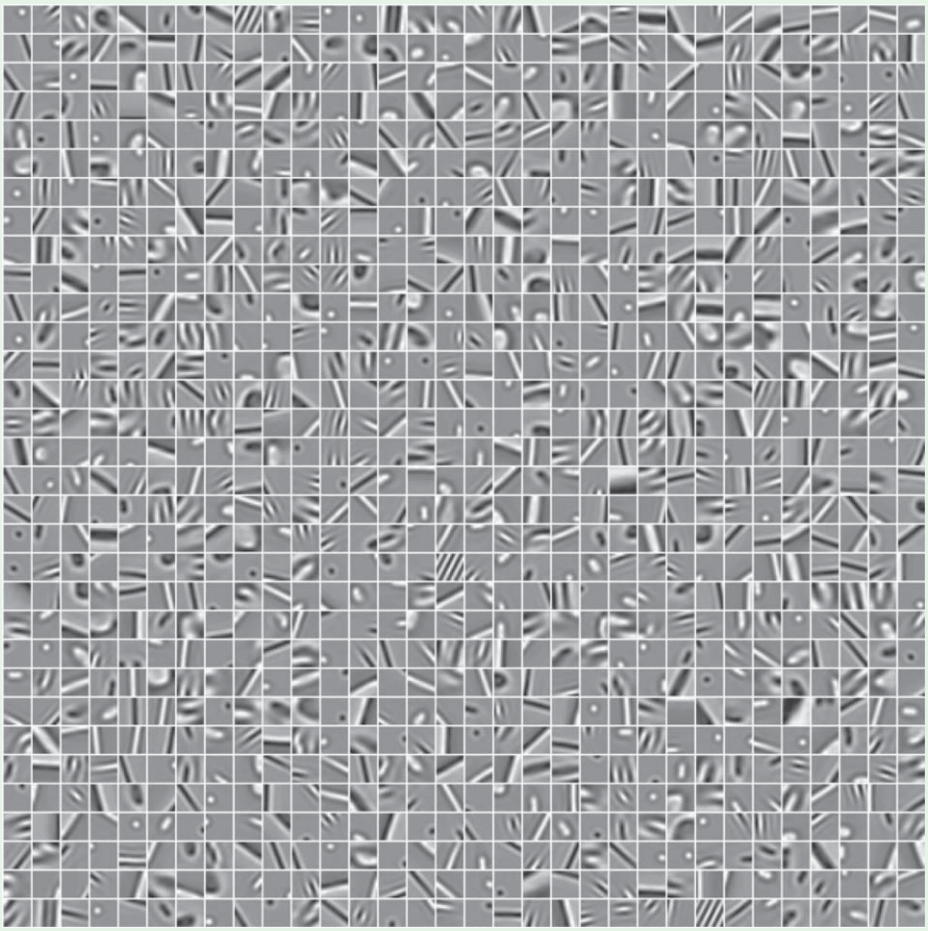
For high-dimensional vectors $\mathbf{a}$, the computation of the integral in (4) is extremely difficult. To simplify the problem and solve the ML optimization, Olshausen and Field introduced two main assumptions. First, the distribution $P(\mathbf{a})$ is assumed to be a product of Laplacian distributions for each coefficient, or equivalently that the coefficients a_i are independent. The Laplacian distribution is peaked at zero and presents a heavy tail, which nicely fits the probability distributions of coefficients a_i when the signal decomposition is sparse. Choosing the prior distribution on $\mathbf{a}$ to be tightly peaked at zero permits to approximate the integral in (4) only by its value at the maximum of $P(\mathbf{y}|\mathbf{a},\mathbf{\Phi})P(\mathbf{a})$. The second assumption is that the approximation noise η can be modeled as a Gaussian zero-mean noise. Under these two assumptions, the optimization problem in (4) can be reduced to an energy minimization problem

$$\begin{aligned}\mathbf{\Phi}^* &= \arg\min_{\Phi,\mathbf{a}} E(\mathbf{y},\mathbf{a}|\mathbf{\Phi}) \\ &= \arg\min_{\mathbf{\Phi},\,\mathbf{a}}[\|\mathbf{y}-\mathbf{\Phi a}\|_2^2 + \lambda\|\mathbf{a}\|_1],\end{aligned} \tag{5}$$

where the energy function is defined as $E(\mathbf{y},\mathbf{a}|\mathbf{\Phi}) = -\log[P(\mathbf{y}|\mathbf{a},\mathbf{\Phi})P(\mathbf{a})]$. To take into account statistics of different images, the dictionary is usually learned by minimizing an average energy $\langle E(\mathbf{y}_i,\mathbf{a}_i|\mathbf{\Phi})\rangle$ over a set of randomly chosen images $\{\mathbf{y}_i\}$. The casted optimization problem can be solved by iterating between two steps. In the first step, $\mathbf{\Phi}$ is kept constant and the energy function is minimized with respect to a set of coefficient vectors $\{\mathbf{a}_i\}$. This inference step is essentially the sparse approximation problem defined by (3). It can be solved by convex optimization for each $\mathbf{y}_i$. The second step is called the learning step. It keeps the coefficients $\{\mathbf{a}_i\}$ constant, while performing the gradient descent on $\mathbf{\Phi}$ to minimize the average energy. Since the first step is computationally expensive, the probabilistic dictionary learning methods usually work with small image patches, i.e., the size of $\mathbf{y}_i$ is typically below 32×32 pixels. The algorithm iterates between the sparse approximation and the dictionary learning steps until convergence. This alternating optimization process does not necessarily find the global optimum solution of the considered problem. However, it has been shown to converge to a dictionary with atoms that resemble the receptive fields of simple neurons in V1. A ten times overcomplete dictionary learned on 16×16 image patches [13] is illustrated in Figure 2. The biggest part of the learned dictionary consists of atoms that are localized, oriented and bandpass. Interestingly, these types of features represent well the oriented edges in images. Moreover, the dictionary contains atoms that are center-surround and gratings, which better approximate textures in images. Dictionary learning here clearly meets our objectives: it identifies the most important building blocks in natural images, which permit to approximate the signals by a sparse series of causes or components. It also permits to build an interesting bridge between sparse image representation methods and the properties of the human visual cortex, which is undoubtedly a very efficient encoder for natural images.

The probabilistic inference approach in overcomplete dictionary learning has subsequently been adopted by other researchers. The two-step optimization structure has been preserved in most of these works, and the modifications usually appeared in either the sparse approximation step, or the dictionary update step, or in both. For example, the method of optimal directions (MOD) algorithm [14] optimizes iteratively the same objective ML function as in sparse coding. However, it uses the OMP algorithm to find a sparse vector $\mathbf{a}$ and introduces a closed-form solution for the dictionary update step. The two modifications render the MOD approach faster compared to the method of Olshausen and Field, but still does not guarantee to find the globally optimal solution. Moreover, it is not guaranteed to converge, neither to decrease the objective function at each iteration. The maximum a posteriori (MAP) dictionary learning method [15] belongs also to the family of two-step iterative algorithms based on probabilistic inference. Instead of maximizing the likelihood $P(\mathbf{y}|\mathbf{\Phi})$, the MAP method maximizes the posterior probability $P(\mathbf{\Phi}, \mathbf{a}|\mathbf{y})$. This essentially reduces to the same two-step algorithm, where dictionary update includes an additional constraint on the dictionary that can be for example the unit Frobenius norm of $\mathbf{\Phi}$ or the unit l_2 norm of all atoms in the dictionary. The sparse approximation step is here performed with FOCUSS [10]. Finally, the majorization method can also be used to minimize the objective function in both sparse approximation and dictionary update steps [16]. The sparse approximation step then reduces to the use of an iterative thresholding algorithm.

Naturally, the two assumptions introduced in the sparse coding method represent constraints that can be modified or even removed to learn better dictionaries or to extend the method to other signal models. Lewicki and Sejnowski have modified the first assumption and proposed a new way to approximate the integral in (4) with a Gaussian around the posterior estimate of the coefficient vector $\mathbf{a}$. This changes the update rule in the learning step [17]. They have shown that the ML dictionary learning method with the new estimate for $P(\mathbf{y}|\mathbf{\Phi})$ learns dictionaries that improve the efficiency of sparse coding. The efficiency is measured here in terms of the entropy of data given the overcomplete dictionary. This method actually represents a generalization of the independent component analysis (ICA) method to overcomplete dictionaries. On the other hand, one can also modify the second assumption on the existence of Gaussian noise. When the noise term is zero (i.e., $\eta = 0$), the sparse representation step



[FIG2] **Overcomplete dictionary learned with sparse coding from a large data set of 16 × 16 natural image patches. [Used with permission from SPIE (B. A. Olshausen, C. F. Cadieu, and D. K. Warland, "Learning real and complex overcomplete representations from the statistics of natural images," *Proc. SPIE,* vol. 7446, 2009).]**

is performed using the exact ℓ_1 sparse optimization [18]. In general, convergence is not guaranteed for the ℓ_1-constrained methods, although it can be proved in some conditions [19]. One could also introduce smoother sparsity priors to obtain more stable solutions. For example, the ℓ_1 constraint is replaced by a Kullback-Leibler (KL) divergence in [20], which shows that the sparsity is preserved, while the KL-regularization leads to efficient convex inference and stable coefficient vectors (i.e., stable representations).

Finally, fast online learning algorithms have been proposed recently [19]. As most of the learning methods based on alternate solutions of the sparse coding and dictionary updates steps use the whole training set at each iteration, these algorithms become rapidly expensive when the data set is large and mostly inappropriate for dynamic systems where data evolve over time. Online learning overcomes this limitation by increasing progressively the training set. An alternate optimization of sparse coding and dictionary update steps is performed with a subset of the training data. This subset is then augmented with a new training sample. The alternate optimization is run again on the new training data with the outcome of the previous iteration as initialization. The online algorithm repeats these iterations until all training data have been used. The resulting solution converges with efficient learning performance and drastically lower computational complexity.

CLUSTERING-BASED METHODS

A slightly different family of dictionary learning techniques is based on vector quantization (VQ) achieved by K-means clustering. The VQ approach for dictionary learning has been first proposed by Schmid-Saugeon and Zakhor in MP-based video coding [21]. Their algorithm optimizes a dictionary given a set of image patches by first grouping patterns such that their distance to a given atom is minimal, and then by updating the atom such that the overall distance in the group of patterns is minimal. The implicit assumption here is that each patch can be represented by a single atom with a coefficient equal to one, which reduces the learning procedure to a K-means clustering. Since each patch is represented by only one atom, the sparse approximation step becomes trivial.

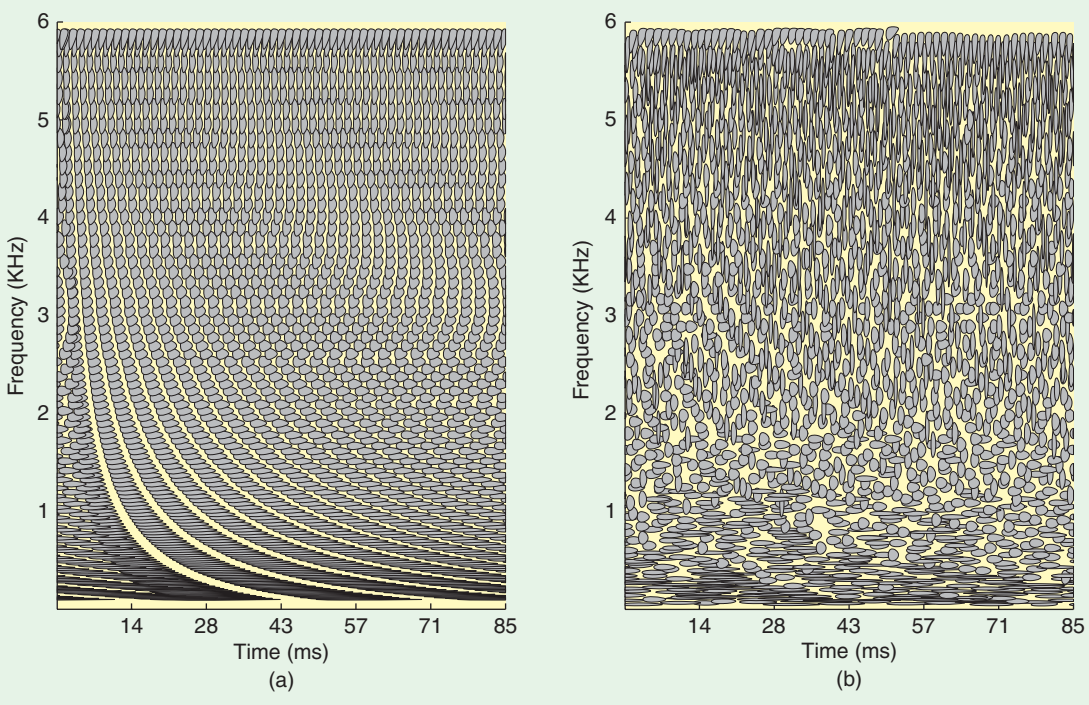
A generalization of the K-means algorithm for dictionary learning, called the K-SVD algorithm, has been proposed by Aharon et al. [5]. After the sparse approximation step with OMP, the dictionary update is performed by sequentially updating each column of $\mathbf{\Phi}$ using a singular value decomposition (SVD) to minimize the approximation error. The update step is hence a generalized K-means algorithm since each patch can be represented by multiple atoms and with different weights. This algorithm is not guaranteed to converge in general. However, in practice, dictionaries learned with K-SVD have shown excellent performance in image denoising. Figure 1(f) shows how K-SVD finds almost all 20 letters as the underlying causes of noisy letter samples. In this example, the sparse approximation step has been implemented by OMP, so it converges to a local minimum where letters "R" and "P" are not successfully separated.

LEARNING DICTIONARIES WITH SPECIFIC STRUCTURES

Many applications do not necessitate general forms of dictionary atoms but can rather benefit from a dictionary that is a set of parametric functions. In contrary to the generic dictionaries above, the advantages of parametric dictionaries reside in the short description of the atoms. The generating function and the atom parameters are sufficient for building the dictionary functions. This is quite beneficial in terms of memory requirements, communication costs or implementation complexity in practical applications.

Such generating functions can be built on prior knowledge about the form of signal causes or the target task. For example, some perceptual criteria can drive the choice of the generating functions in building the dictionary atoms, when the objective is to reconstruct data that are eventually perceived by the human auditory or visual system. Learning in such parametric dictionaries reduces to the problem of learning the parameters for one or more generating functions. Equivalently, it consists in finding a good discrete parametrization that leads to efficient sparse signal approximations. Parametric dictionaries are usually structured, so one can enforce some desired dictionary properties during learning such as minimal dictionary coherence; for example, one can optimize a parametric dictionary such that it gets close to an equiangular tight frame (ETF). In [22], a dictionary for audio signals is learned based on a Gammatone generating function, which has been shown to have similarities with the human auditory system. The method learns a dictionary with good coherence properties, which tiles the time-frequency plane more uniformly than the original Gammatone filter bank. The resulting dictionaries are shown in Figure 3.

Priors or models of the underlying signal causes can also lead to imposing properties such as shift-invariance [23] or multiscale [24] characteristics of the atoms. Such constraints typically limit the search space in the dictionary optimization problem, but lead to more accurate or task-friendly representations. Similarly, the target dictionary might present specific characteristics in particular recovery problems, such as a block-based structure [25], or orthogonality between subspaces [26]. These requirements



[FIG3] Time-frequency representations of structured dictionaries for audio signal representation. It can be observed that the learned dictionary (b) provides a more uniform tiling of the time-frequency plane than the original dictionary (a) designed from a Gammatone filter bank. This corresponds to a smaller coherence than in the original dictionary. Figure used with permission from [22].

considerably affect the design of learning strategies as well as the approximation performance.

APPLICATIONS OF DICTIONARY LEARNING

Dictionary learning for sparse signal approximation has found successful applications in several domains. For example, it has been applied to medical imaging and representation of audio and visual data. We overview here some of the main applications in these directions.

MEDICAL IMAGING

Dictionary learning has the interesting potential to reveal a priori unknown statistics of certain types of signals captured by different measurement devices. An important example are medical signals, such as electroencephalogram (EEG), electrocardiography (ECG), magnetic resonance imaging (MRI), functional MRI (fMRI), and ultrasound tomography (UST) where different physical causes produce the observed signals. It is crucial, however, that representation, denoising, and analysis of these signals are performed in the right signal subspace, such that the underlying physical causes of the observed signals can be identified. Learning of components in ECG signals facilitates ventricular cancellation and atrial modeling in the ECG of patients suffering from atrial fibrillation [27]. Overcomplete dictionaries learned from MRI scans of breast tissues have been shown to provide an excellent representation space for reconstructing images of breast tissue obtained by the UST scanner [28], which drastically reduces the imaging cost compared to MRI. Moreover, standard breast screening techniques, such as the X-ray projection mammography and computed tomography can potentially exploit highly sparse representations in learned dictionaries [29]. Analysis of other signals, such as neural signals obtained by EEG, multielectrode arrays, or two-photon microscopy could also largely benefit from adapted representations obtained by dictionary learning methods.

REPRESENTATION OF AUDIO AND VISUAL DATA

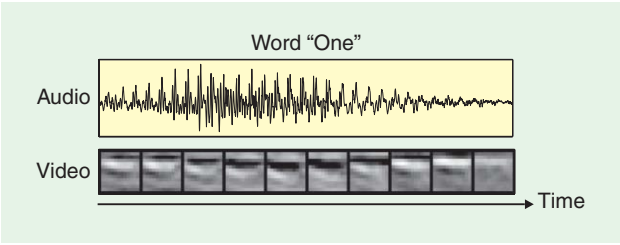
Dictionary learning has introduced significant progress in denoising of speech [30] and images [5], and in audio coding and source separation [16], [31], where it is very important to capture the underlying causes or the most important constitutive components of the target signals. The probabilistic dictionary learning framework has been also proposed for modeling natural videos. These methods explicitly model the separation of the invariant signal part given by the image content and the varying part represented by the motion. Learning under these separation constraints can be achieved using the bilinear model [32], [33], or the phase coding model [34]. In addition to learning the dictionary elements for the visual content, these methods also learn the sparse components of the invariant part (e.g., translational motion).

There exist many examples in nature where a physical process is observed or measured under different conditions. This results in sets of correlated signals whose common part corresponds to the underlying physical cause. However, different observation conditions introduce variability in the measured signals, such that the common cause is usually difficult to extract. Dictionary learning methods based on ML and MAP can be extended by modifying the objective function such that the learning procedures identify the proper subspace for the joint analysis of multiple signals. This permits to learn the underlying causes under different observation conditions. Such modified learning procedures have been applied to audio-visual signals [35] and to multiview imaging [36]. The synchrony between audio and visual signals is exploited in [35] to extract and learn the components of their generating cause that is human speech. A multimodal dictionary is learned with elements that have an audio part and a video part corresponding to the movement of the lips that generate the audio signal. An example of the learned atom for the word "one" is shown in Figure 4. One important contribution of this work certainly lies in its benefits towards understanding and modeling the integration of audio and visual sensory information in the cortex.

In stereo vision, the same three-dimensional (3-D) scene is observed from different viewpoints, which produce correlated multiview images. Due to the projective properties of light rays, the correlation between multiview images has to comply with epipolar geometry constraints. Dictionaries can be learned such that they efficiently describe the content of natural stereo images and simultaneously permit to capture the geometric correlation between multiview images [36]. The correlation between images is modeled by the local atom transforms, which is made feasible by the use of geometric dictionaries built on scaling, rotation and shifts of a generating function. Learning is based on an ML objective that includes the probability that left image $\mathbf{y}_L$ and right image $\mathbf{y}_R$ are well represented by a dictionary Φ, and the probability that corresponding image components in different views satisfy the epipolar constraint

$$\mathbf{\Phi}^* = \arg\max_{\Phi}[\log P(\mathbf{y}_L, \mathbf{y}_R, D = 0|\mathbf{\Phi})], \tag{6}$$

where $D = 0$ denotes the event when the epipolar geometry is satisfied. This ML objective leads to an energy minimization learning method, where the energy function has three terms: image approximation error term (for both stereo images), the sparsity term, and the multiview geometry term. Dictionary learning is performed in two steps: sparse approximation step



[FIG4] Learned audio-visual atom representing the word "one." Figure used with permission from [35].

with the multiview MP algorithm [36], and dictionary update step with the conjugate gradient method. An illustrative example of a sparse decomposition of two stereo image patches with three correlated learned stereo atoms is shown in Figure 5. Learned stereo dictionaries can be applied to the joint or distributed coding of multiple correlated views or to the analysis and understanding of the geometry in 3-D scenes [36].

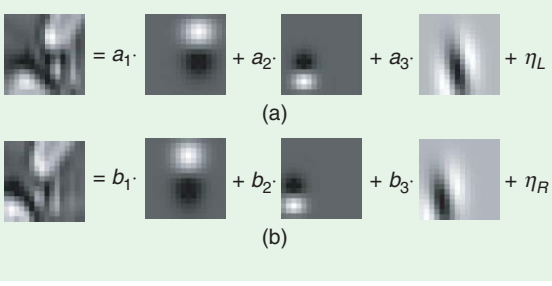
The above illustrations demonstrate the benefits of sparse approximations with learned dictionaries in very diverse applications. One of the main advantages of dictionary learning is that it allows for representing the underlying causes of signals or the main components of data. This is very important for proper understanding and analysis of data that are often the result of noisy measurements of physical processes.

LEARNING FOR CLASSIFICATION

DIMENSIONALITY REDUCTION AND CLASSIFICATION

Dimensionality reduction has been described so far from a pure approximation perspective, where a subspace or a dictionary is computed to explain the observed data with a sparse representation. Alternatively, dimensionality reduction can also target the analysis of data with the objective of distinguishing between different classes of signals or physical processes and to beat the curse of dimensionality and scale. Low-dimensional problems generally involve less complex and more efficient algorithms. The reduced subspace emphasizes in this case the most relevant information in the signal and permits to distinguish between different classes of observations.

Dimensionality reduction for signal analysis finds numerous applications in diverse domains such as sensor networks, computer vision, data mining, machine learning, or information retrieval. We can distinguish two main types of algorithms for computing the reduced subspace: the discriminative methods and the reconstructive methods that are illustrated in Figure 6. The main objective of the discriminative method is to find a mapping or an embedding between the original data space and a reduced dimension subspace, where data can then be efficiently analyzed or classified. This mapping can be either linear (e.g., linear discriminant analysis (LDA) [37]) or nonlinear (e.g., locally linear embedding (LLE) [38], Isomap [39]). The objective of the mapping is to clearly separate the data from different classes in the low-dimensional subspaces. The discriminative methods however aim at pure discrimination objectives and do not necessarily rely on the computation of meaningful features or specific components of the signal. These methods become unfortunately quite vulnerable to noise in the data, to missing data, or to imperfect testing conditions.



[FIG5] Sparse decomposition of a stereo image pair with three correlated learned stereo atoms. (a) Left image and its atoms. (b) Right image and its atoms. Stereo atoms in the two views (three right-most columns) are correlated by local geometric transforms that obey epipolar geometry constraints.

The reconstructive methods try to compute representations that enable analysis and labeling of the data and simultaneously capture its constitutive components to provide robustness to impairments. We focus here on representations that use linear subspaces as opposed to more generic manifold methods. The most common low-rank approximation methods used in signal analysis are based on ICA, PCA, or part-based representations such as nonnegative matrix factorization (NMF) algorithms [40]. The role of a dimensionality reduction algorithm consists here in simplifying the signal to its most meaningful components, such that it can be efficiently characterized in the reduced subspace. For example, PCA maximizes the variance of the data projected on the reduced subspace, which eventually reinforces the discrimination capabilities of the subspace representation. In most reconstructive methods, the projected data are eventually labeled based on nearest neighbor or nearest subspace criteria. However, the basis vectors that define the reduced dimension subspace might unfortunately be holistic, of global support, or with long description length. In the next sections, we describe methods that build linear subspaces from redundant dictionaries of functions with fine adaptation to the data under consideration toward effective signal classification.

SUBSPACE SELECTION FOR CLASSIFICATION

Dimensionality reduction can first be achieved by selecting a subset of functions from a large, fixed dictionary that is used for the analysis of particular signals. These functions then determine a subspace of reduced dimension, where classification can be performed by computing the nearest neighbor points among the projected data. A simple method to build such a subspace consists in modifying the sparse approximation methods described in the previous sections, such that the objective function is augmented with a discrimination term that represents the separability properties of the projection subspace. One can thus select a subset of functions in a dictionary (represented by the matrix $\mathbf{\Phi}$), which approximate the data samples and simultaneously encourage the separability of data in different classes. In other words, the reduced subspace can be computed by solving a problem like

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} [\|\mathbf{y} - \mathbf{\Phi}\mathbf{a}\|_2^2 + \gamma J(\mathbf{\Phi}, \mathbf{a})], \tag{7}$$

where the term $J(\mathbf{\Phi}, \mathbf{a})$ measures the separability of the different classes when data is represented by atoms in $\mathbf{\Phi}$ and coefficients $\mathbf{a}$. It typically tries to maximize the variance

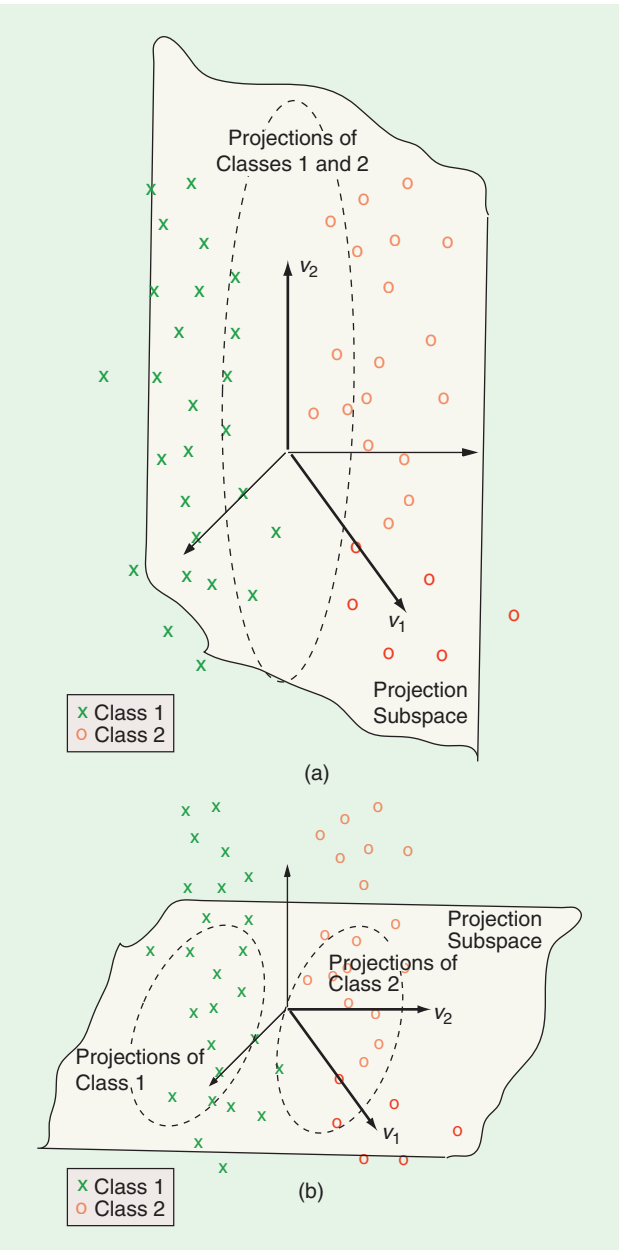
between the active atoms from $\mathbf{\Phi}$ that represent signals in different classes. The reduced subspace used for classification is finally formed by the subset of atoms in $\mathbf{\Phi}$ whose corresponding coefficients in $\mathbf{a}^*$ are nonzero. The subset selection problem can be interpreted as the inference step in the dictionary learning methods when the objective function is modified to include a discriminative term. Finally, the weight parameter γ controls the tradeoff between approximation and classification performance in the reduced dimension subspace. A subset of $\mathbf{\Phi}$ that solves the problem posed in (7) can be determined by iterative supervised atom selection built on OMP for example [41]. The idea mainly consists in selecting greedily the atoms from the dictionary that lead to the best tradeoff between approximation of the training data and discrimination between classes.

The minimum of the joint objective function above can be achieved with several distinct sets of functions: finding the best subspace for classification becomes nontrivial due to the redundancy of the dictionary. However, good subspaces for reconstructive dimensionality reduction are characterized by sparsity properties, where only a few significant components participate in the representation of the data. The method of sparse representation for signal classification in [42] thus explicitly includes sparsity constraints in the dimensionality reduction process. The reduced subspace is determined here by a simultaneous sparse approximation algorithm built on OMP, where the data separability term $J(\mathbf{\Phi},\mathbf{a})$ is given by a Fisher's discrimination criterion used in LDA. The reduced dimensionality subspace is therefore chosen as a compromise between approximation of data within classes, discrimination of data in different classes, and sparsity of the data representation as determined by an optimization problem of the following generic form:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} [\|\mathbf{y} - \mathbf{\Phi a}\|_2^2 + \gamma_1 \|\mathbf{a}\|_0 + \gamma_2 J(\mathbf{\Phi}, \mathbf{a})]. \quad (8)$$

SUPERVISED DICTIONARY LEARNING

An important advantage of redundant dictionaries for classification is that signal analysis can be performed with functions that are likely to match the data characteristics in different classes of signals. Similarly to data approximation problems, data analysis applications can further benefit from dictionary learning methods. The previous section describes subspace selection methods from predefined dictionaries. However, learning can improve the classification performance, as it leads to a better adaptation of the dictionary by enforcing sparsity in the representation of data in the different classes. The atoms in a dictionary $\mathcal{D}$ that is computed with dictionary learning methods generally capture the most important constitutive components of the signals. They naturally permit to classify the data into the corresponding linear subspace as shown in [5], for example. However, there is no guarantee that the subspace built on a learned dictionary $\mathbf{\Phi}$ is truly optimal for classification, as it targets efficient representation but not necessarily class separability. For example, one may define a set of functions that are good to (sparsely) approximate signals in a face image data set. However, there is no good reason why this same set of functions is also the best one for distinguishing different persons in this data set.



[FIG6] Illustration of dimensionality reduction of a two-class data set, by projection on a linear subspace defined by vectors (v_1,v_2). (a) Purely reconstructive methods compute a representative subspace where the projections of the data are close to the original data points. The approximation of data by their projections is optimized, but the classification of the projected data is not trivial. (b) Purely discriminative methods compute the reduced dimensionality subspace so that the classification can be done efficiently from the data projections. Data approximation is quite poor in this case, which results in low robustness to data impairments. The optimal subspace has to be the result of a tradeoff between approximation and separability.

Dictionary learning methods should rather be modified so that they become simultaneously reconstructive (for robustness to noise) and discriminative (for efficient classification with the learned dictionary). The addition of a discriminative term into the dictionary learning algorithms requires supervision, where labels of training data are used to ensure that the data representation is sufficiently different in each class. It can be achieved by modifying the sparse coding step in the learning algorithms, so that it optimizes an objective function that favors the sparsest representation of a given signal and simultaneously the representation that is also the most different from the one of signals in other data classes. The supervised dictionary learning problem can be cast as a mixed formulation that minimizes the average value of the sparse approximation errors over different classes and also enforces discrimination between classes. For example, the dictionary optimization problem can be written as

$$\mathbf{\Phi}^* = \arg\min_{\Phi,\mathrm{a}} [\|\mathrm{y} - \mathbf{\Phi}\mathrm{a}\|_2^2 + \gamma_1\|\mathrm{a}\|_1 + \gamma_2 \mathrm{C}(\mathrm{a}, \mathbf{\Phi}, \theta)], \quad (9)$$

where the function $C(\mathrm{a}, \mathbf{\Phi}, \theta)$ is a discrimination term that depends on the dictionary, the coefficient vectors, and the parameters θ of the model used for classification. Since the dictionary is learned, alternate inference and learning steps have to be used in solving (9). In contrary, the subspace selection problem in the section "Classification Subspace Selection" is solved only within the inference step. Note that the discrimination term is specific to the chosen classifier through the parameters θ so that the learning problem becomes highly dependent on the classification method and unfortunately nonconvex. Still, it can be solved efficiently by fixed-point continuation methods [43] when the classifier is based on logistic regression methods.

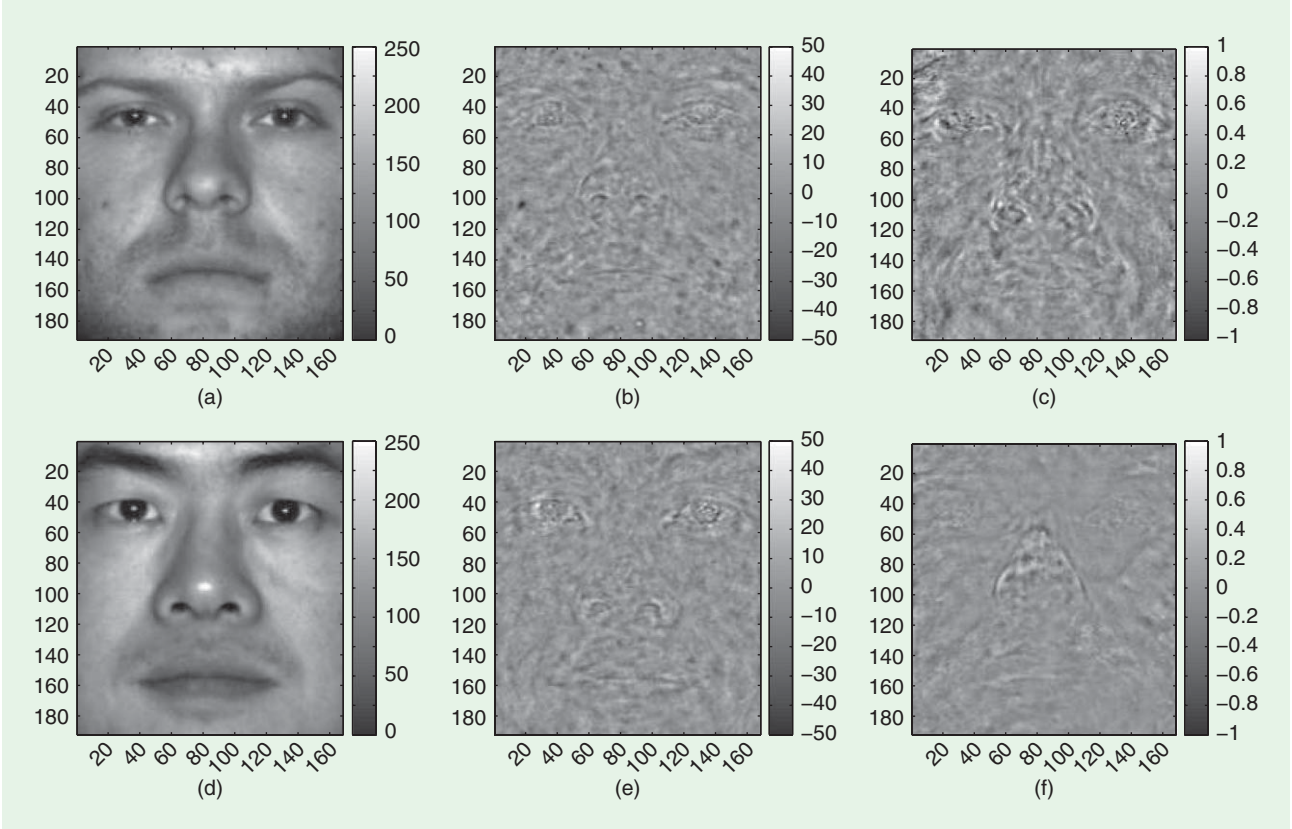
The use of one learned dictionary for all the data classes leads to a straightforward classification stage where the dictionary vectors and the coefficients in the signal representation are used directly to make classification decisions. Alternatively, one may want to improve the discrimination by building a distinctive projection subspace for each data class. Classification is then performed by selecting the subspace that is the nearest to the test signal, or equivalently the subspace that leads to the best representation of the test signal. A simple way to build adaptive dictionaries for each class is to use the signals in the training set for the class dictionary. Sparsity constraints are then rather applied within the classification process, where the sparsest representation of the test signal determines its class label. For example, Wright et al. [44] have proposed a face recognition method that uses training face images as dictionaries and an l_1 sparse optimization method in the classification stage. The authors show that the recognition task can be successfully accomplished even using random features at first. Furthermore, the algorithm is robust to a certain amount of noise due to the sparsity constraints.

It is often preferable, however, to construct adapted dictionaries that can lead to an efficient classification process based on simple subspace projections. The construction of class dictionaries can be performed with learning methods where the sparse coding step in the iterative learning algorithms is modified, so that sparse coding is computed independently within each class. Such a sparse coding stage can be implemented by class-supervised versions of simultaneous pursuit algorithms, for example, where a joint sparse representation of the training data is selected independently in each class. The subsequent dictionary update step further favors the reconstruction of signals with the functions selected in the modified sparse coding step. If the update step is based on an SVD algorithm, it simply leads to a supervised version of the K-SVD algorithm [45], where the K-SVD learning algorithm becomes adapted to classification tasks. As supervised dictionary learning should intuitively lead to subspaces that are good for approximating data in their own class but bad for representing data from any other class, the subspaces can also be computed with a hierarchical process that ensures that features selected in different dictionaries have only a minimal correlation [46]. Alternatively, global softmax discriminative functions can enforce that the learned dictionaries are better for representing data of their classes than data from any other classes. Such a discrimination can be achieved by modifying the dictionary update steps in the learning process with a modified version of MOD/K-SVD algorithm whose role is thus extended to ensuring data separability with the updated dictionary in addition to good approximation properties [47].

Finally, discrimination in dictionary learning can also be achieved by enforcing incoherency between the subspaces that represent data in different classes and not only by minimizing the correlation between the features in different subspaces. It relies on the intuition that some features might be relatively good in representing data in different classes, but several features taken together form a subspace that is mostly good in approximating data from the corresponding class. For example, the subspace formed by noses and eyes of persons in different classes are incoherent, even if these persons have similar eyes or the same nose. With the assumption that the residue of the subspace projection is minimal in the correct class, incoherent subspaces can be designed by an alternate projection method [48]. It builds on the natural conditions that the interplay between features of different classes should be small, while the interaction of training data with features in the correct class should be clearly higher than the interaction with features representing any other class (see Figure 7). With minimal assumptions on the signal models or sparsity features, such a dictionary learning method reaches state-of-the-art performance on a face classification experiment.

CONCLUSIONS

The goal of dimensionality reduction is to find efficient, low-dimensional data representations within the large dimensional space where the observed data lies. This article has presented some of the recent results supporting the idea that these representations are sparse within an overcomplete dictionary of atoms or subspaces. In this context, the methods for dictionary



[FIG7] Images of two subjects: Parts (a) and (d) show original projections onto the span of features from their own class. Parts (b) and (e) show projections onto the span of features of the (c) and (f) wrong class. The representation of signal with the subspace of the proper class is clearly more relevant than the representation with a subspace of another class: the scales and positions of projection components are close to the original signal. Figure used with permission from [48].

learning have much to offer since they are able to adapt the data representation to the underlying causes of the observations. We have given a broad overview of the main dictionary learning algorithms and shown their usage in various applications, such as audio-visual coding and stereo image approximation. We have also discussed the discriminative power of sparse representations and outlined the large potential benefits of dictionary learning in classification and face recognition applications.

Many challenges are still open in dictionary learning. Understanding the underlying causes of signals or the relevant information in observations becomes more challenging when the training samples are imperfect. In many applications, the training samples are noisy, distorted by the sensing process, or simply incomplete like in the case of occlusions in multiview imaging. The last example particularly makes us question the validity of linear representation models in vision where we usually encounter nonlinearities such as occlusions. Linear models also become invalid in advanced applications like medical imaging where the acquisition methods are typically nonlinear. In all these situations, dictionary learning still faces critical research questions. Similarly, signal analysis may require more complex models than linear subspaces for efficient classification. One can build dictionaries to be used in the definition of manifold models or graph-based representations that could potentially handle transformation-invariant classifications problems. In general, dictionaries offer a very flexible and powerful way to represent relevant information in high-dimensional signals. However, the proper modeling of the complex underlying causes of observations poses many exciting questions about the proper construction of these dictionaries.

ACKNOWLEDGMENTS

This work has been partly supported by the Swiss National Science Foundation under grant PBELP2-127847. The authors would like to thank the editors and anonymous reviewers for their insightful comments that have greatly helped to improve the quality of the work. Special thanks go also to Bruno Olshausen, Pierre Vandergheynst, Sofia Karygianni, Effrosyni Kokiopoulou, and Elif Vural for the careful reading of the manuscript and helpful feedbacks. We would also like to thank Gianluca Monaci, Fritz Sommer, Laurent Daudet, and Karin Schnass for providing some of the figures used in this article.

AUTHORS

Ivana Tošić (ivana@berkeley.edu) received the Dipl.Ing. degree in telecommunications from the University of Niš, Serbia, and the Ph.D. degree in computer and communication sciences from the Swiss Federal Institute of Technology (EPFL), Lausanne,

Switzerland. She is currently a postdoctoral researcher at the Redwood Center for Theoretical Neuroscience, University of California at Berkeley, United States, where she works on the intersection of image processing and computational neuroscience domains. She was awarded the Swiss National Science Foundation fellowship for prospective researchers. Her research interests include representation and coding of the plenoptic function, distributed source coding, binocular vision, and 3-D object representation. She is a Member of the IEEE.

Pascal Frossard (pascal.frossard@epfl.ch) received the M.S. and Ph.D. degrees, both in electrical engineering, from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1997 and 2000, respectively. Between 2001 and 2003, he was a member of the research staff at the IBM T.J. Watson Research Center, Yorktown Heights, New York. Since 2003, he has been an assistant professor at EPFL, where he heads the Signal Processing Laboratory (LTS4). His research interests include image representation and coding, visual information analysis, distributed image processing and communications, and media streaming systems. He is a Senior Member of the IEEE.

REFERENCES

[1] I. Jolliffe, *Principal Component Analysis.* Springer-Verlag: New York, 1986.

[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[3] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," *in Sensory Communication*, W. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, ch. 13, pp. 217–234.

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

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

[6] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[7] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[8] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B (Method.)*, vol. 58, no. 1, pp. 267–288, 1996.

[10] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[11] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[12] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE* (Special Issue on Applications of Sparse Representation and Compressive Sensing), to be published.

[13] B. A. Olshausen, C. F. Cadieu, and D. K. Warland, "Learning real and complex overcomplete representations from the statistics of natural images," *Proc. SPIE*, vol. 7446, 2009.

[14] K. Engan, S. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'99)*, 1999.

[15] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, 2003.

[16] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2178–2191, 2009.

[17] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.

[18] M. D. Plumbley, "Dictionary learning for L1-exact sparse coding," *Lect. Notes Comput. Sci.*, vol. 4666, pp. 406–413, 2007.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, to be published.

[20] D. Bradley and J. Bagnell, "Differentiable sparse coding," *Proc. NIPS*, vol. 11, pp. 19–60, Jan. 2009.

[21] P. Schmid-Saugeon and A. Zakhor, "Dictionary design for matching pursuit and application to motion-compensated video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 880–886, 2004.

[22] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4800–4810, 2009.

[23] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Shift-invariant dictionary learning for sparse representations: Extending K-SVD," in *Proc. European Signal Processing Conf.*, vol. 4, 2008.

[24] P. Sallee and B. A. Olshausen, "Learning sparse multiscale image representations," in *Proc. Conf. Neural Information Processing Systems*, 2003.

[25] K. Engan, K. Skretting, and J. Husøy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Dig. Signal Process.*, vol. 17, no. 1, pp. 32–49, 2007.

[26] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[27] B. Mailhé, R. Gribonval, F. Bimbot, M. Lemay, P. Vandergheynst, and J.-M. Vesin, "Dictionary learning for the sparse modelling of atrial fibrillation in ECG signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'09)*, 2009.

[28] I. Tošić, I. Jovanović, P. Frossard, M. Vetterli, and N. Durić, "Ultrasound tomography with learned dictionaries," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'10)*, 2010.

[29] C. K. Abbey, J. N. Sohl-Dickstein, B. A. Olshausen, M. P. Eckstein, and J. M. Boone, "Higher-order scene statistics of breast images," in *Proc. Soc. Photo-Optical Instrumentation Engineers Conf. Series (SPIE'09)*, vol. 7263, 2009.

[30] M. G. Jafari and M. D. Plumbley, "Speech denoising based on a greedy adaptive dictionary algorithm," in *Proc. European Signal Processing Conf.*, 2009.

[31] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, no. 99, pp. 1–11, 2009.

[32] D. B. Grimes and R. P. N. Rao, "Bilinear sparse coding for invariant vision," *Neural Comput.*, vol. 17, no. 1, pp. 47–73, 2005.

[33] B. A. Olshausen, C. Cadieu, B. J. Culpepper, and D. K. Warland, "Bilinear models of natural images," in *Proc. SPIE Conf. Human Vision and Electronic Imaging*, 2007.

[34] C. Cadieu and B. A. Olshausen, "Learning transformational invariants from time-varying natural images," in *Proc. Conf. Neural Information Processing Systems*, 2008.

[35] G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audio-visual data," *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1898–1910, 2009.

[36] I. Tošić and P. Frossard, "Dictionary learning for stereo image representation," *IEEE Trans. Image Process.*, to be published.

[37] A. Webb, *Statistical Pattern Recognition*, 2nd ed. Hoboken, NJ: Wiley, 2002.

[38] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[39] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[40] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 11–126, 1994.

[41] E. Kokiopoulou and P. Frossard, "Semantic coding by supervised dimensionality reduction," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 806–818, 2008.

[42] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Conf. Neural Information Processing Systems*, 2007.

[43] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Supervised dictionary learning," in *Proc. Conf. Neural Information Processing Systems*, 2008.

[44] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[45] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," *IMA Preprint*, 2007.

[46] A. Destrero, C. De Mol, F. Odone, and A. Verri, "A sparsity-enforcing method for learning face features," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 188–201, 2009.

[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[48] K. Schnass and P. Vandergheynst, "A union of incoherent spaces model for classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'10)*, 2010.

[SP]

Lawrence Carin, Richard G. Baraniuk, Volkan Cevher, David Dunson, Michael I. Jordan, Guillermo Sapiro, and Michael B. Wakin

Learning Low-Dimensional Signal Models

A Bayesian approach based on incomplete measurements

Sampling, coding, and streaming even the most essential data, e.g., in medical imaging and weather-monitoring applications, produce a data deluge that severely stresses the available analog-to-digital converter, communication bandwidth, and digital-storage resources. Surprisingly, while the ambient data dimension is large in many problems, the relevant information in the data can reside in a much lower dimensional space. This observation has led to several important theoretical and algorithmic developments under different low-dimensional modeling frameworks, such as compressive sensing (CS) [1], [2], matrix completion [3], [4], and general factor-model representations [5], [6]. These approaches have enabled new measurement systems, tools, and methods for information extraction from dimensionality-reduced or incomplete data. A key aspect of maximizing the potential of such techniques is to develop appropriate data models. In this article, we investigate this challenge from the perspective of nonparametric Bayesian analysis.

Before detailing the Bayesian modeling techniques, we review the form of measurements. Specifically, we consider measurement systems based on dimensionality reduction, where we linearly project the signal of interest into a lower-dimensional space via

$$\boldsymbol{y} = \Phi \boldsymbol{x} + \delta. \tag{1}$$

The signal is $\boldsymbol{x} \in \mathbb{R}^d$, the measurements are $\boldsymbol{y} \in \mathbb{R}^{d'}$, Φ is a $d' \times d$ matrix with $d' < d$, and δ accounts for noise.

Such a projection process loses signal information in general, since Φ has a nontrivial null space. Hence, there has been significant interest over the last few decades in finding dimensionality reductions that preserve as much information as possible in the incomplete measurements $\boldsymbol{y}$ about certain signals $\boldsymbol{x}$. One way to preserve information is for Φ to provide a stable embedding that approximately preserves pairwise distances between all signals in some set of interest. In some cases, this property allows the recovery of $\boldsymbol{x}$ from its measurement $\boldsymbol{y}$.

Digital Object Identifier 10.1109/MSP.2010.939733
Date of publication: 17 February 2011

Geometric data models, such as sparsity, union of subspaces, manifolds, and mixture of factor analyzers (MFAs), are at the core of low-dimensional modeling frameworks [7]. For instance, given a signal $\boldsymbol{x} \in \mathbb{R}^d$ and an appropriate basis $\Psi \in \mathbb{R}^{d \times d}$, we can transform the signal as $\boldsymbol{x} = \Psi\theta$, where θ is sparse or can be well approximated as such, that is, it has only a few nonzero elements. CS exploits this fact to recover signals from its compressive samples $\boldsymbol{y} \in \mathbb{R}^{d'}$, which are dimensionality-reducing, nonadaptive random measurements. According to CS theory, the number of measurements for stable recovery is proportional to the signal sparsity (hence, $d' \ll d$) rather than to its Fourier bandwidth, as dictated by the Shannon/Nyquist theorem. While signal recovery at such measurement rates is impressive, significant improvements can be achieved through the generalization of sparsity; for instance, union-of-subspace models encode dependencies among sparse coefficients; manifold models exploit smooth variations in the signals; and MFAs combine the strength of both models using a mixture of low-rank Gaussians [5], [7], [8].

The existing results in signal recovery from compressive or incomplete measurements of the type discussed in the "Stable Embeddings" section are predicated upon the knowledge of the appropriate low-dimensional signal model; a signal recovery algorithm relies on this model to locate the correct signal among all possible signals that can generate the same measurement.

In this article we consider the more difficult, but more broadly applicable, problem for which we must first learn the signal model from a set of training data. One can use this learned model to subsequently recover the underlying signal from compressive measurements. There are also examples for which we jointly learn the underlying model and recover the high-dimensional data, without any a priori training data; specifically, this is done when considering the image-interpolation problem (closely related to matrix completion), for which the underlying image is recovered based on the measurement of a small subset of pixels uniformly selected at random.

The tools and methods used to tackle the rich problems associated with learning low-dimensional signal models are based on probabilistic, nonparametric Bayesian techniques. By nonparametric, we mean that the number of parameters within the probabilistic models is unspecified beforehand. While it has been historically challenging to find workable prior distributions in the parameter space for such problems, we leverage the beta process (BP), Bernoulli process, Dirichlet process (DP), and Indian buffet process (IBP). We observe that these distributions provide a nice scaffold for analytically managing posterior distributions, given a set of training samples as well as observations. Additionally, we develop performance bounds for recovering high-dimensional data based on incomplete measurements. We present several examples of how this technology may be used in practice in CS, in matrix completion (when we recover a full low-rank matrix based on a small number of randomly sampled matrix elements), and in image interpolation based on highly incomplete measurements. These applications are of significant practical importance; for example, matrix-completion techniques are of interest for automatic recommendation systems (e.g., for movies, music, and books).

STABLE EMBEDDINGS

We consider several classes of low-dimensional models for which the dimensionality reduction process (1) is stable. This means that we not only have the information-preservation guarantee that $\Phi\boldsymbol{x}_1 \neq \Phi\boldsymbol{x}_2$ holds for all signal pairs $\boldsymbol{x}_1, \boldsymbol{x}_2$ belonging to the model set but also the guarantee that if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are far apart in $\mathbb{R}^d$, then their respective projections $\Phi\boldsymbol{x}_1$ and $\Phi\boldsymbol{x}_2$ are also far apart in $\mathbb{R}^{d'}$. This latter guarantee ensures the robustness of the dimensionality-reduction process to noise δ.

A requirement on the matrix Φ that combines both information preservation and stability properties for a signal model is the so-called ϵ-stable embedding property

$$(1-\epsilon)\,\|\,\boldsymbol{x}_1-\boldsymbol{x}_2\,\|_2^2 \leq \|\,\Phi\boldsymbol{x}_1-\Phi\boldsymbol{x}_2\,\|_2^2 \leq (1+\epsilon)\,\|\,\boldsymbol{x}_1-\boldsymbol{x}_2\,\|_2^2\,, \tag{2}$$

which must hold for all $\boldsymbol{x}_1, \boldsymbol{x}_2$ in the model set. The interpretation is simple: a stable embedding approximately preserves the Euclidean distances between all points in a signal model.

A dimensionality reduction $\boldsymbol{y} = \Phi\boldsymbol{x}$ from $\mathbb{R}^d$ down to $\mathbb{R}^{d'}$, $d' < d$, cannot hope to preserve all of the information in all signals $\boldsymbol{x} \in \mathbb{R}^d$, since it is impossible to guarantee that $\Phi x_1 \neq \Phi x_2$ holds for all signal pairs $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. This is because there are infinitely many $\boldsymbol{x} + \boldsymbol{x}'$, with $\boldsymbol{x}'$ from the $(d - d')$-dimensional null space of Φ, which yield exactly the same measurement $\boldsymbol{y}$. However, by restricting our attention only to signals from a low-dimensional model that occupies a subset of $\mathbb{R}^d$, such an information-preservation guarantee becomes possible, meaning that we can uniquely identify/recover any signal $\boldsymbol{x}$ in the model from its measurement $\boldsymbol{y}$.

Let us review the three deterministic model classes that have been shown to support stable dimensionality reduction. First, a sparse signal $\boldsymbol{x} \in \mathbb{R}^d$ can be represented in terms of just $k \ll d$ nonzero coefficients in the basis expansion $\boldsymbol{x} = \Psi\theta$, where Ψ is a fixed basis. Concisely, we say that $\|\,\theta\,\|_{\ell_0} = k$, where ℓ_0 is the pseudonorm that merely counts the nonzero entries in $\boldsymbol{x}$. The set of all sparse signals Σ_k is the union of $\binom{d}{k}$, k-dimensional canonical subspaces in $\mathbb{R}^d$, aligned with the coordinate axes of the basis Ψ. For sparse signals, the stable embedding property (2) corresponds to the restricted isometry property (RIP) [9]. Although the design of such a stable embedding is an nondeterministic polynomial time (NP) complete problem, in general, it has been shown that any independent identically distributed (i.i.d.) sub-Gaussian random matrix Φ stably embeds Σ_k into $\mathbb{R}^{d'}$ with high probability as long as $d' = O(k \log(d/k))$ [10].

Second, a structured sparse signal is not only sparse but also has correlated coefficients, such that it lies on one of a subset of the $\binom{d}{k}$ subspaces of Σ_k [8]. As a result, a random dimensionality reduction Φ is stable for a commensurately smaller value of d' than for a conventional sparse signal.

Third, an ensemble of articulating signals often live on a manifold, in particular, when the family of signals $\{\boldsymbol{x}_\theta : \theta \in \Theta\}$ is smoothly parameterized by a k-dimensional parameter vector θ [11]. The manifold dimension k is equal to the number of degrees of freedom of the articulation. It has been shown that a random dimensionality reduction Φ stably embeds a

k-dimensional smooth manifold from $\mathbb{R}^d$ into $\mathbb{R}^{d'}$ as long as $d' = O(k \log(d))$ [12].

Given a stable embedding of the form (1), a number of techniques have been developed to recover a (structured) sparse signal of interest $\boldsymbol{x}$ from the measurements $\boldsymbol{y}$, including various sparsity-promoting convex optimizations [1], [2], [13], greedy algorithms [14], [15], and Bayesian approaches [16]–[18]. Recently, algorithms have also been developed that recover signal manifolds from randomized measurements [19]. The challenge this article addresses concerns learning the underlying signal models, particularly union-of-subspace and manifold models, with this learning performed nonparametrically based on the available data. The MFA model discussed later is a statistical form of the union-of-subspace data model, and the MFA may also be used to approximate a manifold. Once these models are learned, they may be used in algorithms that seek to recover high-dimensional data based on low-dimensional compressive measurements.

LEARNING CONCISE SIGNAL MODELS

In this section, we assume that we may not have access to the model but instead to training data representative of the signals of interest. Our goal is to learn a concise signal model from this data, enabling stable signal recovery. We design these models in a statistical manner, using nonparametric Bayesian techniques.

UNION-OF-SUBSPACE MODEL FOR SPARSE SIGNALS

Assume access to a set of N training data $\{\boldsymbol{x}_n\}_{n=1,N}$. Our goal is to infer a concise model for $\{\boldsymbol{x}_n\}_{n=1,N}$ appropriate for recovering high-dimensional data from compressive measurements. Further, we would like to learn the model parameters nonparametrically (e.g., without having to set the dimensionality of the subspaces or the number of mixture components). We express each $\boldsymbol{x}_n \in \mathbb{R}^d$ as

$$\boldsymbol{x}_n = \mathbf{A}(\boldsymbol{c}_n \circ \boldsymbol{b}_n) + \epsilon_n, \tag{3}$$

where $\boldsymbol{c}_n \in \mathbb{R}^K$, $\boldsymbol{b}_n \in \{0,1\}^K$, and $\circ$ denotes a pointwise or Hadamard vector product. The columns of the matrix $\mathbf{A} \in \mathbb{R}^{d \times K}$ define a dictionary, and in many cases $K > d$, such that $\mathbf{A}$ may be overcomplete. Because of the binary nature of $\boldsymbol{b}_n$ and because $\|\boldsymbol{b}_n\|_{\ell_0} < d$, each $\boldsymbol{x}_n$ is represented by a subset of the columns of $\mathbf{A}$ (defining a subspace); ϵ_n is meant to represent the portion of $\boldsymbol{x}_n$ not contained within the aforementioned subspace.

If we assume that the components of ϵ_n are drawn from $\mathcal{N}(0, \alpha_0^{-1}\mathbf{I}_d)$, where $\mathbf{I}_d$ represents the d-dimensional identity matrix, and if $\boldsymbol{c}_n \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}_K)$, then, after integrating out $\boldsymbol{c}_n$, $\boldsymbol{x}_n$ is drawn from

$$\boldsymbol{x}_n \sim \mathcal{N}(0, \alpha^{-1}\mathbf{A}\Lambda_n\mathbf{A}^T + \alpha_0^{-1}\mathbf{I}_d), \tag{4}$$

where $\Lambda_n = \text{diag}(\boldsymbol{b}_n)$ is a binary diagonal matrix. Therefore, if the columns of $\mathbf{A}\Lambda_n$ are linearly independent, and if r_n represents the number of nonzero components in $\boldsymbol{b}_n$, then $\boldsymbol{x}_n$ is drawn from a zero-mean Gaussian with approximate rank r_n (approximate because $\alpha_0^{-1}\mathbf{I}_d$, with generally small α_0^{-1}, is added to the rank-r_n $\alpha^{-1}\mathbf{A}\Lambda_n\mathbf{A}^T$). Note that priors like $\boldsymbol{c}_n \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}_K)$, and other similar priors considered later, are typically selected for modeling convenience; the inferred posterior on such model parameters are not in general as simple as the prior (e.g., they are typically not Gaussian).

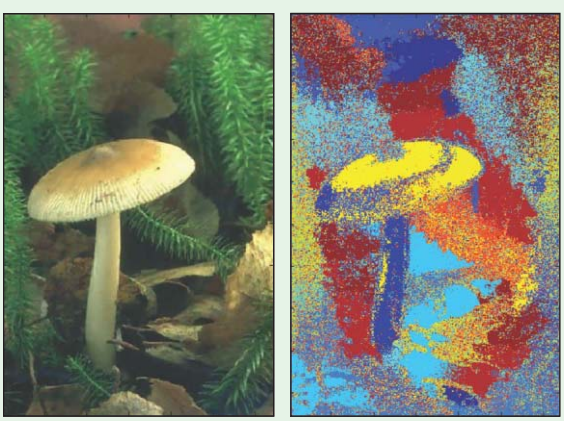
One of our objectives is to learn the dictionary matrix $\mathbf{A}$, and in a Bayesian setting, we place a prior on it. Specifically, a convenient prior is to draw the kth column of $\mathbf{A}$, $\boldsymbol{a}_k$, i.i.d. as

$$\boldsymbol{a}_k \sim \mathcal{N}\left(0, \frac{1}{d}\mathbf{I}_d\right), \quad k = 1, \ldots, K, \tag{5}$$

such that each column has a unit-expected norm, and the columns have zero-expected correlation. One typically also places gamma priors on the precisions α and α_0 (these priors are selected because of model conjugacy [20]).

The final part of the model involves placing a prior on the sparse matrix $\mathbf{B} \in \{0,1\}^{N \times K}$, with the nth row defined by $\boldsymbol{b}_n$; the cumulative set of binary vectors $\{\boldsymbol{b}_n\}_{n=1,N}$ defines the total number of columns needed from $\mathbf{A}$. The prior we will employ for $\{\boldsymbol{b}_n\}_{n=1,N}$ is the beta-Bernoulli process, which is closely connected to the IBP [21] developed by Griffiths and Ghahramani; this is discussed in detail in the "Completely Random Measures" section. At this point, we simply assume that an appropriate prior for $\mathbf{B}$ may be constituted.

As a first look at an application, to be discussed further in the "Applications" section, in Figure 1, $\{\boldsymbol{x}_n\}_{n=1,N}$ correspond to N patches of pixels from a red, green, and blue (RGB) image. In this problem, only 20% of the pixels are observed, selected uniformly at random, and the model is used to infer the missing pixels in the image. In this analysis, the incomplete data (image) are used as model inputs to infer all model parameters, and importantly, $\mathbf{A}$ and $\{\boldsymbol{c}_n \circ \boldsymbol{b}_n\}_{n=1,N}$. Note that the columns of $\mathbf{A}$



(a) (b)

[FIG1] Recovery of an RGB image based on measuring 20% of the voxels, uniformly at random. (a) Recovered image (PSNR = 29.73), (b) local usage of BP dictionary elements, where the color denotes a specific usage of a subset of dictionary elements. (Results courtesy of J. Paisley.)

have the same support as $\boldsymbol{x}_n$, and hence, they may be used to infer missing pixel values via $\mathbf{A}(c_n \circ \boldsymbol{b}_n)$. It is important that the pixels are missing at random; if the same pixel is missing in all $\{\boldsymbol{x}_n\}_{n=1,N}$, then it is impossible to learn the corresponding components in $\mathbf{A}$ (the row of $\mathbf{A}$ corresponding to this missing pixel cannot be inferred). Because the pixels are missing at random, information about a missing pixel may be inferred by using the information from a similar patch elsewhere in the image. Hence, the simultaneous (collaborative) analysis of all $\{\boldsymbol{x}_n\}_{n=1,N}$ allows one to infer information about the missing pixels by exploiting the observed versions of that pixel from similar patches (we are exploiting self-similarity between image patches, which is typical of natural imagery). Inferring interrelationships between incomplete patches with complementary missing pixels is consequently critical to model success.

MFA MODEL FOR SIGNAL ENSEMBLES

In the aforementioned model, all data share the same dictionary defined by the columns of $\mathbf{A}$, but each sample $\boldsymbol{x}_n$ generally employs a subset of the dictionary elements, defined by the binary vector $\boldsymbol{b}_n$. When the number of samples N is large, it can be expected that many of the $\boldsymbol{b}_n$ will be the same or similar, defining a union of subspaces. This can be statistically represented as a mixture of Gaussians with covariance matrices that are nearly of low rank.

Specifically, we generalize (4) as

$$\boldsymbol{x}_n \sim \sum_{m=1}^{M} \nu_m \mathcal{N}(\mu_m, \alpha_m^{-1}\mathbf{A}_m\Lambda_m\mathbf{A}_m^T + \alpha_0^{-1}\mathbf{I}_d), \tag{6}$$

where $\sum_{m=1}^{M}\nu_m = 1$, $\Lambda_m = \text{diag}(\boldsymbol{b}_m)$, with $\boldsymbol{b}_m \in \{0,1\}^K$, again a binary vector that selects particular columns of $\mathbf{A}_m$ to define the subspace spanned by the mth mixture component [in (3), there is a separate binary vector $\boldsymbol{b}_n$ for each sample $\boldsymbol{x}_n$, and now there is a related binary vector $\boldsymbol{b}_m$ associated with the mixture component m]. Note that the number of nonzero components in $\mathbf{A}_m$ may vary with m, implying that the dimensionality of the mixture components need not be the same. If for each m the number of nonzero components of $\boldsymbol{b}_m$ is small (i.e., $\| \boldsymbol{b}_m \|_{\ell_0} \ll d$), then each mixture component $\mathcal{N}(\mu_m, \alpha_m^{-1}\mathbf{A}_m\Lambda_m\mathbf{A}_m^T + \alpha_0^{-1}\mathbf{I}_d)$ defines a relatively low-dimensional pancake in $\mathbb{R}^d$, with the number of principal dimensions in the mth associated subspace defined by $\| \boldsymbol{b}_m \|_{\ell_0}$. The means μ_m locate the center of each pancake, and these are assumed to be drawn from $\mathcal{N}(0, \beta_m^{-1}\mathbf{I}_d)$, with a gamma prior placed on β_m (again due to conjugacy).

The model (6) is called MFA [22], and the nonzero columns of $\mathbf{A}_m\Lambda_m$ define the factor loadings associated with the mth mixture component. When building an MFA model, a natural question concerns how many mixture components M are appropriate for the training data $\{x_n\}_{n=1,N}$. One may use model-selection techniques to choose a single setting of M. Perhaps the most widely employed approach for choosing M is the Bayesian information criteria (BIC) [23]–[25]. Alternatively, later we consider nonparametric modeling, which yields a posterior distribution on M, and inference essentially performs model averaging across a weighted set of models with different M. This is implemented via DP [26], as summarized in the "Completely Random Measures" section.

Note that in (3) the $\boldsymbol{b}_n$ selects a subset of the columns of $\mathbf{A}$ for the representation of $\boldsymbol{x}_n$, and one may expect that different $\boldsymbol{x}_n$ will (partially) share the usage of these columns. In the mixture model of (6), $\boldsymbol{b}_m$ selects which subset of the columns of $\mathbf{A}_m$ are used for the mth mixture component; $\boldsymbol{b}_m$ therefore defines the dimensionality and subspace of this mixture component. In general, the subspaces spanned by $\mathbf{A}_m\Lambda_m$ and $\mathbf{A}_{m'}\Lambda_{m'}$ are different. Hence, (3) implies that $\boldsymbol{x}_n$ are drawn from partially overlapping subspaces, without an explicit clustering; (6) explicitly clusters the data, with the data in cluster m spanned by the nonzero columns of $\mathbf{A}_m\Lambda_m$. The representation in (6) is of most interest when one wishes to approximate a data manifold as a mixture of low-rank Gaussians, with the number of mixture components and their characteristics (e.g., ranks) inferred by the data.

