IEEE
Signal Processing MAGAZINE
Volume 35 | Number 6 | November 2018

EFFICIENT SIGNAL PROCESSING

Theory and Applications

Model Selection Techniques

Sub-Nyquist Radar Systems

Privacy-Aware Smart Metering

Deep Convolutional Neural Networks

Sliding Discrete Fourier Transform with Kernel Windowing

IEEE Signal Processing Society®

IEEE

IEEE Collabratec™

Bright Minds. Bright Ideas.

Introducing IEEE Collabratec™

The premier networking and collaboration site for technology professionals around the world.

IEEE Collabratec is a new, integrated online community where IEEE members, researchers, authors, and technology professionals with similar fields of interest can **network** and **collaborate**, as well as **create** and manage content.

Featuring a suite of powerful online networking and collaboration tools, IEEE Collabratec allows you to connect according to geographic location, technical interests, or career pursuits.

You can also create and share a professional identity that showcases key accomplishments and participate in groups focused around mutual interests, actively learning from and contributing to knowledgeable communities. All in one place!

Network.
Collaborate.
Create.

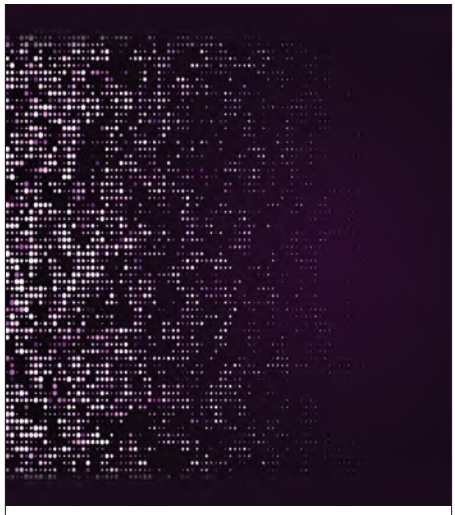
Learn about IEEE Collabratec at
ieee-collabratec.ieee.org

IEEE

Contents

Volume 35 | Number 6 | November 2018

FEATURES



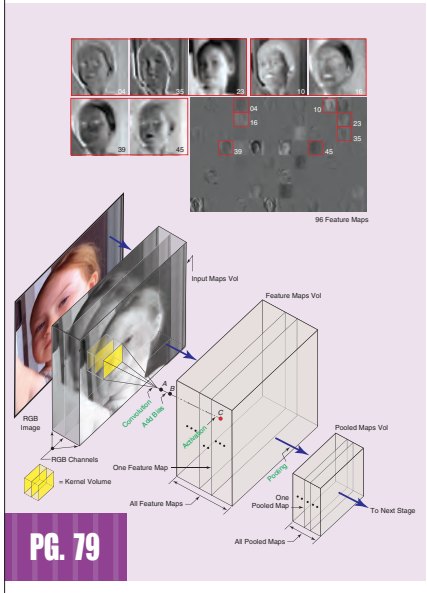
ON THE COVER

Three feature articles have been selected for this issue of the magazine. The first article examines model selection algorithms for which efficiency/sparsity is a factor; the second article presents low-complexity, efficient methods for radar; and the third article looks at intelligent methods for the smart grid.

COVER IMAGE: ©ISTOCKPHOTO.COM/SELIM DÖNMEZ

PG. 8

COLUMNS



PG. 79

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. **Individual copies:** IEEE Members US$20.00 (first copy only), nonmembers US$241.00 per copy. **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.2018.2866004

DEPARTMENTS

©ISTOCKPHOTO.COM/MARTINLISNER

The 44th IEEE International Conference on Acoustics, Speech, and Signal Processing will be held 12–17 May 2019 in Brighton, United Kingdom.

IEEE Signal Processing Magazine

EDITOR-IN-CHIEF
Robert W. Heath, Jr.—The University of Texas at Austin, U.S.A.

AREA EDITORS

Feature Articles
Matthew McKay—Hong Kong University of Science and Technology, Hong Kong SAR of China

Special Issues
Namrata Vaswani—Iowa State University, U.S.A.

Columns and Forum
Roberto Togneri—The University of Western Australia

e-Newsletter
Ervin Sejdic—University of Pittsburgh, U.S.A.

Social Media and Outreach
Tiago Henrique Falk—INRS, Canada

Special Initiatives
Andres Kwasinski—Rochester Institute of Technology, U.S.A.

EDITORIAL BOARD
Daniel Bliss—Arizona State University, USA
Danijela Cabric—University of California, Los Angeles
Volkan Cevher—École polytechnique fédérale de Lausanne, Switzerland
Mrityunjoy Chakraborty—Indian Institute of Technology, Kharagpur, India
George Chrisikos—Qualcomm, Inc., U.S.A.
Elza Erkip—New York University, U.S.A.
Alfonso Farina—Leonardo S.p.A., Italy
Clem Karl—Boston University, U.S.A.
C.-C. Jay Kuo—University of Southern California, U.S.A.
Erik Larsson—Linköping University, Sweden
David Love—Purdue University, USA
Maria G. Martini—Kingston University, U.K.
Helen Meng—City University of Hong Kong, Hong Kong SAR of China
Meinard Mueller—Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
Alejandro Ribeiro—University of Pennsylvania, U.S.A.
Douglas O'Shaughnessy—INRS Universite de Recherche, Canada
Osvaldo Simeone—Kings College London, U.K.
Milica Stojanovic—Northeastern University, USA
Ananthram Swami, Army Research Labs, U.S.A.
Jong Chul Ye—KAIST, South Korea
Qing Zhao—Cornell University, USA
Josiane Zerubia—INRIA Sophia-Antipolis Mediterranee, France

ASSOCIATE EDITORS—COLUMNS AND FORUM
Ivan Bajic—Simon Fraser University, Canada
Balázs Bank—Budapest University of Technology and Economics, Hungary
Panayiotis (Panos) Georgiou—University of Southern California, U.S.A.
Hana Godrich—Rutgers University, U.S.A.
Rodrigo Capobianco Guido—São Paulo State University, Brazil
Yuan-Hao Huang—National Tsing Hua University, Taiwan
Euee Seon Jang—Hanyang University, Republic of Korea
Vishal Patel—Rutgers University, U.S.A.
Christian Ritz—University of Wollongong, Australia
Changshui Zhang—Tsinghua University, China
H. Vicky Zhao—Tsinghua University, China

ASSOCIATE EDITORS—e-NEWSLETTER
Csaba Benedek—Hungarian Academy of Sciences, Hungary
Yuhong Liu—Penn State University at Altoona, U.S.A.
Andreas Merentitis—University of Athens, Greece
Michael Muma—TU Darmstadt, Germany
Le Yang—Harbin Institute of Technology, China
Xiaorong Zhang—San Francisco State University, U.S.A.

ASSOCIATE EDITOR—SOCIAL MEDIA/OUTREACH
Guijin Wang—Tsinghua University, China

IEEE SIGNAL PROCESSING SOCIETY
Ali H. Sayed—President
Ahmed Tewfik—President-Elect
Fernando Pereira—Vice President, Conferences
Nikos D. Sidiropoulos—Vice President, Membership
Sergio Theodoridis—Vice President, Publications
Walter Kellerman—Vice President, Technical Directions

IEEE SIGNAL PROCESSING SOCIETY STAFF
William Colacchio—Senior Manager, Publications and Education Strategy and Services
Rebecca Wollman—Publications Administrator

IEEE PERIODICALS MAGAZINES DEPARTMENT
Jessica Welsh, *Managing Editor*
Geraldine Krolin-Taylor, *Senior Managing Editor*
Janet Dudar, *Senior Art Director*
Gail A. Schnitzer, *Associate Art Director*
Theresa L. Smith, *Production Coordinator*
Mark David, *Director, Business Development - Media & Advertising*
Felicia Spagnoli, *Advertising Production Manager*
Peter M. Tuohy, *Production Director*
Kevin Lisankie, *Editorial Services Director*
Dawn M. Melley, *Staff Director, Publishing Operations*

Digital Object Identifier 10.1109/MSP.2018.2866005

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.

Robert W. Heath, Jr. | Editor-in-Chief | rheath@utexas.edu

Making Papers, Code, and Data Accessible

There have been three key revolutions in the way that research has become accessible: publishing, code, and data. The second and third revolutions are still taking place, particularly driven by the rise of machine-learning and artificial intelligence research in the last decade. When I started my research career in 1995, the World Wide Web was still in its infancy. The popular Netscape browser had just been launched. Search engines were not widely used. While many academics owned e-mail addresses, few had web pages. If they did, they were not kept current. If you wanted to look at a paper, you had to make the trek to the dusty library stacks. Or perhaps your research group maintained hard copies of journals and conference proceedings. In short, papers were available, but it was tedious to find them and obtain a copy for reading.

The first main revolution in accessibility came with the World Wide Web. Research groups began creating and maintaining research pages with publication lists. Articles subsequently appeared in electronic format, becoming available through databases like IEEE *Xplore*. Search engines began crawling such data sources so that research queries could be processed instantly, rather than enduring a long wait for the results of a careful library search. Publication times also decreased. Conference proceedings appeared online more quickly with each year. The submission-to-publication time for journals was reduced through online review management software and by having the authors take more responsibility for the formatting of their papers. Researchers also started sharing preprints of their research through their websites or through preprint servers. As a result, research has become easier to access, more widely available, and much more current.

The second big revolution in accessibility has come through sharing code. When I was a graduate student, it was part of the learning process to reproduce the results from another paper. I still believe this is an important part of research. But I do see the value of sharing code for research reproducibility. As algorithms and simulations become more complicated, having code available on the web enables other authors to more easily replicate the simulations in prior work and to justify their innovations. It encourages innovation by avoiding incremental research that does not improve upon what has already been performed. Of course, the authors who share code also benefit. Researchers cite the paper that describes the algorithm (or they should do so), which improves the usual publication metrics.

Having code available on the web enables other authors to more easily replicate the simulations in prior work and to justify their innovations.

I believe we are still in the early stages of making code available. In the past, most researchers were posting code somewhere on their webpage and including a link. More recently, researchers have been making code available online at places like GitHub to encourage more open-source development and refinement. Finally, the IEEE has also joined the game through Code Ocean. This allows for source code to be linked with the paper in IEEE *Xplore*, and it is supported by *IEEE Signal Processing Magazine* (*SPM*). In a quick survey of my own research, I realize that I have not put much source code online. But I am encouraging my students to make code available. In summary, I think the community is making good progress in this area, although code is not shared for every paper or in the same way and is hard to search.

The third main revolution in accessibility has been sharing data. When I was a graduate student working in signal processing for communications, there was not much interest in data. If you had the source code for a paper, you could reproduce the necessary data to make simulation comparisons, e.g., bit error rate or mean squared error plots. I do remember trying to do some audio experiments from a record, but this required receiving a data tape in the mail (and, in any case, I was never able to get my algorithms to work on the data). With the substantial interest

Digital Object Identifier 10.1109/MSP.2018.2869962
Date of publication: 13 November 2018

We encourage authors to make code available through Code Ocean and IEEE DataPort or through other means, as you prefer.

now in data-driven signal processing, sharing data has never been so relevant. In *SPM*'s March 2018 "President's Message" column, IEEE Signal Processing Society President Ali H. Sayed's article "Big Ideas or Big Data?" makes the case that signal processing and data science have always been intimately connected [1].

I think that we are still in the early stages of making data accessible. Many researchers make data sets available on their web pages. But such links may expire as their careers change. IEEE DataPort offers an option to make data sets available through a subscription model similar to IEEE *Xplore*. These fees help to offset the costs of long-term storage and accessibility of very large data sets. Much work on big data happens outside of IEEE publications, and it is unclear to me if there are similar offerings targeted toward researchers in other communities.

SPM is working on enhanced accessibility of our content. The magazine's articles are in IEEE *Xplore*, and some content (like the "From the Editor" column) is available for free from the IEEE Signal Processing Society's website. We encourage authors to make code available through Code Ocean and IEEE DataPort or through other means, as you prefer. If you submit to *SPM*, I hope that you will see at least the selfish value of sharing and consider making your code and data available to future researchers.

Reference

[1] A. H. Sayed, "Big ideas or big data?" *IEEE Signal Process. Mag.*, vol. 35, no. 2, pp. 5–6, 2018.

SP

Ali H. Sayed | IEEE Signal Processing Society President | ali.sayed@epfl.ch

Twinkle, Twinkle, Little Star

The title of this editorial is borrowed from a popular children's lullaby from the 1800s, which reads "Twinkle, twinkle, little star, how I wonder what you are!" It reminds me of the vast expanse of unexplored space (and science) that lie before us.

The human race has always been fascinated by space—and who would not be? Its shining stars continually challenge us to get closer and unravel their mysteries. Civilizations old and new have been defined by their relationship with space and by their contribution to astronomy.

This past August, NASA launched its first mission to explore a star. It will travel for six long years and explore the atmosphere of the sun at a safe distance of almost 4 million miles. Another Japanese spacecraft, with rovers built in cooperation with German and French space centers, will be exploring the surface of a 1-km-wide asteroid after traveling for more than three years. Earlier, in 2003 and 2011, NASA launched the rovers Spirit, Opportunity, and Curiosity to explore areas on the surface of the planet Mars. These efforts are fantastic examples of creative feats of engineering. Imagine flying robotic machines into far-away planets or asteroids in dark space, landing them on predetermined spots, and controlling them remotely. Significant ingenuity drives these accomplishments.

Digital Object Identifier 10.1109/MSP.2018.2869963
Date of publication: 13 November 2018

Our scientific community should be proud of these achievements. It is not a secret that signal processing theory and methods have been deeply entrenched in space exploration since its early days, providing powerful tools for collecting, transmitting, and processing data. The least-squares method itself, and its famous recursive version, are the outcome of a data fitting exercise by Gauss in 1795 while trying to predict the location of the comet Ceres from past rudimentary telescope measurements. More recently, in a lecture given by the French mathematician Yves Meyer (of wavelets fame) at EPFL in Switzerland in September 2017, the speaker's opening statement was to show how "signal processing has played a role in the detection of gravitational waves!"

It is not a secret that signal processing theory and methods have been deeply entrenched in space exploration since its early days.

The observation of these waves is considered one of the most important discoveries of recent times [1]. For the uninformed reader, the existence of gravitational waves, which amount to invisible ripples in the space–time fabric, was predicted in the early 1900s, but their detection has remained elusive for more than a century until their discovery in February 2016. It is no wonder that three of the scientists involved in the discovery were awarded the 2017 Nobel Prize in Physics almost instantly. The gravitational waves they detected resulted from the collision of two black holes a mere 1.3 billion years ago! Space exploration at this level has often enabled the discovery, testing, and validation of deep scientific theories including Einstein's theory on how planets and stars distort space and time. Experimental validation of scientific theories is a precious exercise because it tests our hypotheses, deepens our understanding, and propels us to explore more confidently. Space exploration has also led to many technological advances that have benefited humanity right here on Earth.

Still, and oddly enough, we have been shamefully less successful at exploring our own planet Earth. According to the National Ocean Service of the U.S. Department of Commerce [2], oceans cover 70% of the surface of our planet, and yet about 80% of them remain unexplored and their floors largely unmapped to accurate measures. Stated another way, we are ignorant about half of the planet on which we live! Imagine if all the water covering our oceans and seas were to disappear, what would you get? You would be left with vast expanses of land. If you were to drive your car through this wilderness, you will be on your own for almost half of the earth's surface; no online maps would be available to guide you!

There are, of course, many reasons why we have not explored our oceans

more vigorously. Besides the extreme environment that one encounters as we move deeper into the oceans, and the more limited resources available for ocean exploration, humans appear to have a natural fascination for space exploration. Just observe how some of the most successful entrepreneurs of our times have marched almost by inertia toward commercial enterprises to explore spaceflight opportunities. These include, according to data from *Wikipedia*, companies such as Blue Origin founded in 2000 by Jeff Bezos (of Amazon), SpaceX founded in 2002 by Elon Musk (of Tesla), and Virgin Galactic founded in 2004 by Richard Branson (of Virgin Group). To a lesser degree, some entrepreneurs have ventured into exploring the deep sea, including the impressive 2012 Deep Sea Challenger submersible of Canadian filmmaker James Cameron (of the *Titanic* film), and the Schmidt Ocean Institute founded in 2009 by the former Google chairman Erich Schmidt.

That said, whether in space or on our planet, we readily identify a frontier calling out for attention. There is a clear need for the development of more science, technology, and methods for the exploration of extreme environments. Signal processing scientists can and will play an important role in enabling these developments. Why? Because, by training, we are experts at drawing inferences about unobservable variables from indirect measurements. There are many success stories, including scattering methods for detecting layer boundaries in geophysics or oil exploration applications, and noninvasive imaging techniques such as MRIs for biomedical applications, and sonar technology. In fact, this latter technology is already one of the main techniques used to map the ocean floor up to 100 meter resolution. However, only 10% of the oceans' floors have been mapped by the technology and the mapping that exists for the remaining surface has a poor resolution in the order of 5 km [3]. With this state of affairs, one can better understand why it has been such a daunting task to locate the aircraft of Malaysia Airlines flight 370, which tragically disappeared back in March 2014. Imagine how much discovery is awaiting us in the unexplored oceans: new materials, precious metals or minerals that may have gone undiscovered, species with wondrous biomechanisms that may motivate new technologies, and even undiscovered substances that may lead to new medical treatments.

Whether in space or on our planet, we readily identify a frontier calling out for attention.

These facts are humbling. We pride ourselves on the technological advances of the 21st century, such as the ability to track (whether legal or not, right or wrong) every online click and every cell phone user, and yet we still cannot locate a missing aircraft! Even more humbling, there is so much we do not know right here on Earth. I am always amazed at the discovery of new species. We are desperately looking for the tiniest forms of life on remote planets, and yet we continue to be ignorant of the full biodiversity that encircles us. According to [4], it is estimated that 18,000 new species are named every year. And we are not talking about tiny species. In 2018, a new species of the great apes was discovered called the *Pongo Tapanuliensis* orangutan (only 800 of them are left in the Indonesian island of Sumatra). According to the United Nations Environmental Programme [5], it is believed that about 150–200 species become extinct every day. How many of these extinct species belong to a group that we may not have discovered yet?

Even while working on this column, it was announced on CNN's website in August 2018 that a "Never before seen Amazon tribe" has been spotted on drone video. Isn't that astonishing? We are referring to spotting unseen human beings, like you and me, on planet Earth in 2018! The indigenous people spotted in this video live in a large protected area in the Javari Valley in Brazil. Almost a week later, the same CNN website announced the discovery of an 85-mile long deep-sea coral reef off the east coast of the United States; one of the most technologically advanced nations on Earth! All of this was hiding in plain sight.

You can now understand why I feel frustrated and surprised when someone asks, "what else is there to do?" Their argument is that we live in the 21st century and our "advanced" civilization has attained so much sophistication in its technology from the online revolution, to intelligent machines, to deep space exploration, that there is not much more to discover. Some use this argument broadly referring to science in general, and others are more specific and target our signal processing discipline. Luckily enough, there is so much we do not know and may not even come to understand fully. There are so many unanswered questions, and so much opportunity for new methods in science, including in signal and information processing, that the path forward is limitless. We

have only explored the tiniest fraction of space and a fraction of planet Earth. Not to mention many other areas of exploration in biology, basic and natural sciences, social sciences, and so forth.

Discovery never ends. The uninformed sees an obstacle where there is a wall. The scientist wants to see through the wall or jump over it. This is also true at a more abstract level. I would assume that many of you have shared a similar experience with me. When I derive a new result, I often sit back in awe wondering at how "the more we learn, the less we actually know!" In other words, similar to how this new result was hiding in some invisible space waiting for someone to discover it, many more discoveries are awaiting their chance to be brought forward for all of us to admire. How many more unknowns are there? Enough to keep our curious minds busy for ages.

We have only explored the tiniest fraction of space and a fraction of planet Earth.

The human race has always been fascinated by exploration including in many literary works. Jules Verne's 1870 classic *Twenty Thousand Leagues Under the Sea* chronicled the adventures of a fictional submarine and its exploration of the world's oceans. Interestingly, his book was preceded by Verne's 1864 earlier classic *Journey to the Center of the Earth*. That is another frontier yet to be explored. Exploration and discovery will never end. For as long as we look up into the skies, we will continue to wonder at the twinkling stars and the mysteries that lie beyond them.

Once, two new graduate students walked into my office showing interest in joining my research team. One student had just completed his undergraduate studies while the second student had completed his master-level studies. I printed a research article and asked them to return in a week to present it to me. The undergraduate student was understandably concerned that the other student is better prepared to read the article given his more advanced studies. I assured them that my criterion to judge their presentations would be different. The undergraduate student would need to convince me that he understood what *was* written in the paper, while the master student would need to convince me that he understood what *was not* written in the paper (such as discussing any assumptions or approximations that could be relaxed or are limiting). The students were expected to approach and critique the paper from different perspectives. Even here, in this simple exercise of reading a paper, one can find opportunities to push knowledge and discovery further.

References

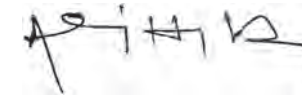
[1] N. Drake and M. Greshko, "Gravitational waves 101," National Geographic, Oct. 16, 2017. [Online]. Available: https://news.nationalgeographic.com/2017/10/what-are-gravitational-waves-ligo-astronomy-science/. Accessed on: Sept. 18, 2018.

[2] [Online]. Available: https://oceanservice.noaa.gov/facts/exploration.html. Accessed on: Sept. 18, 2018.

[3] J. Copley, "Mapping the deep, and the real story behind the 95% unexplored oceans," Online post, University of Southampton, Oct. 2014. [Online]. Available: http://moocs.southampton.ac.uk/oceans/2014/10/04/mapping-the-deep-and-thereal-story-behind-the-95-unexplored-oceans/. Accessed on: Sept. 18, 2018.

[4] K. Loria, "Scientists just discovered these 10 bizarre and beautiful animal species that show what it takes to survive on Earth against the odds," Business Insider, May 23, 2018. [Online]. Available: http://uk.businessinsider.com/new-animal-species-top-10-2018-5?r=US&IR=T. Accessed on: Sept. 18, 2018.

[5] J. Vidal, "Protect nature for world economic security, warns UN biodiversity chief," The Guardian, Aug. 16, 2010. [Online]. Available: https://www.theguardian.com/environment/2010/aug/16/nature-economic-security. Accessed on: Sept. 18, 2018.



SP

John Edwards

Something to Talk About: Signal Processing in Speech and Audiology Research

Promising investigations explore new opportunities in human communication

Speech, the expression of thoughts and feelings by articulating sounds, is an ability so taken for granted that few people bother to think about how complex and nuanced the process actually is. Yet, as more devices gain the ability to listen to and interpret what speakers are saying, speech and audiology technologies are attracting the interest of a growing number of academic researchers. Signal processing is now playing a critical role in making speech detection and recognition more accurate, flexible, and reliable for use in a wide range of research and everyday applications.

Singing mice

Vocalization plays a critical role in social communication across many species. Male mice, for example, generate ultrasonic vocalizations (USVs) in the presence of females. Both male and female mice "sing" during friendly social encounters.

Although mice are extensively used for research into autism and other areas, studying their USVs has long challenged experts. Fortunately, researchers may soon have access to a sophisticated new investigatory technology. A joint collaboration between the Children's Hospital Los Angeles and the University of Southern California's (USC's) Viterbi School of Engineering has led to a new signal processing tool that aims to enable unbiased, data-driven analyses of mouse vocalizations.

"Signal processing methods hold the promise of offering objective, scalable, and reproducible means for characterizing animal behavior, such as communication patterns," says Shrikanth Narayanan, the Niki and C.L. Max Nikias Chair in Engineering at USC. "This opens up tremendous possibilities for researchers in scaling up and accelerating research in many domains, such as neurosciences, genetics, and pharmacology." Narayanan is an electrical engineering professor, computer scientist, and trained linguist who oversees the school's signal analysis and interpretation laboratory, which developed the software.

The interdisciplinary project focused on creating accurate and efficient high-throughput computational methods and tools for discovering and extracting social communication profiles from the USVs of mice. "We are automating procedures that are often done manually," Narayanan states. "Specifically, using a fully automated and unsupervised signal processing approach, Mouse Ultrasonic Profile ExTraction (MUPET) measures, learns, and compares syllable patterns in USVs, enabling users to assess the fine details of syllable production objectively and use it across large numbers of mouse strains and experimental conditions."

The new signal processing tool (Figure 1), featuring a graphical interface, is designed to offer rapid, automated, and unsupervised analysis of ultrasonic mouse vocalizations. With a time and date stamp attached to the vocalizations, the researchers believe that that their tool will prove useful in correlating vocalizations with video-recorded behavioral interactions, allowing additional information to be mined from mouse models relevant to the social deficits experienced by people with autism.

"Our MUPET approach employs unbiased discovery of hundreds of unnamed syllable patterns, while other approaches and tools that are available generate a smaller number of named categories based on predefined rules," Narayanan says. The team is now offering MUPET as an open-access software tool to the research community.

"Signal processing is core to my laboratory's work, to analyze and understand the science of human and, sometimes, animal communication and to develop technologies that support and enhance human experiences," Narayanan explains. "In research, I deal with many types of signals, such as audio, video, movement sensors, and physiological data streams."