A related model is the mixture of probabilistic principal component analyzers (MPPCA) framework of Tipping and Bishop [27]; MPPCA is similar to the proposed MFA, but in [27], one must set the dimensionality (rank) of each mixture component as well as the number of mixtures, where this is inferred via nonparametric Bayesian inference. In [27], the authors achieve a point estimate of model parameters via expectation maximization (EM), where we estimate a full posterior density function on model parameters.

MANIFOLD MODELS FOR SIGNAL ENSEMBLES

One intriguing use of the MFA model in (6) is for data living along a nonlinear k-dimensional manifold in $\mathbb{R}^d$. Locally, a k-dimensional manifold can be well approximated by its tangent plane, with the quality of this approximation depending on the local curvature of the manifold. Therefore, an MFA model as in (6) may be considered a candidate for manifold-modeled data, where the mean vectors μ_m roughly correspond to points sampled from the manifold, the columns of $\mathbf{A}_m\Lambda_m$ roughly span the k-dimensional local tangent spaces, the thickness parameter α_0^{-1} depends on the manifold curvature, and the weights ν_m reflect the density of the data across the manifold [28].

When an MFA model is used for recovering data of this type from compressive measurements, one will expect the recovered signal to draw only from a small number of MFA components. The recovered signal is therefore an affine combination of the columns of the few active $\mathbf{A}_m\Lambda_m$. This is reminiscent of the classical CS problem in which an unknown signal must be recovered as a sparse superposition of vectors from some dictionary. Indeed, one could alternatively formulate the MFA recovery program using CS techniques [5], in which $\widehat{x}$ is recovered as a sparse superposition of the columns of $\mathbf{A}_m$. A key consideration in this formulation, however, is that the set of selected columns may draw from only a few MFA components; this requirement is closely related to the notion of block sparsity that has been studied in CS. An example application of this framework is presented in Figure 2.

MATRIX COMPLETION

As a final model, consider a matrix $\mathbf{M} \in \mathbb{R}^{d\times N}$ with $N \geq d$ (this can always be achieved by matrix transpose). Let the N columns

[FIG2] Sparse signal recovery performed on an MFA model, inferred based on the training data using DP and BP. Column (a) shows the original data, and columns (b)–(g) represent CS recovery based on random compressive measurements. The results show the performance when the number of CS measurements are 5%, . . . , 30% of the total number of pixels in the image. In columns (b)–(g), the left figure employs the CS-recovery algorithm in [29], which does not exploit the MFA, and the right image is based on the learned MFA. (Results courtesy of M. Chen.)

of $\mathbf{M}$ constitute the set of vectors $\{x_n\}_{n=1}^N$, where x_n' is the nth column manifested with randomly selected missing entries (in this problem $x_n' = \Phi_n x_n$, where the rows of Φ_n are randomly selected rows of the $d \times d$ identity matrix, with Φ_n different for each n). If the matrix $\mathbf{M}$ is such that its columns satisfy the properties inherent to (4), specifically that each x_n resides approximately within a subspace defined by the columns in matrices of the form $\mathbf{A}\Lambda_n$, then the data-recovery technique discussed in the "Pixel/Voxel Recovery Via Union of Subspaces" section may be applied directly to achieve matrix completion.

It is of interest to examine how such a procedure is related to conventional matrix-completion frameworks based on low-rank constructions [3], [30], [31]. In this context, assume that the matrix may be expressed as

$$\mathbf{M} = \sum_{k=1}^{d} \lambda_k b_k u_k v_k^T + \mathbf{E}, \tag{7}$$

where $\lambda_k \in \mathbb{R}$, $b_k \in \{0, 1\}$, $u_k \in \mathbb{R}^d$, and $v_k \in \mathbb{R}^N$. We may again draw $u_k \sim \mathcal{N}(0, (1/d)\mathbf{I}_d)$, $v_k \sim \mathcal{N}(0, (1/N)\mathbf{I}_N)$, $\lambda_k \sim \mathcal{N}(0, \alpha^{-1})$, and each component of $\mathbf{E}$ drawn i.i.d. from $\mathcal{N}(0, \alpha_0^{-1})$. A binary sparseness-promoting prior (see the "Completely Random Measures" section) may again be employed to define the binary vector $\boldsymbol{b} = (b_1, \ldots, b_d)$, thereby imposing a preference for low-rank constructions. Note that, in this case, because it is assumed

that each column of $\mathbf{M}$ is drawn from the same linear subspace, there is only a single $\boldsymbol{b}$ (so in this case, we do not use a BP); however, this model may be clearly generalized to the nonlinear case by making $\boldsymbol{b}$ a function of index n, as in (3). This shows that the matrix completion problem is closely linked to the inference of missing pixels in images.

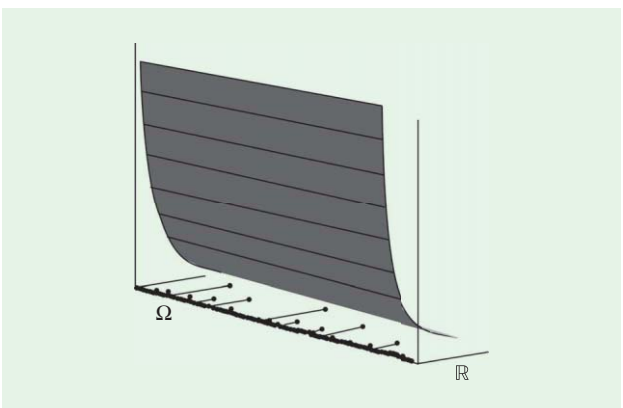
Considering (7), let v_{kn} represent the nth component of v_k. Then the nth column of $\mathbf{M}$ may be expressed as

$$\boldsymbol{x}_n = \mathbf{U}(\boldsymbol{c}_n \circ \boldsymbol{b}) + \epsilon_n, \tag{8}$$

where ϵ_n represents the nth column of $\mathbf{E}$, $\mathbf{U} \in \mathbb{R}^{d \times d}$ has columns defined by $\boldsymbol{u}_k$, and $\boldsymbol{c}_n = (\lambda_1 v_{1n}, \ldots, \lambda_d v_{dn})$. Therefore, matrix completion based on a sparseness constraint of the form in (7) represents each column of $\mathbf{M}$ as being drawn from a single linear subspace spanned by the columns of $\mathbf{U}$, while the union-of-subspace construction [32] in (3), applied to the N columns of $\mathbf{M}$, is nonlinear in that each column $\boldsymbol{x}_n$ in general has its own subspace defined by the binary vector $\boldsymbol{b}_n$.

BAYESIAN NONPARAMETRIC INFERENCE

Based on the earlier discussions, learning a model for concise representation of high-dimensional data requires the ability to infer the dimensionality of the subspace data reside in, with this defined by the number of columns needed in $\mathbf{A}$ and $\mathbf{U}$. Further, in the context of MFA model, we require a means of inferring an appropriate number of mixture components. The former problem will be addressed using the beta-Bernoulli process. The latter will be addressed via DP. These nonparametric models represent special cases of a more general concept, the completely random measure. Later, we first review the completely random measure, and then we show three examples for which it may be applied: for the BP, gamma process, and DP. Finally, we explain how BP may be combined with Bernoulli process to place a prior on the aforementioned matrix $\mathbf{B}$ (to infer the dimensionality of the subspace in which a signal resides), and how the DP may be used to infer the appropriate number of mixture components in MFA. For a thorough discussion of nonparametric Bayesian methods, the interested reader is referred to [33].



[FIG3] Construction of a completely random measure on Ω from a nonhomogeneous Poisson process on $\Omega \otimes \mathbb{R}$.

COMPLETELY RANDOM MEASURES

The key idea of Bayesian nonparametrics is easily stated: one replaces classical finite-dimensional prior distributions with general stochastic processes. Recall that a stochastic process is an indexed collection of random variables, where the index set may be infinite; thus, by using stochastic processes as priors, we introduce an open-ended number of degrees of freedom in a model. For this idea to be useful in practical models, it is necessary for these stochastic processes to have simplifying properties, and, in particular, it is necessary that they combine in simple ways with the likelihoods that arise in common statistical models so that posterior inference is feasible. One general approach to designing such stochastic processes is to make use of the notion of completely random measures, a class of objects that embody a simplifying independence assumption. We begin by presenting a general framework of completely random measures, and then we show how to derive some particularly useful stochastic processes—BP and DP—from this framework. When learning MFA, BP is used to infer the number of factor loadings (equivalently the rank) for each mixture component, while DP is used to infer the number of mixture components.

Letting Ω denote a measurable space endowed with a sigma algebra $\mathcal{A}$, a random measure G is a stochastic process whose index set is $\mathcal{A}$. That is, $G(A)$ is a random variable for each set A in the sigma algebra. A completely random measure G is defined by the additional requirement that whenever A_1 and A_2 are disjoint sets in $\mathcal{A}$, the corresponding random variables $G(A_1)$ and $G(A_2)$ are independent [34].

Kingman [34] presented a way to construct completely random measures based on the nonhomogeneous Poisson process. The construction runs as follows (see Figure 3 for a graphical depiction). Consider the product space $\Omega \otimes \mathbb{R}$ and place a product measure η on this space. Treating η as the rate measure for a nonhomogeneous Poisson process, draw a sample $\{(\omega_i, p_i)\}$ from this Poisson process. From this sample, form a measure on Ω in the following way:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}, \tag{9}$$

where δ_{ω_i} corresponds to a unit-point measure concentrated at the parameter/atom ω_i. We refer to $\{\omega_i\}$ as the atoms of the measure G and $\{p_i\}$ as the weights.

Clearly, the random measure defined in (9) is completely random, because the Poisson process assigns independent mass to disjoint sets. The interesting fact is that all completely random processes can be obtained this way (up to a deterministic component and Brownian motion).

BETA PROCESS

BP is an example of a completely random measure. In this case, we define the rate measure η as a product of an arbitrary measure B_0 on Ω and an improper beta distribution on $(0, 1)$

$$\eta(d\omega, dp) = cp^{-1}(1-p)^{c-1} dp\, B_0(d\omega), \tag{10}$$

where $c > 0$. Note that the expression $cp^{-1}(1-p)^{c-1}$ integrates to infinity; this has the consequence that a countably infinite number of points are obtained from the Poisson process.

We denote a draw from BP as

$$B \sim \mathrm{BP}(c, B_0), \tag{11}$$

where $c > 0$ is referred to as a concentration parameter and B_0 as the base measure.

For further details on this derivation of BP, see [35]. For an alternative derivation that does not make use of the framework of completely random measures, see [36]. Additional work on the applications of BP can be found in [37]–[39].

GAMMA PROCESS

As a second example, let the rate measure be a product of base measure G_0 and an improper gamma distribution

$$\eta(d\omega, dp) = cp^{-1}e^{-cp}dp\, G_0(d\omega). \tag{12}$$

Again the density on p integrates to infinity, yielding a countably infinite number of atoms. The resulting completely random measure is known as the gamma process. We write

$$G \sim \mathrm{GaP}(c, G_0) \tag{13}$$

to denote a draw from the gamma process. Note that the weights $\{p_i\}$ lie in $(0, \infty)$, and their sum is again finite.

DIRICHLET PROCESS

It is also of interest to consider random measures that are obtained from completely random measures by normalization. For example, returning to the rate measure defining the gamma process in (12), let $\{(\omega_i, p_i)\}$ denote the points obtained from the corresponding Poisson process. Form a random probability measure as follows:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}, \tag{14}$$

where $\pi_i = p_i / \sum_{j=1}^{\infty} p_j$. This is called DP [26]. We denote a draw from DP as $G \sim \mathrm{DP}(\alpha, H_0)$, where $\alpha = G_0(\Omega)$ and $H_0 = G_0/\alpha$.

APPLICATION TO DATA MODELS

From the "Learning Concise Signal Models" section, there are two principal modeling objectives: 1) an ability to infer the number of mixture components needed in an MFA and 2) the capacity to infer the number of needed factor loadings and their characteristics. Item 2) is related to inferring the binary matrix $\mathbf{B}$ discussed in the "Union-of-Subspace Model for Sparse Signals" section. First, considering 1), recall from above that a draw from DP may be expressed as $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$, where because of the aforementioned normalization $\sum_{i=1}^{\infty} \pi_i = 1$ and $\pi_i \geq 0$. In this application, the atoms $\{\omega_i\}_{i=1,\infty}$ correspond to the candidate mixture model parameters $(\alpha_m, \mu_m, \mathbf{A}_m, \Lambda_m)$ used in the mixture model of (6). Specifically, we constitute the following generative process for the data $\{\boldsymbol{x}_n\}_{n=1,N}$ when these data are assumed to be drawn from MFA:

$$\begin{aligned} \boldsymbol{x}_n &\sim f(\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n, \alpha_0), \\ (\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n) &\sim G, \\ G &\sim \mathrm{DP}(\alpha, G_0), \end{aligned} \tag{15}$$

where $f(\cdot)$ represents the Gaussian distribution in (6). In this case, the base measure G_0 from which ω_i is drawn corresponds to a factorized prior for the set of parameters $(\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n)$, with the individual components of that prior, as defined in the "Mixture of Factor Analyzers Model for Signal Ensembles" section. A gamma prior is also placed on α_0. Note that because of the form of $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$ with $\sum_{i=1}^{\infty} \pi_i = 1$ and $\pi_i \geq 0$, the set of parameters $\{\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n\}_{n=1,N}$ are characteristic of being drawn from a mixture model. With probability π_i, any particular set $(\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n)$ corresponds to ω_i. Therefore, although there are an infinite set of atoms in $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$, at most, N of them will be used in the generative process, and typically fewer than N are needed, as often the same atom ω_i is shared among multiple data samples $\{\boldsymbol{x}_n\}_{n=1,N}$. We therefore manifest a clustering of $\{\boldsymbol{x}_n\}_{n=1,N}$ (with data in the same cluster sharing a particular model parameter ω_i), and the model posterior density function allows inference of the number of mixture components needed. Therefore, in principle, the number of mixture components is unbounded, while in practice, the model allows one to infer the finite number of mixture components needed to represent the data.

There is a so-called Chinese restaurant process (CRP) viewpoint of DP. The data are viewed as customers, and the clusters are tables, with the dish associated with a given table manifested by the associated model parameters. One may explicitly draw from this CRP by marginalizing out DP draw G [33].

We now consider BP as a prior for the binary vectors $\{\boldsymbol{b}_n\}_{n=1,N}$ in the model (4); these binary vectors are also of interest in the matrix-completion problem in the "Matrix Completion" section of the "Learning Concise Signal Models" section. Recall that a draw from BP may be expressed as $G = \sum_{i=1}^{\infty} \pi_i \delta_{w_i}$, where each $\pi_i \in (0, 1)$. In this case, each atom ω_i corresponds to a potential column of the matrix $\mathbf{A}$ in (4) or a potential column of $\mathbf{U}$ in (8). Therefore, in this case, the base measure B_0 in BP corresponds to the same prior used for the columns $\mathbf{A}$ or $\mathbf{U}$. The random variable π_i defines the probability that the ith column of $\mathbf{A}$ or $\mathbf{U}$ is used to represent the data of interest. Specifically, the data sample $\boldsymbol{x}_n$ selects from among dishes in a buffet, where the dishes correspond to the columns of $\mathbf{A}$ or $\mathbf{U}$. With probability π_i data, $\boldsymbol{x}_n$ selects the ith column of $\mathbf{A}$ or $\mathbf{U}$, with the respective column denoted by the atom ω_i. Therefore, π_i defines the parameter of a Bernoulli distribution, with which a particular $\boldsymbol{x}_n$ decides which ω_i to use for data representation. This beta-Bernoulli process therefore defines a binary matrix $\mathbf{B} \in \{0, 1\}^{N \times \infty}$, where each row corresponds to a particular data sample $\boldsymbol{x}_n$ and the columns correspond to specific atoms $\{\omega_i\}_{i=1,\infty}$; hence, the rows of $\mathbf{B}$ are defined by $\{\boldsymbol{b}_i\}_{i=1,N}$, and sample $\boldsymbol{x}_n$ selects atom/dish ω_i if the ith component of $\boldsymbol{b}_n$ is equal to one. While $\mathbf{B}$ has an infinite number of columns in principle, it can be shown that only a finite number of columns in each row will have nonzero values [21]; in practice, one may truncate the model to K columns/atoms for large K.

The beta-Bernoulli process yields a so-called IBP [33] if the BP draw is marginalized out. In this construction, the data are again customers, and the model parameters are dishes at a buffet. Each customer sequentially selects parameters from the buffet, where the binary vector $\boldsymbol{b}_n$ for customer/data n defines which dishes/parameters are selected; if the kth component of $\boldsymbol{b}_n$ is one, then the kth parameter is used by data n, and if the kth component of $\boldsymbol{b}_n$ is zero, the kth parameter is not used.

POSTERIOR INFERENCE

Markov chain Monte Carlo (MCMC) procedures provide the dominant approach to inference with random measures. In such methods, one approximates the posterior distribution of all model parameters in terms of a set of parameter-vector samples. These samples yield an ensemble of models, and the relative frequency of samples approximates the posterior distribution. In this manner, one need not explicitly compute the high-dimensional integrals that would be required of a direct evaluation of the posterior distribution.

A special case of MCMC is Gibbs sampling, for which samples from the posterior distribution are drawn by sequentially sampling conditional distributions. By appropriate design of the model, of the form discussed earlier, these conditional distributions may often be expressed analytically. As an example of such samplers, consider DP in particular. Working the marginal distribution embodied in CRP, the core problem is to sample the seating assignment of a single customer conditioning on the seating assignments of the remaining customers. By exchangeability, one can pretend that this customer is the last to arrive in the restaurant, and the contribution of the prior to the seating assignment becomes the following rule: the customer sits at a table with the probability proportional to the number of customers at that table. Multiplying this prior by a likelihood term, one obtains a conditional probability that can be sampled. Similarly, in models based on BP, one can work with the marginal distribution embodied in IBP, and sampling the sparse binary vector associated with a data point by pretending that that data point is the last to arrive in the restaurant.

The insight of exploiting exchangeability in inference for random measures is due to Escobar [40], and a large literature has emerged. See Neal [41] for a thorough discussion in the case of DP. Another direction of research on inference has involved working directly with the random measures rather than the marginals obtained from these random measures. There have been two main approaches: 1) truncate the random measure by limiting the random measure to a fixed number of atoms that is larger than any value expected to arise during sampling [42] and 2) use slice sampling to adaptively truncate the random measure [43]. See [44] for a discussion of these methods in the setting of BP and for pointers to literature on variational approaches to inference for random measures.

APPLICATIONS

To illustrate the broad applicability and high performance of the earlier approach to learning concise signal models, we consider several representative examples.

PIXEL/VOXEL RECOVERY VIA UNION OF SUBSPACES

We first consider an application of the union-of-subspaces model from (4), in which it is assumed that the data $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1,N}$ are constituted from pixels/voxels in an image. Specifically, each of the N image patches is defined by a set of contiguous pixels, with $\boldsymbol{x}_n \in \mathbb{R}^d$ representing data from the nth patch (it is possible that patches may overlap). For a color image, one often considers $d = 8 \cdot 8 \cdot 3 = 192$, corresponding to the RGB components of an 8×8 image patch.

From (3), note that $\boldsymbol{x}_n$ is defined by the matrix $\mathbf{A} \in \mathbb{R}^{d \times K}$, by the sparse vector $\boldsymbol{c}_n \circ \boldsymbol{b}_n$ (with r_n nonzero components), and by $\epsilon_n \in \mathbb{R}^d$. We assume that ϵ_n may be made negligibly small, which implies $\alpha_0 \gg \alpha$ (to be demonstrated in the experiments). Since $\mathbf{A}$ is shared for all vectors in $\mathcal{D}$, the total number of real-model components needed is $dK + \sum_{n=1}^{N} r_n$, if the $\{\boldsymbol{b}_n\}_{n=1,N}$ are known (i.e., if it is assumed we know which columns of $\mathbf{A}$ are associated with each $\boldsymbol{x}_n$, with this clearly impossible in practice). Nevertheless, under these assumptions, note that we have Nd real numbers in $\mathcal{D}$ available for computation of $dK + \sum_{n=1}^{N} r_n$ real-model parameters. Therefore, if $N \gg K$ and $r_n \ll d$, and if we process all $\{\boldsymbol{x}_n\}_{n=1,N}$ jointly to infer the cumulative set of model parameters (exploiting the fact that $\mathbf{A}$ is shared among all), it appears that we have more data in $\mathcal{D}$ than needed. Further, it would appear that we have enough data to infer, which among the r_n columns of $\mathbf{A}$ are needed for the representation of each $\boldsymbol{x}_n$ (therefore, we do not need a priori access to $\{\boldsymbol{b}_n\}_{n=1,N}$).

On the basis of the earlier observations, researchers have recently assumed access to only a subset of components of each $\boldsymbol{x}_n$, with the observed components selected uniformly at random [38], [45], [46]. For example, rather than measuring all $d = 192$ contiguous pixels in an $8 \times 8 \times 3$ patch, a fraction of the pixels is measured, with the measured subset of pixels selected uniformly at random. Let $\mathcal{D}' = \{\boldsymbol{x}'_n\}_{n=1,N}$ represent a modified form of $\mathcal{D}$, with each $\boldsymbol{x}'_n$ defined by a fraction of the components of each $\boldsymbol{x}_n$, with the observed samples selected uniformly at random. Processing all of the data in $\mathcal{D}'$ jointly (collaboratively), it has been demonstrated that for real, natural images, one may indeed recover the missing data accurately, even when downsampling $\mathcal{D}$ significantly. Further, the compressive measurements may be performed very simply: by just randomly sampling/measuring the pixels/voxels in existing cameras (no need to develop new compressive-sampling cameras). An example is shown in Figure 1.

Note that this looks like CS [47], [48], in that a small subset of measurements are performed, with the full data recovered based upon the exploitation of properties of the signal (that the signals live in a low-dimensional subspace of $\mathbb{R}^d$). However, in CS, it is typically assumed that projection-type measurements are performed and that the signal is sparse in an underlying known basis or frame. The projections should be incoherent with the basis vectors [49], and for a discrete cosine transform (DCT)-type basis, one could use delta-function-like projections (selecting random components of each $\boldsymbol{x}_n$), like those considered earlier. However, in an earlier collaborative-filtering framework, we not only perform random sampling but also infer the underlying union of subspaces in which the signals reside, as

defined by the columns of $\mathbf{A}$, thereby matching the signal subspace to the observed data adaptively. In fact, as discussed in the "Matrix Completion" section of the "Applications" section, collaborative filtering for image recovery is closer to the field of matrix completion [3] than it is to CS.

The model in (4) is well suited for recovering the missing components of $\mathcal{D}$ from $\mathcal{D}'$ [38]. Specifically, when performing computations for the posterior distribution of the model parameters, the likelihood function represented by $\prod_{n=1x}^{N} \mathcal{N}(\boldsymbol{x}_n'|0, \alpha^{-1}\mathbf{A}\Lambda_n\mathbf{A}^T + \alpha_0^{-1}\mathbf{I}_d)$ is simply evaluated at the pixels for which data are observed. As discussed in the "Posterior Inference" section, a Gibbs sampler may be implemented, and from this, one may obtain an approximation to all model parameters. Hence, the posterior probability of each $\boldsymbol{x}_n$ in $\mathcal{D}$, based on observed $\mathcal{D}'$ and model hyperparameters Θ, may be expressed as

$$p(\boldsymbol{x}_n|\mathcal{D}', \Theta) = \int_{\mathbf{A}} \int_{\alpha} \int_{\Lambda_n} \int_{\alpha_0} \mathcal{N}(\boldsymbol{x}_n|0, \alpha^{-1}\mathbf{A}\Lambda_n\mathbf{A}^T + \alpha_0^{-1}\mathbf{I}_d) \times p(\mathbf{A}, \alpha, \Lambda_n, \alpha_0|\mathcal{D}', \Theta), \quad (16)$$

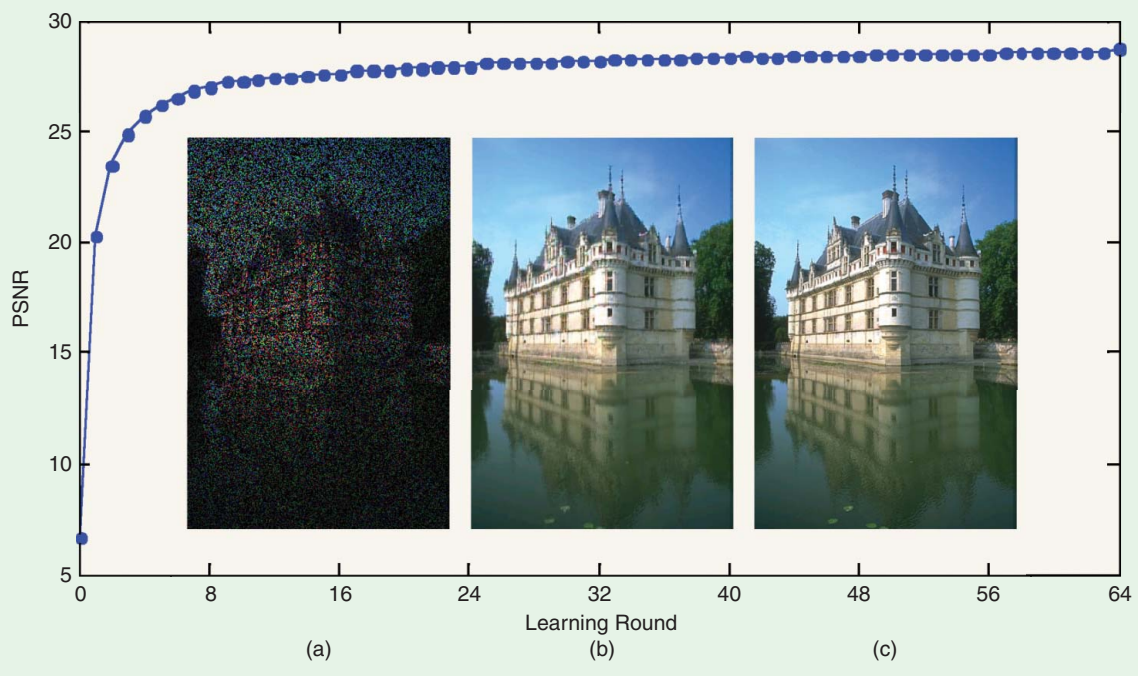
with the posterior $p(\mathbf{A}, \alpha, \Lambda_n, \alpha_0|\mathcal{D}', \Theta)$ approximated via samples from the Gibbs computations, and the integrals are approximated as sums.

A model like that in (3) or (6) has a posterior on model parameters that is invariant to exchanging the order of the data $\{\boldsymbol{x}_n\}_{n=1,N}$. In other words, any permutation of the order of the data will yield exactly the same inferred model parameters. This implies that the model is not utilizing all available prior information, since if one reconstituted an image after permuting the order of $\{\boldsymbol{x}_n\}_{n=1,N}$, very distinct images are manifested (recall that each $\boldsymbol{x}_n$ corresponds to an $8 \times 8 \times 3$ patch of contiguous pixels, and reordering these patches causes significant changes to the overall image). It is therefore desirable to impose within the model that if $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ are spatially proximate, they will likely employ similar factors (manifested by similar binary factor-selection vectors $\boldsymbol{b}_n$ and $\boldsymbol{b}_{n'}$).

Toward this end, we utilize DP in an additional manner (beyond within the MFA) to exploit spatial information. Specifically, we cluster the image patches spatially using a DP and impose that if two patches are spatially proximate, they are likely to be drawn from the same Gaussian mixture component and from the spatial mixture component. Figure 1(b) uses a different color to represent each Gaussian mixture component, and effective spatial segmentation is realized. One may therefore envision extending this framework for simultaneous image recovery and segmentation based on randomly subsampled images.

As another example of this type, consider Figure 4. In this example, rather than processing all possible (overlapping) image patches at once, we select a subset of them for analysis; the approximate posterior on model parameters so inferred is used as a prior for the next randomly selected subset of patches for analysis. In Figure 4, the peak signal-to-noise ratio (PSNR) curve shows how the model performance improves as we consider more data in a sequential manner. Each analysis of a subset of the image patches is termed a learning round.

Theoretically, one would expect to need thousands of Gibbs iterations to achieve convergence. However, our experience is that even a single iteration in each of the above B^2 rounds yields



[FIG4] **Inpainting results. The curve shows the PSNR as a function of the 64 Gibbs learning rounds. Part (a) shows is the test image, with 80% of the RGB pixels missing, (b) is the result 64 after Gibbs rounds (final result), and (c) is the original uncontaminated image. (Results courtesy of M. Zhou.)**

good results. In Figure 4, we show the PSNR as a function of each of the 64 rounds discussed earlier. For Gibbs rounds 16, 32, and 64, the corresponding PSNR values were 27.64, 28.20, and 28.66 dB. For this example, we used $K = 256$. This example was considered in [50]; the best results reported was a PSNR of 29.65 dB. However, to achieve those results, a training data set was employed for initialization [50]; the BP results are achieved with no a priori training data.

SIGNAL RECOVERY FROM MFAs AND MANIFOLDS

Assume that it is known a priori that the data of interest are drawn from an MFA of the form in (6), with the MFA learned offline based upon training data $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1,N}$. We now wish to measure a single new $\boldsymbol{x} \in \mathbb{R}^d$, under the assumption that $\boldsymbol{x}$ is drawn from the same MFA [5]. Since the MFA may be used to approximate a manifold, we may also consider the case for which $\boldsymbol{x}$ is drawn from a known manifold [12]. On the basis of this prior knowledge, we wish to measure $\boldsymbol{y} \in \mathbb{R}^{d'}$, with $d' < d$, and ideally with $d' \ll d$; based on the measured $\boldsymbol{y}$, $\boldsymbol{x}$ is recovered.

It is assumed that $\boldsymbol{y} = \Phi\boldsymbol{x}$, where $\Phi \in \mathbb{R}^{d' \times d}$ is a projection matrix typically defined randomly. In the "Bayesian Nonparametric Inference" section, we discuss the desired properties of Φ and the connection of such to the characteristics of MFA. In a statistical sense, to recover $\boldsymbol{x}$ from $\boldsymbol{y}$, we desire $p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})/p(\boldsymbol{y})$, under the assumption that $p(\boldsymbol{x})$ is the known MFA of the form in (6). Assuming that the compressive measurements are noisy, we may express $\boldsymbol{y} = \Phi\boldsymbol{x} + \delta$, where $\delta \in \mathbb{R}^{d'}$ represents additive noise. If $\delta \sim \mathcal{N}(0, \beta_0^{-1}\mathbf{I}_{d'})$, then we have $p(\mathbf{y}|\boldsymbol{x}) = \mathcal{N}(\Phi\boldsymbol{x}, \beta_0^{-1}\mathbf{I}_{d'})$. If β_0 is known, under the MFA assumption for $p(\boldsymbol{x})$, the expression $p(\boldsymbol{x}|\boldsymbol{y})$ may also be expressed analytically in terms of a mixture of Gaussians. If needed, we may also infer β_0 by placing a (conjugate) gamma prior on it. Therefore, under the assumption of an MFA model for $p(\boldsymbol{x})$, one may readily constitute a statistical estimate of $\boldsymbol{x}$ based on the observed $\boldsymbol{y}$, with performance bounds discussed in the "Performance Guarantees" section. An example of CS recovery for images that live on a union of subspaces is shown in Figure 2; we are unaware of such CS inversion being performed by any previous method.

MATRIX COMPLETION

As our final example, we consider the problem of matrix completion, as applied to movie-rating matrices. We tested the union-of-subspace construction on the widely employed 10M MovieLens dataset (10,681 movies by 71,567 users). Table 1 shows the results for both r_a and r_b partitions provided with the data, in which ten ratings per user are held out for testing. One of the best competing algorithms is the Gaussian process latent-variable model (GP-LVM) [51]. Averaged over both partitions, the GP-LVM reports the root-mean-square error (RMSE) of 0.8740 ± 0.0278 using a ten-dimensional latent space, while the baselines of our approaches achieve average RMSEs of 0.8539 ± 0.0298 and 0.8499 ± 0.0250. In this example, we employed the model in (3) in two constructions. In Table 1, we show the results when the vectors $\boldsymbol{x}_n$ correspond to the user-dependent rankings of all movies (user profile), and with $\boldsymbol{x}_n$ corresponding to the movie-dependent rankings manifested by all people (movie profile). In other words, one construction is in terms of the rows of the ranking matrix and the other construction is in terms of the columns, with the state-of-the-art results manifested in each case. While the Bayesian models may readily be extended to integer-observed matrices via a probit or logistic link function, the integer values are simply approximated as real numbers.

[TABLE 1] RMSE OF UNION-OF-SUBSPACE MODEL ON 10 MILLION MOVIELENS DATA.

METHODS	r_a PARTITION	r_b PARTITION
USER PROFILE	0.8749 ± 0.0009	0.8328 ± 0.0004
MOVIE PROFILE	0.8676 ± 0.0006	0.8323 ± 0.0002

RESULTS COURTESY OF M. ZHOU.

These results were computed using a Gibbs sampler, with a truncated BP implementation with $K = 256$ dishes. One Gibbs iteration required 150 s on a 2.53-GHz E5540 Xeon processor, using nonoptimized MATLAB software. The results in Table 1 correspond to 50 burn-in iterations and 100 collection iterations.

PERFORMANCE GUARANTEES

The BP and DP nonparametric methods may be used to infer an MFA based upon the given training data. Once this model is so learned, MFA may be assumed known and can be used in the inversion of subsequent compressive measurements. An example of such MFA learning and the subsequent utilization within CS signal recovery was presented in Figure 2. It is of interest to examine the performance guarantees based on CS measurements and a known MFA model (learned based on training data, using nonparametric techniques of the type discussed earlier). It should be emphasized that the underlying MFA for general data is typically not identifiable or unique. This implies that multiple MFAs may provide similar generative models for the underlying data of interest. For the following bounds, we assume that one learned MFA is used to perform CS inversion, and this model provides an accurate statistical representation of the data; it is for such a learned MFA that the bounds are constituted.

BOUNDS FOR MFAs

Recall our expression for compressive measurements $\boldsymbol{y} = \Phi\boldsymbol{x} + \delta \in \mathbb{R}^{d'}$ of a signal $\boldsymbol{x} \in \mathbb{R}^d$ as in (1). As discussed in the "Signal Recovery from MFAs and Manifolds" section, if we assume that $\boldsymbol{x}$ is drawn from an MFA of the form in (6), whose parameters are known (based on training data), then the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ can be expressed as a mixture of Gaussians [5]. Using this model, we obtain an analytical expression for the mean estimate of $\boldsymbol{x}$

$$\widehat{\boldsymbol{x}} = \sum_{m=1}^{M} \widehat{\nu}_m \widehat{\boldsymbol{x}}_m, \tag{17}$$

where

$$\widehat{\nu}_m = \frac{\nu_m \mathcal{N}(\boldsymbol{y}; \Phi\mu_m, \beta_0^{-1}\mathbf{I}_{d'} + \Phi\Omega_m\Phi^T)}{\sum_{\ell=1}^{M} \nu_\ell \mathcal{N}(\boldsymbol{y}; \Phi\mu_\ell, \beta_0^{-1}\mathbf{I}_{d'} + \Phi\Omega_\ell\Phi^T)}$$

represents an estimated mixture weight of the mth component,

$$\hat{x}_m = \Omega_m \Phi^T (\beta_0^{-1} \mathbf{I}_{d'} + \Phi \Omega_m \Phi^T)^{-1} (y - \Phi \mu_m) + \mu_m$$

equals the signal estimate that would be recovered if only component m was present in the MFA, and $\Omega_m = \alpha_m^{-1} \mathbf{A}_m \Lambda_m \mathbf{A}_m^T + \alpha_0^{-1} \mathbf{I}_d$ represents the covariance matrix of the mth component in the MFA.

We can also consider the situation where $\alpha_0^{-1} \to 0$ and $\beta_0^{-1} \to 0$, in which case the matrix inverse in (17) should be treated as a pseudoinverse.

For an MFA being used for manifold-modeled data, an analogous requirement to the stable embedding is that (2) holds for $\mu_{m_1} - \mu_{m_2}$ for all $1 \le m_1, m_2 \le M$ and that (2) also holds for all vectors in colspan$(\mathbf{A}_m \Lambda_m)$ for all $1 \le m \le M$. From the Johnson-Lindenstrauss lemma, we know that when Φ is generated randomly with i.i.d. Gaussian or sub-Gaussian entries, the former property holds with high probability as long as $d' = O(\log(M)\epsilon^{-2})$, and using similar arguments, the latter property also holds as long as $d' = O((k + \log(M))\epsilon^{-2})$ [10].

Under the assumption that these two conditions are met, we can establish certain guarantees [28] about the performance of the mean estimator (17) when recovering a signal x that is drawn from the manifold.

For example, supposing that $\beta_0^{-1} \to 0$, the isometry property for colspan $(\mathbf{A}_m \Lambda_m)$ discussed essentially guarantees that $\| x - \hat{x}_m \|_2$ is a combination of the two error terms, one depending on the size of $x - \mu_m$ when projected onto colspan$(\mathbf{A}_m \Lambda_m)$, and one depending on the size of $x - \mu_m$ when projected orthogonal to colspan$(\mathbf{A}_m \Lambda_m)$.

The size of α_0^{-1} controls the balance between these two terms, and by choosing sufficiently small α_0^{-1}, we can ensure that the dependence on the first of these terms is small; this allows us to guarantee that $\| x - \hat{x}_m \|_2$ is small for any signal x living near mixture component m.

Analysis of the recovery error $\| x - \hat{x} \|_2$ for a multicomponent mean estimator (17) is more involved but can proceed based on the observation that for any $m_0 \in \{1, 2, \ldots, M\}$, we can write $\| x - \hat{x} \|_2 \le \| x - \hat{x}_{m_0} \|_2 + \sum_{m \ne m_0} \hat{v}_m \| x - \hat{x}_m \|_2$. One conclusion that can be drawn from this is that if the mixture centers $\{\mu_m\}$ are well separated in $\mathbb{R}^d$ and remain well separated in $\mathbb{R}^{d'}$ (as discussed earlier), then for a signal x living near mixture component m_0, all $\hat{v}_m$ will be small for $m \ne m_0$, thus $\| x - \hat{x} \|_2$ will be small.

We refer the interested readers to [28] for a more detailed analysis.

BOUNDS ON MATRIX COMPLETION

Earlier, we considered bounds for CS measurements and an underlying MFA model, with the example results; in Figure 2, the CS projection matrix Φ was defined by draws from a Gaussian distribution. In Figure 1, rather than taking such random-projection measurements, we observed a small subset of pixels, with these selected at random. This is closely related to the matrix-completion problem, for which we briefly review theoretical guarantees. We also showed the experimental results in the "Applications" section for the matrix-completion problem, using nonparametric Bayesian techniques.

Several recent papers have examined the problem of recovering a low-rank matrix from just a fraction of its entries. As in the "Matrix Completion" section of the "Learning Concise Signal Models" section, let us consider a matrix $\mathbf{M} \in \mathbb{R}^{d \times N}$ with $N \ge d$ and let us suppose that $\mathbf{M}$ has rank r. Because such a matrix has only $(d + N - r)r$ degrees of freedom [4], it seems natural that one may be able to recover the matrix when observing far less than all of its dN entries. We denote by $m \ll dn$ the number of available entries.

A recent approach for recovering the missing entries of $\mathbf{M}$ involves solving a convex optimization problem, wherein one seeks the matrix $\mathbf{M}'$ having the smallest nuclear norm, such that $\mathbf{M}'$ agrees with $\mathbf{M}$ at the m observed entries. (The nuclear norm of a matrix equals the sum of its singular values.) As an example, it has been shown that, with high probability, nuclear norm minimization recovers the matrix $\mathbf{M}$ exactly supposing that $m \sim CNr \log^6 N$ and that the locations of the m observed positions are drawn uniformly at random [52]; the constant C in this expression depends on the coherence of the singular vectors of $\mathbf{M}$, implying that some matrices are easier to recover than others. Similar statements [4] have also been made for matrix recovery in terms of a generalization of RIP from CS, which is discussed in the "Stable Embeddings" section. Like signal recovery in CS, matrix completion has also been shown to be robust to noise in the observed entries [53].

To the best of our knowledge, there do not exist bounds available for the alternative form of matrix completion discussed in the "Matrix Completion" section of the "Learning Concise Signal Models" section, in which each column of the matrix is defined by a unique subspace and thus conventional rank minimization techniques will not be appropriate. This problem includes conventional rank-based matrix recovery as a special case (when the columns happen to share a common subspace), however, it is likely to be more difficult to solve in general, both in terms of the requisite number of observations and in terms of algorithmic complexity.

CONCLUSIONS

While the dimensionality of data used for visualization by humans (e.g., imagery and video) may be very large, the underlying information content in the data may be relatively low. We have reviewed addressing this problem through the representation of data in terms of the underlying (low-dimensional) manifold or union of subspaces on which it resides. By exploiting this low-dimensional representation, one may significantly reduce the quantity of data that need be measured from a given scene (or needed within a general data matrix), manifesting compressive or incomplete measurements. There are several technical challenges that must be addressed, including development of models to learn the underlying low-dimensional latent space. In this article, we have examined such learning from a nonparametric Bayesian viewpoint, with the example results presented for CS of signals

that reside on a union of subspaces, image interpolation, and matrix completion. We have also reviewed the theoretical results on the accuracy of data recovery for such problems.

Concerning future research, note that the discussion in the "Bayesian Nonparametric Inference" section on completely random measures is quite general, with the DP and beta-Bernoulli processes considered here as special cases. It is of interest to consider more general nonparametric models. For example, such models may be replaced by generalized forms, which yield power-law behavior in the number of clusters and dictionary elements as a function of the quantity of data. Such power-law behavior may be better matched to the properties of real data, such as images, video, and general matrices. There are early and promising studies that have examined this power-law construction [54], [55].

ACKNOWLEDGMENTS

Many graduate students contributed to the ideas and results reviewed in this article. The authors particularly acknowledge the contributions of Minhua Chen, Armin Eftekhari, John Paisley, and Mingyuan Zhou. The authors also thank the reviewers for a careful reading of the original version of this article and suggestions that led to a significantly improved final article.

AUTHORS

Lawrence Carin (lcarin@ece.duke.edu) earned his B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1985, 1986, and 1989, respectively. He is the William H. Younger Professor in the Electrical and Computer Engineering Department at Duke University, a cofounder of Signal Innovations Group, where he is the director of technology, and a Fellow of the IEEE. His recent research addresses applications of nonparametric Bayesian methods to general signal analysis, with an emphasis on sensing.

Richard G. Baraniuk (richb@rice.edu) is the Victor E. Cameron Professor of Electrical and Computer Engineering at Rice University. His work on the Rice single-pixel compressive camera was selected by *MIT Technology Review* as a "TR10 Top 10 Emerging Technology" for 2007. He is a Fellow of the IEEE and the American Association for the Advancement of Science and has received national young investigator Awards from the National Science Foundation and the Office of Naval Research, the Rosenbaum Fellowship from the Isaac Newton Institute of Cambridge University, the ECE Young Alumni Achievement Award from the University of Illinois, and the Wavelet Pioneer Award from SPIE. He is the founder of Connexions. His research interests lie in new theory, algorithms, and hardware for sensing and signal processing.

Volkan Cevher (volkan.cevher@epfl.ch) received his B.Sc. degree (valedictorian) in electrical engineering from Bilkent University in 1999, and his Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology in 2005. He was a research scientist at the University of Maryland, College Park (2006–2007) and at Rice University (2008–2009). Currently, he is an assistant professor at Ecole Polytechnique Federale de Lausanne, with a joint appointment at Idiap Research Institute and a faculty fellow at Rice University. His research interests include signal processing theory, machine learning, graphical models, and information theory.

David Dunson (dunson@stat.duke.edu) is the professor of statistical science at Duke University. He is a fellow of the American Statistical Association and the Institute of Mathematical Statistics. He is the winner of the 2007 Mortimer Spiegelman Award for the top public health statistician, the 2010 Myrto Lefkopoulou Distinguished Lectureship at Harvard University, and the 2010 COPSS Presidents' Award for the top statistician under 41. His research focuses on developing Bayesian statistical methods motivated by high-dimensional and complex data sets.

Michael I. Jordan (jordan@eecs.Berkeley.edu) is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. He received the IEEE Neural Networks Pioneer Award in 2006, the SIAM Activity Group on Optimization Prize in 2008, and the ACM/AAAI Allen Newell Award in 2009. He was named as Neyman Lecturer and Medallion Lecturer by IMS. He is a Fellow of AAAS, IMS, IEEE, AAAI, and ASA. In 2010, he was named to the National Academy of Engineering and the National Academy of Sciences. His research has focused on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines, and applications to problems in computational biology, information retrieval, signal processing, and speech recognition.

Guillermo Sapiro (guille@umn.edu) received his B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees from the Department of Electrical Engineering at Technion, Israel Institute of Technology, in 1989, 1991, and 1993, respectively. After postdoctoral research at MIT, he became a member of technical staff at the research facilities of HP Labs in Palo Alto, California. He is currently with the Department of Electrical and Computer Engineering at the University of Minnesota, where he holds the position of Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. He received the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for postdoctoral studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, the NSF Career Award in 1999, and the National Security Science and Engineering Faculty Fellowship in 2010. He is the funding editor-in-chief of *SIAM Journal on Imaging Sciences*.

Michael B. Wakin (mwakin@mines.edu) earned his Ph.D. degree in electrical engineering from Rice University in 2007. He was an NSF mathematical sciences postdoctoral research fellow at the California Institute of Technology. Before joining the Division of Engineering at the Colorado School of Mines, he was an assistant professor at the University of Michigan. His research interests include sparse, geometric, and manifold-based

models for signal and image processing, approximation, compression, CS, and dimensionality reduction. He is a recipient of the Hershel M. Rich Invention Award from Rice University and DARPA Young Faculty Award.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2005.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, 2006.

[3] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, pp. 717–772, 2008.

[4] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," submitted for publication.

[5] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Trans. Signal Processing*, 2010.

[6] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. Int. Conf. Machine Learning*, 2009 .

[7] R. Baraniuk, V. Cevher, and M. Wakin, "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective," in *Proc. IEEE*, to be published.

[8] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde (2008). Model-based compressive sensing. Preprint [Online]. Available: http://dsp.rice.edu/cs

[9] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, 2005.

[10] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Construct. Approx.*, vol. 28, pp. 253–263, 2008.

[11] D. L. Donoho and C. Grimes, "Image manifolds which are isometric to Euclidean space," *J. Math. Imag. Comput. Vision*, vol. 23, no. 1, pp. 5–24, July 2005.

[12] R. Baraniuk and M. Wakin, "Random projections of smooth manifolds," *Foundat. Comput. Math.*, vol. 9, no. 1, pp. 51–77, 2009.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scient. Comput.*, vol. 20, p. 33, 1998.

[14] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Ann. Asilomar Conf. Signals, Systems, and Computers*, 1993 .

[15] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, to be published.

[16] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[17] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[18] P. Schniter, L. C. Potter, and J. Ziniel, "Fast Bayesian matching pursuit," in *Proc. Information Theory and Applications Workshop*, 2008, pp. 326–333.

[19] M. B. Wakin, "Manifold-based signal recovery and parameter estimation from compressive measurements," *Preprint*, 2008.

[20] J. Bernardo and A. Smith, *Bayesian Theory*. New York: Wiley, 2004.

[21] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Proc. Advances in Neural Information Processing Systems*, 2005.

[22] Z. Ghahramani and M. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2000 .

[23] J. Berger, J. Ghosh, and N. Mukhopadhyay, "Approximation and consistency of Bayes factors as model dimension grows," *J. Statist. Plan. Infer.*, vol. 112, pp. 241–258, 2003.

[24] S. Press and K. Shigemasu, "A note on choosing the number of factors," *Commun. Statist. Theory Methods*, vol. 28, pp. 1653–1670, 1999.

[25] S. Lee and X. Song, "Bayesian selection on the number of factors in a factor analysis model," *Behaviormetrika*, vol. 29, pp. 23–39, 2002.

[26] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statistics*, vol. 1, pp. 209–230, 1973.

[27] M. Tipping and C. Bishop, "Mixture of principal component analyzers," *Neural Comput.*, vol. 11, pp. 443–482, 1999.

[28] A. Eftekhari and M. B. Wakin, "Performance bounds for compressive sensing with a mixture of factor analyzers," School of Mines, Tech. Rep. 2010.

[29] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, pp. 3488–3497, 2009.

[30] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization with MCMC," in *Proc. Advances in Neural Information Processing Systems*, 2008.

[31] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Advances in Neural Information Processing Systems*, 2008.

[32] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," *IEEE Trans. Inform. Theory*, vol. 12, pp. 1338–1351, 2008.

[33] N. Hjort, C. Holmes, P. Muller, and S. Walker, *Bayesian Nonparametrics*. Cambridge, MA: Cambridge Univ. Press, 2010.

[34] J. F. C. Kingman, "Completely random measures," *Pacific J. Math.*, vol. 21, no. 1, pp. 59–78, 1967.

[35] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Proc. Int. Workshop Artificial Intelligence and Statistics*, 2007, vol. 11. .

[36] N. L. Hjort, "Nonparametric Bayes estimators based on beta processes in models for life history data," *Ann. Stat.*, vol. 18, no. 3, pp. 1259–1294, 1990.

[37] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. Int. Conf. Machine Learning (ICML)*, 2009.

[38] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for sparse image representations," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2009 .

[39] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky, "Sharing features among dynamical systems with beta processes," in *Proc. Advances in Neural Information Processing (NIPS) 22*. Cambridge, MA: MIT Press, 2010.

[40] M. D. Escobar, "Estimating normal means with a Dirichlet process prior," *J. Amer. Stat. Assoc.*, vol. 89, pp. 268–277, 1994.

[41] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *J. Comput. Graph. Stat.*, vol. 9, pp. 249–265, 2000.

[42] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priros," *J. Amer. Stat. Assoc. Theory Methods*, vol. 96, no. 453, pp. 161–173, 2001.

[43] S. G. Walker, "Sampling the Dirichlet mixture model with slices," *Commun. Stat. Simulat. Comput.*, vol. 36, p. 45, 2007.

[44] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," *Bayesian Nonparametrics: Principles and Practice*. Cambridge: U.K.: Cambridge Univ. Press, 2010.

[45] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2008.

[46] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Machine Learning (ICML)*, 2009.

[47] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, 2006.

[48] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, pp. 21–30, 2008.

[49] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, vol. 346, pp. 589–592, 2008.

[50] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Processing*, vol. 17, 2008.

[51] N. Lawrence and R. Urtasun, "Non-linear matrix factorization with Gaussian processes," in *Proc. Int. Conf. Machine Learning*, 2009.

[52] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, pp. 2053–2080, 2009.

[53] E. J. Candès and Y. Plan, "Matrix completion with noise," in *Proc. IEEE*, 2010.

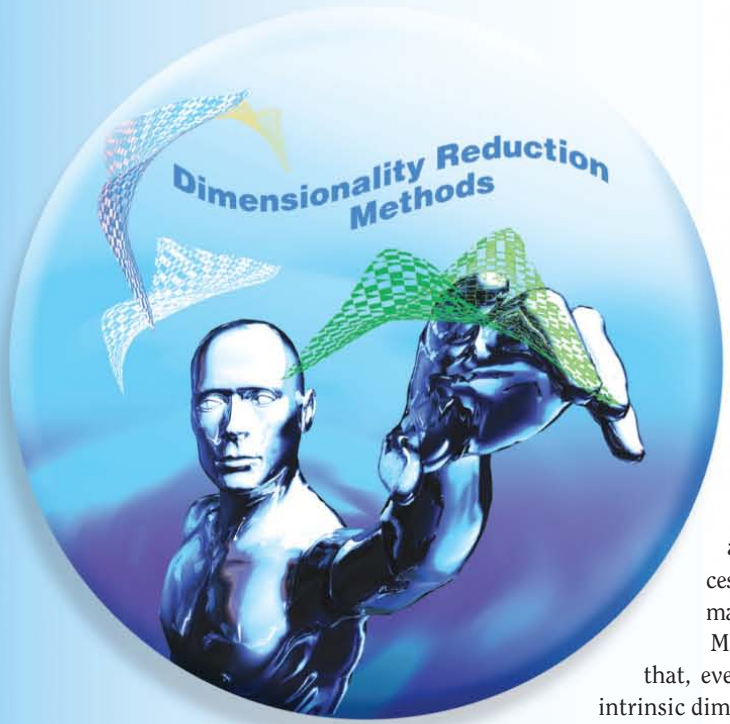
[54] S. Goldwater, T. Griffiths, and M. Johnson, "Interpolating between types and tokens by estimating power-law generators," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2006.

[55] Y. Teh and D. Gorur, "Indian buffet processes with power-law behavior," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2009. SP

René Vidal

Subspace Clustering

[Applications in motion segmentation and face clustering]



© DIGITAL STOCK & LUSPHIX

The past few years have witnessed an explosion in the availability of data from multiple sources and modalities. For example, millions of cameras have been installed in buildings, streets, airports, and cities around the world. This has generated extraordinary advances on how to acquire, compress, store, transmit, and process massive amounts of complex high-dimensional data.

Many of these advances have relied on the observation that, even though these data sets are high dimensional, their intrinsic dimension is often much smaller than the dimension of the ambient space. In computer vision, for example, the number of pixels in an image can be rather large, yet most computer vision models use only a few parameters to describe the appearance, geometry, and dynamics of a scene. This has motivated the development of a number of techniques for finding a low-dimensional representation of a high-dimensional data set. Conventional techniques, such as principal component analysis (PCA), assume that the data are drawn from a single low-dimensional subspace of a high-dimensional space. Such approaches have found widespread applications in many fields, e.g., pattern recognition, data compression, image processing, and bioinformatics.

In practice, however, the data points could be drawn from multiple subspaces, and the membership of the data points to the subspaces might be unknown. For instance, a video sequence could contain several moving objects, and different subspaces might be needed to describe the motion of different objects in the scene. Therefore, there is a need to simultaneously cluster the data into multiple subspaces and find a low-dimensional subspace fitting each group of points. This problem, known as subspace clustering, has found numerous applications in computer vision (e.g., image segmentation [1], motion segmentation [2], and face clustering [3]), image processing (e.g., image representation and compression [4]), and systems theory (e.g., hybrid system identification [5]).

Digital Object Identifier 10.1109/MSP.2010.939739
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

A number of approaches to subspace clustering have been proposed in the past two decades. A review of methods from the data-mining community can be found in [6]. This article will present methods from the machine learning and computer vision communities, including algebraic methods [7]–[10], iterative methods [11]–[15], statistical methods [16]–[20], and spectral clustering-based methods [7], [21]–[27]. We review these methods, discuss their advantages and disadvantages, and evaluate their performance on the motion segmentation and face-clustering problems.

A NUMBER OF APPROACHES TO SUBSPACE CLUSTERING HAVE BEEN PROPOSED IN THE PAST TWO DECADES.

THE SUBSPACE CLUSTERING PROBLEM

Consider the problem of modeling a collection of data points with a union of subspaces, as illustrated in Figure 1. Specifically, let $\{\boldsymbol{x}_j \in \mathbb{R}^D\}_{j=1}^N$ be a given set of points drawn from an unknown union of $n \geq 1$ linear or affine subspaces $\{S_i\}_{i=1}^n$ of unknown dimensions $d_i = \dim(S_i)$, $0 < d_i < D$, $i = 1, \ldots, n$. The subspaces can be described as

$$S_i = \{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} = \boldsymbol{\mu}_i + U_i \boldsymbol{y}\}, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^D$ is an arbitrary point in subspace S_i that can be chosen as $\boldsymbol{\mu}_i = \boldsymbol{0}$ for linear subspaces, $U_i \in \mathbb{R}^{D \times d_i}$ is a basis for subspace S_i, and $\boldsymbol{y} \in \mathbb{R}^{d_i}$ is a low-dimensional representation for point $\boldsymbol{x}$. The goal of subspace clustering is to find the number of subspaces n, their dimensions $\{d_i\}_{i=1}^n$, the subspace bases $\{U_i\}_{i=1}^n$, the points $\{\boldsymbol{\mu}_i\}_{i=1}^n$, and the segmentation of the points according to the subspaces.

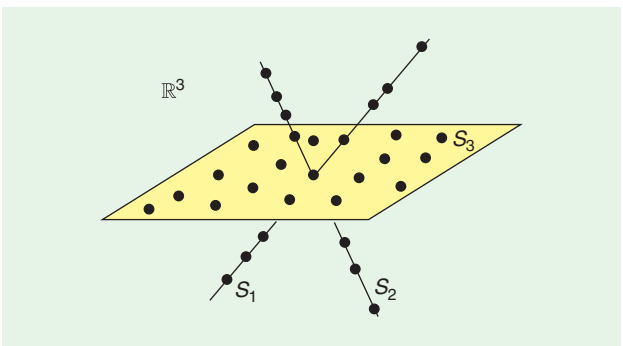
When the number of subspaces is equal to one, this problem reduces to finding a vector $\boldsymbol{\mu} \in \mathbb{R}^D$, a basis $U \in \mathbb{R}^{D \times d}$, a low-dimensional representation $Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N] \in \mathbb{R}^{d \times N}$, and the dimension d. This problem is known as PCA [28]. (The problem of matrix factorization dates back to the work of Beltrami [29] and Jordan [30]. In the context of stochastic signal processing, PCA is also known as Karhunen-Loeve transform [31]. In the applied statistics literature, PCA is also known as Eckart-Young decomposition [32].) PCA can be solved in a remarkably simple way: $\boldsymbol{\mu} = (1/N)\sum_{j=1}^N \boldsymbol{x}_j$ is the mean of the data points (U, Y) can be obtained from the rank-d singular value decomposition (SVD) of the (mean-subtracted) data matrix $X = [\boldsymbol{x}_1 - \boldsymbol{\mu}, \boldsymbol{x}_2 - \boldsymbol{\mu}, \ldots, \boldsymbol{x}_N - \boldsymbol{\mu}] \in \mathbb{R}^{D \times N}$ as

$$U = \mathcal{U} \quad \text{and} \quad Y = \Sigma \mathcal{V}^\top, \quad \text{where} \quad X = \mathcal{U} \Sigma \mathcal{V}^\top, \tag{2}$$

and d can be obtained as $d = \text{rank}(X)$ with noise-free data or using model-selection techniques when the data are noisy [28].

When $n > 1$, the subspace clustering problem becomes significantly more difficult due to a number of challenges.

- First, there is a strong coupling between data segmentation and model estimation. Specifically, if the segmentation of the data is known, one could easily fit a single subspace to each group of points using standard PCA. Conversely, if the subspace parameters were known, one could easily find the data points that best fit each subspace. In practice, neither the segmentation of the data nor the subspace parameters are known, and one needs to solve both problems simultaneously.
- Second, the distribution of the data inside the subspaces is generally unknown. If the data within each subspace are distributed around a cluster center and the cluster centers for different subspaces are far apart, the subspace clustering problem reduces to the simpler and well-studied central clustering problem. However, if the distribution of the data points in the subspaces is arbitrary, the subspace clustering problem cannot be solved by central clustering techniques. In addition, the problem becomes more difficult when many points lie close to the intersection of two or more subspaces.
- Third, the position and orientation of the subspaces relative to each other can be arbitrary. As we will show later, when the subspaces are disjoint or independent, the subspace clustering problem can be solved more easily. However, when the subspaces are dependent, the subspace clustering problem becomes much harder. (n linear subspaces are disjoint if every two subspaces intersect only at the origin. n linear subspaces are independent if the dimension of their sum is equal to the sum of their dimensions. Independent subspaces are disjoint, but the converse is not always true. n affine subspaces are disjoint, independent, if so are the corresponding linear subspaces in homogeneous coordinates.)
- The fourth challenge is that the data can be corrupted by noise, missing entries, and outliers. Although robust estimation techniques for handling such nuisances have been developed for the case of a single subspace, the case of multiple subspaces is not well understood.
- The fifth challenge is model selection. In classical PCA, the only parameter is subspace dimension, which can be found by searching for the subspace of the smallest dimension



[FIG1] A set of sample points in $\mathbb{R}^3$ drawn from a union of three subspaces: two lines and a plane.

that fits the data with a given accuracy. In the case of multiple subspaces, one can fit the data with N different subspaces of dimension one, i.e., one subspace per data point, or with a single subspace of dimension D. Obviously, neither solution is satisfactory. The challenge is to find a model-selection criteria that favors a small number of subspaces of small dimensions.

In what follows, we present a number of subspace clustering algorithms and show how they try to address these challenges.

SUBSPACE CLUSTERING ALGORITHMS

ALGEBRAIC ALGORITHMS

We first review two algebraic algorithms for clustering noise-free data drawn from multiple linear subspaces, i.e., $\boldsymbol{\mu}_i = 0$.

The first algorithm is based on linear algebra, specifically matrix factorization, and is provably correct for independent subspaces. The second one is based on polynomial algebra and is provably correct for both dependent and independent subspaces.

Although these algorithms are designed for linear subspaces, in the case of noiseless data, they can also be applied to affine subspaces by using homogeneous coordinates, thus interpreting an affine subspace of dimension d in $\mathbb{R}^D$ as a linear subspace of dimension $d+1$ in $\mathbb{R}^{D+1}$. (The homogeneous coordinates of $\boldsymbol{x} \in \mathbb{R}^D$ are given by $[\boldsymbol{x}^\top 1]^\top \in \mathbb{R}^{D+1}$.)

Also, while these algorithms operate under the assumption of noise-free data, they provide great insights into the geometry and algebra of the subspace clustering problem. Moreover, they can be extended to handle moderate amounts of noise.

MATRIX FACTORIZATION-BASED ALGORITHMS

These algorithms obtain the segmentation of the data from a low-rank factorization of the data matri X. Hence, they are a natural extension of PCA from one to multiple independent linear subspaces.

Specifically, let $X_i \in \mathbb{R}^{D \times N_i}$ be the matrix containing the N_i points in subspace i. The columns of the data matrix can be sorted according to the n subspaces as $[X_1, X_2, \ldots, X_n] = X\Gamma$, where $\Gamma \in \mathbb{R}^{N \times N}$ is an unknown permutation matrix. Because each matrix X_i is of rank d_i, it can be factorized as

$$X_i = U_i Y_i \quad i = 1, \ldots, n, \tag{3}$$

where $U_i \in \mathbb{R}^{D \times d_i}$ is an orthogonal basis for subspace i and $Y_i \in \mathbb{R}^{d_i \times N_i}$ is the low-dimensional representation of the points with respect to U_i. Therefore, if the subspaces are independent, then $r \triangleq \text{rank}(X) = \sum_{i=1}^{n} d_i \leq \min\{D, N\}$ and

$$X\Gamma = [U_1, U_2, \cdots, U_n] \begin{bmatrix} Y_1 & & & \\ & Y_2 & & \\ & & \ddots & \\ & & & Y_n \end{bmatrix} \triangleq UY, \tag{4}$$

where $U \in \mathbb{R}^{D \times r}$ and $Y \in \mathbb{R}^{r \times N}$. The subspace clustering problem is then equivalent to finding a permutation matrix Γ, such that $X\Gamma$ admits a rank-r factorization into a matrix U and a block diagonal matrix Y. This idea is the basis for the algorithms of Boult and Brown [7], Costeira and Kanade [8], and Gear [9], which compute Γ from the SVD of X [7], [8] or from the row echelon canonical form of X [9].

Specifically, the Costeira and Kanade algorithm proceeds as follows. Let $X = \mathcal{U}\Sigma\mathcal{V}^\top$ be the rank-r SVD of the data matrix, i.e., $\mathcal{U} \in \mathbb{R}^{D \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, and $\mathcal{V} \in \mathbb{R}^{N \times r}$. Also, let

$$Q = \mathcal{V}\mathcal{V}^\top \in \mathbb{R}^{N \times N}. \tag{5}$$

As shown in [2] and [33], the matrix Q is such that

$$Q_{jk} = 0 \text{ if points } j \text{ and } k \text{ are in different subspaces.} \tag{6}$$

In the absence of noise, (6) can be used to obtain the segmentation of the data by applying spectral clustering to the eigenvectors of Q [7] (see the "Spectral Clustering-Based Methods" section) or by sorting and thresholding the entries of Q [8], [34]. For instance, [8] obtains the segmentation by maximizing the sum of the squared entries of Q in different groups, while [34] finds the groups by thresholding a subset of the rows of Q. However, as noted in [33] and [35], this thresholding process is very sensitive to noise. Also, the construction of Q requires knowledge of the rank of X, and using the wrong rank can lead to very poor results [9].

Wu et al. [35] use an agglomerative process to reduce the effect of noise. The entries of Q are first thresholded to obtain an initial oversegmentation of the data. A subspace is then fit to each group G_i, and two groups are merged when the distance between their subspaces is below a threshold. A similar approach is followed by Kanatani et al. [33], [36], except that the geometric Akaike information criterion [37] is used to decide when to merge the two groups.

Although these approaches indeed reduce the effect of noise, in practice, they are not effective because the equation $Q_{jk} = 0$ holds only when the subspaces are independent. In the case of dependent subspaces, one can use the subset of the columns of $\mathcal{V}$ that do not span the intersections of the subspaces. Unfortunately, we do not know which columns to choose a priori. Zelnik-Manor and Irani [38] propose to use the top columns of $\mathcal{V}$ to define Q. However, this heuristic is not provably correct. Another issue with factorization-based algorithms is that, with a few exceptions, they do not provide a method for computing the number of subspaces, n, and their dimensions, $\{d_i\}_{i=1}^n$. The first exception is when n is known. In this case, d_i can be computed from each group after the segmentation has been obtained. The second exception is for independent subspaces of equal dimension d. In this case $\text{rank}(X) = nd$, hence we may determine n when d is known or vice versa.

GENERALIZED PCA

Generalized PCA (GPCA; see [10] and [39]) is an algebraic-geometric method for clustering data lying in (not necessarily independent) linear subspaces. The main idea behind GPCA is that one can fit a union of n subspaces with a set of polynomials of degree n, whose derivatives at a point give a vector normal to the subspace containing that point. The segmentation of the

data is then obtained by grouping these normal vectors using several possible techniques.

The first step of GPCA, which is not strictly needed, is to project the data points onto a subspace of $\mathbb{R}^D$ of dimension $r = d_{\max} + 1$, where $d_{\max} = \max\{d_1, \ldots, d_n\}$. (The value of r is determined using model-selection techniques when the subspace dimensions are unknown.) The rationale behind this step is as follows. Since the maximum dimension of each subspace is $d_{\max}$, a projection onto a generic subspace of $\mathbb{R}^D$ of dimension $d_{\max} + 1$ preserves the number and dimensions of the subspaces with probabilty one. As a by-product, the subspace clustering problem is reduced to clustering subspaces of dimension at most $d_{\max}$ in $\mathbb{R}^{d_{\max}+1}$. As we shall see, this step is very important to reduce the computational complexity of the GPCA algorithm. With an abuse of notation, we will denote the original and projected subspaces as S_i, and the original and projected data matrix as

GENERALIZED PCA IS AN ALGEBRAIC-GEOMETRIC METHOD FOR CLUSTERING DATA LYING IN (NOT NECESSARILY INDEPENDENT) LINEAR SUBSPACES.

$$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N} \text{ or } \mathbb{R}^{r \times N}. \tag{7}$$

The second step is to fit a homogeneous polynomial of degree n to the (projected) data. The rationale behind this step is as follows. Imagine, for instance, that the data came from the union of two planes in $\mathbb{R}^3$, each one with normal vector $\boldsymbol{b}_i \in \mathbb{R}^3$. The union of the two planes can be represented as a set of points, such that $p(\boldsymbol{x}) = (\boldsymbol{b}_1^\top \boldsymbol{x})(\boldsymbol{b}_2^\top \boldsymbol{x}) = 0$. This equation is nothing but the equation of a conic of the form

$$c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_1 x_3 + c_4 x_2^2 + c_5 x_2 x_3 + c_6 x_3^2 = 0. \tag{8}$$

Imagine now that the data came from the plane $\boldsymbol{b}^\top \boldsymbol{x} = 0$ or the line $\boldsymbol{b}_1^\top \boldsymbol{x} = \boldsymbol{b}_2^\top \boldsymbol{x} = 0$. The union of the plane and the line is the set of points, such that $p_1(\boldsymbol{x}) = (\boldsymbol{b}^\top \boldsymbol{x})(\boldsymbol{b}_1^\top \boldsymbol{x}) = 0$ and $p_2(\boldsymbol{x}) = (\boldsymbol{b}^\top \boldsymbol{x})(\boldsymbol{b}_2^\top \boldsymbol{x}) = 0$. More generally, data drawn from the union of n subspaces of $\mathbb{R}^r$ can be represented with polynomials of the form $p(\boldsymbol{x}) = (\boldsymbol{b}_1^\top \boldsymbol{x}) \cdots (\boldsymbol{b}_n^\top \boldsymbol{x}) = 0$, where the vector $\boldsymbol{b}_i \in \mathbb{R}^r$ is orthogonal to S_i. Each polynomial is of degree n in $\boldsymbol{x}$ and can be written as $\boldsymbol{c}^\top \nu_n(\boldsymbol{x})$, where $\boldsymbol{c}$ is the vector of coefficients and $\nu_n(\boldsymbol{x})$ is the vector of all monomials of degree n in $\boldsymbol{x}$. There are

$$M_n(r) = \binom{n+r-1}{n}$$

independent monomials; hence, $\boldsymbol{c} \in \mathbb{R}^{M_n(r)}$.

In the case of noiseless data, the vector of coefficients $\boldsymbol{c}$ of each polynomial can be computed from

$$\boldsymbol{c}^\top [\nu_n(\boldsymbol{x}_1), \nu_n(\boldsymbol{x}_2), \cdots, \nu_n(\boldsymbol{x}_N)] \stackrel{\Delta}{=} \boldsymbol{c}^\top V_n = \boldsymbol{0}^\top \tag{9}$$

and the number of polynomials is simply the dimension of the null space of V_n. While in general the relationship between the number of subspaces, n, their dimensions, $\{d_i\}_{i=1}^n$, and the number of polynomials involves the theory of Hilbert functions [40], in the particular case where all the dimensions are equal to d and $r = d + 1$, there is a unique polynomial that fits the data. This fact can be exploited to determine both n and d. For example, given d, n can be computed as

$$n = \min\{i : \text{rank}(V_i) = M_i(r) - 1\}. \tag{10}$$

In the case of data contaminated with small-to-moderate amounts of noise, the polynomial coefficients (9) can be found using least squares—the vectors $\boldsymbol{c}$ are the left singular vectors of V_n corresponding to the smallest singular values. To handle larger amounts of noise in the estimation of the polynomial coefficients, one can resort to techniques from robust statistics [20] or rank minimization [41]. Model-selection techniques can be used to determine the rank of V_n and, hence, the number of polynomials, as shown in [42]. Model-selection techniques can also be used to determine the number of subspaces of equal dimensions in (10), as shown in [10]. However, determining n and $\{d_i\}_{i=1}^n$ for subspaces of different dimensions from noisy data remains a challenge. The reader is referred to [43] for a model-selection criteria called minimum effective dimension, which measures the complexity of fitting n subspaces of dimensions $\{d_i\}_{i=1}^n$ to a given data set within a certain tolerance, and to [40] and [42] for algebraic relationships among n, $\{d_i\}_{i=1}^n$ and the number of polynomials, which can be used for model-selection purposes.

The last step is to compute the normal vectors $\boldsymbol{b}_i$ from the vector of coefficients $\boldsymbol{c}$. This can be done by taking the derivatives of the polynomials at a data point. For example, if $n = 2$, we have $\nabla p(\boldsymbol{x}) = (\boldsymbol{b}_2^\top \boldsymbol{x})\boldsymbol{b}_1 + (\boldsymbol{b}_1^\top \boldsymbol{x})\boldsymbol{b}_2$. Thus, if $\boldsymbol{x}$ belongs to the first subspace, then $\nabla p(\boldsymbol{x}) \sim \boldsymbol{b}_1$. More generally, in the case of n subspaces, we have $p(\boldsymbol{x}) = (\boldsymbol{b}_1^\top \boldsymbol{x}) \cdots (\boldsymbol{b}_n^\top \boldsymbol{x})$ and $\nabla p(\boldsymbol{x}) \sim \boldsymbol{b}_i$ if $\boldsymbol{x} \in S_i$. We can use this result to obtain the set of all normal vectors to S_i from the derivatives of all the polynomials at $\boldsymbol{x} \in S_i$. This gives us a basis for the orthogonal complement of S_i from which we can obtain a basis U_i for S_i. Therefore, if we knew one point per subspace, $\{\boldsymbol{y}_i \in S_i\}_{i=1}^n$, we could compute the n subspace bases $\{U_i\}_{i=1}^n$ from the gradient of the polynomials at $\{\boldsymbol{y}_i\}_{i=1}^n$ and then obtain the segmentation by assigning each point $\{\boldsymbol{x}_j\}_{j=1}^N$ to its closest subspace. A simple method for choosing the points $\{\boldsymbol{y}_i\}_{i=1}^n$ is to select any data point as $\boldsymbol{y}_1$ to obtain the basis U_1 for the first subspace S_1. After removing the points that belong to S_1 from the data set, we can choose any of the remaining data points as $\boldsymbol{y}_2$ to obtain U_2, hence S_2, and then repeat this process until all the subspaces are found. In the "Spectral Clustering-Based Methods" section, we will describe an alternative method based on spectral clustering.

The first advantage of GPCA is that it is an algebraic algorithm; thus, it is computationally cheap when n and d are small. Second, intersections between subspaces are automatically allowed; hence, GPCA can deal with both independent and

dependent subspaces. Third, in the noiseless case, it does not require the number of subspaces or their dimensions to be known beforehand. Specifically, the theory of Hilbert functions may be used to determine n and $\{d_i\}$, as shown in [40].

The first drawback of GPCA is that its complexity increases exponentially with n and $\{d_i\}$. Specifically, each vector $\boldsymbol{c}$ is of dimension $O(M_n(r))$, while there are only $O(r\sum_{i=1}^{n}(r-d_i))$ unknowns in the n sets of normal vectors. Second, the vector $\boldsymbol{c}$ is computed using least squares; thus, the computation of $\boldsymbol{c}$ is sensitive to outliers. Third, the least-squares fit does not take into account nonlinear constraints among the entries of $\boldsymbol{c}$ (recall that $p(\boldsymbol{x})$ must factorize as a product of linear factors). These issues cause the performance of GPCA to deteriorate as n increases. Fourth, the method in [40] to determine n and $\{d_i\}_{i=1}^{n}$ does not handle noisy data. Fifth, while GPCA can be applied to affine subspaces by using homogeneous coordinates, in our experience, this does not work very well when the data are contaminated with noise.

A VERY SIMPLE WAY OF IMPROVING THE PERFORMANCE OF ALGEBRAIC ALGORITHMS IN THE CASE OF NOISY DATA IS TO USE ITERATIVE REFINEMENT.

ITERATIVE METHODS

A very simple way of improving the performance of algebraic algorithms in the case of noisy data is to use iterative refinement. Intuitively, given an initial segmentation, we can fit a subspace to each group using classical PCA. Then, given a PCA model for each subspace, we can assign each data point to its closest subspace. By iterating these two steps, we can obtain a refined estimate of the subspaces and segmentation. This is the basic idea behind the K-planes [11] algorithm, which generalizes the K-means algorithm [44] from data distributed around multiple cluster centers to data drawn from multiple hyperplanes. The K-subspaces algorithm [12], [13] further generalizes K-planes from multiple hyperplanes to multiple affine subspaces of any dimensions and proceeds as follows. Let $w_{ij}=1$ if point j belongs to subspace i and $w_{ij}=0$ otherwise. Referring back to (1), assume that the number of subspaces n and the subspace dimensions $\{d_i\}_{i=1}^{n}$ are known. Our goal is to find the points $\{\boldsymbol{\mu}_i\in\mathbb{R}^D\}_{i=1}^{n}$, the subspace bases $\{U_i\in\mathbb{R}^{D\times d_i}\}_{i=1}^{n}$, the low-dimensional representations $\{Y_i\in\mathbb{R}^{d_i\times N_i}\}_{i=1}^{n}$, and the segmentation of the data $\{w_{ij}\}_{i=1,\ldots,n}^{j=1,\ldots,N}$. We can do so by minimizing the sum of the squared distances from each data point to its own subspace

$$\min_{\{\mu_i\},\{U_i\},\{y_i\},\{w_{ij}\}} \quad \sum_{i=1}^{n}\sum_{j=1}^{N} w_{ij}\,\|\,\boldsymbol{x}_j-\boldsymbol{\mu}_i-U_i\boldsymbol{y}_j\,\|^2$$
$$\text{subject to} \quad w_{ij}\in\{0,1\} \text{ and } \sum_{i=1}^{n} w_{ij}=1. \tag{11}$$

Given $\{\boldsymbol{\mu}_i\}$, $\{U_i\}$, and $\{\boldsymbol{y}_j\}$, the optimal value for w_{ij} is

$$w_{ij}=\begin{cases}1 & \text{if } i=\arg\min\limits_{k=1,\ldots,n}\|\boldsymbol{x}_j-\boldsymbol{\mu}_k-U_k\boldsymbol{y}_j\|^2\\ 0 & \text{else}\end{cases}. \tag{12}$$

Given $\{w_{ij}\}$, the cost function in (11) decouples as the sum of n cost functions, one per subspace. Since each cost function is identical to that minimized by standard PCA, the optimal values for $\boldsymbol{\mu}_i$, U_i, and $\boldsymbol{y}_j$ are obtained by applying PCA to each group of points. The K-subspaces algorithm then proceeds by alternating between assigning points to subspaces and reestimating the subspaces. Since the number of possible assignments of points to subspaces is finite, the algorithm is guaranteed to converge to a local minimum in a finite number of iterations.

The main advantage of K-subspaces is its simplicity since it alternates between assigning points to subspaces and estimating the subspaces via PCA. Another advantage is that it can handle both linear and affine subspaces explicitly. The third advantage is that it converges to a local optimum in a finite number of iterations. However, K-subspaces suffers from a number of drawbacks. First, its convergence to the global optimum depends on a good initialization. If a random initialization is used, several restarts are often needed to find the global optimum. In practice, one may use any of the algorithms described in this article to reduce the number of restarts needed. We refer the reader to [22] and [45] for two additional initialization methods. Second, K-subspaces is sensitive to outliers, partly due to the use of the ℓ_2-norm. This issue can be addressed using a robust norm, such as the ℓ_1-norm, as done by the median K-flat algorithm [15]. However, this results in a more complex algorithm, which requires solving a robust PCA problem at each iteration. Alternatively, one can resort to nonlinear minimization techniques, which are only guaranteed to converge to a local minimum. Third, K-subspaces requires n and $\{d_i\}_{i=1}^{n}$ to be known beforehand. One possible avenue to be explored is to use the model-selection criteria for mixtures of subspaces proposed in [43]. We refer the reader to [45] and [46] for a more detailed analysis of some of the aforementioned issues.

STATISTICAL METHODS

The approaches described so far seek to cluster the data according to multiple subspaces using mostly algebraic and geometric properties of a union of subspaces. While these approaches can handle noise in the data, they do not make explicit assumptions about the distribution of data inside the subspaces or about the distribution of noise. Therefore, the estimates they provide are not optimal, e.g., in a maximum likelihood (ML) sense. This issue can be addressed by defining a proper generative model for the data, as described next.

MIXTURE OF PROBABILISTIC PCA

Resorting back to the geometric PCA model (1), probabilistic PCA (PPCA) [47] assumes that the data within a subspace S is generated as

$$\boldsymbol{x}=\boldsymbol{\mu}+U\boldsymbol{y}+\boldsymbol{\epsilon}, \tag{13}$$

where $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are independent zero-mean Gaussian random vectors with covariance matrices I and $\sigma^2 I$, respectively. Therefore,

$\boldsymbol{x}$ is also Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma = UU^\top + \sigma^2 I$. It can be shown that the ML estimate of $\boldsymbol{\mu}$ is the mean of the data, and ML estimates of U and σ can be obtained from the SVD of the data matrix X.

PPCA can be naturally extended to a generative model for a union of subspaces $\cup_{i=1}^n S_i$ by using a mixture of PPCA (MPPCA) model [16]. Let $G(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ be the probability density function of a D-dimensional Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ. MPPCA uses a mixture of Gaussians model

$$p(\boldsymbol{x}) = \sum_{i=1}^{n} \pi_i G(\boldsymbol{x}; \boldsymbol{\mu}_i, U_i U_i^\top + \sigma_i^2 I), \qquad \sum_{i=1}^{n} \pi_i = 1, \tag{14}$$

where the parameter π_i, called the mixing proportion, represents the a priori probability of drawing a point from subspace S_i. The ML estimates of the parameters of this mixture model can be found using expectation maximization (EM) [48]. EM is an iterative procedure that alternates between data segmentation and model estimation. Specifically, given initial values $(\tilde{\boldsymbol{\mu}}_i, \tilde{U}_i, \tilde{\sigma}_i, \tilde{\pi}_i)$ for the model parameters, in the E-step, the probability that $\boldsymbol{x}_j$ belongs to subspace i is estimated as

$$\tilde{p}_{ij} = \frac{G(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \tilde{U}_i \tilde{U}_i^\top + \tilde{\sigma}_i^2 I)\tilde{\pi}_i}{p(\boldsymbol{x}_j)}, \tag{15}$$

and in the M-step, the $\tilde{p}_{ij}$s are used to recompute the subspace parameters using PPCA. Specifically, π_i and $\boldsymbol{\mu}_i$ are updated as

$$\tilde{\pi}_i = \frac{1}{N}\sum_{j=1}^{N} \tilde{p}_{ij} \text{ and } \tilde{\boldsymbol{\mu}}_i = \frac{1}{N\tilde{\pi}_i}\sum_{j=1}^{N} \tilde{p}_{ij}\boldsymbol{x}_j, \tag{16}$$

and σ_i and U_i are updated from the SVD of

$$\tilde{\Sigma}_i = \frac{1}{N\tilde{\pi}_i}\sum_{j=1}^{N} \tilde{p}_{ij}(\boldsymbol{x}_j - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \tilde{\boldsymbol{\mu}}_i)^\top. \tag{17}$$

These two steps are iterated until convergence to a local maxima of the log-likelihood. Notice that MPPCA can be seen as a probabilistic version of K-subspaces that uses soft assignments $p_{ij} \in [0, 1]$ rather than hard assignments $w_{ij} = \{0, 1\}$.

As in the case of K-subspaces, the main advantage of MPPCA is that it is a simple and intuitive method, where each iteration can be computed in closed form by using PPCA. Moreover, the MPPCA model is applicable to both linear and affine subspaces and can be extended to accommodate outliers [49] and missing entries in the data points [50]. However, an important drawback of MPPCA is that the number and dimensions of the subspaces need to be known beforehand. One way to address this issue is to put a prior on these parameters, as shown in [51]. A second drawback is that MPPCA is not optimal when the data inside each subspace or the noise is not Gaussian.

AN IMPORTANT DRAWBACK OF MPPCA IS THAT THE NUMBER AND DIMENSIONS OF THE SUBSPACES NEED TO BE KNOWN BEFOREHAND.

A third drawback is that MPPCA often converges to a local maximum; hence, a good initialization is critical. The initialization problem can be addressed by using any of the methods described earlier for K-subspaces. For example, the multistage learning (MSL) algorithm [17] uses the factorization method of [8] followed by the agglomerative refinement steps of [33] and [36] for initialization.

AGGLOMERATIVE LOSSY COMPRESSION

The agglomerative lossy compression (ALC) algorithm [18] assumes that the data are drawn from a mixture of degenerate Gaussians. However, unlike MPPCA, ALC does not aim to obtain an ML estimate of the parameters of the mixture model. Instead, it looks for the segmentation of the data that minimizes the coding length needed to fit the points with a mixture of degenerate Gaussians up to a given distortion.

Specifically, the number of bits needed to optimally code N independent identically distributed (i.i.d.) samples from a zero-mean D-dimensional Gaussian, i.e., $X \in \mathbb{R}^{D \times N}$, up to a distortion δ can be approximated as $[(N + D)/2] \log_2 \det(I + (D/\delta^2 N)XX^\top)$. Thus, the total number of bits for coding a mixture of Gaussians can be approximated as

$$\sum_{i=1}^{n} \frac{N_i + D}{2} \log_2 \det\left(I + \frac{D}{\delta^2 N_i} X_i X_i^\top\right) - N_i \log_2\left(\frac{N_i}{N}\right), \tag{18}$$

where $X_i \in \mathbb{R}^{D \times N_i}$ is the data from subspace i, and the last term is the number of bits needed to code (losslessly) the membership of the N samples to the n groups.

The minimization of (18) over all possible segmentations of the data is, in general, an intractable problem. ALC deals with this issue by using an agglomerative clustering method. Initially, each data point is considered as a separate group. At each iteration, two groups are merged if doing so results in the greatest decrease of the coding length. The algorithm terminates when the coding length cannot be further decreased. Similar agglomerative techniques have been used [52], [53], though with a different criterion for merging subspaces.

ALC can naturally handle noise and outliers in the data. Specifically, it is shown in [18] that outliers tend to cluster either as a single group or as small separate groups depending on the dimension of the ambient space. Also, in principle, ALC does not need to know the number of subspaces and their dimensions. In practice, however, the number of subspaces is directly related to the parameter δ. When δ is chosen to be very large, all the points could be merged into a single group. Conversely, when δ is very small, each point could end up as a separate group. Since δ is related to the variance of the noise, one can use statistics on the data to determine δ (see [22] and [33] for possible methods). When the number of subspaces is known, one can run ALC for several values of δ, discard the values of δ that give the wrong number of subspaces, and choose the δ that results in the segmentation with the smallest

coding length. This typically increases the computational complexity of the method. Another disadvantage of ALC, perhaps the major one, is that there is no theoretical proof for the optimality of the agglomerative procedure.

RANDOM SAMPLE CONSENSUS

Random sample consensus (RANSAC) [54] is a statistical method for fitting a model to a cloud of points corrupted with outliers in a statistically robust way. More specifically, if d is the minimum number of points required to fit a model to the data, RANSAC randomly samples d points from the data, fits a model to these d points, computes the residual of each data point to this model, and chooses the points whose residual is below a threshold as the inliers. The procedure is then repeated for d sample points, until the number of inliers is above a threshold, or enough samples have been drawn. The outputs of the algorithm are the parameters of the model and the labeling of inliers and outliers.

RANDOM SAMPLE CONSENSUS IS A STATISTICAL METHOD FOR FITTING A MODEL TO A CLOUD OF POINTS CORRUPTED WITH OUTLIERS IN A STATISTICALLY ROBUST WAY.

In the case of clustering subspaces of equal dimension d, the model to be fit by RANSAC is a subspace of dimension d. Since there are multiple subspaces, RANSAC proceeds in a greedy fashion by fitting one subspace at a time as follows:

1) Apply RANSAC to the original data set and recover a basis for the first subspace along with the set of inliers. All points in other subspaces are considered as outliers to the first subspace.
2) Remove the inliers from the current data set and repeat Step 1 to find the second subspace and so on until all the subspaces are recovered.
3) For each set of inliers, use PCA to find an optimal basis for each subspace. Segment the data into multiple subspaces by assigning each point to its closest subspace.

The main advantage of RANSAC is its ability to handle outliers explicitly. Also, notice that RANSAC does not require the subspaces to be independent, because it computes one subspace at a time. Moreover, RANSAC does not need to know the number of subspaces beforehand. In practice, however, determining the number of subspaces depends on the user-defined thresholds. An important drawback of RANSAC is that its performance deteriorates quickly as the number of subspaces n increases, because the probability of drawing d inliers reduces exponentially with the number of subspaces. Therefore, the number of trials needed to find d points in the same subspace grows exponentially with the number and dimension of the subspaces. As shown in [55], this issue can be addressed by introducing a nonuniform prior in the sampling strategy so that points in the same subspace are more likely to be chosen than points in different subspaces. Another critical drawback of RANSAC is that it requires the dimension of the subspaces to be known and equal. In the case of subspaces of different dimensions, one could start from the largest to the smallest dimension or vice versa. However, those procedures suffer from a number of issues, as discussed in [20].

SPECTRAL CLUSTERING-BASED METHODS

Spectral clustering algorithms (see [56] for a review) are a very popular technique for clustering high-dimensional data. These algorithms construct an affinity matrix $A \in \mathbb{R}^{N \times N}$, whose (j,k)th entry measures the similarity between points j and k. Ideally, $A_{jk} = 1$ if points j and k are in the same group and $A_{jk} = 0$ if points j and k are in a different group. A typical measure of similarity is $A_{jk} = \exp(-\text{dist}_{jk}^2)$, where dist_{jk} is some distance between points j and k. Given A, the segmentation of the data is obtained by applying the K-means algorithm to the eigenvectors of a matrix $L \in \mathbb{R}^{N \times N}$ formed from A. Specifically, if $\{U_j\}_{j=1}^N$ are the eigenvectors of L, then $n \ll N$ eigenvectors are chosen and stacked into a matrix $V \in \mathbb{R}^{N \times n}$. The K-means algorithm is then applied to the rows of V. Typical choices for L are the affinity matrix itself, $L = A$, the Laplacian, $L = \text{diag}(A\mathbf{1}) - A$, where $\mathbf{1}$ is the vector of all ones, and the normalized Laplacian, $L_{\text{sym}} = I - \text{diag}(A\mathbf{1})^{-1/2} A \text{diag}(A\mathbf{1})^{-1/2}$. Typical choices for the eigenvectors are the top n eigenvectors of the affinity or the bottom n eigenvectors of the (normalized) Laplacian, where n is the number of groups.

One of the main challenges in applying spectral clustering to the subspace clustering problem is to define a good affinity matrix. This is because two points could be very close to each other but lie in different subspaces (e.g., near the intersection of two subspaces). Conversely, two points could be far from each other but lie in the same subspace. As a consequence, one cannot use the typical distance-based affinity.

In what follows, we review some of the methods for building a pairwise affinity for points lying in multiple subspaces. The first two methods (factorization and GPCA) are designed for linear subspaces, though they can be applied to affine subspaces by modifying the affinity or using homogeneous coordinates. The remaining methods can handle either linear or affine subspaces.

FACTORIZATION-BASED AFFINITY

Interestingly, one of the first subspace clustering algorithms is based on both matrix factorization and spectral clustering. Specifically, the algorithm of Boult and Brown [7] obtains the segmentation of the data from the eigenvectors of the matrix $Q = \mathcal{V}\mathcal{V}^\top$ in (6). Since these eigenvectors are the singular vectors of X, the segmentation is obtained by clustering the rows of $\mathcal{V}$. However, recall that the affinity $A_{jk} = Q_{jk}$ has a number of issues. First, it is not necessarily the case that $A_{jk} \approx 1$ when points i and j are in the same subspace. Second, the equation $Q_{jk} = 0$ is sensitive to noise, and it is valid only for independent subspaces.

GPCA-BASED AFFINITY

As noticed in [2] and [57], the GPCA algorithm can also be used to define an affinity between two points. Specifically, recall that an estimate $\hat{S}_j$ of the subspace passing through the point x_j can

be obtained from the derivatives of the polynomials $p(\boldsymbol{x})$ at $\boldsymbol{x}_j$. Let θ_{jk}^m be the mth principal angle between $\hat{S}_j$ and $\hat{S}_k$, for $j,k = 1,\ldots,N$. One can use these angles to define an affinity as

$$A_{jk} = \prod_{m=1}^{\min(d_j,d_k)} \cos^2(\theta_{jk}^m). \tag{19}$$

Notice that this affinity is applicable only to linear subspaces, because it only captures the similarity between the subspace bases. To see this, notice that when two affine subspaces are parallel to each other, all their principal angles are equal to zero; hence, A_{jk} is equal to one not only for points j and k in the same subspace, but also for points j and k in two different subspaces. Therefore, in the case of data drawn from affine subspaces, A_{jk} needs to be modified to also incorporate an appropriate distance between points j and k. We will discuss ways to do this in the next paragraph. Given a pairwise affinity, GPCA finds the segmentation of the data by applying spectral clustering to the normalized Laplacian.

LOCAL SUBSPACE AFFINITY AND SPECTRAL LOCAL BEST-FIT FLATS

The local subspace affinity (LSA) [21] and spectral local best-fit flats (SLBF) [22] algorithms are based on the observation that a point and its nearest neighbors (NNs) often belong to the same subspace. Therefore, we can fit an affine subspace $\hat{S}_j$ to each point j and its d-NNs using, e.g., PCA. In practice, we can choose $K \geq d$ NNs; hence, d does not need to be known exactly: we only need an upper bound.

Then, if two points j and k lie in the same subspace S_i, their locally estimated subspaces $\hat{S}_j$ and $\hat{S}_k$ should be the same, while if the two points lie in different subspaces, $\hat{S}_j$ and $\hat{S}_k$ should be different. Therefore, we can use a distance between $\hat{S}_j$ and $\hat{S}_k$ to define an affinity between the two points.

The first (optional) step of the LSA and SLBF algorithms is to project the data points onto a subspace of dimension $r = \text{rank}(X)$ using the SVD of X. With noisy data, the value of r is determined using model-selection techniques. In the case of data drawn from linear subspaces, the LSA algorithm projects the resulting points in $\mathbb{R}^r$ onto the hypersphere $\mathbb{S}^{r-1}$.

The second step is to compute the K-NNs of each point j and to fit a local affine subspace $\hat{S}_j$ to the point and its neighbors. LSA assumes that K is specified by the user and finds K-NN using the angle between the two data points or as a metric. PCA is then used to fit the local subspace $\hat{S}_j$. The subspace dimension d_j is then determined using model-selection techniques. SLBF determines both the number of neighbors K_j and the subspace $\hat{S}_j$ for each point j automatically. It does so by searching for the smallest value of K_j that minimizes a certain fitting error.

The third step of LSA is to compute an affinity matrix as

$$A_{jk} = \exp\left[-\sum_{m=1}^{\min(d_j,d_k)} \sin^2(\theta_{jk}^m)\right], \tag{20}$$

where θ_{jk}^m is the mth principal angle between the estimated subspaces $\hat{S}_j$ and $\hat{S}_k$. As in the case of the GPCA-based affinity in (19), the affinity in (20) is applicable only to linear subspaces. SLBF addresses this issue by using the affinity

$$A_{jk} = \exp(-\hat{d}_{jk}/2\sigma_j^2) + \exp(-\hat{d}_{jk}/2\sigma_k^2), \tag{21}$$

where σ_j measures how well point j and its K_j-NNs are fit by $\hat{S}_j$, $\hat{d}_{jk} = \sqrt{\text{dist}(\boldsymbol{x}_j,\hat{S}_k)\text{dist}(\boldsymbol{x}_k,\hat{S}_j)}$, and $\text{dist}(\boldsymbol{x},S)$ is the Euclidean distance from point $\boldsymbol{x}$ to subspace S. Notice that this affinity uses the distance from points to subspaces; thus, it is applicable to both linear and affine subspaces. Given a pairwise affinity, LSA and SLBF find the segmentation of the data by applying spectral clustering to the normalized Laplacian.

The LSA and SLBF algorithms have two main advantages when compared with GPCA. First, outliers are likely to be rejected, because they are far from all the points, and so they are not considered as neighbors of the inliers. Second, LSA requires only $O(nd_{\max})$ data points, while GPCA needs $O(M_n(d_{\max}+1))$. On the other hand, LSA has two main drawbacks. First, the neighbors of a point could belong to a different subspace. This is more likely to happen near the intersection of two subspaces. Second, the selected neighbors may not span the underlying subspace. Thus, K needs to be small enough so that only points in the same subspace are chosen and large enough so that the neighbors span the local subspace. SLBF resolves these issues by choosing the size of the neighborhood automatically.

Notice also that both GPCA and LSA are based on a linear projection followed by spectral clustering. While in principle both algorithms can use any linear projection, GPCA prefers to use the smallest possible dimension $r = d_{\max}+1$, so as to reduce the computational complexity. On the other hand, LSA uses a slightly larger dimension $r = \text{rank}(X) \leq \sum d_i$. This is because if the dimension of the projection is too small [less than $\text{rank}(X)$], the projected subspaces become dependent. While in theory, LSA can handle both independent and dependent subspaces, the projection increases the dimension of the intersection of two subspaces; hence, many of the data points could be projected close to the intersection. As a consequence, LSA does not perform as well with dependent subspaces, as the experiments will show. Another major difference between LSA and GPCA is that LSA fits a subspace locally around each projected point, while GPCA uses the gradient of a polynomial that is globally fit to the projected data.

LOCALLY LINEAR MANIFOLD CLUSTERING

The locally linear manifold clustering (LLMC) algorithm [23] is also based on fitting a local subspace to a point and its K-NNs. Specifically, every point j is written as an affine combination of all other points $k \neq j$. The coefficients w_{jk} are found in closed form by minimizing the cost

$$\sum_{j=1}^{N} \| \boldsymbol{x}_j - \sum_{k\neq j} w_{jk}\boldsymbol{x}_k \|^2 = \| (I-W)X^\top \|_F^2, \tag{22}$$

where $\|X\|_F^2 = \sum X_{ij}^2$ is the Frobenius norm of X, subject to $\sum_{k\neq j} w_{jk} = 1$ and $w_{jk} = 0$ if $\boldsymbol{x}_k$ is not a K-NN of $\boldsymbol{x}_j$. Then, the affinity matrix and the matrix L are built as

$$A = W + W^\top - W^\top W \quad \text{and} \quad L = (I - W)^\top (I - W). \tag{23}$$

It is shown in [23] that when every point and its K-NNs are always in the same subspace, then there are vectors $\boldsymbol{v}$ in the null space of L with the property that $v_j = v_k$ when points j and k are in the same subspace. However, these vectors are not the only vectors in the null space of L; hence, spectral clustering is not directly applicable. In this case, a procedure for properly selecting linear combinations of the eigenvectors of L is needed, as discussed in [23].

SSC RELIES ON THE FACT THAT A POINT IN A LINEAR OR AFFINE SUBSPACE OF DIMENSION *D* CAN ALWAYS BE WRITTEN AS A LINEAR OF AFFINE COMBINATION OF *D* OR *D* + 1 DATA IN THE SAME SUBSPACE.

A first advantage of LLMC is its robustness to outliers. This is because, as in the case of LSA and SLBF, outliers are often far from the inliers, hence it is unlikely that they are chosen as neighbors of the inliers. Another important advantage of LLMC is that it is also applicable to nonlinear subspaces, while all the other methods discussed so far are only applicable to linear (or affine) subspaces. However, LLMC suffers from the same disadvantage of LSA, namely, that it has problems with points near the intersections, because it is not always the case that a point and its K-NNs are in the same subspace. Also, properly choosing the number of NNs is a challenge. These issues could be resolved by choosing the neighborhood automatically, as done by SLBF. Finally, even though, in theory, LLMC can handle both dependent and independent subspaces, in practice, it does not perform as well with dependent subspaces for the same reasons as for LSA.

SPARSE SUBSPACE CLUSTERING

Sparse subspace clustering (SSC) [24], [25] is also based on the idea of writing a data point as a linear or affine combination of neighboring data points. However, while LSA, SLBF, and LLMC use the angular or Euclidean distance between two points to choose the K-NNs, SSC uses the principle of sparsity to choose any of the remaining data points ($N - 1 \gg K$) as a possible neighbor. Specifically, SSC relies on the fact that a point in a linear or affine subspace of dimension d can always be written as a linear or affine combination of d or $d + 1$ data points in the same subspace. Therefore, if we write a data point $\boldsymbol{x}_j \in S_i$ as a linear or affine combination of all other $N - 1$ data points $\{\boldsymbol{x}_k\}_{k\neq j}$ drawn from $\cup_{i=1}^n S_i$ with $d_i = \dim(S_i)$, then a sparse linear or affine combination can be obtained by choosing d_i or $d_i + 1$ nonzero coefficients corresponding to points from S_i. This sparse linear or affine combination $\boldsymbol{x}_j = \sum_{k\neq j} w_{jk}\boldsymbol{x}_k$ can be found by minimizing the number of nonzero coefficients w_{jk}, subject to $\sum w_{jk} = 1$ in the case of affine subspaces. Since this problem is combinatorial, the SSC algorithm solves the following simpler ℓ_1 optimization problem instead

$$\min_{\{w_{jk}\}} \sum_{k\neq j} |w_{jk}| \text{ s.t. } \boldsymbol{x}_j = \sum_{k\neq j} w_{jk}\boldsymbol{x}_k \left(\text{and} \sum_{k\neq j} w_{jk} = 1\right). \tag{24}$$

It is shown in [24] and [25] that, when the subspaces are either independent or disjoint, the solution to the optimization problem in (24) is such that $w_{jk} = 0$ only if points j and k are in different subspaces. In other words, a *sparse representation* is obtained, where each point is written as a linear or affine combination of a few points in its own subspace.

In the case of data contaminated by noise, the SSC algorithm does not attempt to write a data point as an exact linear or affine combination of other points. Instead, a penalty in the ℓ_2-norm of the error is added to the ℓ_1 norm. Specifically, the sparse coefficients are found by solving the problem

$$\min_{\{w_{jk}\}} \sum_{k\neq j} |w_{jk}| + \lambda \| \boldsymbol{x}_j - \sum_{k\neq j} w_{jk}\boldsymbol{x}_k \|^2 \left(\text{s.t.} \sum_{k\neq j} w_{jk} = 1\right), \tag{25}$$

where $\lambda > 0$ is a parameter. Obviously, different solutions for $\{w_{jk}\}$ will be obtained for different choices of the parameter λ. However, we are not interested in the specific values of w_{jk}: all what matters is that, for each point j, the top nonzero coefficients come from points in the same subspace.

In the case of data contaminated with outliers, the SSC algorithm assumes that $\boldsymbol{x}_j = \sum_{k\neq j} w_{jk}\boldsymbol{x}_k + \boldsymbol{e}_j$, where the vector of outliers $\boldsymbol{e}_j$ is also sparse. The sparse coefficients and the outliers are found by solving the problem

$$\min_{\{w_{jk}\},\{\boldsymbol{e}_j\}} \sum_{k\neq j} |w_{jk}| + \| \boldsymbol{e}_j \|_1 + \lambda \| \boldsymbol{x}_j - \sum_{k\neq j} w_{jk}\boldsymbol{x}_k - \boldsymbol{e}_j \|^2 \tag{26}$$

subject to $\sum_{k\neq j} w_{jk} = 1$ in the case of affine subspaces.

Given a sparse representation for each data point, the pairwise affinity matrix is defined as

$$A = |W| + |W^\top|. \tag{27}$$

The segmentation is then obtained by applying spectral clustering to the Laplacian.

The SSC algorithm presents several advantages with respect to all the algorithms discussed so far. With respect to factorization-based methods, the affinity in (27) is very robust to noise. This is because the solution changes continuously with the amount of noise. Specifically, with moderate amounts of noise, the top nonzero coefficients will still correspond to points in the same subspace. With larger amounts of noise, some of the nonzero coefficients will come from other subspaces. These mistakes can be handled by spectral clustering, which is also robust to noise (see [56]). With respect to GPCA, SSC is more robust to outliers because, as in the case of LSA, SLBF, and LLMC, it is

very unlikely that a point in a subspace will write itself as a linear combination of a point that is very far from all of the subspaces. Also, the computational complexity of SSC does not grow exponentially with the number of subspaces and their dimensions. Nonetheless, it requires solving N optimization problems in $O(N)$ variables, as per (24), (25), or (26), hence, it can be slow. With respect to LSA and LLMC, the great advantage of SSC is that the neighbors of a point are automatically chosen, without having to specify the value of K. Moreover, the dimension of the individual subspaces does not need to be known beforehand and can be estimated from the number of nonzero coefficients. More importantly, the SSC algorithm is provably correct for independent [24] and disjoint [25] subspaces; hence, its performance is not affected when the NNs of a point (in the traditional sense) do not come from the same subspace containing that point. Another advantage of SSC over GPCA is that it does not require the data to be projected onto a low-dimensional subspace. A possible disadvantage of SSC is that it is provably correct only in the case of independent or disjoint subspaces. However, the experiments will show that SSC performs well also for dependent subspaces.

LOW-RANK REPRESENTATION

This algorithm [26] is very similar to SSC, except that it aims to find a low-rank representation (LRR) instead of a sparse representation. Before explaining the connection further, let us first rewrite the SSC algorithm in a matrix form. Specifically, recall that SSC requires solving N optimization problems in $O(N)$ variables, as per (24). These N optimization problems can be written as a single optimization problem in $O(N^2)$ variables as

$$\min_{\{w_{jk}\}} \sum_{j=1}^{N} \sum_{k \neq j} |w_{jk}| \text{ s.t. } \boldsymbol{x}_j = \sum_{k \neq j} w_{jk} \boldsymbol{x}_k \left(\text{and } \sum_{k \neq j} w_{jk} = 1 \right). \quad (28)$$

This problem can be rewritten in matrix form as

$$\min_{W} \| W \|_1 \text{ s.t. } X = XW^\top, \text{diag}(W) = \mathbf{0} \text{ (and } W\mathbf{1} = \mathbf{1}). \quad (29)$$

Similarly, in the case of data contaminated with noise, the N optimization problems in (25) can be written as

$$\min_{W,E} \| W \|_1 + \lambda \| E \|_F^2 \text{ s.t. } X = XW^\top + E, \text{diag}(W) = \mathbf{0} \text{ (and } W\mathbf{1} = \mathbf{1}). \quad (30)$$

The LRR algorithm aims to minimize rank(W) instead of $\| W \|_1$. Since this rank-minimization problem is nondeterministic polynomial (NP) time hard, the authors replace the rank of W by its nuclear norm $\| W \|_* = \sum \sigma_i(W)$, where $\sigma_i(W)$ is the ith singular value of W. In the case of noise-free data drawn from linear (affine) subspaces, this leads to the following (convex) optimization problem

$$\min_{W} \| W \|_* \quad \text{s.t.} \quad X = XW^\top \text{ (and } W\mathbf{1} = \mathbf{1}). \quad (31)$$

It can be shown that when the data are noise free and drawn from independent linear subspaces, the optimal solution to (31) is given by the matrix Q of the Costeira and Kanade algorithm, as defined in (5). Recall from (6) that this matrix is such that $Q_{jk} = 0$ when points j and k are in different subspaces, hence it can be used to build an affinity matrix.

In the case of data contaminated with noise or outliers, the LRR algorithm solves the (convex) optimization problem

$$\min_{W} \| W \|_* + \lambda \| E \|_{2,1} \text{ s.t. } X = XW^\top + E \text{ (and } W\mathbf{1} = \mathbf{1}), \quad (32)$$

where $\| E \|_{2,1} = \sum_{k=1}^{N} \sqrt{\sum_{j=1}^{N} |E_{jk}|^2}$ is the $\ell_{2,1}$ norm of the matrix of errors E. Notice that this problem is analogous to (30), except that the ℓ_1 and Frobenius norms are replaced by the nuclear and $\ell_{2,1}$ norms, respectively.

The LRR algorithm proceeds by solving the optimization problem in (32) using an augmented Lagrangian method. The optimal W is used to define an affinity matrix A as in (27). The segmentation of the data is then obtained by applying spectral clustering to the normalized Laplacian.

One of the main attractions of LRR is that it provides a theoretical justification for the Costeira and Kanade algorithm. A second advantage is that, similarly to SSC, the optimization problem is convex. One drawback of LLR is that it is provably correct only in the case of noiseless data drawn from independent subspaces. Another drawback is that the optimization problem involves $O(N^2)$ variables.

SPECTRAL CURVATURE CLUSTERING

The methods discussed so far choose a data point plus d NNs (LSA, SLBF, LLMC) or d sparse neighbors (SSC), fit an affine subspace to each of these N groups of $d+1$ points, and build a pairwise affinity by comparing these subspaces. In contrast, multiway clustering techniques such as [27], [58], and [59] are based on the observation that a minimum of $d+1$ points are needed to define an affine subspace of dimension d (d for linear subspaces). Therefore, they consider $d+2$ points, build a measure of how likely these points are to belong to the same subspace, and use this measure to construct an affinity between two points.

Specifically, let $X_{d+2} = \{\boldsymbol{x}_{j_\ell}\}_{j=1}^{d+2}$ be $d+2$ randomly chosen data points. One possible affinity is the volume of the $(d+1)$-simplex formed by these points, vol(X_{d+2}), which is equal to zero if the points are in the same subspace. However, one issue with this affinity is that it is not invariant to data transformations, e.g., scaling of the $d+2$ points. The spectral curvature clustering (SCC) algorithm [27] is based on the concept of polar curvature, which is also zero when the points are in the same subspace. The multiway affinity $\mathcal{A}_{j_1, j_2, \ldots, j_{d+2}}$ is defined as

$$\exp\left(-\frac{1}{2\sigma^2} \text{diam}^2(X_{d+2}) \sum_{\ell=1}^{d+2} \frac{(d+1)!^2 \text{vol}^2(X_{d+2})}{\prod_{\substack{1 \leq m \leq d+2 \\ m \neq \ell}} \| \boldsymbol{x}_{j_m} - \boldsymbol{x}_{j_\ell} \|^2} \right) \quad (33)$$

if $j_1, j_2, \ldots, j_{d+2}$ are distinct and zero otherwise, where diam(X_{d+2}) is the diameter of X_{d+2}. Notice that this affinity is

invariant to scaling of the data points while the volume is not. A pairwise affinity matrix is then defined as

$$A_{jk} = \sum_{j_2,\ldots,j_{d+1}\in\{1,\ldots,N\}} \mathcal{A}_{j,j_2,\ldots,j_{d+2}} \mathcal{A}_{k,j_2,\ldots,j_{d+2}}. \quad (34)$$

This requires computing $O(N^{d+2})$ entries of $\mathcal{A}$ and summing over $O(N^{d+1})$ elements of $\mathcal{A}$. Therefore, the computational complexity of SCC grows exponentially with the dimension of the subspaces. A practical implementation of SCC uses a fixed number c of $(d+1)$-tuples ($c \ll N^{d+1}$) for each point to build the affinity A. A choice of $c \approx c_0 n^{d+2}$ is suggested in [27], which is much smaller but still exponential in d. In practice, the method appears to be not too sensitive to the choice of c but more importantly to how the $d+1$ points are chosen. Reference [27] argues that a uniform sampling strategy does not perform well, because many samples could contain subspaces of different dimensions. To avoid this, two stages of sampling are performed. The first stage is used to obtain an initial clustering of the data. In the second stage, the initial clusters are used to guide the sampling and thus obtain a better affinity. Given A, the segmentation is obtained by applying spectral clustering to the normalized Laplacian. One difference of SCC with respect to the previous methods is that SCC uses a procedure for initializing K-means based on maximizing the variance among all possible combinations of K rows of V.

One advantage of SCC (and also of SSC) over LSA, SLBF, and LLMC is that it uses points from the entire data set to define the affinity between two points, while LSA, SLBF, and LLMC restrict themselves to K-NNs. This ultimately results in better affinities because it is less likely that they are built using points from different subspaces. One advantage of SCC over factorization-based methods and GPCA is that it can handle noisy data drawn from both linear and affine subspaces. Another advantage of SCC over GPCA is that it does not require the data to be projected onto a low-dimensional subspace. Also, when the data are sampled from a mixture of distributions concentrated around multiple affine subspaces, SCC performs well with overwhelming probability, as shown in [60]. Finally, SCC can be extended to nonlinear manifolds by using kernel methods [61]. However, the main drawbacks of SCC are that it requires sampling of the affinities to reduce the computational complexity and that it requires the subspaces to be of known and equal dimension d. In practice, the algorithm can still be applied to subspaces of different dimensions by choosing $d = d_{\max}$, but the effect of this choice on the definition of spectral curvature remains unknown.

ONE ADVANTAGE OF SCC OVER FACTORIZATION-BASED METHODS AND GPCA IS THAT IT CAN HANDLE NOISY DATA DRAWN FROM BOTH LINEAR AND AFFINE SUBSPACES.

APPLICATIONS IN COMPUTER VISION

MOTION SEGMENTATION FROM FEATURE POINT TRAJECTORIES

Motion segmentation refers to the problem of separating a video sequence into multiple spatiotemporal regions corresponding to different rigid-body motions. Most existing motion segmentation algorithms proceed by first extracting a set of point trajectories from the video using standard tracking methods. As a consequence, the motion segmentation problem is reduced to clustering these point trajectories according to the different rigid-body motions in the scene.

The mathematical models needed to describe the motion of the point trajectories vary depending on the type of camera projection model. Under the affine model, all the trajectories associated with a single rigid motion live in a three-dimensional (3-D) affine subspace. To see this, let $\{x_{fj} \in \mathbb{R}^2\}_{j=1,\ldots,N}^{f=1,\ldots,F}$ denote the two-dimensional (2-D) projections of N 3-D points $\{X_j \in \mathbb{R}^3\}_{j=1}^{N}$ on a rigidly moving object onto F frames of a moving camera. The relationship between the tracked feature points and their corresponding 3-D coordinates is

$$x_{fj} = A_f \begin{bmatrix} X_j \\ 1 \end{bmatrix}, \quad (35)$$

where $A_f \in \mathbb{R}^{2\times 4}$ is the affine motion matrix at frame f. If we form a matrix containing all the F tracked feature points corresponding to a point on the object in a column, we get

$$\begin{bmatrix} x_{11} \cdots x_{1N} \\ \vdots \quad \vdots \\ x_{F1} \cdots x_{FN} \end{bmatrix}_{2F\times N} = \begin{bmatrix} A_1 \\ \vdots \\ A_F \end{bmatrix}_{2F\times 4} \begin{bmatrix} X_1 \cdots X_N \\ 1 \cdots 1 \end{bmatrix}_{4\times N}. \quad (36)$$

We can briefly write this as $W = MS^\top$, where $M \in \mathbb{R}^{2F\times 4}$ is called the motion matrix and $S \in \mathbb{R}^{N\times 4}$ is called the structure matrix. Since $\text{rank}(M) \le 4$ and $\text{rank}(S) \le 4$ we get

$$\text{rank}(W) = \text{rank}(MS^\top) \le \min(\text{rank}(M), \text{rank}(S)) \le 4. \quad (37)$$

Moreover, since the last row of $S^\top$ is one, the feature point trajectories of a single rigid-body motion lie in an affine subspace of $\mathbb{R}^{2F}$ of dimension at most three.

Assume now that we are given N trajectories of n rigidly moving objects. Then, these trajectories lie in a union of n affine subspaces in $\mathbb{R}^{2F}$. The 3-D motion segmentation problem is the task of clustering these N trajectories into n different groups such that the trajectories in the same group represent a single rigid-body motion. Therefore, the motion segmentation problem reduces to clustering a collection of point trajectories according to multiple affine subspaces.

In what follows, we evaluate a number of subspace clustering algorithms on the Hopkins155 motion segmentation database, which is available online at http://www.vision.jhu.edu/data/hopkins155 [57]. The database consists of 155 sequences of two and three motions, which can be divided into three main categories: checkerboard, traffic, and articulated sequences. The checkerboard sequences contain multiple objects moving independently and arbitrarily in 3-D space, hence the motion trajectories are

expected to lie in independent affine subspaces of dimension three. The traffic sequences contain cars moving independently on the ground plane, hence the motion trajectories are expected to lie in independent affine subspaces of dimension two. The articulated sequences contain motions of people, cranes, etc., where object parts do not move independently, and so the motion subspaces are expected to be dependent. For each sequence, the trajectories are extracted automatically with a tracker and outliers are manually removed. Therefore, the trajectories are corrupted by noise but do not have missing entries or outliers. Figure 2 shows sample images from videos in the database with the feature points superimposed.

MOTION SEGMENTATION REFERS TO THE PROBLEM OF SEPARATING A VIDEO SEQUENCE INTO MULTIPLE SPATIOTEMPORAL REGIONS CORRESPONDING TO DIFFERENT RIGID-BODY MOTIONS.

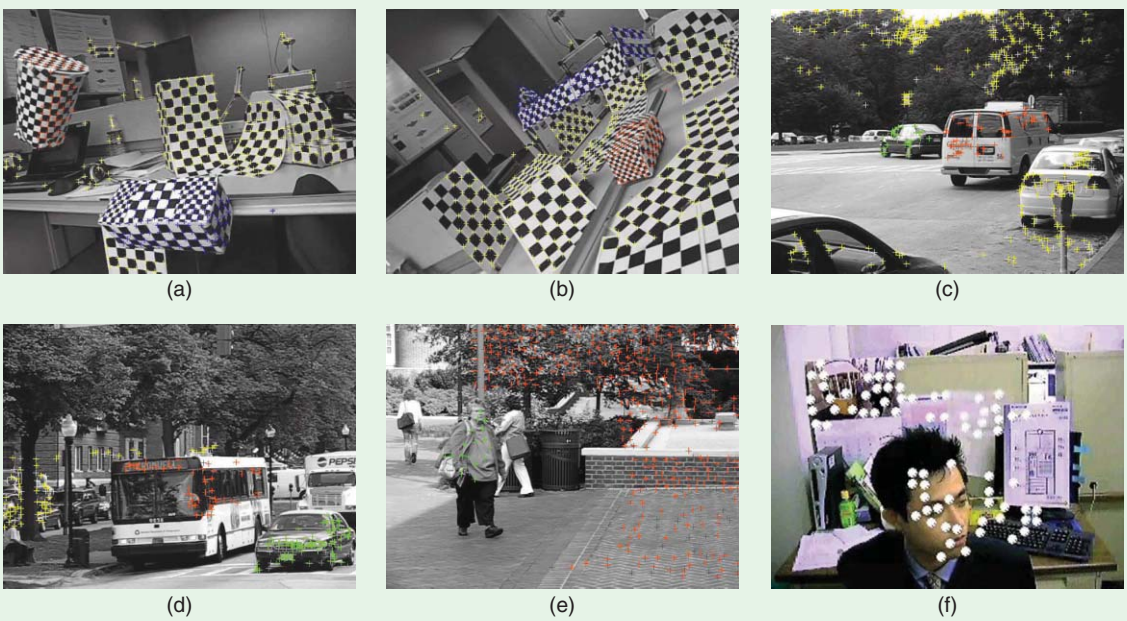
To make our results comparable to those in the existing literature, for each method we apply the same preprocessing steps described in their respective articles. Specifically, we project the trajectories onto a subspace of dimension $r \leq 2F$ using either PCA (GPCA, RANSAC, LLMC, LSA, ALC, and SCC) or a random projection matrix (SSC) whose entries are drawn from a Bernoulli (SSC-B) or normal (SSC-N) distribution. Historically, there have been two choices for the dimension of the projection: $r = 5$ and $r = 4n$. These choices are motivated by algebraic methods, which model 3-D affine subspaces as four-dimensional (4-D) linear subspaces. Since $d_{\max} = 4$, GPCA chooses $r = d_{\max} + 1 = 5$, while factorization methods use the fact that for independent subspaces $r = \text{rank}(X) = 4n$. In our experiments, we use $r = 5$ for GPCA and RANSAC and $r = 4n$ for GPCA, LLMC, LSA, SCC, and SSC. For ALC, r is chosen automatically for each sequence as the minimum r such that $r \geq 8 \log(2F/r)$. We will refer to this choice as the sparsity preserving (sp) projection. We refer the reader to [62] for more recent work that determines the dimension of the projection automatically. Also, for the algorithms that make use of K-means, either a single restart is used when initialized by another algorithm (LLMC, SCC), or ten restarts are used when initialized at random (GPCA, LLMC, LSA). SSC uses 20 restarts.

For each algorithm and each sequence, we record the classification error defined as

$$\text{Classification error} = \frac{\text{number of misclassified points}}{\text{total number of points}} \times 100\%. \quad (38)$$

Table 1 reports the average and median misclassification errors, and Figure 3 shows the percentage of sequences for which the classification error is below a given percentage of misclassification. More detailed statistics with the classification errors and computation times of each algorithm on each of the 155 sequences can be found at http://www.vision.jhu.edu/data/hopkins155/.

By looking at the results, we can draw the following conclusions about the performance of the algorithms tested.



[FIG2] Sample images from some sequences in the database with tracked points superimposed: (a) 1R2RCT_B, (b) 2T3RCRT, (c) cars3, (d) cars10, (e) people2, and (f) kanatani3.

[TABLE 1] CLASSIFICATION ERRORS OF SEVERAL SUBSPACE CLUSTERING ALGORITHMS ON THE HOPKINS 155 MOTION SEGMENTATION DATABASE.

	TWO MOTIONS								THREE MOTIONS							
	CHECK. (78)		TRAFFIC (31)		ARTICUL. (11)		ALL (120)		CHECK. (26)		TRAFFIC (7)		ARTICUL. (2)		ALL (35)		ALL (155)	
	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN
GPCA (4,5)	6.09	1.03	1.41	0.00	2.88	0.00	4.59	0.38	31.95	32.93	19.83	19.55	16.85	16.85	28.66	28.26	10.34	2.54
GPCA (4N-1,4N)	4.78	0.51	1.63	0.00	6.18	3.20	4.10	0.44	36.99	36.26	39.68	40.92	29.62	29.62	37.11	37.18	11.55	1.36
RANSAC (4,5)	6.52	1.75	2.55	0.21	7.25	2.64	5.56	1.18	25.78	26.00	12.83	11.45	21.38	21.38	22.94	22.03	9.76	3.21
LSA (4,5)	8.84	3.43	2.15	1.00	4.66	1.28	6.73	1.99	30.37	31.98	27.02	34.01	23.11	23.11	29.28	31.63	11.82	4.00
LSA (4,4N)	2.57	0.27	5.43	1.48	4.10	1.22	3.45	0.59	5.80	1.77	25.07	23.79	7.25	7.25	9.73	2.33	4.94	0.90
LLMC (4,5)	4.85	0.00	1.96	0.00	6.16	1.37	4.22	0.00	9.06	7.09	6.45	0.00	5.26	5.26	8.33	3.19	5.15	0.00
LLMC (4,4N)	3.96	0.23	3.53	0.33	6.48	1.30	4.08	0.24	8.48	5.80	6.04	4.09	9.38	9.38	8.04	4.93	4.97	0.87
LLMC-G (4,5)	4.34	0.00	2.13	0.00	6.16	1.37	3.95	0.00	8.87	7.09	5.62	0.00	5.26	5.26	8.02	3.19	4.87	0.00
LLMC-G (4,4N)	2.83	0.00	3.61	0.00	5.94	1.30	3.32	0.00	8.20	5.26	6.04	4.60	8.32	8.32	7.78	4.93	4.37	0.53
MSL	4.46	0.00	2.23	0.00	7.23	0.00	4.14	0.00	10.38	4.61	1.80	0.00	2.71	2.71	8.23	1.76	5.03	0.00
ALC (4,5)	2.56	0.00	2.83	0.30	6.90	0.89	3.03	0.00	6.78	0.92	4.01	1.35	7.25	7.25	6.26	1.02	3.76	0.26
ALC (4,5P)	1.49	0.27	1.75	1.51	10.70	0.95	2.40	0.43	5.00	0.66	8.86	0.51	21.08	21.08	6.69	0.67	3.37	0.49
SCC (3, 4)	2.99	0.39	1.20	0.32	7.71	3.67	2.96	0.42	7.72	3.21	0.52	0.28	8.90	8.90	6.34	2.36	3.72	
SCC (3, 4N)	1.76	0.01	0.46	0.16	4.06	1.69	1.63	0.06	6.00	2.22	1.78	0.42	5.65	5.65	5.14	1.67	2.42	
SCC (3, 2F)	1.77	0.00	0.63	0.14	4.02	2.13	1.68	0.07	6.23	1.70	1.11	1.40	5.41	5.41	5.16	1.58	2.47	
SCC (4, 5)	2.31	0.25	0.71	0.26	5.05	1.08	2.15	0.27	5.56	2.03	1.01	0.47	8.97	8.97	4.85	2.01	2.76	
SCC (4, 4N)	1.30	0.04	1.07	0.44	3.68	0.67	1.46	0.16	5.68	2.96	2.35	2.07	10.94	10.94	5.31	2.40	2.33	
SCC (4, 2F)	1.31	0.06	1.02	0.26	3.21	0.76	1.41	0.10	6.31	1.97	3.31	3.31	9.58	9.58	5.90	1.99	2.42	
SLBF (3, 2F)	1.59	0.00	0.20	0.00	0.80	0.00	1.16	0.00	4.57	0.94	**0.38**	0.00	2.66	2.66	3.63	0.64	1.66	
SSC-B (4,4N)	**0.83**	0.00	0.23	0.00	1.63	0.00	**0.75**	0.00	4.49	0.54	0.61	0.00	1.60	1.60	3.55	0.25	1.45	0.00
SSC-N (4,4N)	1.12	0.00	**0.02**	0.00	**0.62**	0.00	0.82	0.00	**2.97**	0.27	0.58	0.00	**1.42**	0.00	**2.45**	0.20	**1.24**	0.00

All algorithms use two parameters (d,r), where d is the dimension of the subspaces and r is the dimension of the projection. Affine subspace clustering algorithms treat subspaces as 3-D affine subspaces, i.e., $d = 3$, while linear subspace clustering algorithms treat subspaces as four-dimensional linear subspaces, i.e., $d = 4$. The dimensions of the projections are $r = 5$, $r = 4n$, where n is the number of motions, and $r = 2F$, where F is the number of frames. ALC uses an sp dimension for the projection. All algorithms use PCA to perform the projection, except for SSC that uses a random projection with entries drawn from SSC-B or SSC-N distribution. The results for GPCA correspond to the spectral clustering-based GPCA algorithm. LLMC-G denotes LLMC initialized by the algebraic GPCA algorithm.

GPCA

To avoid using multiple polynomials, we use an implementation of GPCA based on hyperplanes in which the data are interpreted as a subspace of dimension $r - 1$ in $\mathbb{R}^r$, where $r = 5$ or $r = 4n$.

For two motions, GPCA achieves a classification error of 4.59% for $r = 5$ and 4.10% for $r = 4n$. Notice that GPCA is among the most accurate methods for the traffic and articulated sequences, which are sequences with dependent motion subspaces. However, GPCA has higher errors on the checkerboard sequences, which constitute a majority of the database. This result is expected because GPCA is best designed for dependent subspaces. Notice also that increasing r from 5 to $4n$ improves the results for checkerboard sequences but not for the traffic and articulated sequences. This is also expected because the rank of the data matrix should be high for sequences with full-dimensional and independent motions (checkerboard) and low for sequences with degenerate (traffic) and dependent (articulated) motions. This suggests that using model selection to determine a different value of r for each sequence should improve the results.

For three motions, the results are completely different with a segmentation error of 29–37%. This is expected because the number of coefficients fitted by GPCA grows exponentially with the number of motions, while the number of feature points remains of the same order. Furthermore, GPCA uses a least-squares method for fitting the polynomial, which neglects nonlinear constraints among the coefficients. The number of nonlinear constraints neglected also increases with the number of subspaces.

RANSAC

The results for this purely statistical algorithm are similar to what we found for GPCA. In the case of two motions, the results are a bit worse than those of GPCA. In the case of three motions, the results are better than those of GPCA but still quite far from those of the best-performing algorithms. This is expected because, as the number of motions increases, the probability of drawing a set of points from the same group reduces significantly. Another drawback of RANSAC is that its performance varies between two runs on the same data. Our experiments report the average performance by more than 1,000 trials for each sequence.

LSA

When the dimension of the projection is chosen as $r = 5$, this algorithm performs worse than GPCA. This is because the points in different subspaces are closer to each other when $r = 5$, and so a point from

a different subspace is more likely to be chosen as an NN. GPCA, on the other hand, is not affected by points near the intersection of the subspaces. The situation is completely different when $r = 4n$. In this case, LSA clearly outperforms GPCA and RANSAC, achieving an error of 3.45% for two groups and 9.73% for three groups. These errors could be further reduced by using model selection to determine the dimension of each subspace. Another important thing to observe is that LSA performs better on the checkerboard sequences, but has larger errors than GPCA on the traffic and articulated sequences. This confirms that LSA has difficulties with dependent subspaces.

LLMC

The results of this algorithm also represent a clear improvement over GPCA and RANSAC, especially for three motions. The only cases where GPCA outperforms LLMC are for traffic and articulated sequences. This is expected because LLMC is not designed to handle dependent subspaces. Unlike LSA, LLMC is not significantly affected by the choice of r, with a classification error of 5.15% for $r = 5$ and 4.97% for $r = 4n$. Notice also that the performance of LLMC improves when initialized with GPCA to 4.87% for $r = 5$ and 4.37% for $r = 4n$. However, there are a few sequences for which LLMC performs worse than GPCA even when LLMC is initialized by GPCA. This happens for sequences with dependent motions, which are not well handled by LLMC.

MSL

By looking at the average classification error, we can see that MSL, LSA, and LLMC have a similar accuracy. Furthermore, their segmentation results remain consistent when going from two to three motions.

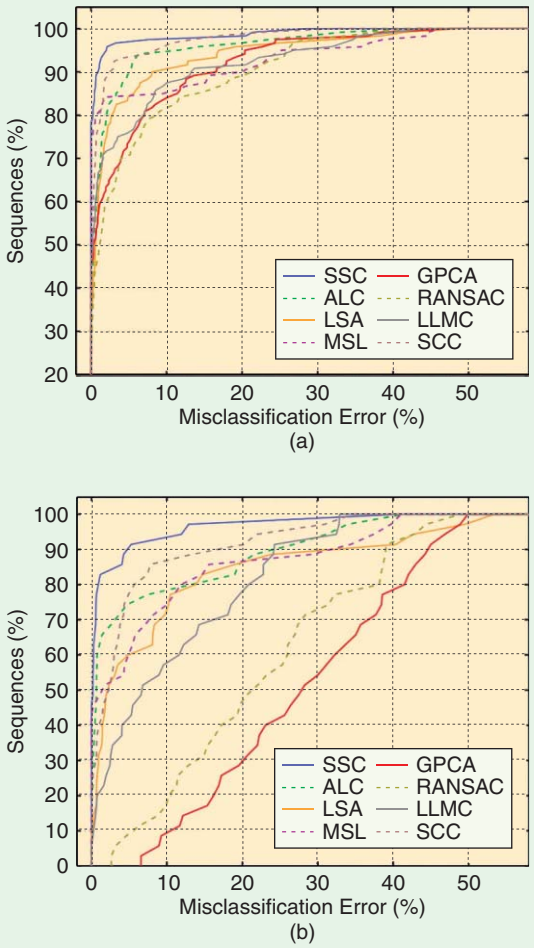
However, sometimes the MSL method gets stuck in a local minimum. This is reflected by high classification errors for some sequences, as can be seen by the long tails in Figure 3.

ALC

This algorithm represents a significant increase in performance with respect to all previous algorithms, especially for the checkerboard sequences, which constitute the majority of the database. However, ALC does not perform very well on the articulated sequences. This is because ALC typically needs the samples from a group to cover the subspace with sufficient density, while many of the articulated scenes have very few feature point trajectories. With regard to the projection dimension, the results indicate that, overall, ALC performs better with an automatic choice of the projection rather than with a fixed choice of $r = 5$. One drawback of ALC is that it needs to be run about 100 times for different choices of the distortion parameter δ to obtain the right number of motions and the best segmentation results.

SCC

This algorithm performs even better than ALC in almost all motion categories. The only exception is for the articulated sequences with three motions. This is because these sequences contain few trajectories for the sampling strategy to operate correctly. Another advantage of SCC with respect to ALC is that it is not very sensitive to the choice of the parameter c (number of sampled subsets), while ALC needs to be run for several choices of the distortion parameter δ. Notice also that the performance of SCC is not significantly affected by the dimension of the projection $r = 5, r = 4n$, or $r = 2F$.