The approach involves several signal processing methods, including audio preprocessing and signal conditioning, spectral-domain feature extraction using a gammatone-scaled filterbank, automatic syllable segmentation and clustering (i.e., k-means), and automatic repertoire generation. "We are continuing to explore new signal processing and machine-learning approaches for

Digital Object Identifier 10.1109/MSP.2018.2847564
Date of publication: 13 November 2018

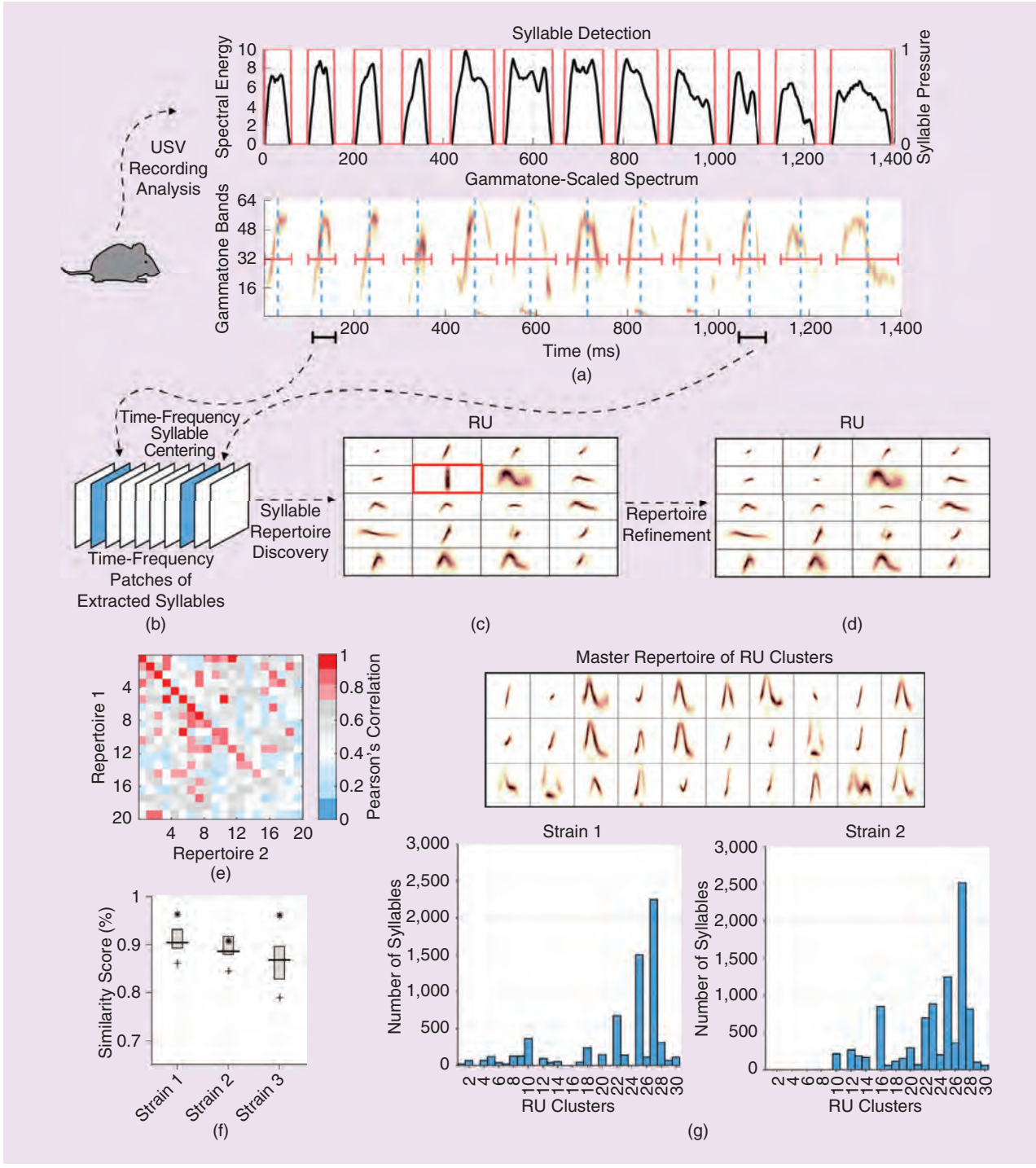


FIGURE 1. The computational framework for syllable repertoire learning and repertoire analysis functions. (a) The mouse USV recordings are loaded into the MUPET tool, and the syllable detector segments individual syllables by measuring the power spectrum (black lines) in the ultrasonic range and comparing it with a noise threshold. The regions of vocalized activity/nonactivity (red boxes, top panel) are used to extract the syllable shapes to form the gammaton filter ultra sonic vocalization (GFUSV) spectral representation (bottom panel). The center (dashed blue line) and duration (red horizontal line) of the GFUSV and key spectrotemporal features are automatically measured. (b) During processing, the extracted syllable shapes are centered along the time and frequency axes and subsequently vectorized before being stacked into a data matrix. Iterative clustering is then performed with (c) and (d). The algorithmic output (c) is a collection of exemplar repertoire units (RUs) (i.e., cluster centroids), which show the average shape of the different syllables that recur in the data set. RUs learned from noise [the red box in (c)] are removed during syllable repertoire refinement in (d). (e) and (f) The MUPET compares the shapes of RUs from different repertoires using two similarity metrics. (e) The cross repertoire similarity matrix gives the Pearson correlations between RU pairs from two different repertoires, which are sorted from highest to lowest shape similarity (see diagonal), regardless of frequency of RU use in each repertoire. (f) The cross repertoire similarity boxplot gives the Pearson correlations between collections of RUs, which represent the top 5%, 25%, 50%, 75%, and 95% of most the frequently used RUs in each repertoire. (g) To compare the frequency of use of similar and unique RU types across different data sets, the MUPET performs a cluster analysis of RU types to generate a master repertoire of RU clusters (top panel). The MUPET provides information on the frequency of use of each RU cluster, enabling the user to identify shared and unique RU types and usage across strains or conditions (bottom panels). (Figure courtesy of the Signal Analysis and Interpretation Laboratory, Viterbi School of Engineering, USC.)

discovering primitives, or basic units, in the recorded data and how they temporally pattern," Narayanan says. "We are also extending the methods … to look at multimodal data, such as video, in conjunction with USVs, modeling the temporal patterning of the primitives, and [creating] predictive models using these features to predict outcomes. The potential of signal processing and machine learning as both tools for providing scale and efficiency as well as novel discovery in biomedical sciences is exciting," she states. "We are also looking at applying these methods to analyzing behavior in other species."

Recording the inaudible

Researchers at the Coordinated Science Laboratory at the University of Illinois have created a unique type of sound that is entirely inaudible to people at 40 kHz or higher, yet can be detected by virtually any microphone. The sound combines multiple tones that, when interacting with the microphone's circuitry, create what researchers describe as a shadow, a sound that microphones can easily detect.

The research team, including Ph.D. degree students Nirupam Roy and Sheng Shen, as well as Prof. Romit Roy Choudhury and Prof. Haitham Hassanieh, anticipates the development of multiple commercial and government applications based on their work.

"We show that these high-frequency sounds can be designed to become recordable by unmodified microphones while remaining inaudible to humans," Shen states. The technology focuses on exploiting nonlinearities in microphone hardware.

"We design the sound and play it on a speaker such that, after passing through the microphone's nonlinear diaphragm and power amplifier, the signal creates a shadow in the audible frequency range," Shen explains. The shadow can be configured to carry data bits, adding an acoustic, yet inaudible, communication channel to current microphone technology. "We designed a system called *BackDoor* that develops the technical building blocks for harnessing this opportunity."

BackDoor can utilize a microphone's entire spectrum for communication purposes, as shown in Figure 2. "Thus, Internet-of-Things devices could find an alternative channel for communication, reducing the growing load on Bluetooth," Shen says. Museums and shopping malls, for instance, could use BackDoor to power acoustic beacons that broadcast information about nearby artworks or products. Other applications include the live watermarking of concert music, stealthily tagging songs, and enhanced navigation systems. "Various ultrasound ranging schemes that compute the time of flight of signals could benefit from the substantially higher bandwidth in BackDoor," he observes.

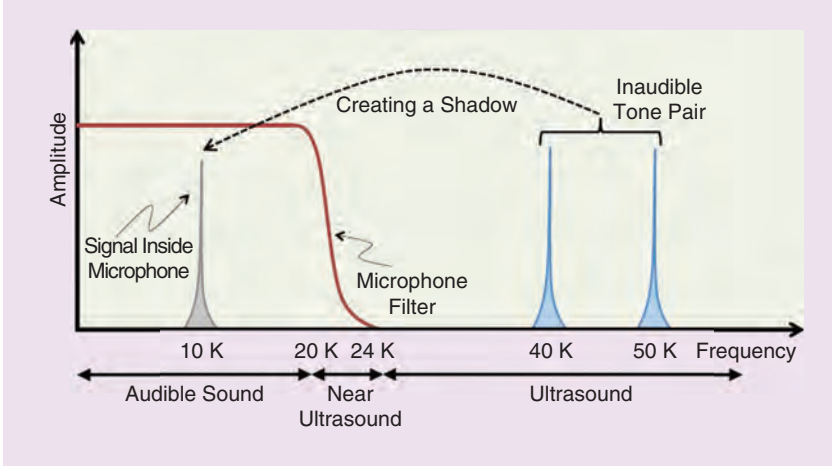


FIGURE 2. BackDoor, developed by the Coordinated Science Laboratory at the University of Illinois, utilizes a microphone's entire spectrum to create an alternative communication channel. A video presentation of the technology can be viewed at https://youtu.be/_FrKySibcb8. (Figure courtesy of the Coordinated Science Laboratory, University of Illinois.)

BackDoor also has the potential to be misused. Shen observes that inaudible jammers can use the technology to disable hearing aids and cell phones without being detected. "For example, during a robbery, the perpetrators can prevent people from making 911 calls by silently jamming all of the phones' microphones," he explains.

Shen notes that an attacker could also design an inaudible BackDoor signal to mimic a human voice. "This can empower an adversary to stand on the road and silently control Amazon Echo and Google Home-like devices in people's homes," Shen warns. "A voice command like, 'Alexa, open the garage door,' can be a serious threat."

Signal processing played an important role in resolving several significant challenges BackDoor's developers faced. "The nonlinearities we intend to exploit are not unique to the microphone; they are also present in speakers that transmit the sounds," Shen notes. "As a result, the speaker also produces a shadow within the audible range, making its output audible to humans."

The researchers addressed this issue by using multiple speakers and then isolating the signals in frequency across the speakers. "We show, both analytically and empirically, that none of these isolated sounds create a shadow as they pass through the speaker's diaphragm and amplifier," he reports. "However, once these sounds arrive and combine nonlinearly inside the microphone, the shadow emerges within the audible range."

Another important challenge the researchers resolved is allowing standard modulation and coding schemes to be used directly in communication applications. "We show how appropriate frequency modulation, combined with inverse filtering, resonance alignment, and ringing mitigation, is needed to boost achievable data rates," Shen says.

Finally, for security applications, jamming requires transmitting noisy signals that cover the entire audible frequency range. "With audible jammers, this requires speakers to operate at very high volumes," Shen explains. "We leverage the adaptive gain control in microphones, in conjunction with selective frequency

distortion, to improve jamming at modest power levels."

Shen is confident that BackDoor will eventually find many useful applications. "Nonlinearity is typically an enemy," he says. "We are beginning to think there is a way to make nonlinearity a friend."

A cognitive hearing aid

Hearing-impaired people often struggle to follow conversations in busy, noisy environments, such as crowded restaurants and offices. Although current hearing aids are generally useful for suppressing background noise, they're relatively helpless at assisting a listener detect who is talking in a conversation being conducted between multiple individuals. A cognitive hearing aid that constantly monitors its user's brain activity to determine whether a subject is conversing with a specific speaker would effectively solve this problem.

Tapping into deep neural network (DNN) models, researchers at Columbia University's Fu Foundation School of Engineering and Applied Science claim they have made a breakthrough in auditory attention decoding (AAD) methods and are coming closer to making cognitively controlled hearing aids a reality (Figure 3). The research, led by Nima Mesgarani, an associate professor of electrical engineering, was conducted in collaboration with Columbia University Medical Center's Department of Neurosurgery, Hofstra-Northwell School of Medicine, and the Feinstein Institute for Medical Research.

"This work combines the state of the art from two disciplines: speech signal processing and auditory attention decoding," Mesgarani says. "We developed an end-to-end system that receives as input a single audio channel containing a mixture of speakers heard by a listener along with the listener's neural signals." The system then automatically separates the individual speakers, determines which speaker is being listened to, and amplifies that speaker's voice to assist the listener. The entire process is completed in fewer than 10 s.

"We came up with the idea of a cognitively controlled hearing aid after we

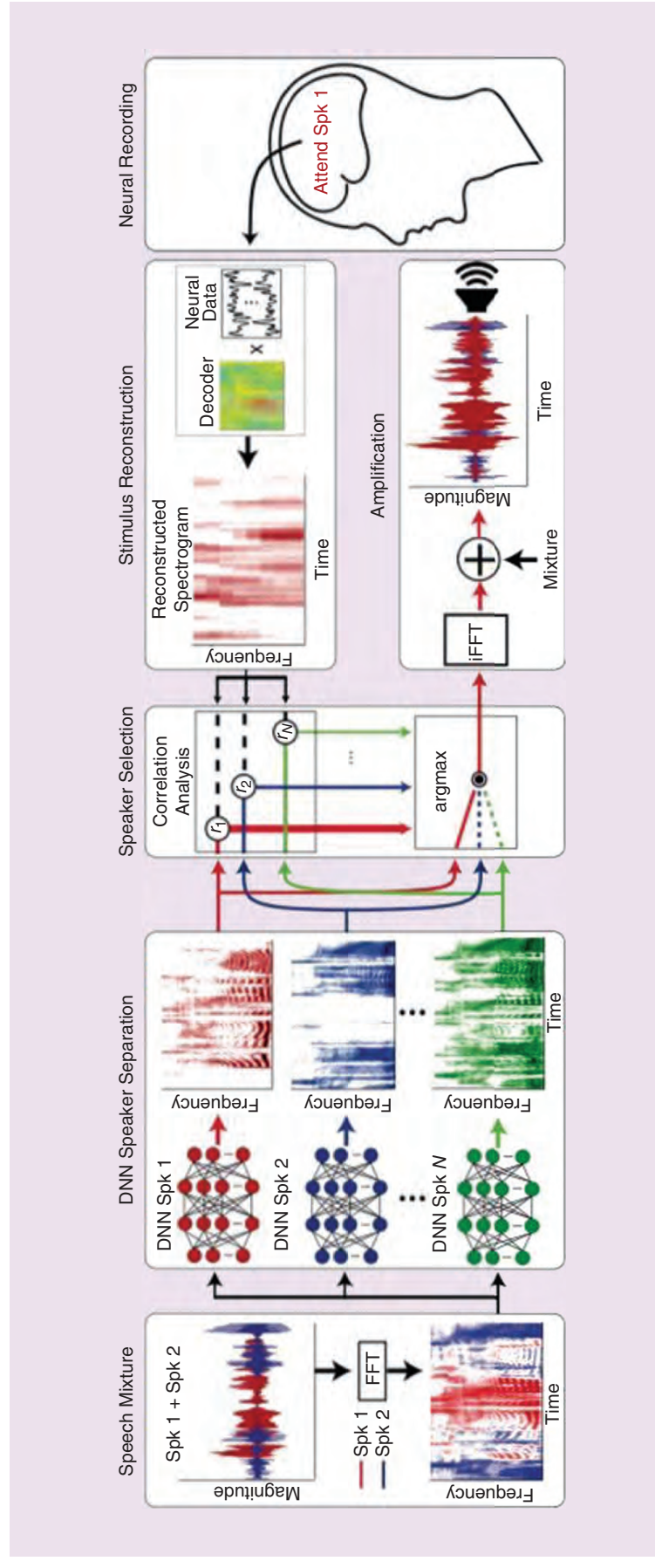


FIGURE 3. A cognitively controlled assistive hearing device can automatically amplify one speaker among many. To do so, a DNN automatically separates each of the speakers from the mixture and compares each speaker with the neural data from the user's brain. The speaker that best matches the neural data is then amplified to assist the user. IFFT: inverse Fourier transform; Spk: speaker. (Figure courtesy of Nima Mesgarani/Columbia Engineering.)

demonstrated it was possible to decode the attended target of a listener using neural responses in the listener's brain [via] invasive neural recordings," Mesgarani reports. "Later, we showed that we could decode attention with noninvasive methods as well."

Several other research groups are also working on cognitively controlled hearing aid technologies, Mesgarani notes. "However, our current study is a breakthrough in removing a major obstacle toward real-world implementation of this idea, which is to remove the need to have clean sources."

The system works with two distinct signal types: neural and acoustic. The neural signal is recorded at a high sampling frequency. Wavelet decomposition is then used to isolate the signal's high-frequency component, since that is the most relevant part of the signal for attention decoding. "The high-frequency part of the neural signal reflects mostly the neural spiking activity in the brain near the electrodes," Mesgarani explains. "We then use a Hilbert transform to estimate the envelope of the high-frequency components."

On the acoustic side, the incoming sound is decomposed into different frequencies using a Fourier transform. The sound's magnitude then goes through several neural network models that are designed to separate the signal of a particular speaker among many. The networks in effect modify the magnitude of the Fourier transform, which is then combined with the original phase to perform an inverse Fourier transform to reconstruct the modified audio. Signal correlation analysis is used to find the similarity of the brain signals with the acoustic signal of each neural network; the separated audio that most resembles the neural activity is chosen and subsequently amplified.

The team recently tested the technology by using invasive electrocorticography recordings obtained from neurological subjects undergoing epilepsy surgery. The recordings enabled the researchers to locate the precise regions of the auditory cortex that contribute to AAD. Using this information, they discovered that their system decoded the attention of the listener and amplified the voice he or she wanted to hear using just the mixed audio.

Several technical barriers must still be overcome before a commercial version of the technology can be brought to market. Leading the list is the development of new algorithms to process local sounds and synthesize which voice is ideal for the listener to hear based on the engaged task. The project also needs to find a way to provide sufficient computational power to implement the sophisticated technology inside a small, wearable device.

"All of these are active areas of research and have seen significant improvements in recent years," Mesgarani notes. "There is no theoretical reason prohibiting the implementation of this technology in an actual hearing aid and, in fact, several hearing aid companies have already started researching this idea and expressed interest in our approach."

Author

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

John Edwards

Signal Processing Leads to New Clinical Medicine Approaches

Innovative methods promise improved patient diagnoses and treatments

Popular consumer and business technologies, such as smartphones, tablets, wearable devices, and sophisticated photoimaging—all driven or supported by signal processing—are leading to a generation of powerful new diagnostic tools designed to help physicians working in clinical medicine. In Rochester, New York, for instance, a team of engineers and clinicians at the Rochester Institute of Technology (RIT) and the University of Rochester Medical Center (URMC) is developing a video-based smartphone/tablet-based health app (Figure 1) that is designed to serve as a clinical tool to assess atrial fibrillation (AF), a heart-rhythm disorder that afflicts millions of people worldwide. Co-project leaders are Gill Tsouri, an associate professor of electrical engineering in RIT's Kate Gleason College of Engineering, who is developing both the app and its video system algorithm, and Jean Philippe Couderc, a biomedical engineer and assistant director of the University of Rochester Heart Research Follow-Up Program Lab, who will head the clinical trials.

The prime task of the app is to detect AF in high-risk populations. "In order to detect AF, we monitor the heart rate and its variability," Tsouri says. "This means that other applications that rely on these biometrics are possible, too, such as monitoring stress, providing biofeedback, and detecting other types of arrhythmias." The app's video plethysmography technology is designed to replace sensors utilizing skin contact, such as pulse oximeters, with noncontact video cameras. "It takes advantage of subtle blushing in skin color as blood is being pumped by the heart to and from the

Digital Object Identifier 10.1109/MSP.2018.2865889
Date of publication: 13 November 2018

face," Tsouri reports. He continues, "Our basic approach in this project is to have the frontal camera (on the smartphone or tablet) take periodic recordings of the face of the user to track cardiac activity and use it to detect and monitor AF." The monitoring occurs in the background, leaving the user free to do other things, such as read e-mail or watch a movie. "This approach has significant advantages, among them the ability to provide an AF monitoring service to virtually anyone just by providing a downloadable app," Tsouri notes. "The physician does not have to rely on patient compliance to receive the required measurements."

The approach is also inexpensive since there is no need for a dedicated sensor. Cardiac-monitoring technologies that are currently available rely on cumbersome and costly dedicated sensors that require either continuous skin contact or active subject participation. Signal processing is used to detect the subject's face in the video images, extract signals from the images, and preprocess the signals using filtering and detrending. "We then apply our signal processing algorithms to extract cardiac signals and activity from the preprocessed signals," Tsouri says. "We also use signal processing algorithms to identify and mitigate the effects of motion, shaking, and varying ambient light conditions." He goes on, "Typically, we use face detection algorithms to identify the face in the video image; extract the red, green, and blue pixels from the face; and use basic signal processing to generate three signals corresponding to the three colors." Tsouri further explains, "We then apply our signal processing algorithms to infer cardiac activity based on these three signals."

For the critical step of extracting a cardiac signal from red-green-blue (RGB) signals, the researchers favor color conversion over blind deconvolution. "We noticed in past research that the pulsating heart is expressed better in color spaces with the trace hue," Tsouri says. "Conversion from RGB to hue is much simpler than applying blind deconvolution methods, and it provides an immediate signal per frame, unlike blind deconvolution that relies on processing a block of frames thereby introducing high complexity and latency."

To test the approach and its algorithms, the team conducted a large-scale clinical study with partners at the URMC, led by Couderc. The study involved 300 subjects, who each received a tablet containing the app for a period of two weeks. Cardiac activity was monitored by both the app and a U.S. Food and Drug Administration-approved electrocardiogram patch attached to the subject's chest as a reference source. Tsouri notes that the research is still in the learning phase. "We would need to improve our algorithms based on the data we receive to make sure we obtain reliable measurements." He cautions that the project should be perceived as just one element in a growing digital healthcare trend that is rapidly moving clinicians from disease-based medicine to preventative medicine. "The technology we are developing provides a low-cost and easily accessible cardiac-monitoring solution that relies on the smart devices people already have in their possession," Tsouri states, noting that the approach can help detect cardiac problems well before they manifest as a disease and leverage the smart devices' Internet connection to send data to cloud services. "With the growing proliferation of smartphones and tablets, this technology can become ubiquitous and accessible to all populations regardless of their geographical location and social and financial status."

Mood monitor

Rose T. Faghih, an assistant professor of electrical and computer engineering at the University of Houston, Texas (Figure 2), is investigating whether wrist-worn wearable devices, similar to models offered by Fitbit and Apple, can be used to monitor stress and related conditions. She believes they can. "We are developing algorithms to monitor mental-stress-related arousal and fatigue using measured galvanic skin response (GSR) and cortisol, respectively," Faghih says. While numerous consumer-level stress-tracking wearables already exist, they can't interpret brain activity related to stress. "The ones currently available track heart rate as an indicator of stress," she notes.

The project uses state-space modeling and Bayesian filtering methods to extract stress from skin conductance and cortisol measurements. "We also

FIGURE 1. Engineers and clinicians at RIT and URMC are developing a video-based smartphone/tablet-based health app. The software is designed to serve as a clinical tool to assess AF, a heart-rhythm disorder that afflicts millions of people worldwide. (Photo courtesy of the RIT and URMC.)

FIGURE 2. Rose T. Faghih, assistant professor of electrical and computer engineering at the University of Houston, Texas, is investigating whether wrist-worn wearable devices, similar to models offered by Fitbit and Apple, can be used to monitor stress and other emotions. (Photo courtesy of the University of Houston, Texas.)

use compressed sensing methods in this research due to the sparse nature of the secretory events that underlie the measured signals," Faghih explains. While sweat-based cortisol wearable devices have been developed in research laboratories, such sensors have not yet been integrated into commercial smart watches. "Once smart watches on the market integrate a cortisol sensor…our algorithms can be worked into the existing smart watches to monitor fatigue," Faghih says.

Decoding brain states using wrist-worn wearable devices also promises to transform how mental-stress-related conditions are diagnosed and treated. "The transformative tool sets developed could be used by individuals in tracking their brain dynamics-related fatigue and arousal using wearable devices and by clinicians for monitoring patients' physical health," Faghih explains. "Our method relates a person's internal stress state to the probability that tiny bursts of sweat are released by the body, resulting in the occurrence of spikes in a skin conductance signal," she says. "Relating the internal stress state to this spike occurrence probability, we perform Bayesian filtering and smoothing to extract stress from skin conductance." Faghih further adds, "Increases in spike probability indicate an increase in stress." The stress level is then quantified in terms of an offline certainty level that indicates when the spike probability is above its baseline. "In our analysis, we noted high stress levels when subjects are involved in active cognitive tasks, while a gradual decline occurs when they're only engaged passively, such as when watching a clip from a horror movie," Faghih explains.

The GSR algorithms could soon be worked into existing smart watches. "For example, Microsoft Band already measures conductivity of the wearer's GSR to determine if the individual is wearing the band," Faghih says. "Our algorithms can use the measured GSR data collected by Microsoft Band to examine brain activity."

Examining brain connectivity

The TrueBrainConnect project takes another approach to analyzing brain activity. Hosted by Charité–Universitatsmedizin Berlin, Germany, and its Center for Advanced Neuroimaging, the project aims to systematically study connections between different areas of the brain, potentially drawing conclusions regarding possible disease patterns. A team of researchers, headed by Stefan Haufe, a computer scientist and machine-learning expert, plans to generate complex models capable of forecasting various mental states.

Funded by the European Research Council, TrueBrainConnect's primary goal is to enable the reliable analysis of functional brain connectivity. Using electroencephalography (EEG) and magnetoencephalography (MEG), the technology promises to help clinicians and others determine which brain regions send information to which other brain regions during a given mental task. "By relating such information to task parameters, behavior, and potentially clinical variables, using classical statistics and machine learning, we hope to better understand how the brain works in health and disease," Haufe states.

Two-thirds of the project, Haufe notes, is focused on the development of data analysis methods. "This work touches the fields of inverse problems, machine learning, and, of course, signal processing, such as statistical source separation, spectral decompositions, and analysis of causal interactions between time series," he explains. The rest of the project is devoted to its potential clinical applications. "We hope that the developed methods are useful to derive neurophysiologically interpretable predictions (and) diagnoses of disease states," Haufe adds.

The project is based on the premise that many neurological disorders have roots that are well established before the onset of any symptoms or before they produce noticeable changes in brain structure or behavior. Team members believe that these disorders reveal their presence through irregularities in the way different areas of the brain communicate with each other. In the hope of improving the prognosis of pathological conditions affecting the brain, the researchers are developing a method (Figure 3) capable of reliably estimating

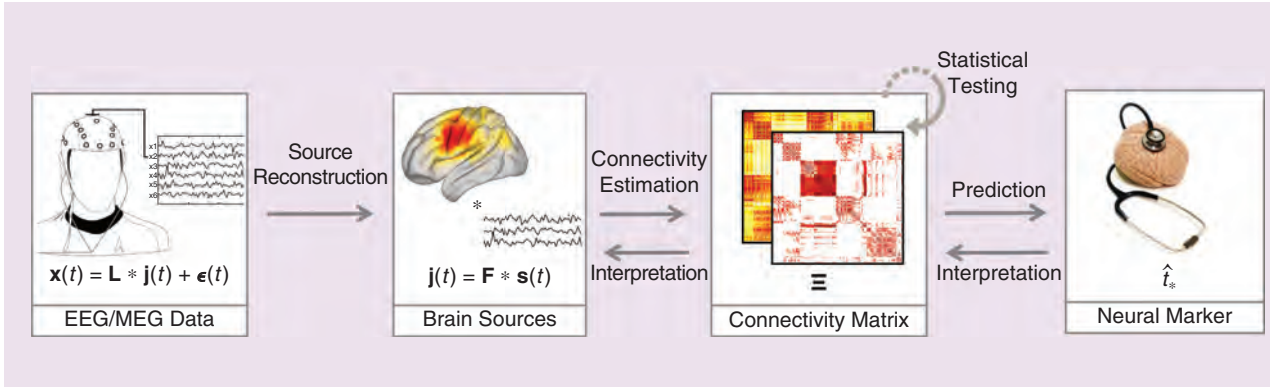


FIGURE 3. The planned methodology of TrueBrainConnect, an approach for reliably estimating and localizing brain interactions with the goal of improving the prognosis of pathological conditions affecting the brain. (Image courtesy of Stefan Haufe.)

and localizing brain interactions. "Most importantly, we could show that some of the most established analysis methods have serious shortcomings, which renders them problematic for the purpose of EEG/MEG analysis," Haufe says.

Existing techniques for analyzing neuroimaging data are not yet sufficiently developed, and robust conclusions cannot be drawn from them. The researchers believe that new signal processing and machine-learning techniques will allow them to make precise determinations regarding brain signal sources and actual nerve-cell interactions. The researchers are primarily focusing on EEG data, Haufe says, because it is the only neuroimaging technique that is noninvasive, has a high temporal resolution (which enables the study of functional interactions at the temporal scale of actual neuronal activity), is inexpensive, mobile, and offers direct measurements of neuronal activity. "These advantages also make it an attractive candidate for clinical diagnosis," he observes.