[FIG3] Percentage of sequences for which the classification error is less than or equal to a given percentage of misclassification. The algorithms tested are GPCA (4,5), RANSAC (4,5), LSA (4,4n), LLMC (4,4n), MSL, ALC (4,sp), SCC (4,4n), and SSC-N (4,4n). (a) Two motions. (b) Three motions.

SSC

This algorithm performs extremely well not only for checkerboard sequences, which have independent and fully dimensional motion subspaces, but also for traffic and articulated sequences, which are the bottleneck of almost all existing methods, because they contain degenerate and dependent motion subspaces. This is surprising because the algorithm is provably correct only for independent or disjoint subspaces. Overall, the performance of

[TABLE 2] MEAN PERCENTAGE OF MISCLASSIFICATION ON CLUSTERING YALE FACE B DATA SET.

n	2	3	4	5	6	7	8	9	10
GPCA	0.0	49.5	0.0	26.6	9.9	25.2	28.5	30.6	19.8
SCC	0.0	0.0	0.0	1.1	2.7	2.1	2.2	5.7	6.6
SSC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	4.6
SLBF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.9
ALC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

SSC is not very sensitive to the choice of the projection (Bernoulli versus normal), though SSC-N gives slightly better results. We have also observed that SSC is not sensitive to the dimension of the projection ($r = 5$ versus $r = 4n$ versus $r = 2F$) or the parameter λ.

SLBF

This algorithm performs extremely well for all motion sequences. Its performance is essentially on par with that of SSC. We refer the reader to [22] for additional experiments.

FACE CLUSTERING UNDER VARYING ILLUMINATION

Given a set of images $\{I_j \in \mathbb{R}^D\}_{j=1}^N$ of n different faces taken from the same viewpoint under varying illumination conditions, the face clustering problem consists of clustering the images according to the identity of the person. For a Lambertian object, it has been shown that the set of all images taken under all lighting conditions forms a cone in the image space, which can be well approximated by a low-dimensional subspace [3]. Therefore, the face clustering problem reduces to clustering a set of images according to multiple subspaces.

Table 2 shows the experiments from [22], which evaluate the performance of the GPCA, ALC, SCC, SLBF, and SSC algorithms on the face clustering problem. The experiments are performed on the Yale faces B database, which is available at http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html. This database consists of $10 \times 9 \times 64$ images of ten faces taken under nine different viewpoints and 64 different illumination conditions. Nine subsets containing the images of the frontal views of the following $n = 2, \ldots, 10$ individuals are considered: {5, 8}, {1, 5, 8}, {1, 5, 8, 10}, {1, 4, 5, 8, 10}, {1, 2, 4, 5, 8, 10}, {1, 2, 4, 5, 7, 8, 10}, {1, 2, 4, 5, 7, 8, 9, 10}, {1, 2, 3, 4, 5, 7, 8, 9, 10}, and {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. For computational efficiency, the images are downsampled to 120×160 pixels. Since this number is still large compared with the dimension of the subspaces, PCA is used to project the images onto a subspace of dimension $r = 5$ for GPCA and $r = 20$ for ALC, SCC, SLBF, and SSC. In all cases, the dimension of the subspaces is set to $d = 2$.

By looking at the results, we can draw the following conclusions about the performance of the algorithms tested.

GPCA

This algorithm does not perform very well. This is attributed to the fact that it is very hard to distinguish faces from only five dimensions. While one could have chosen to project the faces to a space of larger dimension, GPCA cannot handle a large number of variables, especially as the number of groups increases.

SCC

This algorithm performs better than GPCA, achieving a perfect classification for $n \leq 4$. However, as n increases from 5 to 10, the classification error ranges from 1.1% to 6.6%.

SSC

This algorithm performs very well, achieving perfect classification for $n \leq 8$ and classification errors of 2.4% and 4.6% for $n = 9$ and $n = 10$, respectively.

SLBF

This algorithm performs very well, slightly better than SSC. It achieves perfect classification for $n \leq 8$ and errors of 1.2% and 0.9% for $n = 9$ and $n = 10$, respectively.

ALC

This algorithm performs extremely well, achieving 100% accuracy in all cases. However, this requires using the algorithm from [22] to set the parameter δ. When multiple values of δ are chosen, the error goes up to 50% for $n = 2$ and stays at 0% in other cases, as reported in [22].

Although these experiments show very promising results, we believe there is still plenty of room for improvement. For example, the face clustering problem is more challenging from nonfrontal faces, thus it would be natural to evaluate the algorithms for nonfrontal faces and see if their performance deteriorates. Also, many of the images in the Yale faces B database contain not only faces but also background, which can facilitate the clustering of the images using the background intensities. Thus, it would be natural to evaluate the algorithms on the cropped images and see if their performance deteriorates. Finally, one could also explore several choices for the subspace dimensions d and for the dimension of the projection D.

CONCLUSIONS AND FUTURE DIRECTIONS

Over the past few decades, significant progress has been made in clustering high-dimensional data sets distributed around a collection of linear and affine subspaces. This article presented a review of such progress, which included a number of existing subspace clustering algorithms together with an experimental evaluation on the motion segmentation and face clustering problems in computer vision.

While earlier algorithms were designed under the assumptions of perfect data and knowledge of the number of subspaces and their dimensions, throughout the years algorithms started to handle noise, outliers, data with missing entries, unknown number of subspaces, and unknown dimensions.

In the case of noiseless data drawn from linear subspaces, the theoretical correctness of existing algorithms is well studied.

Some algorithms are provably correct for independent subspaces, others are provably correct for disjoint subspaces, and others are able to handle an unknown number of subspaces of unknown dimensions in an arbitrary configuration. However, a theoretical analysis of the applicability of many methods to affine subspaces in the noiseless case is still due.

IN OUR VIEW, THE GRAND CHALLENGE FOR THE NEXT DECADE WILL BE TO DEVELOP CLUSTERING ALGORITHMS FOR DATA DRAWN FROM MULTIPLE NONLINEAR MANIFOLDS.

In the case of noisy data, the theoretical correctness of existing algorithms is largely untouched. To the best of our knowledge, the first works in this direction are [45] and [60]. By and large, most existing algorithms assume that the number of subspaces and their dimensions are known. While some algorithms can provide estimates for these quantities, their estimates come with no theoretical guarantees. In our view, the development of theoretically sound algorithms for finding the number of subspaces and their dimension in the presence of noise and outliers is a very important open challenge.

On the other hand, it is important to mention that most existing algorithms operate in a batch fashion. In real-time applications, it is important to cluster the data as it is being collected, which motivates the development of online subspace clustering algorithms. The works of [15] and [63] are two examples in this direction.

Finally, in our view, the grand challenge for the next decade will be to develop clustering algorithms for data drawn from multiple nonlinear manifolds. The works of [64]–[67] have already considered the problem of clustering quadratic, bilinear, and trilinear surfaces using algebraic algorithms designed for noise-free data. The development of methods that are applicable to more general manifolds with corrupted data is still at its infancy.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation (NSF) IIS-0447739 and Office of Naval Research (ONR) N000140510836. The author thanks Prof. Yi Ma, Prof. Richard Hartley, Dr. Alvina Goh, Dr. Shankar Rao, Mr. Ehsan Elhamifar, and Mr. Roberto Tron for their numerous contributions to this work. The author also thanks Prof. Gilad Lerman for organizing a wonderful workshop on multimanifold data analysis at the Institute for Mathematics and its Applications of the University of Minnesota and his students Mr. Guangliang Chen and Mr. Teng Zhang for answering numerous questions about their work. The author also thanks an anonymous reviewer for his/her very insightful comments, which have significantly improved this manuscript.

AUTHOR

René Vidal (rvidal@jhu.edu) received his B.S. degree in electrical engineering (highest honors) from the Pontificia Universidad Católica de Chile in 1997 and his M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley in 2000 and 2003, respectively. He was a research fellow at the National ICT Australia in 2003 and joined Johns Hopkins University in 2004 as a faculty member in the Department of Biomedical Engineering and Center for Imaging Science. He was coeditor of the book *Dynamical Vision* and has coauthored more than 100 articles in biomedical image analysis, computer vision, machine learning, hybrid systems, and robotics. He is a recipient of the 2009 ONR Young Investigator Award, the 2009 Sloan Research Fellowship, the 2005 NFS CAREER Award, and the 2004 Best Paper Award Honorable Mention at the European Conference on Computer Vision. He also received the 2004 Sakrison Memorial Prize for completing an exceptionally documented piece of research, the 2003 Eli Jury Award for outstanding achievement in the area of systems, communications, control, or signal processing, the 2002 Student Continuation Award from National Aeronautics and Space Administration, the 1998 Marcos Orrego Puelma Award from the Institute of Engineers of Chile, and the 1997 Award of the School of Engineering of the Pontificia Universidad Católica de Chile to the best graduating student of the school. He is a Member of the IEEE and the Association for Computing Machinery.

REFERENCES

[1] A. Yang, J. Wright, Y. Ma, and S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 212–225, 2008.

[2] R. Vidal, R. Tron, and R. Hartley, "Multiframe motion segmentation with missing data using power factorization and GPCA," *Int. J. Comput. Vis.*, vol. 79, no. 1, pp. 85–105, 2008.

[3] J. Ho, M. H. Yang, J. Lim, K.C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 11–18.

[4] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multi-scale hybrid linear models for lossy image representation," *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3655–3671, 2006.

[5] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *Proc. Conf. Decision and Control*, 2003, pp. 167–172.

[6] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 90–105, 2004.

[7] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Proc. IEEE Workshop Motion Understanding*, 1991, pp. 179–186.

[8] J. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vis.*, vol. 29, no. 3, pp. 159–179, 1998.

[9] C. W. Gear, "Multibody grouping from motion images," *Int. J. Comput. Vis.*, vol. 29, no. 2, pp. 133–150, 1998.

[10] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 12, pp. 1–15, 2005.

[11] P. S. Bradley and O. L. Mangasarian, "k-plane clustering," *J. Global Optim.*, vol. 16, no. 1, pp. 23–32, 2000.

[12] P. Tseng, "Nearest q-flat to m points," *J. Optim. Theory Applicat.*, vol. 105, no. 1, pp. 249–252, 2000.

[13] P. Agarwal and N. Mustafa, "k-means projective clustering," in *Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems*, 2004, pp. 155–165.

[14] L. Lu and R. Vidal, "Combined central and subspace clustering on computer vision applications," in *Proc. Int. Conf. Machine Learning*, 2006, pp. 593–600.

[15] T. Zhang, A. Szlam, and G. Lerman, "Median k-flats for hybrid linear modeling with many outliers," in *Proc. Workshop Subspace Methods*, 2009, pp. 234–241.

[16] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, 1999.

[17] Y. Sugaya and K. Kanatani, "Geometric structure of degeneracy for multi-body motion segmentation," in *Proc. Workshop Statistical Methods in Video Processing*, 2004, pp. 13–25.

[18] Y. Ma, H. Derksen, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy coding and compression," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 9, pp. 1546–1562, 2007.

[19] S. Rao, R. Tron, Y. Ma, and R. Vidal, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[20] A. Y. Yang, S. Rao, and Y. Ma, "Robust statistical estimation and segmentation of multiple subspaces," in *Proc. Workshop 25 Years of RANSAC*, 2006.

[21] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. European Conf. Computer Vision*, 2006, pp. 94–106.

[22] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1927–1934.

[23] A. Goh and R. Vidal, "Segmenting motions of different types by unsupervised manifold clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[24] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[25] E. Elhamifar and R. Vidal, "Clustering disjoint subspaces via sparse representation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2010, pp. 1926–1929.

[26] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Machine Learning*, 2010.

[27] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. Comput. Vis.*, vol. 81, no. 3, pp. 317–330, 2009.

[28] I. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[29] E. Beltrami, "Sulle funzioni bilineari," *Giornale di Math. Battaglini*, vol. 11, pp. 98–106, 1873.

[30] M. C. Jordan, "Mémoire sur les formes bilinéaires," *J. Math. Pures Appliqués*, vol. 19, pp. 35–54, 1874.

[31] H. Stark and J.W. Woods, *Probability and Random Processes with Applications to Signal Processing*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 2001.

[32] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[33] K. Kanatani, "Motion segmentation by subspace separation and model selection," in *Proc. IEEE Int. Conf. Computer Vision*, 2001, vol. 2, pp. 586–591.

[34] N. Ichimura, "Motion segmentation based on factorization method and discriminant criterion," in *Proc. IEEE Int. Conf. Computer Vision*, 1999, pp. 600–605.

[35] Y. Wu, Z. Zhang, T. S. Huang, and J. Y. Lin, "Multibody grouping via orthogonal subspace decomposition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, vol. 2. pp. 252–257.

[36] K. Kanatani and C. Matsunaga, "Estimating the number of independent motions for multibody motion segmentation," in *Proc. European Conf. Computer Vision*, 2002, pp. 25–31.

[37] K. Kanatani, "Geometric information criterion for model selection," *Int. J. Comput. Vis.*, vol. 26, no. 3, pp. 171–189, 1998.

[38] L. Zelnik-Manor and M. Irani, "On single-sequence and multi-sequence factorizations," *Int. J. Comput. Vis.*, vol. 67, no. 3, pp. 313–326, 2006.

[39] Y. Ma, A. Yang, H. Derksen, and R. Fossum, "Estimation of subspace arrangements with applications in modeling and segmenting mixed data," *SIAM Rev.*, vol. 50, no. 3, pp. 413–458, 2008.

[40] H. Derksen, "Hilbert series of subspace arrangements," *J. Pure Appl. Algebra*, vol. 209, no. 1, pp. 91–98, 2007.

[41] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps, "GPCA with denoising: A moments-based convex approach," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[42] A. Yang, S. Rao, A. Wagner, Y. Ma, and R. Fossum, "Hilbert functions and applications to the estimation of subspace arrangements," in *Proc. IEEE Int. Conf. Computer Vision*, 2005.

[43] K. Huang, Y. Ma, and R. Vidal, "Minimum effective dimension for mixtures of subspaces: A robust GPCA algorithm and its applications," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 631–638.

[44] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[45] A. Aldroubi and K. Zaringhalam, "Nonlinear least squares in $\mathbb{R}^N$," *Acta Appl. Math.*, vol. 107, no. 1-3, pp. 325–337, 2009.

[46] A. Aldroubi, C. Cabrelli, and U. Molter, "Optimal non-linear models for sparsity and sampling," *J. Fourier Analysis Applicat.*, vol. 14, no. 5-6, pp. 793–812, 2008.

[47] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Royal Stat. Soc.*, vol. 61, no. 3, pp. 611–622, 1999.

[48] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[49] C. Archambeau, N. Delannay, and M. Verleysen, "Mixtures of robust probabilistic principal component analyzers," *Neurocomputing*, vol. 71, no. 7–9, pp. 1274–1282, 2008.

[50] A. Gruber and Y. Weiss, "Multibody factorization with uncertainty and missing data using the EM algorithm," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 1, pp. 707–714.

[51] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. Int. Conf. Machine Learning*, 2009, pp. 777–780.

[52] A. Leonardis, H. Bischof, and J. Maver, "Multiple eigenspaces," *Pattern Recognit.*, vol. 35, no. 11, pp. 2613–2627, 2002.

[53] Z. Fan, J. Zhou, and Y. Wu, "Multibody grouping by inference of multiple subspaces from high-dimensional data using oriented-frames," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 1, pp. 91–105, 2006.

[54] M. A. Fischler and R. C. Bolles, "RANSAC random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[55] J. Yan and M. Pollefeys, "Articulated motion segmentation using RANSAC with priors," in *Proc. Workshop Dynamical Vision*, 2005, pp. 75–85.

[56] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[57] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[58] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie, "Beyond pairwise clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2005, vol. 2, pp. 838–845.

[59] V. Govindu, "A tensor decomposition for geometric grouping and segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 1150–1157.

[60] G. Chen and G. Lerman, "Foundations of a multi-way spectral clustering framework for hybrid linear modeling," *Foundat. Comput. Math.*, vol. 9, no. 5, pp. 517–558, 2009.

[61] G. Chen, S. Atev, and G. Lerman, "Kernel spectral curvature clustering (KSCC)," in *Proc. Workshop Dynamical Vision*, 2009.

[62] F. Lauer and C. Schnörr, "Spectral clustering of linear subspaces for motion segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2009.

[63] R. Vidal, "Online clustering of moving hyperplanes," in *Proc. Neural Information Processing Systems, NIPS*, 2006.

[64] R. Vidal, Y. Ma, S. Soatto, and S. Sastry, "Two-view multibody structure from motion," *Int. J. Comput. Vis.*, vol. 68, no. 1, pp. 7–25, 2006.

[65] R. Vidal and Y. Ma, "A unified algebraic approach to 2-D and 3-D motion segmentation," *J. Math. Imag. Vis.*, vol. 25, no. 3, pp. 403–421, 2006.

[66] R. Vidal and R. Hartley, "Three-view multibody structure from motion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 2, pp. 214–227, 2008.

[67] S. Rao, A. Yang, S. Sastry, and Y. Ma, "Robust algebraic segmentation of mixed rigid-body and planar motions from two views," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 425–446, 2010. [SP]

Arta A. Jamshidi, Michael J. Kirby, and Dave S. Broomhead

Geometric Manifold Learning

Using dimension reduction and skew radial basis functions

© DIGITAL STOCK & LUSHPIX

We present algorithms for analyzing massive and high-dimensional data sets motivated by theorems from geometry and topology. Optimization criteria for computing data projections are discussed and skew radial basis functions (sRBFs) for constructing nonlinear mappings with sharp transitions are demonstrated. Examples related to modeling dynamical systems, including hurricane intensity and financial time series prediction, are presented. The article represents an overview of the authors' and collaborators' work in manifold learning.

WHY MANIFOLDS?

Over the last decade there has been a surge of interest in the modeling of large data sets using manifold theory; see, e.g., [36]–[40]. Abstract concepts such as isometric embeddings, homeomorphisms, diffeomorphisms, and Betti numbers began gradually creeping into talks and discussions at conferences concerned with developing tools for the analysis of massive high-dimensional data sets.

In concert with this surge were preliminary results that indicated manifolds could capture the essence of data in a way that fundamentally outperformed linear methodologies, the purpose of which is to essentially describe things that are flat. As a result, algorithms capable of extracting geometric or topological descriptions of data have become of widespread interest to theoreticians as well as practioners in the field; see, e.g., [44].

In what follows, we primarily focus on presenting our own efforts related to developing algorithms that extract geometric information in large data sets. Along the way, we will highlight the appropriate theoretical background and intuition concerning the deeper mathematical ideas.

MANIFOLD THEORY

Manifolds are, in some sense, well known to us, and at the same time, they are one of the most intriguing topics in mathematics. There are only two one-dimensional manifolds, i.e., the line and

Digital Object Identifier 10.1109/MSP.2010.939550
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

the circle [10]. Every one-dimensional connected set can be mapped homeomorphically to one of these one dimensional manifolds. (A homeomorphism is a mapping that has a continuous inverse.) The situation for two-dimensional manifolds is only slightly more complicated. Compact manifolds in two dimensions are characterized by the number of their holes, or genus. A two-dimensional manifold with no holes is a sphere; with one hole a torus and so on. In contrast, three-dimensional manifolds have no simple characterization. We still have a three-dimensional sphere and a three-dimensional torus, but, in general, three-dimensional manifolds are an area of intense research. Grigori Perelman recently proved the Poincare conjecture, a century old problem about three-dimensional manifolds.

MANIFOLDS ARE, IN SOME SENSE, WELL KNOWN TO US, AND AT THE SAME TIME, THEY ARE ONE OF THE MOST INTRIGUING TOPICS IN MATHEMATICS.

Loosely speaking, we can think of a manifold as a nonlinear object that looks locally linear. Indeed, a local approximation to a manifold is given by its tangent space, a vector space with the same dimension as the manifold.

The rate at which the manifold departs from this point of tangency is a measure of how nonlinear, or curved, the manifold is. If we move about the manifold, constructing local approximations as we go, we will observe that the dimension of each tangent space is exactly the same, i.e., a fixed integer m. For every point on this m-dimensional manifold there is a nonlinear mapping from the m-dimensional Euclidean space, i.e., the tangent space, to the manifold. In this sense, manifolds can be modeled locally as mappings on tangent spaces. Although we will not pursue this direction here, we note that this approach is of considerable interest [35], [34].

Manifolds may be modeled mathematically in a variety of ways. For example, consider x as n-dimensional and suppose

$$f(x) = 0,$$

where f is a scalar valued function. The locus of points x for which this equation is satisfied may be used to represent an $(n-1)$-dimensional manifold; each additional constraint may reduce the dimension of the manifold by one. Alternatively, the graph of a function $(x, g(x))$ is another model for a manifold; we will discuss the latter model in more detail below.

In contrast to subspaces of vector spaces, submanifolds of manifolds are not closed under addition. In other words, if you take two points in a subspace, the sum of these two points is also in the subspace. However, the sum of two points on a submanifold generally is not contained in the submanifold. For example, the sum of two points in a plane remain in the plane while the sum of two points on a unit circle do not live on the unit circle. Evidence suggests that the set of digital images of human faces reside on a manifold rather than a subspace. If faces did lie in a subspace, then eigenfaces, which are just an appropriately weighted sum of faces, should look human! See Figure 1 for confirmation. See [1] for details.

One of the first issues that arises when one is attempting to model data as one or more manifolds is whether the data really has a manifold structure. One of the most effective approaches we have found for doing this is based on an ϵ-ball scaling algorithm [9]. Here one picks a point from the data at random and then probes it with a ball, or local neighborhood. As the ball of radius ϵ grows, one looks at the data contained in it and performs a singular value decomposition. The set of singular values that scale linearly with the epsilon ball span the tangent space and hence determine the dimension of the manifold. If one repeats this calculation over and over and the same fixed dimension is obtained, this is excellent evidence that the data resides on a manifold.

It can also been argued that the data of interest need not actually live on a manifold for this mathematical framework to be useful. If the data does not reside on a manifold, we may invoke the idea of a manifold model in the same sense that we use lines of best fit to approximate data that is not really on a line. Also, as we see at the end of this article, we can take any data and build special manifolds that can be exploited for knowledge discovery.

[FIG1] Eigenfaces are the weighted sum of the digital images in the database. Since they do not appear human, we may conclude that the data set is not closed under addition. This data does not reside in a subspace but appears to lie on a manifold. This figure originally appeared in [1] and is reprinted with permission from John Wiley & Sons, Inc.

MANIFOLDS AND NONLINEAR DATA REDUCTION

We assume that the data is collected in an n-dimensional ambient (or data-acquisition) space and that the actual topological, i.e., local, dimension of the data is in general much smaller than n. The task is to determine a representation of the data that exploits its geometric structure and is consequently greatly simplified.

WHITNEY'S THEOREM IS MORE THAN A THEORETICAL STATEMENT ABOUT THE CONDITIONS UNDER WHICH A COPY OF THE MANIFOLD MAY BE FOUND.

As our motivating example, we consider a closed curve that passes through each of the standard basis vectors e_i where this vector consists of $n-1$ zeros and a one in the ith component. It is clear that standard techniques such as the singular value decomposition, or principal component analysis, suggest that the basis dimension of the space is n while the topological dimension of the curve is only one.

Closed curves of data arise in nature whenever a periodic phenomenon is being observed or modeled. They may be modeled by recognizing that there exists a **homeomorphism** between such a data set and the unit circle in $\mathbb{R}^2$.

BOTTLENECK MANIFOLDS AND NEURAL NETWORKS

This section concerns nonlinear data reduction to a bottleneck manifold using a technique proposed in [43]. It is based on the idea that we can compose mappings G and H such that they behave as the identity, i.e.,

$$x = (H \circ G)(x),$$

where

$$y = G(x) \in \mathbb{R}^m$$

reveals the data in a space of dimension $m < n$. In one example, we reduced the dimension of a closed curve in $\mathbb{R}^{20}$ to the unit circle, i.e., $m = 2$. The drawback to this approach is that the dimension m is unknown in general and the methods for computing G and H are based on expensive nonlinear optimization routines that determine the parameters for G and H such that their composition approximates the identity. Further, the mapping is not unique given

$$x = (H \circ F \circ F^{-1} \circ G)(x)$$

for an invertible function F.

DATA AS THE GRAPH OF A FUNCTION

The bottleneck neural network idea suggests the representation of data as the graph of a function, i.e., the set of points $(\hat{p}, f(\hat{p}))$ where

$$\hat{q} = f(\hat{p}), \tag{1}$$

and $\hat{p}$ is a point in the domain $U \subset \mathbb{R}^d$ of the function and the point $\hat{q}$ in the range $V \subset \mathbb{R}^{n-d}$. The challenge here is that the data initially resides in an n-dimensional ambient, or acquisition, space and a useful domain U as well as its dimension are unknown.

However, if we can solve these problems we may express the reconstruction of the original data in terms of two components, one linear and one nonlinear, specifically,

$$x = \Phi\hat{p} + (I - \Phi)f(\hat{p}), \tag{2}$$

where Φ is a matrix of d basis vectors spanning the domain of f. The nonlinear representation of the data comes from writing the residuals of the projection onto the range of Φ in terms of a function of the projected data values.

Conditions governing the existence of this function f when the data reside on a manifold are addressed in the section "Whitney's Theorem." An algorithm for computing the basis that has especially nice properties is summarized in the section "Bilipschitz Maps Are Good Projections."

WHITNEY'S THEOREM

Whitney's 1930s embedding theorem [12] from differential topology assures the existence of the mapping above for data sampled from a manifold. Roughly speaking, it states that an m-dimensional manifold can be realized in a Euclidean space of dimension $2m + 1$. In other words, a potentially low-dimensional "copy" of the manifold can be recovered and that a bijection exists between the original and its copy. The statement of this theorem is independent of the original dimension of the data n.

Whitney's theorem is more than a theoretical statement about the conditions under which a copy of the manifold may be found. It actually provides a blueprint for the construction of such a representation. Whitney shows that projecting data is admissable as long as all bad directions are avoided. Such directions are defined by the secant set is

$$\Sigma = \left\{ \frac{x-y}{\|x-y\|} : \forall x, y \in \mathcal{M}, x \neq y \right\}.$$

In general, it is intuitive that one should not project data along a secant of the data if one seeks to invert the mapping. Two points projected in this way will be identified as a single point in the image of the projection and there is no way to recover the original points.

In summary, Whitney's theorem indicates that data sampled from an m-dimensional manifold may be mapped linearly using a projection matrix of rank $2m + 1$ and that this mapping has a nonlinear inverse that losslessly recovers the data. Whitney's theorem describes the set of admissible projections as open and dense, i.e., they are not hard to find. However, Whitney's theorem does not prescribe a recipe for constructing these projections in any algorithmic sense. Further, only the existence of

the nonlinear inverse is proven and no technique for computing it is provided.

We outline practical approaches for solving each of these problems, i.e., data reduction and reconstruction, in the sections "Bilipschitz Maps Are Good Projections" and "Data Fitting with Radial Basis Functions," respectively.

THE BILIPSCHITZ MAPPING THEOREM IN CONJUNCTION WITH WHITNEY'S THEOREM SUGGESTS AN INTERESTING APPROACH FOR CONSTRUCTING GOOD PROJECTIONS OF DATA.

BILIPSCHITZ MAPS ARE GOOD PROJECTIONS

A function $f(x)$ is said to be bilipschitz on S if for all $x, y \in S$ holds that

$$a\|x-y\| \le \|f(x)-f(y)\| \le b\|x-y\|,$$

where a is the injectivity parameter and b is the Lipschitz constant. The bilipschitz mapping theorem [6] states that if a function $f: S \to T$ is bilipschitz, then

$$\dim(S) = \dim(T),$$

where the dimensions being preserved includes both topological and fractal dimensions, i.e., the data need not reside on a manifold.

Assume that the data set of interest is a discrete sampling of a compact m-dimensional submanifold of a raw vector space of high dimension q where typically $q \gg 2m+1$. We would like to reduce the dimension of the data such that we have a copy of the data that could be, if needed, mapped back to the original data. In other words, the projection should be lossless.

The bilipschitz mapping theorem in conjunction with Whitney's theorem suggests an interesting approach for constructing good projections of data. We use the term *good* here rather than optimal, given that we envision cases where the iterative approach for determining the projection terminates early when it has achieved sufficient quality. See [8], [7], and [11] for additional details.

First, the right inequality in the bilipschitz condition is satisfied automatically if $f(x)$ is taken to be a projection $\mathbb{P}$, given that projections only shrink (or maintain) the length of secants. Similarly, the condition

$$\|\mathbb{P}\hat{k}\| \ge a > 0, \tag{3}$$

for all unit secants $\hat{k} \in \Sigma$ ensures that the mapping is bilipschitz. Moreover, it has been shown that the Lischitz constant of the nonlinear inverse to $\mathbb{P}$ is Lipschitz with Lipschitz constant $1/a$. This provides a strong indication that solving

$$\mathbb{P}^* = \arg\max_{\mathbb{P}} \min_{\hat{k}\in\Sigma} \|\mathbb{P}\hat{k}\|$$

is an excellent way to determine a reduction of the data set as the inverse at the heart of the reconstruction is optimally smooth. An algorithm for solving this optimization problem was proposed in [7] and extensions to noisy data considered in [8]. A related differentiable objective function that prohibits small projected secants,

$$F(\mathbb{P}) = \frac{1}{|\Sigma|}\sum_{\hat{k}\in\Sigma}\frac{1}{\|\mathbb{P}\hat{k}\|}, \tag{4}$$

has also proven to be very useful. Since we are seeking optimal subspaces, tools from geometric optimization can be used; see [11] and references therein for details.

It is interesting to distinguish the bilipschitz condition from the special case of an isometric, or distance preserving, mapping where

$$\|f(x)-f(y)\| = \|x-y\|.$$

These mappings have also received considerable attention in the literature recently [39], [40].

TAKENS' THEOREM AND TIME-DELAYED EMBEDDING

Takens' theorem [13] provides a theoretical foundation for the reconstruction of an m-dimensional manifold when only a scalar time series is observable. We assume that the actual time series $x(t) = (x_1(t), \ldots, x_n(t))$ generated by a dynamical system

$$\frac{dx}{dt} = f(x),$$

trace out solutions on a manifold of dimension m that cannot be observed directly. Takens' theorem is a result about observability—in the sense of control theory—of finite-dimensional, nonlinear dynamical systems. Imagine that we have a scalar measurement that we make on the system; let's write this as a function $y: \mathbb{R}^n \to \mathbb{R}$. We shall assume that the restriction of this function to the manifold is sufficiently smooth. If at some time t we make a measurement, the value that we obtain is $y(t) = y(x(t))$. A sequence of such measurements made at equally spaced times gives a scalar time series: $(\ldots, y(t-\delta t), y(t), y(t+\delta t), \ldots)$, where δt is the sampling interval.

Takens' theorem is based on the construction of a delay map from such a scalar time series

$$\hat{x} = (y(t), y(t-\delta t), \ldots, y(t-(k-1)\delta t)).$$

Here $\hat{x}$ should be seen as the image of the point $x(t)$ under a map from the m-dimensional manifold to $\mathbb{R}^k$. Takens' theorem amounts to the statement that if $k > 2m$ then, generically, this map is an embedding of the manifold. This statement of Takens' theorem assumes that the vector field f and the sampling time satisfy certain reasonable—indeed, generic—constraints (see [Sect. 2, 21] for more details). Here, *generically* means that

within the set of all C^2 real-valued measurement functions on the manifold, the set which gives an embedding is open and dense. Embedding means that the image of the manifold in $\mathbb{R}^k$ is a copy of the original with the same geometric structure—the image is the same as the original up to a smooth (nonlinear) change of coordinates.

RECENTLY, THERE HAS BEEN CONSIDERABLE RESEARCH CONCERNING SO-CALLED SPECIAL MANIFOLDS AND THEIR APPLICATIONS TO THE ANALYSIS OF LARGE DATA SETS.

An application to classification can be envisioned as follows: assume that $x(t)$ belongs to class C_+ if for some smooth, real-valued function h, $h(x(t)) > 0$ and belongs to C_- if $h(x(t)) \leq 0$. Further, assume that only the scalar $y(x(t))$ is observed. In general, it is impossible to tell the class of $x(t)$ from $y(x(t))$, however, we may use Takens' theorem and a delay embedding to recover this class information. This idea is potentially useful for classification of time series such as those that arise in electroencephalography (EEG) analysis [42]. If EEG activity in the brain has coherent structure that can be exploited via time-delay embeddings, this approach is effectively a form of super resolution. The application of Takens' theorem to the time series prediction problem is discussed, e.g., in [14].

MANIFOLD RECONSTRUCTION

Here we describe a recently developed algorithm for building nonlinear models for multivariate data that is particularly useful for modeling time evolving trajectories on manifolds. In general, the techniques described here are suitable for finding the nonlinear inverse required to represent the data as a graph of a function.

DATA FITTING WITH RADIAL BASIS FUNCTIONS

The mappings described in the previous section are of the form

$$f : U \subset \mathbb{R}^n \to \mathrm{V} \subset \mathbb{R}^m, \tag{5}$$

where we assume that, in general, both m and n may be greater than 1. In the data fitting problem, we assume that we have samples $x^{(k)} \in U$ and $y^{(k)} \in V$ are indexed by k and related via f, a nonlinear function, as $y^{(k)} = f(x^{(k)})$.

A standard radial basis function seeks to represent $f(x)$ as

$$f(x) = \sum_{i=1}^{N_c} w_i \phi(\|x - c_i\|_{W_i}), \tag{6}$$

where x is an input pattern, ϕ is an RBF centered at location c_i, w_i denotes the weight for ith RBF and N_c is the number of functions being used. The term W denotes the parameters in the weighted inner product $\|x\|_W = \sqrt{x^T W x}$.

RBFs have been widely used for data approximation. Multiquadrics, i.e., functions of the form $\phi(r) = \sqrt{1 + r^2}$, were introduced by [15] for modeling scattered data on topographical surfaces. Thin plate splines, $\phi(r) = r^2 \ln r$, were introduced by [16] for surface interpolation. Gaussian RBFs, $\phi(r) = \exp(-r^2/\sigma^2)$, were proposed by [2] for data fitting and classification. There is a significant literature treating theoretical aspects of RBFs including universal approximation theorems; see, e.g., [4] and [18]–[20]. A large number of algorithms have been proposed in the literature for computing the model parameters and are reviewed in [26]. The research monographs [22]–[24] contain additional references across theory, algorithms, and applications.

Alternatively, we also consider the sRBFs introduced in [25]–[27], which are defined as

$$f(x) = \sum_{i=1}^{N_c} w_i z(x\,; \nu_i) \phi(\|x - c_i\|_{W_i}). \tag{7}$$

In the above equation, the function $z(x; \nu_i)$ is a skew component that makes the representation nonradial, ν_i contains the parameters for the symmetric breaking term. This has the advantage of being able to represent asymmetric data, e.g., data near boundaries, much more efficiently. An example of an sRBF is the skew-Gaussian RBF

$$f(x) = \sum_{i=1}^{N_c} w_i \exp(-\|x - c_i\|_{W_i}^2) \int_{-\infty}^{-\lambda_i^T (x - c_i)} \exp(-y^2) dy, \tag{8}$$

where λ_i is a vector of skew parameters. In this article, we employ a truncated cosine function in the same fashion as a Hanning filter to produce an RBF with compact support [28], i.e.,

$$f(x) = \sum_{i=1}^{N_c} w_i (\cos(\|x - c_i\|_{W_i} \pi) + 1) H(1 - \|x - c_i\|_{W_i}), \tag{9}$$

where H is the Heaviside function. To create an sRBF here, we employ the Arctan function as the skewing term z. These functions, taken together, result in the Arctan-Hanning sRBF.

Of course the issue now is to determine the parameters associated with the data fitting problem. Of particular importance are the locations of the fitting functions $\{c_i\}$ and the width of the basis $\{W_i\}$ as well as the number of fitting functions N_c.

There are a wide range of algorithms that have appeared in the literature for determining these parameters, e.g., [17]. Here we used an iterative approach that tests whether the residuals are independent and identically distributed (i.i.d.) noise at each step. If not, then additional functions are added until the i.i.d. test is passed with at least 95% confidence. For more details concerning the algorithms used to generate the results presented here, see [26], [29], [25], and [30].

EXAMPLES

In this section, we provide several applications of the theoretical and algorithmic ideas presented above.

THE PRINGLE DATA SET

In this example, we illustrate the representation of data, a topological circle, on a manifold as the graph of a function. We employ the pringle data set, generated as the solution to a

system of ordinary differential equation as described in [11].

The task is to construct a mapping from an (x, y) value in the plane to its corresponding z value on the curve. Thus, we are fitting the graph of a function from $\mathbb{R}^2$ to $\mathbb{R}$.

TAKENS' THEOREM PROVIDES A THEORETICAL FOUNDATION FOR THE RECONSTRUCTION OF AN M-DIMENSIONAL MANIFOLD WHEN ONLY A SCALAR TIME SERIES IS OBSERVABLE.

Figure 2(a) shows the location and shape of the first (of four RBFs) that is generated by the algorithm to model the data; at that point the i.i.d. stopping criteria is satisfied. The training data (consisting of 101 points, almost two cycles) and the RBFs are displayed together to illustrate how the algorithm has fit the RBFs to the data. A plot of the output of the model and testing data set consisting of 500 points (almost nine cycles) are shown in Figure 2(b). The fact that the solution is periodic will clearly illustrate the need for spatial as well as temporal windowing of the data.

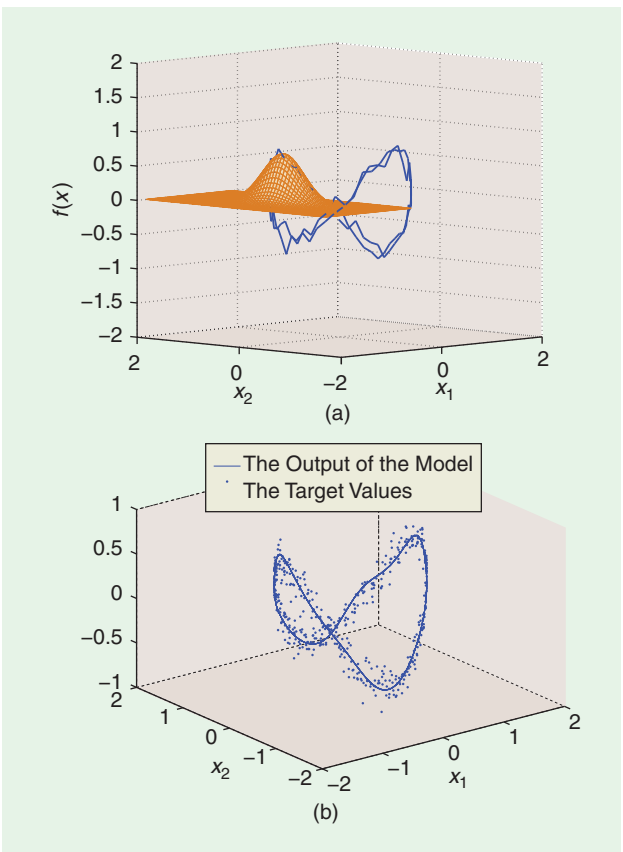
PREDICTING HURRICANE INTENSIFICATION

In this example, we predict the instantaneous maximum intensity of a hurricane. The data used to build the model was generated by a numerical simulation of an axisymmetric model as described in [31] and references therein. The available data represents maximum tangential wind speed with intervals of 5 s. We have employed a time delay embedding scheme for computing the embedding dimension of the time series [32]. The goal is to predict one step ahead using the previous four values. In the spirit of Takens' theorem, this problem could be formulated as

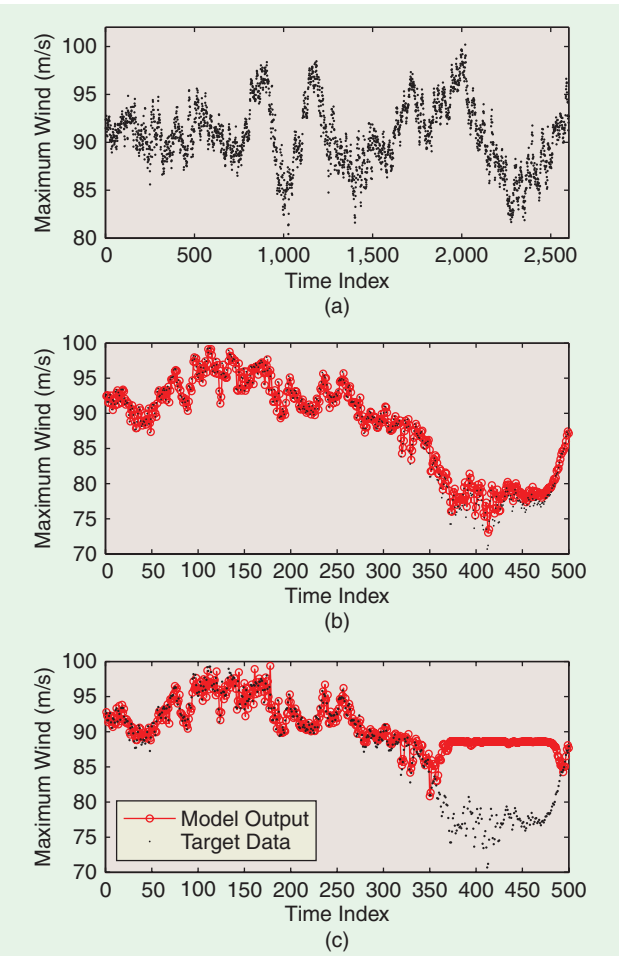
$$x_{n+1} = f(x_n, x_{n-1}, x_{n-2}, x_{n-3}).$$

Alternatively we aim at constructing an sRBF mapping as

$$f : (x_n, x_{n-1}, x_{n-2}, x_{n-3}) \in \mathcal{M} \subset \mathbb{R}^4 \rightarrow x_{n+1} \in \mathbb{R},$$



[FIG2] The solution to the dynamical system is corrupted with Gaussian noise with standard deviation (STD) of 0.1. (a) The primary first RBF allocated by the algorithm and the training data set. (b) The testing data set and the output of the four-mode model. Originally presented in [29]. Figure reprinted with permission by the Society for Industrial and Applied Mathematics.



[FIG3] The performance of the symmetric and sRBFs on hurricane intensity prediction [25]. (a) The training and validation data sets. (b) The testing set and the output of the one-mode sRBF. (c) The testing set and the output of the three-mode symmetric RBF. Figure reprinted with permission from the Society for Industrial and Applied Mathematics.

where $\mathcal{M}$ is the set of all possible points from which the data is sampled. Due to the sharp transitions in the maximum wind speed, the above map experiences discontinuous behavior, which makes the map more difficult to learn using symmetric basis functions. Figure 3(a) shows the maximum wind speed of a hurricane in a time series format.

In this study, we model the steady state behavior of the dynamics. The data set consists of 1,801 training, 800 validation, and 500 testing data points taken sequentially in time. Figure 3(a) shows the training and validation data sets. Figure 3(b) and (c) shows the predicted values using data adapted Erf-Gaussian sRBFs and symmetric Gaussian RBFs, respectively. Note that the asymmetric fit was complete using one RBF, and the root mean square error (RMSE) value of the testing data is 1.3036. Confidence levels of 96.22% and 95.62% were achieved on training and validations data sets, respectively. The symmetric RBF training stopped based on 95.37% confidence criteria on a validation set with seven RBFs. A confidence level of 95% was achieved on the training using three RBFs. We found that the three-mode RBF model preformed as well or better than the seven-mode model, and we report results only for this. The RMSE on the test set using the three-mode RBF was 5.7918. Hence, in this example, a one mode sRBF mode performs better than a three-mode RBF model. In this example, the parameters are optimized by employing BFGS; see [26] for details and references.

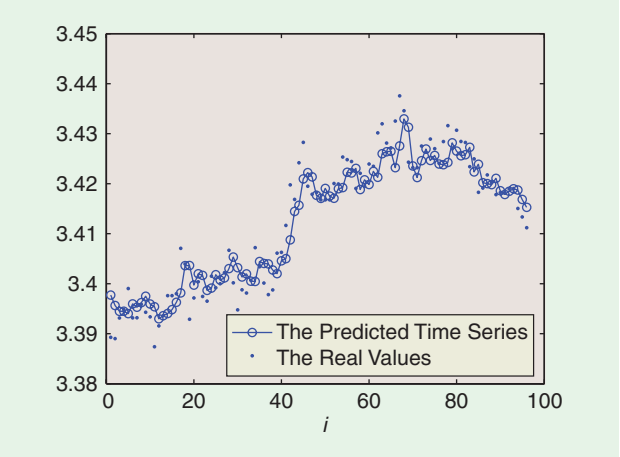
PREDICTING EXCHANGE RATE DATA

This data set consists of the Deutsche Mark/French Franc (daily) exchange rates over 701 days. As mentioned in [33], this data set has irregular nonstationary components due to government intervention in the European exchange rate mechanism. A window of five previous values can be used as input, hence forming an interesting data set to make a mapping from $\mathbb{R}^5$ to $\mathbb{R}$. Figure 4 shows the output of the resulting model (one-step prediction values) and the target (market) values for the test data. The modeling process terminated with a model of order three with 97.17% confidence. The model produces the RMSE of 0.0043, $NPE_1 = 0.2760$ and $NPE_2 = 0.1033$. The results for this exchange rate data reported in [33] show a model fit with 11 RBFs and NPE_1 of 0.336.

Although most of the applications shown above form mappings from $\mathbb{R}^m$ to $\mathbb{R}$, we have also presented multivariate results on the pringle and Mackey-Glass data sets in [30].

BEYOND DATA MANIFOLDS

Recently, there has been considerable research concerning so-called special manifolds and their applications to the analysis of large data sets; see, e.g., [41]. Now, a point on a data set actually is a set of data points. For example, consider the variation of illumination on an object such as a human face. The set of images collected under a variation in illumination contains information concerning the identity of the subject that goes beyond an individual snapshot. So, if we can encode this information in a useful way one might suspect the resulting recognition algorithm could be very useful.



[FIG4] The one-step prediction of the exchange rate data using a three-mode RBF constructed using spatiotemporal balls [29]. Figure reprinted with permission from the Society for Industrial and Applied Mathematics.

The setting of the Grassmann manifold, or Grassmannian allows one to encode data as a subspace. The set of k-dimensional subspaces of $\mathbb{R}^n$, referred to as $G(n, k)$, may be used to encode sets of data such as images. Classification of these sets of sets may be achieved using standard norms defined on Grassmannians. Recognition rates using this approach have proved compelling; see [41] and references therein.

CONCLUSIONS

A wide range of problems, e.g., in machine learning, optimal control, mathematical modeling of physical systems, biological systems, human behavior, voice recognition, failure prediction, and image processing, often require the construction of relationships from observed data. When this data has a special manifold structure, we can exploit ideas from geometry and topology to motivate new analysis techniques. In this article, we have focused our applications on time-series modeling where there is some explicit or implicit manifold structure. The data fitting methodology does not require the tuning of ad hoc parameters and, by virtue of their skewness, the sRBFs are suitable for learning data with asymmetric behavior or singularities such as jumps and sharp transitions.

ACKNOWLEDGMENTS

Arta Jamshidi was partially supported by NSF grant ATM-530884. Michael Kirby acknowledges partial support from NSF grants DMS-0915262, ATM-530884, and DOD/AFOSR grant FA9550-08-1-0166. We would also like to thank John Persing for providing the hurricane simulation data used in this article.

AUTHORS

Arta A. Jamshidi (arta@princeton.edu) holds a B.Sc. degree (Summa Cum Laude) and an M.Sc. degree in electrical and computer engineering from the University of Tehran and

Colorado State University (CSU), respectively. He earned his M. Sc. and Ph.D. degrees in mathematics in 2004 and 2008 from CSU. He completed a post doctoral research term at Imperial College London working on imaging systems. Currently, he is a post doctoral researcher at Princeton University working on space-time modeling of multiscale dynamics. His research interests include computational and applied mathematics, geometric methods for analysis of large data sets, optimization and mathematical modeling, nonlinear signal and image processing, and their applications. His articles have been published in SIAM and IEEE journals.

Michael J. Kirby (kirby@math.colostate.edu) is a professor in the Department of Mathematics, with a joint appointment in the Department of Computer Science, Colorado State University. He received an S.B. degree in mathematics from the Massachusetts Institute of Technology in 1984 and a Ph.D. degree in applied mathematics from Brown University in 1988. He was an Alexander von Humboldt Fellow at the Institute fuer Informationsverarbeitung in Tuebingen, Germany. He authored a book on geometric data analysis and has written over 50 papers in this area. His interests include the geometry of massive data sets and developing algorithms to exploit it. He is currently the director of the Pattern Analysis Lab at Colorado State University.

Dave S. Broomhead (D.S.Broomhead@manchester.ac.uk) is a professor of applied mathematics in the School of Mathematics at the University of Manchester and director of the Centre for Interdisciplinary Computational and Dynamical Analysis that was created in 2007 within the University of Manchester. His original appointment in Manchester was to a chair in 1995 following a 12-year period working in the scientific civil service. His major research interest is in nonlinear dynamical systems theory and in the application of these ideas to interdisciplinary projects.

REFERENCES

[1] M. Kirby, *Geometric Data Analysis*. New York: Wiley, 2001.

[2] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," *Complex Syst.*, vol. 2, pp. 321–355, 1988.

[3] M. J. D. Powell, "Radial basis functions for multivariable interpolation: A review," in *Algorithms for Approximation*, J. C. Mason and M. G. Cox, Eds. Oxford: Clarendon, 1987, pp. 143–167.

[4] M. J. D. Powell, "The theory of radial basis functions in 1990," in *Wavelets, Subdivision Algorithms, and Radial Basis Functions, vol. II: Advances in Numerical Analysis*, W. Light, Ed. London, U.K.: Oxford Univ. Press, 1992, pp. 105–210.

[5] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, no. 6, pp. 716–723, 1974.

[6] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. New York: Wiley, 1990.

[7] D. S. Broomhead and M. Kirby, "A new approach for dimensionality reduction: Theory and algorithms," *SIAM J. Appl. Math.*, vol. 60, no. 6, pp. 2114–2142, 2000.

[8] D. S. Broomhead and M. Kirby, "The Whitney reduction network: A method for computing autoassociative graphs," *Neural Comput.*, vol. 13, pp. 2595–2616, 2001.

[9] D. S. Broomhead, R. Jones, and G. P. King, "Topological dimension and local coordinates from time series data," *J. Phys. A, Math. Gen.*, vol. 20, no. 9, pp. L563–L569, 1987.

[10] D. Barden and C. Thomas, *An Introduction to Differential Manifolds*. London: Imperial College Press, 2003.

[11] D. S. Broomhead and M. Kirby, "Large dimensionality reduction using secant-based projection methods: The induced dynamics in projected systems," *Nonlinear Dyn.*, vol. 41, pp. 47–67, 2005.

[12] H. Whitney, "Differential manifolds," *Ann. Math.*, vol. 37, no. 3, pp. 645–680, July 1936.

[13] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence,* vol, 898, D. Rand and L.-S. Young, Eds. Warwick, 1980, pp. 366–381.

[14] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Phys. Rev. Lett.*, vol. 45, no. 9, pp. 712–716, 1980.

[15] R. L. Hardy, "Multiquadric equations of topography and other irregular surfaces," *J. Geophys. Res.*, vol. 76, no. 8, pp. 1905–1915, 1997.

[16] R. L. Harder and R. N. Desmarais, "Interpolation using surface splines," *J. Aircraft*, vol. 9, no. 2, pp. 189–191, 1972.

[17] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[18] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, vol. 3, pp. 246–257, 1991.

[19] J. Park and I. W. Sandberg, "Approximation and radial-basis-function networks," *Neural Comput.*, vol. 5, pp. 305–316, 1993.

[20] R. Schaback and H. Wendland, "Characterization and construction of radial basis functions," in *Multivariate Approximation and Applications, Eilat Proceedings*, N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2000, pp. 1–24.

[21] J. P. Huke and D. S. Broomhead, "Embedding theorems for non-uniformly sampled dynamical systems," *Nonlinearity*, vol. 20, no. 9, pp. 2205–2244, 2007.

[22] P. V. Lee and S. Haykin, *Regularized Radial Basis Function Networks Theory and Applications*. New York: Wiley, 2001.

[23] M. D. Buhmann, *Radial Basis Functions*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[24] H. Wendland, *Scattered Data Approximation*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[25] A. A. Jamshidi and M. J. Kirby, "Skew-radial basis function expansions for empirical modeling," *SIAM J. Sci. Comput.*, vol. 31, no. 6, pp. 4715–4743, Feb. 2010.

[26] A. A. Jamshidi, "Modeling spatio-temporal systems with skew radial basis functions: Theory, algorithms and applications," Ph.D. dissertation, Dept. Math., Colorado State Univ., Fort Collins, CO, 2008.

[27] A. A. Jamshidi and M. J. Kirby, "Skew-radial basis functions for modeling edges and jumps," in *Proc. 8th IMA Int. Conf. Mathematics in Signal Processing*, Cirencester, U.K., Dec. 2008, pp. 51–54.

[28] A. A. Jamshidi and M. J. Kirby, "Examples of compactly supported functions for radial basis approximations," in *Proc. 2006 Int. Conf. Machine Learning; Models, Technologies and Applications*, H. R. Arabnia, E. Kozerenko, and S. Shaumyan, Eds. Las Vegas: CSREA, June 2006, pp. 155–160.

[29] A. A. Jamshidi and M. J. Kirby, "Towards a black box algorithm for nonlinear function approximation over high-dimensional domains," *SIAM J. Sci. Comput.*, vol. 29, no. 3, pp. 941–963, May 2007.

[30] A. A. Jamshidi and M. J. Kirby, "Modeling multivariate time series on manifolds with skew radial basis functions," *Neural Comput*, vol. 23, no. 1, pp. 97–123, Jan. 2011.

[31] J. Persing and M. T. Montgomery, "Hurricane superintensity," *J. Atm. Sci.*, vol. 60, no. 19, pp. 2349–2371, 2003.

[32] M. Casdagli, D. D. Jardins, S. Eubank, J. D. Farmer, J. Gibson, N. Hunter, and J. Theiler, "Nonlinear modeling of chaotic time series: Theory and applications," in *Applied Chaos*, J. H. Kim and J. Stringer, Eds. New York: Wiley, 1992, pp. 335–380.

[33] A. MacLachlan, "An improved novelty criterion for resource allocating networks," in *Proc. IEE 5th Int. Conf. Artificial Neural Networks*, Cambridge, U.K.: IEE, July 1997, pp. 48–52.

[34] D. R. Hundley, "Local nonlinear modeling via neural charts," Ph.D. dissertation, Dept. Math., Colorado State Univ., Fort Collins, CO, 1998.

[35] M. Brand, "Charting a manifold," in *Proc. Neural Information Processing Systems 15 (NIPS'2002)*, Vancouver, Canada, Dec. 9–14, 2002, Cambridge, MA: MIT Press, pp. 961–968.

[36] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[37] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[38] S. Roweis, L. Saul, and G. Hinton, "Global coordination of local linear models," in *Proc. Neural Information Processing Systems 15 (NIPS'2002)*, Vancouver, Canada, Dec. 9–14, 2002, Cambridge, MA: MIT Press, pp. 889–896.

[39] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[40] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 10, pp. 5591–5596, 2003.

[41] J. R. Beveridge, B. Draper, J. M. Chang, M. Kirby, H. Kley, and C. Peterson, "Principal angles separate subject illumination spaces in YDB and CMU-PIE," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 2, pp. 351–363, 2009.

[42] M. Kirby and C. Anderson, "Geometric analysis for the characterization of nonstationary time series," in *Perspectives and Problems in Nonlinear Science: A Celebratory Volume in Honor of Larry Sirovich,* E. Kaplan, J. E. Marsden, and K. R. Sreenivasan, Eds. New York: Springer-Verlag, Mar. 2003, ch. 8, pp. 263–292.

[43] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE J.*, vol. 37, no. 2, pp. 233–243, 1991.

[44] G. Carlsson, "Topology and data," *Bull. Amer. Mat. Soc. B*, vol. 46, no. 2, pp. 255–308, 2009.

[SP]

Paul Honeine and Cédric Richard

Preimage Problem in Kernel-Based Machine Learning

[An intimate connection with the dimensionality-reduction problem]

Kernel machines have gained considerable popularity during the last 15 years, making a breakthrough in nonlinear signal processing and machine learning, thanks to extraordinary advances. This increased interest is undoubtedly driven by the practical goal of being able to easily develop efficient nonlinear algorithms. The key principle behind this, known as the kernel trick, exploits the fact that a great number of data-processing techniques do not explicitly depend on the data itself but rather on a similarity measure between them, i.e., an inner product. To provide a nonlinear extension of these techniques, one can apply a nonlinear transformation to the data, mapping them onto some feature space. According to the kernel trick, this can be achieved by simply replacing the inner product with a reproducing kernel (i.e., positive semidefinite symmetric function), the latter corresponds to an inner product in the feature space. One consequence is that the resulting nonlinear algorithms show significant performance improvements over their linear counterparts with essentially the same computational complexity.

While the nonlinear mapping from the input space to the feature space is central in kernel methods, the reverse mapping from the feature space back to the input space is also of primary interest. This is the case in many applications, including kernel principal component analysis (PCA) for signal and image denoising. Unfortunately, it turns out that the reverse mapping generally does not exist and only a few elements in the feature space have a valid preimage in the input space. The preimage problem consists of finding an approximate solution by identifying data in the input space based on their corresponding features in the high-dimensional feature space. It is essentially a dimensionality-reduction problem, and both have been intimately connected in their historical evolution, as studied in this article.

Digital Object Identifier 10.1109/MSP.2010.939747
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

AN INTRODUCTORY EXAMPLE: KERNEL PCA FOR DENOISING

LINEAR DENOISING WITH PCA

In general, some correlations exist among data, thus techniques for dimensionality reduction or the so-called feature extraction provide a way to confine the initial space to a subspace of lower dimensionality. The PCA, also known as Karhunen-Loève transformation, is one of the most widely used dimensionality-reduction techniques. Conventional PCA seeks principal directions that capture the highest variance in the data. Mutually orthonormal, these directions define the subspace, exhibiting information rather than noise, providing the optimal linear transformation. Here, the optimality is in the sense of least-mean-square reconstruction error. For instance, in data compression and manifold learning, much information is conserved by projecting onto the directions of highest variance, while in denoising, directions with small variance are dropped. These schemes are mathematically equivalent, and we use here a denoising schema without loss of generality.

Consider an input space $\mathcal{X}$ endowed by the inner product $\langle \cdot, \cdot \rangle$; for instance, a vectorial space with the Euclidean inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \boldsymbol{x}_i^\top \boldsymbol{x}_j$. Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ denote a set of available data (observations) from $\mathcal{X}$. PCA techniques seek the axes that maximize the mean variance of the projected data under the unit-norm constraint, namely, $\psi_1, \psi_2, \ldots, \psi_k$ by maximizing $(1/n)\sum_{i=1}^{n} |\langle \boldsymbol{x}_i, \psi_\ell \rangle|^2$ subject to $\langle \psi_\ell, \psi_{\ell'} \rangle = \delta_{\ell\ell'}$ for all $\ell, \ell' = 1, 2, \ldots, k$. In this expression, the Kronecker delta is defined as $\delta_{\ell\ell'} = 1$ if $\ell = \ell'$, and $\delta_{\ell\ell'} = 0$ otherwise. Solving this constrained optimization problem using the Lagrangian provides the following problem:

$$\lambda_\ell \psi_\ell = \boldsymbol{C} \psi_\ell, \tag{1}$$

where λ_ℓ defines the amount of variance captured by ψ_ℓ, and $\boldsymbol{C}$ is the covariance matrix of the data. In other words, $(\lambda_\ell, \psi_\ell)$ is the eigenvalue–eigenvector of the covariance matrix, data assumed zero-mean. Furthermore, eigenvectors lie in the span of the data, since for every $\ell = 1, 2, \ldots, k$ we have

$$\psi_\ell = \frac{1}{\lambda_\ell} \boldsymbol{C} \psi_\ell = \frac{1}{\lambda_\ell n} \sum_{i=1}^{n} \langle \boldsymbol{x}_i, \psi_\ell \rangle \boldsymbol{x}_i.$$

The eigenvectors associated with the largest eigenvalues provide a relevant low-dimensional subspace. As a consequence, we are interested in elements from this relevant subspace. This is the case, for instance, in data denoising, where the projection of a given noisy data onto this subspace provides its noise-free counterpart. Therefore, the latter can be written as an expansion of the eigenvectors, namely, for a noisy data $\tilde{\boldsymbol{x}}$, we get the denoised $\psi = \sum_{i=1}^{k} \langle \tilde{\boldsymbol{x}}, \psi_i \rangle \psi_i$, and from the aforementioned expression, as a linear expansion in terms of the available data, by taking the form

$$\psi = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i.$$

KERNEL PCA FOR NONLINEAR DENOISING

To provide a natural nonlinear extension of PCA, a nonlinear mapping is applied to the data as a preprocessing stage, prior to applying the PCA algorithm. Let $\phi(\cdot)$ be the nonlinear transformation mapping data from the input space $\mathcal{X}$ to some feature space $\mathcal{H}$. Then problem (1) essentially remains the same, with the covariance matrix associated to the transformed data. From the linear expansion with respect to the latter, the resulting principal axes take the form

$$\psi_\ell = \sum_{i=1}^{n} \langle \phi(\boldsymbol{x}_i), \psi_\ell \rangle_{\mathcal{H}} \, \phi(\boldsymbol{x}_i), \tag{2}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in the feature space $\mathcal{H}$. In this space, each feature ψ_ℓ lies in the span of the mapped input data, with the coefficients given by the ℓth eigenvector of the eigenproblem

$$n \lambda_\ell \boldsymbol{\alpha}_\ell = \boldsymbol{K} \boldsymbol{\alpha}_\ell, \tag{3}$$

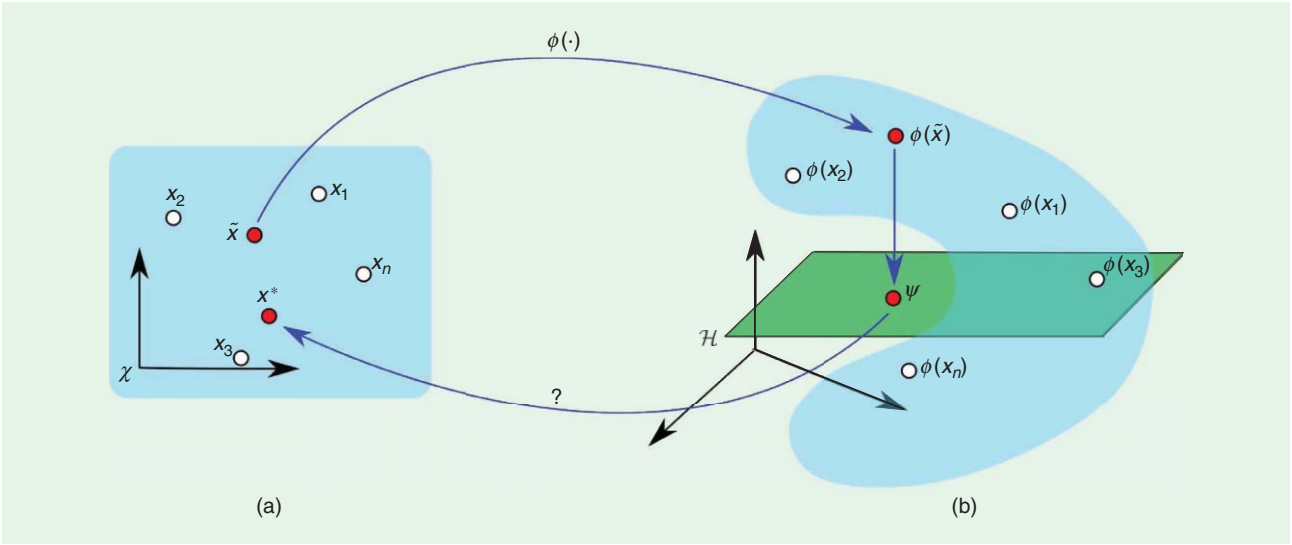
where $\boldsymbol{K}$ is the so-called Gram matrix with entries $\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}}$, for $i, j = 1, 2, \ldots, n$. As illustrated here, the expansion coefficients require only the evaluation of the inner products. Without the need to exhibit the mapping function, this information can be easily exploited for a large class of nonlinearities by substituting the inner product with a positive semidefinite kernel function. This argument is the kernel trick, which provides a nonlinear counterpart of the classical PCA algorithm, the so-called kernel PCA [1].

Consider the denoising application using kernel PCA. For a given $\tilde{\boldsymbol{x}}$, its nonlinear transformation $\phi(\tilde{\boldsymbol{x}})$ is projected onto the subspace spanned by the most relevant principal axes, providing the denoised pattern. The latter can be written as a linear expansion of the k principal axes, $\psi_1, \psi_2, \ldots, \psi_k$, with

$$\psi = \sum_{i=1}^{k} \langle \boldsymbol{x}, \psi_i \rangle \, \psi_i. \tag{4}$$

Equivalently, the denoised pattern can also be written as a linear expansion of the n images of the training data, namely $\psi = \sum_{i=1}^{n} \alpha_i \, \phi(\boldsymbol{x}_i)$, where the expansion in (2) is used. In practice, one is interested in representing the denoised pattern in the input space, as illustrated in Figure 1. It turns out that most elements of the feature space, including the denoised patterns, are not valid images, i.e., the result of applying the map to some input data. To get the denoised counterpart in the original input space, one needs to operate an approximation scheme, i.e., estimate $\boldsymbol{x}^*$ such that its image $\phi(\boldsymbol{x}^*)$ is as close as possible to ψ.

Beyond this kernel-PCA example, the kernel trick is well known in the machine-learning community. It provides flexibility to derive nonlinear techniques based on linear ones, with the data being implicitly mapped into a feature space. This space is given by the span of the mapped data, i.e., all the linear expansions of mapped data. The price to pay is that, in general, not each element of the space is necessarily the image of some data.



[FIG1] Kernel machines map the input space [blue region in (a)] into a higher-dimensional space [blue region in (b)]. The reproducing kernel Hilbert space (rkHs) $\mathcal{H}$ is defined as the completion of the span of the mapped input data, with elements written as a linear expansion of mapped data. However, not each element of $\mathcal{H}$ is necessarily the image of some input data. The preimage problem consists of going back to the input space, e.g., to represent in the input space elements of the rkHs (e.g., the effect of projecting onto a subspace).

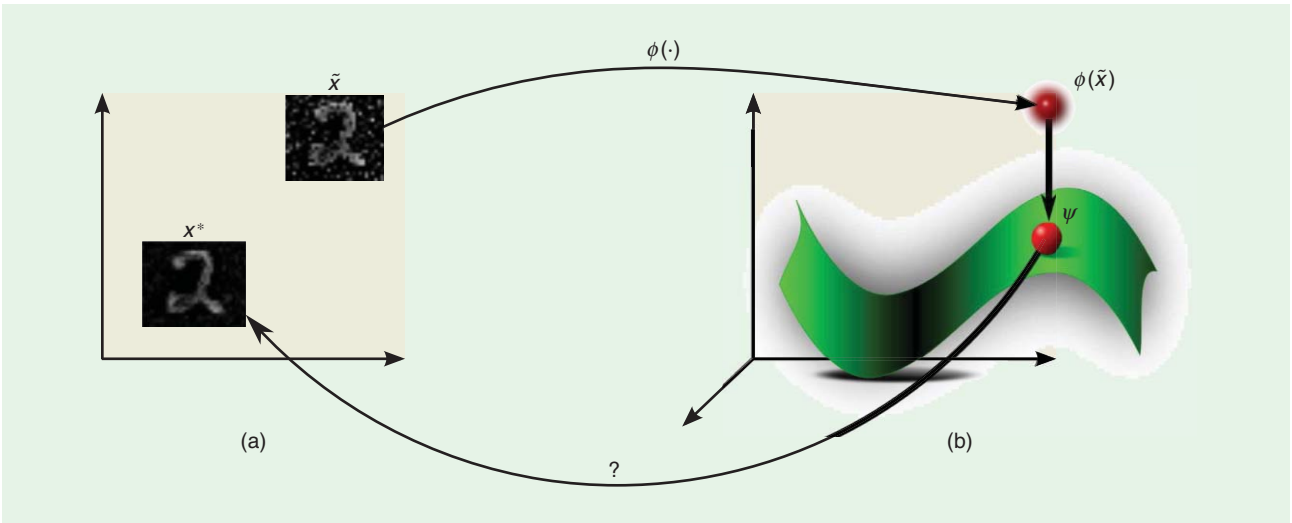
This is the case of most elements in the feature space, since they can be written as

$$\psi = \sum_{i=1}^{n} \alpha_i \, \phi(x_i),$$

as illustrated earlier with either a principal axe ψ_ℓ or a denoised feature ψ. To give proper interpretation for these components, one should define the way back from the feature into the input space. This is the preimage problem in kernel-based machine learning, as illustrated in Figure 2.

KERNEL-BASED MACHINE LEARNING

In the past 15 years or so, a novel breakthrough for artificial neural networks has been achieved in the field of pattern recognition and classification within the framework of kernel-based machine learning. They have gained wide popularity owing to the theoretical guarantees regarding performance and low computational complexity in nonlinear algorithms. Pioneered by Vapnik's support vector machines (SVMs) for classification and regression [2], kernel-based methods are nonlinear algorithms that can be adapted to an extensive class of nonlinearities. As a consequence, they have found numerous applications, including



[FIG2] Schematic illustration of the preimage problem for pattern denoising with kernel PCA. While dimensionality reduction through orthogonal projection is performed in the (b) feature space, a preimage technique is required to recover the denoised pattern in the (a) input space.

classification [3], regression [4], time-series prediction [5], novelty detection [6], image denoising [7], and bioengineering [8], to name just a few (see, e.g., [9] for a review).

REPRODUCING KERNELS AND rkHs

Originally proposed by Aizerman et al. in [10], the kernel trick provides an elegant mathematical means to derive powerful nonlinear variants of classical linear techniques. Most well-known statistical (linear) techniques can be formulated as an inner product between pairs of data. Thus, applying any nonlinear transformation to the data can only impact the values of the resulting inner products. Therefore, one does not need to compute such a transformation explicitly for a large class of nonlinearities. Instead, one only needs to replace the inner product operator with an appropriate kernel, i.e., a symmetric hermitian function. The only restriction is that the latter defines an inner product in some space. A sufficient condition for this is ensured by Mercer's theorem [11], which may be stated as follows: any positive semidefinite kernel can be expressed as an inner product in some space, where the positive semidefiniteness of a kernel $\kappa: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is determined by the property

$$\sum_{i,j} \alpha_i \, \alpha_j \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0,$$

for all $\alpha_i, \alpha_j \in \mathbb{R}$ and $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. Furthermore, the Moore-Aronszajn theorem [12] states that, to any positive semidefinite kernel κ corresponds a unique rkHs, whose inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, usually called reproducing kernel, is κ itself.

The one-to-one correspondence between rkHs and positive semidefinite functions has proved to be quite useful in numerous fields (see [13] and references therein). Since the pioneering work of Aronszajn [12], reproducing kernels and rkHs formalism have been increasingly used, especially, after being selected for the resolution of interpolation problems by Parzen [14], Kailath [15], and Wahba [16]. An rkHs is a Hilbert space of functions for which point evaluations are bounded and where the existence and uniqueness of the reproducing kernel are guaranteed by the Riesz representation theorem. In fact, let $\mathcal{H}$ be a Hilbert space of functions defined on some compact $\mathcal{X}$, for which the evaluation $\psi(\boldsymbol{x})$ of the function $\psi \in \mathcal{H}$ is bounded for all $\boldsymbol{x} \in \mathcal{X}$. By this theorem, there exists a unique function $\phi(\boldsymbol{x}) \in \mathcal{H}$ such as $\psi(\boldsymbol{x}) = \langle \psi, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}}$. Also denoted $\kappa(\cdot\,, \boldsymbol{x})$, this function has the following popular property:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}}, \tag{5}$$

for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. Moreover, the distances can be easily evaluated using the kernel trick, since the distance between two elements can be given using only kernel values, with

$$\begin{aligned} \|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\|_{\mathcal{H}}^2 &= \langle \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j), \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}} \\ &= \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\,\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) + \kappa(\boldsymbol{x}_j, \boldsymbol{x}_j), \end{aligned} \tag{6}$$

where $\| \cdot \|_{\mathcal{H}}$ denotes the norm in the rkHs.

The inherent modularity of reproducing kernels allows scaling-up linear algorithms into nonlinear ones, adapting kernel-based machines to tackle a large class of nonlinear tasks. Kernels are commonly defined on vectorial spaces, $\mathcal{X}$ endowed with the Euclidean inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \boldsymbol{x}_i^\top \boldsymbol{x}_j$ and the associated norm $\|\boldsymbol{x}_i\|$. They can be easily adapted to operate on images, e.g., in face recognition or image denoising. They are not restricted to vectorial inputs but can be naturally designed to measure similarities between sets, graphs, strings, and text documents [9]. As illustrated in Table 1, most of the kernels used in the machine-learning literature can be divided into two categories: projective kernels are functions of inner product, such as the polynomial kernel, and radial kernels (also known by isotropic kernels) are functions of distance, such as the Gaussian kernel. These kernels implicitly map the data into a high-dimensional space, even infinite dimensional for the latter.

[TABLE 1] COMMONLY USED KERNELS IN MACHINE LEARNING, WITH PARAMETERS $c > 0$, $p \in \mathbb{N}_+$, AND $\sigma > 0$.

KERNELS	EXPRESSIONS
PROJECTIVE	
MONOMIAL	$(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)^p$
POLYNOMIAL	$(c + \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)^p$
EXPONENTIAL	$\exp(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle / 2\sigma^2)$
SIGMOID (PERCEPTRON)	$\tanh(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle / \sigma + c)$
RADIAL	
GAUSSIAN	$\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / 2\sigma^2)$
LAPLACIAN	$\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\| / 2\sigma^2)$
MULTIQUADRATIC	$\sqrt{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + c}$
INVERSE MULTIQUADRATIC	$1/\sqrt{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + c}$

THE REPRESENTER THEOREM

In machine learning, inferences are focused on the estimation of the structure of some data, based on a set of available data. Given n observations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, and eventually the corresponding labels, $y_1, y_2, \ldots, y_n$, one seeks a function that minimizes a fitness error over the data, with some control of its complexity (i.e., functional norm). To this end, we consider the rkHs associated to the reproducing kernel as the hypothesis space from which the optimal is determined. The rkHs associated to κ can be identified, modulo certain details, with a space of functions defined by a linear combination of the functions $\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_n)$. Its flexibility efficiently allows for solving optimization problems, owing to the (generalized) representer theorem. Originally derived by Kimeldorf and Wahba for splines in [17], it was recently generalized to kernel-based machine learning in [18], including SVM and kernel PCA, as follows in Theorem 1.

THEOREM 1 (REPRESENTER THEOREM)

For any function $\psi \in \mathcal{H}$ minimizing a regularized cost function of the form

$$\sum_{i=1}^{n} f\big(y_i, \psi(\boldsymbol{x}_i)\big) + \eta\, g(\|\psi\|_{\mathcal{H}}^2),$$

with $f(\cdot,\cdot)$ some loss function and $g(\cdot)$ a strictly monotonic increasing function on $\mathbb{R}_+$, can be written as an image expansion in terms of the available data, namely,

$$\psi = \sum_{i=1}^{n} \alpha_i \, \phi(\boldsymbol{x}_i). \tag{7}$$

This theorem shows that, even in an infinite-dimensional rkHs, one only needs to work in the subspace spanned by the n images of the training data.

Before we proceed further, we examine the effectiveness of this theorem on two machine-learning techniques: first, consider the kernel PCA where the projected variance is maximized, namely $\psi_1, \psi_2, \ldots, \psi_k = \arg\max_{\psi} (1/n) \sum_i |\langle \boldsymbol{x}_i, \psi \rangle|^2$, under the orthonormality constraint $\langle \psi_\ell, \psi_{\ell'} \rangle_{\mathcal{H}} = \delta_{\ell\ell'}$ for all $\ell, \ell' = 1, 2, \ldots, k$. As derived in the introductory example, one only needs to solve the eigenproblem (3), involving only n unknowns for each principal axes. These unknowns correspond to the weighting coefficients in the expansion (7). Second, we consider a regression problem known as ridge regression. In this case, the mean squared error is minimized, with

$$\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} |y_i - \psi(\boldsymbol{x}_i)|^2 + \eta \|\psi\|_{\mathcal{H}}^2, \tag{8}$$

where the first term is the fitness error while the second one controls the complexity of the solution (known as Tikhonov regularization). By substituting (7) into (8), we get the optimization problem

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \eta \, \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha},$$

with $\boldsymbol{\alpha} = [\alpha_1 \, \alpha_2 \, \cdots \, \alpha_n]^\top$ and $\boldsymbol{y} = [y_1 \, y_2 \, \cdots \, y_n]^\top$. The optimal weighting coefficients are obtained by solving the linear system

$$(\boldsymbol{K} + \eta \boldsymbol{I}) \, \boldsymbol{\alpha} = \boldsymbol{y}, \tag{9}$$

where $\boldsymbol{I}$ is the identity matrix.

Such models as a sum of basis functions have been extensively studied in the literature, for instance, in interpolation problems [19], and more recently, in machine learning [20]. To illustrate this theorem, take for instance, the Gaussian kernel investigated in [21] for interpolation in two dimensions (2-D). For this kernel, we can think about the map $\phi(\boldsymbol{x}_i): \boldsymbol{x}_i \mapsto \exp(-\|\cdot - \boldsymbol{x}_i\|^2 / 2\sigma^2)$ that transforms each input data into a Gaussian bump centered on that point. Clearly, the representer theorem (Theorem 1) states that the optimal solution is a linear combination of Gaussians centered on the available input data. However, it is well known that the sum of Gaussians centered at different points cannot be written as a single Gaussian. Thus, the solution ψ in (7) cannot be a Gaussian sitting on some arbitrary data; in other words, it is not a valid image of some $\boldsymbol{x} \in \mathcal{X}$, using the map $\phi(\cdot)$ associated to the Gaussian kernel. Finding an input $\boldsymbol{x}^*$ whose image can approximate the function ψ is the preimage problem.

SOLVING THE PREIMAGE PROBLEM

A problem is ill posed if at least one of the following three conditions that characterize the well-posed problems in the sense of Hadamard is violated: 1) a solution exists, 2) it is unique, and 3) it continuously depends on the data (also known as stability condition). Unfortunately, identifying the preimage is generally an ill-posed problem. This is an outcome of the higher dimensionality of the feature space compared with the input space. As a consequence, most elements of ψ in the rkHs might not have a preimage in the input space, i.e., there may not exist an $\boldsymbol{x}^*$ such that $\phi(\boldsymbol{x}^*) = \psi$. Moreover, even if $\boldsymbol{x}^*$ exists, it may not be unique. To circumvent this difficulty, one seeks an approximate solution, i.e., $\boldsymbol{x}^*$ whose map $\phi(\boldsymbol{x}^*)$ is as close as possible to ψ.

Consider a pattern ψ in the feature space $\mathcal{H}$, obtained by any kernel-based machine, e.g., a principal axe or a denoised pattern obtained from kernel PCA. By virtue of Theorem 1, let $\psi = \sum_{i=1}^{n} \alpha_i \, \phi(\boldsymbol{x}_i)$. The preimage problem consists of the following optimization problem

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} \left\| \sum_{i=1}^{n} \alpha_i \, \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}) \right\|_{\mathcal{H}}^2. \tag{10}$$

Equivalently, from the kernel trick, $\boldsymbol{x}^*$ minimizes the objective function

$$\Xi(\boldsymbol{x}) = \kappa(\boldsymbol{x}, \boldsymbol{x}) - 2 \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}, \boldsymbol{x}_i), \tag{11}$$

where the term independent of $\boldsymbol{x}$ has been dropped.

As opposed to this functional formalism, one may also adopt a vectorwise representation, with elements in the rkHs given by their coordinates with respect to an orthogonal basis. Taking, for instance, the basis defined by the kernel PCA, as given in (4), each $\psi \in \mathcal{H}$ is represented vectorwise with $[\langle \psi, \psi_1 \rangle \, \langle \psi, \psi_2 \rangle \, \cdots \, \langle \psi, \psi_k \rangle]^\top$, thus defining a k-dimensional representation. In such a case, the Euclidean distance between the latter and the one obtained from the image of $\boldsymbol{x}^*$ is minimized. This is essentially a classical dimensionality-reduction problem, connecting the preimage problem to the historical evolution of dimensionality-reduction techniques. This is emphasized next, providing a survey on a large variety of methods.

THE EXACT PREIMAGE WHEN IT EXISTS

Suppose that there exists an exact preimage of ψ, i.e., $\boldsymbol{x}^*$ such that $\phi(\boldsymbol{x}^*) = \psi$, then the optimization problem in (10) results into that preimage. Furthermore, the preimage can be easily computed when the kernel is an invertible function of $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, such as some projective kernels including the polynomial kernel with odd degree and the sigmoid kernel (see Table 1). Let $h: \mathbb{R} \to \mathbb{R}$ define the inverse function such that $h(\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. Then, given any orthonormal basis in the input space $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N\}$, every element $\boldsymbol{x} \in \mathcal{X}$ can be written as

$$\boldsymbol{x} = \sum_{j=1}^{N} \langle \boldsymbol{e}_j, \boldsymbol{x} \rangle \, \boldsymbol{e}_j = \sum_{j=1}^{N} h(\kappa(\boldsymbol{e}_j, \boldsymbol{x})) \, \boldsymbol{e}_j.$$

As a consequence, the exact preimage x^* of some pattern $\psi = \sum_{i=1}^{n} \alpha_i \phi(x_i)$, namely $\phi(x^*) = \psi$, can be expanded as

$$x^* = \sum_{j=1}^{N} h\Big(\sum_{i=1}^{n} \alpha_i \kappa(e_j, x_i)\Big) e_j.$$

THE STRUCTURE OF THE KERNEL FUNCTIONS PROVIDES USEFUL INSIGHTS TO DERIVE MORE APPROPRIATE OPTIMIZATION TECHNIQUES.

Likewise, when the kernel is an invertible function of the distance, such as radial kernels, a similar expression can be derived by using the polarization identity $4\langle x^*, e_j\rangle = \|x^* + e_j\|^2 - \|x^* - e_j\|^2$ [22].

Clearly, such a simple derivation for the preimage is only valid under the crucial assumption that the preimage x^* exists. Unfortunately, for a large class of kernels, there are no exact preimages. Rather than seeking the exact preimage, we consider an approximate preimage by solving the optimization problem in (10). In what follows, we present several strategies for solving this problem. We first review the techniques based on classical optimization schemes and then present learning-based techniques by incorporating additional prior information.

GRADIENT DESCENT TECHNIQUES

Gradient descent is one of the simplest optimization techniques. It requires computing the gradient of the objective function (11), denoted as $\nabla_x \Xi(x^*)$. In its simplest form, the current guess x_t^* is updated into x_{t+1}^* by stepping into the direction opposite to the gradient, with

$$x_{t+1}^* = x_t^* - \eta_t \nabla_x \Xi(x_t^*),$$

where η_t is a step size parameter, often optimized using a line-search procedure. As an alternative to the gradient descent, one may use more sophisticated techniques, such as Newton's method. Unfortunately, the objective function is inherently nonlinear and clearly nonconvex. Thus, a gradient descent algorithm must be run many times with several starting values, hoping that a feasible solution will be among the local minima obtained over the runs.

FIXED-POINT ITERATION METHOD

The structure of the kernel functions provides useful insights to derive more appropriate optimization techniques beyond classical gradient descent. More precisely, the gradient of expression (11) has a closed-form expression for most kernels. By setting this expression to zero, this greatly simplifies the optimization scheme, resulting in a fixed-point iterative technique. Taking for instance the Gaussian kernel [7], the objective function in (11) becomes

$$-2\sum_{i=1}^{n} \alpha_i \exp(-\|x - x_i\|^2/2\sigma^2),$$

with its gradient

$$\nabla_x \Xi(x) = -\frac{2}{\sigma^2}\sum_{i=1}^{n} \alpha_i \exp(-\|x - x_i\|^2/2\sigma^2)(x - x_i).$$

We get the preimage by setting this gradient to zero, which results in a fixed-point iterative expression

$$x_{t+1}^* = \frac{\sum_{i=1}^{n} \alpha_i \kappa(x_t^*, x_i) x_i}{\sum_{i=1}^{n} \alpha_i \kappa(x_t^*, x_i)},$$

with $\kappa(x_t^*, x_i) = \exp(-\|x_t^* - x_i\|^2/2\sigma^2)$. Similar expressions can be derived for most kernels, such as the polynomial kernel of degree p [23] with

$$x_{t+1}^* = \sum_{i=1}^{n} \alpha_i \left(\frac{\langle x_t^*, x_i\rangle + c}{\langle x_t^*, x_t^*\rangle + c}\right)^{p-1} x_i.$$

Unfortunately, the fixed-point iterative technique still suffers from local minima and tends to be unstable. The numerical instability especially occurs when the value of the denominator decreases to zero. To prevent this situation, a regularized solution can be easily formulated, as studied in [24].

An interesting fact about the fixed-point iterative method is that the resulting preimage lies in the span of the available data, taking the form $x^* = \sum_i \beta_i x_i$ for some coefficients $\beta_1, \beta_2, \ldots, \beta_n$ to be determined. Thus, the search space is controlled, as opposed to gradient-descent techniques that explore the entire space. We further exploit information from available training data and their mapped counterparts, as discussed later.

LEARNING THE PREIMAGE MAP

To find the preimage map, a learning machine is constructed, with training elements from the feature space and estimated values in the input space, as follows: we seek to estimate a function Γ^* with the property that $\Gamma^*(\phi(x_i)) = x_i$, for $i = 1, 2, \ldots, n$. Then, ideally, $\Gamma^*(\psi)$ should give x^*, the preimage of ψ. To make the problem computationally tractable, two issues are considered in [25] and [26]. First, the function is defined on a vector space. This can be done by representing vectorwise any $\psi \in \mathcal{H}$ with $[\langle\psi, \psi_1\rangle \langle\psi, \psi_2\rangle \cdots \langle\psi, \psi_k\rangle]^\top$, using an orthogonal basis obtained from kernel PCA. Second, the preimage map Γ^* is decomposed into $\dim(\mathcal{X})$ functions to estimate each component of x^*. From these considerations, we seek functions $\Gamma_1^*, \Gamma_2^*, \ldots, \Gamma_{\dim(\mathcal{X})}^*$, with $\Gamma_m^*: \mathbb{R}^k \to \mathbb{R}$. Each of these functions is obtained by solving the optimization problem

$$\Gamma_m^* = \arg\min_{\Gamma} \sum_{i=1}^{n} f([x_i]_m, \Gamma(\psi)) + \eta g(\|\Gamma\|^2),$$

where $f(\cdot, \cdot)$ is some loss function and $[\cdot]_m$ denotes the mth component operator. By taking for instance the distance as a loss function, we get

$$\Gamma_m^* = \arg\min_{\Gamma} \frac{1}{n}\sum_{i=1}^{n} \big|[x_i]_m - \Gamma(\psi)\big|^2 + \eta\|\Gamma\|^2.$$

This optimization problem can be easily solved by a matrix-inversion scheme in analogy to the ridge-regression problem (8)

and its linear system (9). This learning approach is further investigated in the literature, incorporating neighborhood information [27] and regularization with a penalized learning [28]. All these methods are based on a set of available data in the input space and the associated images in the rkHs. The method discussed next carries this concept further by exploring pairwise distances in both spaces.

AN INTERESTING FACT ABOUT FIXED-POINT ITERATIVE METHOD IS THAT THE RESULTING PREIMAGE LIES IN THE SPAN OF THE AVAILABLE DATA.

MULTIDIMENSIONAL SCALING-BASED TECHNIQUES

As illustrated in the earlier preimage-learning approach, the preimage map seeks data in the input space based on their associated images in the rkHs. Essentially, this is a low-dimensional embedding of objects from a high-dimensional space. This problem has received a lot of attention in multivariate statistics under the framework of multidimensional scaling (MDS) [29]. MDS techniques mainly embed data in a low-dimensional space by preserving pairwise distances. This approach has been applied with success to solve the preimage problem [23]. Consider each distance in the rkHs $\delta_i = \|\psi - \phi(\boldsymbol{x}_i)\|_{\mathcal{H}}$ and its counterpart in the input space $\|\boldsymbol{x}^* - \boldsymbol{x}_i\|$. Ideally, these distances are preserved, namely

$$\|\boldsymbol{x}^* - \boldsymbol{x}_i\|^2 = \|\psi - \phi(\boldsymbol{x}_i)\|_{\mathcal{H}}^2, \tag{12}$$

for every $i = 1, 2, \ldots, n$. It is easy to verify that if there exists an i such that $\psi = \phi(\boldsymbol{x}_i)$, then we get the preimage $\boldsymbol{x}^* = \boldsymbol{x}_i$ (Figure 3).

One way to solve this problem is to minimize the mean-square error between these distances, with

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{n} \left| \|\boldsymbol{x} - \boldsymbol{x}_i\|^2 - \|\psi - \phi(\boldsymbol{x}_i)\|_{\mathcal{H}}^2 \right|^2.$$

To solve this optimization problem, a fixed-point iteration method is proposed by setting the gradient of the aforementioned expression to zero, resulting in the expression

$$\boldsymbol{x}^* = \frac{\sum_{i=1}^{n} (\|\boldsymbol{x}^* - \boldsymbol{x}_i\|^2 - \delta_i^2)\boldsymbol{x}_i}{\sum_{i=1}^{n} (\|\boldsymbol{x}^* - \boldsymbol{x}_i\|^2 - \delta_i^2)}.$$

Another approach to solve this problem is to separately consider the identities (12), resulting in n equations

$$2\langle \boldsymbol{x}^*, \boldsymbol{x}_i \rangle = \langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle + \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle - \delta_i^2,$$

for $i = 1, 2, \ldots, n$. In these expressions, the unknown also appears on the right-hand side, with $\langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle$. This unknown quantity can be easily identified in the case of centered data, since taking the average of both sides results in

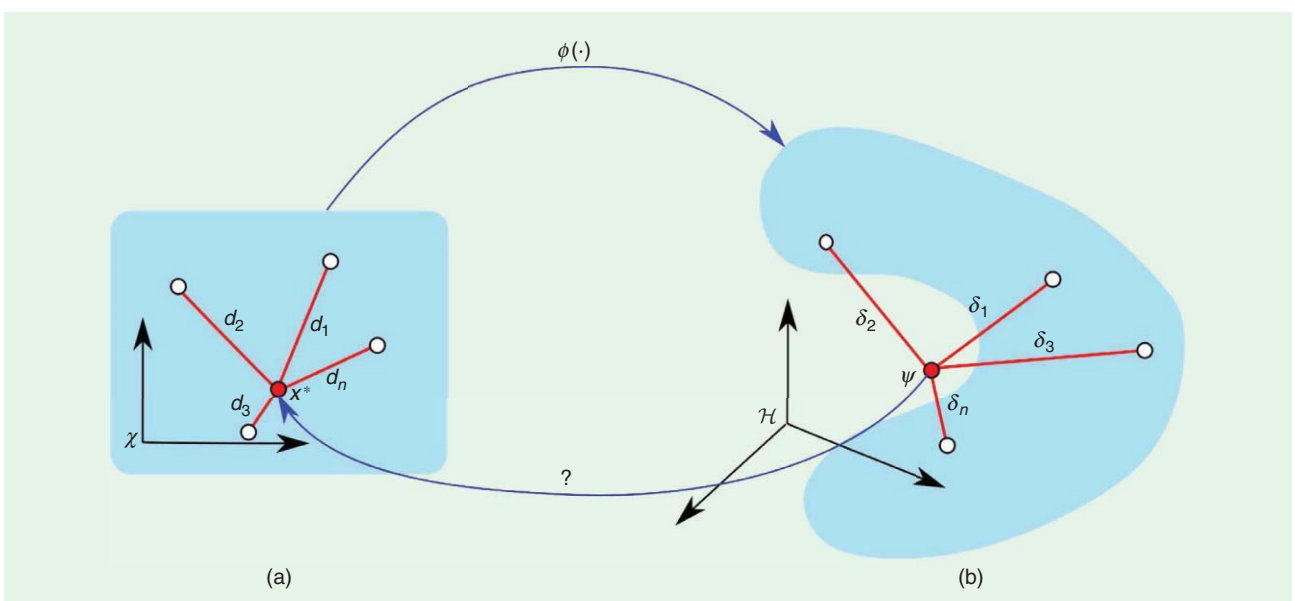
$$\langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle = \frac{1}{n} \sum_{i=1}^{n} \left(\delta_i^2 - \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle \right).$$

Let ϵ be the vector having all its entries equal to $(1/n) \sum_{i=1}^{n} (\delta_i^2 - \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle)$ then, in matrix form, we have

$$2X^\top \boldsymbol{x}^* = \operatorname{diag}(X^\top X) - \left[\delta_1^2 \, \delta_2^2 \, \cdots \, \delta_n^2\right]^\top + \epsilon,$$

where $X = [\boldsymbol{x}_1 \, \boldsymbol{x}_2 \, \cdots \, \boldsymbol{x}_n]$ and $\operatorname{diag}(\cdot)$ is the diagonal operator with $\operatorname{diag}(X^\top X)$ the column vector with entries $\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle$. The unknown preimage is obtained using the least-squares solution, namely

$$\boldsymbol{x}^* = \frac{1}{2} (XX^\top)^{-1} X \left(\operatorname{diag}(X^\top X) - [\delta_1^2 \, \delta_2^2 \, \cdots \, \delta_n^2]^\top \right),$$



[FIG3] Schematic illustration of the MDS-based technique where the preimage is identified from pairwise distances in both (a) input and (b) feature spaces.

where the term $(XX^\top)^{-1}X\epsilon$ becomes zero, thanks to the assumption of centered data.

To keep this technique tractable in practice, only a certain neighborhood is considered in the preimage estimation, in the same spirit as the locally linear embedding scheme in dimensionality reduction [30]. This approach opened the door to a range of other techniques, borrowed from dimensionality reduction and manifold learning literature [31].

CONFORMAL MAP APPROACH

In addition to the distance-preserving method of MDS, one may also propose a preimage method by preserving inner product measures. Using such a strategy, the angular measure is also preserved, since $\boldsymbol{x}_i^\top \boldsymbol{x}_j / \|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|$ defines the cosine of the angle between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the Euclidean input space. For this reason it is called the conformal map approach. A recent technique to solve the preimage problem based on the conformal map has been presented in [32]. To this end, a coordinate system in the rkHs is constructed with an isometry with respect to the input space. We emphasize the fact that the model is not coupled with any constraint on the coordinate functions, as opposed to the orthogonality between the functions resulting from the kernel PCA.

By virtue of Theorem 1, each of the n coordinate functions can be written as a linear expansion of the available images, namely $\Psi_\ell = \sum_{i=1}^n \theta_{\ell,i}\, \phi(\boldsymbol{x}_i)$, for $\ell = 1, 2, \ldots, n$, with unknown weights to be determined, rearranged in a matrix Θ. Therefore, the coordinates of any element of the rkHs can be obtained by a projection onto these coordinate functions, thus any $\phi(\boldsymbol{x}_i)$ can be represented with the n coordinates in $\boldsymbol{\Psi}_{\boldsymbol{x}_i} = [\langle \Psi_1, \phi(\boldsymbol{x}_i)\rangle \, \langle \Psi_2, \phi(\boldsymbol{x}_i)\rangle \, \cdots \, \langle \Psi_k, \phi(\boldsymbol{x}_i)\rangle]^\top$. Ideally, the inner products are preserved in both coordinate system and Euclidean input space, specifically

$$\boldsymbol{\Psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\Psi}_{\boldsymbol{x}_j} = \boldsymbol{x}_i^\top \boldsymbol{x}_j, \tag{13}$$

for all $i, j = 1, 2, \ldots, n$. This can be solved by minimizing the fitness error over all pairs,

$$\min_{\Psi_1, \ldots, \Psi_n} \sum_{i,j=1}^n \left| \boldsymbol{x}_i^\top \boldsymbol{x}_j - \boldsymbol{\Psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\Psi}_{\boldsymbol{x}_j} \right|^2 + \eta \sum_{\ell=1}^n \|\Psi_\ell\|_{\mathcal{H}}^2,$$

where the second term incorporates regularization. This can be written in matrix form as

$$\min_{\boldsymbol{\Theta}} \frac{1}{2} \|X^\top X - K\boldsymbol{\Theta}^\top \boldsymbol{\Theta} K\|_F^2 + \eta \, \mathrm{tr}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta} K),$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and $\|\cdot\|_F$ the Frobenius norm, i.e., the root of sum of squared (absolute) values of all its elements, or equivalently $\|M\|_F^2 = \mathrm{tr}(M^\top M)$. By taking the derivative of this expression with respect to $\boldsymbol{\Theta}^\top \boldsymbol{\Theta}$, one obtains

$$\boldsymbol{\Theta}^\top \boldsymbol{\Theta} = K^{-1}\left(X^\top X - \eta K^{-1}\right) K^{-1}. \tag{14}$$

Now we are in a position to determine the preimage of some $\psi = \sum_{i=1}^n \alpha_i \, \phi(\boldsymbol{x}_i)$. Its coordinates associated to the system of coordinate functions $\Psi_1, \Psi_2, \ldots, \Psi_n$ are given by

$$\langle \psi, \Psi_\ell \rangle_{\mathcal{H}} = \sum_{i,j=1}^n \theta_{\ell,i}\, \alpha_j \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

for $\ell = 1, 2, \ldots, n$. By preserving the inner products in both spaces, ideally the model in (13) can be extended to ψ, resulting in

$$X^\top \boldsymbol{x}^* = K\boldsymbol{\Theta}^\top \boldsymbol{\Theta} K \boldsymbol{\alpha}.$$

By combining this expression with (14), we get the simplified expression $X^\top \boldsymbol{x}^* = (X^\top X - \eta K^{-1})\, \boldsymbol{\alpha}$, whose least square solution is

$$\boldsymbol{x}^* = (XX^\top)^{-1} X (X^\top X - \eta K^{-1})\, \boldsymbol{\alpha}.$$

It is worth noting that this expression is independent of the kernel type under investigation.

Furthermore, this technique can be easily extended to identify the preimages of a set of elements in rkHs, since the term between parentheses needs to be computed only once. In fact, this is a matrix-completion scheme like the one studied in [33]. This corresponds to completing an inner-product matrix based on another Gram matrix, the matrix of kernel values.

SCOPE OF APPLICATION OF THE PREIMAGE PROBLEM

In this section, we present some application examples that involve solving the preimage problem. Our first experiments are with kernel PCA on toy data and are mainly intended to illustrate the preimage problem. Then we provide a comparative study of the several methods presented in this article on image denoising problem. Finally, we show how the preimage can be required in other applications beyond kernel PCA. To this end, we consider a problem of autolocalization of sensors in wireless sensor networks.

SOME APPLICATIONS OF KERNEL PCA WITH PREIMAGE

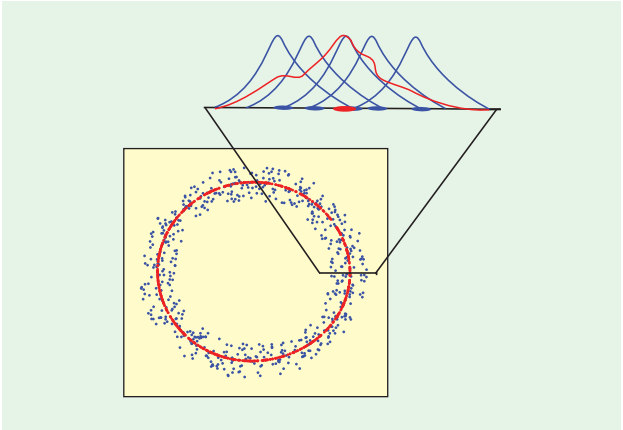
FEATURE EXTRACTION

The first illustration considered here is the use of kernel PCA on synthetic data to provide a visual illustration of PCA versus kernel PCA for feature extraction. The data distribution takes the form of a ring in 2-D, with an inner diameter of two and an outer diameter of three. Within this region, $n = 600$ training data were generated, as illustrated in Figure 4 with blue dots. To extract the most relevant feature, two methods were used: on the one hand, the conventional PCA and on the other hand kernel PCA with a preimage step. The PCA technique provided linear axes by solving the eigenvector problem and thus did not capture the circular shape of the data. This is illustrated by projecting the data onto the first principal axis, given by red curve in Figure 4(a). The kernel PCA was applied using a Gaussian kernel with bandwidth $\sigma = 2$, the principal axes being defined by a sum of n Gaussian functions in an infinite-dimensional

feature space. A preimage method was required to derive the axes, or representations of these axes, within the input space. As shown in Figure 4(b), this technique captured the nonlinear feature in the original space.

As described at the beginning of this article, when we introduced the preimage problem with the Gaussian kernel, each data is mapped into a Gaussian bump centered around it. By taking the sum of these Gaussians with some optimized weighting coefficients, we get the principal distribution whose mean, if it exists, provides the preimage. It is worth noting that the definition of a mean only exists and makes sense for Gaussian like curves and not for a sum of Gaussians centered at different points. A schematic illustration of the preimage problem is given in Figure 5, taking only a (unidirectional) radial cut in the ring-distributed data. The data obtained by solving the preimage problem can be interpreted as the center of the distribution Gaussian that best approximates the sum of Gaussians.

In this application, a fixed-point iterative technique was used. Next, we give a comparative study of several techniques given in this article by considering the image-denoising problem.



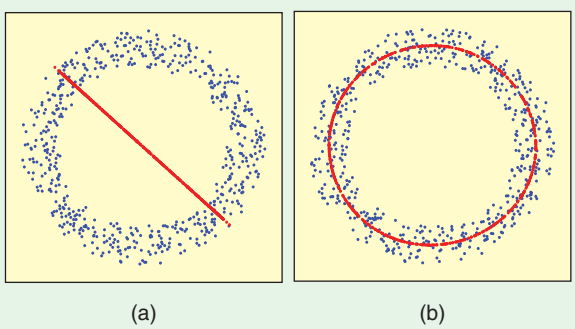
[FIG5] Schematic illustration of the preimage problem with the Gaussian kernel, where the profile corresponds to a radial cut in the ring-distributed data. From the sum of Gaussians (red curve), the preimage corresponds to the mean value of the distribution (blue dots).

IMAGE DENOISING

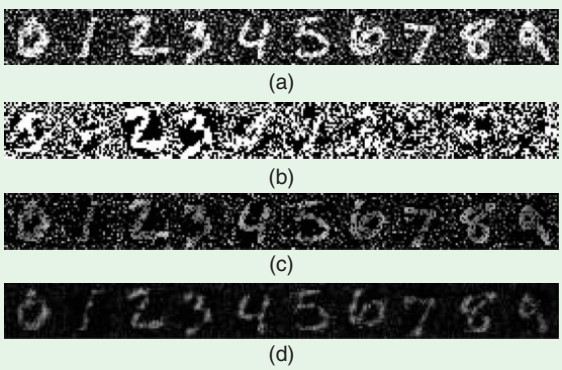
In this section, we illustrate the results obtained in a problem of real-image denoising, using three techniques: the fixed-point iterative method, MDS-based technique, and conformal map approach. The images consist of the modified National Institute of Standards and Technology (NIST) database of handwritten digits [34], corresponding to handwritten digits, from 0 to 9, in (almost) binary 28-by-28 pixels. From a machine-learning point of view, each image can be represented as a point in a 28×28 dimensional space. The original images were corrupted by adding a zero-mean white Gaussian noise with variance 0.2. In the training stage, a set of 1,000 images, 100 of each digit, were used to train the kernel PCA, retaining only 100 leading principal axes. We used the Gaussian kernel for the three algorithms, with the bandwidth set to $\sigma = 10^5$.

To illustrate the ability of this method for image denoising, another set of ten images, one for each digit, was considered under the same noise conditions. These images are illustrated in Figure 6(a), with the results obtained with the (b) fixed-point iterative, (c) MDS-based, and (d) conformal methods. For such applications, the fixed-point iterative algorithm was found to be inappropriate, even with a large number of iterations (here 10,000 iterations were used). To take advantage of prior knowledge, the same training data set was used for learning the reverse map. Realistic results were obtained using the MDS-based method. It is obvious that the conformal algorithm achieved better denoised results. For this simulation, the regularization parameter was set to $\eta = 10^{-9}$.

In an attempt to provide a measure of computational requirements, we considered the (average) total CPU time of each algorithm. These algorithms were implemented on a MATLAB running on a MacBook Pro Duo Core to offer a comparative study. With 10,000 iterations, the fixed-point iterative algorithm required a total CPU time of up to 1 h. The MDS-based and conformal algorithms required 5 min and 1.5 s, respectively.



[FIG4] Denoising data distributed on a ring, using (a) classical PCA and (b) kernel PCA with preimage. The extracted feature is (a) linear and (b) circular.



[FIG6] Application to handwritten digit denoising with kernel PCA, using several preimage methods presented in this article. (a) Ten digits corrupted by noise. (b) Fixed-point iterative method. (c) MDS-based method. (d) Conformal method.

AUTOLOCALIZATION IN WIRELESS SENSOR NETWORKS

With recent technological advances in both electronics and wireless communications, low-power and low-cost tiny sensors have been developed for monitoring physical phenomena and tracking applications. Densely deployed in the inspected environment with efficiently designed distributed algorithms, wireless ad hoc networks seem to offer several opportunities. They were successfully employed in many situations, ranging from military applications such as battleground supervision to civilian applications such as habitat monitoring and healthcare surveillance (see [35], [36] and references therein). While these sensors are often randomly deployed, e.g., for monitoring inhospitable habitats and disaster areas, information captured by each sensor remains obsolete as long as it stays unaware of its location. Implementing a self-localization device, such as a global positioning system receiver, at each sensor device may be too expensive and too power hungry for the desired application with battery-powered devices. As a consequence, only a small fraction of the sensors may be location aware, the so-called anchors or beacons. The other sensors have to estimate their locations by exchanging some information with its neighbors.

For this purpose, each sensor determines a ranging (distance) with other sensors, from intersensor measurements such as the received signal strength indication (RSSI), the connectivity, the hop count, and the time difference of arrival. Most methods used for autolocalization in sensor networks are based on either MDS techniques or semidefinite programming (for a survey, see [37] and [38]), identifying a function that links the ranging between sensors to their locations. However, if the data are not intersensor distances or are linked to coordinates by an unknown nonlinear function, e.g., using the RSSI measurements or the estimated covariance sensor data [39], linear techniques such as MDS and PCA fail to accurately estimate the locations. Once again, the kernel machines provide an elegant way to overcome this drawback.

Here, we describe the method proposed in [40]. The main idea can be described in three stages. In the first stage, we construct the reproducing kernel and its associated rkHs that best describes the anchor pairwise similarities. In the second stage, a nonlinear manifold is designed from similarities between anchor–sensor measurements by applying the kernel PCA technique. The final stage consists of estimating the coordinates of nonanchor sensors by applying a preimage technique on their projections onto the manifold. Next, we describe these three stages before presenting the experimental results.

Consider a network of N sensor nodes, with n location-aware anchors and $N-n$ sensors of unknown location, living in a p-dimensional space, e.g., $p=2$ for localization in a plane. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be the coordinates of the ith sensor, rearranged such that indices $i=1,2,\ldots,n$ correspond to anchors. Let $\tilde{K}(i,j)$ be the intersensor similarity between sensors i and j, such as RSSI.

KERNEL SELECTION FROM INTERANCHOR SIMILARITIES

As a model of similarity measurements, the appropriate reproducing kernel should be chosen and tuned up, which allows a physical meaning of the results obtained from the kernel PCA (next stage). The alignment criterion [41] provides a measure of similarity between the reproducing kernel and target function, e.g., between a Gaussian kernel and RSSI measurements. Maximizing the alignment $\mathcal{A}(K,\tilde{K})$ provides the optimal-reproducing kernel, faithful to the interanchor measurements, where

$$\mathcal{A}(K,\tilde{K}) = \frac{\langle K,\tilde{K}\rangle_F}{\sqrt{\langle K,K\rangle_F \langle \tilde{K},\tilde{K}\rangle_F}},$$

with $\langle\cdot,\cdot\rangle_F$ as the Frobenius inner product between two matrices. Taking, for instance, the Gaussian kernel, the optimization problem is reduced to finding the optimal bandwidth. In practice, this optimization problem is solved at each anchor, using only the information from its neighborhood.

KERNEL PCA UPON ANCHORS

After identifying the reproducing kernel adapted to the measurements, a kernel-PCA approach is applied to provide the most relevant subspace of the associated rkHs. Classical kernel PCA is computed by a diagonalization scheme, which may be computationally expensive for in-network processing. An alternative approach can be done using an iterative scheme, such as the kernel-Hebbian algorithm [42] (we refer the reader to [40] for its implementation in wireless sensor networks).

PREIMAGE FOR LOCATION ESTIMATION

For each sensor, we represent its image in the rkHs associated to the kernel, maximizing the alignment criterion. The image is projected onto the manifold, obtained using kernel PCA with anchor pairwise similarities. The problem of estimating the coordinates from that representation is the preimage problem.

EXPERIMENTAL RESULTS

The first batch of experiments was carried out on simulated measurements. For this purpose, we considered a network of sensors measuring some physical phenomena, e.g., temperature, atmospheric pressure, or luminance. In a static field, we assumed that measurements were jointly generated from a normal distribution, with decreasing correlations between measurements as a function of the distance between sensors. This information was used as a local similarity measure between sensors [39]. More precisely, we considered the spherical model, commonly used in environmental and geological sciences [43], defined by a covariance of the form $\zeta(\|\boldsymbol{x}_i-\boldsymbol{x}_j\|)$ with

$$\zeta(u) = \begin{cases} 1-\frac{3}{2d}u+\frac{1}{2d^3}u^3 & \text{for } 0\le u\le d; \\ 0 & \text{for } d<u, \end{cases}$$

where d denotes the cutoff distance, and fixed to $d = 60$ in our experiments. The profile of the spherical model is illustrated in Figure 7. The experiments consisted of 100 sensors, from which 20 were anchors with known locations, randomly spread over a 100-by-100 square region. For each sensor, 200 measurements were collected, and the Gaussian kernel was considered. Figure 8 illustrates the localization results obtained with this method.

In a second experiment, real measurements of RSSI were collected from an indoor experiment at the Motorola facility in Plantation, Florida. The environment is a 14-by-13 m office area, partitioned by cubicle walls (height = 1.8 m). The network consisted of 40 unknown-location sensors and four anchors near the corners. The experimental settings are described more in detail in [44] (see also http://www.eecs.umich.edu/~hero/localize/). For each sensor i, we collected the RSSI associated to it in a 44-dimensional vector, denoted by $\boldsymbol{u}_i$. The intersensor similarity between sensors is given by the matrix $\tilde{K}$, defined between sensors i and j by

$$\tilde{K}(i,j) = \exp\left(-\|\boldsymbol{u}_i - \boldsymbol{u}_j\|^2/200\right).$$

The Gaussian kernel was considered, with its bandwidth optimized by maximizing the alignment. The proposed method gives a root-mean-square location error over the 40 sensors of 2.13 m each. This should be compared to the maximum-likelihood estimator studied in [44] (that turned out to be biased), having a root-mean-square location error of 2.18 m.

FINAL REMARKS

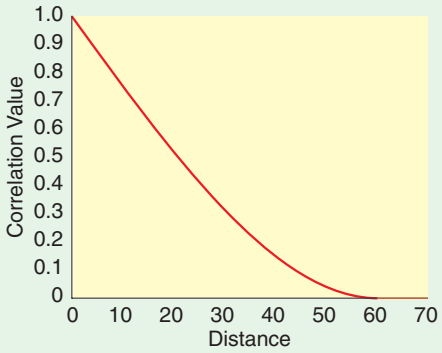
This article presented the preimage problem in machine learning, providing an overview of the state-of-the-art methods and approaches for solving such a problem. Our aim was to show how this problem is intimately related to dimensionality reduction issues, borrowing and enhancing ideas derived from dimensionality reduction and manifold learning. Throughout this article, we studied this problem for kernel PCA and provided a comparative study of several methods for image denoising. We extended the range of application of the preimage problem to another context, sensor autolocalization in wireless sensor networks.

By interpreting the processing in the feature space to the original input space, this strategy opens the way to a range of diverse signal-processing problems. These problems are nonlinear kernel-based formulations of classical signal processing methods, including the independent component analysis [45] and the Kalman filter [46]. Another area of application is the preimage problem on structured spaces, including biological sequence analysis in bioinformatics [47] and string analysis in natural language [48]. In the latter, the authors derived a preimage solution for a string kernel, using a graph-theoretical formulation. All these promising areas of application of the preimage problem open an avenue for future work.

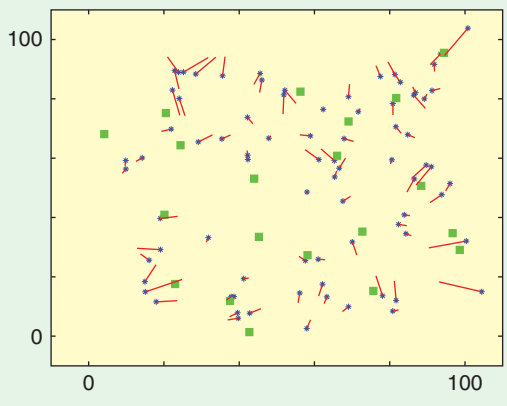
AUTHORS

Paul Honeine (paul.honeine@utt.fr) received his Dipl.-Ing. degree in mechanical engineering in 2002 and his M.Sc. degree in industrial control in 2003, both from Lebanese University in Lebanon. In 2007, he received the Ph.D. degree from the University of Technology of Troyes, France. Since September 2008, he has been an assistant professor at the Institut Charles Delaunay (UMR CNRS 6279) at the University of Technology of Troyes, France. He is the coauthor of the 2009 Best Paper Award at the IEEE Workshop on Machine Learning for Signal Processing (MLSP). His research interests include nonstationary signal analysis, nonlinear adaptive identification, and machine learning.

Cédric Richard (cedric.richard@unice.fr) is a full professor at Observatoire de la Côte d'Azur, University of Nice, Sophia-Antipolis, France. He is a junior member of the Institut Universitaire de France. He was an associate editor for *IEEE Transactions on Signal Processing* (2006–2010). He is a member of the Signal Processing Theory and Methods Technical Committee of the IEEE Signal Processing Society and is the author of more than 100 papers. He received the 2009 Best Paper Award at the IEEE Workshop on MLSP. His current research interests include statistical signal processing and machine learning.



[FIG7] Profile of the spherical model as a function of the distance. The cutoff distance is set to $d = 60$.



[FIG8] Estimated locations of 80 sensors (asterisk) based on 20 anchors of known positions (green squares), with error to real position represented by a line (—).

REFERENCES

[1] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[2] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[3] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. Müller, "Fisher discriminant analysis with kernels," in *Advances in Neural Networks for Signal Processing*, Y. H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. San Mateo, CA: Morgan Kaufmann, 1999, pp. 41–48.

[4] R. Rosipal and L. Trejo, "Kernel partial least squares regression in reproducing kernel hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, Dec. 2002.

[5] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Trans. Signal Processing*, vol. 57, no. 3, pp. 1058–1067, Mar. 2009.

[6] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA, MIT Press, vol. 12, 2000.

[7] S. Mika, B. Schölkopf, A. Smola, K. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," in *Proc. 1998 Conf. Advances in Neural Information Processing Systems II*. Cambridge, MA: MIT Press, 1999, pp. 536–542.

[8] G. Camps-Valls, J. L. Rojo-Alvarez, and M. Martinez-Ramon Manuel, Eds., *Kernel Methods in Bioengineering, Signal and Image Processing*. Hershey, PA: IGI Publishing, 2007.

[9] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge: U.K.: Cambridge Univ. Press, 2004.

[10] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automat. Remote Contr.*, vol. 25, pp. 821–837, 1964.

[11] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Roy. Soc. Lond. Philos. Trans. A*, vol. 209, pp. 415–446, Jan. 1909.

[12] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[13] D. Alpay, Ed., *Reproducing Kernel Spaces and Applications*, (ser. Operator Theory: Advances and Applications). Cambridge, MA: Birkhäuser, 2003, vol. 143.

[14] E. Parzen, "Statistical inference on time series by RKHS methods," in *Proc. 12th Biennial Seminar*, R. Pyke, Ed. Montreal, Canada: Canadian Mathematical Congress, 1970, pp. 1–37.

[15] T. Kailath, "RKHS approach to detection and estimation problems—I: Deterministic signals in gaussian noise," *IEEE Trans. Inform. Theory*, vol. 17, no. 5, pp. 530–549, Sept. 1971.

[16] G. Wahba, *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Math (SIAM), 1990.

[17] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," *J. Math. Anal. Applicat.*, vol. 33, no. 1, pp. 82–95, 1971.

[18] B. Schölkopf, R. Herbrich, and R. Williamson, "A generalized representer theorem," Royal Holloway College, Univ. London, U.K., Tech. Rep. NC2-TR-2000-81, 2000.

[19] C. A. Micchelli, "Interpolation of scattered data: Distance matrices and conditionally positive definite functions," *Construct. Approx.*, vol. 2, no. 1, pp. 11–22, Dec. 1986.

[20] F. Girosi, M. Jones, and T. Poggio, "Priors stabilizers and basis functions: From regularization to radial, tensor and additive splines," CBCL, MIT, Cambridge, MA, Tech. Rep. AIM-1430, CBCL-075, 1993.

[21] I. P. Schagen, "Interpolation in two dimensions—A new technique," *J. Inst. Math. Applicat.*, vol. 23, no. 1, pp. 53–59, 1979.

[22] B. Schölkopf, "Support vector learning," Ph.D. dissertation, Technischen Universität, Berlin, Germany, 1997.

[23] J. T. Kwok and I. W. Tsang, "The pre-image problem in kernel methods," in *Proc. 20th Int. Conf. Machine Learning (ICML)*. Washington, DC: AAAI Press, Aug. 2003, pp. 408–415.

[24] T. J. Abrahamsen and L. K. Hansen, "Input space regularization stabilizes pre-images for kernel PCA de-noising," in *Proc. IEEE Workshop Machine Learning for Signal Processing*, Grenoble, France, 2009, pp. 1–6.

[25] G. Bakir, J. Weston, and B. Schölkopf, "Learning to find pre-images," in *Neural Information Processing Systems 2003*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, vol. 16, pp. 449–456.

[26] G. Bakir, "Extension to kernel dependency estimation with applications to robotics," Ph.D. dissertation, Tech. Univ., Berlin, Germany, Nov. 2005.

[27] W.-S. Zheng and J.-H. Lai, "Regularized locality preserving learning of pre-image problem in kernel principal component analysis," in *Proc. 18th Int. Conf. Pattern Recognition (ICPR)*. Washington, DC: IEEE Computer Society, 2006, pp. 456–459.

[28] W.-S. Zheng, J. H. Lai, and P. C. Yuen, "Penalized preimage learning in kernel principal component analysis," *IEEE Trans. Neural Networks*, vol. 21, no. 4, pp. 551–570, 2010.

[29] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, 2nd ed. (ser. Monographs on Statistics and Applied Probability). London, Chapman and Hall, Sept. 2000.

[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec. 22, 2000.

[31] P. Etyngier, F. Ségonne, and R. Keriven, "Shape priors using manifold learning techniques," in *Proc. 11th IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[32] P. Honeine and C. Richard, "Solving the pre-image problem in kernel machines: A direct method," in *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, Sept. 2009, pp. 1–6.

[33] Y. Yamanishi and J.-P. Vert. (2007). Kernel matrix regression. Tech. Rep. arXiv:q-bio/0702054v1 [Online] Available: http://arxiv.org/abs/q-bio/0702054v1

[34] Y. Lecun and C. Cortes. (1998). The MNIST database of handwritten digits [Online]. Available: http://yann.lecun.com/exdb/mnist

[35] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, "Instrumenting the world with wireless sensor networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Los Alamitos, CA: IEEE Computer Society, 2001, vol. 4, pp. 2033–2036.

[36] C. S. Raghavendra and K. Sivalingam, Eds., *Proc. 2nd ACM Int. Workshop Wireless Sensor Networks and Applications*. San Diego, CA: ACM, Sept. 2003.

[37] J. Bachrach and C. Taylor, "Localization in sensor networks," *Handbook of Sensor Networks*, I. Stojmenovic, Ed. New Jersey, Wiley, pp. 277–310, 2005.

[38] G. Mao, B. Fidan, and B. Anderson, "Wireless sensor network localization techniques," *Comput. Networks*, vol. 51, no. 10, pp. 2529–2553, 2007.

[39] N. Patwari and A. O. Hero, "Manifold learning algorithms for localization in wireless sensor networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2004, vol. 3. pp. 857–860.

[40] M. Essoloh, C. Richard, H. Snoussi, and P. Honeine, "Distributed localization inwireless sensor networks as a pre-image problem in a reproducing kernel hilbert space," in *Proc. European Conf. Signal Processing (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008, pp. 1–5.

[41] N. Cristianini, A. Elisseeff, J. Shawe-Taylor, and J. Kandola, "On kernel target alignment," in *Proc. Neural Information Processing Systems (NIPS)*, 2002, pp. 367–373.

[42] K. Kim, M. Franz, and B. Schölkopf, "Iterative kernel principal component analysis for image modeling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 9, pp. 1351–1366, 2005.

[43] T. Gneiting, "Compactly supported correlation functions," Environmental Protection Agency, NRCSE, Seatle, WA, Tech. Rep. NRCSE-TRS No. 045, May 2000.

[44] N. Patwari, A. O. Hero, M. Perkins, N. S. Correal, and R. J. O'Dea, "Relative location estimation in wireless sensor networks," *IEEE Trans. Signal Processing*, vol. 51, no. 8, pp. 2137–2148, Aug. 2003.

[45] J. Yang, X. Gao, D. Zhang, and J.-Y. Yang, "Kernel ICA: An alternative formulation and its application to face recognition," *Pattern Recognit.*, vol. 38, no. 10, pp. 1784–1787, Oct. 2005.

[46] L. Ralaivola and F. D'Alché-Buc, "Time series filtering, smoothing and learning using the kernel Kalman filter," in *Proc. Int. Joint Conf. Neural Networks*, 2005, vol. 3. pp. 1449–1454.

[47] S. Sonnenburg, A. Zien, P. Philips, and G. Ratsch, "Poims: positional oligomer importance matrices—Understanding support vector machine-based signal detectors," *Bioinformatics*, vol. 24, no. 13, pp. i6–14, July 2008.

[48] C. Cortes, M. Mohri, and J. Weston, "A general regression technique for learning transductions," in *Proc. 22nd Int. Conf. Machine Learning (ICML)*. New York, ACM, 2005, pp. 153–160. [SP]

Kevin M. Carter, Raviv Raich, William G. Finn, and Alfred O. Hero, III

Information-Geometric Dimensionality Reduction

Statistical manifold of probability distributions

We consider the problem of dimensionality reduction and manifold learning when the domain of interest is a set of probability distributions instead of a set of Euclidean data vectors. In this problem, one seeks to discover a low-dimensional representation, called embedding, that preserves certain properties such as distance between measured distributions or separation between classes of distributions. This article presents the methods that are specifically designed for low-dimensional embedding of information-geometric data, and we illustrate these methods for visualization in flow cytometry and demography analysis.

DIMENSIONALITY REDUCTION

High-dimensional data visualization and interpretation have become increasingly important for data mining, information retrieval, and information discrimination applications, arising in areas such as search engines, security, and biomedicine. The explosion in sensing and storage capabilities has generated a vast amount of high-dimensional data and led to the development of many algorithms for feature extraction and visualization, known variously as dimensionality reduction, manifold learning, and factor analysis.

Dimensionality reduction strategies fall in two categories: supervised task-driven approaches and unsupervised geometry-driven approaches. Supervised task-driven approaches reduce data dimension according to the optimization of a performance criterion that depends on both reduced data and ground truth, e.g., class labels. Examples include linear discriminant analysis (LDA) [1], supervised principal components [2], and multi-instance dimensionality reduction [3]. Unsupervised geometry-driven approaches perform

Digital Object Identifier 10.1109/MSP.2010.939536
Date of publication: 17 February 2011

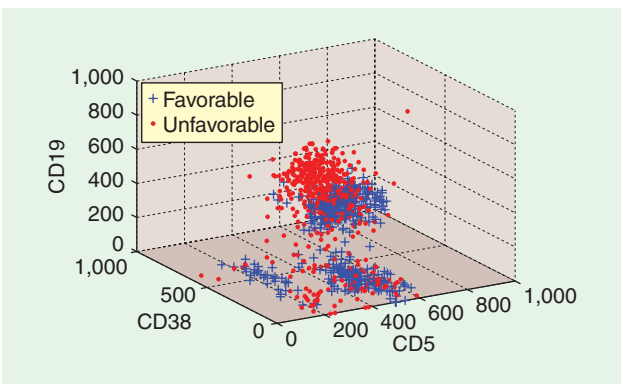
dimension reduction without ground truth and try to preserve geometric properties such as distances or angles between data points. Examples include principal component analysis (PCA), multidimensional scaling (MDS) [4], and ISOMAP [5]. Most of these approaches use Euclidean distances between sample points to drive the dimensionality-reduction algorithm.

It has been recognized that these Euclidean algorithms can be generalized to non-Euclidean spaces by replacing the Euclidean distance metric with a more general dissimilarity measure. In particular, when the data samples are probability distributions, use of an information divergence such as Kullback-Leibler (KL) instead of Euclidean distance leads to a class of information-geometric algorithms for dimensionality reduction [6], [7]. In this article, we motivate and explain the application of information-geometric dimensionality reduction (IGDR) for two real-world applications.

IGDR operates on a statistical manifold of probability distributions instead of the geometric manifold of Euclidean data points. When such distributional information can be extracted from the data, IGDR results in significant improvements in information retrieval, visualization, and classification performance [6]–[10]. This improvement can be understood from the point of view of information-theoretic bounds: information divergence is generally more relevant to statistical discrimination performance rather than Euclidean distance.

For example, for binary classification, the minimum probability of error converges to zero at an exponential rate, with the rate constant equal to KL information divergence between the distributions of the data over each class [11]. The KL divergence is not a function of the Euclidean distances between data points unless these distributions are spherical Gaussian. Therefore, as it preserves information divergence, in many cases, IGDR can produce more informative dimension reductions than classical Euclidean approaches.

Implementation of information-geometric versions of PCA, ISOMAP, and others is often not as straightforward as the Euclidean counterparts, which are frequently convex and solvable as generalized eigenvalue problems. Nonetheless, as shown in this article, the added complexity of implementation can be well worth the effort. We illustrate the power of IGDR by presenting generalizations of ISOMAP, PCA, and LDA. These implementations are called Fisher information nonparametric embedding (FINE) [6], information-preserving components analysis (IPCA) [9], and information-maximizing components analysis (IMCA) [12], respectively. Each of these algorithms solves a well-posed optimization problem over the information-geometric embedding of each sample point's distribution.



[FIG1] In clinical flow cytometry, diagnoses and prognoses are made through the analysis of high-dimensional point clouds, the measurement space of selected biomarkers.

Probability distributions and information divergence can arise as useful targets for dimensionality reduction in several ways. In image-retrieval applications, the most discriminating properties of an image may be invariants, such as the relative frequency of occurrence of different textures, colors, or edge features. The histogram of these relative frequencies is a probability distribution that is specific to the particular image, up to scale, translation, rotation, or other unimportant spatial transformations. Dimensionality reduction on these probabilities can accelerate retrieval speed, without negatively affecting precision or recall rates. Furthermore, visualization of the database, for example, as manifested by clusters of similar images, can be useful for understanding the database complexity or for comparing the different databases.

In other applications, each object in the database is itself stored as a cloud of high-dimensional points, and the shape of this point cloud is what naturally differentiates the objects. For example, in the flow cytometry application, discussed in the "Flow Cytometry" section, the objects are different patients, and the data points are vector attributes of a population of the patient's blood cells, and it is the shape of the point cloud that is of interest to the pathologist. This is demonstrated in Figure 1, where we compare the point clouds, with respect to three biomarkers, of two patients with favorable and unfavorable prognoses. Another example, discussed in the "Crime in the 1990s" section, is spatiodemographic analysis of crime data, where the analyst is interested in comparing the patterns of crime in different cities based on distributions of community and law enforcement characteristics.

All the algorithms presented here are available for download as MATLAB code on our reproducible research Web site [13].

DISTANCE ON STATISTICAL MANIFOLDS

Information geometry is a field that has emerged from the study of geometrical constructs on manifolds of probability distributions. These investigations analyze probability distributions as geometrical structures in a Riemannian space. Using tools and methods deriving from differential geometry, information geometry is applicable to information theory, probability theory, and statistics. (For a more thorough introduction on information geometry, see [14] and [15].) As most dimensionality reduction techniques are designed to either preserve pairwise sample distances (unsupervised) or maximize between-class distances (supervised), it is first necessary to understand the principles of distance in information geometry. Similar to points on a Riemannian manifold in Euclidean space, probability density functions (PDFs) that share a parameterization lie on a statistical manifold. A statistical manifold may be viewed as a set $\mathcal{M}$, whose elements are probability distributions. The coordinate system of this manifold is equivalent to the parameterization of PDFs. For example, a d-variate Gaussian distribution is entirely defined by its mean vector μ and covariance matrix Σ, leading to a $d + d(d+1)/2$-dimensional statistical

manifold, which is of a higher dimension than the dimension d of a sample realization $X \sim \mathcal{N}(\mu, \Sigma)$ from this distribution.

For a parametric family of probability distributions on a statistical manifold, it is possible to define a Riemannian metric using the Fisher information metric, which measures the amount of information a random variable contains in reference to an unknown parameter θ. This metric may then be used to compute the Fisher information distance $D_F(p_1, p_2)$ between two distributions $p(x; \theta_1)$, $p(x; \theta_2) \in \mathcal{M}$. This distance is the length of the shortest path—the geodesic—on $\mathcal{M}$, connecting coordinates θ_1 and θ_2.

Although the Fisher information distance cannot be exactly computed without a priori knowledge about the parameterization θ of the manifold, the distance between PDFs p_1 and p_2 may be approximated with a variety of pseudometrics such as KL divergence

$$\mathrm{KL}(p_1 \| p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx. \tag{1}$$

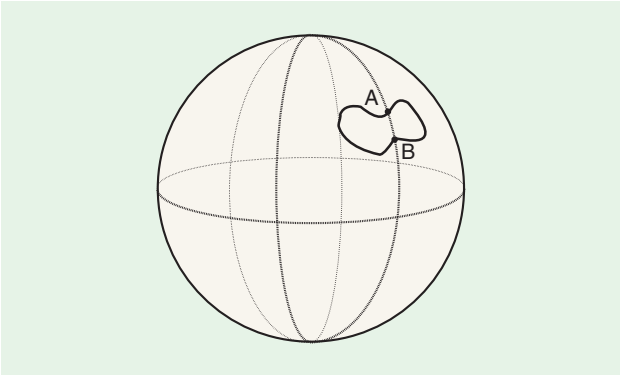
KL divergence is very important in information theory and is commonly referred to as the relative entropy of one PDF to another. As the pair of densities approaches each other, the KL divergence is a good approximation to the Fisher information distance between them [14]:

$$\sqrt{2\mathrm{KL}(p_1 \| p_2)} \to D_F(p_1, p_2)$$

as $p_1 \to p_2$. (More precisely, $2\mathrm{KL}(p_1 \,\|\, p_2) = D_F^2(p_1, p_2)(1 + O(\| p_1 - p_2 \|))$, where $\| p_1 - p_2 \|$ denotes the L_2 norm of the difference between the densities.) This allows for a data-driven approximation of the Fisher information distance, through the use of the empirically determined PDFs in the absence of information about the Fisher information metric. Although KL divergence is not a symmetric measure, we can add symmetry by defining $D_{\mathrm{KL}}(p_1, p_2) = \mathrm{KL}(p_1 \,\|\, p_2) + \mathrm{KL}(p_2 \,\|\, p_1)$, which maintains similar convergence properties. We note that there are several other metrics that approximate the Fisher information distance, such as the Hellinger and cosine distances. Although for brevity, we utilize KL divergence throughout this article. For additional measures of probabilistic distance and details on their computation for empirical data, we refer the reader to [16] and [17].