Many of the world's most prevalent diseases are age-related neurological conditions, such as Parkinson's disease and dementia. "These conditions do not have cures yet, and even diagnosis is typically possible only at late disease stages," Haufe notes. Current diagnostic tools, such as nuclear imaging and tissue probes, tend to be expensive and/or invasive, which limits their use in preemptive medical screening. The TrueBrainConnect project is based on the hypothesis that many brain diseases are characterized, if not caused by, impaired communication between different brain sites. "If we would be able to see such early signs of miscommunication using EEG, this would open the door to early diagnosis," Haufe reports.

The biggest challenge facing the researchers is the fact that using the EEG tends to be a very difficult way to handle the signals. "The EEG data analysis pipeline is very complex, and nearly all parts require more or less advanced signal processing," Haufe says. Although the EEG is the oldest neuroimaging technique—developed in the late 1920s—and there has been a recent surge of research on using the approach for neuroimaging, there has been no breakthrough yet. "My hypothesis is that there is still a lack of appropriate methods for extracting the relevant nontrivial brain dynamics related to clinical brain states," Haufe notes.

Haufe says that the researchers are testing all of their approaches with extensive numerical simulations. "Based on these simulations, we choose the approach that best recovers the ground truth with as little variance as possible," he adds, noting that, over the years, some understanding inevitably emerges about which approaches work and don't work and why. Handling data complexity is a key challenge for the team. "We are dealing with multivariate time series sampled at several hundred hertz and evaluated at potentially several thousand brain sites; thus, a data set can easily comprise several gigabytes of memory," Haufe explains. "Having memory- and runtime-efficient algorithms are, therefore, critical, especially when interactions between brain sites need to be analyzed," he notes.

An even larger challenge affecting the analysis of EEG and other neurophysiological time series data is the low signal-to-noise ratio (SNR) of brain signals as well as the mixing of brain signals into EEG channels and reconstructed sources. "The latter can also be seen as a problem of low SNR—brain activity of interest versus other brain processes and artifacts—or as a problem of spatially correlated noise," Haufe says. It is the correlated noise that makes the data analysis so difficult, he notes. "Correlated noise can effectively obfuscate the interpretation of multivariate prediction models, which would be important in order to understand."

Challenges aside, Haufe believes that his team's research could ultimately lead to advancements in many different clinical areas. "In general, the developed methods could be useful for all disciplines that use EEG/MEG data," he states. It is also possible that some of the yet-to-be-developed methods will find applications in other domains, e.g., the methods for estimating nontrivial interactions between time series or for solving electromagnetic inverse problems, Haufe notes.

Author

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

SP

Model Selection Techniques

An overview

©ISTOCKPHOTO.COM/GREMLIN

Jie Ding, Vahid Tarokh, and Yuhong Yang

In the era of big data, analysts usually explore various statistical models or machine-learning methods for observed data to facilitate scientific discoveries or gain predictive power. Whatever data and fitting procedures are employed, a crucial step is to select the most appropriate model or method from a set of candidates. Model selection is a key ingredient in data analysis for reliable and reproducible statistical inference or prediction, and thus it is central to scientific studies in such fields as ecology, economics, engineering, finance, political science, biology, and epidemiology. There has been a long history of model selection techniques that arise from researches in statistics, information theory, and signal processing. A considerable number of methods has been proposed, following different philosophies and exhibiting varying performances. The purpose of this article is to provide a comprehensive overview of them, in terms of their motivation, large sample performance, and applicability. We provide integrated and practically relevant discussions on theoretical properties of state-of-the-art model selection approaches. We also share our thoughts on some controversial views on the practice of model selection.

Digital Object Identifier 10.1109/MSP.2018.2867638
Date of publication: 13 November 2018

Why model selection

Vast developments in hardware storage, precision instrument manufacturing, economic globalization, and so forth have generated huge volumes of data that can be analyzed to extract useful information. Typical statistical inference or machine-learning procedures learn from and make predictions on data by fitting parametric or nonparametric models (in a broad

sense). However, there exists no model that is universally suitable for any data and goal. An improper choice of model or method can lead to purely noisy discoveries, severely misleading conclusions, or disappointing predictive performances. Therefore, a crucial step in a typical data analysis is to consider a set of candidate models (referred to as the *model class*) and then select the most appropriate one. In other words, model selection is the task of selecting a statistical model from a model class, given a set of data. We may be interested, e.g., in the selection of

- variables for linear regression
- basis terms, such as polynomials, splines, or wavelets in function estimation
- order of an autoregressive (AR) process
- number of components in a mixture model
- most appropriate parametric family among a number of alternatives
- number of change points in time series models
- number of neurons and layers in neural networks
- best choice among logistic regression, support vector machine, and neural networks
- best machine-learning techniques for solving real-world data challenges on an online competition platform.

There have been many overview papers on model selection scattered in the communities of signal processing [1], statistics [2], machine learning [3], epidemiology [4], chemometrics [5], and ecology and evolution [6]. Despite the abundant literature on model selection, existing overviews usually focus on derivations, descriptions, or applications of particular model selection principles. In this article, we aim to provide an integrated understanding of the properties and practical performances of various approaches by reviewing their theoretical and practical advantages, disadvantages, and relations.

Vast developments in hardware storage, precision instrument manufacturing, economic globalization, and so forth have generated huge volumes of data that can be analyzed to extract useful information.

Some basic concepts

Notation

We use $\mathcal{M}_m = \{p_{\theta_m} : \theta_m \in \mathcal{H}_m\}$ to denote a model (in the formal probabilistic sense), which is a set of probability density functions to describe the data $z_1, \ldots, z_n$. Here, $\mathcal{H}_m$ is the parameter space associated with $\mathcal{M}_m$. A model class, $\{\mathcal{M}_m\}_{m \in \mathbb{M}}$, is a collection of models indexed by $m \in \mathbb{M}$. The number of models (or the cardinality of $\mathbb{M}$) can be fixed or depend on the sample size n. For each model $\mathcal{M}_m$, we denote by d_m the dimension of the parameter in model $\mathcal{M}_m$. Its log-likelihood function is written as $\theta_m \mapsto \ell_{n,m}(\theta_m) = \log p_{\theta_m}(z_1, \ldots, z_n)$, and the maximized log-likelihood value is

$$\ell_{n,m}(\hat{\theta}_m), \text{ with } \hat{\theta}_m = \operatorname*{argmax}_{\theta_m \in \mathcal{H}_m} p_{\theta_m}(z_1, \ldots, z_n), \tag{1}$$

the maximum likelihood estimator (MLE) under model $\mathcal{M}_m$. We will write $\ell_{n,m}(\hat{\theta}_m)$ as $\hat{\ell}_{n,m}$ for simplicity. We use p_* and E_* to denote the true data-generating distribution and expectation with respect to the true data-generating distribution, respectively. In the parametric framework, there exists some $m \in \mathbb{M}$ and some $\theta_* \in \mathcal{H}_m$ such that p_* is exactly p_{θ_*}. In the nonparametric framework, p_* is excluded in the model class. We sometimes call a model class $\{\mathbb{M}_m\}_{m \in \mathcal{M}}$ *well-specified* (respectively, misspecified) if the data-generating process is in a parametric (respectively nonparametric) framework. We use $\to_p$ and $\to_d$ to denote convergence in probability and in distribution (under p_*), respectively. We use $\mathcal{N}(\mu, V)$ to denote a Gaussian distribution of mean μ and covariance V, χ^2_d to denote a chi-squared distribution with d degrees of freedom, and $\|\cdot\|_2$ to denote the Euclidean norm. The word *variable* is often referred to as the *covariate* in a regression setting.

A typical data analysis can be thought of as consisting of two steps.

- *Step 1*: For each candidate model $\mathcal{M}_m = \{p_{\theta_m}, \theta_m \in \mathcal{H}_m\}$, fit all of the observed data to that model by estimating its parameter $\theta_m \in \mathcal{H}_m$.
- *Step 2*: Once we have a set of estimated candidate models $p_{\hat{\theta}_m} (m \in \mathbb{M})$, select the most appropriate one for either interpretation or prediction.

We note that not every data analysis and its associated model selection procedure formally rely on probability distributions. Examples of model-free methods are nearest-neighbor learning, certain reinforcement learning, and expert learning. Before we proceed, it is helpful to first introduce the following two concepts: the model fitting and the best model.

The model fitting

The fitting procedure (also called *parameter estimation*) given a certain candidate model $\mathcal{M}_m$ is usually achieved by minimizing the following (cumulative) loss:

$$\tilde{\theta}_m = \operatorname*{arg\,min}_{\theta_m \in \mathcal{H}_m} \sum_{t=1}^{n} s(p_{\theta_m}, z_t). \tag{2}$$

In (2), each p_{θ_m} represents a distribution for the data, and $s(\cdot, \cdot)$, referred to as the *loss function* (or *scoring function*), is used to evaluate the goodness of fit between a distribution and the observation. A commonly used loss function is the logarithmic loss

$$s(p, z_t) = -\log p(z_t), \tag{3}$$

the negative logarithm of the distribution of z_t. Then, (2) produces the MLE for a parametric model. For time series data, (3) is written as $-\log p(z_t | z_1, \ldots, z_{t-1})$, and the quadratic loss $s(p, z_t) = \{z_t - E_p(z_t | z_1, \ldots, z_{t-1})\}^2$ is often used, where the expectation is taken over the joint distribution p of $z_1, \ldots, z_t$.

The best model

Let $\hat{p}_m = p_{\hat{\theta}_m}$ denote the estimated distribution under model $\mathcal{M}_m$. The predictive performance can be assessed via the out-sample prediction loss, defined as

$$E_*(s(\hat{p}_m, Z)) = \int s(\hat{p}_m(z), z)p_*(z)dz, \quad (4)$$

where Z is independent with and identically distributed as the data used to obtain $\hat{p}_m$. Here, Z does not have the subscript t as it is the out-sample data used to evaluate the predictive performance. There can be a number of variations to this in terms of the prediction loss function [8] and time dependency. In view of this definition, the best model can be naturally defined as the candidate model with the smallest out-sample prediction loss, i.e.,

$$\hat{m}_0 = \arg\min_{m \in \mathbb{M}} E_*(s(\hat{p}_m, Z)).$$

In other words, $\mathcal{M}_{\hat{m}_0}$ is the model whose predictive power is the best offered by the candidate models. We note that the best is in the scope of the available data, the class of models, and the loss function.

In a parametric framework, typically the true data-generating model, if not too complicated, is the best model. In this vein, if the true density function p_* belongs to some model $\mathcal{M}_m$ or, equivalently, $p_* = p_{\theta_*}$ for some $\theta_* \in \mathcal{H}_m$ and $m \in \mathbb{M}$, then we seek to select such $\mathcal{M}_m$ (from $\{\mathcal{M}_m\}_{m \in \mathbb{M}}$) with probability going to one as the sample size increases, which is called *consistency in model selection*. In addition, the MLE of p_{θ_m} for $\theta_m \in \mathcal{H}_m$ is known to attain Cramer–Rao lower bound asymptotically. In a nonparametric framework, the best model depends on the sample size—typically the larger the sample size, the larger the dimension of the best model because more observations can help reveal weak variables (whose effects are relatively small) that are out of reach at a small sample size. As a result, the selected model is sensitive to the sample size, and selection consistency becomes statistically unachievable. We revisit this point in the "Because All Models Are Wrong, Why Pursue Consistency in Selection?" section.

In this article, we aim to provide an integrated understanding of the properties and practical performances of various approaches by reviewing their theoretical and practical advantages, disadvantages, and relations.

We note that the aforementioned equivalence between the best model and the true model may not hold for regression settings where the number of independent variables is large relative to the sample size. Here, even if the true model is included as a candidate, its dimension may be too high to be appropriately identified based on relatively small data. Then the parametric framework becomes practically nonparametric. We will emphasize this point in the "An Illustration on Fitting and the Best Model" section.

Goals of data analysis and model selection

There are two main objectives in learning from data. One is for scientific discovery, understanding of the data-generation process, and interpretation of the nature of the data. A scientist, e.g., may use the data to support a physical model or identify genes that clearly promote early onset of a disease. Another objective of learning from data is for prediction, i.e., to quantitatively describe future observations. Here the data scientist does not necessarily care about obtaining an accurate probabilistic description of the data. Of course, one may also be interested in both directions.

In tune with the two different objectives, model selection can also have two directions: model selection for inference and model selection for prediction. The first one is intended to identify the best model for the data, which hopefully provides a reliable characterization of the sources of uncertainty for scientific insight and interpretation. And the second is to choose a model as a vehicle to arrive at a model or method that offers top performance. For the former goal, it is crucially important that the selected model is not too sensitive to the sample size. For the latter, however, the selected model may simply be the lucky winner among a few close competitors, yet the predictive performance can still be (nearly) the best possible. If so, the model selection is perfectly fine for the second goal (prediction), but the use of the selected model for insight and interpretation may be severely unreliable and misleading. Associated with the first goal of model selection for inference or identifying the best candidate is the following concept of selection consistency.

Definition 1

A model selection procedure is consistent if the best model is selected with probability going to one as $n \to \infty$. In the context of variable selection, in practical terms, model selection consistency is intended to mean that the important variables are identified and their statistical significance can be ascertained in a follow-up study of a similar sample size but the rest of the variables cannot. In many applications, prediction accuracy is the dominating consideration. Even when the best model as defined earlier cannot be selected with high probability, other models may provide asymptotically equivalent predictive performance. The following asymptotic efficiency property demands that the loss of the selected model or method is asymptotically equivalent to the smallest among all of the candidates.

Definition 2

A model selection procedure is asymptotically efficient if

$$\frac{\min_{m \in \mathbb{M}} \mathcal{L}_m}{\mathcal{L}_{\hat{m}}} \to_p 1 \quad \text{as } n \to \infty, \quad (5)$$

where $\hat{m}$ is the selected model, $\mathcal{L}_m = E_*(s(\hat{p}_m, Z)) - E_*(s(p_*, Z))$ is the adjusted prediction loss, and $\hat{p}_m$ denotes the estimated density function under model m.

The subtraction of $E_*(s(p_*, Z))$ allows for better comparison of competing model selection methods. Another property often used to describe model selection is minimax-rate optimality, which will be elaborated on in the "Theoretical Properties of the Model Selection Criteria" section. A related but different school of thought is the structural risk minimization in the literature of statistical learning theory. In that context,

a common practice is to bound the out-sample prediction loss using in-sample loss plus a positive term (e.g., a function of the Vapnik–Chervonenkis dimension [9] for a classification model). The major difference of the current setting compared with that in statistical learning is the (stronger) requirement that the selected model should exhibit prediction loss comparable to the best offered by the candidates. In other words, the positive term plus the in-sample loss should asymptotically approach the true out-sample loss (as sample size goes to infinity).

The goals of inference and prediction as assessed in terms of asymptotic efficiency of model selection can often be well aligned in a parametric framework, although there exists an unbridgeable conflict when a minimax view is taken to assess the prediction performance. We will elaborate on this and related issues in the "War and Peace—Conflicts Between AIC and BIC and Their Integration" section.

In light of all of the preceding discussions, we note that the task of model selection is primarily concerned with the selection of $\mathcal{M}_m$ ($m \in \mathbb{M}$), because once m is identified, the model fitting part is straightforward. Thus, the model selection procedure can also be regarded as a joint estimation of both the distribution family $(\mathcal{M}_m)$ and the parameters in each family $(\theta_m \in \mathcal{H}_m)$.

A model class $\{\mathcal{M}_m\}_{m\in\mathbb{M}}$ is nested if smaller models are always special cases of larger models. For a nested model class, the model selection is sometimes referred to as the *order selection problem*. The task of model selection in its broad sense can also refer to *method* (or *modeling procedure*) *selection*, which we shall revisit in the "Modeling Procedure Selection" section.

An illustration on fitting and the best model

We provide a synthetic experiment to illustrate the general rules that 1) better fitting does not imply better predictive performance, and 2) the predictive performance is optimal at a candidate model that typically depends on both the sample size and the unknown data-generating process. As a result, an appropriate model selection technique is important to single out the best model for inference and prediction in a strong, practically parametric framework or to strike a good balance between the goodness of fit and model complexity on the observed data to facilitate optimal prediction in a practically nonparametric framework.

Example 1

Suppose that a set of time series data $\{z_t : t = 1, \ldots, n\}$ is observed, and we specify an AR model class with order at most d_n. Each model of dimension (or order) k ($k = 1, \ldots, d_n$) is in the form of

$$z_t = \sum_{i=1}^{k} \psi_{k,i} z_{t-i} + \varepsilon_t, \tag{6}$$

referred to as the AR(k), where $\psi_{k,i} \in \mathbb{R}$ ($i = 1, \ldots, k$), $\psi_{k,k} \neq 0$, and ε_ts are independent Gaussian noises with zero mean and variance σ^2. Adopting quadratic loss, the parameters $\psi_{k,1}, \ldots, \psi_{k,k}$ can be estimated by the method of least squares. When the data-generating model is unknown, one critical problem is the identification of the (unknown) order of the AR model. We need to first estimate parameters with different orders $1, \ldots, d_n$ and then select one of them based on a certain principle.

Experiment

In this experiment, we first generate time series data using each of the following three true data-generating processes, with the sample sizes $n = 100, 500, 2{,}000, 3{,}000$. We then fit the data using the model class in Example 1, with maximal order $d_n = 15$.

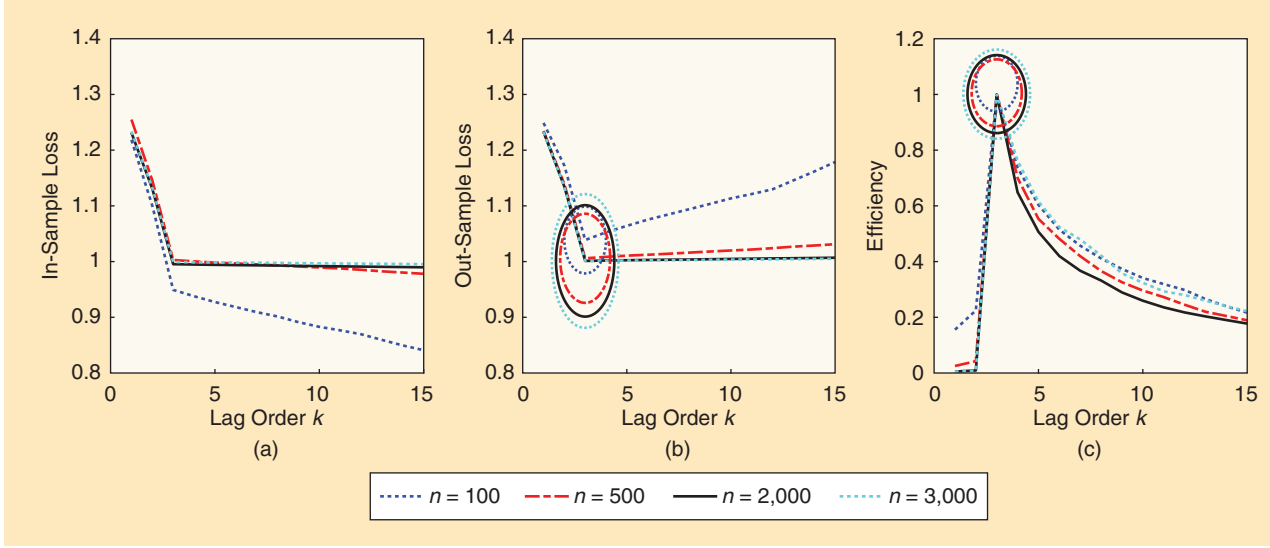


FIGURE 1. The parametric framework: the best predictive performance is achieved at the true order three. (a) The in-sample loss for each sample size n monotonically decreases as the order (model complexity) increases. (b) The predictive performance is only optimal at the true order (circled). (c) The most efficient model (circled) is therefore the true model.

1) *Parametric framework*: The data are generated in the way described by (6) with true order $k_0 = 3$ and parameters $\psi_{3,\ell} = 0.7^{\ell}$ $(\ell = 1, 2, 3)$.

Suppose that we adopt the quadratic loss in Example 1. Then we obtain the average in-sample loss

$$\hat{e}_k = (n-k)^{-1} \sum_{t=k+1}^{n} \left(z_t - \sum_{i=1}^{k} \hat{\psi}_{k,i} z_{t-i}\right)^2.$$

In Figure 1(a), we plot $\hat{e}_k$ against k for $k = 1, \ldots, d_n$, averaged over 50 independent replications. The curve for each sample size n is monotonically decreasing, because larger models fit the same data better. We compute and plot in Figure 1(b) the out-sample prediction loss in (4), which is equivalent to $E_*(s(\hat{p}_k, Z_t)) = E_*\left(Z_t - \sum_{i=1}^{k} \hat{\psi}_{k,i} Z_{t-i}\right)^2$ in this example. The above expectation is taken over the true stationary distribution of an independent process of Z_t. (An alternative definition is based on the same-realization expectation that calculates the loss of the future of an observed time series [10].) The curves in Figure 1(b) show that the predictive performance is only optimal at the true order.

Under the quadratic loss, we have $E_*(s(p_*, Z_t)) = \sigma^2$, and the asymptotic efficiency (Definition 2) requires that

$$\frac{\min_{k=1,\ldots,d_n} E_*\left(Z_t - \sum_{i=1}^{k} \hat{\psi}_{k,i} Z_{t-i}\right)^2 - \sigma^2}{E_*\left(Z_t - \sum_{j=1}^{\hat{k}} \hat{\psi}_{\hat{k},j} Z_{t-j}\right)^2 - \sigma^2} \tag{7}$$

converges to one in probability. To describe how the predictive performance of each model deviates from the best possible, we define the efficiency of each model of order k' to be the quantity in (7) with $\hat{k}$ being replaced with k' $(k' = 1, \ldots, d_n)$. Note that the concepts of efficiency and asymptotic efficiency in model selection are reminiscent of their counterparts in parameter estimation. We plot the efficiency of each candidate model in Figure 1(c). Similarly to Figure 1(b), the curves here show that the true model is the most efficient model. We note that the minus-σ^2 adjustment of out-sample prediction loss in the above definition makes the property highly nontrivial to achieve (see, e.g., [11]–[13]). Consider, e.g., the comparison between AR(2) and AR(3) models, with the AR(2) being the true data-generating model. It can be proved that without subtracting σ^2, the ratio (of the mean square prediction errors) for each of the two candidate models approaches one; by subtracting σ^2, the ratio for AR(2) still approaches one, whereas the ratio for AR(3) approaches 2/3.

2) *Nonparametric framework*: The data are generated by the moving average (MA) model $z_t = \varepsilon_t - 0.8\varepsilon_{t-1}$, with ε_t being independent standard Gaussian.

Similarly to case 1, we plot the results in Figure 2. Different from case 1, the predictive performance is optimal at increasing model dimensions as n increases. In such a nonparametric framework, the best model is sensitive to the sample size, so that pursuing an inference of a fixed good model becomes unrealistic. The model selection task aims to select a model that is asymptotically efficient [see Figure 2(c)]. Note that Figure 2(b) and (c) is drawn based on the information of the underlying true model, which is unavailable in practice; hence, we need a model selection method to achieve the asymptotic efficiency.

3) *Practically nonparametric framework*: The data are generated in the same way as in case 1, except that $k_0 = 10$.

We plot the results in Figure 3. For $n = 2{,}000,\ 3{,}000$, the sample sizes are large enough to support the evidence of a true model with a relatively small model dimension. Similarly to experiment 1, this is a parametric framework in which the optimal predictive performance is achieved at the true model. For $n = 100,\ 500$, where the sample sizes are not large enough

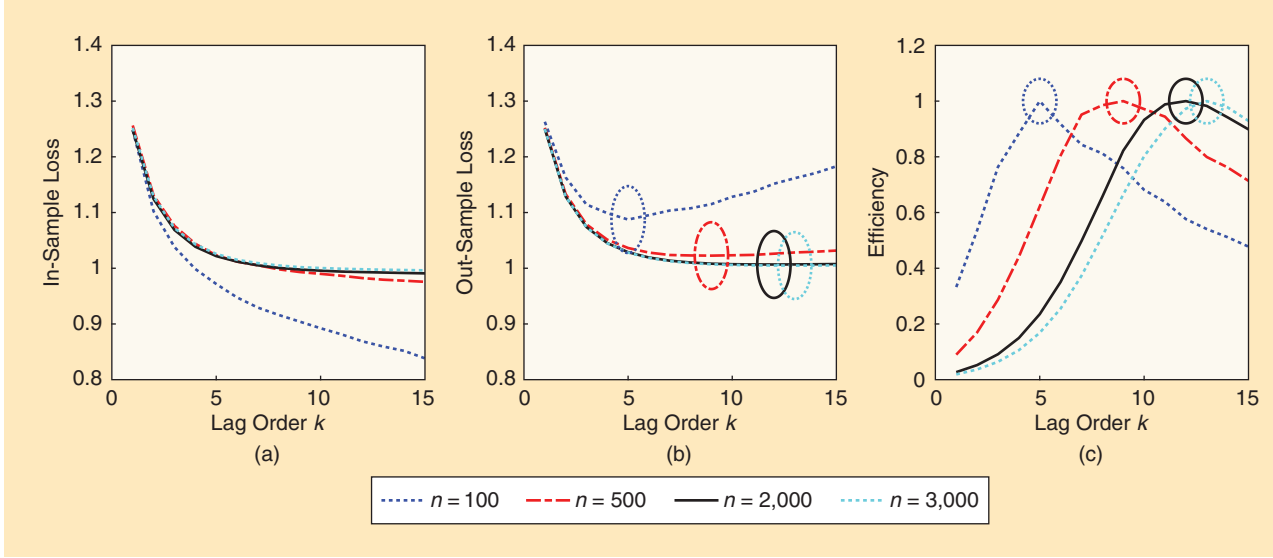


FIGURE 2. The nonparametric framework: the best predictive performance is achieved at an order that depends on the sample size. (a) The in-sample loss for each sample size n monotonically decreases as the order (model complexity) increases. (b) The predictive performance is optimal at increasing orders (circled) as n increases. (c) The order of the most efficient model (circled) therefore increases as n increases.

compared to the true model dimension, however, fitting too many parameters actually causes an increased variance that diminishes the predictive power. In such a scenario, even though the true model is included as a candidate, the best model is not the true model, and it is unstable for small or moderate sample sizes as if in a nonparametric setting. In other words, the parametric framework can turn into a practically nonparametric framework in the small data regime. It can also work the other way around, i.e., for a true nonparametric framework, for a large range of sample sizes (e.g., 100–2,000), a relatively small parametric model among the candidates continues to be the best model [14].

Principles and approaches from various philosophies or motivations

A wide variety of model selection methods have been proposed in the past few decades, motivated by different viewpoints and justified under various circumstances. Many of them originally aimed to select either the order in an AR model or a subset of variables in a regression model. We review some of the representative approaches in these contexts in this section.

Information criteria based on likelihood functions

Information criteria generally refer to model selection methods that are based on likelihood functions and applicable to parametric model-based problems. Here we introduce some information criteria whose asymptotic performances are well understood.

Akaike information criterion (AIC) is a model selection principle proposed by Akaike [15]. A detailed derivation of it from an information theoretic perspective can be found in [1]. Briefly speaking, the idea is to approximate the out-sample prediction loss by the sum of the in-sample loss and a correction term. We refer to [1] for a detailed derivation of this correction term. In the typical setting where the loss is logarithmic, the AIC procedure is to select the model $\mathcal{M}_m$ that minimizes

$$\mathrm{AIC}_m = -2\hat{\ell}_{n,m} + 2d_m, \tag{8}$$

where $\hat{\ell}_{n,m}$ is the maximized log likelihood of model $\mathcal{M}_m$ given n observations as defined in (1), and d_m is the dimension of model $\mathcal{M}_m$. It is clear that more complex models (with larger d_m) will suffer from larger penalties.