As the two densities p_1 and p_2 in (1) become more dissimilar, the KL-divergence approximation of the Fisher information distance becomes weak. Additionally, when PDFs are constrained to form a submanifold of interest, the straight-shot distance is no longer an accurate description of the manifold distance. This is illustrated in Figure 2, in which we represent a one-dimensional submanifold that occupies a subspace of the two-dimensional hypersphere. The Fisher information distance is equal to the shortest path along the submanifold (curvy line) and is not equal to the distance on the full manifold, that is, the portion of a great circle on a hypersphere connecting the two points. Hence, there are situations in which standard approximations of the information distance do not converge to the true distance, and it is necessary to approximate the geodesic along the manifold.

Using a connected graph, we may define the path between p_1 and p_2 as a series of connected segments. The geodesic distance

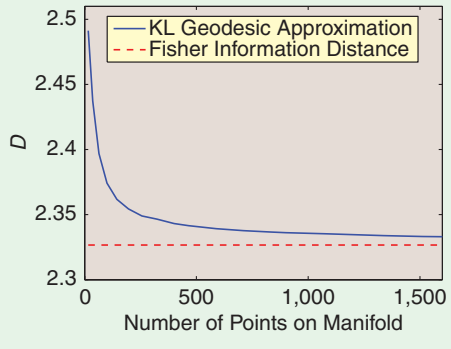


[FIG2] Given a one-dimensional submanifold (the curvy dark line) of interest lying on a two-dimensional sphere manifold, the Fisher information distance is the shortest path connecting the points A and B along the one-dimensional submanifold, rather than the length of a portion of the great circle connecting the points on the sphere.

may then be approximated as the sum of the lengths of those segments. Specifically, given the collection of N PDFs $\mathcal{P} = \{p_1, \ldots, p_N\}$ and using the KL divergence as an approximation of the Fisher information distance, we can now define an approximation function G for all pairs of PDFs

$$G(p_1, p_2; \mathcal{P}) = \min_{M, \mathcal{P}} \sum_{i=1}^{M-1} D_{\mathrm{KL}}(p_{(i)}, p_{(i+1)}), \quad p_{(i)} \to p_{(i+1)} \ \forall i. \tag{2}$$

Intuitively, this estimate calculates the length of the shortest path between points in a connected graph on the well-sampled manifold, and as such $G(p_1, p_2; \mathcal{P}) \to D_F(p_1, p_2)$ as $N \to \infty$. Empirically, (2) may be solved with Dijkstra's shortest-path algorithm. This is similar to the manner in which ISOMAP [5] approximates the distances on Euclidean manifolds. Figure 3 illustrates this approximation by comparing the KL graph approximation to the actual Fisher information distance for the univariate Gaussian case. As the manifold is more densely sampled (uniformly sampling over the range of mean and variance parameters for this simulation), the approximation converges to the true Fisher information distance.



[FIG3] Convergence of the graph approximation of the Fisher information distance using KL divergence. As the manifold is more densely sampled, KL divergence approaches the Fisher information distance.

DIMENSIONALITY REDUCTION IN THE DENSITY SPACE

Consider the collection of PDFs $\mathcal{P} = \{p_1, \ldots, p_N\}$ lying on some statistical manifold $\mathcal{M}$. By performing dimensionality reduction in the space of probability densities, one wishes to reconstruct $\mathcal{M}$ using only the information available in $\mathcal{P}$. Specifically, the aim is to find an embedding $A : p(x) \rightarrow y$, where $y \in \mathbb{R}^m$. This is a similar setting to traditional manifold learning algorithms that aim to reconstruct Riemannian manifolds based on a finite sampling, extended to the properties of statistical manifolds.

By performing dimensionality reduction on a family of PDFs, we are better able to both visualize and classify data. To obtain a lower-dimensional embedding, we calculate the pairwise KL divergences within $\mathcal{P}$. In problems of practical interest, however, the parameterization of the probability densities is usually unknown. We are instead given a family of data sets $\mathcal{X} = \{X_1, \ldots, X_N\}$, in which we may assume that each data set X_i is a realization of some underlying probability distribution to which we do not have knowledge of the parameters. As such, we rely on nonparametric techniques to estimate both the probability density and KL divergence. For the purpose of this article, we implement kernel density estimation methods, although other estimation methods are also applicable.

In a previous work [6], we developed an algorithm for dimensionality reduction in the density space that we called FINE. By assuming that each data set is a realization of an underlying PDF, and each of those distributions lie on a manifold with some natural parameterization, then this embedding can be viewed as an embedding of the actual manifold into Euclidean space. We illustrate the FINE algorithm in Figure 4.

Through information geometry, FINE enables the joint embedding of multiple data sets X_i into a single low-dimensional Euclidean space. By viewing each $X_i \in \mathcal{X}$ as a realization of $p_i \in \mathcal{P}$, we reduce the numerous samples in X_i to a single point. The dimensionality of the statistical manifold may be significantly less than that of the Euclidean realization. MDS methods reduce the dimensionality of p_i from the Euclidean dimension to the dimension of the statistical manifold on which it lies.

ADDING APPLICATION-SPECIFIC CONSTRAINTS

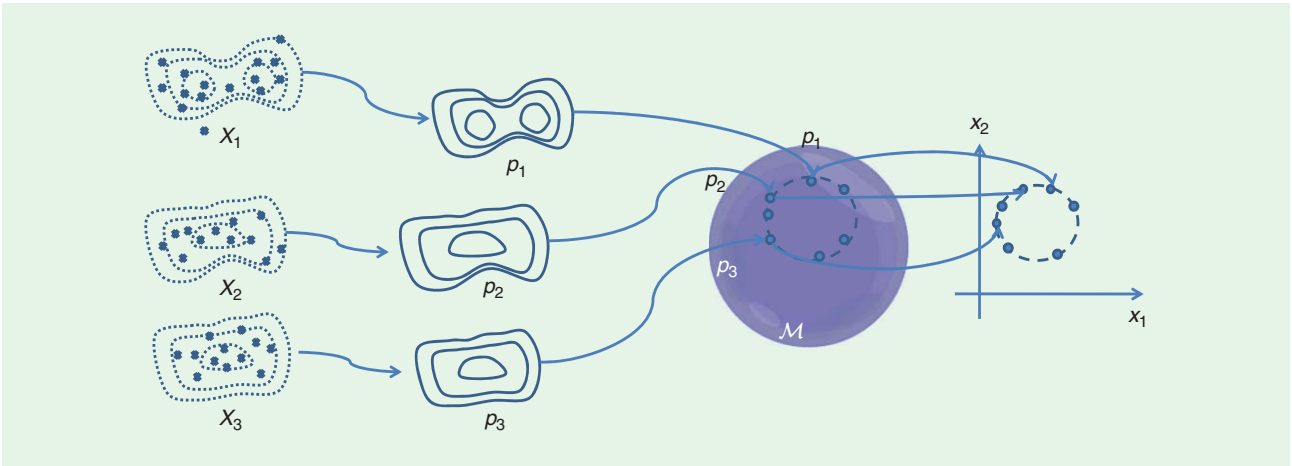
FINE was developed to be applied to the general case of dimensionality reduction in the space of PDFs, making no assumptions on the data distributions or the geometry of the underlying statistical manifolds. However, there are several applications where known intrinsic properties may be exploited when performing information-geometric dimensionality reduction. By incorporating these properties into algorithm constraints, one may be able to obtain improved performance.

Lee et al. [18] have demonstrated the use of IGDR for image segmentation, using multinomial distributions as points that lie on an n-simplex (or projected onto an $(n+1)$-dimensional sphere). By framing their problem as such, they are able to exploit the properties of such a manifold, using the cosine distance as an exact computation of the Fisher information distance and using linear methods (PCA) of dimensionality reduction. They have shown very promising results for the problem of image segmentation.

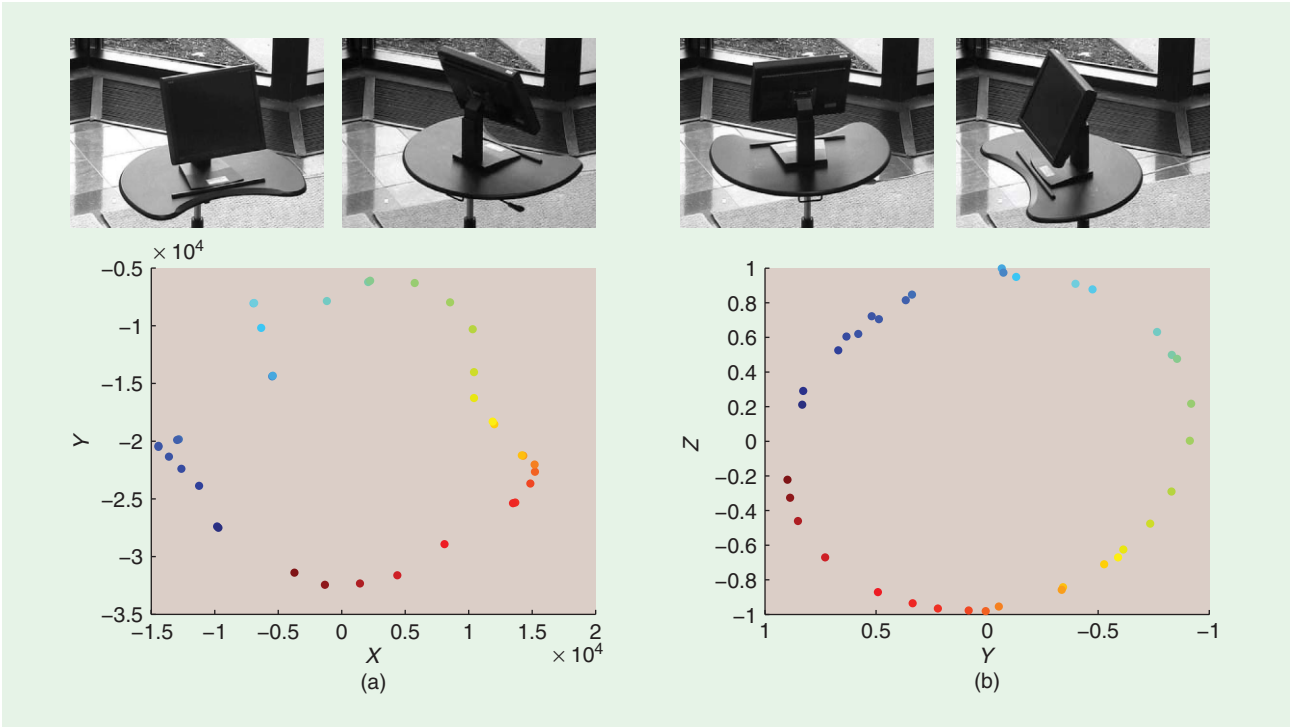
If there exists a priori knowledge that the geometry of the underlying manifold is that of a (hyper)sphere, adding such a constraint results in an improved embedding. In [8], we presented a special case of FINE, which we called spherical Laplacian information maps (SLIM), which restricted the final embedding to constrain all points to lie on the surface of a sphere. SLIM is useful when the user wants to preserve the spherical geometry of the ambient space, for example, when the dimensionality reduction is used to extract object pose trajectories from video. This is illustrated in Figure 5, where we embed the rotation of an object captured by a stationary camera with SLIM and PCA. Each of the 36 images was featured as a multinomial distribution over the pixel space prior to embedding. While PCA discerns the order of the change in angle, it does not properly identify the shape of the trajectory (i.e., circular) as SLIM does.

DIMENSIONALITY REDUCTION IN THE SAMPLE SPACE

For many learning methods, it is often desirable to reduce the dimensionality of X, finding a transformation $A : X \rightarrow Y$, where $Y = [y_1, \ldots, y_n]$ and each $y_i \in \mathbb{R}^m$, $m < d$. Typically, each set



[FIG4] FINE: first, a PDF p_i is estimated for each data set X_i. Then, an information-geometric metric is used to learn the geometry of the manifold of PDFs from pairwise distance measurements. Finally, a Euclidean embedding from the manifold $\mathcal{M}_x$ to $\mathbb{R}^d$ is obtained, associating each original data set X_i with its embedded point in Euclidean space x_i.



[FIG5] The embedding of an object captured at various rotation points with SLIM and PCA. SLIM preserves the spherical nature of the intrinsic manifold. (a) PCA. (b) SLIM.

would be reduced in an individual manner; if there is a deemed relationship between the sets, it has generally been approached as a classification problem in which each signal X_i is considered as a set of points belonging to class i. An example of this situation would be supervised dimensionality reduction with Fisher's LDA [1].

Viewing this problem from an information-geometric perspective presents a different vantage point rather than considering each X_i to be a collection of points in a specific class; let us generalize the relationship between sets X_i and X_j. Specifically, consider the case for which each X_i is a realization of some unknown generating function p_i, in which p_i and p_j may or may not be equivalent. This agrees with the standard classification problem in which each p_i represents a class PDF, but it also allows for the different relationships between PDFs. Specifically, rather than having a number class equal to the number of data sets N, there may be significantly fewer classes $M \ll N$, in which M is unknown and no labels are available. In this generalized scenario, dimensionality reduction is desirable for the purpose of classification, feature extraction, and/or visualization.

Let us illustrate with a simple example. Every ten years, the U.S. census is conducted, generating a collection of data about U.S. residents, such as height, weight, income, ethnicity, and education level. Let us now partition the data such that each county within the same state is represented by its own set X. Standard methods of feature extraction will find the features that best describe each county on an individual level. We are interested in determining the most important features when comparing all counties at the same time. While median income may not be a distinguishing characteristic within a single county and may not be recognized as such when solely extracting features from that individual county, it would be quite informative when comparing all counties across the state.

The construct of comparison across data sets can be directly abstracted to the biomedical field, where it is necessary to compare patients who have been analyzed with the same set of features and identify which of those features best distinguishes the patient corpus. We have presented a method of IGDR, which we refer to as IPCA, to solve this problem for flow cytometry data [9]. IPCA aims to find the optimal transformation of PDFs $A : p(x) \rightarrow p(y)$. By preserving the KL divergence, the estimated PDFs generating the data sets, IPCA ensures that the low-dimensional representation maintains the similarities between data sets that are contained in the full-dimensional data, minimizing the loss of information.

With some abuse of notation, we will further refer to $D_{\mathrm{KL}}(p_i, p_j)$ as $D_{\mathrm{KL}}(X_i, X_j)$, recalling that the KL divergence is calculated with respect to PDFs and not realizations. We define the IPCA projection matrix $A \in \mathbb{R}^{m \times d}$, in which A reduces the dimension of X from d to m ($m \leq d$), such that

$$D_{\mathrm{KL}}(AX_i, AX_j) = D_{\mathrm{KL}}(X_i, X_j), \ \forall i,j. \tag{3}$$

This can be formulated as an optimization problem

$$A = \arg \min_{A:AA^T=I} J(A), \tag{4}$$

where I is the identity matrix and $J(A)$ is some cost function designed to implement (3). Note that we include the optimization constraint $AA^T = I$ to ensure that our projection is orthonormal, which keeps the data from scaling or skewing, as that would undesirably distort

the data. Let $D(\mathcal{X})$ be a dissimilarity matrix such that $D_{ij}(\mathcal{X}) = D_{\mathrm{KL}}(X_i, X_j)$, and $D(\mathcal{X};A)$ is a similar matrix where the elements are perturbed by A, i.e., $D_{ij}(\mathcal{X};A) = D_F(AX_i, AX_j)$. This formulation results in the following cost function:

$$J(A) = \sum_i \sum_j W_{ij}(D_{ij}(\mathcal{X}) - D_{ij}(\mathcal{X};A))^2, \tag{5}$$

where W_{ij} is some weighting factor.

The weights W_{ij} can be selected based on $D_{ij}(\mathcal{X})$ to de-emphasize the influence of certain pairs (i,j) on embedding. For example, the nearest neighbor (NN) weights of $W_{ij} = 1$ for some k-NN and $W_{ij} = 0$ will eliminate far-flung interactions for which KL divergence is a poor approximation to the Fisher metric. The use of heat kernel weights, similar to Laplacian eigenmaps [19], will have a more gradual effect. These functions will ensure that more weight is given to preserve the pairwise distances of close PDFs. Although the choice of cost-weighting function is dependent on the problem, the overall projection method ensures that the similarity between data sets is maximally preserved in the desired low-dimensional space, allowing for comparative learning between sets.

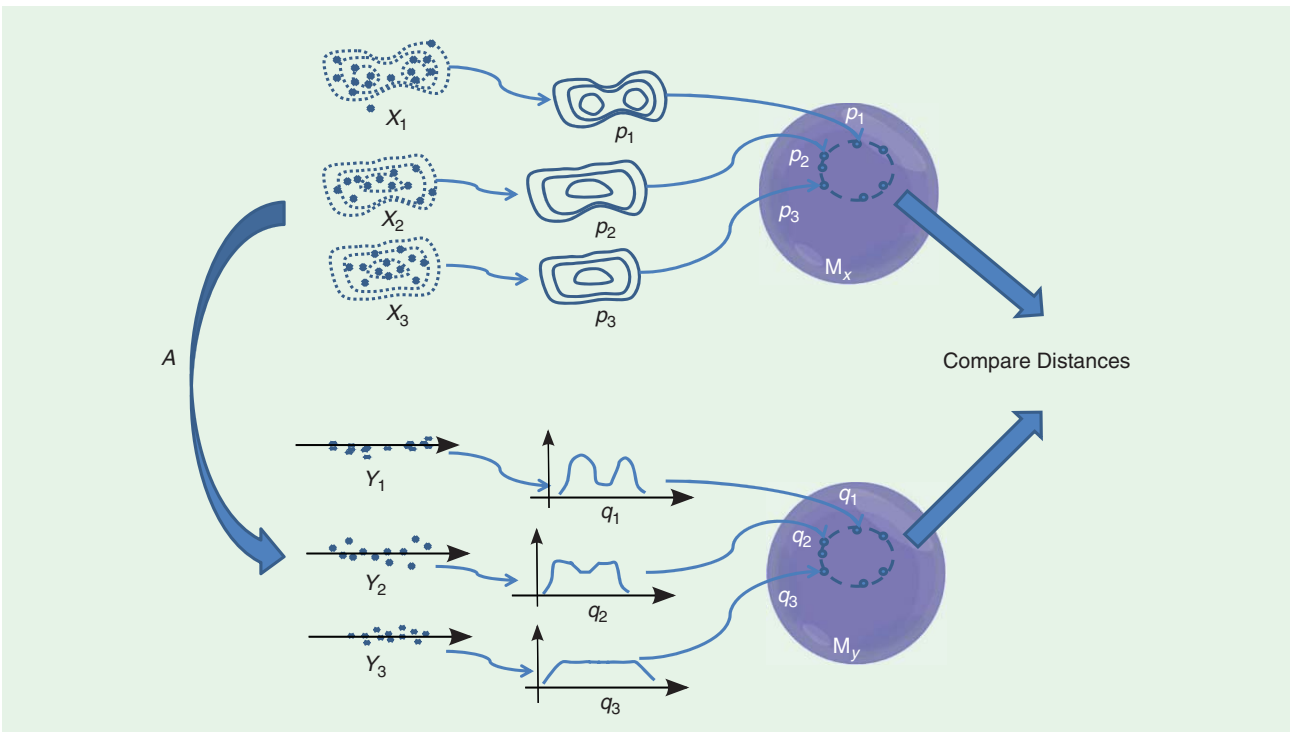
We illustrate the IPCA and IMCA (see the "Supervised Learning" section) in Figure 6. While we omit the details in this article (see [9] and [17]), the cost function (5) may be minimized with various convex optimization techniques; we utilize gradient descent with random initializations for A. There are computational issues with gradient methods, namely, local extrema. We find the global minimum by computing IPCA over several random initializations and taking the resultant A that minimizes the cost function. In most applications we have tested, this method has been very effective, and we have found that most random initializations of A converge to the same minimum.

Recall that the information distance is entirely defined by those areas of input space in which the PDFs differ. As the IPCA preserves the information distance between probability distributions, A is going to be highly weighted toward the variables that contribute the most to that distance. Hence, the loading vectors of A give a ranking of the discriminative value of each variable in the full-dimensional feature space. This form of variable selection is useful in exploratory data analysis.

SUPERVISED LEARNING

As mentioned previously, when developing ICPA, we generalized the relationship between PDFs such that they may or may not represent unique classes in a classification task. We presented an IPCA in the scenario for which sample classification is not the desired task, but we now extend the methods to supervised dimensionality reduction.

The Chernoff performance bound on classification error is used to bound the probability of error based on the probabilistic distance between classes. The Chernoff distance is a single-parameter class of probabilistic distances, and as the distance increases, the probability of misclassification decreases. A special member of



[FIG6] IPCA/IMCA: first, a PDF p_i is estimated for each data set X_i. Simultaneously, a PDF q_i is estimated for each data set $Y_i = AX_i$. Then, an information-geometric metric is used to learn the geometry of the manifold $\mathcal{M}_x$ of PDFs p_is and manifold $\mathcal{M}_y$ from PDFs q_is from pairwise distance measurements. Finally, an objective is calculated to compare the geometry of the two manifolds $\mathcal{M}_x$ and $\mathcal{M}_y$. For IPCA, we consider the minimization of the sum of squared differences between each pairwise distance on $\mathcal{M}_x$ and its equivalent in $\mathcal{M}_y$. For IMCA, we consider the maximization of the sum of distances in $\mathcal{M}_y$.

the class of Chernoff distances, known as the Bhattacharya distance between PDFs, converges to the Fisher information distance, similar to the KL divergence. It is therefore natural to find a form of dimensionality reduction, which will maximize the information distance between class PDFs, as this will enable control of the error probability.

This information-geometric approach fits into the IPCA framework. Consider the following theorem:

THEOREM 1

Let RVs $X, X' \in \mathbb{R}^d$ have PDFs f_X and $f_{X'}$, respectively. Using the $m \times d$ matrix A satisfying $AA^T = I_m$, construct RVs $Y, Y' \in \mathbb{R}^m$ such that $Y = AX$ and $Y' = AX'$. The following relation holds:

$$D_{\mathrm{KL}}(f_X, f_{X'}) \geq D_{\mathrm{KL}}(f_Y, f_{Y'}),$$

where f_Y and $f_{Y'}$ are the PDFs of Y, Y', respectively.

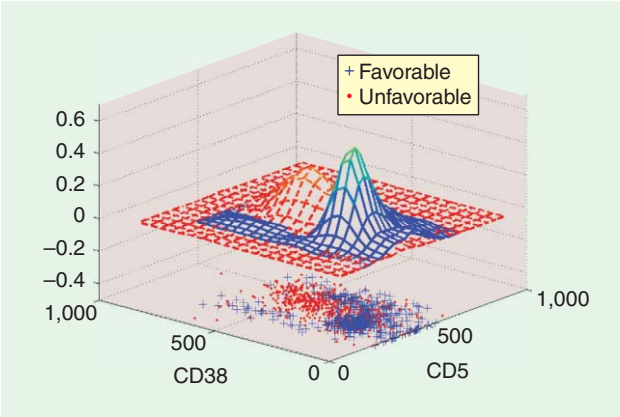
The proof of this theorem may be found in [17] and states that the KL divergence cannot be increased through an orthonormal transform of the input space. This is intuitive as an orthonormal transform is simply a rotation that cannot increase distance. As such, maximizing the information distance between PDFs in a low-dimensional space is directly related to the preserving said distance, albeit with a different formulation.

The first difference is in the setup of the data. We now specify that $\mathcal{X} = \{X_1, \ldots, X_N\}$, where X_i consists of all points $x \in \mathbb{R}^d$ in class C_i, estimating the PDF of X_i as $p_i(x)$. Our objective function for the supervised scenario undergoes a slight modification to become

$$A = \arg \max_{A:AA^T = I} \sum_i \sum_j W_{ij} D_{ij}(\mathcal{X}; A)^2. \tag{6}$$

We refer to this modified algorithm as IMCA [12]. By maximizing the information distance between class PDFs, we not only ensure an optimal performance bound on classification error but also preserve the natural information geometry between classes. This fact is critical when class PDFs are not linearly separable (e.g., such is the assumption of standard LDA). Note that the optimization of the IMCA cost function may be done in a similar fashion to that of IPCA. In fact, for the two-class problem, IPCA and IMCA are identical. For our purpose, we use gradient ascent, as the objective is now a maximization, and the calculation/code is quite similar. Note that we may still use the IMCA projection matrix for variable selection, with the knowledge that the variables with the highest weights are those that contain the most discriminative value, which is critical for classification tasks.

It is explicitly worth pointing out that IMCA is similar to LDA. In fact, if the classes are Gaussian, IMCA would result in an orthogonal version of LDA. Recall that LDA assumes Gaussian classes and maximizes the between-class covariance while minimizing the within-class spread. This would maximize the information distance between the classes. Hence, IMCA can be viewed as a generalized and orthogonal version of LDA, which does not make assumptions on the class distribution.



[FIG7] The point-cloud method of analyzing flow cytometry data is parallel to the analysis of the marginal densities of the data distributions.

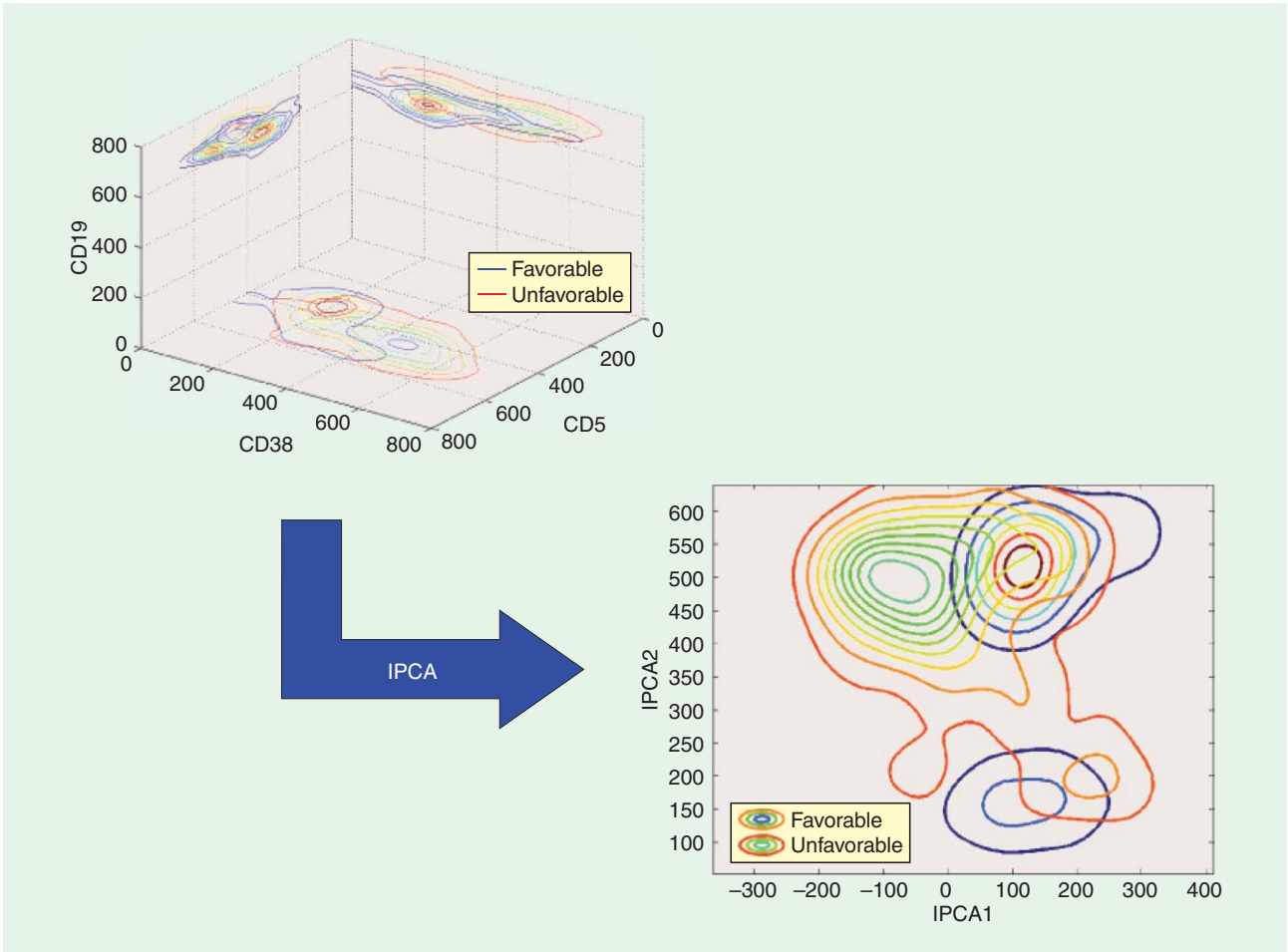
FLOW CYTOMETRY

In clinical flow cytometry, pathologists gather readings of fluorescent markers and light scatter individual blood cells from a patient sample, leading to a characteristic multidimensional distribution that, depending on the panel of markers selected, may be distinct for a specific disease entity. Clinical pathologists generally interpret results in the form of two-dimensional scatter plots in which the axes each represent one of the many cell characteristics analyzed; the multidimensional nature of flow cytometry is routinely underutilized in practice. Given the manner in which analysis is performed on point clouds, pathologists are actually performing a visual density analysis, as illustrated in Figure 7. Here we demonstrate the similar marginal densities (with respect to two biomarkers) of patients with differing prognosis. This enables the utilization of IGDR methods to provide a single-analysis space for pathologists.

We present a study of chronic lymphocytic leukemia (CLL) patients using IPCA to find a low-dimensional space that preserves the differentiation between patients with good and poor prognoses (i.e., favorable and unfavorable immunophenotypes). Using a collection of 23 patients diagnosed with CLL (courtesy of the Department of Pathology at the University of Michigan), we define $\mathcal{X} = \{X_1, \ldots, X_{23}\}$, where each X_i was analyzed with by the series of markers in Table 1. We use IPCA to determine the optimal information-preserving projection space and illustrate this projection in Figure 8. This image shows the three-dimensional measurement space of markers CD5, CD38, and

[TABLE 1] CLL ANALYSIS MARKERS AND THEIR CORRESPONDING IPCA LOADING WEIGHTS.

MARKER	LOADING
FORWARD LIGHT SCATTER	0.1843
SIDE LIGHT SCATTER	0.1044
CD5	0.6270
CD38	0.8420
CD45	0.7228
CD19	0.5750



[FIG8] Contour plots (i.e., PDFs) for three of the six analysis dimensions for CLL prognosis. The data for these patients are then transformed by IPCA, yielding a simple and easily discernable two-dimensional analysis space. The patients chosen are the most similar favorable and unfavorable prognosis CLL patients.

CD19, comparing two very similar patients with differing prognoses. It should be clear that IPCA provides a projection space for which discerning prognosis is simplified.

In Table 1, we also display the loading weights of each of the markers in the IPCA projection matrix. This is done by taking the vector norm of each column in the 2×6 IPCA matrix. Note that CD38 has the largest loading value; reference [20] has shown that patients whose leukemic cells are strong expressers of CD38 have significantly worse survival outcome. We also identify the possibility that CD45 and CD19 expression are also areas that may help prognostic ability, which is an area for further investigation.

Using FINE to embed the data (Figure 9) for comparative visualization, we see that the different prognosis groups are very similar, although decent clusters are formed when labels are applied. These clusters are not well separated, which further illustrates the difficulties in forming an appropriate prognosis. There are also issues of sample size as a larger database of patients may lead to a more clear separation of clusters. Nonetheless, IPCA and FINE were able to appropriately identify the important markers for assigning prognosis and group patients accordingly with respect to immunophenotype. For additional details on this and other studies of FINE and IPCA with flow cytometry, we refer the reader to [9] and [21].

CRIME IN THE 1990s

We next illustrate IDGR to the analysis of crime indicators from the 1990 U.S. census data. This data will be used to illustrate how information geometry can be used to discover which community and law enforcement features may be indicative of the level of crime seen in the said community. We obtained the data from the University of California at Irvine (UCI) Machine Learning Repository [22], which is described in an abbreviated fashion.

The data combines socioeconomic data from the 1990 U.S. census, law enforcement data from the 1990 U.S. Law Enforcement Management and Administrative Statistics (LEMAS) survey, and crime data from the 1995 Federal Bureau of Investigations Uniform Crime Reports (FBI UCR). Attributes were picked if there was any plausible connection to crime ($N = 122$), plus the attribute to be predicted [per capita violent crimes (PCVC)]. The variables included in the data set involve the community, such as the percent of the population considered urban and the median family income, and involving law enforcement, such as per capita

number of police officers and percent of officers assigned to drug units. The PCVC variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault.

All numeric data were normalized in the decimal range [0.00–1.00] using an unsupervised, equal-interval binning method. Attributes retain their distribution and skew (hence, for example, the population attribute has a mean value of 0.06, because most communities are small). An attribute described as mean people per household is actually the normalized (0–1) version of that value.

Since this data set was developed to identify potential crime indicators, the natural partitioning comes by grouping communities by the PCVC indicator variable. We note that while the data set contains 122 features, 22 of those features were only available for a small minority of communities, so we removed them from the set. This left us with a data set consisting of 1,993 communities measured by 100 features. We omit the full feature list, which can be found in [22]; however, we will make explicit note of some selected features shortly.

A DISTINCT DIFFERENCE

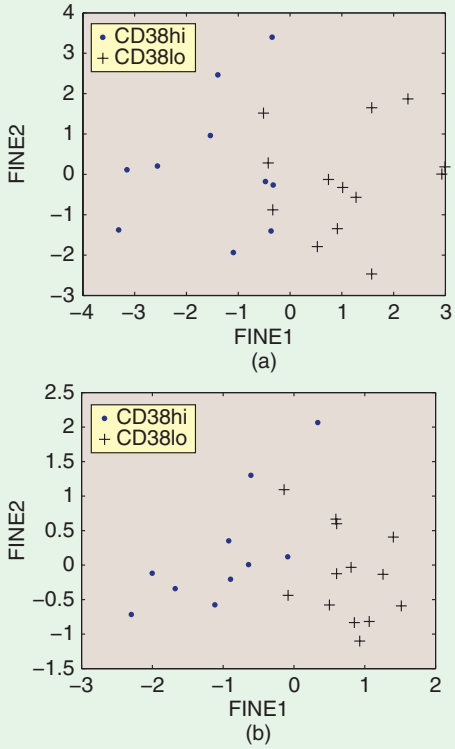
Although it is intuitive to think that communities with high rates of violent crime contain inherent differences than those on the opposite end of the spectrum, it is worth noting that none of the measured features are directly related to crime. Hence, it is worthwhile to first confirm our initial intuition. Additionally, if these features are truly indicative of violent crime, it is reasonable to expect a smooth gradient of change in the features from one end of the spectrum to the other. For example, if a low-median family income indicates the potential for a high amount of crime, and vice versa, then it should be expected that a mid-range median family income should correspond to mid-range crime rates.

We set up this study by grouping communities with respect to their PCVC values. Recall that the range of PCVC is $[0.00, 1.00]$, with the distribution being illustrated in the histogram of Figure 10. As this distribution is highly nonuniform, we use nonuniform bin ranges to group the communities, intended to keep each bin with roughly the same number of samples. This leaves us with a set of 29 crime-based groupings $\mathcal{X} = \{X_1, \ldots, X_{29}\}$, each consisting of between 50 and 122 sample points.

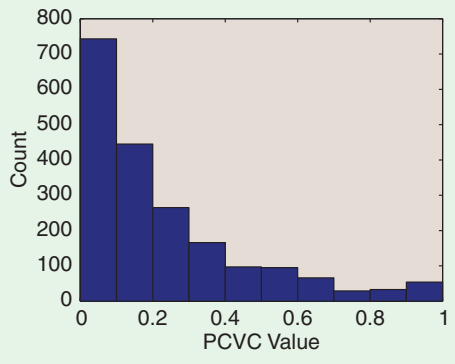
Using kernel density estimation to approximate group PDFs, we embed each crime grouping into a two-dimensional space with FINE. The embedding results can be seen in Figure 11, where each two-dimensional sample point represents a collection of communities whose maximum PCVC value is identified by the plot color. It is clear that our intuition was correct; there exists a smooth and continuous gradient of increasing crime rate. This leads to the natural conclusion that the collection of measured features (or some subset thereof) does indeed contain predictive indicators of violent crime rates.

PREDICTING CRIME AND DISCRIMINATING FEATURES

Given the confirmation that the chosen features do indeed contain predictive value, we now test the classification capabilities when using IGDR as a preprocessing step. Specifically, we look to find the optimal subspace for classifying a community as having either



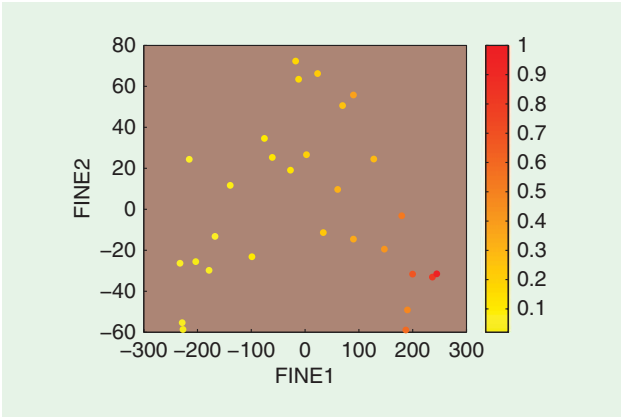
[FIG9] Comparison of CLL patient embeddings, obtained with FINE, using (a) the full-dimensional and (b) IPCA projection matrix. The patients with a poor immunophenotype (CD38hi) are generally well clustered against those with a favorable immunophenotype (CD38lo) in both embedddings.



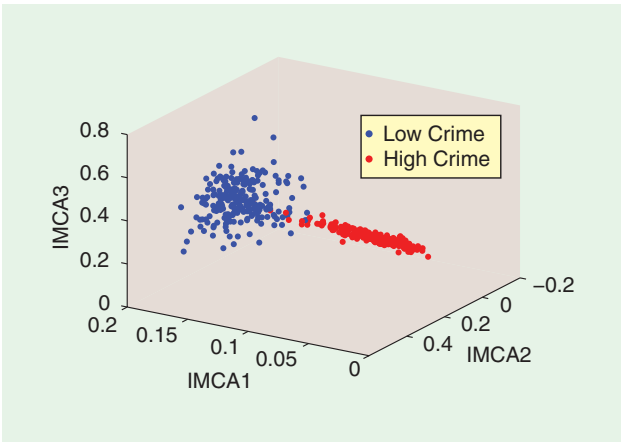
[FIG10] Histogram of the PCVC statistic for the measured communities.

low or high rates of violent crime. This sets up as a two-class problem, and we determine the low-crime class as those communities having a PCVC value of 0.03 or less, and the high-crime class contains communities with a PCVC value greater than 0.53. These thresholds were chosen such that each class roughly contained the same number of samples (226 and 239, respectively).

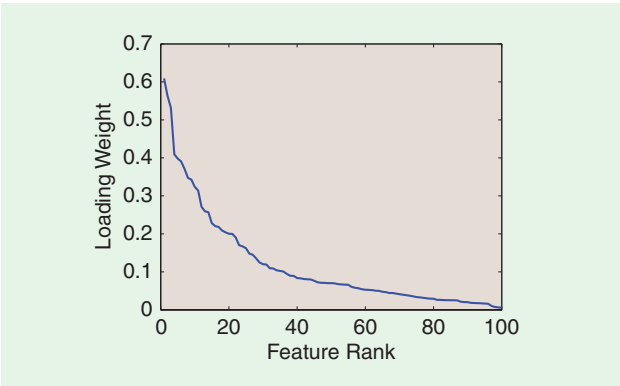
Given that this is a classification problem, we use IMCA to determine our optimal orthonormal projection matrix. Note that we stress the orthonormality constraint here, as using Fisher's



[FIG11] Embedding the crime-based community groupings with FINE. The color of each sample corresponds to the maximum PCVC rate within the group.



[FIG12] The IMCA projection of communities based on the classes defined by low and high PCVC values.



[FIG13] The rankings of the 100 variables in an IMCA projection matrix.

LDA, which does not result in orthogonality, may seem appropriate for this task. If classification was the only desirable task, then LDA would be sufficient. However, we also intend to analyze the projection matrix for variable selection, for which orthogonality becomes a necessity. The LDA projection is useful, however, as it gives us a means for initialization; we make the LDA matrix orthogonal with the classical Gram-Schmidt orthogonalization algorithm and initialize our IMCA gradient methods with the resultant matrix.

[TABLE 2] THE FIVE MOST AND LEAST DISCRIMINATING FEATURES FOR PREDICTING HIGH OR LOW RATES OF VIOLENT CRIME.

TOP FIVE VARIABLES
—POPULATION FOR COMMUNITY
—RENTAL HOUSING-MEDIAN RENT
—NUMBER OF PEOPLE LIVING IN AREAS CLASSIFIED AS URBAN
—PERCENTAGE OF POPULATION WHO ARE DIVORCED
—MEDIAN HOUSEHOLD INCOME
BOTTOM FIVE VARIABLES
—PER CAPITA INCOME FOR PEOPLE WITH HISPANIC HERITAGE
—PERCENT OF OFFICERS ASSIGNED TO DRUG UNITS
—PER CAPITA INCOME FOR PEOPLE WITH ASIAN HERITAGE
—LAND AREA IN SQUARE MILES
—MEDIAN YEAR HOUSING UNITS BUILT

We choose to perform our analysis in an $m = 3$-dimensional projection space for two reasons—the ability to visualize the data and the three-dimensional space optimized our objective, obtaining the maximum separation between classes for $m \in [2, 7]$. After obtaining the 3×100 IMCA matrix A, we project the data from each class into the same space and perform our classification task. The projected data are shown in Figure 12. It is interesting that the low-crime communities show much more variation than the high-crime communities, which exhibit a tight cluster even though the range of PCVC value was much larger.

To test classification performance, we use a simple linear classifier and perform leave-one-out cross-validation over all samples in the set. The results yield a 1.29% classification error: one low-crime and five high-crime communities were misclassified. For comparison sake, we note that PCA, an orthonormal unsupervised method, results in a 3.44% error rate, and LDA yielded a 1.51% error rate. Recall that LDA does not have the orthogonal constraint, yet IMCA still results in (slightly) better classification performance. In all cases, the projection data was projected to three dimensions.

We now use the IMCA matrix A to identify the most discriminating features. To do so, we calculate the L_2 norm of the vector of weights for each of the 100 features (columns) of the 3×100 projection matrix A. After sorting in descending order, we plot these ranks in Figure 13. This shows that there are several features that offer some discriminative value and many more that offer very little. In Table 2, we report the five most and least discriminating features. We preface these results by recalling that this data was from a 1990 census and 1995 crime report. Obviously, much has changed since this data was reported, but the results do appear logical.

CONCLUSIONS

In this article, we have presented IGDR, an information-geometric framework for dimensionality reduction. In contrast to standard Euclidean approaches to manifold learning, which aim to reconstruct a Riemannian submanifold of Euclidean space, our objective is to learn statistical manifolds. We have shown that when the data produces realizations of PDFs lying on a statistical manifold, we can perform information-driven dimensionality reduction in both density space and sample space. These techniques were

illustrated on the problem of flow cytometry analysis, showing the ability to find a subspace in which a pathologist can better diagnose CLL patients. We were also able to compare patients one to another in a single low-dimensional embedding space. We also applied IGDR to a crime and community data set, identifying community indicators of violent crime and accurately clustering and classifying communities with high or low crime rates. The power of using information geometry for dimensionality reduction has just begun to be explored, and we hope this article will lead to further extensions and applications.

ACKNOWLEDGMENTS

This research was supported in part by ARO grant W911NF-09-1-0310, NSF grant CCF 0830490, and an AFRL ATR Center grant through SIG Inc. The authors also thank Christine Kim of the University of Michigan, who helped in collecting the data used for Figure 5 while a summer intern at AFRL.

AUTHORS

Kevin M. Carter (kmcarter@umich.edu) received the B.Eng. (cum laude) degree in computer engineering from the University of Delaware in 2004. He received an M.S. degree in electrical engineering in 2006 and a Ph.D. degree in electrical engineering in 2009, both at the University of Michigan. He is currently a member of the technical staff at MIT Lincoln Laboratory, working on problems of cyber security, network traffic analysis, and anomaly detection. He is a Student Member of the IEEE. His research interests include manifold learning, with specific focuses on statistical manifolds, dimensionality reduction, intrinsic dimension estimation, statistical signal processing, machine learning, and pattern recognition.

Raviv Raich (raich@eecs.oregonstate.edu) received the B.Sc. and M.Sc. degrees from Tel Aviv University, Tel Aviv, Israel, in 1994 and 1998, respectively, and the Ph.D. degree from the Georgia Institute of Technology in 2004, all in electrical engineering. From 2004 to 2007, he was a postdoctoral fellow at the University of Michigan. Since fall 2007, he has been an assistant professor in the School of Electrical Engineering and Computer Science, Oregon State University. He is a Member of the IEEE. His research interests are statistical signal processing, machine learning, dimensionality reduction, probabilistic modeling and inference, manifold learning, and sparse signal reconstruction.

William G. Finn (wgfinn@umich.edu) received an M.D. degree from the University of Michigan in 1989 and completed residency training in anatomic and clinical pathology at Northwestern University in 1994 and is board certified. He has served on the faculty of the University of Michigan since 1998, where he is a professor in the Department of Pathology, associate director of clinical laboratories, and medical director of Clinical Hematology Laboratory. He serves on the board of directors of the American Society for Clinical Pathology and the International Society for Laboratory Hematology and on the executive committee of the Society for Hematopathology.

Alfred O. Hero, III (hero@umich.edu), received the B.S. degree from Boston University in 1980 and the Ph.D. degree from Princeton University in 1984, both in electrical engineering. Since 1984 he has been with the University of Michigan, where he is the R. Jamison and Betty Williams Professor of Engineering. He is a Fellow of the IEEE and has received several awards for his research and professional activities. He was president of the IEEE Signal Processing Society and is currently the IEEE Division IX director. His research focus is the development of information-theoretic approaches to imaging, signal processing, and large-scale data analysis.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[3] Y. Y. Sun, M. K. Ng, and Z. H. Zhou, "Multi-instance dimensionality reduction," in *Proc. 24th AAAI Conf. Artificial Intelligence*, 2010, pp. 587–592.

[4] T. Cox and M. Cox, *Multidimensional Scaling*. London: Chapman & Hall, 1994.

[5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[6] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Fine: Fisher information non-parametric embedding," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 11, pp. 2093–2098, Nov. 2009.

[7] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," vol. 11, pp. 451–490, Feb. 2010.

[8] K. M. Carter, R. Raich, and A. O. Hero, "Spherical Laplacian information maps (slim) for dimensionality reduction," in *Proc. IEEE Inter. Conf. Statistical Signal Processing*, Aug. 2009.

[9] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Information preserving component analysis: Data projections for flow cytometry analysis," *IEEE J. Select. Topics Signal Processing (Special Issue on Digital Image Processing Techniques for Oncology)*, vol. 3, no. 1, pp. 148–158, Feb. 2009.

[10] J. Peltonen, "Visualization by linear projections as information retrieval," in *Proc. 7th Int. Workshop Advances in Self-Organizing Maps*, pp. 237–245, 2009.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York: Springer-Verlag, 1998.

[12] K. M. Carter, R. Raich, and A. O. Hero III, "An information geometric approach to supervised dimensionality reduction," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2009, pp. 1829–1832.

[13] Information geometric dimensionality reduction toolbox [Online]. Available: http://tbayes.eecs.umich.edu/kmcarter/igdr/index.html

[14] R. Kass and P. Vos, *Geometrical Foundations of Asymptotic Inference* (Wiley Series in Probability and Statistics). New York: Wiley, 1997.

[15] S. Amari and H. Nagaoka, *Methods of Information Geometry*. (Translations of Mathematical Monographs, vol. 191). Oxford: Oxford Univ. Press, 2000.

[16] S. K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 6, pp. 917–929, June 2006.

[17] K. M. Carter, "Dimensionality reduction on statistical manifolds," Ph.D. dissertation, Univ. Michigan, Jan. 2009.

[18] S.-M. Lee, A. L. Abbott, and P. A. Araman, "Dimensionality reduction and clustering on statistical manifolds," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2007, pp. 1–7.

[19] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*, vol. 14. T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002.

[20] R. N. Damle, T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S. L. Allen et al., "Ig v gene mutation status and cd38 expression as novel prognostic indicators in chronic lymphocytic leukemia," *Blood*, vol. 95, no. 7, pp. 1840–1847, 1999.

[21] W. G. Finn, K. M. Carter, R. Raich, and A. O. Hero, "Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: Treating flow cytometry data as high-dimensional objects," *Cytometry B*, vol. 76, no. 1, pp. 1–7, Jan. 2009.

[22] UCI Machine Learning Repository: Communities and crime data set [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime [SP]

applications CORNER

Samuel Kaski and Jaakko Peltonen

Dimensionality Reduction for Data Visualization

Dimensionality reduction is one of the basic operations in the toolbox of data analysts and designers of machine learning and pattern recognition systems. Given a large set of measured variables but few observations, an obvious idea is to reduce the degrees of freedom in the measurements by representing them with a smaller set of more "condensed" variables. Another reason for reducing the dimensionality is to reduce computational load in further processing. A third reason is visualization. "Looking at the data" is a central ingredient of exploratory data analysis, the first stage of data analysis where the goal is to make sense of the data before proceeding with more goal-directed modeling and analyses. It has turned out that although these different tasks seem alike, their solution requires different tools. In this article, we show that dimensionality reduction for data visualization can be represented as an information retrieval task, where the quality of visualization can be measured by precision and recall measures and their smoothed extensions. Furthermore, we show that visualization can be optimized to directly maximize the quality for any desired tradeoff between precision and recall, yielding very well-performing visualization methods.

HISTORY

Each multivariate observation $\mathbf{x}_i = [x_{i1}, \ldots x_{\text{in}}]^T$ is a point in an n-dimensional space. A key idea in dimensionality reduction is that if the data lies in a d-dimensional $(d < n)$ subspace of the n-dimensional space, and if we can identify the subspace, then there exists a transformation that loses no information and allows the data to be represented in a d-dimensional space. If the data lies in a (linear) subspace, then the transformation is linear and more generally the data may lie in a d-dimensional (curved) manifold and the transformation is nonlinear.

Among the earliest methods are so-called multidimensional scaling (MDS) methods [1] that try to position data points into a d-dimensional space such that their pairwise distances are preserved as well as possible. If all pairwise distances are preserved, it can be argued that the data manifold has been identified (up to some transformations). In practice, data of course are noisy, and the solution is found by minimizing a cost function such as the squared loss between the pairwise distances, $E_{\text{MDS}} = \sum_{i,j}(d(\mathbf{x}_i, \mathbf{x}_j) - \text{d}(\mathbf{x}_i', \mathbf{x}_j'))^2$, where the $d(\mathbf{x}_i, \mathbf{x}_j)$ are the original distances between the points $\mathbf{x}_i$ and $\mathbf{x}_j$, and the $d(\mathbf{x}_i', \mathbf{x}_j')$ are the distances between their representations $\mathbf{x}_i'$ and $\mathbf{x}_j'$ in the d-dimensional space.

MDS comes in several types that differ in the specific form of cost function and additional constraints on the mapping, and some of the choices give familiar methods such as principal components analysis or Sammon's mapping as special cases.

Neural computing methods are other widely used families of manifold embedding methods. The so-called autoencoder networks (see, e.g., [2]) pass the data vector through a lower-dimensional bottleneck layer in a neural network that aims to reproduce the original vector. The activities of the neurons in the bottleneck layer give the coordinates on the data manifold. Self-organizing maps (see [3]), on the other hand, directly learn a discrete representation of a low-dimensional manifold by positioning weight vectors of neurons along the manifold; the result is a discrete approximation to principal curves or manifolds, a nonlinear generalization of principal components [4].

In the year 2000, a new manifold learning boom began after the publication of two papers in *Science* showing how to learn nonlinear data manifolds. Locally linear embedding [5] made, as the name reveals, locally linear approximations to the nonlinear manifold. The other, called Isomap [6], is essentially MDS tuned to work along the data manifold. After the manifold has been learned, distances will be computed along the manifold. But plain MDS tries to approximate distances of the data space that do not follow the manifold, and hence plain MDS will not work in general. That is why Isomap starts by computing distances along the data manifold, approximated by a graph connecting neighbor points. Since only neighbors are connected, the connections are likely to be on the same part of the manifold instead of jumping across gaps to different branches; distances along the neighborhood graph are thus decent approximations of distances along the data manifold known as "geodesic distances."

A large number of other approaches have been introduced for the learning of manifolds during the past ten years, including methods based on spectral graph theory and on simultaneous variance maximization and distance preservation.

CONTROVERSY

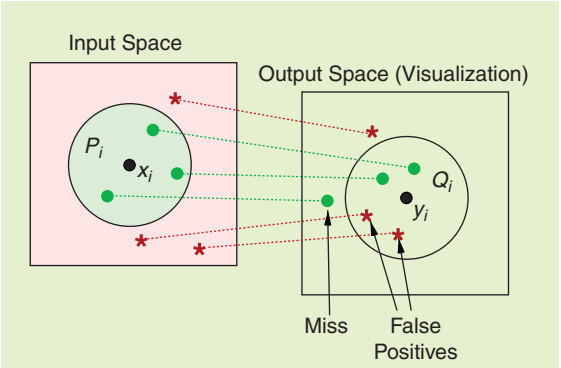
Manifold learning research has been criticized for lack of clear goals. Many papers introduce a new method and only show its performance by nice images of how it learns a toy manifold. A famous example is the "Swiss roll," a two-dimensional data sheet curved in three dimensions

Digital Object Identifier 10.1109/MSP.2010.940003
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

into a Swiss roll shape. Many methods have been shown capable of unrolling the Swiss roll, but few have been shown to have real applications, success stories, or even to quantitatively outperform alternative methods.

One reason why quantitative comparisons are rare is that the goal of manifold embedding has not always been clearly defined. In fact, manifold learning may have several alternative goals depending on how the learned manifold will be used. We focus on one specific goal, data visualization, intended for helping analysts to look at the data and find related observations during exploratory data analysis.



[FIG1] A visualization can have two kinds of errors (from [9]). When a neighborhood P_i in the high-dimensional input space is compared to a neighborhood Q_i in the visualization, false positives are points that appear to be neighbors in the visualization but are not in the original space; misses (which could also be called false negatives) are points that are neighbors in the original space but not in the visualization.

Data visualization is traditionally not a well-defined task either. But it is easy to observe empirically [7] that many of the manifold learning methods are not good for data visualization. The reason is that they have been designed to find a d-dimensional manifold if the inherent dimensionality of data is d. For visualization, the display needs to have $d = 2$ or $d = 3$; that is, the dimensionality may need to be reduced beyond the inherent dimensionality of data.

NEW PRINCIPLE

It is well known that a high-dimensional data set cannot in general be faithfully represented in a lower-dimensional space, such as the plane with $d = 2$. Hence a visualization method needs to choose what kinds of errors to make. The choice naturally should depend on the visualization goal; it turns out that under a specific but general goal the choice can be expressed as an interesting tradeoff, as we will describe below.

When the task is to visualize which data points are similar, the visualization can have two kinds of errors (Figure 1): it can miss some similarities (i.e., it can place similar points far apart as false negatives) or it can bring dissimilar data points close together as false positives. If we know the cost of each type of error, the visualization can be optimized to minimize the total cost. Hence, once the user gives the relative cost of misses and false positives, it fixes visualization to be a well-defined optimization task. It turns out [8], [9] that under simplifying assumptions the two costs turn into precision and recall, standard measures between which a user-defined tradeoff is made in information retrieval.

Hence, the task of visualizing which points are similar can be formalized as a task of visual information retrieval, that is, retrieval of similar points based on the visualization. The visualization can be optimized to maximize information retrieval performance, involving as an unavoidable element a tradeoff between precision and recall. In summary, visualization can be made into a rigorous modeling task, under the assumption that the goal is to visualize which data points are similar.

When the simplifying assumptions are removed, the neighborhoods are allowed to be continuous-valued probability distributions p_{ij} of point j being a neighbor of point i. Then it can be shown that suitable analogs of precision and recall are distances between the neighborhood distributions p in the input space and q on the display. More specifically, the Kullback-Leibler divergence $D(p_i, q_i)$ reduces under simplifying assumptions to recall and $D(q_i, p_i)$ to precision. The total cost is then

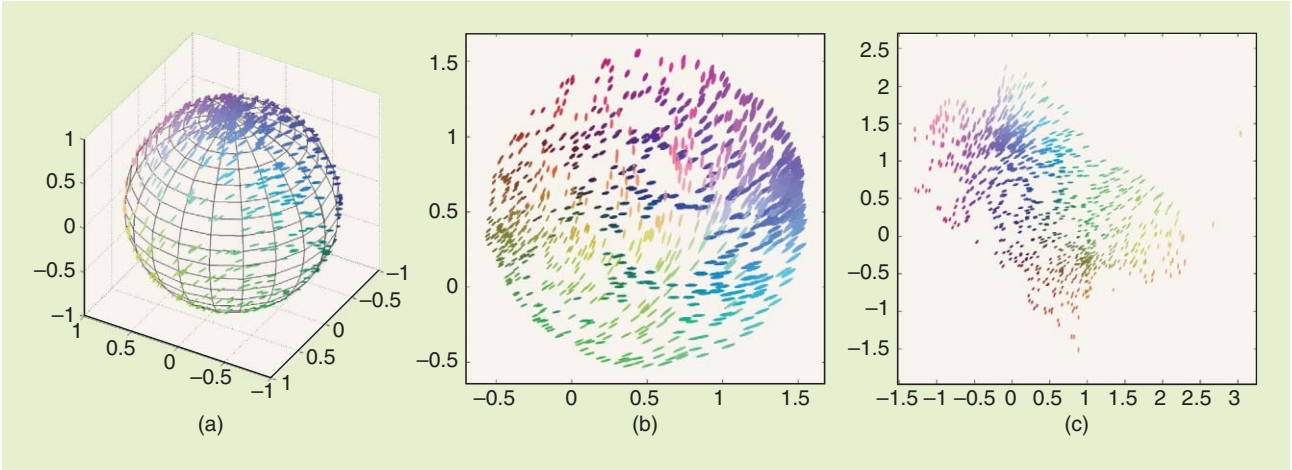
$$E = \lambda \sum_i D(p_i, q_i) + (1 - \lambda) \sum_j D(q_i, p_i), \quad (1)$$

where λ is the relative cost of misses and false positives. The display coordinates of all data points are then optimized to minimize this total cost; several nonlinear optimization approaches could be used; we have simply used conjugate gradient descent. This method has been called NeRV for neighbor retrieval visualizer [8], [9]. When $\lambda = 1$ the method reduces to stochastic neighbor embedding [10], an earlier method that we now see maximizes recall.

Visualization of a simple data distribution makes the meaning of the tradeoff between precision and recall more concrete. When visualizing the surface of a three-dimensional sphere in two dimensions, maximizing recall squashes the sphere flat (Figure 2) whereas maximizing precision "peels" the surface open. Both solutions are good but have different kinds of errors.

Both nonlinear and linear visualizations can be optimized by minimizing (1). The remaining problem is how to define the neighborhoods p; in the absence of more knowledge, symmetric Gaussians or more heavy-tailed distributions are justifiable choices. An even better alternative is to derive the neighborhood distributions from probabilistic models that encode our knowledge of the data, both prior knowledge and what was learned from data.

Deriving input similarities from a probabilistic model has recently been done in Fisher information nonparametric embedding [11], where the similarities (distances) approximate Fisher information distances (geodesic distances where the local metric is defined by a Fisher information matrix) derived from nonparametric probabilistic models. In related earlier work [12], [13], approximated geodesic distances were computed in a "learning metric" derived using Fisher information matrices for a conditional class probability model. In all these works, though, the distances were given to standard visualization methods, which have not been designed



[FIG2] Tradeoff between precision and recall in visualizing a sphere (from [9]). (a) The three-dimensional location of points on the three-dimensional sphere is encoded into colors and glyph shapes. (b) Two-dimensional visualization that maximizes recall by squashing the sphere flat. All original neighbors remain close-by but false positives (false neighbors) from opposite sides of the sphere also become close-by. (c) Visualization that maximizes precision by peeling the sphere surface open. No false positives are introduced but some original neighbors are missed across the edges of the tear.

for a clear task of visual information retrieval. In contrast, we will combine the model-based input similarities to the rigorous precision-recall approach to visualization. Then the whole procedure corresponds to a well-defined modeling task where the goal is to visualize which data points are similar. We will next discuss this in more detail in two concrete applications.

APPLICATION 1: VISUALIZATION OF GENE EXPRESSION COMPENDIA FOR RETRIEVING RELEVANT EXPERIMENTS

In the study of molecular biological systems, behavior of the system can seldom be inferred from first principles either because such principles are not known yet or because each system is different. The study must be data driven. Moreover, to make research cumulative, new experiments need to be placed in the context of earlier knowledge. In the case of data-driven research, a key part of that is retrieval of relevant experiments. An earlier experiment, a set of measurements, is relevant if some of the same biological processes are active in it, either intentionally or as side effects.

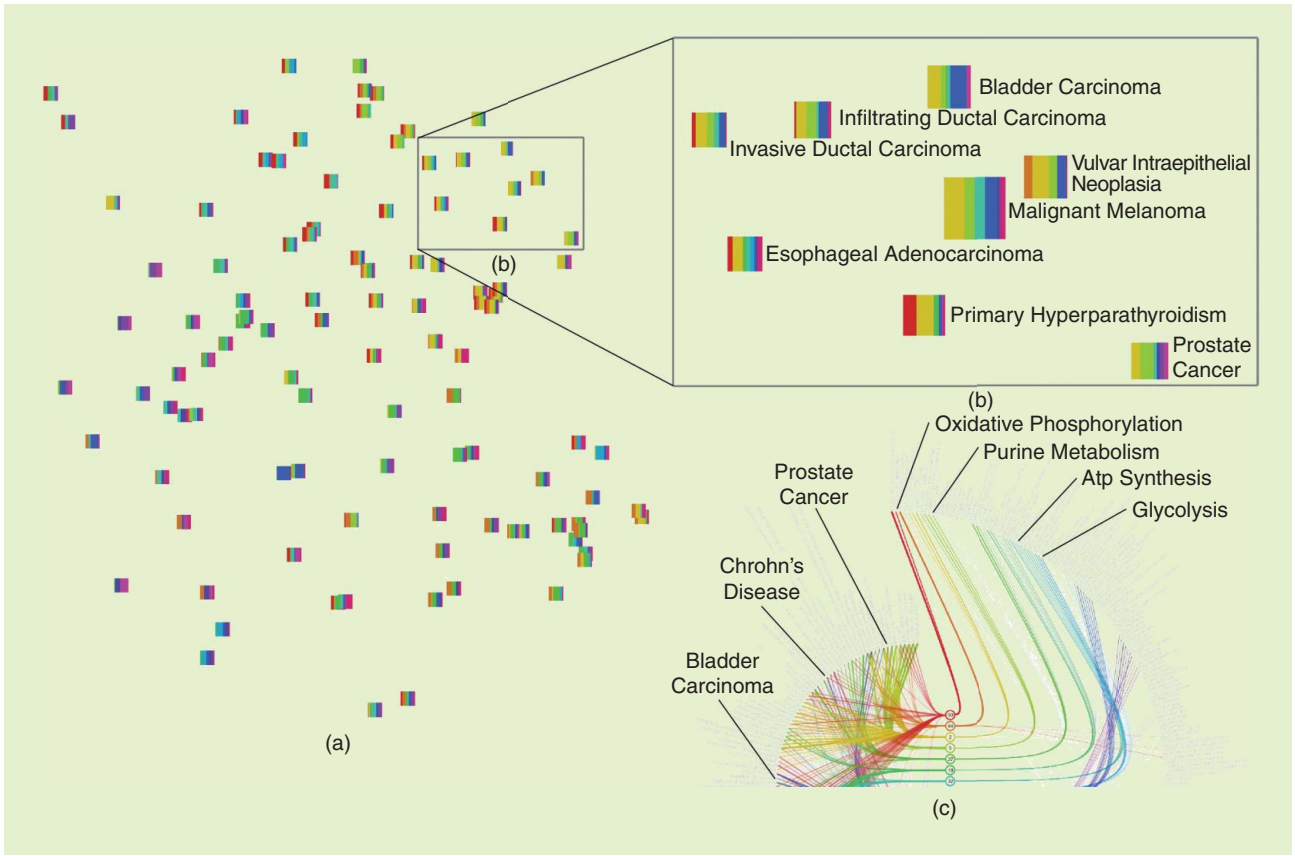
In molecular biology it has become standard practice to store experimental data in repositories such as ArrayExpress of the European Bioinformatics Institute (EBI). Traditionally, experiments are sought from the repository based on metadata annotations only, which works well when searching for experiments that involve well-annotated and well-known biological phenomena. In the interesting case of studying and modeling new findings, more data-driven approaches are needed, and information retrieval and visualization based on latent variable models are promising tools [14].

Let's assume that in experiment i data $\mathbf{g}_i$ have been measured; in the concrete case below $\mathbf{g}_i$ will be a differential gene expression vector, where g_{ij} is expression level of gene or gene set j compared to a control measurement. Now if we fit to the compendium a model that generates a probability distribution over the experiments, $p(\mathbf{g}_i, \mathbf{z}_i|\theta)$, where the θ are parameters of the model which we will omit below and $\mathbf{z}$ are latent variables, this model can be used for retrieval and visualization as explained below. This modeling approach makes sense in particular if the model is constructed such that the latent variables have an interpretation as activities of latent or "underlying" biological processes which are manifested indirectly as the differential gene expression.

Given the model, relevance can be defined in a natural way as follows: The likelihood of experiment i being relevant for an earlier experiment j is $p(\mathbf{g}_i|\mathbf{g}_j) = \int p(\mathbf{g}_i|\mathbf{z})\mathrm{p}(\mathbf{z}|\mathbf{g}_j)\mathrm{d}\mathbf{z}$. That is, the experiment is relevant if it is likely that the measurements have arisen as products of the same unknown biological processes $\mathbf{z}$. This definition of relevance can now be used for retrieving the most relevant experiments, and, moreover, the definition can be used as the natural probability distribution p in (1) to construct a visual information retrieval interface (Figure 3); in this case the data are 105 microarray experiments from the ArrayExpress database, comparing pathological samples such as cancer tissues to healthy samples.

The above visual information retrieval idea was explained in abstract concepts, applicable to many data sources. In the gene expression retrieval case of Figure 3, the data were expressions of a priori defined gene sets, quantized into counts, and the probabilistic model was the discrete principal component analysis model, also called latent Dirichlet allocation, and in the context of texts called a topic model. The resulting relevances can directly be given as inputs to NeRV; in Figure 3 a slightly modified variant of the relevances was used, details in [14].

In summary, fitting a probabilistic latent variable model to the data produces a natural relevance measure that



[FIG3] A visual information retrieval interface to a collection of microarray experiments visualized as glyphs on a plane (from [14]). (a) Glyph locations have been optimized by NeRV so that relevant experiments are close by. For this experiment data, relevance is defined by the same data-driven biological processes being active, as modeled by a latent variable model (component model). (b) Enlarged view with annotations; each color bar corresponds to a biological component or process, and the width tells the activity of the component. These experiments are retrieved as relevant for the melanoma experiment shown in the center. (c) The biological components (nodes in the middle) link the experiments (left) to sets of genes (right) activated in them.

can then be plugged as a similarity measure into the visualization framework. Everything from start to finish is then based on rigorous choices.

APPLICATION 2: VISUALIZATION OF GRAPHS

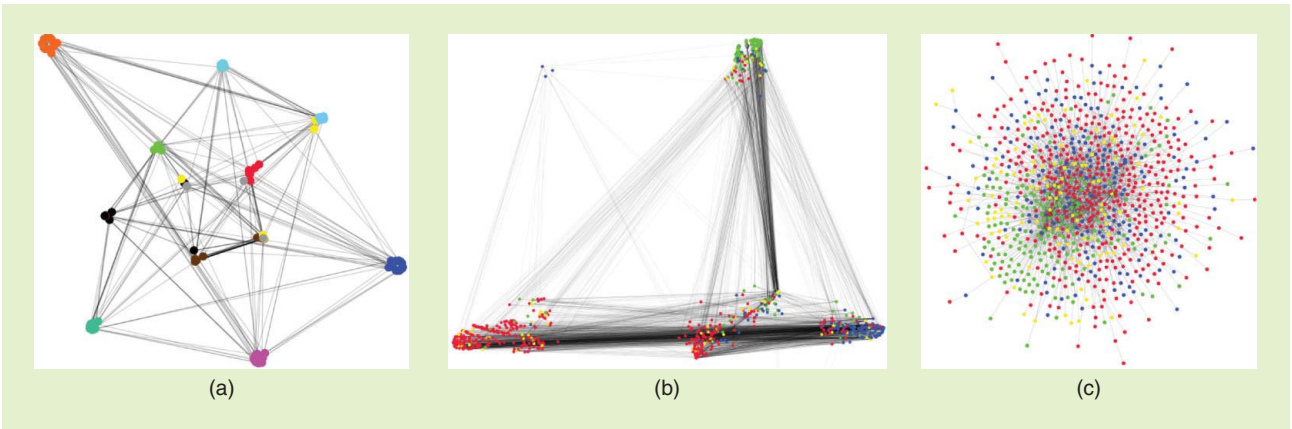
Graphs are a natural representation of data in several fields where visualizations are helpful, such as social networks analysis, interaction networks in molecular biology, and citation networks. In a sense, graphs are high-dimensional structured data where nodes are points and all other nodes are dimensions; the value of the dimension is the type or strength of the link.

There exist many graph drawing algorithms, including string analogy-based methods such as Walshaw's algorithm [15] and spectral methods [16]. Most of them focus explicitly or implicitly on local properties of graphs, drawing nodes linked by an edge close together but avoiding overlap. That works well for simple graphs but for large and complicated ones additional principles are needed to avoid the famous "hairball" visualizations.

A promising direction forward is to learn a probabilistic latent variable model of the graph, in the hope of capturing its central properties, and then focus on visualizing those properties. In the case of graphs, the data to be modeled are which other nodes a node links to. But as the observed links in a network may be stochastic (noisy) measurements such as gene interaction measurements, it makes sense to assume that the links are a sample from an underlying link distribution and learn a probabilistic latent variable model to model the distributions. The similarity of two nodes is then naturally evaluated as similarity of their link distributions. The rest of the visualization can proceed as in the previous section, with experiments replaced by graph nodes.

Figure 4 shows sample graphs visualized based on a variant of discrete principal components analysis or latent Dirichlet allocation suitable for graphs. With this link distribution-based approach, NeRV places nodes close-by on the display if they link to similar other nodes, with similarity defined as similarity of link distributions. This has the nice side-result that links form bundles where all start nodes are similar and all end nodes are similar.

In summary, the idea is to use any prior knowledge in choosing a suitable model for the graph, and after that all steps of the visualization follow naturally and rigorously from start to finish. In



[FIG4] Visualizations of graphs. (a) U.S. college football teams (nodes) and who they played against (edges). The visual groups of teams match the 12 conferences arranged for yearly play (shown with different colors). (b), (c) Word adjacencies in the works of Jane Austen. The nodes are words, and edges mean the words appeared next to each other in the text. NeRV visualization in (b) shows visual groups that reveal syntactic word categories: adjectives, nouns, and verbs shown in blue, red, and green. The edge bundles reveal disassortative structure that matches intuition, for example, verbs are adjacent in text to nouns or adjectives and not to other verbs. Earlier graph layout methods (Walshaw's algorithm shown in (c) fails to reveal the structure.) (Figure from [17], © ACM, 2010, used with permission).

the absence of prior knowledge flexible machine learning models such as the discrete principal components analysis above can be learned from data.

CONCLUSIONS

We have discussed dimensionality reduction for a specific goal, data visualization, which has been so far defined heuristically. Recently it has been suggested that a specific kind of data visualization task, that is, visualization of similarities of data points, could be formulated as a visual information retrieval task, with a well-defined cost function to be optimized. The information retrieval connection further reveals that a tradeoff between misses and false positives needs to be made in visualization as in all other information retrieval. Moreover, the visualization task can be turned into a well-defined modeling problem by inferring the similarities using probabilistic models that are learned to fit the data.

A free software package that solves nonlinear dimensionality reduction as visual information retrieval, with a method called NeRV, is available at http://www.cis.hut.fi/projects/mi/software/dredviz/.

AUTHORS

Samuel Kaski (samuel.kaski@tkk.fi) is a professor of computer science at Aalto University and director of the Helsinki Institute for Information Technology (HIIT). He studies machine learning, in particular multisource machine learning, with applications in bioinformatics, neuroinformatics and proactive interfaces.

Jaakko Peltonen (jaakko.peltonen@tkk.fi) is a postdoctoral researcher and docent at Aalto University, Department of Information and Computer Science. He received the D.Sc. degree from Helsinki University of Technology in 2004. He is an associate editor of *Neural Processing Letters* and has served in program committees of 11 conferences. He studies generative and information theoretic machine learning especially for exploratory data analysis, visualization, and multisource learning.

REFERENCES

[1] I. Borg and P. Groenen, *Modern Multidimensional Scaling*. New York: Springer-Verlag, 1997.

[2] G. Hinton, "Connectionist learning procedures," *Artif. Intell.*, vol. 40, pp. 185–234, 1989.

[3] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin: Springer-Verlag, 2001.

[4] F. Mulier and V. Cherkassky, "Self-organization as an iterative kernel smoothing process," *Neural Comput.*, vol. 7, pp. 1165–1177, 1995.

[5] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] J. Venna and S. Kaski, "Comparison of visualization methods for an atlas of gene expression data sets," *Inform. Visual.*, vol. 6, no. 2, pp. 139–154, 2007.

[8] J. Venna and S. Kaski, "Nonlinear dimensionality reduction as information retrieval," in *Proc. AISTATS*07, the 11th Int. Conf. on Artificial Intelligence and Statistics, JMLR Workshop and Conf. Proc.*, vol. 2, M. Meila and X. Shen, eds. 2007, pp. 572–579.

[9] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *J. Mach. Learn. Res.*, vol. 11, pp. 451–490, Feb. 2010.

[10] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 833–840.

[11] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero III, "FINE: fisher information nonparametric embedding," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 11, pp. 2093–2098, 2009.

[12] S. Kaski, J. Sinkkonen, and J. Peltonen, "Bankruptcy analysis with self-organizing maps in learning metrics," *IEEE Trans. Neural Netw.*, vol. 12, no. 4, pp. 936–947, 2001.

[13] J. Peltonen, A. Klami, and S. Kaski, "Improved learning of Riemannian metrics for exploratory analysis," *Neural Netw.*, vol. 17, no. 8-9, pp. 1087–1100, 2004.

[14] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski, "Probabilistic retrieval and visualization of biologically relevant microarray experiments," *Bioinformatics*, vol. 25, no. 12, pp. i145–i153, 2009.

[15] C. Walshaw, "A multilevel algorithm for force-directed graph drawing," in *GD'00: Proc. 8th Int. Symp. on Graph Drawing*, London, UK. New York: Springer-Verlag, 2001, pp. 171–182.

[16] K. M. Hall, "An r-dimensional quadratic placement algorithm," *Manage. Sci.*, vol. 17, no. 3, pp. 219–229, 1970.

[17] J. Parkkinen, K. Nybo, J. Peltonen, and S. Kaski, "Graph visualization with latent variable models," in *Proc. MLG-2010, the 8th Workshop on Mining and Learning with Graphs*. New York, NY: ACM, 2010, pp. 94–101 DOI: http://doi.acm.org/10.1145/1830252.1830265.

[SP]

dsp EDUCATION

Çağatay Candan

On the Eigenstructure of DFT Matrices

The discrete Fourier transform (DFT) not only enables fast implementation of the discrete convolution operation, which is critical for the efficient processing of analog signals through digital means, but it also represents a rich and beautiful analytical structure that is interesting on its own. A typical senior-level digital signal processing (DSP) course involves a fairly detailed treatment of DFT and a list of related topics, such as circular shift, correlation, convolution operations, and the connection of circular operations with the linear operations [1]. Despite having detailed expositions on DFT, most DSP textbooks (including advanced ones) lack discussions on the eigenstructure of the DFT matrix. Here, we present a self-contained exposition on such.

Our goals are to study the eigenvalues and eigenvectors of the DFT matrix, to determine the multiplicity of the eigenvalues, to define the invariant subspaces under DFT mapping, to construct the projectors to the invariant subspaces and to underline some connections between invariant subspaces and other transforms.

(We believe that this discussion can be followed by most of the signal processing community, including advanced undergraduate students. The concepts used in this discussion are mostly elementary and available in standard linear algebra textbooks. A comprehensive knowledge of linear spaces is not required but would be highly beneficial to fully interpret some of the results. These notes have been prepared as an assignment for supplementary reading material on the DFT.)

Digital Object Identifier 10.1109/MSP.2010.940004
Date of publication: 17 February 2011

DESCRIPTION OF THE PROBLEM

Let $\mathbf{F}$ be a $N \times N$ unitary DFT matrix:

$$[\mathbf{F}]_{k,n} = \frac{1}{\sqrt{N}} e^{-j\frac{2\pi}{N}nk}.$$

In the equation above, $[\mathbf{F}]_{k,n}$ denotes the matrix entry in the kth row and nth column of the matrix $\mathbf{F}$. We assume both k and n run from 0 to $N-1$, following the literature on the DFT.

THE EIGENVALUES OF A MATRIX ARE, BY DEFINITION, THE ROOTS OF ITS CHARACTERISTIC POLYNOMIAL.

Different from the conventional definition given in [1], the definition above includes a scaling factor of $1\sqrt{N}$. This factor is required to make the matrix $\mathbf{F}$ unitary. From the theory of matrices, we know that the unitary matrices satisfy the relation $\mathbf{F}^H\mathbf{F} = \mathbf{I}$ (another form of Parseval's relation) and have unit norm eigenvalues and have a complete orthogonal set of eigenvectors [2]. Our goal is to study the eigenstructure of $\mathbf{F}$ matrices by finding the eigenvalues and their multiplicity, invariant subspaces and projectors to the invariant subspaces.

EIGENVALUES, EIGENSPACES AND PROJECTORS TO EIGENSPACES

The eigenvalues of a matrix are, by definition, the roots of its characteristic polynomial. Here we do not calculate the characteristic polynomial explicitly but relate the powers of $\mathbf{F}$ to the characteristic polynomial. Let $\mathbf{J}$ denote the second power of the matrix $\mathbf{F}$, that is $\mathbf{J} = \mathbf{F}^2$. The entries of matrix $\mathbf{J}$ can be calculated as follows:

$$[\mathbf{J}]_{k,n} = \sum_{d=0}^{N-1} [\mathbf{F}]_{k,d}[\mathbf{F}]_{d,n} = \frac{1}{N}\sum_{d=0}^{N-1} e^{-j\frac{2\pi}{N}(n+k)d} = \delta[(n+k)_N].$$

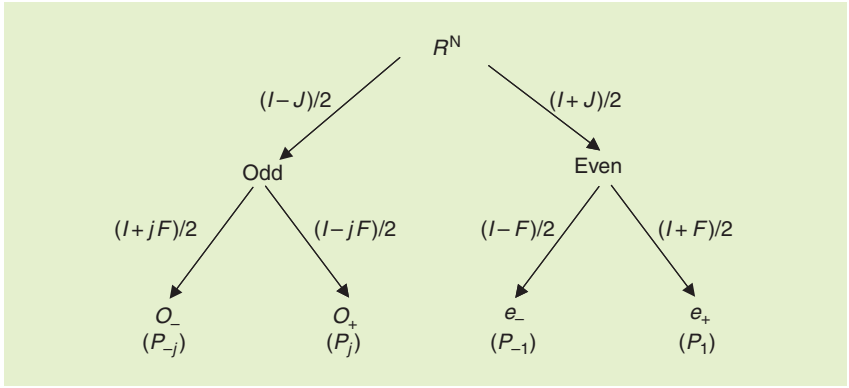
The notation of $(\cdot)_N$ indicates the modulo N reduction of $(\cdot)$, that is $(n+k)_N \equiv (n+k) \bmod N$. It can be seen that the $\mathbf{J}$ matrix is a permutation matrix that maps $x[n] \to x[(-n)_N]$. The $\mathbf{J}$ matrix is called a coordinate inversion or reflection matrix in the literature.

Two coordinate inversion operations executed in a row can be denoted by $\mathbf{J}^2$ or $\mathbf{F}^4$. Since two coordinate inversions result in the identity mapping, $\mathbf{F}^4$ is equal to $\mathbf{I}$. If $\mathbf{e}_k$ is an eigenvector of the $\mathbf{F}$ matrix with the eigenvalue λ_k, then the vector $\mathbf{F}^4\mathbf{e}_k$ should be equal to $\mathbf{F}^4\mathbf{e}_k = \lambda_k^4\mathbf{e}_k$, by the eigenvector definition. Using the identity $\mathbf{F}^4 = \mathbf{I}$ along with the last relation results in the conclusion that the possible values of λ must satisfy $\lambda_k^4 = 1$. Hence the list of possible eigenvalues for the DFT matrix is $\lambda_k = \{1, -1, j, -j\}$.

Having established the list of possible eigenvalues, we construct a $p_1(\lambda)$ polynomial having roots at $\{-1, j, -j\}$ and taking the value of 1 at $\lambda = 1$. Hence, this polynomial takes the value of zero for all except one of the eigenvalues of DFT matrix. This polynomial can be explicitly written as follows

$$p_1(\lambda) = \frac{1}{4}(\lambda^2+1)(\lambda+1) = \frac{1}{4}(\lambda^3+\lambda^2+\lambda+1).$$

1053-5888/11/$26.00©2011IEEE



[FIG1] Decomposition of *N* dimension space into even–odd sub-spaces.

When $\mathbf{F}$ is substituted for λ in $p_1(\lambda)$, we get the $\mathbf{P}_1$ matrix:

$$\mathbf{P}_1 = \frac{1}{4}(\mathbf{F}^3 + \mathbf{F}^2 + \mathbf{F} + \mathbf{I}).$$

When the $\mathbf{P}_1$ matrix is multiplied from right with an eigenvector of the DFT matrix having eigenvalue λ_k, the resultant vector is as given:

$$\mathbf{P}_1\mathbf{e}_k = \begin{cases} \mathbf{0} & \lambda_k = \{-1, j, -j\} \\ \mathbf{e}_k & \lambda_k = 1 \end{cases}.$$

The last relation shows that the eigenvectors of DFT with the eigenvalue of 1 pass through $\mathbf{P}_1$ without any change (mapped to itself) and the other eigenvectors are projected to the zero vector, i.e., elements of null space. Since the eigenvectors of DFT are complete, i.e., span N dimensional space, $\mathbf{e}_k$ vectors form a complete set of eigenvectors for the $\mathbf{P}_1$ matrix. From this information, we can deduce that the matrix $\mathbf{P}_1$ has only two eigenvalues that can be either 0 or 1. This leads to the conclusion that $\mathbf{P}_1$ is a projection matrix [2].

The projection matrices satisfy the relation $\mathbf{P}^2 = \mathbf{P}$. Among the projection matrices, the matrices with the property $\mathbf{P}^T = \mathbf{P}$ are called the orthogonal projectors. With these facts, we can confirm that the matrix $\mathbf{P}_1$ is an orthogonal projector to the range space of DFT eigenvectors having the eigenvalue of 1.

Following the same route, we can write the projectors to four eigenspaces as follows:

$$\mathbf{P}_1 = \frac{1}{4}(\mathbf{F}^3 + \mathbf{F}^2 + \mathbf{F} + \mathbf{I})$$

$$\mathbf{P}_{-1} = \frac{1}{4}(-\mathbf{F}^3 + \mathbf{F}^2 - \mathbf{F} + \mathbf{I})$$

$$\mathbf{P}_j = \frac{1}{4}(j\mathbf{F}^3 - \mathbf{F}^2 - j\mathbf{F} + \mathbf{I})$$

$$\mathbf{P}_{-j} = \frac{1}{4}(-j\mathbf{F}^3 - \mathbf{F}^2 + j\mathbf{F} + \mathbf{I}).$$

Below we present a summary of our current findings along with some new, but easy to establish, results on $\mathbf{P}_k$ matrices:

- $\mathbf{P}_k$ matrices are orthogonal projectors, i.e., $\mathbf{P}_k^2 = \mathbf{P}_k$ and $\mathbf{P}_k^T = \mathbf{P}_k$.
- The projection matrices are complementary ($\mathbf{P}_k\mathbf{P}_l = 0$, $k \neq l$).
- The direct sum of the projection subspaces is $\mathcal{R}^N$.
- The projection subspaces are invariant under DFT, that is, $\mathbf{F}\mathbf{P}_k = \mathbf{P}_k\mathbf{F} = \lambda_k\mathbf{P}_k$.
- $\mathbf{P}_1 + \mathbf{P}_{-1}$ is the projector to the space spanned by even vectors, that is, $\mathbf{E} = \mathbf{P}_1 + \mathbf{P}_{-1} = 1/2(\mathbf{I} + \mathbf{J})$ and $\mathbf{E}\{x[n]\} = 1/2(x[n] + x[(-n)_N])$.
- $\mathbf{P}_j + \mathbf{P}_{-j}$ is the projector to the space spanned by odd vectors, that is, $\mathbf{O} = \mathbf{P}_j + \mathbf{P}_{-j} = 1/2(\mathbf{I} - \mathbf{J})$ and $\mathbf{O}\{x[n]\} = 1/2(x[n] - x[(-n)_N])$.

The results given above can be verified by algebraic multiplication and addition of $\mathbf{P}_k$ matrices. However, we would like to encourage readers not to interpret these results algebraically, but through the concepts of linear spaces, e.g., subspace, range space, and null space. As an example, $\mathbf{P}_1$ is the projector to the space spanned by the eigenvectors with the eigenvalue of one, that is,

$$\mathbf{P}_1 = \sum_{k=1}^{m_1} \mathbf{e}_k^1 (\mathbf{e}_k^1)^T. \qquad (1)$$

Here m_1 is the multiplicity of the eigenvalue and $\mathbf{e}_k^1$ is the kth eigenvector with the eigenvalue of one. The first and second results given above immediately follow from the definition in (1) and the orthogonality of the eigenvectors with different eigenvalues. The third result is due to the completeness of the eigenvectors. The other results can be interpreted similarly with a little bit of effort.

Up to this point we have studied how to construct the projection matrices for the invariant subspaces of the DFT matrix. It is well known that DFT maps even sequences to even sequences and odd sequences to odd sequences. Hence the subspace of even sequences and odd sequences are invariant under DFT. Here we generalize the invariance property of even and odd subspaces. We show that a vector in $\mathbf{P}_k$ space is mapped to another vector in $\mathbf{P}_k$ space. With this interpretation we can say that $\mathbf{P}_k$ matrices partition even and odd subspaces into two, as shown in Figure 1.

THE MULTIPLICITY OF EIGENVALUES

The eigenvalue multiplicity problem of DFT matrices is known to be a difficult problem. We present a solution to the eigenvalue multiplicity problem using an equally difficult result known as the Gaussian sum. The Gaussian sum identity is given below:

Guassian sum:

$$\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} e^{j\frac{2\pi}{N}n^2} = \begin{cases} 1+j & N=4m \\ 1 & N=4m+1 \\ 0 & N=4m+2 \\ j & N=4m+3 \end{cases}.$$

The proof of this result took Gauss two years [3]. Since the original proof of Gauss, it is an ongoing challenge among mathematicians to present new, possibly better, proofs of this result. Interested readers can find four different proofs of Mertens, Kronecker, Schur, and Gauss in [3]. In 1972, J. McClellan solved the eigenvalue multiplicity problem using elementary means [4]. McClellan's solution can be considered as another proof of the Gaussian sum and resides at the intersection of pure and applied

[TABLE I] EIGENVALUE MULTIPLICITY OF $N \times N$ DFT MATRIX.

N	$\lambda = 1$	$\lambda = -1$	$\lambda = J$	$\lambda = -J$
4M	M+1	M	M−1	M
4M+1	M+1	M	M	M
4M+2	M+1	M+1	M	M
4M+3	M+1	M+1	M	M+1

mathematics as noted in [5]. McClellan is also known for an optimal filter design technique (Parks-McClellan algorithm) and a multidimensional filter design technique through mapping (McClellan transform) to the DSP community. Here we do not attempt to prove the Gaussian sum and but use the relation for the solution of the DFT eigenvalue multiplicity problem.

It can be noted that the trace of the matrix $\mathbf{P}_k$ is equal to the multiplicity of the eigenvalue with value λ_k. This can be justified from equation (1) by using the identity $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$. Another justification can be given by noting that the projection matrices have eigenvalues either zero or one. Therefore the trace, which is the sum of the eigenvalues, is equal to the number of eigenvalues with value one.

The trace of the projection matrices can be written as follows:

$$\text{trace}\{\mathbf{P}_1\} = \frac{1}{4}\left\{\frac{2}{\sqrt{N}}\sum_{n=0}^{N-1}\cos\left(\frac{2\pi}{N}n^2\right) + \text{trace}\{\mathbf{J}\} + N\right\}$$

$$\text{trace}\{\mathbf{P}_{-1}\} = \frac{1}{4}\left\{-\frac{2}{\sqrt{N}}\sum_{n=0}^{N-1}\cos\left(\frac{2\pi}{N}n^2\right) + \text{trace}\{\mathbf{J}\} + N\right\}$$

$$\text{trace}\{\mathbf{P}_j\} = \frac{1}{4}\left\{\frac{2}{\sqrt{N}}\sum_{n=0}^{N-1}\sin\left(\frac{2\pi}{N}n^2\right) - \text{trace}\{\mathbf{J}\} + N\right\}$$

$$\text{trace}\{\mathbf{P}_{-j}\} = \frac{1}{4}\left\{\frac{2}{\sqrt{N}}\sum_{n=0}^{N-1}\sin\left(\frac{2\pi}{N}n^2\right) - \text{trace}\{\mathbf{J}\} + N\right\}.$$

Using the Gaussian summation, the trace of each matrix can be easily calculated and the eigenvalue multiplicity of DFT matrices can be found as shown in Table 1 (we note that the trace of the $N \times N$ **J** matrix is equal to one and two for odd and even values of N, respectively).

EIGENVECTORS OF DFT MATRIX

The eigenvector set of DFT matrices for $N \geq 4$ are not unique due to the eigenvalue multiplicity problem as shown in Table 1. The table indicates that there are infinitely many eigenvector sets of DFT matrix.

An eigenvector set of DFT can be easily constructed using projection matrices. Since the projection spaces are invariant under DFT operation, that is, $\mathbf{FP}_k = \lambda_k \mathbf{P}_k$, the columns of projection matrix $\mathbf{P}_k$ are the eigenvectors of DFT. Unfortunately, this eigenvector set does have the orthogonality property. If the orthogonality of eigenvectors is desired, one can apply the Gram-Schmidt procedure over the columns of $\mathbf{P}_k$. This operation can be done with a few lines of MATLAB code as shown below:

WE HAVE EXAMINED THE STRUCTURE OF DFT EIGENSPACES AND USED THE PROJECTORS TO THE INVARIANT SPACES TO ESTABLISH SOME CONNECTIONS WITH THE RELATIVES OF THE FOURIER TRANSFORM.

```
>> N=7; F = 1/
   sqrt(N)*dftmtx(N);
>> P1 = 0.25 * (F^3 + F^2 +
   F + eye(N));
>> E1 = orth(P1);
```

To get a distinct set of orthogonal eigenvectors with eigenvalue of 1, we can modify the last line as follows:

```
>> E1=orth(P1*randn(N,N))
```

An alternative approach is to define a commuting matrix **K** through an arbitrary but a full-rank matrix **M** as shown below:

$$\mathbf{K} = \mathbf{M} + \mathbf{FMF}^{-1} + \mathbf{F}^2\mathbf{MF}^{-2} + \mathbf{F}^3\mathbf{MF}^{-3}.$$

It is easy to show that matrices **K** and **F** commute; that is, $\mathbf{FK} = \mathbf{KF}$, therefore **K** and **F** have a common eigenvector set, [7, p. 52]. In other words, by finding the eigenvectors of **K** matrix, we can also get the eigenvectors of the DFT matrix. This technique has been applied to derive the eigenvectors of the DFT with some desirable features. In [8] and [9], the discrete equivalents of Hermite-Gaussian functions (which are the eigenfunctions of continuous Fourier transform) are defined by a proper choice of **M** matrix.

EXTENSIONS

Up to this point we have presented results on a one-dimensional conventional DFT matrix. In this section, we extend the earlier results to non-conventional DFT matrices and multidimensional DFT matrices, and establish some connections with other relatives of the Fourier transform.

EIGENSTRUCTURE OF OFFSET DFT

The offset DFT is a generalization of the conventional DFT. Its definition is given as follows:

$$[\mathbf{F}_{a,b}]_{k,n} = \frac{1}{\sqrt{N}} e^{-j\frac{2\pi}{N}(k-a)(n-b)}.$$

The offset DFT has two parameters (a and b) that can be freely selected. It can be shown that the offset DFT matrix is unitary and reduces to the conventional DFT when $a = b = 0$ [10]. The special case of $a = b = 1/2$, is called an odd-time odd-frequency DFT and was studied in [11]. The eigenstructure of the offset DFT has been shown to be closely related to ordinary DFT for the special case of $a = b = 1/2$, [12]. The other cases are a little more complicated and studied under the categories of $a + b =$ integer and $a + b \neq$ integer. Further details can be found in [13].

EIGENSTRUCTURE OF MULTIDIMENSIONAL DFT

By definition, the multidimensional DFT is a separable transformation. Hence a two-dimensional DFT operation can be interpreted as the cascade

application of a one-dimensional DFT to the columns of the input (a matrix) followed by the application of DFT to the rows of the resultant matrix. The separability property aids in identifying the eigenstructure of multidimensional DFT.

It can be noted that the following $M \times N$ rank-1 matrix is an eigenmatrix of two dimensional DFT with the eigenvalue $\lambda_x\lambda_y$:

$$\mathbf{E} = \mathbf{e}_x\mathbf{e}_y^{\mathrm{T}}. \tag{2}$$

Here $\mathbf{e}_x$ is an eigenvector of $M \times M$ one-dimensional DFT matrix with the eigenvalue λ_x and $\mathbf{e}_y$ is an eigenvector of $N \times N$ one-dimensional DFT matrix with the eigenvalue λ_y. From this discussion, it can be noted that the set of eigenvalues of a two-dimensional DFT is identical to the corresponding set of a one-dimensional transform. The results on a two-dimensional DFT can be easily extended to multidimensions. More details can be found in [14].

RELATIONS TO OTHER TRANSFORMS

The projectors to the invariant spaces of the DFT can be useful to characterize other relatives of the Fourier transform. The following lines show the relation between the projectors and DFT, Hartley transform, identity, and coordinate inversion operations respectively:

$$\mathbf{F} = \mathbf{P}_1 - \mathbf{P}_{-1} + j\mathbf{P}_{\mathrm{j}} - j\mathbf{P}_{-\mathrm{j}}$$

$$\mathbf{H} = \mathbf{P}_1 - \mathbf{P}_{-1} - \mathbf{P}_{\mathrm{j}} + \mathbf{P}_{-\mathrm{j}}$$

$$\mathbf{I} = \mathbf{P}_1 + \mathbf{P}_{-1} + \mathbf{P}_{\mathrm{j}} + \mathbf{P}_{-\mathrm{j}}$$

$$\mathbf{J} = \mathbf{P}_1 + \mathbf{P}_{-1} - \mathbf{P}_{\mathrm{j}} - \mathbf{P}_{-\mathrm{j}}.$$

It can be noted that the projectors define an algebra for the relatives of the Fourier transform. As an illustrative example, the transformation formed by the cascade application of a Hartley transform and a DFT transform, that is, an $\mathbf{FH}$ matrix can be expressed in terms of projectors as follows:

$$\begin{aligned}\mathbf{FH} &= (\mathbf{P}_1 - \mathbf{P}_{-1} + j\mathbf{P}_{\mathrm{j}} - j\mathbf{P}_{-\mathrm{j}}) \\ &\quad \times (\mathbf{P}_1 - \mathbf{P}_{-1} - \mathbf{P}_{\mathrm{j}} + \mathbf{P}_{-\mathrm{j}}) \\ &= \mathbf{P}_1 + \mathbf{P}_{-1} - j\mathbf{P}_{\mathrm{j}} - j\mathbf{P}_{-\mathrm{j}} \\ &= \frac{1}{2}(\mathbf{I} + \mathbf{J}) - \frac{j}{2}(\mathbf{I} - \mathbf{J}) \\ &= \mathbf{E} - j\mathbf{O}.\end{aligned}$$

From this result, we can conclude that the cascade operation of Hartley and DFT is equivalent to expressing even and odd parts of the input and combining them together as the real and imaginary parts of the output.

The fractional powers or any other function of $\mathbf{F}$ can also be defined through the projectors. We illustrate the idea on the square root of a DFT matrix. The square root or one half power of a DFT matrix can be defined as follows $\mathbf{F}^{\frac{1}{2}} \triangleq \sqrt{1}\mathbf{P}_1 + \sqrt{-1}\mathbf{P}_{-1} + \sqrt{j}\mathbf{P}_{\mathrm{j}} + \sqrt{-j}\mathbf{P}_{-\mathrm{j}}$. Since the square root operation is one-to-many, that is $\sqrt{1} = \{1, -1\}$, the proposed definition is not unique unless a branch-cut for every square root is specified. A possible definition is $\mathbf{F}^{\frac{1}{2}} \triangleq \mathbf{P}_1 + j\mathbf{P}_{-1} + \frac{1+j}{2}\mathbf{P}_{\mathrm{j}} + \frac{1-j}{2}\mathbf{P}_{-\mathrm{j}}$. One can easily note that $\mathbf{F}^{\frac{1}{2}}\mathbf{F}^{\frac{1}{2}} = \mathbf{F}$ as expected. More information on the fractional Fourier transform and the details of the definition multiplicity problem can be found in [15].

THE PROJECTORS TO THE INVARIANT SPACES OF THE DFT CAN BE USEFUL TO CHARACTERIZE OTHER RELATIVES OF THE FOURIER TRANSFORM.

CONCLUSIONS

We have examined the structure of DFT eigenspaces and used the projectors to the invariant spaces to establish some connections with the relatives of the Fourier transform. The presented results are heavily based on the properties of projectors that can also be of interest on their own due to their strong algebraic structure and important geometric interpretations.

AUTHOR

Çağatay Candan (ccandan@metu.edu.tr) is with the Electrical and Electronics Engineering Department of Middle East Technical University (METU), Ankara, Turkey.

REFERENCES

[1] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1999.

[2] G. Strang, Linear Algebra and Its Applications. *Pacific Grove*, CA: Brooks/Cole, 1988.

[3] E. Landau, Elementary Number Theory. New York: Chelsea Publishing Co., 1966.

[4] J. McClellan and T. Parks, "Eigenvalue and eigenvector decomposition of the discrete Fourier transform," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 20, no. 1, pp. 66–74, 1972.

[5] L. Auslander and R. Tolmieri, "Is computing with the finite Fourier transform pure or applied mathematics?" *Bull. Am. Math. Soc.*, vol. 1, no. 6, pp. 847–897, 1979.

[6] C. Candan, "Discrete fractional Fourier transform," M.S. thesis, Bilkent Univ., Ankara, Turkey, 1998.

[7] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. London, U.K.: Oxford Univ. Press, 1988.

[8] C. Candan, "On higher order approximations for Hermite Gaussian functions and discrete fractional Fourier transforms," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 699–702, 2007.

[9] B. Dickinson and K. Steiglitz, "Eigenvectors and functions of the discrete Fourier transform," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 30, no. 1, pp. 25–31, 1982.

[10] G. Bongiovanni, P. Corsini, and G. Frosini, "One-dimensional and two-dimensional generalised discrete Fourier transforms," *IEEE Trans. Acoustics, Speech Signal Process.*, vol. 24, pp. 97–99, Feb. 1976.

[11] G. Bonnerot and M. Bellanger, "Odd-time odd-frequency discrete Fourier transform for symmetric real-valued series," *Proc. IEEE*, vol. 64, pp. 392–393, Mar. 1976.

[12] C.-C. Tseng, "Eigenvalues and eigenvectors of generalized DFT, generalized DHT, DCT-IV and DST-IV matrices," *IEEE Trans. Signal Process.*, vol. 50, pp. 866–877, Apr. 2002.

[13] S.-C. Pei and J.-J. Ding, "Generalized eigenvectors and fractionalization of offset DFTs and DCTs," *IEEE Trans. Signal Process.*, vol. 52, pp. 2032–2046, July 2004.

[14] N. K. Bose, "Eigenvectors and eigenvalues of 1-D and n-D DFT matrices," *Int. J. Electr. Commun.*, vol. 55, no. 2, pp. 131–133, 2001.

[15] H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform*. New York: Wiley, 2001. [SP]

David V. Anderson

Storytelling—The Missing Art in Engineering Presentations

It happened again. I had just finished sitting through the presentation and defense of a student's Ph.D. proposal, and it simply couldn't be described as a success. Why?! The student had prepared well, followed the basic rules for a good presentation, spoken clearly, and kept within the allotted time; yet something was missing. The fact that something was missing was even clearer as members of the committee struggled to find the significance of the completed and proposed work. Having first-hand knowledge of this student's work, I knew that it was well done and that he had made useful contributions to his field. Plots and tables showing the effectiveness of his approach were included in the presentation. One might expect him to have received congratulatory pats on the back instead of skepticism and doubt.

This student fell victim to a common problem in engineering and perhaps in other disciplines as well. He forgot the story that made his research exciting, and in his desire to impress, turned the presentation into a series of plots, equations, and facts that left the audience nearly comatose. That "something missing" was the storytelling.

WHY STORYTELLING?

The human brain is not a computer. Although it might simplify education if we could simply dump all necessary raw facts directly into a person's brain, that is still the stuff of science fiction. Throughout the ages, people have taught their children and others using stories. Stories convey not only information but experience and wisdom, and they excite sympathetic reactions within the listeners that enable them to apply the learned concepts to other situations. Our brains are simply better equipped to understand and retain narrative and information in context than to retain bare facts.

Technical topics, however, may seem to defy a narrative or "storytelling" approach. For example, precision is often paramount, and such precision may best be represented in terms of equations, plots, and tables of data. Do not despair, you too can enchant your audiences with gripping tales of multidimensional analysis and measurement techniques.

STORIES CONVEY NOT ONLY INFORMATION BUT EXPERIENCE AND WISDOM, AND THEY EXCITE SYMPATHETIC REACTIONS WITHIN THE LISTENERS THAT ENABLE THEM TO APPLY THE LEARNED CONCEPTS TO OTHER SITUATIONS.

STORYTELLING IN ENGINEERING?

"But I only have 20 minutes to present my paper! How can I fit in a plot, character development, and a battle of good versus evil, and still have time to get to my results?"

Let me assure you, it is possible. A good story must have plot, characters, and dramatic appeal; and this goes for stories in engineering as well. Instead of villains, we have problems—they could be noise, theoretical limits, or practical limitations. To set up the drama, we can relate failed attempts to solve a problem or potential benefits from being the first to find a solution. Acting the role of the protagonist is our newly developed method or mathematical approach. And, for a plot, we can relate the saga and suspense of seeking a solution through forests of data and impenetrable mathematics.

Now, before you simply turn all your presentations into fairy tales, let us look a little more into these concepts. First, when talking about presentations and storytelling in the same breath, most people think of stories that are included within a presentation. These stories can introduce, punctuate, clarify, or emphasize points of a presentation or lecture. As an undergraduate student taking thermodynamics, I had resigned myself to a dreary semester of looking up thermodynamic properties of select substances in countless tables. Then one day, we had a guest lecturer who explained how a coal-fired power plant works—it essentially has a flame 12 stories high! Suddenly, heat transfer became much more interesting.

This type of storytelling, also known as "sharing real-life experiences," is extremely valuable, but it is not what we are talking about in this discussion on storytelling. So, let us go on to see how we can make a compelling story out of a simple technical presentation itself.

HOW TO TURN PRESENTATIONS INTO STORIES

"Effective storytelling is a fine and beautiful art. A well-developed and presented story can cut across age barriers and will hold the interest and reach its listeners. Stories will be remembered long after other orations" [1].

It is easier to be a comedian if you find humor in the vicissitudes of life. Likewise, stories spring up from a well of

Digital Object Identifier 10.1109/MSP.2011.940239
Date of publication: 17 February 2011

fascination with life and your surroundings. The key is to convey that fascination to those around you. This takes practice and attention and you will need to develop your own style. But to get you started, here are eight suggestions to help you enliven your presentations and make them memorable and effective:

- Share the love.
- Know your audience.
- Pay attention to the big picture.
- Learn the history.
- Try to explain it to a nontechnical friend.
- Follow a pattern of tension and resolution.
- Practice.
- Don't overdo it.

SHARE THE LOVE

This is all about conveying your "fascination" for your topic. Why are you involved in this work? What makes it interesting to you? I hope that answers to those questions spring readily to mind, otherwise you may have more problems than just making a good presentation. If answers to these questions do not come easily, then you may need to think more about them as you are actually doing the work. It will make your work more rewarding and productive.

Now, once you remember why you love your topic, you are ready for the next step, which is to share your love of your topic or material with the audience. Ask yourself—what might other people find interesting? Give this some thought. I have endured more than one presentation in which I know the presenter has a love for his or her topic but did not share that with the audience. Perhaps this was for one of several reasons: one, they were afraid that others would not find it interesting and so did not even try; two, they were unaware that their audience wouldn't automatically feel the same way they did; or three, they were under the impression that technical presentations should be dry. Technical presentations should not be dry. They should be interesting and compelling, and the presenter should exhibit enthusiasm for the topic. It also helps substantially if you know how to connect with your audience.

KNOW YOUR AUDIENCE

"Know your audience" is a proverb in public speaking. What does it mean to know your audience? To begin with, you should know what the audience wants from your presentation. Whether you give them what they want may be a different discussion (they may want free sports cars), but you should try to understand their expectations as a starting point. Then, estimate their ability to understand your presentation, their level of endurance, and how your topic relates to their primary interests. The single thing that often sets apart an excellent teacher from a poor teacher is the ability to understand and connect with the audience.

Before your presentation, learn what you can about your audience. Visit with audience members, if possible, beforehand. During the presentation, observe what resonates with the audience and what does not, then adapt accordingly. This will take practice, but the effort made in learning these skills will pay off for the rest of your life.

WHEN LOOKING AT THE "BIG PICTURE," IT IS OFTEN HELPFUL TO ZOOM OUT EVEN FURTHER THAN JUST YOUR OWN WORK.

PAY ATTENTION TO THE BIG PICTURE

A good novel will have a main plot and it may have many subplots, but it will not have random unrelated points. Some things just do not contribute to good stories or good presentations. Isolated details fall into that category. When deciding if something should be included in a presentation, ask yourself how it contributes to the overall story. If it is important, then it should be properly incorporated into your story. If it does not contribute, then leave it out.

Consider, for example, how you might discuss a detail in the simplification of a mathematical expression. When replacing e^{κ} with $1+\kappa$ you probably had a good reason; but you must ask yourself how it contributes to the main message or big picture. We will assume that it is important and that it does contribute because if it was not important, then you would not have included it in your presentation. Since it is important, then it is worth presenting properly.

How do you present details properly? You first explain the context or motivation that makes the detail relevant to the rest of the story. Only after presenting the motivation should you share the detail—it will then fit nicely into the narrative in the place prepared for it. Back to our e^{κ} approximation, you might ask yourself why you made the approximation and why the audience would care. How did that change affect the larger system? You might explain that no closed-form solution to the original equation is known but that by approximating $e^{\kappa} \approx 1+\kappa$ it is possible to generate a closed form solution. Then you could explain that the approximation is significant for them to remember because it limits the range over which the approximate solution is valid; thereby limiting the scope of the solution.

A good rule of thumb for keeping the big picture in mind is to provide context or motivation before presenting details so that the audience can easily understand the significance and relative importance of each point that you make. If a detail is not worth that effort, then it is likely not worthwhile including in your presentation.

LEARN THE HISTORY

When looking at the "big picture," it is often helpful to zoom out even further than just your own work. Are you the first person to look at this topic? If so, then either you are a genius of the first order or you picked a topic no one cares about. If you are not the first person to look at the topic (as is more common), then be prepared to set the background

for your work by discussing the successes and shortcomings of those who went before. Build it up so that now you will be presenting the culminating chapter in the great saga that began with those pioneers in your field. Your audience will naturally be more interested in your efforts, and it will make it that much easier for them to understand its significance. As any sports fan knows, a game is much more interesting to watch when you know the teams and players and their struggles and triumphs.

TRY TO EXPLAIN IT TO A NONTECHNICAL FRIEND

Once you have the big picture and the backstory, it is time to test it out on a willing subject. This is not a practice run of your presentation but more like an elevator pitch. (An elevator pitch is a short but compelling overview on some topic that could presumably be given in the brief time during which two people are together in an elevator.) The goal is to explain your subject so that a nontechnically inclined friend finds it interesting and is not lost among details or esoteric terms. If you succeed, then you have likely established the overall story. If you did not succeed, it is time to reevaluate your approach and then try again (hopefully your friend is patient).

FOLLOW A PATTERN OF TENSION AND RESOLUTION

A good comedian never reveals the punch line until the joke is set up. Similarly, do not show results until you have audience in a state of anticipation. (This is closely related to the method for including details discussed above.) In practice, this means that when giving a presentation, your audience should be eager for the next slide. Consider these two examples:

1) "Next we used a nearest-neighbor approach to remove measurement noise on the critical axis. The plot is on the next slide."

2) "Now that we had discovered how to capture the data, we needed to find a way to minimize the error on the critical axis while leaving all other dimensions unchanged. However, no one has ever been able to do so. The difficulty is that…. After trying the usual approaches, we realized that by plotting the data versus the nearest-neighbor density, it might be possible to remove the effects of most of the measurement error. The plot on the next slide shows the average error before and after processing using this approach. If we were successful, the plot would look like …."

This example is a bit long winded because it combines several examples of how to build anticipation. There are many variations on this theme. The main thing is to bring the listeners along with you as you recreate the tension and thrill associated with finding a new approach or uncovering new truths. I still remember clearly a concept taught in one of my graduate classes. Prof. Monty Hayes was discussing methods of estimating the spectrum of a signal. What made it memorable was the historical walk through the topic with enough discussion of the advantages and disadvantages of each newly discovered method that the students could make the intuitive leap to anticipate the next discovery. This recreated the thrill of discovery for each listener and created a deeper interest in and for the subject matter.

PRACTICE

A good joke can be made or destroyed by the delivery—the timing, level of detail, and phrasing are essential. A presentation can also sink or swim on delivery. Practice is essential for identifying and correcting awkward parts of your presentation. Furthermore, to implement some of the suggestions in this article, such as following a pattern of tension and release, you must know what slide comes next at each point. Without practice, it becomes much more difficult to have a smooth and natural flow that interweaves your story and the supporting slides.

DON'T OVERDO IT

Warning—As when using spices in cooking, these techniques should always be used with judgment and moderation. Remember that the purpose of a technical presentation is to convey certain material, not to entertain. Although it may be possible to do both, care must be taken to avoid giving a presentation with plenty of style but insufficient substance. The use of storytelling techniques as described in this article should be subtle tweaks to your presentation style and may not be all used at once.

These suggestions are likely to improve your technical presentations, but they cannot be used as a substitute for good basic skills. See, for example, "Effective Communication: Excellence in a Technical Presentation" by Wayne Padgett and Mark Yoder [2].

THE END OF THE STORY

Back to the student at the beginning of the story. (Sidenote: several of my students have read this article, and they each think that the person at the beginning is them. Although the events are real, I am specifically not writing this with any one person in mind.) By the time of his next presentation he had learned the basic concepts of storytelling, and his presentation was accompanied by congratulations and pats on the back.

AUTHOR

David V. Anderson (anderson@gatech.edu) is with the School of Electrical and Computer Engineering at the Georgia Institute of Technology.

REFERENCES

[1] B. McWilliams. Effective storytelling: a manual for beginners [Online]. Available: http://www.eldrbarry.net/roos/eest.htm

[2] W. T. Padgett and M. A. Yoder, "Effective communication: excellence in a technical presentation," *IEEE Signal Process. Mag.*, vol. 25, pp. 124–127, Mar. 2008.

[SP]

dsp TIPS&TRICKS

Richard Lyons

Reducing FFT Scalloping Loss Errors Without Multiplication

"DSP Tips and Tricks" introduces practical design and implementation signal processing algorithms that you may wish to incorporate into your designs. We welcome readers to submit their contributions. Contact Associate Editors Rick Lyons (R.Lyons@ieee.org) or Clay Turner (clay@claysturner.com).

This article discusses the estimation of time-domain sinewave peak amplitudes based on the fast Fourier transform (FFT) data. Such an operation sounds simple, but the scalloping loss characteristic of FFTs complicates the procedure. Here we present novel multiplier-free methods to accurately estimate sinewave amplitudes, based on FFT data, that greatly reduce scalloping loss problems.

FFT SCALLOPING LOSS REVISITED

There are many applications that require the estimation of a time-domain sinewave's peak amplitude based on FFT data. Such applications include oscillator and analog-to-digital converter performance measurements, as well as standard total harmonic distortion (THD) testing. However, the scalloping loss inherent in FFTs creates an uncertainty in such time-domain peak amplitude estimations. This section provides a brief review of FFT scalloping loss.

As most of you know, if we perform an N-point FFT on N real-valued time-domain samples of a discrete sinewave, whose frequency is an integer multiple of f_s/N (f_s is the sample rate in hertz), the peak magnitude of the sinewave's positive-frequency spectral component will be

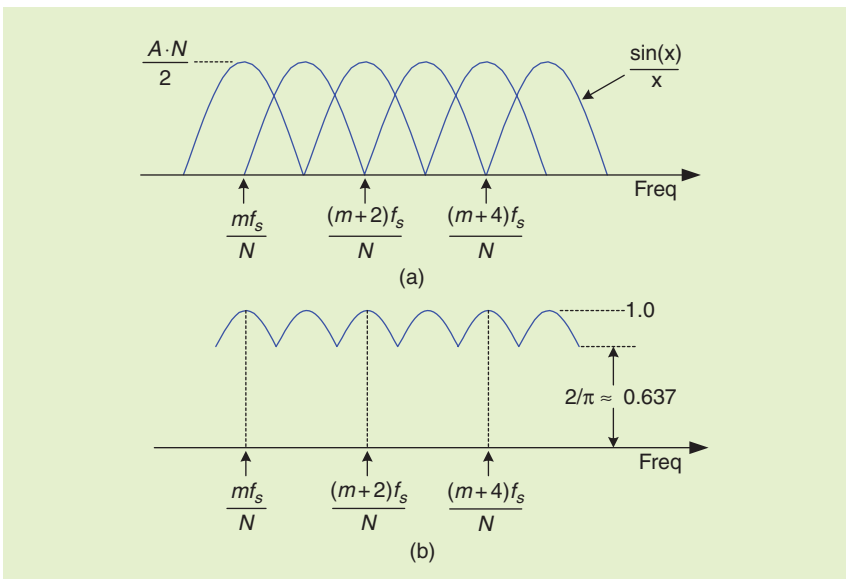
$$M = \frac{A \cdot N}{2}, \tag{1}$$

where A is the peak amplitude of the time-domain sinewave. That phrase "whose frequency is an integer multiple of f_s/N" means that the sinewave's frequency is located exactly at one of the FFT's bin centers.

Now, if an FFT's input sinewave's frequency is between two FFT bin centers (equal to a noninteger multiple of f_s/N), the FFT magnitude of that spectral component will be less that the value of M in (1). Figure 1 illustrates this behavior. Figure 1(a) shows the frequency responses of individual FFT bins where, for simplicity, we show only the main lobes (no side lobes) of the FFT bins' responses. What this means is that if we were to apply a sinewave to an FFT and scan the frequency of that sinewave over multiple bins, the magnitude of the FFT's largest normalized magnitude sample value will follow the curve in Figure 1(b). That curve describes what is called the "scalloping loss" of an FFT [1].

(As an aside, the word scallop is not related to my favorite shellfish. As it turns out, some window drapery, and tablecloths, do not have linear borders. Rather they have a series of circular segments, or loops, of fabric defining their decorative borders. Those loops of fabric are called scallops.)

What Figure 1(b) tells us is that if we examine the N-point FFT magnitude sample of an arbitrary-frequency, peak amplitude = A sinewave, that spectral component's measured peak magnitude M_{peak} can be in anywhere in the range of:



[FIG1] FFT frequency magnitude responses: (a) individual FFT bins and (b) overall FFT response.

Digital Object Identifier 10.1109/MSP.2010.939845
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

$$\frac{0.637 \cdot A \cdot N}{2} \le M_{\text{peak}} \le \frac{A \cdot N}{2} \quad (2)$$

depending on the frequency of that sinewave. This is shown as the *rectangular window* curve in Figure 2, where the maximum scalloping error occurs at a frequency midpoint between two FFT bins. The variable M in Figure 2 is the M from (1). So if we want to determine a sinewave's time-domain peak amplitude A, by measuring its maximum FFT spectral peak magnitude M_{peak}, our estimated value of A, from (1), using

$$A = \frac{2M_{\text{peak}}}{N} \quad (3)$$

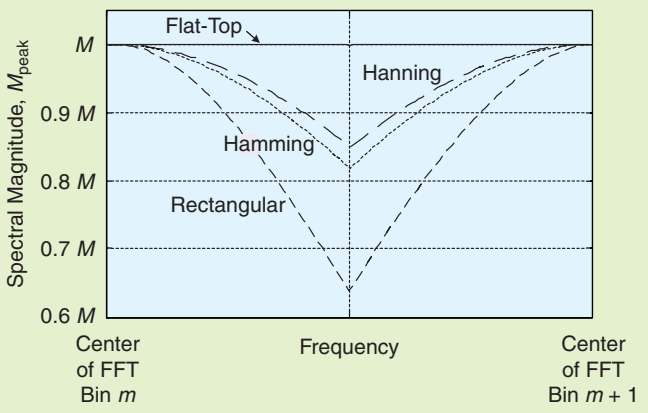
can have an error as great as 36.3%. In many spectrum analysis applications such a large potential error, equivalent to 3.9 dB, is unacceptable. As shown by the magnitude-normalized curves in Figure 2, Hanning and Hamming windowing of the FFT input data reduce the unpleasant frequency-dependent fluctuations in a measured spectral M_{peak} value but not nearly enough to satisfy many applications.

One solution to this frequency-dependent, FFT-based, measured amplitude uncertainty is to multiply the original N time-domain samples by an N-sample flat-top window function and then perform a new FFT on the windowed data. Flat-top window functions are designed to overcome the scallop loss inherent in rectangular-windowed FFTs. While such a flat-top-windowed FFT technique will work, there are more computationally efficient methods to solve our signal peak amplitude estimation uncertainty problem.

FREQUENCY-DOMAIN CONVOLUTION

Because multiplication in the time domain is equivalent to convolution in the frequency domain, we can convert rectangular-windowed (no windowing) FFT samples to windowed-FFT samples by way of convolution. For example, consider an N-point $w(n)$ window sequences whose time-domain samples are generated using

$$w(n) = \sum_{k=0}^{K-1} (-1)^k h_k \cos(2\pi kn/N), \quad (4)$$



[FIG2] **Windowed-FFT, bin-to-bin, frequency magnitude responses.**

where the $w(n)$ sequence's generating polynomial has an integer K number of h_k coefficients.

Many window functions, including Hanning, Hamming, Blackman, and flat-top, are generated using (4). One popular flat-top window sequence, generated using (4), is Matlab's **flattopwin**(N) routine where the h_k polynomial coefficients are [2]

$$\begin{aligned} h_0 &= 0.2156, h_1 = 0.4160, \\ h_2 &= 0.2781, h_3 = 0.0836, \\ h_4 &= 0.0069. \end{aligned} \quad (5)$$

(Very similar flat-top window generating coefficients are recommended in [3].) Thus in implementing frequency-domain convolution, to compute a single flat-top windowed $X_{\text{ft}}(m)$ spectral sample from rectangular-windowed $X(m)$ spectral samples, we would compute

$$\begin{aligned} X_{\text{ft}}(m) = {} & \frac{h_4}{2}X(m-4) - \frac{h_3}{2}X(m-3) \\ & + \frac{h_2}{2}X(m-2) - \frac{h_1}{2}X(m-1) \\ & + h_0X(m) - \frac{h_1}{2}X(m+1) \\ & + \frac{h_2}{2}X(m+2) - \frac{h_3}{2}X(m+3) \\ & + \frac{h_4}{2}X(m+4), \end{aligned} \quad (6)$$

where $X(m)$ is the rectangular-windowed FFT sample having the largest magnitude, and m is the FFT's frequency-domain sample index.

If we apply (6) to rectangular-windowed $X(m)$ FFT samples and compute the flat-top windowed maximum FFT spectral peak magnitude $M_{\text{peak}} = |X_{\text{ft}}(m)|$, the estimated value of A from (3) will have an error of no more than 0.0166 dB. Such a small error is represented by the very flat, nearly ideal, solid curve labeled as flat-top in Figure 2.

That appealing flat-top curve in Figure 2 is the good news associated with the frequency-domain flat-top window convolution in (6). The bad news is that each computation of an $X_{\text{ft}}(m)$ sample requires, assuming we combine terms having identical coefficients, 18 real multiplies and 16 real additions. In what follows, we show how to drastically reduce the computational workload in computing an $X_{\text{ft}}(m)$ sample.

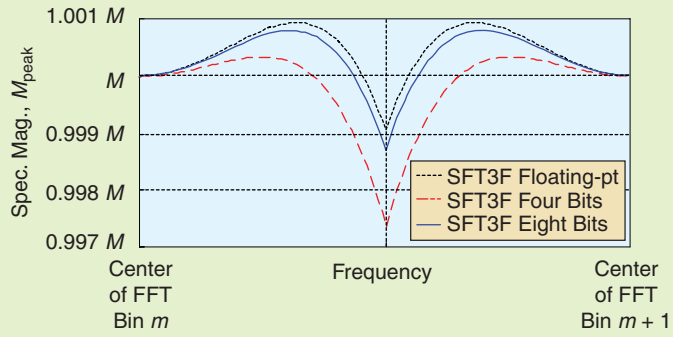
IMPROVED CONVOLUTION COEFFICIENTS

Reference [4], which discusses many different sets of window generating-polynomial coefficients, presents the following useful set of flat-top window coefficients

$$h_0 = 0.26526, h_1 = 0.5, h_2 = 0.23474 \quad (7)$$

collectively called the SFT3F coefficients. Thus to obtain a single flat-top windowed $X_{\text{ft}}(m)$ spectral sample from rectangular windowed $X(m)$ samples, based on the SFT3F coefficients in (7), we compute

$$\begin{aligned} X_{\text{ft}}(m) = {} & \frac{h_2}{2}X(m-2) - \frac{h_1}{2}X(m-1) \\ & + h_0X(m) - \frac{h_1}{2}X(m+1) \\ & + \frac{h_2}{2}X(m+2). \end{aligned} \quad (8)$$



[FIG3] Bin-to-bin frequency magnitude response of SFT3F coefficients.

The frequency-dependent scalloping loss of the floating-point SFT3F coefficients is shown as the black dotted curve in Figure 3. The variable M in Figure 3 is the M from (1). The computation of an $X_{ft}(m)$ sample using (8) results in an estimated value of A, from (3), having a scalloping error in the range of –0.0082 dB to +0.0082 dB. We call the coefficients in (7) "improved" because the computation in (8) requires only ten real multiplies and eight real additions.

Notice that flat-top window coefficients, such as those (7), have the interesting characteristic that they have both a scalloping loss and a scalloping gain versus frequency. (Compare the black dotted curve in Figure 3 to the lossy Hanning, Hamming, and rectangular curves in Figure 2 whose M_{peak} values are always less than M.)

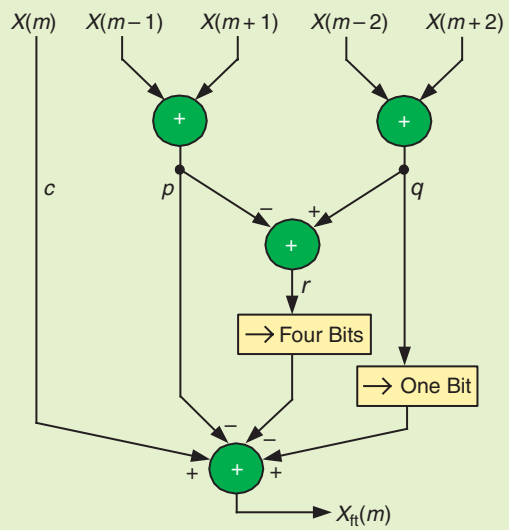
FURTHER COMPUTATIONAL IMPROVEMENTS

We can take three steps to further reduce the computational workload of computing an $X_{ft}(m)$ sample using (8).

FIRST STEP

If we divide the coefficients in (7) by the first coefficient, h_0, we obtain the new coefficients

$$h_0 = 1.0,\ h_1 = 1.88494,\ h_2 = 0.88494. \quad (9)$$



[FIG4] Multiplier-free scalloping loss compensation using 4-b coefficients in (14).

The coefficients in (9) eliminate the amplitude gain loss of the flat-top coefficients in (8) without changing their scalloping loss compensation performance. Given the flat-top window generating polynomial coefficients in (9), computing an $X_{ft}(m)$ sample proceeds as

$$\begin{aligned} X_{ft}(m) = \mathrm{X}(m) &- \frac{1.88494}{2} \\ &\times [X(m-1) + X(m+1)] \\ &+ \frac{0.88494}{2} \\ &\times [X(m-2) + X(m+2)]. \end{aligned} \quad (10)$$

The coefficients in the convolution expression in (10) are

$$\begin{aligned} g_0 &= 1.0, \\ g_1 &= -\frac{1.88494}{2} = -0.94247, \\ g_2 &= \frac{0.88494}{2} = 0.44247. \end{aligned} \quad (11)$$

SECOND STEP

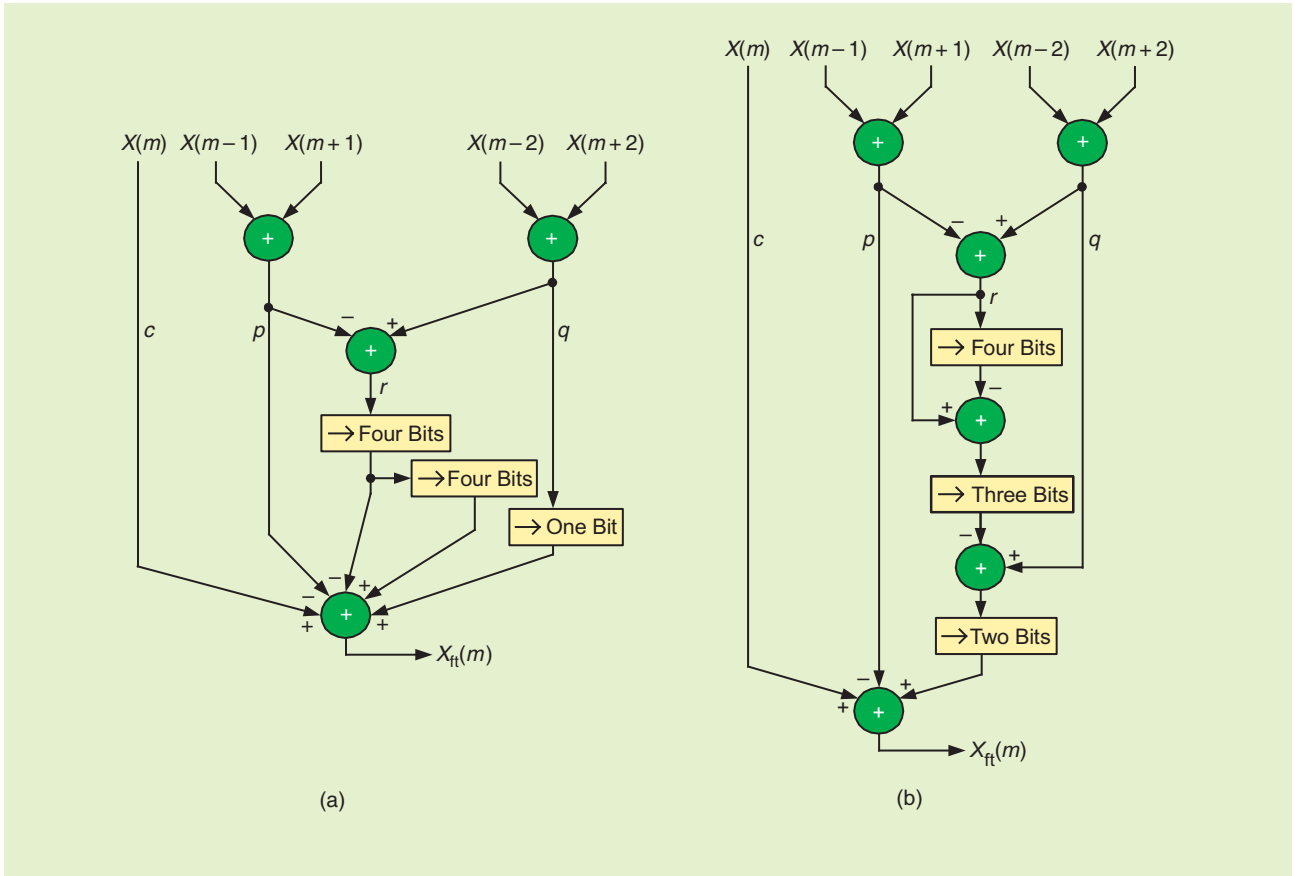
Next, we convert the coefficients in (11) to binary representation to simplify our processing by replacing the multiplications in (10) by arithmetic right-shifts. Doing so, the nonunity coefficients in (11) become

$$\begin{aligned} g_1 &= -0.94247 \\ &= -0.1111\ 0001\ 0100\ 0101\ldots \\ g_2 &= 0.44247 \\ &= 0.0111\ 0001\ 0100\ 0101\ldots. \end{aligned} \quad (12)$$

The leftmost sequence of three consecutive zeros in coefficients g_1 and g_2 suggest that we can represent those coefficients with four fractional bits without inducing too much truncation error.

To simplify our equations, let's represent our five unwindowed frequency-domain samples in (10) with

$$\begin{aligned} c &= X(m) \\ p &= X(m-1) + X(m+1) \\ q &= X(m-2) + X(m+1) \\ r &= q - p. \end{aligned}$$



[FIG5] Scalloping loss compensation using 8-b coefficients: (a) initial implementation and (b) reduced truncation error implementation.

Those assignments convert (10), using the first four fractional bits for g_1 and g_2 in (12), to

$$X_{ft}(m) = c - \frac{1}{2}p + \frac{1}{4}r + \frac{1}{8}r + \frac{1}{16}r, \quad (13)$$

allowing us to replace the multiplications in (10) with binary right-shifts. However, rather than implement the four separate binary right-shifts in (13), we can use canonical signed digit (CSD) notation to further streamline our computations. Using CSD, (13) becomes

$$X_{ft}(m) = c - p + \frac{1}{2}q - \frac{1}{16}r, \quad (14)$$

which is equivalent to, but simpler to compute than, (13). The signal flow implementation of (14) is given in Figure 4, and its performance is shown as the red dashed curve in Figure 3.