In the task of AR order selection, it is also common to use

$$\mathrm{AIC}_k = n \log \hat{e}_k + 2k \tag{9}$$

for the model of order k, where $\hat{e}_k$ is the average in-sample loss based on the quadratic loss. In fact, (9) can be derived from (8) by assuming that AR noises are Gaussian and by regarding ARs of different orders as $\{\mathcal{M}_m\}_{m\in\mathbb{M}}$. A predecessor of AIC is the final prediction error criterion (FPE) [16] (also by Akaike). An extension of AIC is the Takeuchi's information criterion [17], derived in a way that allows model misspecification, but it is rarely used in practice due to its computational complexity. In the context of generalized estimating equations for correlated response data, a variant of AIC based on quasi-likelihood is derived in [18].

Finite-sample corrected AIC (AICc) [19] was proposed as a corrected version of the AIC for small-sample study. It selects the model that minimizes

$$\mathrm{AICc}_m = \mathrm{AIC}_m + \frac{2\{d_m+1\}\{d_m+2\}}{n-d_m-2}.$$

Unless the sample size n is small compared with model dimension d_m, there is little difference between AICc and

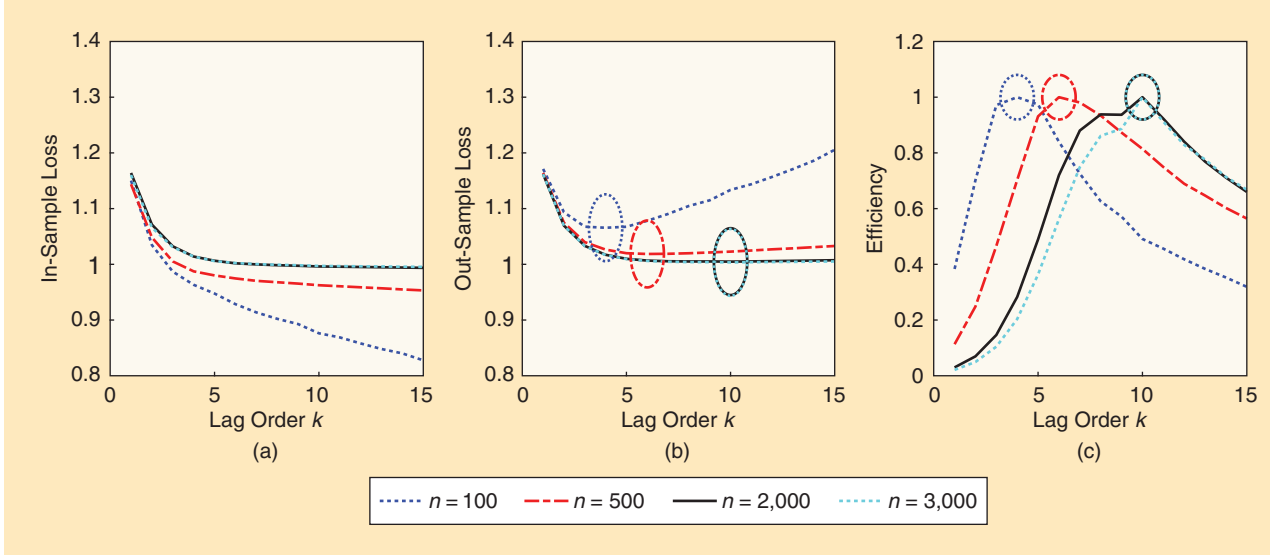


FIGURE 3. The practically nonparametric framework: the best predictive performance is achieved at an order that depends on the sample size in the small data regime. (a) The in-sample loss for each sample size n monotonically decreases as the order (model complexity) increases. (b) The predictive performance is optimal at increasing orders (circled) as n increases in a certain range. (c) The order of the most efficient model (circled) therefore depends on n in a certain range.

AIC. Another modified AIC that replaces the constant two with a different positive number has also been studied in [20].

Bayesian information criterion (BIC) [21] is another popular model selection principle. It selects the model m that minimizes

$$\mathrm{BIC}_m = -2\hat{\ell}_{n,m} + d_m \log n. \tag{10}$$

The only difference with AIC is that the constant two in the penalty term is replaced with the logarithm of the sample size. The original derivation of BIC by Schwarz turned out to have a nice Bayesian interpretation, as its current name suggests.

To see the interpretation, we assume that $z_1, \ldots, z_n$ are the realizations of independent, identically distributed random variables, and $\pi(\cdot)$ is any prior distribution on θ that has dimension d. We let $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(z_i)$ be the log-likelihood function and $\hat{\theta}_n$ the MLE of θ. Based on classical Bayesian asymptotics we have (see, e.g., [22, eq. (1.5)]) under regularity conditions

$$\begin{aligned}\int_{\mathbb{R}^d} \pi(\hat{\theta}_n + n^{-\frac{1}{2}} r) \exp(\ell_n(\hat{\theta}_n + n^{-\frac{1}{2}} r) - \ell_n(\hat{\theta}_n)) dr \\ \to {}_p \pi(\theta_*)(2\pi)^{d/2} \{\det E_*(-\nabla^2_\theta \log p_{\theta_*}(z))\}^{-1/2}\end{aligned} \tag{11}$$

as $n \to \infty$, for some constant θ_*. Note that the right-hand side of (11) is a constant that does not depend on n and the left-hand side of (11) equals

$$p(z_1, \ldots, z_n) \exp(-\ell_n(\hat{\theta}_n) + \frac{d}{2} \log n). \tag{12}$$

Therefore, selecting a model with the largest marginal likelihood $p(z_1, \ldots, z_n)$ (as advocated by Bayesian model comparison) is asymptotically equivalent to selecting a model with the smallest BIC in (10). It is interesting to see that the marginal likelihood of a model does not depend on the imposed prior at all in the large sample limit. Intuitively speaking, this is because, in the integration of $p(z_1, \ldots, z_n) = \int_\theta \pi(\theta) p_\theta(z_1, \ldots, z_n) d\theta$, the mass is concentrated around $\hat{\theta}_n$ with radius $O(n^{-1/2})$ and dimension d, so its value is proportional to the maximized likelihood value multiplied by the volume approximately at the order of $n^{-d/2}$, which is in line with (12).

Hannan and Quinn (HQ) criterion [23] was proposed as an information criterion that achieves strong consistency in AR order selection. In other words, if the data are truly generated by an AR model of fixed order k_0, then the selected order k converges almost surely to k_0 as the sample size goes to infinity. We note that strong consistency implies (the usual) consistency. In general, this method selects a model by minimizing $\mathrm{HQ}_m = -2\hat{\ell}_{n,m} + 2c\, d_m \log\log n$ (for any constant $c > 1$). It can be proved under some conditions that any penalty no larger than $2d_m \log\log n$ is not strongly consistent [23]; therefore, HQ employs the smallest possible penalty to guarantee strong consistency.

Bridge criterion (BC) [24], [25] is a recently proposed information criterion that aims to bridge the advantages of both AIC and BIC in the asymptotic regime. It selects the model $\mathcal{M}_m$ that minimizes $\mathrm{BC}_m = -2\hat{\ell}_{n,m} + c_n(1 + 2^{-1} + \cdots + d_m^{-1})$ (with the suggested $c_n = n^{2/3}$) over all of the candidate models whose dimensions are no larger than $d_{m_{\mathrm{AIC}}}$, the dimension of the model selected by AIC. Note that the penalty is approximately $c_n \log d_m$, but it is written as a harmonic number to highlight some of its nice interpretations. Its original derivation was motivated by a recent finding that the information loss of underfitting a model of dimension d using dimension $d-1$ is asymptotically χ_1^2/d for large d, assuming that nature generates the model from a noninformative uniform distribution over its model space (in particular the coefficient space of all stationary autoregressions) [24, Appendix A]. BC was proved to perform similarly to AIC in a nonparametric framework and similarly to BIC in a parametric framework. We further discuss BC in the "War and Peace—Conflicts Between AIC and BIC and Their Integration" section.

Methods from other perspectives

In addition to information criteria, some other model selection approaches have been motivated from either Bayesian, information-theoretic, or decision-theoretic perspectives.

Bayesian posterior probability is commonly used in Bayesian data analysis. Suppose that each model $m \in \mathbb{M}$ is assigned a prior probability $p(\mathcal{M}_m) > 0$ (such that $\sum_{m \in \mathbb{M}}(\mathcal{M}_m) = 1$), interpreted as the probability that model $\mathcal{M}_m$ contains the true data-generating distribution p_*. Such a prior may be obtained from scientific reasoning or knowledge from historical data. For each $m \in \mathbb{M}$, we also introduce a prior, with density $\theta_m \mapsto p_m(\theta_m)$ $(\theta_m \in \mathcal{H}_m)$, and a likelihood of data $p_m(z \mid \theta_m)$, where $z = [z_1, \ldots, z_n]$. A joint distribution on $(z, \theta_m, \mathcal{M}_m)$ is therefore well defined, based on which various quantities of interest can be calculated. We first define the marginal likelihood of model $\mathcal{M}_m$ by

$$p(z \mid \mathcal{M}_m) = \int_{\mathcal{H}_m} p_m(z \mid \theta_m) p_m(\theta_m) d\theta_m. \tag{13}$$

Based on (13), we obtain the following posterior probabilities on models by Bayes formula:

$$p(\mathcal{M}_m \mid z) = \frac{p(z \mid \mathcal{M}_m) p(\mathcal{M}_m)}{\sum_{m' \in \mathbb{M}} p(z \mid \mathcal{M}_{m'}) p(\mathcal{M}_{m'})}. \tag{14}$$

The maximum a posteriori approach [26] would select the model with the largest posterior probability.

Bayes factors are also popularly adopted for Bayesian model comparison, defined for a pair of models $(\mathcal{M}_m, \mathcal{M}_{m'})$ by

$$B_{m,m'} = \frac{p(\mathcal{M}_m \mid z)}{p(\mathcal{M}_{m'} \mid z)} \Big/ \frac{p(\mathcal{M}_m)}{p(\mathcal{M}_{m'})} = \frac{p(z \mid \mathcal{M}_m)}{p(z \mid \mathcal{M}_{m'})}.$$

The model $\mathcal{M}_m$ is favored over $\mathcal{M}_{m'}$ if $B_{m,m'} > 1$. Bayes factors remove the impact of prior probabilities on the models from the selection process to focus on the ratio of marginal likelihoods. Compared with the Bayesian posterior probability, Bayes factors are appealing when it is difficult to formulate prior probabilities on models.

Bayesian marginal likelihood, defined in (13), also referred to as the *evidence* or *model evidence*, is a quantity naturally motivated by Bayes factors. In the presence of multiple models, the one with the largest Bayesian marginal likelihood is favored over all other models in terms of the Bayes factor. Moreover, it can be seen that the model with the highest marginal likelihood is the model with the highest posterior probability given that the Bayesian prior probabilities on models are all equal. Interestingly, this Bayesian principle using marginal likelihood is asymptotically equivalent to the BIC (as we have introduced in the "Information Criteria Based on Likelihood Function" section). In practice, the preceding Bayesian model selection methods can be computationally challenging. Calculation of the quantities in (13) and (14) are usually implemented using Monte Carlo methods, especially sequential Monte Carlo (for online data) and Markov chain Monte Carlo (for batch data) (see, e.g., [27]). It is worth noting that improper or vague priors on the parameters of any candidate model can have a nonnegligible impact on the interpretability of marginal likelihood and Bayes factors in the nonasymptotic regime, and that has motivated some recent research on Bayesian model selection (see, e.g., [28] and the references therein).

Interestingly, the LOO was shown to be asymptotically equivalent to either AIC/Takeuchi's information criterion under some regularity conditions.

The minimum message length (MML) principle [29] was proposed from an information-theoretic perspective. It favors the model that generates the shortest overall message, which consists of a statement of the model and a statement of the data concisely encoded with that model. Specifically, this criterion aims to select the model that minimizes

$$-\log p(\theta) - \log p(x \mid \theta) + \frac{1}{2}\log \mid I(\theta) \mid + \frac{d}{2}(1 + \log \kappa_d),$$

where $p(\theta)$ is a prior, $p(x \mid \theta)$ is the likelihood function, $I(\theta) = \int \{\partial \log p(x \mid \theta) / \partial \theta\}^2 p(x \mid \theta) dx$ is the Fisher information, d is the dimension of θ, and k_d is the so-called optimal quantizing lattice constant that is usually approximated by $\kappa_1 = 1/12$. A detailed derivation and application of MML can be found in [30].

The minimum description length (MDL) principle [31]–[34] describes the best model as the one that leads to the best compression of a given set of data. It was also motivated by an information-theoretic perspective (which is similar to MML). Different from MML, which is in a fully Bayesian setting, MDL avoids assumptions on prior distribution. Its predictive extension, referred to as the *predictive minimum description length criterion* (*PMDL*), is proposed in [35]. One formulation of the principle is to select the model by minimizing the stochastic complexity $-\log p_{\theta_1}(z_1) - \sum_{t=2}^{n} \log p_{\theta_t}(z_t \mid z_1, \ldots, z_{t-1})$, in which θ_t's are restricted to the same parameter space (with the same dimension). Here, each $\theta_t (t > 1)$ is the MLE calculated using $z_1, \ldots, z_{t-1}$, and $p_{\theta_1}(\cdot)$ can be an arbitrarily chosen prior distribution. The above PMDL criterion is also closely related to the prequential (or predictive sequential) rule [36] from a decision-theoretic perspective.

Deviance information criterion (DIC) [37] was derived as a measure of Bayesian model complexity. Instead of being derived from a frequentist perspective, DIC can be thought of as a Bayesian counterpart of AIC. To define DIC, a relevant concept is the deviance under model m: $D_m(\theta) = -2 \log p_m(y \mid \theta) + C$, where C does not depend on the model being compared. Also, we define the effective number of parameters of the model to be $p_D = E_{\theta \mid z} D_m(\theta) - D_m(E_{\theta \mid z}(\theta))$, where $E_{\theta \mid z}(\cdot)$ is the expectation taken over θ conditional on all of the observed data z under model $\mathcal{M}_m$. Then the DIC selects the model $\mathcal{M}_m$ that minimizes

$$\mathrm{DIC}_m = D_m(E_{\theta \mid z}(\theta)) + 2p_D. \tag{15}$$

The conceptual connection between DIC and AIC can be readily observed from (15). The MLE and model dimension in AIC are replaced with the posterior mean and effective number of parameters, respectively, in DIC. Compared with AIC, DIC enjoys some computational advantage for comparing complex models whose likelihood functions may not even be in analytic forms. In Bayesian settings, Markov chain Monte Carlo tools can be utilized to simulate posterior distributions of each candidate model, which can be further used to efficiently compute DIC in (15).

Methods that do not require parametric assumptions

Cross validation (CV) [38], [39] is a class of model selection methods widely used in machine-learning practice. CV does not require the candidate models to be parametric, and it works as long as the data are permutable and one can assess the predictive performance based on some measure. A specific type of CV is the delete-1 CV method [40] [or leave-one-out (LOO)]. The idea is as follows. For brevity, let us consider a parametric model class as before. Recall that we wish to select a model $\mathcal{M}_m$ with as small out-sample loss $E_*(s(p_{\tilde{\theta}_m}, Z))$ as possible. Its computation involves an unknown true data-generating process, but we may approximate it by $n^{-1} \sum_{i=1}^{n} s(p_{\hat{\theta}_{m,-i}}, z_i)$, where $\hat{\theta}_{m,-i}$ is the MLE under model $\mathcal{M}_m$ using all of the observations except z_i. In other words, given n observations, we leave each one observation out in turn and attempt to predict that data point by using the $n-1$ remaining observations, and we record the average prediction loss over n rounds. Interestingly, the LOO was shown to be asymptotically equivalent to either AIC/Takeuchi's information criterion under some regularity conditions [40].

In general, CV works in the following way. It first randomly splits the original data into a training set of n_t data $1 \le n_t \le n-1$ and a validation set of $n_v = n - n_t$ data; each candidate model is then trained from the n_t data and validated on the remaining data (i.e., to record the average validation loss). This procedure is independently replicated a few times (each with a different validation set) to reduce the variability caused by splitting. Finally, the model with the

smallest average validation loss is selected, and it is retrained using the complete data for future prediction.

A special type of CV is the so-called k-fold CV (with k being a positive integer). It randomly partitions data into k subsets of (approximately) equal size; each model is trained on $k-1$ folds and validated on the remaining one fold. The procedure is repeated k times, and the model with the smallest average validation loss is selected. The k-fold CV is perhaps more commonly used than LOO, partly due to the large computational complexity involved in LOO. The holdout method, as often used in data competitions (e.g., Kaggle competition), is also a special case of CV. It does data splitting only once, one part as the training set and the remaining part as the validation set. We note that there exist fast methods, such as generalized cross validation (GCV) [90], as surrogates to LOO to reduce the computational cost. Some additional discussion on CV will be provided in the "Clarification of Some Misconceptions" section.

Methods proposed for specific types of applications

There have been some other criteria proposed for specific types of applications, mostly for time series or linear regression models.

The predictive least-squares (PLS) principle proposed by Rissanen [41] is a model selection criterion based on his PMDL principle. PLS aims to select the stochastic regression model by minimizing the accumulated squares of prediction errors (in a time-series setting), defined as

$$\mathrm{PLS}_m = \sum_{t=t_0+1}^{n} (y_t - x_{m,t}^T \beta_{m,t-1})^2,$$

where y_t is each response variable, $x_{m,t}$ is the vector of covariates corresponding to model m, and $\beta_{m,t-1}$ is the least-squares estimate of model $\mathcal{M}_m$ based on data before time t. The time index t_0 is the first index such that β_t is uniquely defined. Conceptually, PLS is not like AIC and BIC, which select the model that minimizes a loss plus a penalty. It seems more like the counterpart of LOO in sequential contexts. Interestingly, it has been proved that PLS and BIC are asymptotically close, both strongly consistent in selecting the data-generating model (in a parametric framework) [42]. Extensions of PLS where the first index t_0 is a chosen sequence indexed by n have also been studied. It has been shown, e.g., that PLS with $t_0/n \to 1$ shares the same asymptotic property of AIC under some conditions (see, e.g., [43, Example 8]).

Generalized information criterion (GIC_{λ_n}) [12], [44] represents a wide class of criteria whose penalties are linear in model dimension. It aims to select the regression model $\mathcal{M}_m$ that minimizes

$$\mathrm{GIC}_{\lambda_n,m} = \hat{e}_m + \frac{\lambda_n \hat{\sigma}_n^2 d_m}{n}.$$

Here, $\hat{\sigma}_n^2$ is an estimator of σ^2, the variance of the noise, and $\hat{e}_m = n^{-1} \| y - \hat{y}_m \|_2^2$ is the mean square error between the observations and least-squares estimates under regression model $\mathcal{M}_m$. λ_n is a deterministic sequence of n that controls the tradeoff between the model fitting and model complexity. If we replace $\hat{\sigma}_n^2$ with $(n-d_m)^{-1} n\hat{e}_m$, it can be shown under mild conditions that minimizing GIC_{λ_n} is equivalent to minimizing [12, p. 232]

$$\log \hat{e}_m + \frac{\lambda_n d_m}{n}. \tag{16}$$

In this case, $\lambda_n = 2$ corresponds to AIC, and $\lambda_n = \log n$ corresponds to BIC. Mallows's C_p method [45] is a special case of GIC with $\hat{\sigma}_n^2 \triangleq (n-d_{\bar{m}})^{-1} n\hat{e}_{\bar{m}}$ and $\lambda_n = 2$, where $\bar{m}$ indexes the largest model that includes all of the covariates.

Theoretical properties of the model selection criteria

Theoretical examinations of model selection criteria have centered on several properties: consistency in selection, asymptotic efficiency, and minimax-rate optimality. Selection consistency targets the goal of identifying the best model or method on its own for scientific understanding, statistical inference, insight, or interpretation. Asymptotic efficiency and minimax-rate optimality (defined in Definition 3, which follows) are in tune with the goal of prediction. Before we introduce the theoretical properties, it is worth mentioning that many model selection methods can also be categorized into two classes according to their large-sample performances, represented by AIC and BIC. In fact, it has been known that AICc, FPE, and GCV are asymptotically close to AIC, whereas Bayes factors, HQ, and the original PLS are asymptotically close to BIC. For some other methods, such as CV and GIC, their asymptotic behavior usually depends on the tuning parameters. GIC_{λ_n} is asymptotically equivalent to AIC when $\lambda_n = 2$ and to BIC when $\lambda_n = \log n$. In general, any sequence of λ_n satisfying $\lambda_n \to \infty$ would exhibit the consistency property shared by BIC. As a corollary, the C_p method (as a special case of GIC_2) is asymptotically equivalent to AIC. For CV with n_t training data and n_v validation data, it is asymptotically similar to AIC when $n_v/n_t \to 0$ (including the LOO as a special case) and to BIC when $n_v/n_t \to \infty$ [12, eq. (4.5)].

In general, AIC and BIC have served as the golden rules for model selection in statistical theory during their existence. Though cross validations or Bayesian procedures have also been widely used, their asymptotic justifications are still rooted in frequentist approaches in the form of AIC, BIC, and so forth. Therefore, understanding the asymptotic behavior of AIC and BIC is crucial in both theory and practice. We thus focus on the properties of AIC and BIC in the rest of this section and the "War and Peace—Conflicts Between AIC and BIC and Their Integration" section. It is remarkable that the asymptotic watershed of AIC and BIC (and their closely related methods) simply lies in whether the penalty is a fixed, well-chosen constant or goes to infinity as a function of n.

First of all, AIC is proved to be minimax-rate optimal for a range of variable selection tasks, including the usual subset selection and order selection problems in linear regression and nonparametric regression based on series expansion with such bases as polynomials, splines, or wavelets (see, e.g., [46] and the references therein). Consider, e.g., the minimax risk of estimating the regression function $f \in \mathcal{F}$ under the squared error

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^{n} E_*(\hat{f}(x_i)-f(x_i))^2, \tag{17}$$

where $\hat{f}$ is over all estimators based on the observations and $f(x_i)$ is the expectation of the ith response variable (or the ith value of the regression function) conditional on the ith vector of variables x_i. Each x_i can refer to a vector of explanatory variables, or polynomial basis terms, and so on. For a model selection method ν, its worst-case risk is $\sup_{f\in\mathcal{F}} R(f,\nu,n) = n^{-1}\sum_{i=1}^{n} E_*\{\hat{f}_\nu(x_i)-f(x_i)\}^2$, with $\hat{f}_\nu$ being the least-squares estimate of f under the variables selected by ν.

Definition 3

A method ν is said to be minimax-rate optimal over $\mathcal{F}$ if $\sup_{f\in\mathcal{F}} R(f,\nu,n)$ converges at the same rate as the minimax risk in (17).

Another good property of AIC is that it is asymptotically efficient [as defined in (5)] in a nonparametric framework (see, e.g., [11] and [47]). In other words, the predictive performance of its selected model is asymptotically equivalent to the best offered by the candidate models (even though it is sensitive to the sample size).

BIC, on the other hand, is known to be consistent in selecting the smallest true data-generating model in a parametric framework (see, e.g., [12] and [23]). Suppose, e.g., that the data are truly generated by an AR(2) and the candidate models are AR(2), AR(3), and an MA model that is essentially AR(∞). Then AR(2) is selected with probability going to one as the sample size tends to infinity. MA(1) is not selected because it is a wrong model, and AR(3) is not selected because it overfits [even though it nests AR(2) as its special case]. Moreover, it can be proved that the consistency of BIC also implies that it is asymptotically efficient in a parametric framework [12], [24]. We will elaborate more on the theoretical properties of AIC and BIC in the next section.

War and peace—Conflicts between AIC and BIC and their integration

In this section, we review some research advances in the understanding of AIC, BIC, and related criteria. The choice of AIC and BIC to focus on here because they represent two cornerstones of model selection principles and theories. We are only concerned with the settings where the sample size is larger than the model dimension. Details of the following discussions can be found in such original papers as [11], [12], [24], [47]–[49], and the references therein.

Recall that AIC is asymptotically efficient for the nonparametric framework and is also minimax optimal [46]. In contrast, BIC is consistent and asymptotically efficient for the parametric framework. Despite the good properties of AIC and BIC, they have their own drawbacks. AIC is known to be inconsistent in a parametric framework where there are at least two correct candidate models. As a result, AIC is not asymptotically efficient in such a framework. If data are truly generated by an AR(2), e.g., and the candidate models are AR(2), AR(3), and so forth, then AR(2) cannot be selected with probability going to one by AIC as the sample size increases. The asymptotic probability of it being selected can actually be analytically computed [48]. BIC, on the other hand, does not enjoy the properties of minimax-rate optimality and asymptotic efficiency in a nonparametric framework [12], [50].

Why do AIC and BIC work in those ways? Theoretical arguments in those aspects are highly nontrivial and have motivated a vast literature since the formulations of AIC and BIC. Here we provide some heuristic explanations. For AIC, its formulation in (8) was originally motivated by searching the candidate model p that is the closest in Kullback–Leibler (KL) divergence (denoted by D_{KL}) from p to the data-generating model p_*. Because $\min_p D_{KL}(p_*, p)$ is equivalent to $\min_p E_*(-\log p)$ for a fixed p_*, AIC is expected to perform well in minimizing the prediction loss. But AIC is not consistent for a model class containing a true model and at least one oversized model, because fitting the oversized model would only reduce the first term $-2\hat{\ell}_{n,m}$ in (8) by a random quantity that is approximately chi-square distributed (by, e.g., Wilks's theorem [51]), whereas the increased penalty on the second item $2d_m$ is at a constant level, which is not large enough to suppress the overfitting gain in fitness with high probability. Selection consistency of BIC in a parametric framework is not surprising due to its nice Bayesian interpretation (see the "Principles and Approaches from Various Philosophies or Motivations" section). However, its penalty $d_m \log n$ in (10) is much larger than the $2d_m$ in AIC, so it cannot enjoy the predictive optimality in a typical nonparametric framework (if AIC already does so).

There has been a debate between AIC and BIC in model selection practice, centering on whether the data-generating process is in a parametric framework or not.

To briefly summarize, for asymptotic efficiency, AIC (respectively, BIC) is only suitable in nonparametric (respectively, parametric) settings. Figure 4 illustrates the two situations. There has been a debate between AIC and BIC in model selection practice, centering on whether the data-generating process is in a parametric framework or not. The same debate was sometimes raised under other terminology. In a parametric (respectively, nonparametric) framework, the true data-generating model is often said to be well specified (respectively, misspecified) or finite (respectively, infinite) dimensional. (To see a reason for such terminology, consider, e.g., the regression analysis using polynomial basis function as covariates. If the true regression function is indeed a polynomial, then it can be parameterized with a finite number of parameters; if it is an exponential function, then it cannot be parameterized with any finite dimensional parameter.) Without prior knowledge on how the observations were generated, determining which method to use becomes very challenging. It naturally motivates the following fundamental question: Is it possible to have a method that combines the strengths of AIC and BIC?

The combining of strengths can be defined in two ways. First, can the properties of minimax-rate optimality and consistency be shared? Unfortunately, it has been theoretically shown under rather general settings that there exists no model selection method that achieves both optimality and consistency simultaneously [49]. For any model selection procedure to be consistent, i.e., it must behave suboptimally in terms of minimax rate of convergence in the prediction loss. Second, can the properties of asymptotic efficiency and consistency be shared? In contrast to minimax-rate optimality, which allows the true data-generating model to vary, asymptotic efficiency is in a pointwise sense, meaning that the data are already generated by some fixed (unknown) data-generating model. Therefore, the asymptotic efficiency is a requirement from a more optimistic view and thus weaker in some sense than the minimaxity. Recall that consistency in a parametric framework is typically equivalent to asymptotic efficiency [12], [24]. Clearly, if an ideal method can combine asymptotic efficiency and consistency, it achieves asymptotic efficiency in both parametric and nonparametric frameworks. That motivated an active line of recent advances in reconciling the two classes of model selection methods [24], [43], [52].

In particular, the new model selection method BC was recently proposed (see the "Principles and Approaches from Various Philosophies or Motivations" section) to simultaneously achieve consistency in a parametric framework and asymptotic efficiency in both (parametric and nonparametric) frameworks. The key idea of BC is to impose a BIC-like heavy penalty for a range of small models but to alleviate the penalty for larger models if more evidence is supporting an infinite dimensional true model. In that way, the selection procedure is automatically adaptive to the appropriate setting (either parametric or nonparametric). A detailed statistical interpretation of how BC works in both theory and practice and how it relates to AIC and BIC is elaborated in [24].

Moreover, in many applications, data analysts would like to quantify to what extent the framework under consideration can be practically treated as parametric, or, in other words, how likely the postulated model class is well specified. This motivated the concept of the parametricness index (PI) [14], [24], which assigns a confidence score to model selection. One definition of PI, which we shall use in the following experiment, is this quantity on $[0,1]$:

$$\mathrm{PI}_n = \left| d_{m_{\mathrm{BC}}} - d_{m_{\mathrm{AIC}}} \right| \Big/ \left(\left| d_{m_{\mathrm{BC}}} - d_{m_{\mathrm{AIC}}} \right| + \left| d_{m_{\mathrm{BC}}} - d_{m_{\mathrm{BIC}}} \right| \right)$$

if the denominator is not zero, and $\mathrm{PI}_n = 1$ otherwise. Here, d_{m_ν} is the dimension of the model selected by the method ν. Under some conditions, it can be proved that $\mathrm{PI}_n \to_p 1$ in a parametric framework and $\mathrm{PI}_n \to_p 0$ otherwise.

Experiments

We now revisit Example 1 in the "An Illustration on Fitting and the Best Model" section and numerically demonstrate the performances of different methods based on 100 replications and $n = 500$. For each of the three cases, we compute the means and standard errors of the efficiency [defined in (7)], dimension of the selected model, and PI and summarize them in Table 1. In case 1, BIC and BC perform much better than AIC in terms of efficiency, and PI is close to 1. This is expected from theory, as we are in a parametric setting. In cases 2 and 3, which are (practically) nonparametric, BC performs similarly to AIC, much better than BIC, and PI is closer to zero.

In practice, AIC seems more widely used compared with BIC, perhaps mainly due to the thinking that all models are wrong and minimax-rate optimality of AIC offers more robustness in adversarial settings than BIC. Nevertheless, the parametric setting is still of vital importance. First of all, being consistent in selecting the true model if it is really among the candidates is certainly mathematically appealing, and a nonparametric

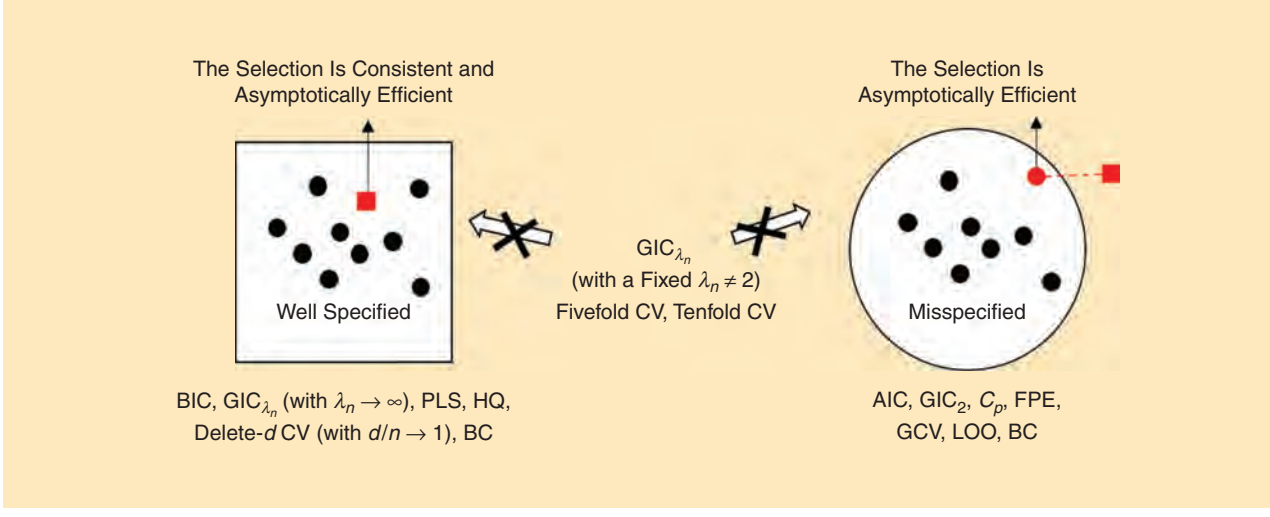


FIGURE 4. A graph illustrating a parametric setting where the model class (by large square) includes the true data-generating model (by small red square) and a nonparametric setting where the model class (by large circle) excludes the true data-generating model, along with an asymptotically efficient model (by red circle) in the second case. It also lists some popular methods suitable for either situation and a class of GIC and CV that are asymptotically suboptimal for regression models.

framework can be a practically parametric framework. More importantly, when decisions need to be made on the use of certain variables, the concept of consistency that avoids over-selection of variables is practically very important. If medical researchers need to decide if certain genes should be further studied in costly experiments, e.g., the protection of overfitting of BIC avoids recommending variables that are hard to be justified statistically in a follow-up study, whereas AIC may recommend quite a few variables that may have some limited predictive values but their effects are too small to be certain with the limited information in the data for decision making and inference purposes.

The war between AIC and BIC originates from two fundamentally different goals: one to minimize certain loss for prediction purpose and the other to select the best model for inference purpose. A unified view on reconciling two such different goals wherever possible is a fundamental issue in model selection, and it remains an active line of research. We have witnessed some recent advances in that direction, and we expect more discoveries to flourish in the future.

High-dimensional variable selection

The methods introduced in the "Principles and Approaches from Various Philosophies or Motivations" section were designed for small models, where the dimension d_n is often required to be $o(\sqrt{n})$ in technical proofs. In this section, we elaborate on high-dimensional regression variable selection, an important type of model selection problems in which d_n can be comparable with or even much larger than n. To alleviate the difficulties, the data-generating model is often assumed to be a well-specified linear model, i.e., one of the following candidate models.

Each model $\mathcal{M}$ assumes that $y = \Sigma_{i\in\mathcal{M}}\beta_i x_i + \varepsilon$, with ε being random noises. Here, with a slight abuse of notation, we have also used $\mathcal{M}$ to denote a subset of $\{1, \ldots, d_n\}$, and each data point is written as $z = [y, x_1, \ldots, x_{d_n}]$, with y being the observed response and x_i being the (either fixed or random) covariates. Here, d_n instead of d is used to highlight that the number of candidate variables may depend on the sample size n.

The variable selection problem is also known as *support recovery* or *feature selection* in different literature. The mainstream idea to select the subset of variables is to either solve a penalized regression problem or iteratively pick up significant variables. The proposed methods differ from each other in terms of how they incorporate unique domain knowledge (e.g., sparsity, multicollinearity, group behavior) or what desired properties (e.g., consistency in coefficient estimation, consistency in variable selection) to achieve. The list of methods we will introduce is far from complete. Wavelet shrinkage, iterative thresholding, Dantzig selector, ℓ_q-regularization with $q \in (0, 1)$ (see, e.g., [53]–[57]), e.g., will not be covered.

Penalized regression for variable selection

In a classical setting, a model class is first prescribed to data analysts (either from scientific reasoning or from exhaustive search over d_n candidate variables), and then a criterion is used to select the final model (by applying any properly chosen method explained in the "Principles and Approaches from Various Philosophies or Motivations" section). When there is no ordering of variables known in advance and the number of variables d_n is small, one may simply search over 2^{d_n} possible subsets and perform model selection. But it is usually computationally prohibitive to enumerate all possible subsets for large d_n, especially when d_n is comparable with or even larger than the sample size n. Note also that the problem of obtaining a sparse representation of signal y through some chosen basis x_i (say polynomial, spline, or wavelet basis) usually falls under the framework of variable subset selection as well (but with a different motivation). Such a representation can be practically useful in, e.g., compressing image signals, locating radar sources, or understanding principal components.

Suppose that we have response Y_n and design matrix X_n whose entries are n observations of $[y, x_1, \ldots, x_{d_n}]$. For high-dimensional regression, a popular solution is to consider the following penalized regression that amalgamates variable selection and prediction simultaneously in operation. Solve

$$\hat{\beta} = \arg\min_{\beta}\left\{\| Y_n - X_n\beta \|_2^2 + \sum_{j=1}^{d_n} p(|\beta_j|; \lambda, \gamma)\right\}, \tag{18}$$

and let $\{i : \hat{\beta}_i \neq 0\}$ be the selected subset of variables. Here, the $p(\beta; \lambda, \gamma)$ is a penalty function of β with tuning parameters λ, γ (which are usually determined by cross validation). It is crucial that the penalty function is not differentiable at $\beta = 0$ so that the resulting solution becomes sparse when λ gets large.

Least absolute shrinkage and selection operator (LASSO) [58] in the form of $p(\beta; \lambda) = \lambda |t|$ is perhaps the most commonly used penalty function. Here, λ is a tuning parameter that controls the strength of the penalty term. Increasing λ leads to fewer variables selected. In practice, data analysts can either 1) numerically sweep over a range of λ or 2) use the least-angle regression method [59] to find all of the possible candidate models (also called the *solution paths*) and then

Table 1. The AR order selection: The average efficiency, dimension, and PI (along with standard errors).

		AIC	BC	BIC
Case 1	Efficiency	0.78 (0.04)	0.93 (0.02)	0.99 (0.01)
	Dimension	3.95 (0.20)	3.29 (0.13)	3.01 (0.01)
	PI		0.93 (0.03)	
Case 2	Efficiency	0.77 (0.02)	0.76 (0.02)	0.56 (0.02)
	Dimension	9.34 (0.25)	9.29 (0.26)	5.39 (0.13)
	PI		0.13 (0.03)	
Case 3	Efficiency	0.71 (0.02)	0.67 (0.02)	0.55 (0.02)
	Size	6.99 (0.23)	6.61 (0.26)	4.02 (0.10)
	PI		0.35 (0.05)	

select the model with the best cross-validation performance. In a time series setting where LASSO solutions need to be continuously updated, fast online algorithms have been proposed (e.g., in [60]). Given that the data are truly generated by a linear model, tight prediction error bounds have been established for LASSO. Though originally designed for linear regression, LASSO has been also extended to a wide range of statistical models, such as generalized linear models (see [61] and the references therein).

Smoothly clipped absolute deviation (SCAD) [62] is another penalized regression that can correct the bias in LASSO estimates that comes from the ℓ_1-penalty function being unbounded. It was also shown to exhibit oracle property, meaning that, as the sample size and model dimension go to infinity, all and only the true variables will be identified with probability going to one, the estimated parameters converge in probability to the true parameters, and the usual asymptotic normality holds as if all of the irrelevant variables have already been excluded. More discussions on such an oracle property will be included in the "Clarification of Some Misconceptions" section. The penalty of SCAD is in the form of

$$p(\beta;\lambda,\gamma)=\begin{cases}\lambda|t| & \text{if } |t|\le\lambda\\ \dfrac{2\gamma\lambda|t|-t^2-\lambda^2}{2(\gamma-1)} & \text{if } \lambda<|t|\le\gamma\lambda\\ \dfrac{\lambda^2(\gamma+1)}{2} & \text{if } |t|>\gamma\lambda\end{cases}.$$

In choosing a parsimonious set of variables, LASSO tends to overshrink the retained variables. In the SCAD penalty, the idea is to let λ and γ jointly control that the penalty first suppresses insignificant variables as LASSO does and then tapers off to achieve bias reduction. The tuning parameters in SCAD can be chosen by sweeping over a range of them and then applying cross validation.

Minimax concave penalty (MCP) [63] in the form of

$$p(\beta;\lambda,\gamma)=\begin{cases}\lambda|t|-\dfrac{t^2}{2\gamma} & \text{if } |t|\le\gamma\lambda\\ \dfrac{\gamma\lambda^2}{2} & \text{if } |t|>\gamma\lambda\end{cases}$$

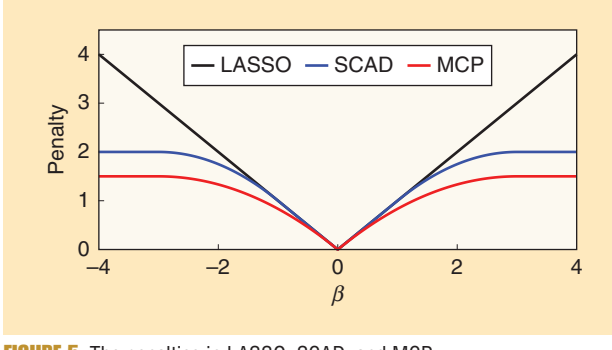


FIGURE 5. The penalties in LASSO, SCAD, and MCP.

is a penalized regression that works in a similar way as SCAD. Under some conditions, MCP attains minimax convergence rates in probability for the estimation of regression coefficients. Figure 5 illustrates the penalties in LASSO, SCAD, and MCP for $\lambda=1$ and $\gamma=3$.

Elastic net [64] in the form of $p(\beta;\lambda)=\lambda_1|t|+\lambda_2 t^2$ is proposed to address several shortcomings of LASSO when the covariates are highly correlated. The solution $\hat{\beta}$ of the elastic net penalty exhibits mixed effects of the LASSO and ridge penalties. Recall that ridge regression in the form of $p(\beta;\lambda)=\lambda t^2$ introduces bias to the regression estimates to reduce the large variances of ordinary least-squares estimates in the case of multicollinearity, and that LASSO tends to select a sparse subset. Interestingly, under elastic net, highly correlated covariates will tend to have similar regression coefficients. This property, distinct from LASSO, is appealing in many applications when data analysts would like to find all of the associated covariates rather than selecting only one from each set of strongly correlated covariates.

A unified view on reconciling two such different goals wherever possible is a fundamental issue in model selection, and it remains an active line of research.

Group LASSO [65] is another penalty introduced to restrict that all of the members of each predefined group of covariates are selected together. Different from (18), the penalty of the regression is not a sum of n terms but is replaced with $\lambda\Sigma_{j=1}^{r}\|\beta_{I_j}\|_2$, where β_{I_j} is a subvector of β indexed by I_j (the jth group), and $I_1,\ldots,I_r$ form a partition of $\{1,\ldots,n\}$. It can be proved that $\hat{\beta}_{I_j}$ is restricted to be vanishing together for each j [65]. The groups are often predefined using prior knowledge.

Adaptive LASSO [66] has been introduced to overcome the inconsistency in variable selection of LASSO. It replaces the penalty in (18) with $\lambda\Sigma_{j=1}^{d_n}|\tilde{\beta}_j|^{-u}|\beta_j|$, where $\tilde{\beta}_j$ is referred to as a *pilot estimate* that can be obtained in various ways (e.g., by least squares for $d_n<n$ or univariate regressions for $d_n\ge n$). Adaptive LASSO was shown to exhibit the aforementioned oracle property. The adaptive LASSO can be solved by the same efficient algorithm for solving the LASSO, and it can be easily extended for generalized linear models as well.

In addition to the preceding penalized regression, a class of alternative solutions is known as *greedy algorithms* (or *stepwise algorithms*), which select a set of variables by making locally optimal decisions in each iteration.

Orthogonal matching pursuit (OMP) [67], [68], also referred to as the *forward stepwise regression algorithm*, is a very popular greedy algorithm that also inspired many other greedy algorithms. The general idea of OMP is to iteratively build a set of variables that are the most relevant to the response. It works in the following way. In each iteration, the variable most correlated with the current residual (in absolute value) is added to the subset (which is initialized as the empty set). Here, the residual represents the component of the observation vector y not in the linear span of the selected variables. Stopping criteria that guarantee good asymptotic properties, such as consistency in variable selection, remain an active line of research.

The OMP algorithm can sequentially identify all of the significant variables with high probability under some conditions, such as weak dependences of the candidate variables (see, e.g., [69] and [70] and the references therein).

Least-angle regression (LARS) [59] is a greedy algorithm for stepwise variable selection. It can also be used for computing the solution paths of LASSO. Different from OMP, it does not permanently maintain a variable once it is selected into the model. Instead, it only adjusts the coefficient of the most correlated variable until that variable is no longer the most correlated with the recent residual. Briefly speaking, LARS works in the following way. It starts with all coefficients β_i being zeros. In each iteration, it looks for the variable x_i most correlated with the current residual r and increases its coefficient β_i in the direction of the sign of its correlation with y. Once some other variable x_j has the same correlation with r as x_i has, it increases β_i and β_j in the direction of their joint least squares until another variable has the same correlation with the residual. The procedure is repeated until all of the variables are in the model or the residuals have become zero.

The war between AIC and BIC originates from two fundamentally different goals: one to minimize certain loss for prediction purpose and the other to select the best model for inference purpose.

Properties of the penalized regression methods

Theoretical examinations of the penalized regression methods have mainly focused on the properties of tight prediction error bounds and consistency in selection. These asymptotic properties are mostly studied by assuming a parametric framework, i.e., data are truly generated by a linear regression model. Analysis for nonparametric, high-dimensional regression models has been also investigated in terms of oracle inequalities for prediction loss [71] and nonlinear additive models [72], [73].

The goal for prediction in high-dimensional regression focuses the control of the prediction loss (usually squared loss) bound so that it eventually vanishes even for a very large number of variables d_n (compared with the sample size n). Suppose that data are generated, e.g., by $Y_n = X_n\beta_* + \varepsilon$, where $Y_n \in \mathbb{R}^n$, $\beta_* \in \mathbb{R}^{d_n}$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $\|\beta_*\|_0$ denote the number of nonzero entries in β_*. Then, under certain restricted eigenvalue assumptions [71], there exist some constants $c_1 > 2\sqrt{2}$ and $c_2 > 0$ such that the LASSO solution satisfies $n^{-1}\|X_n\beta_* - X_n\hat{\beta}\|_2^2 \le c_2\sigma^2\|\beta_*\|_0 n^{-1}\log d_n$ with probability at least $1 - d_n^{1-c_1^2/8}$, if we choose $\lambda = c_1\sigma\sqrt{n\log d_n}$. Note that the choice of λ depends on an unknown $c_1\sigma$ that, though it does not scale with n, can have an effect for small sample size. Notably, the number of variables d_n is allowed to be much larger than n to admit a good predictive performance, as long as $\log d_n$ is small compared with n. Similar tight bounds can be obtained by making other assumptions on β_* and X_n.

Selection consistency, as before, targets the goal of identifying the significant variables for scientific interpretation. The property of asymptotic efficiency we introduced before is rarely considered in high-dimensional regressions, because it is implied by selection consistency in the parametric setting. For any vector $\beta \in \mathbb{R}^{d_n}$, let $r(\beta)$ denote the indicator vector of β such that for any $j = 1, \ldots, d_n$, $r_i(\beta) = 0$ if $\beta_i = 0$, and $r_i(\beta) = 1$ otherwise. Selection consistency requires that the probability of $r(\hat{\beta}) = r(\beta)$ converges in probability to one (as $n \to \infty$). Under various conditions, such as fixed design or random design matrices, consistency of LASSO in estimating the significant variables has been widely studied under such various technical conditions as sparsity, restricted isometry [74], mutual coherence [75], irrepresentable condition [76], and restricted eigenvalue [71], which create theoretical possibilities to distinguish the true subset of variables from all of the remaining subsets for large n.

At the same time, it has been known that LASSO is not generally consistent in parameter/coefficient estimation. This motivates the methods, such as SCAD, MCP, adaptive LASSO, and so forth, that correct the estimation bias of LASSO. These three methods are also known to enjoy the so-called oracle property. The oracle property is perhaps more widely considered than selection consistency for high-dimensional regression analysis, because the penalized regression methods target simultaneous parameter estimation and prediction loss control. An oracle estimator [62] must be consistent in variable selection and parameter estimation, and satisfy 1) the sparsity condition, meaning that $P_*\{r(\hat{\beta}) = r(\beta)\} \to 1$ as $n \to \infty$, where the inequality is componentwise; and 2) the asymptotic normality $\sqrt{n}(\hat{\beta}_S - \beta_S) \to_d \mathcal{N}(0, I^{-1}(\beta_S))$, where S is the support set of β, β_S is the subvector of β_* indexed by S, and $I(\beta_S)$ is the Fisher information knowing S in advance. Intuitively speaking, an oracle estimator enjoys the properties achieved by the MLE knowing the true support. We will revisit the oracle property in the "Controversy over the Oracle Property" section.

Practical performance of penalized regression methods

With the huge influx of high-dimensional regression data, the penalized regression methods have been widely applied for sparse regression where a relatively small (or tiny) number of variables are selected out of a large number of candidates. In applications with a gene expression type of data, e.g., although the number of subjects may be only tens or hundreds, a sparse set of genes is typically selected out of thousands of choices. This has created a lot of excitement, with thousands of publications of such research and applications. This celebrated sparsity feature of penalized regression methods has generated an optimistic view that, even with, e.g., fewer than a hundred observations, the modern variable selection tool can identify a sparse subset out of thousands or even many more variables as the set of the most important ones for the regression problem. The estimated model is often readily used for data-driven discoveries.

There is little doubt that penalized regression methods have produced many successful results for the goal of prediction

(see, e.g., [77]). As long as a proper cross validation is done for tuning parameter selection, the methods can often yield good predictive performance. However, given the challenge of high dimension and diverse data sources, the different penalized regression methods may have drastically different relative performance for various data sets. Therefore, proper choice of a method is important, to which end cross validation may be used, as will be presented in the next section.

For the goal of model selection for inference, however, the picture is much less promising. Indeed, many real applications strongly suggest that the practice of using the selected model for understanding and inference may be far from reliable. It has been reported that the selected variables from these penalized regression methods are often severely unstable, in the sense that the selection results can be drastically different under a tiny perturbation of data (see [78] and the references therein). Such high uncertainty damages reproducibility of the statistical findings [79]. Overall, being overly optimistic about the interpretability of high-dimensional regression methods can lead to spurious scientific discoveries.

The fundamental issue still lies in the potential discrepancy between inference and prediction, which is also elaborated in the "War and Peace—Conflicts Between AIC and BIC and Their Integration" and "Controversy over the Oracle Property" sections. If data analysts know in advance that the true model is exactly (or close to) a stable low-dimensional linear model, then the high-dimensional methods with the aforementioned oracle property may produce stable selection results not only good for prediction but also for inference purposes. Otherwise, the produced selection is so unstable that analysts can only focus on prediction alone. In practice, data analysts may need to utilize data-driven tools, such as model averaging [80], resampling [81], and confidence set for models [82], or model selection diagnostic, such as the parametricness index introduced in the "War and Peace—Conflicts Between AIC and BIC and Their Integration" section, to make sure the selected variables are stable and properly interpretable. Considerations along these lines also lead to stabilized variable selection methods [81], [83], [84]. The instability of penalized regression also motivated some recent research on postselection inference [85], [86]. Their interesting results in specific settings call for more research for more general applications.

Modeling procedure selection

The discussions in the previous sections have focused on model selection in the narrow sense, where the candidates are models. In this section, we review the use of CV as a general tool for modeling procedure selection, which aims to select one from a finite set of modeling procedures [87]. Multiple modeling procedures such as AIC, BIC, and CV could be used for variable selection, and one of those procedures (together with the model selected by the procedure) is selected using an appropriately designed CV (which is at the second level). Another example is the emerging online competition platforms, such as Kaggle, that compare new problem-solving procedures and award prizes using cross validation. The best procedure is defined in the sense that it outperforms, with high probability, the other procedures in terms of out-sample prediction loss for sufficiently large n (see, e.g., [13, Definition 1]).

There are two main goals of modeling procedure selection. The first is to identify with high probability the best procedure among the candidates. The property of selection consistency is of interest here. The second goal of modeling procedure selection is to approach the best performance (in terms of out-sample prediction loss) offered by the candidates, instead of pinpointing which candidate procedure is the best. Note again that, in case there are procedures that have similar best performances, we do not need to single out the best candidate to achieve the asymptotically optimal performance.

Similarly to model selection, for the task of modeling procedure selection, CV randomly splits n data into n_t training data and n_v validation data (so $n = n_t + n_v$). The first n_t data are used to run different modeling procedures, and the remaining n_v data are used to assess the predictive performance. We will see that, for the first goal, the evaluation portion of CV should be large enough. For the second goal, a smaller portion of the evaluation may be enough to achieve optimal predictive performance.

In the literature, much attention has been focused on choosing whether to use the AIC procedure or BIC procedure for data analysis. For regression variable selection, it has been proved that the CV method is consistent in choosing between AIC and BIC given $n_t \to \infty, n_v/n_t \to \infty$, and some other regularity assumptions [87, Th. 1]. In other words, the probability of BIC being selected goes to one in a parametric framework, and the probability of AIC being selected goes to one otherwise. In this way, the modeling procedure selection using CV naturally leads to a hybrid model selection criterion that builds upon strengths of AIC and BIC. Such a hybrid selection combines some theoretical advantages of both AIC and BIC. This aspect is seen in the context of the "War and Peace—Conflicts Between AIC and BIC and Their Integration" section. The task of classification is somewhat more relaxed compared with the task of regression. To achieve consistency in selecting the better classifier, the splitting ratio may be allowed to converge to infinity or any positive constant, depending on the situation [13]. In general, it is safe to let $n_t \to \infty$ and $n_v/n_t \to \infty$ for consistency in modeling procedure selection.

Closely related to this discussion is the following paradox. Suppose that a set of newly available data is given to an analyst. The analyst would naturally add some of the new data in the training phase and some in the validation phase. Clearly, with more data added to the training set, each candidate modeling procedure is improved in accuracy; with more data added to the validation set, the evaluation is also more reliable. It is tempting to think that improving the accuracy on both training and validation would lead to a sharper comparison between procedures. However, this is not the case. The prediction error estimation and procedure comparison are two different targets.

The cross-validation paradox says that better training and better estimation (e.g., in both bias and variance) of the prediction error by CV together do not imply better modeling procedure selection [13], [87]. Intuitively speaking, when comparing two procedures that are naturally close to each other, the improved estimation accuracy achieved by adopting more observations in the training part only makes the procedures more difficult to be distinguishable. The consistency in identifying the better procedure cannot be achieved unless the validation size diverges fast enough.