Finally, we compute M_{peak}, using (14), as

$$M_{peak} = |X_{ft}(m)| \quad (15)$$

and use that M_{peak} value in (3) to compute our desired A, the peak amplitude of the FFT's time-domain sinewave input. The computation of an M_{peak} value using (14) and (15) results in an estimated value of A, from (3), having a scalloping error in the range of –0.0229 dB to +0.003 dB.

We can achieve 8-b accuracy of our binary coefficients in (12) by adding one more term to the approximation in (14) as

$$X_{ft}(m) = c - p + \frac{1}{2}q - \frac{1}{16}r + \frac{1}{256}r. \quad (16)$$

The signal flow implementation of (16) is given in Figure 5(a), and its performance is shown as the solid blue curve in Figure 3. The computation of an $X_{ft}(m)$ sample using (16) results in an estimated value of A, from (15) and (3), having a scalloping error in the range of –0.0113 dB to +0.0069 dB. That's almost worth writing home about because the performance of the multiplier-free (16) is superior to the multiply-intensive computation in (6).

THIRD STEP

In our relentless pursuit of accuracy, we employ one last binary arithmetic trick to reduce right-shift truncation error. Notice in Figure 5(a) that one of our complex data samples experiences a right-shift by 8 b. To reduce the truncation error of an 8-b right shift, we use Horner's rule to convert (16) to

$$X_{ft}(m) = c - p + \frac{1}{2}\left(q - \frac{1}{8}\left(r - \frac{1}{16}r\right)\right). \quad (17)$$

This way, no data sample experiences a truncation error greater than a 4-b right-shift. The signal flow implementation of (17) is given in Figure 5(b) and its performance is equal to that of (16).

To consolidate what we've covered so far, Table 1 shows the computational

[TABLE 1] COMPUTATIONAL WORKLOAD PER $X_{ft}(m)$ SAMPLE AND PERFORMANCE.

SINGLE COMPUTATION EQUATION	REAL MULTS	REAL ADDS	BINARY RIGHT-SHIFTS	MAX. SCALLOPING ERROR (DB)
(6)	18	16	—	0.0166
(8)	10	8	—	0.0082
(10)	4	6	—	0.0082
(14)	—	12	4	0.0228
(16) AND (17)	—	14	6	0.0113

workload, and error performance in estimating a sinewave amplitude A, of the various scalloping loss compensation methods.

IMPLEMENTATION CONSIDERATIONS

There are two issues to keep in mind when using the above scalloping loss compensation methods.

- The flat-top window frequency-domain convolutions are most useful in accurately measuring the time-domain amplitude of a sinusoidal signal when that signal's spectral component is not contaminated by side lobe leakage from a nearby spectral component. For example, if a positive-frequency spectral component is low in frequency, i.e., located in the first few FFT bins, leakage from the spectral component's corresponding negative-frequency spectral component will contaminate that positive-frequency spectral component. As such, empirical testing indicates that the convolutions in Figures 4 and 5 should not be used for frequencies below the sixth FFT bin or above the $(N/2–5)$th FFT bin.
- The flat-top window frequency-domain convolutions discussed above are most useful when the FFT spectral component being measured is well above the background spectral noise floor.

CONCLUSION

We discussed the inherent scalloping loss uncertainty (potential error) of estimating sinewave peak amplitudes based on FFT spectral data. Then we briefly discussed the performance, and computational workload, of frequency-domain convolution using traditional five-term flat-top window coefficients to drastically reduce sinusoidal peak amplitude measurement uncertainty. Next we demonstrated a little-known three-term flat-top window polynomial that has very good scalloping loss compensation and a reduced computational workload. Finally, we presented a series of binary arithmetic tricks yielding a high-performance, efficient, multiplier-free implementation of scalloping loss compensation. Matlab and C-code implementations of this material are available at: http://www.signalprocessingsociety.org/publications/periodicals/spm/columns-resources/#tips.

ACKNOWLEDGMENT

Special thanks go to Clay Turner, holder of a fifth-degree black belt in the engineering of algebra, for his constructive suggestions that improved the content of this article.

AUTHOR

Richard Lyons (R.Lyons@ieee.org) is a consulting systems engineer and lecturer with Besser Associates in Mountain View, California. He is the author of *Understanding Digital Signal Processing 3/E* (Prentice-Hall, 2010), and editor of, and contributor to, the book *Streamlining Digital Signal Processing, A Tricks of the Trade Guidebook* (IEEE Press/Wiley, 2007).

REFERENCES

[1] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[2] Signal processing toolbox, flat top weighted window. The Mathworks Inc. [Online]. Available: http://www.mathworks.com/access/helpdesk/help/toolbox/signal/flattopwin.html

[3] S. Gade and H. Herlufsen. (1987). Use of weighting functions in DFT/FFT analysis (Part I). *Brüel Kjær Tech. Rev.* [Online]. (3), pp. 19–21. Available: http://www.bksv.com/doc/bv0031.pdf

[4] G. Heinzel, A. Rüdiger, and R. Schilling. (2002, Feb. 15). Spectrum and spectral density estimation by the discrete Fourier transform (DFT), including a comprehensive list of window functions and some new flat-top windows. Max-Planck-Inst. für Gravitationphysik Rep. [Online]. Available: http://www.rssd.esa.int/SP/LISAPATHFINDER/docs/Data_Analysis/GH_FFT.pdf

[SP]

social SCIENCES

Alexander Todorov and
Nikolaas N. Oosterhof

Modeling Social Perception of Faces

In *Baboon Metaphysics*, a detailed investigation of the complexities of baboon life, primatologists Dorothy Cheney and Robert Seyfarth write, "Any way you look at it, most of the problems facing baboons can be expressed in two words: other baboons." This statement applies with even greater force to humans. Navigating the social world requires many cognitive feats, including keeping the identities of countless people straight, as well as the dynamics of their relationships. Our inferences about the social status, beliefs, desires, and intentions of other people determine whether we decide to approach or avoid them, what to say to them, and how to say it. Social complexity is one of the key factors driving brain evolution. Across primates, the size of the neocortex increases with the size of the group, and there is recent evidence that the quality of one's relationships has direct evolutionary benefits [1]. Maintaining suitable relationships requires sophisticated social cognition. At the basis of social cognition are the abilities to represent conspecifics as unique individuals and to perceive their intentions. In light of this, it should not come as a surprise that primates have specialized brain regions for the processing of faces and (in the case of humans) information about others' mental attributes [2].

SOCIAL PERCEPTION OF FACES

The face is our primary source of visual information for identifying people and reading their emotional and mental states. With the exception of prosopagnosics (who are unable to recognize faces) and those suffering from such disorders of social cognition as autism, people are extremely adept at these two tasks. However, our cognitive powers in this regard come at the price of reading too much into the human face. The face is often treated as a window into a person's true nature. References to this belief can be found in all ancient cultures, and the belief has persisted into modern times. The Swiss pastor Johann Kaspar Lavater, who pioneered the pseudoscience of physiognomy, described in detail how to read the true, inner nature of a person from facial features (e.g., "The nearer the eyebrows are to the eyes, the more earnest, deep, and firm the character" [3, p. 59]). Although attempts to characterize personality based on external appearance have largely fallen out of favor in science, the ideas continue to appeal at an intuitive, implicit level. Lavater was probably wrong about most of his specific claims, but research strongly supports his contention that: "Whether they are or are not sensible of it, all men are daily influenced by physiognomy." [3, p. 9]. First, people tend to agree in their social judgments based on faces, indicating that faces provide information that is consistently interpreted [4], [5]. Second, such judgments are made rapidly, without much mental effort: as little as 33 ms exposure to a face is sufficient for people to decide whether a face looks trustworthy or not [6]. Third, recent functional magnetic resonance imaging (fMRI) studies have shown that regions in the brain critical for emotion and decision making are activated when participants look at negatively perceived faces (untrustworthy and aggressive looking) even when the participants have not been asked to evaluate these faces [7]. Thus it appears that our brains automatically categorize faces. Finally, many studies have shown that social judgments based on faces predict important social outcomes, ranging from sentencing decisions to electoral success [8].

THE FACE IS OUR PRIMARY SOURCE OF VISUAL INFORMATION FOR IDENTIFYING PEOPLE AND READING THEIR EMOTIONAL AND MENTAL STATES.

STATISTICAL MODELS FOR FACE REPRESENTATION

Given the agreement in social perception of faces (see Table 1), it should be possible to model this perception. What differences in facial structure lead to appearance-based social inferences? For example, based on what perceptual information do people decide that a face looks trustworthy or untrustworthy? Human faces share the same spatial layout and differences between faces are subtle, making it difficult to characterize what differences trigger specific social inferences. In this respect, data-driven approaches that do not impose any a priori constraints on face perception can be particularly useful for modeling social perception. There are two basic tasks in these approaches: creating a statistical model of face representation and using this model to derive the changes in facial features that lead to corresponding changes in social judgments. There are several statistical approaches for characterizing the commonalities and differences among individual faces. They all

Digital Object Identifier 10.1109/MSP.2010.940006
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

attempt to reduce high-dimensional face representations [e.g., pixel values of photographs, or three-dimensional (3-D) points that define the skin surface] to a lower-dimensional "face space." The dimensions of the face space define abstract, global properties of faces that are not reducible to single features. Here we use the face space model implemented in Facegen (www.facegen.com), a derivative of Blanz and Vetter's work [9]. This model uses 50 dimensions to represent face shape and 50 dimensions to represent face reflectance (brightness, color, and texture variations on the surface map of the face). The face model in Facegen is based on a database of $N = 271$ faces laser-scanned in 3-D and subsequently aligned so that all faces share the same skin surface mesh topology (for details, see [9]). The ith face is represented by a shape vector

$$\vec{s}_i = [x_1, y_1, z_1, \ldots, x_{N_s}, y_{N_s}, z_{N_s}]^T$$

with coordinates for N_s vertices, and a reflectance vector

$$\vec{t}_i = [r_1, g_1, b_1, \ldots, r_{N_t}, g_{N_t}, b_{N_t}]^T$$

with red, green, and blue color values of the N_t pixels in the color bitmap that is projected on the skin surface (in Facegen, $N_s = 2{,}043$ and $N_t = 256 \times 256$).

The face vectors are submitted to a principal component analysis (PCA), a data-driven dimensionality reduction technique that allows for characterizing the most common variations in face shape and face reflectance. In this approach, shape variations are represented by an average face $\vec{s} = 1/N \cdot \sum_m \vec{s}_m$ and a set of $k = 50$ orthogonal principal components (shape eigenfaces) $\vec{v}_1, \ldots, \vec{v}_k$ that have the greatest eigenvalues of the covariance matrix of the face coordinates. The shape of a face can then be approximated by a k dimensional weight vector $\vec{p}_i$, yielding shape coordinates

$$\vec{s}' = \vec{s} + \sum_j \vec{v}_j \cdot \vec{p}_{ij} = \vec{s} + \vec{V} \cdot \vec{p}_i,$$

where $\vec{s}$ is the average shape and $V = [v_1 \cdots v_k]$ the matrix with principal components. Variations in reflectance are treated similarly, also with 50 components.

WITH THE AID OF A STATISTICAL FACE MODEL, IT IS RELATIVELY STRAIGHTFORWARD TO UNCOVER THE VARIATIONS IN THE STRUCTURE OF FACES THAT LEAD TO SPECIFIC SOCIAL JUDGMENTS.

Thus, faces are represented as an average face plus a weighted sum of the principal components (eigenfaces). This gives rise to the concept of face space, which is the space containing the faces that can be represented.

Assuming that shape and reflectance are approximately multinormally distributed, new faces that are plausible in the population can be generated in face space by constructing new weight vectors with random Gaussian values. A practical implication is that a virtually unlimited amount of faces can be generated using this approach, which makes it an attractive alternative to using a database of face photographs. Furthermore, as described in detail below, face properties related to shape and reflectance (the surface map of the face) can be independently manipulated. These qualities of the models allow for the constructions of vectors in face space that approximate specific social judgments and for tests of psychological hypotheses.

[TABLE 1] INTERRATER AGREEMENT AND RELIABILITY OF NINE SOCIAL JUDGMENTS OF EMOTIONALLY NEUTRAL FACES.

JUDGMENT	NUMBER OF RATERS (n)	INTERRATER AGREEMENT (r)	RELIABILITY (α)
DOMINANT	23	.36	.92
THREATENING	21	.26	.87
ATTRACTIVE	35	.23	.91
FRIGHTENING	28	.17	.84
MEAN	27	.17	.83
TRUSTWORTHY	29	.15	.81
EXTROVERTED	33	.14	.84
COMPETENT	44	.11	.84
LIKEABLE	31	.10	.76

RATERS (n) WERE ASKED TO MAKE JUDGMENTS OF 300 RANDOMLY GENERATED FACES ON A SCALE FROM 1 (NOT AT ALL [TRAIT TERM]) TO 9 (EXTREMELY [TRAIT TERM]).

MODELING SOCIAL JUDGMENTS OF FACES

With the aid of a statistical face model, it is relatively straightforward to uncover the variations in the structure of faces that lead to specific social judgments [4], [10], [11]. Here, we describe models of nine different social judgments. The first task is to collect judgments of faces randomly generated by the statistical model and to show that these judgments are reliable. If the judgments are unreliable—there is a low or no agreement among judges—it is futile to try to model these judgments. As a rule of thumb, the reliability of the judgments sets the ceiling of their predictability. The second task is to test whether the statistical model of face representation can account for a meaningful proportion of the variance of these judgments. Assuming that this is the case, the third task is to construct new dimensions in face space that account for the maximum variability in the judgments. These dimensions then can be used to visualize the differences in facial structure that lead to specific judgments (Figures 1–3) and to manipulate faces along these dimensions [10], [11]. Table 1 lists nine different social judgments of 300 faces randomly generated by the statistical model described in the previous section. The most common measure of reliability used in psychological testing is Cronbach's alpha (α). This measure indicates the expected correlation between the ratings of the faces averaged across raters and the ratings of a new sample with the same number of raters. All

judgments show sufficiently high reliability, ranging from .76 to .92. Because Cronbach's α is a function of the sample size of raters and the interrater agreement, it could be a misleading measure of the actual rater agreement (e.g., a large sample of raters with a low agreement will result in reliable judgments). As shown in the third column of Table 1, the interrater agreement varies as a function of the specific judgment. Whereas for some judgments, the agreement is relatively high (e.g., dominance), for others it is relatively low (e.g., likeability). As we show below, this agreement is an important constraint on the ability of statistical models to explain social judgments.

Table 2 lists the proportion of variance of the social judgments accounted for by the shape and reflectance components of the statistical model. Four things should be noted about these data. First, the model does a good job of explaining the variance of judgments. In all cases, the amount of explained variance is statistically significant. Second, there is a high correlation between the amount of variance accounted for by shape components and the amount of variance accounted for by reflectance components ($r = .86$). Third, the variance accounted for by the model that includes both shape and reflectance components is substantially smaller than the sum of the variances accounted for by shape components alone and reflectance components alone. This finding suggests that there is redundancy in shape and reflectance information. For example, a face with a dominant shape is likely to have dominant reflectance. Finally, there is a strong relationship between the inter-rater agreement in judgments (Table 1) and the amount of variance accounted for by shape and reflectance components ($r = .61$ and $r = .82$, respectively). That is, the statistical model better explains judgments for which there is a high interrater agreement. Although this is not surprising, it indicates a sensible behavior of the model.

Before we describe the construction of new social dimensions in face space, we note that introducing nonlinear, quadratic predictors in the statistical models can improve the predictability of social judgments. The quadratic models capture the intuition that extreme faces can be evaluated negatively. In fact, for seven out of nine judgments, the quadratic shape model accounted for significantly more variance than the linear model (Table 3). In contrast to shape, the quadratic reflectance model accounted for significantly more variance only for two judgments (Table 4). This finding is consistent with prior findings on attractiveness showing that averageness is more important for shape than reflectance information [12].

COMPUTING SOCIAL VECTORS IN FACE SPACE

Having shown that the statistical model of face representation accounts for meaningful variance of social judgments, we now describe the construction

[TABLE 2] PROPORTION OF VARIANCE OF SOCIAL JUDGMENTS OF FACES ACCOUNTED FOR BY SHAPE COMPONENTS, REFLECTANCE COMPONENTS, AND SHAPE AND REFLECTANCE COMPONENTS OF STATISTICAL MODEL OF FACE REPRESENTATION.

JUDGMENT	SHAPE	REFLECTANCE	SHAPE AND REFLECTANCE
DOMINANT	.751	.810	.906
THREATENING	.729	.691	.846
ATTRACTIVE	.393	.395	.603
FRIGHTENING	.498	.523	.730
MEAN	.696	.562	.811
TRUSTWORTHY	.486	.381	.640
EXTROVERTED	.692	.524	.800
COMPETENT	.355	.437	.623
LIKEABLE	.358	.329	.559

[TABLE 3] PROPORTION OF VARIANCE OF SOCIAL JUDGMENTS OF FACES ACCOUNTED FOR BY LINEAR AND QUADRATIC SHAPE COMPONENTS OF STATISTICAL MODEL OF FACE REPRESENTATION.

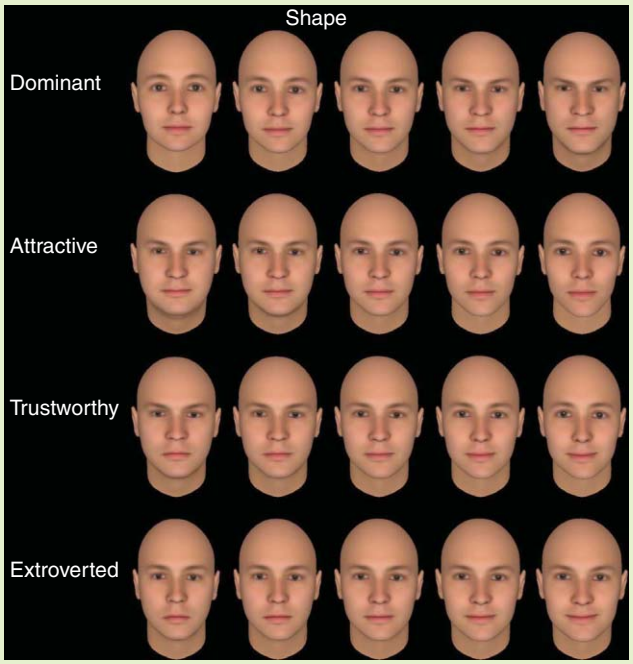
JUDGMENT	NONLINEAR MODEL	CHANGE IN ACCOUNTED VARIANCE	SIGNIFICANCE OF CHANGE
DOMINANT	.824	.073	$p < .008$
THREATENING	.784	.055	$p = .46$
ATTRACTIVE	.632	.239	$p < .0001$
FRIGHTENING	.654	.156	$p < .003$
MEAN	.758	.062	$p = .45$
TRUSTWORTHY	.674	.188	$p < .0001$
EXTROVERTED	.802	.110	$p < .0001$
COMPETENT	.612	.257	$p < .0001$
LIKEABLE	.578	.220	$p < .0002$

THE CHANGE IN ACCOUNTED VARIANCE SHOWS THE DIFFERENCE BETWEEN THE VARIANCE ACCOUNTED FOR BY THE QUADRATIC MODEL AND THE VARIANCE ACCOUNTED FOR BY THE LINEAR MODEL.

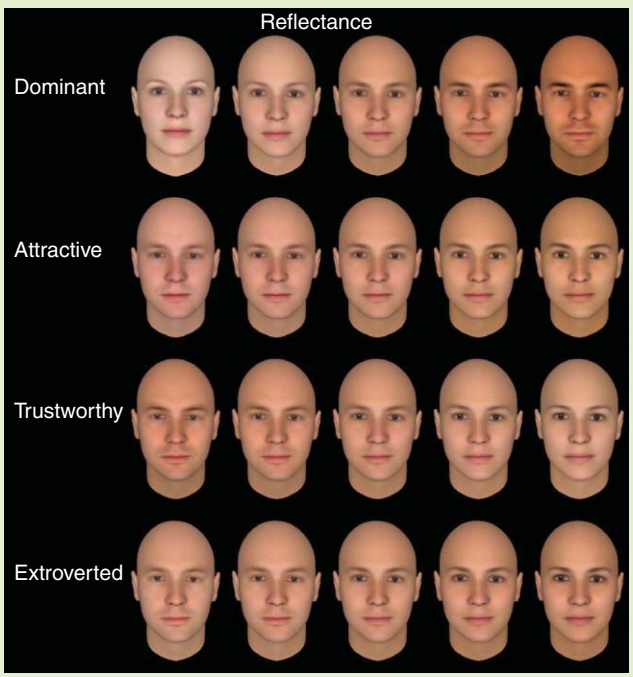
[TABLE 4] PROPORTION OF VARIANCE OF SOCIAL JUDGMENTS OF FACES ACCOUNTED FOR BY LINEAR AND QUADRATIC REFLECTANCE COMPONENTS OF STATISTICAL MODEL OF FACE REPRESENTATION.

JUDGMENT	NONLINEAR MODEL	CHANGE IN ACCOUNTED VARIANCE	SIGNIFICANCE OF CHANGE
DOMINANT	.855	.045	$P = .16$
THREATENING	.739	.048	$P = .90$
ATTRACTIVE	.530	.135	$P = .26$
FRIGHTENING	.627	.104	$P = .30$
MEAN	.646	.084	$P = .58$
TRUSTWORTHY	.502	.121	$P = .54$
EXTROVERTED	.638	.114	$P = .14$
COMPETENT	.597	.160	$P < .015$
LIKEABLE	.524	.195	$P < .010$

THE CHANGE IN ACCOUNTED VARIANCE SHOWS THE DIFFERENCE BETWEEN THE VARIANCE ACCOUNTED FOR BY THE QUADRATIC MODEL AND THE VARIANCE ACCOUNTED FOR BY THE LINEAR MODEL.



[FIG1] Variations of face shape on four social dimensions derived from judgments of dominance, attractiveness, trustworthiness, and extroversion. The perceived value of the faces on the respective dimensions increases from left to right.



[FIG2] Variations of face reflectance on four social dimensions derived from judgments of dominance, attractiveness, trustworthiness, and extroversion. The perceived value of the faces on the respective dimensions increases from left to right.

of new dimensions in face space that account for the maximum variability in the judgments.

Consider a set of randomly generated faces that have been judged on some characteristic [for example, trustworthiness rated on a scale from 1 (untrustworthy) to 9 (trustworthy) averaged over a group of participants]. A normalized linear face control $\vec{\Delta}'$ to manipulate this characteristic is constructed by

$$\vec{\Delta} = P \cdot \vec{r}, \quad \vec{\Delta}' = \vec{\Delta}/\|\vec{\Delta}\|,$$

where P_{ij} is the weight of component j for face i and $\vec{r}$ the ratings vector where the mean has been subtracted.

Intuitively, $\vec{\Delta}'$ can be considered as a normalized vector of correlations between the weights of each face component and the ratings. To justify this approach, consider that the face dimensions are, by construction, independent, and thus the obtained value for $\vec{\Delta}'$ is optimal in the least square sense.

Using the face control $\vec{\Delta}'$, an individual face with component weights $\vec{p}$ can be manipulated by α units by

$$\vec{p}\,' = \vec{p} + \alpha \cdot \vec{\Delta}'.$$

With the average shape $\vec{s}$ and principal component matrix V described earlier, this changes the coordinates of the shape vertex components from

$$\vec{s} = \vec{s} + \alpha \cdot \vec{\Delta}'$$

to

$$\begin{aligned} \vec{s}\,' &= \vec{s} + V \cdot \vec{p}\,' \\ &= \vec{s} + V \cdot (\vec{p} + \alpha \cdot \vec{\Delta}') \\ &= \vec{s} + \alpha \cdot V \cdot \vec{\Delta}', \end{aligned}$$

i.e., coordinates change linearly with changes in α. Reflectance is manipulated similarly. Face controls can be constructed for any face characteristic as long as a rating is associated with each face. Examples of face controls include hooked versus flat nose, masculine versus feminine [9], and the traits presented in this article [10], [11]. These methods uncover structural differences in appearance that predict differences in

social perception. Figure 1 shows shape variations on four dimensions derived from judgments of dominance, attractiveness, trustworthiness, and extroversion, respectively. For each dimension, five versions of a face are shown, manipulated to decrease or increase its value on the respective dimension. For example, as the dominance of the face increases, the face becomes more masculine and mature. As the attractiveness increases, the face becomes thinner with higher cheekbones. As the trustworthiness increases, the face appears to express more positive emotions. As extroversion increases, the face becomes wider and happier. Figure 2 shows reflectance variations on the four dimensions. For example, as the dominance increases, the face becomes darker and more masculine. Similar darkness changes are also detectable for the other social dimensions. Figure 3 shows both shape and reflectance variations on the dimensions.

IMPLICATIONS OF FINDINGS

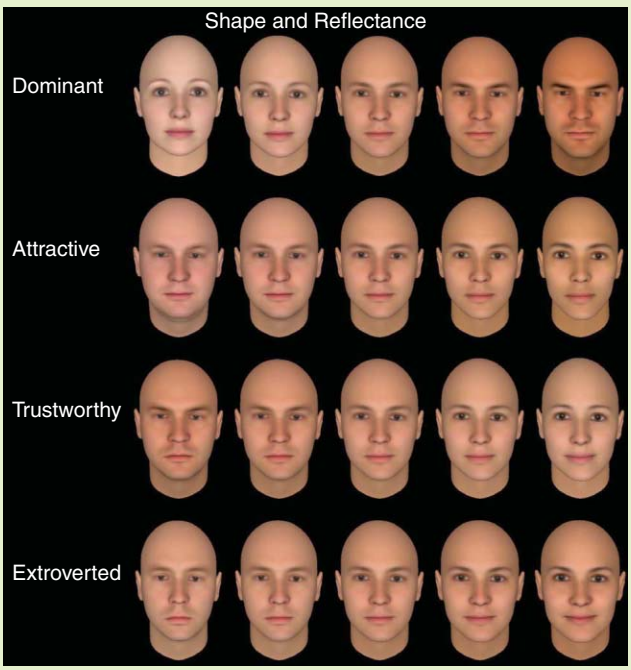
These models of social dimensions can be used to reveal the facial cues that lead to specific social judgments. For example, exaggerating the features that contribute to judgments of emotionally neutral faces reveals the underlying variations that account for these judgments. In the case of trustworthiness, although faces are perceived as emotionally neutral within the range shown in Figure 1, they are perceived as emotionally expressive outside this range [10]. Whereas faces at the extreme negative end of the dimension appear to express anger, faces at the extreme positive end appear to express happiness. In terms of social perception, these models provide clues about the basis of social inferences. Social inferences from facial appearance are based on resemblance to features that have adaptive significance—that is, to successfully navigate the social world, we need to be able to infer the emotional states, gender, and age of others [4], [5]. For example, facial expressions of emotion indicate a person's mental state and provide signals for appropriate behaviors. As a result, people with faces resembling specific emotional expressions, anger, for example, can be mistakenly judged as aggressive. When more sophisticated computer graphics and experimental methods are developed, we will have models that can be used not only to better understand social perception but also to manipulate images and create increasingly complex and lifelike avatars—knowledge that could be used for good or bad purposes. Such models can be used to manipulate images (not only of avatars but also of real people [11]) to induce specific perceptions that could influence potential important decisions ranging from consumer to voting behavior.

THESE MODELS OF SOCIAL DIMENSIONS CAN BE USED TO REVEAL THE FACIAL CUES THAT LEAD TO SPECIFIC SOCIAL JUDGMENTS.

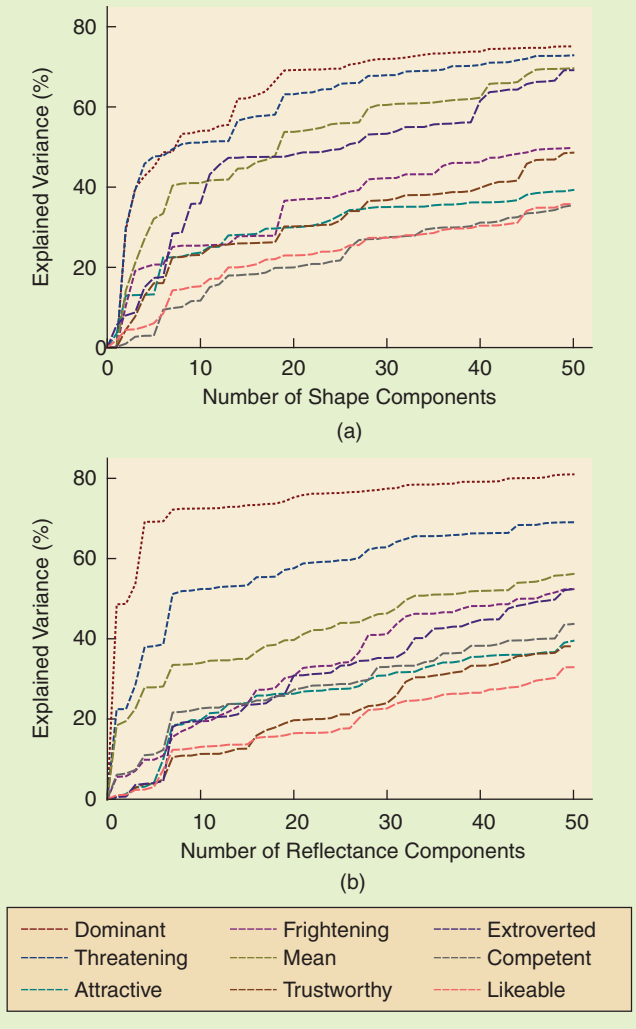
CHALLENGES AND FUTURE DIRECTIONS

One potential issue with the methods for modeling social perception of faces is overfitting. For example, here we used judgments of 300 faces to fit 50 shape and 50 reflectance parameters. Such models can perform well on the modeled set of faces but may fail to generalize to novel faces. In principle, larger training data sets should alleviate such problems.

Another potential approach is to use fewer parameters or face components. As described above, the face components were derived from a PCA, and, hence, each additional component accounts for less and less variance of facial appearance. This suggests that the first few components could capture most of the variance of social judgments. As shown in Figure 4, this is clearly the case. For example, for the shape model, the first ten components



[FIG3] **Variations of face shape and face reflectance on four social dimensions derived from judgments of dominance, attractiveness, trustworthiness, and extroversion. The perceived value of the faces on the respective dimensions increases from left to right.**



[FIG4] Explained variance of nine social judgments of faces as a function of the number of (a) shape and (b) reflectance components in the regression model.

account for more than half of the explained variance of the judgments. For the reflectance model, for many of the judgments, the first five components account for more than half of the explained variance.

Finally, the current models seem to perform rather well. First, judgments of faces manipulated by the models of these judgments agree with the models [10]. Second, the models predict judgments of novel faces. For an unrelated study, we generated another set of 300 faces that varied randomly on shape and were judged on trustworthiness and dominance. The correlations between the predicted trustworthiness and dominance scores and the judgments of these novel faces were .51 and .39 for trustworthiness and dominance, respectively, using a linear shape model, and .67 and .55, respectively, using a quadratic shape model. In principle, the models of social perception could be further improved. In addition to using larger face data sets and relying on the most informative face components, approaches that reduce the dimensionality of social judgments [10] and nonlinear approaches could be particularly fruitful.

ACKNOWLEDGMENTS

This work was supported by the U.S. National Science Foundation (0823749) and the Russell Sage Foundation. We would like to thank Jörn Diedrichsen for helpful comments on an earlier draft and Whitney Shapiro and Sara Verosky for their help.

AUTHORS

Alexander Todorov (atodorov@princeton.edu) is with the Department of Psychology, Princeton University, New Jersey.

Nikolaas N. Oosterhof (n.oosterhof@bangor.ac.uk) is with the Department of Psychology, Bangor University, United Kingdom.

REFERENCES

[1] J. B. Silk, "Social components of fitness in primate groups," *Science,* vol. 317, no. 5843, pp. 1347–1351, 2007.

[2] R. Adolphs, "The social brain: Neural basis of social knowledge," *Annu. Rev. Psychol.,* vol. 60, pp. 693–716, 2009.

[3] J. C. Lavater, *Essays on Physiognomy; for the Promotion of the Knowledge and the Love of Mankind*. London: Gale Group, 1772/1880. Eighteenth Century Collections Online. Abridged from Mr. Holcrofts translation.

[4] A. Todorov, C. P. Said, A. D. Engell, and N. N. Oosterhof, "Understanding evaluation of faces on social dimensions," *Trends Cogn. Sci.*, vol. 12, no. 12, pp. 455–460, 2008.

[5] L. A. Zebrowitz and J. M. Montepare, "Social psychological face perception: Why appearance matters," *Social Personal. Psychol. Compass*, vol. 2, no. 3, pp. 1497–1517, 2008.

[6] A. Todorov, M. Pakrashi, and N. N. Oosterhof, "Evaluating faces on trustworthiness after minimal time exposure," *Social Cogn.*, vol. 27, no. 6, pp. 813–833, 2009.

[7] A. Todorov and A. Engell, "The role of the amygdala in implicit evaluation of emotionally neutral faces," *Social Cogn. Affect. Neurosci.*, vol. 3, no. 4, pp. 303–312, 2008.

[8] C. Y. Olivola and A. Todorov, "Elected in 100 milliseconds: Appearance-based trait inferences and voting," *J. Nonverb. Behav.*, vol. 34, no. 2, pp. 83–110, 2010.

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Computer Graphics and Interactive Techniques*, 1999, pp. 187–194.

[10] N. N. Oosterhof and A. Todorov, "The functional basis of face evaluation," *Proc. Nat. Acad. Sci. USA*, vol. 105, pp. 11087–11092, 2008.

[11] M. Walker and T. Vetter, "Portraits made to measure: Manipulating social judgments about individuals with a statistical face model," *J. Vis.*, vol. 9, no. 11, pp. 1–13, 2009.

[12] A. J. O'Toole, T. Price, T. Vetter, J. C. Bartlett, and V. Blanz, "3D shape and 2D surface textures of human faces: The role of 'averages' in attractiveness and age," *Image Vis. Comput.*, vol. 18, no. 1, pp. 9–19, 1999.

[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.)

2011

MARCH

Data Compression Conference
29–31 March, Snowbird, Utah.
General Chair: James A. Storer
URL: http://www.cs.brandeis.edu/~dcc/

2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'11)
30 March–2 April, Barcelona, Spain.
General Chair: Wiro Niessen
URL: http://www.biomedicalimaging.org/

APRIL

10th International Conference on Information Processing in Sensor Networks (IPSN'11)
12–14 April, Chicago, Illinois.
General Chair: Xenofon Koutsoukos
URL: http://ipsn.acm.org/2011/

MAY

36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)
22–27 May, Prague, Czech Republic.
General Chairs: Petr Tichavský, Jan Cernocký, and Aleš Procházka
URL: http://icassp2011.com/en/welcome

JUNE

12th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAW 2011)
26–29 June, San Francisco, California.
General Chair: Hamid R. Sadjadpour
URL: http://spawc2011.org/

IEEE Statistical Signal Processing Workshop 2011 (SSP)
28–30 June, Nice, France.
General Chair: Cédric Richard
URL: http://www.ssp2011.org/index.html

JULY

2011 IEEE Conference on Multimedia Expo (ICME)
11–15 July, Chicago, Illinois.
General Chairs: Irene Cheng, Gabriel Fernandez, and Haohong Wang
URL: http://www.icme2011.org/

SEPTEMBER

2011 IEEE International Conference on Image Processing (ICIP 2011)
11–14 September, Brussels, Belgium.
General Chair: Benoit Macq
URL: http://www.icip2011.org/

2011 IEEE Thematic Meetings on Signal Processing (THEMES)
11 September, Brussels, Belgium.
URL: http://www.ieee-themes.org/index.html

OCTOBER

2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)
16–19 October, New Paltz, New York.
General Chair: Dan Ellis
URL: http://www.waspaa.com/index.html
Contact: info@waspaa.com

2011 IEEE International Conference on Multimedia Signal Processing (MMSP)
17–19 October, Hangzhou, China.
General Chairs: Wen Gao, Anthony Vetro, and Zhengyou Zhang
URL: http://www.mmsp2011.org/

DECEMBER

2011 Automatic Speech Recognition and Understanding Workshop (ASRU 2011)
11–15 December, Hawaii.
General Chairs: Michael Picheny and David Nahamoo
URL: http://www.asru2011.org/

Fourth International Workshop on Computation Advances in Multisensor Adaptive Processing (CAMSAP)
13–16 December, San Juan, Puerto Rico.
General Cochairs: Aleksandar Dogandzic and Maria Sabrina Greco
URL: http://www.conference.iet.unipi.it/camsap11/

Digital Object Identifier 10.1109/MSP.2010.939081
Date of publication: 17 February 2011

 continued from 128

ash clouds feature fragmentation and aggregation processes and cause backscattering and absorption of incident radiation, transmitted by the radar.

The measured weather radar backscattered power is proportional to the copolar horizontally polarized reflectivity factor Z_H. Microwave scattering from ash particles and from cloud water and ice droplets satisfies the Rayleigh approximation for frequencies up to X band. Under this condition, the simulated radar reflectivity factor Z_H, expressed in $\text{mm}^6 \cdot \text{m}^{-3}$, due to an ensemble of particles p is expressed as the sixth moment of particle size distribution (PSD) N_p as follows [5]:

$$Z_H = \eta_H \frac{\lambda^4}{\pi^5 |K_p|^2} = \int_0^\infty D^6 N_p(D)\, dD = m_6, \tag{1}$$

where η_H is the radar volumetric reflectivity, λ the wavelength, and K_p the dielectric factor of the particle ensemble of category p. It is noted that, keeping constant the ash particle amount, the reflectivity factor is higher for bigger particles. From (1), the variability of ash PSD modulates the radar reflectivity response.

The volcanic ash radar retrieval (VARR) methodology, devoted to quantitative remote sensing of ash cloud properties [4]–[6], includes two steps: i) ash classification and ii) ash estimation. Both steps, applied after an ash cloud detection procedure, are numerical algorithms trained by a physical-electromagnetic forward model, where the main PSD parameters are supposed to be constrained random variables. This is the reason why VARR is sometimes called a model-based supervised technique, whereas the generation of a simulated ash-reflectivity data set by letting PSD parameters vary in a random way can be framed within the so-called Monte Carlo techniques. The input information to current VARR algorithm is the measured reflectivity factor Z_{Hm} available at each radar range bin for a given elevation and azimuth angle. It is worth noting that the measured reflectivity factor Z_{Hm} differs from the simulated (intrinsic) reflectivity factor Z_H due to instrumental noise and calibration, propagation effects, and backscattering modeling errors.

For what concerns the classification step, its aim is related to the possibility to automatically discriminate between ash categories that were defined as fine, coarse, and large sizes. In the overall retrieval scheme, classification may represent a first qualitative output before performing parameter estimation. Maximum a posteriori probability (MAP) criterion can be used to carry out cloud classification in a model-based supervised context. If c is the ash class, then, by using the conditional probability density function (PDF) of a class c and given a measurement of the reflectivity factor Z_{Hm}, the MAP rule is expressed by [4]

$$\hat{c} = \text{Mode}[p(c|Z_{Hm})], \tag{2}$$

where Mode is the modal value of the posterior PDF $p(c|Z_{Hm})$. Assuming a Gaussian probability framework to describe $p(c|Z_{Hm})$ and exploiting the Bayes theorem, then (2) can be transformed into the following expression [4]:

$$\hat{c} = \text{Max}_c \left[-\frac{(Z_{Hm} - m_Z^{(c)})^2}{(\sigma_Z^{(c)})^2} - \ln(\sigma^{(c)}{}_Z)^2 + 2 \ln p(c) \right], \tag{3}$$

where Max_c is the maximum value with respect to c. Computing (3) means to know the reflectivity factor mean $m_Z^{(c)}$ (also called class centroid) and standard deviation $\sigma_Z^{(c)}$ [dBZ] of Z_{Hm} for each ash class c. The prior PDF $p(c)$ can be used to subjectively weight each class as a function of other available information. Ash class perturbations are usually assumed uncorrelated. The statistical characterization of each cloud class can be derived from a simulated synthetic data set where PSD may be either arbitrarily defined or experimentally measured [5], [6].

Within the VARR technique, ash estimation is carried out by means of a regressive approximation of the training data set, as a function of the ash size and concentration class. A way to approach the quantitative retrieval problem is to adopt a statistical parametric model to describe the relation X-Z_{Hm} where X stands for either ash concentration C_a or ash fall-rate R_a [4]–[6]. Assuming a power-law model, we can write the estimated quantity for each class c as

$$\begin{cases} \hat{C}_a^{(c)} = \alpha [Z_{Hm}]^\beta \\ \hat{R}_a^{(c)} = \gamma [Z_{Hm}]^\delta \end{cases}, \tag{4}$$

where "^" indicates estimated quantity, whereas α, β, γ, and δ are the class-dependent regression coefficients. The latter are space-time variant (because they are related to ash cloud microstructure), whereas the synthetic measured reflectivity is simulated by assuming a zero-mean random noise due to instrumental and forward modeling uncertainties. Besides ash concentration, VARR can also provide for each range bin the ash fallout rate (where the terminal ash fall velocity and air updraft are needed).

APPLICATIONS TO VOLCANIC ASH MONITORING

The potential of VARR data processing in observing volcanic ash clouds has been analyzed using some case studies where volcano eruptions happened near an available weather radar:

- the Grímsvötn volcano eruption in 2004, analyzed together with the Icelandic Met Office (IMO), using a C-band weather radar (for details, see [3] and [4])
- the Augustine volcano eruption in 2006, analyzed together with the U.S. Geological Survey Alaska Volcano Observatory, using an S-band weather radar (for details, see [6]).

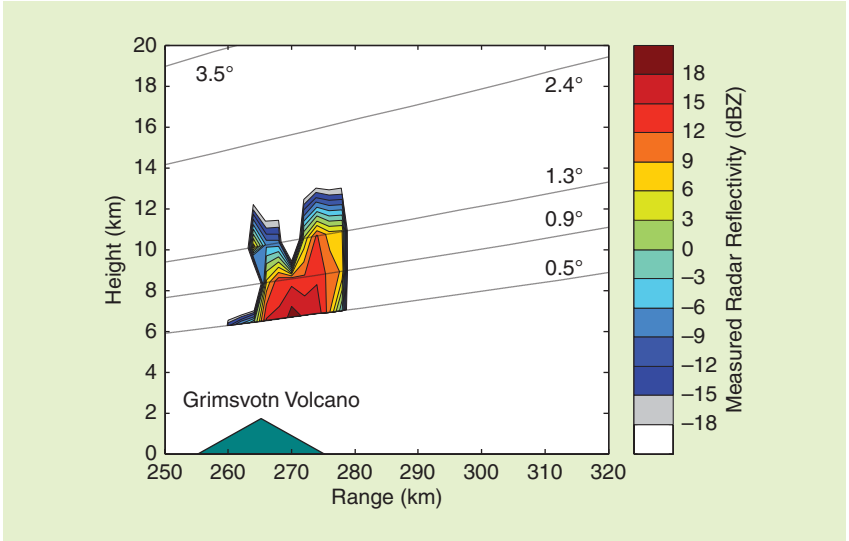
The recent explosive eruption of the Eyjafjalla Icelandic volcano started on 14 April 2010 and ended on 23 May 2010 is under evaluation, together with IMO, using an improved VARR technique.

The Icelandic case study in 2004 may be of particular interest. Grímsvötn is one of the most active volcanoes in Iceland, with a ~62 km^2 caldera covered by 150–250-m-thick ice. Its highest peak, Grímsfjall, on the southern caldera rim, reaches an elevation of

1722 m. Volcanic eruptions, numbering several per century, are water rich because of the ice cover, and they usually persist for days to weeks. The Grímsvötn eruption started in the evening of 1 November 2004 and was observed by a C-band weather radar located in Keflavik, Iceland [3], [4]. The first ash plume detected by the Keflavik radar was at 23:05 UTC (universal time coordinate) on 1 November 2004.

The eruption on the night of 2 November was followed by frequent ash plumes and the last one, detected by the weather radar, was at 08:30 UTC on 3 November. After this time, the ash plume was too low to be detected by the radar (reaching 6 km height or less). Radar volume scans were continuously acquired and data have been made available from 23:00 on 1 November 2004 till 06:00 UTC on 2 November 2004 every half an hour. Reflectivity data were radially averaged to 2 km to increase the measurement sensitivity (equal to about –5 dBZ around 260-km range). Considering the distance of about 260 km between the Keflavik radar and the Grímsvötn volcano, volcanic ash clouds can be detected at heights higher than 6 km using the minimum elevation of 0.5°. This means that the volcanic eruption cloud cannot be detected between the Grímsvötn summit at 1,725 m and 6,000 m altitude.

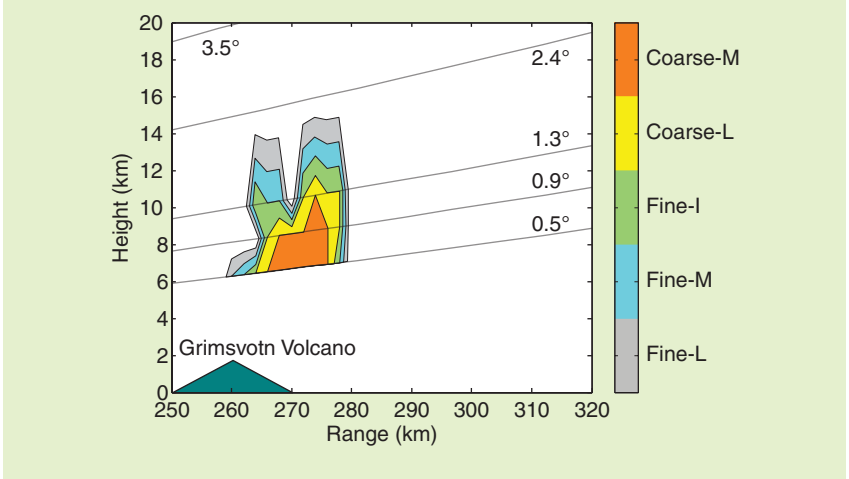
An example of C-band radar imagery can be easily pictured by plotting the so-called range-height indicator (RHI) diagram, illustrated in Figure 1. This figure stresses the fact that volcanic ash clouds can be detected from Keflavik only at heights higher than about 6 km using the minimum elevation of 0.5°. The signal of volcanic cloud is quite evident from the RHI signature with values up to 20 dBZ. If the classification algorithm is applied to radar RHI data, we can detect the ash class distribution displayed in Figure 2. The RHI maps strictly reflect the bimodal spatial structure of reflectivity measurements in Figure 1. Coarse ash particles are dominant in the lower part of volcanic plume, already moved toward northwest.



[FIG1] RHI of the measured horizontally polarized reflectivity (in dBZ) along the radar-vent cross section during the Grímsvötn volcano eruption on 2 November 2004 at 0300 UTC. The measured sector is visualized as a function of distance between the Keflavik radar (64°01′ N, 22°38′ W) and Grímsvötn volcano (64°42′ N, 17°33′ W, schematically indicated by a filled triangle) with elevation angles between 0.5° and 3.5°.

CONCLUSIONS

The possibility of monitoring 24 hours a day, in all weather conditions, at a fairly high spatial resolution and every few minutes after the eruption is the major advantage to using ground-based microwave radar systems. The latter can be crucial systems to monitor the volcanic eruption from its eruption early-stage near the volcano vent, dominated by coarse ash and blocks, to ash-dispersion stage up to few hundreds of kilometers, dominated by transport and evolution of coarse and fine ash particles. Of course, the sensitivity of the ground-based radar measurements will decrease as the ash cloud will be farther so that for distances greater than about 50 km fine ash might become "invisible" to the radar; but, in this respect, radar observations can be complementary to satellite, LIDAR, and aircraft observations. Moreover,



[FIG2] The same as in Figure 1, but for estimated ash class, named as fine-L (fine ash with light concentration), fine-M (fine ash with moderate concentration), fine-I (fine ash with intense concentration), coarse-L (coarse ash with light concentration), and coarse-M (fine ash with moderate concentration). The triangle schematically indicates the volcano vent.

radar-based products such as real-time erupted volcanic ash concentration, height, mass, and volume can be used to initialize dispersion model inputs.

Due to logistics and space-time variability of the volcanic eruptions, a suggested optimal radar system to detect ash cloud could be a portable X-band weather Doppler polarimetric radar. This radar system may satisfy technological, economical, and new scientific requirements to detect ash cloud. The sitting of the observation system, is a problematic tradeoff for a fixed radar system (as the volcano itself may cause a beam obstruction and the ash plume may move in unknown directions), can be easily solved by resorting to portable systems.

Further work is needed to assess the VARR potential using experimental campaign data. Future investigations should be devoted to the analysis of the impact of ash aggregates on microwave radar reflectivity and on the validation of radar estimates of ash amount with ground measurements where available. The last task is not an easy one as the ash fall is dominated by wind advection and by several complicate microphysical processes. This means that what is retrieved within an ash cloud may be not representative of what was collected at ground level in a given area. Spatial integration of ground-collected and radar-retrieved ash amounts may be considered to carry out a meaningful comparison. Preliminary results for the Grímsvötn case study show that the radar-based ash mass retrievals compare well with the deposited ash estimated from in situ ground sampling within the volcanic surrounding area.

AUTHOR

Frank S. Marzano (marzano@die.uniroma1.it) is an associate professor in the Department of Information Engineering, Electronics and Telecommunications (DIET), Sapienza University of Rome, Italy, and vice-director of the Center of Excellence on Remote Sensing and Severe Weather Forecast (CETEMPS), University of L'Aquila, Italy.

REFERENCES

[1] M. Trevelyan. (2010, Apr. 17). Ash cloud over Europe deepens travel chaos. *Reuters* [Online]. Available: http://www.reuters.com/article/idUSTRE63E1TM20100417

[2] A. Tupper, I. Itikarai, M. S. Richards, F. Prata, S. Carn, and D. Rosenfeld, "Facing the challenges of the international airways volcano watch: The 2004/05 eruptions of Manam, Papua New Guinea," *Weather Forecast.*, vol. 22, pp. 175–191, 2007.

[3] F. S. Marzano, S. Barbieri, E. Picciotti, and S. Karlsdóttir, "Monitoring sub-glacial volcanic eruption using C-band radar imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 1, pp. 403–414, 2010.

[4] F. S. Marzano, S. Barbieri, G. Vulpiani, and W. I. Rose, "Volcanic cloud retrieval by ground-based microwave weather radar," *IEEE Trans. Geosci. Remote Sensing*, vol. 44, no. 11, pp. 3235–3246, 2006.

[5] F. S. Marzano, G. Vulpiani, and W. I. Rose, "Microphysical characterization of microwave radar reflectivity due to volcanic ash clouds," *IEEE Trans. Geosci. Remote Sensing*, vol. 44, pp. 313–327, 2006.

[6] F. S. Marzano, S. Marchiotto, S. Barbieri, C. Textor, and D. Schneider, "Model-based weather radar remote sensing of explosive volcanic ash eruption," *IEEE Trans. Geosci. Remote Sensing*, vol. 48, pp. 3591–3607, 2010.

[SP]

[from the **GUEST EDITORS**] continued from page 15

various machine learning and signal processing problems involving NMF, sparse PCA, LARS, OMP, and SOMP: http://www.di.ens.fr/willow/SPAMS/

- Bayesian compressive sensing: http://people.ee.duke.edu/~lcarin/BCS.html
- Orthogonal matching pursuit and KSVD: http://www.cs.technion.ac.il/~ronrubin/software.html
- Low-rank matrix recovery and completion (RPCA): http://perception.csl.uiuc.edu/matrix-rank/home.html

OTHER REFERENCES AND APPLICATIONS

- Compressive Sensing Repository: http://dsp.rice.edu/cs
- Robust Face recognition and others: http://perception.csl.uiuc.edu/recognition/Home.html

ACKNOWLEDGMENTS

We would like to thank all the authors who have contributed to this special issue and all the reviewers who have provided valuable comments. We thank *IEEE Signal Processing Magazine* for supporting this special section, with special thanks to Prof. Dan Schonfeld of UIC for recommending this idea to the editors and the board. We thank IEEE staff, especially Sonal Parikh and Rebecca Wollman, for their professional support with the editorial matters during the preparation of this special issue.

SPECIAL TRIBUTE TO PROF. PARTHA NIYOGI

During the preparation of this special issue, one of the guest editors, Prof. Partha Niyogi of the University of Chicago, passed away. We all have been deeply saddened by the sudden loss of a great scholar, a colleague, and a friend. Prof. Niyogi has made some of the most fundamental contributions to the theory of manifold learning and has been well known as a world leading scientist in this new area. Prof. Niyogi agreed to serve as a guest editor of this special issue despite his medical condition at the time, which had shown his great passion and dedication to this research topic. All the editors are very much honored to have served with him as guest editors on this important special issue during his last days.

REFERENCES

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[2] E. Candèes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, 2005.

[3] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[4] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, pp. 211–218, 1936.

[5] G. Haro, G. Randall, and G. Sapiro, "Stratication learning: Detecting mixed density and dimensionality in high-dimensional point clouds," in *Proc. NIPS*, 2008.

[6] Y. Ma, H. Derksen, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 9, pp. 1546–1562, 2008.

[7] Y. Ma, A. Yang, H. Derksen, and R. Fossum, "Estimation of subspace arrangements with applications to modeling and segmenting mixed data," *SIAM Rev.*, vol. 50, no. 3, 2008.

[8] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 12, pp. 1–15, 2005.

[SP]

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*.

COMPANY	PAGE#	URL	PHONE
Asilomar Conference	9	www.asilomarssc.org	
CAMSAP 2011	5	www.conference.iet.unipi.it/camsap11	
ESC	7	www.embedded.com/sv	
Mathworks	CVR 4	www.mathworks.com/connect	+1 508 647 7040
Mini-Circuits	CVR 2, 3, CVR 3	www.minicircuits.com	+1 718 934 4500
TU Eidhoven	123	www.tue.nl/jobs	

advertising SALES OFFICES

James A. Vick
Staff 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/ SOUTHWEST 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

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/ NORTHWEST/ 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
Heleen Vodegel
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
hv.ieeemedia@ieee.org
Europe, Africa, Middle East

ASIA/FAR EAST/PACIFIC RIM
Susan Schneiderman
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org
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
John Restchack
Phone: +1 212 419 7578;
Fax: +1 212 419 7589
j.restchack@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/TEXAS/CENTRAL CANADA
Darcy Giovingo
Phone: +1 847 498 4520;
Fax: +1 847 498 5911
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
Heleen Vodegel
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
hv.ieeemedia@ieee.org
Europe, Africa, Middle East

Digital Object Identifier 10.1109/MSP.2010.940002

[in the SPOTLIGHT]

Frank S. Marzano

Remote Sensing of Volcanic Ash Cloud During Explosive Eruptions Using Ground-Based Weather Radar Data Processing

The ash ejected into the atmosphere by the Eyjafjalla Icelandic volcano during its recent eruption posed such a threat to flights over much of Europe that the ensuing cancellations resulted in an unprecedented disruption of the European commercial air transportation system [1]. Volcanic ash is not only a significant hazard to aircraft operations but also to public safety from volcanic ash fall at the surface (e.g., [2] and [3]). Given the significance of the hazards posed by volcanic ash, timely detection and tracking of the erupted ash cloud is essential to a successful warning process, particularly during and immediately following an eruptive event. In this article, we will discuss ground-based radar (radio detection and ranging) data processing for ash cloud remote sensing pointing to the physical basis of retrieval algorithms and an example of their application.

MONITORING VOLCANIC ASH CLOUDS

As pointed out by the Volcanic Ash Advisory Centers (VAACs), the largest uncertainty in the ability of numerical models to predict the spread of volcanic ash, and hence to advise aviation regulators, is in observations of the eruption itself: i) knowing how high the ash is being expelled and ii) what concentration of ash is being expelled. Current observations come from a range of sources: satellite (height and spatial distribution of the dispersed ash plume), cloud ceilometers and light detection and ranging (LIDAR) systems (ash cloud height and depth), seismic (volcano activity), and human (ash plume height and shape). Within this list, it should be added the use of ground-based meteorological microwave radars whose new role, within the volcanic ash monitoring network, is the goal of this short contribution.

Real-time and aerial monitoring of a volcano eruption, in terms of its intensity and dynamics, is not always possible by conventional visual inspections. A variety of satellite techniques have been successfully used to track volcanic ash clouds; however, these techniques have certain limitations [2]. As known, these data are subject to limitations in both spatial and temporal resolution. Issues involving the detection of ash clouds using infrared brightness temperature differencing, a commonly used method, have been addressed suggesting several scenarios where effective infrared satellite detection of volcanic ash clouds may be compromised. Ground microwave instrumentation, such as global positioning system (GPS) receivers and wind profiler radars, may play a complementary role, even though their operational utility is limited by the relatively small spatial coverage. On the other hand, ground-based LIDAR optical systems may show a higher sensitivity to ash contents with respect to microwave instruments but counterbalanced by stronger path attenuation effects.

Ground-based microwave radar systems can have a valuable role in volcanic ash cloud monitoring as evidenced by available radar imagery [3], [4]. These systems represent one of the best methods for real-time and areal monitoring of a volcano eruption, in terms of its intensity and dynamics. The possibility of monitoring 24 hours a day, in all weather conditions, at a fairly high spatial resolution (less than few hundreds of meters), and every few minutes after and during the eruption is the major advantage of using ground-based microwave radar systems. They can provide data for determining the ash volume, total mass, and height of eruption clouds.

There are still several open issues about microwave weather radar capabilities to detect and quantitatively retrieve ash cloud parameters [4], [5]. Exploitation of microwave weather radars for volcanic eruption monitoring is fairly limited due to their exclusive use for water clouds and precipitation observations. Several unknowns may also condition the accuracy of radar-derived geophysical products, most of them related to microphysical variability of ash clouds due to particle size distribution, shape, and dielectric composition. Moreover, the aggregation of volcanic ash particles within the eruption column of explosive eruptions may influence the residence time of ash in the atmosphere and the radiative properties of the ash cloud. Numerical experiments are helpful to explore processes occurring in the eruption column. Some advanced ash plume models can simulate the interactions of hydrometeors and volcanic ash and the radar response, including particle formation within a rising eruption column [6].

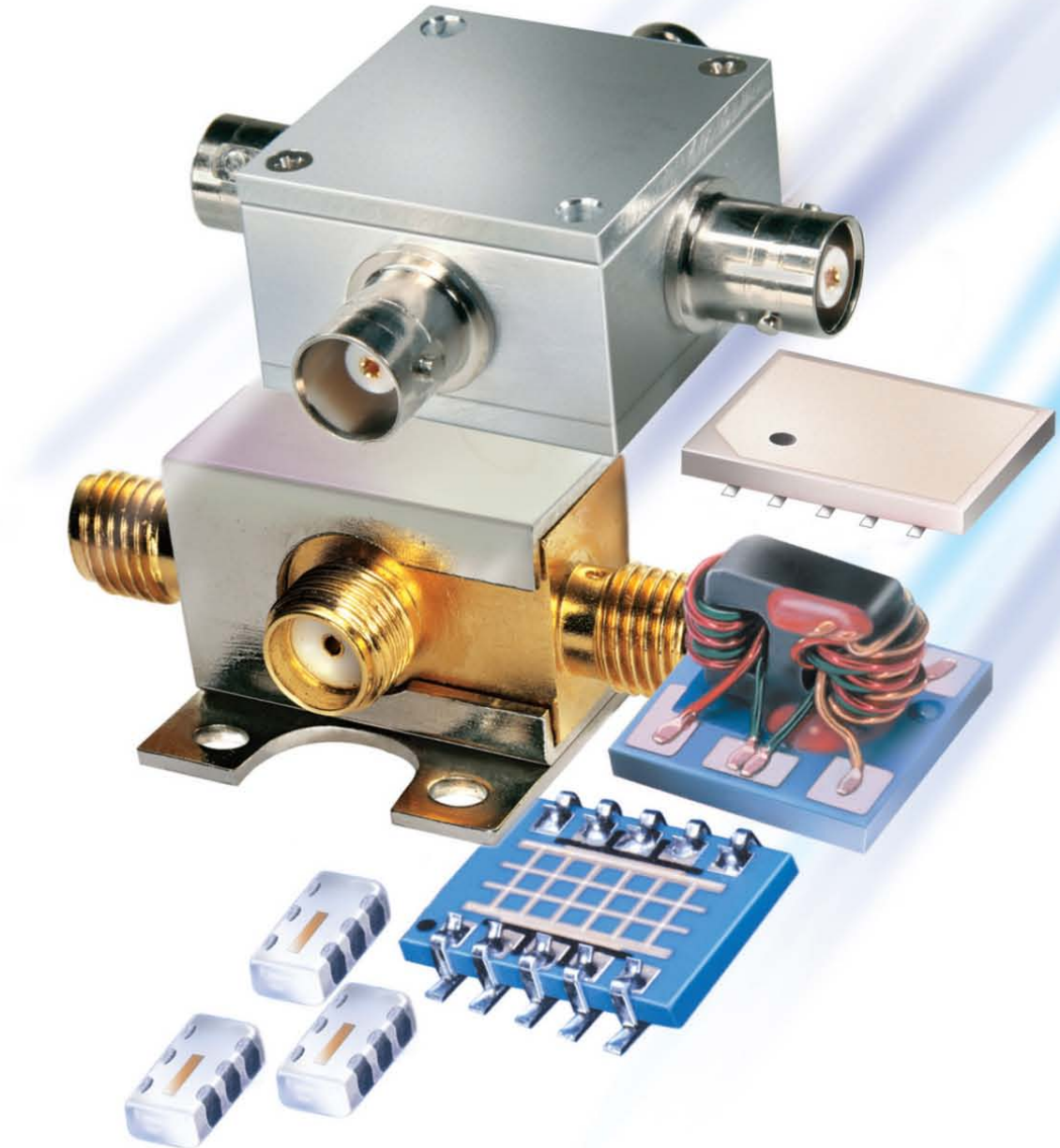
RADAR DATA PROCESSING

Weather radar systems, typically operated at S and C bands, can be used to monitor and measure volcanic eruption parameters, although they were designed to study hydrometeors and rain clouds. Both targets have the same measure principle: both rain clouds and

(continued on page 124)

Digital Object Identifier 10.1109/MSP.2010.939846
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE



Directional/Bi-Directional

LTCC COUPLER FAMILY

IN STOCK $169 From ea. Qty.1000

Mini-Circuits LTCC coupler family offers versatile, low cost solutions for your ***5 kHz to 6 GHz*** needs with rugged connectorized models from .74"x.50" to surface mount couplers from .12"x.06", the smallest in the world! Choose from our 50 & 75 Ω directional and bi-directional couplers with coupling ranging from 6-22 dB and with capability to pass DC. Mini-Circuits offers the world's most highly evolved LTCC technology delivering both minimal insertion loss and high directivity with models handling up to 65 W. All of our couplers are ESD compliant and available as RoHS compliant. For full product details and specifications for all our couplers, go to Mini-Circuits web site and select the best couplers for your commercial, industrial and military requirements.

Mini-Circuits…we're redefining what VALUE is all about!

Mini-Circuits®
ISO 9001 ISO 14001 AS9100 CERTIFIED

P.O. Box 350166, Brooklyn, New York 11235-0003 (718) 934-4500 Fax (718) 332-4661

The Design Engineers Search Engine finds the model you need, Instantly • For detailed performance specs & shopping online see minicircuits.com

U.S. patent 7739260

IF/RF MICROWAVE COMPONENTS

396 rev M

Agilent

Tektronix

LeCroy

Rohde & Schwarz

National Instruments

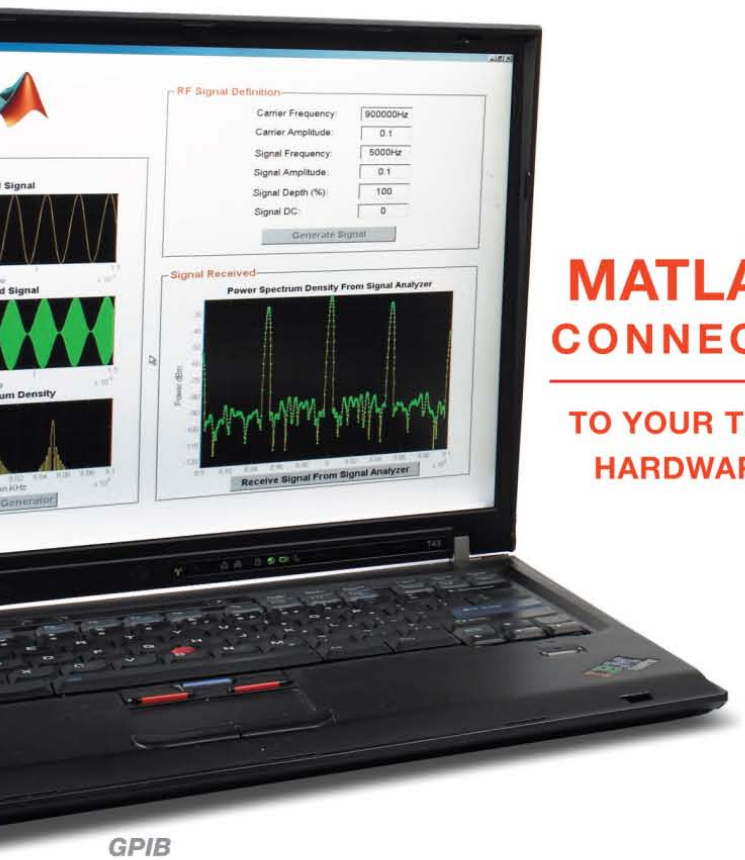
Anritsu

Keithley

Yokogawa

Tabor

Pickering



MATLAB CONNECTS

TO YOUR TEST HARDWARE

GPIB

LXI

IVI

TCP/IP

VISA

USB

UDP

RS-232

Connect to your test equipment directly from MATLAB® using standard communication protocols and hundreds of available instrument drivers.

Analyze and visualize your test results using the full numerical and graphical power of MATLAB.

For more information on supported hardware, visit www.mathworks.com/connect

MathWorks®