Experiments

We illustrate the cross-validation paradox using the synthetic data generated from the linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where $\beta = [1, 2, 0]^{\mathrm{T}}$, and the covariates X_j $(j = 1, 2, 3)$ and noise ε are independent standard Gaussian. Given n observations $(y_i, x_{1,i}, x_{2,i}, x_{3,i})_{i=1,\ldots,n}$, we compare the following two different uses of linear regression. The first is based on X_1 and X_2, and the second is based on all three covariates. Note that, in this experiment, selecting the better procedure is equivalent to selecting a better model. The data-generating model indicates that x_3 is irrelevant for predicting y, so that the first procedure should be better than the second. Suppose that we start with 100 observations. We randomly split the data 100 times, each with 20 training data and 80 validation data, and record which procedure gives the smaller average quadratic loss during validation. We then add 50 new data to the training set and 50 to the validation set and record again which procedure is favored. We continuing doing this until the sample size reaches 500. By running 1,000 independent replications, we summarize the frequency of the first procedure being favored in Table 2. As the paradox suggests, the accuracy of identifying the better procedure does not necessarily increase when more observations are added to both the estimation phase and the validation phase.

Table 2. The cross-validation paradox: More observations in training and evaluation do not lead to higher selection accuracy in selecting the better procedure.

Sample size n	100	200	300	400	500
Training size n_t	20	70	120	170	220
Accuracy	98.3%	94.9%	93.7%	92.3%	92.5%

Clarification of some misconceptions

Pitfall of one-size-fits-all recommendation of data splitting ratio of cross validation

There are widespread general recommendations on how to apply cross validation for model selection. It is stated in the literature, e.g., that tenfold CV is the best for model selection. Such guidelines seem to be unwarranted. First, it mistakenly disregards the goal of model selection. For prediction purposes, LOO is actually preferred in tuning parameter selection for traditional nonparametric regression. In contrast, for selection consistency, tenfold often leaves too few observations in evaluation to be stable. Indeed, fivefold often produces more stable selection results for high-dimensional regression. Second, k-fold CV, regardless of k, in general, is often unstable in the sense that a different dividing of data can produce a very different selection result. A common way to improve performance is to randomly divide the data into k folds several times and use the average validation loss for selection.

For model selection, CV randomly splits n data into n_t training data and n_v validation data. Common practices using fivefold, tenfold, or 30% for validation do not exhibit asymptotic optimality (either consistency or asymptotic efficiency) in simple regression models, and their performances can be very different depending on the goal of applying CV. In fact, it is known that the delete-n_v CV is asymptotically equivalent to GIC_{λ_n} with $\lambda_n = n/(n - n_v) + 1$ for linear regression models under some assumptions [12]. It is also known that GIC_{λ_n} achieves asymptotic efficiency in a nonparametric framework only with $\lambda_n = 2$, and asymptotic efficiency in a parametric framework only with $\lambda_n \to \infty$ (as $n \to \infty$). In this context, from a theoretical perspective, the optimal splitting ratio n_v/n_t of CV should either converge to zero or diverge to infinity to achieve asymptotic efficiency, depending on whether the setting is nonparametric or parametric.

For modeling procedure selection, it is often necessary to let the validation size take a large proportion (e.g., half) to achieve good selection accuracy. In particular, the use of LOO for the goal of comparing procedures is the least trustworthy (see the "Modeling Procedure Selection" section).

Experiment

We show how the splitting ratio can affect CV for model selection using the Modified National Institute of Standards and Technology database [88], which consists of 70,000 images of handwritten digits (from 0 to 9) with 28×28 pixels. We implement six candidate feed-forward neural network models for classification. The first four models have one hidden layer, and the number of hidden nodes are 17, 18, 19, and 20. The fifth model has two hidden layers with 20 and four nodes; the sixth model has three hidden layers with 20, two, and two nodes. Because the true data-generating model for the real data is unavailable, we take 35,000 data (often referred to as the *test data*) out for approximating the true prediction loss and use the remaining data to train and validate. For model selection, we run CV with different n_v/n_t. For each ratio, we compute the average validation loss of each candidate model based on ten random partitions. We then select the model with the smallest average loss and calculate its true predictive loss using the remaining 35,000 data. The results recorded in Table 3 indicate that a smaller splitting ratio n_v/n_t leads to better classification accuracy. This is in line with the existing theory, because the neural network modeling is likely to be of nonparametric nature. This example also provides a complementing message to the cross-validation paradox. At ratio 0.95, the training sample size is too small to represent the full

Table 3. The classification for handwritten digits: Smaller tends to give better predictive performance.

Ratio	0.95	0.9	0.5	0.1	0.05
Accuracy	72.24%	90.28%	91.47%	91.47%	92.99%

sample size, so the ranking of the candidate models estimated from training data can be unstable and deviate from the ranking of models estimated from the full data set.

Because all models are wrong, why pursue consistency in selection?

Because the reality is usually more complicated than a parametric model, perhaps everyone agrees that all models are wrong and the consistency concept of selecting the true model in a parametric framework cannot hold in the rigid sense. One view on such selection consistency is that, in many situations, a stable parametric model can be identified, and it can be treated as the true model. Such an idealization for theoretical investigation with practical implications is no more sinful than deriving theories under nonparametric assumptions. The true judge should be the performance in real applications. The notion of consistency in a nonparametric framework, however, is rarely used in the literature. In fact, it was shown that there does not exist any model selection method that can guarantee consistency in nonparametric regression settings (see, e.g., [25]). This partly explains why the concept of asymptotic efficiency (which is a weaker requirement) is more widely used in nonparametric frameworks.

Controversy over the oracle property

The popular oracle property (as mentioned in the "Properties of the Penalized Regression Methods" section) for high-dimensional variable selection has been a focus in many research publications. However, it has been criticized by some researchers (see, e.g., [89]). At first glance, the oracle property may look very stringent. But we note that its requirement is fundamentally only as stringent as consistency in variable selection. In fact, if all of the true variables can be selected with probability tending to one by any method, then one can obtain MLE or the like restricted to the relevant variables for optimal estimation of the unknown parameters in the model. To our knowledge, there is neither claim nor reason to believe that the original estimator should be better than the refitted one by MLE based on the selected model. Though the oracle property is not theoretically surprising beyond consistency, it is still interesting and nontrivial to obtain such a property with only one stage of regression (as SCAD and MCP do). These methods, when armed with efficient algorithms, may save the computational cost in practice.

It was emphasized in [89] that the oracle estimator does not perform well in a uniform sense for point or interval estimation of the parameters. A price paid for the oracle property is that the risk of any oracle estimator (see [62]) has a supremum that diverges to infinity, i.e.,

$$\sup_{\beta \in \mathbb{R}^p} E_\beta \{ n(\hat{\beta} - \beta)^{\mathrm{T}} (\hat{\beta} - \beta) \} \to \infty,$$

as sample size $n \to \infty$ (see, e.g., [89]). Here, we let E_β denote expectation with respect to the true linear model with coefficients β. In fact, for any consistent model selection method, we can always find a parameter value that is small enough so that the selection method tends to not include it (because it has to avoid overselection), yet the parameter value is big enough so that dropping it has a detrimental effect in rate of convergence (see, e.g., [49] and [90]). Although uniformity and robustness are valid and important considerations, we do not need to overly emphasize such properties. Otherwise, we are unduly burdened to retain not very useful variables in the final model and have to lose the ability in choosing a practically satisfying parsimonious model for interpretation and inference.

Some general recommendations

Model selection, no matter how it is done, is exploratory in nature and cannot be confirmatory. Confirmatory conclusions can only be drawn based on well-designed follow-up studies. Nevertheless, good model selection tools can provide valuable and reliable information regarding explanation and prediction. Obviously there are many specific aspects of the data, nature of the models and practical considerations of the variables in the models, and so on that make each model selection problem unique to some degree. In spite of that, based on the literature and our own experiences, we give some general recommendations.

1) Keep in mind the main objective of model selection. First, if one needs to declare a model for inference, model selection consistency is the right concept to think about. Model selection diagnostic measures need to be used to assess the reliability of the selected model. In a high-dimensional setting, penalized regression methods are typically highly uncertain. For selection stability, when choosing a tuning parameter by cross validation, e.g., fivefold tends to work better than tenfold (see, e.g., [78]). Second, if one's main goal is prediction, model selection instability is less of a concern, and any choice among the best performing models may give a satisfying prediction accuracy. In a parametric framework, consistent selection leads to asymptotic efficiency. In a nonparametric framework, selection methods based on the optimal tradeoff between estimation error and approximation error lead to asymptotic efficiency. When it is not clear if a (practically) parametric framework is suitable, we recommend the use of an adaptively asymptotic efficient method (e.g., the BC criterion).
2) When model selection is for prediction, the minimax consideration gives more protection in the worst case. If one postulates that the nature is adversary, the use of a minimax optimal criterion (e.g., AIC) is safer (than, e.g., BIC).
3) When prediction is the goal, one may consider different types of models and methods and then apply cross validation

to choose one for final prediction. If one needs to know which model or method is really the best, a large enough proportion (e.g., one-third or even half) for validation is necessary. If one just cares about the prediction accuracy and has little interest in declaring the chosen one being the best, the demand on the validation size may be much lessened.

Acknowledgments

This research was funded in part by the Defense Advanced Research Projects Agency under grant W911NF-18-1-0134. We thank Dr. Shuguang Cui and eight anonymous reviewers for giving feedback on the initial submission of the manuscript. We are also grateful to Dr. Matthew McKay and Dr. Osvaldo Simeone for handling the full submission of the manuscript, and to three anonymous reviewers for their comprehensive comments that have greatly improved the article.

Authors

Jie Ding (dingj@umn.edu) received his Ph.D. degree in engineering sciences from Harvard University, Cambridge, in 2017, where he worked as a postdoctoral researcher from January 2017 to December 2018. He was a postdoctoral researcher with the Information Initiative at Duke University, Durham, North Carolina, from January to August 2018. He is currently an assistant professor with the School of Statistics, University of Minnesota, Minneapolis. His research topics include signal processing, statistical inference, data prediction, and machine learning with applications to multimedia processing, human–computer interface, finance, and so on.

Vahid Tarokh (vahid.tarokh@duke.edu) received his Ph.D. degree in electrical engineering from the University of Waterloo, Ontario, Canada, in 1995. He worked at AT&T Labs-Research and AT&T Wireless Services until August 2000 as a member, principal member of technical staff and, finally, as the head of the Department of Wireless Communications and Signal Processing. In September 2000, he joined the Massachusetts Institute of Technology, Cambridge, as an associate professor of electrical engineering and computer science. In June 2002, he joined Harvard University, Cambridge, as a Gordon McKay Professor of Electrical Engineering and Hammond Vinton Hayes Senior Research Fellow. He was named Perkins Professor of Applied Mathematics and Hammond Vinton Hayes Senior Research Fellow of Electrical Engineering in 2005. In January 2018, he joined Duke University as the Rhodes Family Professor of Electrical and Computer Engineering, Computer Science, and Mathematics. From January 2018 to May 2018, he was also a Gordon Moore Distinguished Scholar with the California Institute of Technology, Pasadena. He is a Fellow of the IEEE.

Yuhong Yang (yyang@stat.umn.edu) received his Ph.D. degree in statistics from Yale University, New Haven, Connecticut, in 1996. He first joined the Department of Statistics at Iowa State University, Ames, then moved to the University of Minnesota, Minneapolis, in 2004, where he has been a full professor since 2007. His research interests include model selection, multiarmed bandit problems, forecasting, high-dimensional data analysis, and machine learning. He has published in journals in several fields, including *Annals of Statistics*, *IEEE Transactions on Information Theory*, *Journal of Econometrics*, *Journal of Approximation Theory*, *Journal of Machine Learning Research*, and *International Journal of Forecasting*. He is a fellow of the Institute of Mathematical Statistics.

References

[1] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[2] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 279–290, 2004.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[4] S. Greenland, "Modeling and variable selection in epidemiologic analysis," *Am. J. Public Health*, vol. 79, no. 3, pp. 340–349, 1989.

[5] C. M. Andersen and R. Bro, "Variable selection in regression: A tutorial," *J. Chemometrics*, vol. 24, nos. 11–12, pp. 728–737, 2010.

[6] J. B. Johnson and K. S. Omland, "Model selection in ecology and evolution," *Trends Ecol. Evolut.*, vol. 19, no. 2, pp. 101–108, 2004.

[7] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag, 2003.

[8] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *Ann. Statist.*, vol. 40, no. 1, pp. 561–592, 2012.

[9] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of Complexity*, V. Vovk, H. Papadoopoulos, and A. Gammerman, Eds. New York: Springer-Verlag, 1971, pp. 11–30.

[10] C.-K. Ing and C.-Z. Wei, "Order selection for same-realization predictions in autoregressive processes," *Ann. Statist.*, vol. 33, no. 5, pp. 2423–2474, 2005.

[11] R. Shibata, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Ann. Statist.*, vol. 8, no. 1, pp. 147–164, 1980.

[12] J. Shao, "An asymptotic theory for linear model selection," *Stat. Sinca*, vol. 7, no. 2, pp. 221–242, 1997.

[13] Y. Yang, "Comparing learning methods for classification," *Stat. Sinca*, vol. 16, pp. 635–657, 2006.

[14] W. Liu and Y. Yang, "Parametric or nonparametric? A parametricness index for model selection," *Ann. Statist.*, vol. 39, no. 4, pp. 2074–2102, 2011.

[15] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[16] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math*, vol. 21, no. 1, pp. 243–247, 1969.

[17] K. Takeuchi, "Distribution of informational statistics and a criterion of model fitting," *Suri-Kagaku (Mathematical Sciences)*, vol. 153, pp. 12–18, 1976.

[18] W. Pan, "Akaike's information criterion in generalized estimating equations," *Biometrics*, vol. 57, no. 1, pp. 120–125, 2001.

[19] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[20] P. M. Broersen, "Finite sample criteria for autoregressive order selection," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3550–3558, 2000.

[21] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[22] J. K. Ghosh and R. V. Ramamoorthi, *Bayesian Nonparametrics*. New York: Springer, 2003, p. 36.

[23] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc. Ser. B*, vol. 41, no. 2, pp. 190–195, 1979.

[24] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: A new criterion for autoregression," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4024–4043, 2018.

[25] J. Ding, V. Tarokh, and Y. Yang. (2016). Optimal variable selection in regression models. [Online]. Available: http://jding.org/jie-uploads/2017/11/regression.pdf

[26] P. M. Djuric, "Asymptotic map criteria for model selection," *IEEE Trans. Signal Process*, vol. 46, no. 10, pp. 2726–2735, 1998.

[27] C. Andrieu, P. Djuric, and A. Doucet, "Model selection by MCMC computation," *Signal Process*, vol. 81, no. 1, pp. 19–37, 2001.

[28] S. Shao, P. E. Jacob, J. Ding, and V. Tarokh, 2018. "Bayesian model comparison with the Hyvarinen score: Computation and consistency," *J. Am. Statist. Assoc.* doi: 10.1080/01621459.2018.1518237.

[29] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer J.*, vol. 11, no. 2, pp. 185–194, 1968.

[30] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, 2002.

[31] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[32] J. Rissanen, "Estimation of structure by minimum description length," *Circuits Syst. Signal Process.*, vol. 1, no. 3, pp. 395–406, 1982.

[33] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[34] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.

[35] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[36] A. P. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *J. Roy. Statist. Soc. Ser. A*, vol. 147, no. 2, pp. 278–292, 1984.

[37] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *J. Roy. Statist. Soc. Ser. B*, vol. 64, no. 4, pp. 583–639, 2002.

[38] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[39] S. Geisser, "The predictive sample reuse method with applications," *J. Amer. Statist. Assoc.*, vol. 70, no. 350, pp. 320–328, 1975.

[40] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 44–47, 1977.

[41] J. Rissanen, "A predictive least-squares principle," *IMA J. Math. Control Inform*, vol. 3, no. 2–3, pp. 211–222, 1986.

[42] C.-Z. Wei, "On predictive least squares principles," *Ann. Statist.*, vol, 20, no. 1, pp. 1–42, 1992.

[43] C.-K. Ing, "Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series," *Ann. Statist.*, vol. 35, no. 3, pp. 1238–1277, 2007.

[44] R. Nishii, "Asymptotic properties of criteria for selection of variables in multiple regression," *Ann. Statist.*, vol. 12, no. 2, pp. 758–765, 1984.

[45] C. L. Mallows, "Some comments on CP," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[46] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Relat. Fields*, vol. 113, no. 3, pp. 301–413, 1999.

[47] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, no. 1, pp. 45–54, 1981.

[48] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[49] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation," *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[50] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *Ann. Statist.*, vol, 22, no. 4, pp. 1947–1975, 1994.

[51] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, vol. 9, no. 1, pp. 60–62, 1938.

[52] T. van Erven, P. Grünwald, and S. De Rooij, "Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma," *J. Roy. Statist. Soc. Ser. B*, vol. 74, no. 3, pp. 361–417, 2012.

[53] D. L. Donoho, and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," *Ann. Statist.*, vol. 26, no. 3, pp. 879–921, 1998.

[54] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math*, vol. 57, no. 11, pp. 1413–1457, 2004.

[55] E. Candes, and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[56] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Appl. Comput. Harmon. A*, vol. 26, no. 3, pp. 395–407, 2009.

[57] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[58] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[59] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, 2010.

[60] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The LASSO and Generalizations*. Boca Raton, FL: CRC, 2015.

[61] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[62] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[63] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B*, vol. 67, no. 2, pp. 301–320, 2005.

[64] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[65] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc*, vol. 101, no. 476, pp. 1418–1429, 2006.

[66] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, 1989.

[67] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. IEEE 7th Asilomar Conf. Signals, Systems, Computers*, 1993, pp. 40–44.

[68] C.-K. Ing and T. L. Lai, "A stepwise regression method and consistent model selection for high-dimensional sparse linear models," *Statist. Sinca*, vol. 21, pp. 1473–1513, Oct. 2011.

[69] J. Ding, L. Chen, and Y. Gu, "Perturbation analysis of orthogonal matching pursuit," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 398–410, 2013.

[70] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of LASSO and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[71] J. Lafferty and L. Wasserman, "Rodeo: Sparse, greedy nonparametric regression," *Ann. Statist.*, vol. 36, no. 1, pp. 28–63, 2008.

[72] Q. Han, J. Ding, E. M. Airoldi, and V. Tarokh, "SLANTS: Sequential adaptive nonlinear modeling of time series," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 4994–5005, 2017.

[73] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[74] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[75] P. Zhao and B. Yu, "On model selection consistency of LASSO," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Nov. 2006.

[76] S. Yang, M. Santillana, and S. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proc. Natl. Acad. Sci. U.S.A*, vol. 112, no. 47, pp. 14,473–14,478, 2015.

[77] Y. Nan and Y. Yang, "Variable selection diagnostics measures for high-dimensional regression," *J. Comp. Graph. Statist.*, vol. 23, no. 3, pp. 636–656, 2014.

[78] J. P. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, vol. 2, no. 8, pp. 696–701, 2005.

[79] Y. Yang, "Adaptive regression by mixing," *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 574–588, 2001.

[80] N. Meinshausen and P. Bühlmann, "Stability selection," *J. Roy. Statist. Soc. Ser. B*, vol. 72, no. 4, pp. 417–473, 2010.

[81] D. Ferrari and Y. Yang, "Confidence sets for model selection by f-testing," *Stat. Sinca*, vol. 25, no. 4, pp. 1637–1658, 2015.

[82] C. Lim and B. Yu, "Estimation stability with cross-validation (ESCV)," *J. Comput. Graph. Statist.*, vol. 25, no. 2, pp. 464–492, 2016.

[83] W. Yang and Y. Yang, "Toward an objective and reproducible model choice via variable selection deviation," *Biometrics*, vol. 73, no. 1, pp. 20–30, 2017.

[84] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao, "Valid post-selection inference," *Ann. Statist.*, vol. 41, no. 2, pp. 802–837, 2013.

[85] J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. (2014). Post-selection adaptive inference for least angle regression and the LASSO. [Online]. Available: https://www.researchgate.net/publication/259764783_Post-selection_adaptive_inference_for_Least_Angle_Regression_and_the_Lasso

[86] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *J. Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.

[87] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[88] H. Leeb and B. M. Pötscher, "Sparse estimators and the oracle property, or the return of Hodges estimator," *J. Econom.*, vol. 142, no. 1, pp. 201–211, 2008.

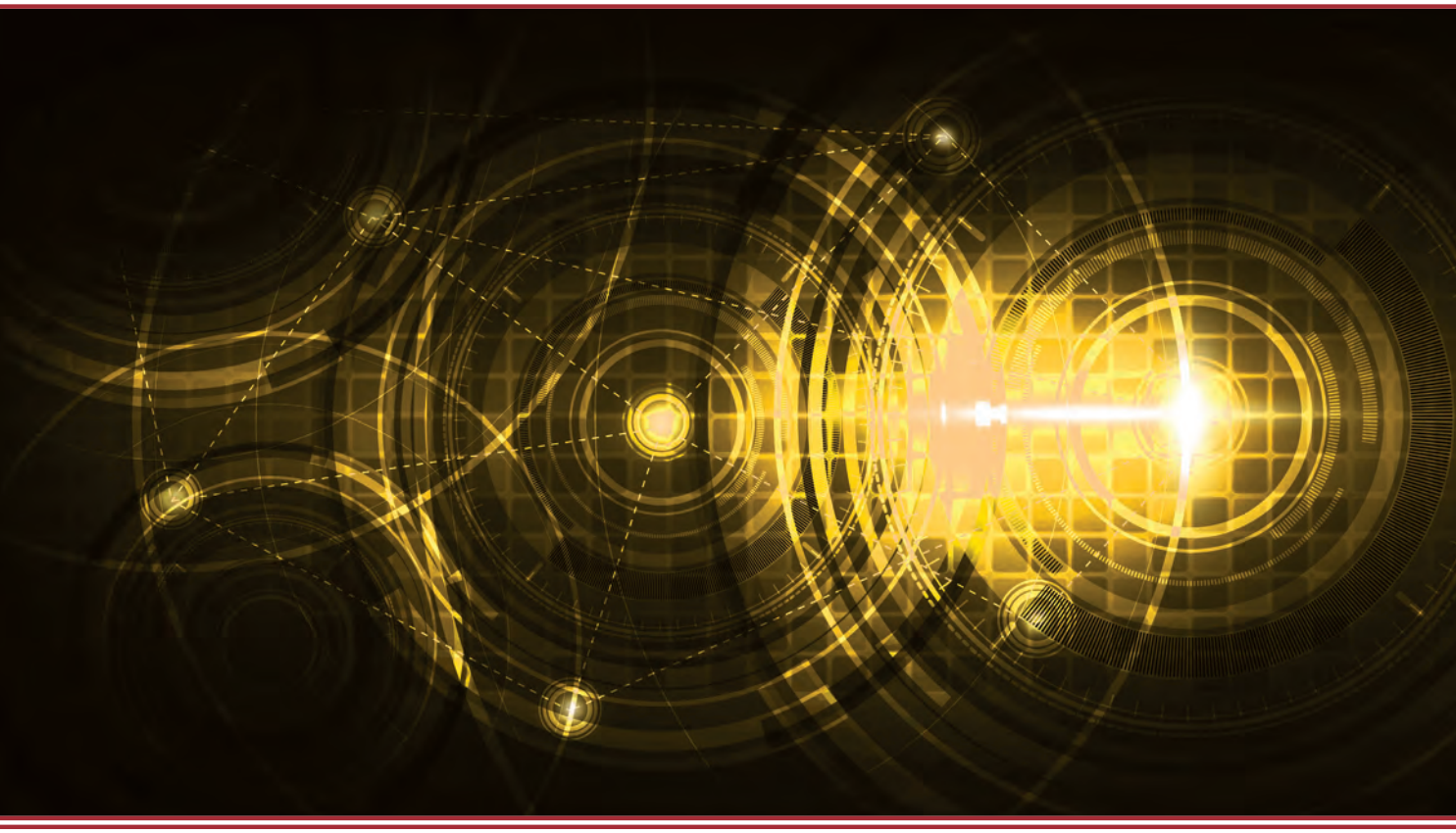
[89] Y. Yang, "Prediction/estimation with simple linear models: Is it really that simple?" *Econom. Theory*, vol. 23, no. 1, pp. 1–36, 2007.

[90] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

SP

Deborah Cohen and Yonina C. Eldar

Sub-Nyquist Radar Systems



Temporal, spectral and spatial compression

Radar is an acronym for "radio detection and ranging." However, the functions of today's radar systems, both in civilian and military applications, go beyond simple target detection and localization; they extend to tracking, imaging, classification, and more and involve different types of radar systems, such as through-the-wall [1], ground-penetration [2], automotive [3], and weather [4]. Although radar technology has been well established for decades, a new line of compressed radars has recently emerged. These aim at reducing the complexity of classic radar systems by exploiting inherent prior information on the structure of the received signal from the targets. The goal of this article is to review these novel sub-Nyquist radars and their potential applications.

Conventional radar systems transmit electromagnetic waves of near-constant power in very short pulses toward the targets of interest. Between outgoing pulses, the radar measures the signal reflected from the targets to determine their presence, range, velocity, and other characteristics. Different systems use different radar waveforms and varying transmit strategies. One of the most popular methods is pulse-Doppler radar, which periodically transmits identical pulses. In contrast, stepped-frequency radars (SFRs) [5] vary the carrier frequency of each pulse. Some systems rely on simple traditional waveforms such as Gaussian pulses while others adopt more complex signals, such as chirps [6], [7]. Each configuration corresponds

Digital Object Identifier 10.1109/MSP.2018.2868137
Date of publication: 13 November 2018

to a certain choice in the complexity-performance tradeoff, between complex waveform and system designs and target detection and estimation.

State-of-the-art radar systems operate with large bandwidths, large coherent processing intervals (CPIs), and high number of antennas in multiple-input, multiple-output (MIMO) settings [8], [9], to achieve high-range velocity and azimuth resolution, respectively. This, in turn, generates large data sets to be sampled, stored, and processed, creating a bottleneck in terms of both analog system complexity, including high-rate analog-to-digital converters (ADCs), and subsequent digital processing [10].

Although the natural application of CS is typically the reduction of the required number of samples to perform a certain signal processing task, it was first used by the radar community to increase a target's parameter resolution.

In the past few years, novel approaches to radar signal processing have emerged that allow radar signal detection and parameter estimation using a much smaller number of measurements than that required by spatial and temporal Nyquist sampling. While temporal sampling refers to taking samples in time intervals determined by the sampling rate, spatial sampling extends this notion to placing transmit and receive antennas, whose locations are governed by the signal wavelength. These works capitalize on the fact that, in most radar applications, the reflectivity scene consists of a small number of strong targets. That is, the reflected signals by only a few targets have high enough power to be detected by the radar receiver. In pulse-Doppler radar, the target scene is often sparse in the joint time–frequency, or ambiguity, domain [5]. In synthetic aperture radar (SAR) [11], the scene is often sparse in the Fourier or wavelet domain, or even in the image domain.

Over the past decade, many works have exploited the inherent sparsity of the target scene to enhance radar-estimation capabilities. These rely on the compressed sensing (CS) [10], [12] framework, brought to the forefront by the works of Candes, Romberg, and Tao [13] and of Donoho [14]. Although the natural application of CS is typically the reduction of the required number of samples to perform a certain signal processing task, it was first used by the radar community to increase a target's parameter resolution [15]–[20]. It was later applied to reduce the number of samples to be processed [21]–[25] and finally to reduce the sampling rate [26], [27] and number of antennas [28] required in radar systems, performing time and spatial compression and alleviating the burden on both the analog and digital sides. In particular, the recently proposed Xampling, i.e., compressed sampling concept [10], [29], has been applied to radar [30]–[32] to break the link between bandwidth, CPI, and the number of antennas on the one hand, and range, Doppler, and azimuth resolution, respectively, on the other hand.

The reviews of compressive radar [22], [33]–[35] mostly deal with radar imaging. The works in [33] and [34] focus on SAR imaging and consider sparsity-based radar imagery using both greedy algorithms, which iteratively recover the sparse target scene, and convex relaxations of sparsity-inducing regularization. The special cases of interferometric, polarimetric, and circular SAR are presented in [33] for both two-dimensional (2-D) and three-dimensional (3-D) images. In [34], diverse SAR applications, such as wide-angle SAR imaging, joint imaging, and autofocusing from data with phase errors, moving targets, analysis, and design of SAR sensing missions, are reviewed. A survey of statistical sparsity-based techniques for radar imagery applications is presented in [35], including superresolution imaging, enhanced-target imaging, autofocusing, and moving-target imaging. The review of [22] presents three applications of CS radars: pulse compression, radar imaging, and airspace surveillance with array antennas. At the time it was written, there was a small number of publications addressing the application of CS to radar, as stated by the authors.

In this article, we focus on nonradar-imaging applications and survey many recent works that exploit CS in different radar systems to achieve various goals. We consider different transmit waveforms and processing approaches, while focusing on pulse-Doppler radar—one of the most popular systems—and its extension to MIMO configurations. Our goal is to review the main impacts of compressed radar on parameter resolution as well as digital and analog complexity. The survey includes fast time compression schemes, which reduce the number of acquired samples per pulse; slow time compression techniques, which decrease the number of pulses; and spatial compression approaches, in which the number of transmit and receive antenna elements is reduced. We show that, beyond a substantial rate reduction, compression may also enable communication and radar spectrum sharing [36]–[38], as elaborated on in [39]. Throughout this article, we consider both theoretical and practical aspects of compressed radar and present hardware prototype implementations [40]–[43] of the theoretical concepts, demonstrating the real-time target parameters' recovery from low-rate samples in pulse-Doppler and MIMO radars.

Radar systems

Radar systems aim to estimate targets' parameters to determine their location and motion. In its simplest form, the radar transmits a single pulse toward targets in one direction and recovers their range, i.e., distance to the radar, which is proportional to the received pulse delay. More elaborate systems are able to provide additional information on the targets. Pulse-Doppler radars transmit several pulses, enabling them to resolve both the targets' ranges and radial velocities, which are proportional to the Doppler frequency. Stepped-frequency-based approaches achieve highly effective bandwidths that increase range resolution, while allowing for narrow instantaneous bandwidth. MIMO radars use several elements both at the transmitter and at the receiver to illuminate the entire target scene and recover targets' azimuths in addition to their ranges and velocities. In this article, we consider the application of compression in

terms of the number of required samples, pulses, and antennas, as well as its impact on different aspects of the radar system, including parameter resolution and system complexity, for several types of radars.

Pulse-Doppler radar

A standard pulse-Doppler radar transceiver detects targets by transmitting a periodic stream of pulses and processing its reflections. The transmitted signal $x_T(t)$ consists of P equally spaced pulses $h(t)$ such that

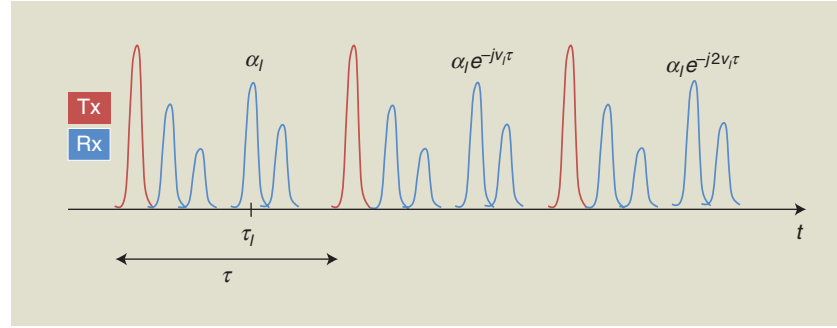


FIGURE 1. The pulse-Doppler radar transmitted and received pulse trains with $P = 3$ pulses and $L = 4$ targets Tx: transmitted; Rx: received.

$$x_T(t) = \sum_{p=0}^{P-1} h(t - p\tau), \quad 0 \le t \le P\tau. \tag{1}$$

The pulse-to-pulse delay τ is the pulse-repetition interval (PRI), and its reciprocal $1/\tau$ is the pulse-repetition frequency (PRF). The entire span of the signal in (1), i.e., $P\tau$, is the CPI. The pulse time support is denoted by T_p, with $0 < T_p < \tau$. The pulse $h(t)$ is typically a known time-limited baseband function with continuous-time Fourier transform (CTFT) $H(f) = \int_{-\infty}^{\infty} h(t) e^{-j2\pi ft}\, dt$ that has negligible energy at frequencies beyond $B_h/2$, where B_h is referred to as the bandwidth of $h(t)$. An example of a transmitted pulse train is illustrated in Figure 1.

It is typically assumed that the target scene is composed of L nonfluctuating point-targets, according to the Swerling-0 model [5]. This is one of the popular models in the radar signal processing literature since, by describing an idealized target, it allows simplifying the radar equations while constituting a fairly good approximation in many applications [6], [7]. Other models, such as Swerling-1, which applies to targets composed of many independent scatters, or fluctuating target models, are beyond the scope of this article. The pulses reflect off the L targets and propagate back to the transceiver. Each target l is defined by three parameters:

1) a time delay $\tau_l = 2r_l/c$, proportional to the target's distance to the radar or range r_l, where c is the speed of light
2) a Doppler-radial frequency $\nu_l = 2\dot{r}_l f_c/c$, proportional to the target's radial velocity to the radar, i.e., the target's velocity radial component $\dot{r}_l$, and the radar's carrier frequency f_c
3) a complex amplitude α_l, proportional to the target's radar cross section (RCS), dispersion attenuation, and other propagation factors.

The targets are defined in the radar radial coordinate system and are typically assumed to lie in the radar unambiguous time–frequency region: delays up to the PRI and Doppler frequencies up to the PRF. When this assumption does not hold, several processing techniques have been proposed that require the transmission of multiple pulse trains with different parameters, e.g., different PRFs. We review this setting in the "Range-Velocity Ambiguity Resolution" section.

Based on the three assumptions A1–A3 presented in "Targets' Assumptions," the received signal can be written as

$$x_R(t) = \sum_{p=0}^{P-1} \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l - p\tau) e^{-j\nu_l p\tau}, \quad 0 \le t \le P\tau. \tag{2}$$

It will be convenient to express $x_R(t)$ as a sum of single frames

$$x_R(t) = \sum_{p=0}^{P-1} x_p(t), \tag{3}$$

where

$$x_p(t) = \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l - p\tau) e^{-j\nu_l p\tau}, \quad 0 \le t \le P\tau. \tag{4}$$

An illustration of a received pulse train is shown in Figure 1 with $L = 4$ targets. In pulse-Doppler radar, the goal is to recover the three L parameters $\{\tau_l, \nu_l, \alpha_l\}$ for $0 \le l \le L-1$ from the received signal $x_R(t)$. In particular, estimating the time delays τ_l and Doppler frequencies ν_l enables an approximation of the targets' distances and radial velocities.

Stepped-radar waveforms

In classic pulse-Doppler radar, high-range resolution requires a large signal bandwidth. This technology bottleneck is partially overcome by stepped-frequency-based waveforms, in which the large bandwidth is obtained sequentially by stepping the frequency of each pulse, keeping the instantaneous bandwidth low. Two popular examples of such waveforms are SFRs and stepped chirps. An SFR [5] system transmits P-narrowband pulses, in which each pulse p has carrier frequency

$$f_p = f_0 + p\Delta_f, \tag{5}$$

for $0 \le p \le P-1$, with f_0 the initial frequency and Δf the frequency increment. The pth-transmitted pulse is a rectangular pulse modulated by its carrier f_p. The corresponding received signal is then of the form

Targets' Assumptions

To simplify the received signal model, the following assumptions of the targets' locations and motions are typically made [5]:

- *A1*: Far targets: The target-radar distance is large compared with the distance change during the coherent processing intervals (CPIs), which allows for constant α_l within the CPI:

$$\dot{r}_l P\tau \ll r_l \Rightarrow \nu_l \ll \frac{f_c \tau_l}{P\tau}. \tag{S1}$$

- *A2*: Slow targets: The constant-Doppler phase during pulse time,

$$\nu_l T_p \ll 1, \tag{S2}$$

and low target velocity allows for constant τ_l during the CPI. This condition holds when the baseband Doppler frequency is smaller than the frequency resolution:

$$\frac{2\dot{r}_l B_h}{c} \ll \frac{1}{P\tau} \Rightarrow \nu_l \ll \frac{f_c}{P\tau B_h}. \tag{S3}$$

- *A3*: Small acceleration: The target velocity remains approximately constant during the CPI allowing for constant ν_l. This condition is satisfied when the velocity change induced by acceleration is smaller than the velocity resolution:

$$\ddot{r}_l P\tau \ll \frac{c}{2f_c P\tau} \Rightarrow \ddot{r}_l \ll \frac{c}{2f_c (P\tau)^2}. \tag{S4}$$

Although these assumptions may seem hard to comply with, they all rely on slow enough relative motion between the radar and its targets. Radar systems tracking people, ground vehicles, and sea vessels usually comply quite easily [6].

In multiple-input, multiple-output settings, two additional assumptions are adopted on the array structure and transmitted waveforms:

- *A4*: Collocated array: The target radar cross sections α_l and θ_l are constant over the array [44].
- *A5*: Narrowband waveform: A small aperture allows τ_l to be constant over the channels:

$$\frac{2Z\lambda}{c} \ll \frac{1}{B_h}. \tag{S5}$$

$$x_p(t) = \sum_{l=0}^{L-1} \alpha_l \text{rect}(t-\tau_l) e^{-j2\pi f_p(t-\tau_l)} e^{j\nu_l p\tau}. \tag{6}$$

To process the received signal, the delay is neglected in the signal envelope because of the narrowband assumption. An SFR traditionally obtains one sample from each received pulse and computes the phase detector output sequence as

$$y_p = \sum_{l=0}^{L-1} \alpha_l e^{j2\pi f_p \tau_l} e^{j\nu_l p\tau}. \tag{7}$$

The phase detector signal y_p can be modeled as the product of the received signal (6) and the reference signal, followed by a low-pass filter (LPF). Conventional processing applies an inverse discrete Fourier transform (DFT) on the output to estimate the targets' time delays τ_l and Doppler frequencies ν_l. The range resolution achieved by SFR is $c/2P\Delta_f$, where $P\Delta_f$ is the total effective bandwidth of the signal over P pulses.

Another popular stepped waveform is the stepped chirp or multifrequency chirp signal. The corresponding transmitted signal is given by

$$x_T(t) = \sum_{p=0}^{P-1} e^{j\phi_p} \text{rect}\Big(\frac{t}{\tau}\Big) e^{j2\pi\left(f_p t + \frac{\gamma}{2}t^2\right)}, \tag{8}$$

where γ is the common chirp rate and f_p and ϕ_p are the frequency and complex phase of the pth subcarrier. The returned signal corresponding to the pth pulse, given by

$$x_p(t) = \sum_{l=0}^{L-1} \alpha_l e^{j\phi_p} \text{rect}\Big(\frac{t-\tau_l}{\tau}\Big) e^{j2\pi\left(f_p(t-\tau_l) + \frac{\gamma}{2}(t-\tau_l)^2\right)}, \tag{9}$$

is dechirped with a reference linear-frequency waveform of fixed frequency equal to the first carrier f_0:

$$m(t) = \text{rect}\Big(\frac{t-\tau_r}{\tau}\Big) e^{-j2\pi\left(f_0 t + \frac{\gamma}{2}t^2\right)}. \tag{10}$$

The receive window is $\tau_r = 2(r_{\max} + r_{\min})/c$, and the reference delay is $t_r = (r_{\max} + r_{\min})/c$, with $r_{\max}$ and $r_{\min}$ as the maximal and minimal ranges, respectively. The resulting dechirped received signal can be written as

$$x_p(t) = \sum_{l=0}^{L-1} \alpha_l e^{j(\phi_p - 2\pi f_p \tau_l)} \text{rect}\Big(\frac{t-\tau_l+t_r}{\tau}\Big) e^{j2\pi(f_p - f_0 - \gamma\tau_l)t}. \tag{11}$$

Classic processing of the received signal includes a DFT operation to recover the targets' delays τ_l.

MIMO pulse-Doppler radar

MIMO radar presents significant potential for advancing state-of-the-art modern radar in terms of flexibility and performance. This configuration [8] combines several antenna elements both at the transmitter and receiver. Unlike phased-array systems, each transmitter radiates a different waveform, which offers more degrees of freedom (DoF) [9]. There are two main configurations of MIMO radar, depending on the location of the transmitting and receiving elements; collocated MIMO

[46], in which the elements are close to each other relative to the signal wavelength, and multistatic MIMO [47], where they are widely separated. In this article, we focus on collocated pulse-Doppler MIMO systems.

Collocated MIMO radar systems exploit waveform diversity, based on mutual orthogonality of the transmitted signals [9]. Consequently, the performance of MIMO systems can be characterized by a virtual array constructed by the convolution of the locations of the transmit and receive antenna locations. In principle, with the same number of antenna elements, this virtual array may be much larger than the array of an equivalent traditional system [48].

The standard approach to collocated MIMO adopts a virtual uniform linear array (ULA) structure [49], where R receivers, spaced by $\lambda/2$ and T transmitters and spaced by $R(\lambda/2)$ (or vice versa), form two ULAs. Here, λ is the signal wavelength. Coherent processing of the resulting TR channels generates a virtual array equivalent to a phased array with TR $(\lambda/2)$−spaced receivers and normalized aperture $Z = TR/2$. Denote by $\{\xi_m\}_{m=0}^{T-1}$ and $\{\zeta_q\}_{q=0}^{R-1}$, the normalized transmitters' and receivers' locations, respectively. For the traditional virtual ULA structure, denote $\zeta_q = q/2$ and $\xi_m = Rm/2$. This standard-array structure and the corresponding virtual array are illustrated in Figure 2 for $R = 3$ and $T = 5$. The circles represent the receivers, and the squares are the transmitters.

Each transmit antenna sends P pulses, such that the mth-transmitted signal is given by

$$s_m(t) = \sum_{p=0}^{P-1} h_m(t - p\tau)e^{j2\pi f_c t}, \quad 0 \le t \le P\tau, \tag{12}$$

where $h_m(t), 0 \le m \le T-1$ are orthogonal pulses with bandwidth B_h and modulated with carrier frequency f_c. For convenience, it is typically assumed that $f_c\tau$ is an integer, so that the initial phase for every pulse $e^{-j2\pi f_c \tau p}$ is canceled in the modulation for $0 \le p \le P-1$ [6].

MIMO radar architectures impose several requirements on the transmitted waveform family. Besides traditional demands from radar waveforms such as low sidelobes, MIMO transmit antennas rely on orthogonal waveforms. In addition, to avoid cross talk between the T signals and form TR channels, the orthogonality condition should be invariant to time shifts, that is $\int_{-\infty}^{\infty} s_i(t)s_j^*(t-\tau_0)dt = \delta(i-j)$ for $i, j \in [0, T-1]$ and for all τ_0. The main waveform families typically considered are time-division multiple access (TDMA), frequency-division multiple access (FDMA), and code-division multiple access (CDMA), respectively. Time-invariant orthogonality is achieved by FDMA and TDMA and approximately achieved by CDMA, as the latter involves overlapping frequency bands [50].

Besides the traditional assumptions on the targets, MIMO systems present additional requirements on the radar array and waveforms with respect to the targets, as described in "Targets' Assumptions." In the MIMO configuration, the goal is to recover the targets' azimuth angles θ_l in addition to their delays τ_l and Doppler shifts ν_l from the received signals.

Current challenges

Standard radar processing samples and processes the received signal at its Nyquist rate B_h. For example, the pulse-Doppler classic radar processing, described in "Classic Pulse-Doppler and Multiple-Input, Multiple-Output Processing," first filters the sampled signal by a matched filter (MF). In modern systems, the MF operation is performed digitally and therefore requires an ADC capable of sampling at rate B_h. Other radar systems similarly require sampling the received signal at its Nyquist rate. The radar bandwidth B_h is inversely proportional to the system fast time, or range resolution, and can thus be hundreds of megahertz or even up to several gigahertz, requiring a high sampling rate and resulting in a large number of samples per pulse $N = \tau B_h$ to process.

The slow time (Doppler) resolution is inversely proportional to the CPI $P\tau$. The Doppler processing stage can be viewed as an MF in the pulse dimension, i.e., slow time domain, to a constant radial velocity target. As such, it increases the signal-to-noise ratio (SNR) by P compared to the SNR of a single pulse [7]. Since an MF is the linear time-invariant system that maximizes SNR, it follows that a factor P increase is optimal for P pulses. A large number of pulses increases resolution and SNR but leads to large time on target and a large total number of samples to process, given by PN.

The required computational power corresponds to P convolutions of a signal of length $N = \tau B_h$ and N-fast Fourier transforms (FFTs) of length P (see "Classic Pulse-Doppler and Multiple-Input, Multiple-Output Processing"). The growing demands for improved estimation accuracy and target separation dictate an ever-growing increase in the signal's bandwidth and CPI. This creates bottlenecks in sampling and processing rates in the fast time (intrapulse) domain and in time on target in the slow time (interpulse) dimension.

In MIMO radar, the additional spatial dimension increases the system's complexity, as may be seen in "Classic Pulse-Doppler and Multiple-Input, Multiple-Output Processing." In such systems, the array aperture determines the azimuth resolution. In a traditional virtual array configuration, the product between the number of transmit and receive antennas scales linearly with the aperture. Consequently, high resolution

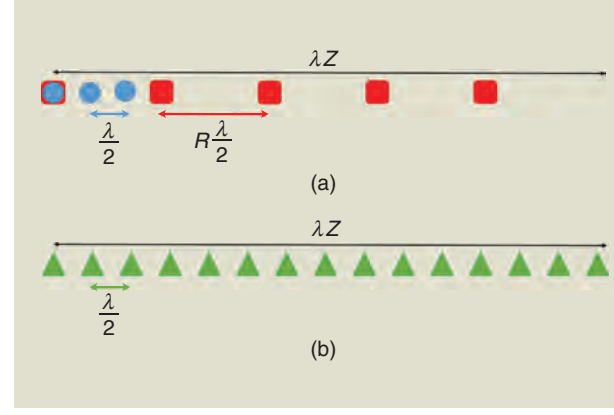


FIGURE 2. An illustration of MIMO arrays: (a) a standard array and (b) a corresponding receiver virtual array [32].

Classic Pulse-Doppler and Multiple-Input, Multiple-Output Processing

The classic methods for radar processing typically consist of the following stages [5], [45]:

1) *Sampling*: Sample each incoming frame $x_p(t)$ at its Nyquist rate B_h, equal to the double-sided bandwidth of $h(t)$, creating the samples $x_p[n], 0 \le n \le N-1$, where $N = \tau B_h$. We assume, for simplicity, that N is an integer.
2) *Matched filter* (*MF*): Apply a standard MF on each frame $x_p[n]$. This results in the outputs $y_p[n] = x_p[n] * h[-n]$, where $h[n]$ is the sampled version of the transmitted pulse $h(t)$ at its Nyquist rate and $*$ is the convolution operation. The time resolution attained in this step is $1/B_h$.
3) *Doppler processing*: For each discrete time n, perform a P-point discrete Fourier transform along the pulse dimension, i.e., $z_n[k] = \mathrm{DFT}_P\{y_p[n]\} = \sum_{p=0}^{P-1} y_p[n] e^{-j2\pi pk/P}$ for $0 \le k \le P$. The Doppler resolution is $1/\mathrm{P}\tau$.
4) *Delay-Doppler map*: Stacking the vectors $\mathbf{z}_n$ and taking absolute value, we obtain a delay-Doppler map $\mathbf{Z} = \mathrm{abs}[\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}] \in \mathbb{R}^{P \times N}$.
5) *Peak detection*: A heuristic detection process, in which knowledge of the number of targets, targets' powers, clutter location, and so on, may help in discovering targets' positions. For example, if we know there are L targets, then we can choose the L-strongest points in the map. Alternatively, constant false alarm (FA) rate detectors determine the power threshold, above which a peak is considered to originate from a target so that a required probability of FA is achieved.

Classic collocated multiple-input, multiple-output radar processing traditionally includes the following stages:

1) *Sampling*: At each receiver $0 \le q \le R-1$, where R denotes the number of receivers, the signal $x_q(t)$ is sampled at its Nyquist rate Btot. In code-division multiple access and time-division multiple access, $B_{\mathrm{tot}} = B_h$ as all waveforms overlap in frequency, whereas in frequency-division multiple access, $B_{\mathrm{tot}} = TB_h$, where B_h denotes the bandwidth of a single waveform in both cases, and T is the number of transmitters.
2) *MF*: The sampled signal is convolved with a sampled version of $h_m(t)$, for $0 \le m \le T-1$. The time resolution attained in this step is $1/B_{\mathrm{tot}}$.
3) *Beamforming*: The correlations between the observation vectors from the previous step and the steering vectors corresponding to each azimuth on the grid defined by the array aperture are computed. The spatial resolution attained in this step is $2/TR$.
4) *Doppler detection*: The correlations between the resulting vectors and Doppler vectors, with Doppler frequencies lying on the grid defined by the number of pulses, are computed. The Doppler resolution is $1/\mathrm{P}\tau$.
5) *Peak detection*: This is similar to classic radar, but detection is performed on the three-dimensional range-azimuth-Doppler map.

requires a large number of antennas, thus increasing the system's complexity in terms of hardware and processing.

In the following sections, we review fast time-compressed radar systems that allow for low-rate sampling and processing of radar signals, regardless of their bandwidth, while retaining the same SNR scaling. We then demonstrate how compression can be extended to the slow time, thereby reducing time on target, and to the spatial dimension allowing one to achieve resolution similar to a filled array but with significantly fewer elements. In reality, the received signal $x_R(t)$ is further contaminated by additive noise and clutter. We will also demonstrate the impact of SNR and clutter on compressed radar system prototypes [30], [40], [54]. Finally, we show how compression and sub-Nyquist sampling may be used to address other challenges, such as communication and radar spectrum sharing.

Increased parameter resolution

In many radar applications, the reflectivity scene consists of a small number L of strong targets. Therefore, CS techniques (see "Compressed Sensing Recovery") are a natural processing tool for radar systems. Shortly after the idea of CS was brought forward by the works of Candes, Romberg, and Tao [13] and of Donoho [14] a decade ago, it was introduced to pulse-Doppler radar [15], [16], [55] and SFR [17].

While CS is typically applied to signal processing tasks to reduce the associated sampling rate [10], earlier papers that applied CS recovery to pulse-Doppler radar and SFR were aimed at increasing delay-Doppler resolution [15]–[17], [20] using Nyquist samples. More recent approaches use CS recovery techniques on low-rate, or sub-Nyquist samples, which enable sampling and processing rate reduction while achieving the same resolution as traditional Nyquist radars. Later in this section, we review radar recovery methods that increase delay-Doppler resolution using CS techniques on Nyquist samples. In the next sections, we consider the application of CS to reduce the fast time-sampling rate and the number of pulses and antennas, while preserving the resolution achieved by Nyquist systems.

In the works of [15]–[17] and [20], the signal is still sampled at its Nyquist rate B_h, but the delay and Doppler resolutions are determined by the CS grid containing $N > \tau B_h$ grid points, rather than the signal's bandwidth and CPI, respectively. The key idea in [15], which adopts a pulse-Doppler radar model, is that the received signal $x_R(t)$ defined in (2) is generally a sparse superposition of time- and frequency-shifted replicas of the transmitted waveforms. The time–frequency plane is discretized into an $N \times N$ grid in which each point represents a unique time–frequency shift $\mathbf{H}_i$, expressed as the product of

Compressed Sensing Recovery

Compressed sensing (CS) [10], [12] is a framework for simultaneous sensing and compression of finite-dimensional vectors, which relies on linear dimensionality reduction. In particular, the field of CS focuses on the recovery problem

$$\mathbf{z} = \mathbf{A}\mathbf{x}, \tag{S6}$$

where $\mathbf{x}$ is an $N \times 1$ sparse vector, i.e., with few nonzero entries, and $\mathbf{z}$ is a vector of measurements of size $M < N$. CS provides recovery conditions and algorithms to reconstruct $\mathbf{x}$ from the low-dimensional vector $\mathbf{z}$.

Two popular CS greedy recovery algorithms, orthogonal matching pursuit (OMP) and iterative hard thresholding (IHT), attempt to solve the optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t. } \mathbf{z} = \mathbf{A}\mathbf{x}, \tag{S7}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm. OMP [51], [52] iteratively proceeds by finding the column of A most correlated to the signal residual $\mathbf{r}$,

$$i = \text{argmax} \, | \, \mathbf{A}^H \mathbf{r} \, | \, , \tag{S8}$$

where the absolute value is computed element-wise and $(\cdot)^H$ is the Hermitian operator. The residual is obtained by subtracting the contribution of a partial estimate $\hat{\mathbf{x}}_\ell$ of the signal at the ℓth iteration, from $\mathbf{z}$, as follows:

$$\mathbf{r} = \mathbf{z} - \mathbf{A}\hat{\mathbf{x}}_\ell. \tag{S9}$$

It is initialized by $\mathbf{r} = \mathbf{z}$. Once the support set is updated by adding the index i, the coefficients of $\hat{\mathbf{x}}_\ell$ over the support set are updated, so as to minimize the residual error.

Other greedy techniques include thresholding algorithms. We focus here on the IHT method proposed in [53]. Starting from an initial estimate $\hat{\mathbf{x}}_0 = 0$, the algorithm iterates a gradient-descent step with step size μ followed by hard thresholding, i.e.,

$$\hat{\mathbf{x}}_\ell = \mathcal{T}(\hat{\mathbf{x}}_{\ell-1} + \mu \mathbf{A}^H(\mathbf{z} - \mathbf{A}\hat{\mathbf{x}}_{\ell-1}), k), \tag{S10}$$

until a convergence criterion is met. Here, $\mathcal{T}(\mathbf{x}, k)$ denotes a thresholding operator on $\mathbf{x}$ that sets all but the k entries of $\mathbf{x}$ with the largest magnitudes to zero, and k is the sparsity level of $\mathbf{x}$ (assumed to be known).

Alternative approaches to greedy recovery are convex-relaxation-based methods using ℓ_1 regularization such as basis pursuit and least absolute shrinkage and selection operator, better known as *LASSO*. Further details on CS recovery conditions and techniques can be found in [10] and [12].

time-shift and frequency-modulation matrices, denoted by $\mathbf{T}^{(\cdot)}$ and $\mathbf{M}^{(\cdot)}$, respectively. In particular,

$$\mathbf{H}_i = \mathbf{M}^{i \bmod N} \mathbf{T}^{\lfloor i/N \rfloor}, \tag{13}$$

where

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & & 1 \\ 1 & 0 & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} 1 & & & 0 \\ & e^{j\frac{2\pi}{N}} & & \\ & & \ddots & \\ 0 & & & e^{j\frac{2\pi}{N}(N-1)} \end{pmatrix}. \tag{14}$$

Here, $\lfloor \cdot \rfloor$ and mod denote the floor and modulation functions, respectively.

The vector $\mathbf{y}$ that concatenates the Nyquist samples of a single pulse $x_p(t)$ can then be expressed as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{s}, \tag{15}$$

where $\mathbf{s}$ is the L-sparse vector of size N^2 whose nonzero entries are the targets' RCS α_l with locations determined by the corresponding time–frequency shift. The ith column, i.e., atom of the $N \times N^2$ matrix $\mathbf{\Phi}$, is given by

$$\mathbf{\Phi}_i = \mathbf{H}_i \mathbf{f}, \tag{16}$$

where the vector $\mathbf{f}$ contains the Nyquist rate samples $h[n]$ of the transmitted signal $h(t)$. The latter is chosen so that the samples correspond to the Alltop sequence $h[n] = \left(1/\sqrt{N}\right) e^{2\pi j n^3/N}$ [56], for some prime $N \geq 5$. This yields a low-coherence matrix $\mathbf{\Phi}$, i.e., a matrix whose columns have small correlation.

The vector $\mathbf{s}$ is reconstructed from $\mathbf{y}$ using CS techniques, as described in "Compressed Sensing Recovery." The time–frequency shifts, determined by the targets' delays and Doppler frequencies, are thus recovered with a resolution of $1/N$.

The CS recovery in [15] is performed without an MF, which reduces performance in low-SNR regimes. Additionally, [15] considers only delay recovery. Alternatively, CS techniques can be performed after applying an analog MF [16] on the pulse-Doppler-received signal (2). The MF output of the pth pulse, sampled at the Nyquist rate $1/B_h$, is given by

$$w_p[k] = \sum_{l=0}^{L-1} \alpha_l e^{j\nu_l \tau_l} e^{j\nu_l p\tau} C_h[k - \tau_l/\tau], \tag{17}$$

where $C_h[k]$ is the discrete autocorrelation function of the transmitted waveform. For each sampling time k, the Nyquist samples have a sparse representation in the frequency (Doppler) domain using a Fourier matrix as a dictionary. A

two-step approach is therefore proposed to apply CS recovery for each k. However, the sidelobes of $C_h[k]$ lead to ambiguity. To avoid this, pairs of Golay complementary sequences x_1 and x_2 of length N, whose correlation functions satisfy

$$C_{x_1}[k] + C_{x_2}[k] = 2N\delta[k], \tag{18}$$

are transmitted alternatively by phased coding of the baseband waveform. This allows for unambiguous delay-Doppler recovery, provided that all of the Doppler coordinates are within the interval $[-\pi/2, \pi/2]$.

CS has also been applied to SFR to increase the range resolution [17]. As in pulse-Doppler radar, the target scene is discretized over an $N \times N$ delay-Doppler map [17]. The outputs of the phase detector (7) are then expressed, as in (15), where $\mathbf{y}$ is the vector of size P with the pth entry given by y_p, and $\mathbf{\Phi}$ is a DFT-based dictionary such that

$$\mathbf{\Phi}_{(p,(i-1)N+k)} = e^{j2\pi f_p \tau_i} e^{j\nu_k p\tau}. \tag{19}$$

The vector $\mathbf{s}$ is then recovered from $\mathbf{y}$ using CS techniques.

The approaches mentioned in this section may increase resolution by taking a large grid size N. However, bounds on N are not discussed, and it is not clear how large it can be. Denser grids reduce the sensitivity of the reconstruction to off-grid targets but increase the computational complexity by a square factor since the dictionaries contain N^2 atoms. More importantly, higher grid dimensions cause a significant increase to the coherence of the CS dictionary, which may degrade recovery performance.

The parameter space discretization, typically used in CS recovery techniques, assumes the targets' delays and Dopplers lie on the predefined grid. Several approaches have been proposed to solve off-grid issues, including grid refinement, which adjusts the detected delay-Doppler peak [32], parameter, perturbation-based, adaptive-sparse reconstruction techniques [21], and sensing matrix perturbation [57]. More references may be found in [58].

Fast time compression

In the works reviewed thus far, sampling and digital processing are still performed at the Nyquist rate. We next consider compressed radar that reduces sampling and processing rates.

Random sampling

Random sampling has been considered in SFR systems by selecting random measurements out of the Nyquist samples [21], [22]. The SFR approach of (1) is adopted in [22], with a random selection of M out of P pulses with different carriers. The sparse representation of the received signal used is a delay-Doppler shifted dictionary [21] similar to [15]. Consider the matrix $\mathbf{\Phi}$ whose ith column is given by

$$\mathbf{\Phi}_i = h(\mathbf{t} - \tau_i) \circ e^{j2\pi\nu_i \mathbf{t}}, \tag{20}$$

where $\mathbf{t}$ is the $N \times 1$ vector containing the sampling instants at the Nyquist rate, i.e., $t_i = i/B_h$, and $\circ$ is the Hadamard product operator. As in [15], the dictionary $\mathbf{\Phi}$ contains N^2 atoms. The Nyquist samples can then be expressed in the form (15), and the compressed samples $\mathbf{z}$ are given by

$$\mathbf{z} = \mathbf{A}\mathbf{y}, \tag{21}$$

where $\mathbf{A}$ is an $M \times N$ matrix, with $M < N$ constructed by randomly selecting M rows of the $N \times N$ identity matrix, which corresponds to the M-selected pulses.

In these approaches, processing is performed at a low rate; however, the random discarding of samples is difficult to implement in a sampling system for the purpose of effectively reducing the sampling rate. Furthermore, the large dictionary size discussed in the previous section remains an issue. Alternative practical radar systems using CS to reduce the sampling rate have been proposed and rely on two main techniques: uniform low-rate sampling using appropriate waveforms and analog preprocessing.

Uniform low-rate sampling

In [26], the authors consider SFR using multifrequency chirps, as described in (8). Low-rate samples are uniformly taken from the received signal (11) at rate $2\gamma\tau_r$, with $\tau_r = 2(r_{\max} - r_{\min})/c$, with γ being the common chirp rate. This results in the aliasing of the multiple sinusoids to baseband with random complex coefficients. Upon discretization of the target range, as denoted by $\mathbf{s}$, the low-rate samples may be modeled as

$$\mathbf{y} = \mathbf{A}\mathbf{s}. \tag{22}$$

Here, the kth column of the sensing matrix $\mathbf{A}$ is the FFT of the samples of (11) for a singular target at range bin k corresponding to a delay of $\tau_l = 2(r_{\min} - k\Delta)/c$, where Δ is the range-discretization step. The targets' delays are therefore recovered from the low-rate uniform sampling of the chirp waveforms.

Random demodulation

Many analog-to-information-conversion systems have been proposed to sample wideband signals at sub-Nyquist rates. Among them, the random demodulator (RD) [59], random-modulation preintegrator (RMPI) [60], and Xampling-based [29] systems have been used for radar applications. All three approaches consider pulse-Doppler radar.

The RD modulates the input signal using a high-rate sequence $p(t)$ created by a pseudorandom number generator, aliasing its frequency content. The random sequence used for demodulation is a square wave, which alternates between the levels ± 1 with equal probability. The mixed output is filtered by a bandpass filter $h_{bp}(t)$, with center frequency f_c and bandwidth $B_{CS} \ll B_h$, and sampled at a low rate, as shown in Figure 3(a).

The RD is adopted in [27] as the analog-mixing front end of a proposed quadrature-compressive-sampling (referred to as *QuadCS* by Liu et al.) system. The mixed and filtered output $y(t)$, shown in Figure 3, is given by

$$y(t) = \int_{-\infty}^{\infty} h_{bp}(\rho) p(t-\rho) x_R(t-\rho)\, \mathrm{d}\rho, \tag{23}$$

where $x_R(t)$ is defined as the real part of (7). The RD samples $y(t)$ at rate $f_s = 1/T_s = f_c/k$, with k an integer satisfying $k \le \lfloor f_c/2B_{CS} \rfloor$. The samples are fed to the quadrature-processing system [61], which extracts the baseband in-phase and quadrature (I and Q) components, respectively, of the radar echoes. As shown in [27], the complex samples of the RD output can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{24}$$

Here, $\mathbf{x}$ is a sparse vector that contains the complex amplitudes α_l at the corresponding delays τ_l, and the (m, p) element of the matrix $\mathbf{A}$ is given by

$$\mathbf{A}_{m,p} = \int_{-\infty}^{\infty} h_{bp}(\rho) e^{-j2\pi f_c \rho} p(mT_s - \rho) h(mT_s - p\tau - \rho) \mathrm{d}\rho. \tag{25}$$

The samples of P pulses are concatenated in a matrix $\mathbf{Y}$ such that each column corresponds to a pulse. The subsequent processing of the quadrature-compressive sampling, referred to as *compressive-sampling pulse Doppler*, is composed of a DFT step on the rows of $\mathbf{Y}$ that acts as an MF in slow time followed by a MF in each column, corresponding to the fast time.

The RMPI is a variant of the RD composed of a parallel set of RD channels driven by a common input, in which each RD uses a distinct pseudorandom binary sequence. A hardware RMPI-based prototype has been implemented in [43] that recovers radar pulses and estimates their amplitude, phase, and carrier frequency. In the next section, we discuss an alternative prototype with a different analog front end, which also recovers the targets' parameters from low-rate samples.

Note that the considerations behind waveform design for CS recovery in the approaches [15]–[17] presented in the previous section are similar to traditional radar requirements. The well-known ambiguity function (AF) impacts CS radar in a way that is similar to traditional radar systems. Indeed, the mutual coherence of the dictionary is linearly related to the highest sidelobe value of the AF [58], [62]. In contrast, we will see in the next section that the CS dictionary of the Xampling method is independent of the waveform, and MF is performed directly on the low-rate samples before parameter recovery.

Fast time Xampling

An alternative sub-Nyquist radar method is the Xampling-based system proposed in [30] and [40]. This approach, which may be used with any transmitted pulse shape, achieves the minimal sampling rate required for target detection, while providing optimal SNR.

The sub-Nyquist analog front end is composed of an ADC that filters the received pulse-Doppler signal (2) to predetermined frequencies before taking pointwise samples. These compressed samples, or "Xamples," contain the information needed to recover the desired signal parameters, i.e., the target' delay-Doppler map. To see this, note that the Fourier-series coefficients of the aligned frames $x_p(t + p\,\tau)$ are given by

$$c_p[k] = \frac{1}{\tau} H[k] \sum_{l=0}^{L-1} \alpha_l e^{-j2\pi k \tau_l/\tau} e^{-j\nu_l p\tau}, \quad 0 \le k \le N-1, \tag{26}$$

where $H[k]$ are the Fourier coefficients of the known transmitted pulse $h(t)$, and $N = B_h\tau$ is the number of Fourier samples. From (26), we see that the unknown parameters $\{\alpha_l, \tau_l, \nu_l\}_{l=0}^{L-1}$ are contained in the Fourier coefficients $c_p[k]$. We now show how the Fourier coefficients $c_p[k]$ may be obtained from low-rate samples of $x_p(t)$ and how the targets' parameters can then be recovered from $c_p[k]$ [30].

The received signals $x_p(t)$ exist in the time domain; thus, there is no direct access to $c_p[k]$. To obtain any arbitrary set of Fourier-series coefficients, the direct multichannel sampling scheme [63] illustrated in Figure 4 can be used. The analog input $x_p(t)$ is split into $k = \lfloor \kappa \rfloor$ channels, where, in each channel k_i with $i \in [0, K-1]$, it is mixed with the harmonic signal $e^{-j2\pi k_i t/\tau}$, integrated over the PRI duration, and then sampled. Xampling thus allows one to obtain an arbitrary set κ out of $N = \tau B_h$ frequency components from K pointwise samples of the received signal after appropriate analog preprocessing. An alternative Xampling method uses the sum-of-sincs filter described in [64]. This class of filters, which consists of a sum-of-sinc function

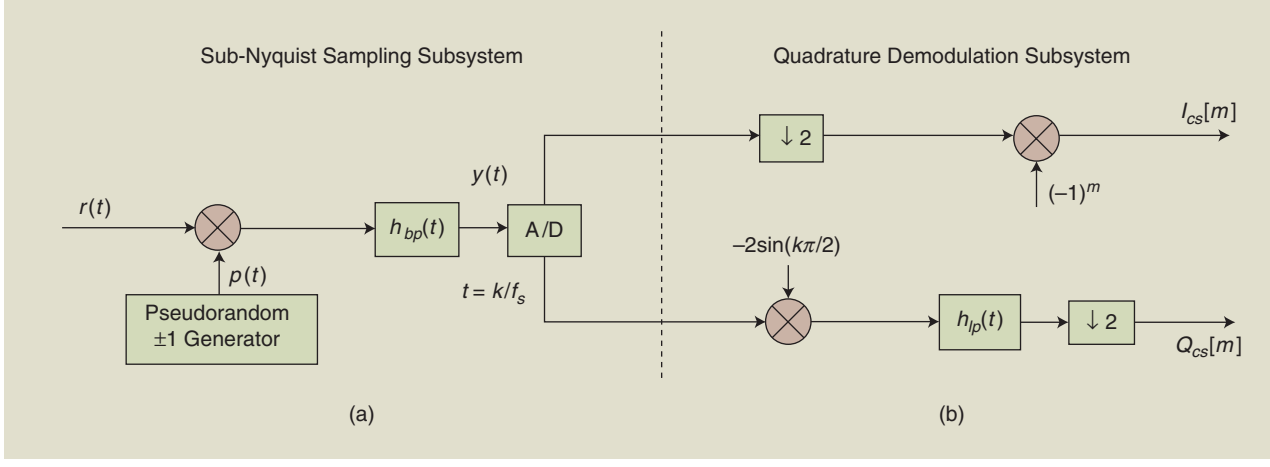


FIGURE 3. A quadrature-compressive-sampling implementation with (a) RD sampling followed by (b) quadrature demodulation [27].

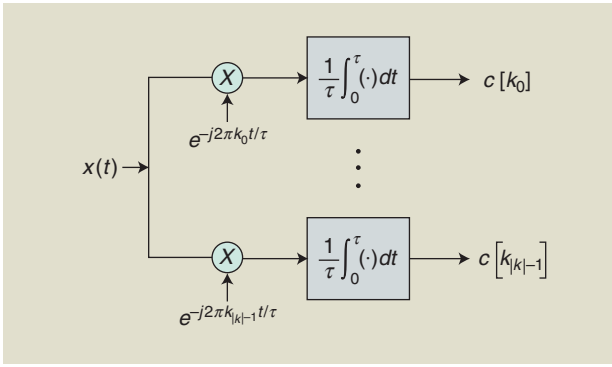


FIGURE 4. A multichannel direct sampling of the Fourier coefficients [63].

in the frequency domain, is a general sampling scheme for arbitrary pulse shapes.

A less expensive and more practical approach for the Fourier-series coefficients acquisition proposed in [40] is based on multiple bandpass filters and is adopted in the Xampling hardware radar prototype described in the next section. This system is composed of a few channels, with each sampling the content of a narrow frequency band of the received signal. Each channel thus yields a group of several consecutive Fourier coefficients. The multiple bandpass constellation has the advantage of acquiring the measurements over a wider frequency aperture. At the same time, it still allows for practical hardware implementation, as detailed in the next section. By widening the frequency aperture, a finer resolution grid may be employed during the recovery process. Moreover, empirical results show that highly distributed frequency samples provide better noise robustness [40]. However, widening the frequency aperture eventually requires increasing the number of samples K; otherwise, recovery performance may degrade. This tradeoff is observed in the experiments presented in [40].

Once a set of Fourier coefficients $c_p[k]$ has been acquired, the delays and Doppler frequencies can be recovered using different techniques. Doppler focusing [30], summarized in "Doppler Focusing," is one approach that has several advantages, as detailed next. This method uses target echoes from all of the pulses to generate a focused pulse at a specific Doppler frequency. It then jointly recovers the delay-Doppler map by reducing the detection problem to a one-dimensional, delay-only estimation. Performing the Doppler focusing operation in frequency results in computing the DFT of the coefficients $c_p[k]$ in the slow time domain:

$$\begin{aligned}\Psi_\nu[k] &= \sum_{p=0}^{P-1} c_p[k]\, e^{j\nu p\tau} \\ &= \frac{1}{\tau} H[k] \sum_{l=0}^{L-1} \alpha_l e^{-j2\pi k\tau_l/\tau} \sum_{p=0}^{P-1} e^{j(\nu-\nu_l)p\tau}. \end{aligned} \tag{27}$$

Note that $\Psi_\nu[k]$ is the Fourier series of $\Phi(t,\nu)$, defined in (S11), with respect to t. Following the same argument as in (S12), we have

$$\Psi_\nu[k] \approx \frac{P}{\tau} H[k] \sum_{l\in\Lambda(\nu)} \alpha_l e^{-j2\pi k\tau_l/\tau}. \tag{28}$$

The resulting equation (27) is a standard delay-estimation problem for each ν and may be solved using multiple techniques [10]. However, improved performance can be obtained by jointly processing the sequences $\{\mathbf{\Phi}_\nu[k]\}$ for different values of ν. Thus, instead of searching separately for each of the delays $\tau_l, l\in\Lambda(\nu)$, the L delays are estimated by jointly processing overall Doppler frequencies.

A particularly convenient method in this case is to employ a matching pursuit-type approach in which the strongest overall peak ν, assuming a single delay, is first found:

$$(\hat{\tau}_l, \hat{\nu}_l) = \underset{\tau_l,\nu_l}{\operatorname{argmax}} \left| \sum_{k\in\kappa} \Psi_{\nu_l}[k]\, e^{j2\pi k\tau_l/\tau} \right|. \tag{29}$$

Once the optimal values $\hat{\tau}_l$ and $\hat{\nu}_l$ are determined, their influence is subtracted from the focused sub-Nyquist samples as

$$\Psi'_\nu[k] = \Psi_\nu[k] - \frac{1}{\tau}\hat{\alpha}_l e^{-j2\pi k\hat{\tau}_l/\tau} \sum_{p=0}^{P-1} e^{j(\nu-\hat{\nu}_l)p\tau}, \tag{30}$$

where

$$\hat{\alpha}_l = \frac{\tau}{P|\kappa|} \sum_{k\in\kappa} \Psi_{\hat{\nu}_l}[k]\, e^{j2\pi k\hat{\tau}_l/\tau}. \tag{31}$$

The same operations are performed iteratively to find all of the desired L peaks. This approach does not require discretization of the targets' parameters, and these are recovered over the continuous domain from a minimal number of samples.

In practice, the search for peaks can be limited to a grid, which enables all of the computations to be carried out using simple FFT operations. Suppose we limit ourselves to the Nyquist grid, i.e., the grid defined by the Nyquist resolution so that $\tau_l/\tau = s_l/N$, where s_l is an integer satisfying $0 \le s_l \le N-1$. Then, (26) is approximately written in vector form as

$$\mathbf{\Psi}_\nu = P\mathbf{H}\mathbf{F}_N^K \mathbf{a}_\nu, \tag{32}$$

where $\mathbf{\Psi}_\nu = [\Psi_\nu[k_0]\ldots\Psi_\nu[k_{K-1}]], k_i\in\kappa$ for $0\le i\le K-1$, $\mathbf{H}$ is a diagonal matrix that contains the Fourier coefficients $H[k]$ of the transmitted waveforms, and $\mathbf{F}_N^K$ is the partial-Fourier matrix that contains the K rows of the $N\times N$ Fourier matrix indexed by k. The entries of the L-sparse vector $\mathbf{a}_\nu$ are the values α_l at the indices s_l for the Doppler frequencies ν_l in the "focus zone," i.e., $|\nu-\nu_l| < \pi/P\tau$. The P equations (32) are simultaneously solved using CS-based algorithms, which, during each iteration, the maximal projection of the observation vectors onto the measurement matrix is retained [30].

Some results comparing different configurations of low-rate sampling and processing are shown in Figure 5 [30]. The recovery performance of the classic processing applied to Nyquist samples is presented as a baseline. Sub-Nyquist approaches, performed at 1/10 of the Nyquist rate, include the same classic processing applied to sub-Nyquist samples, a two-stage CS

recovery method that performs delay and Doppler estimation in parallel, separately [30], and Doppler focusing. It is clearly seen that Doppler focusing applied to random Fourier coefficients, which are widely distributed with high probability leading to a wide aperture, outperforms other sub-Nyquist approaches. The use of consecutive coefficients yields small aperture and poor resolution.

The minimal number of samples required for the perfect recovery of $\{\alpha_l, \tau_l, \nu_l\}$ with L targets in a noiseless environment is $4L^2$, with at least $K \geq 2L$ samples per pulse and at least $P \geq 2L$ pulses.

The Xampling approach has several advantages. First, it recovers the targets' parameters directly from the low-rate samples without requiring sampling at the Nyquist rate. Second, previous CS-based methods typically impose constraints on the radar transmitter, which are not needed here. Indeed, as may be seen in (20) and (25), the CS dictionary depends on samples of the waveform $h(t)$, such that the mutual coherence of the dictionary is linearly related to the highest sidelobe value of the AF [58]. In contrast, the CS dictionary of the Xampling method is independent of the waveform, as shown in (32). Third, in the presence of additive white noise, Doppler focusing achieves an increase in SNR by a factor of P (a detailed analysis may be found in [30]). In addition, this approach can operate at the smallest possible sampling rate for recovering the targets' parameters, as derived in [30]. The minimal number of samples required for the perfect recovery of $\{\alpha_l, \tau_l, \nu_l\}$ with L targets in a noiseless environment is $4L^2$, with at least $K \geq 2L$ samples per pulse and at least $P \geq 2L$ pulses. The

Doppler Focusing

Doppler focusing is a processing technique, suggested in [30], which uses target echoes from different pulses to create a superimposed pulse focused at a particular Doppler frequency. This method allows for joint delay-Doppler recovery of all targets present in the illuminated scene. It results in an optimal signal-to-noise ratio (SNR) boost and may be carried out in the frequency domain, thus enabling sub-Nyquist sampling and processing with the same SNR increase as a matched filter.

The output of Doppler processing can be viewed as a discrete equivalent of the following time shift and modulation operation on the received signal:

$$\begin{aligned}\Phi(t,\nu) &= \sum_{p=0}^{P-1} x_p(t+p\tau)\, e^{i\nu p\tau} \\ &= \sum_{l=0}^{L-1} \alpha_l h(t-\tau_l) \sum_{p=0}^{P-1} e^{i(\nu-\nu_l)p\tau}. \end{aligned} \tag{S11}$$

Consider the sum $g(\nu \mid \nu_l) = \left|\sum_{p=0}^{P-1} e^{i(\nu-\nu_l)p\tau}\right|$. For any given ν, targets with Doppler frequencies ν_l in a bandwidth of $2\pi/P\tau$ around ν will achieve a coherent integration and an SNR increase of approximatively P. On the other hand, since the sum of P equally spaced points covering the unit circle is generally close to zero, targets with ν_l not "in focus" will roughly cancel out. In summary, we have that

$$g(\nu \mid \nu_l) = \sum_{p=0}^{P-1} e^{i(\nu-\nu_l)p\tau} \approx \begin{cases} P & |\nu-\nu_l| < \pi/P\tau \\ 0 & |\nu-\nu_l| \geq \pi/P\tau, \end{cases} \tag{S12}$$

as shown in Figure S1.

We may therefore estimate the sum of exponents in (S11) as

$$\Phi(t,\nu) \approx P \sum_{l\in\Lambda(\nu)} \alpha_l h(t-\tau_l), \tag{S13}$$

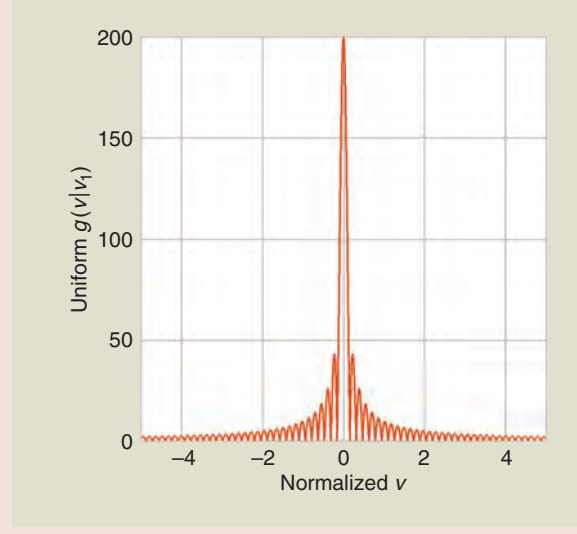


FIGURE S1. The sum of exponents $|g(\nu \mid \nu_l)|$ for $P = 200$, $\tau = 1$s, and $\nu_l = 0$.

where $\Lambda(\nu) = \{l : |\nu - \nu_l| < \pi/P\tau\}$. In other words, the sum is only over the targets whose Doppler shifts are in the interval $|\nu - \nu_l| < \pi/P\tau$.

For each Doppler frequency ν, $\Phi(t,\nu)$ represents a standard pulse-stream model in which the problem is to estimate the unknown delays. Thus, using Doppler focusing, the two-dimensional delay-Doppler recovery problem is reduced to delay-only estimation for a small range of Doppler frequencies, with increased SNR by a factor of P [10]. The Xampling radar of [30] performs Doppler focusing directly on the low-rate samples in the frequency domain, allowing for joint Doppler-delay recovery from the "Xamples."

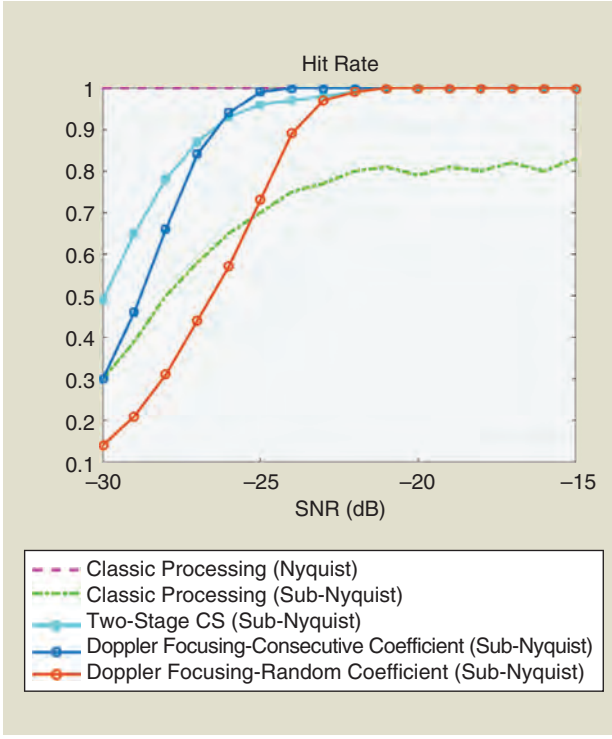


FIGURE 5. The hit rate of classic processing, two-stage CS recovery, and Doppler focusing for a fixed false alarm rate. A *hit* is defined as a delay-Doppler estimate circumscribed by an ellipse around the true target position in the time–frequency plane, with the axes equivalent to ± 3 times the time and frequency Nyquist bins. The two-stage CS recovery separates the delay and Doppler estimation, performing them in parallel [30]. The sub-Nyquist sampling rate was 1/10 of the Nyquist rate [30].

Doppler focusing approach achieves this minimal number of samples. Finally, Doppler focusing is able to deal with certain models of clutter and target dynamic range by adding a simple windowing operation in the sum (27) and by prewhitening in frequency [54].

The Xampling radar was implemented in hardware, as described in the next section, demonstrating real compressed radar capabilities. The hardware prototype is built from off-the-shelf components, which are bandpass filters and low-rate samplers, leading to low hardware complexity.

Hardware prototype

Xampling is used in combination with Doppler focusing in the sub-Nyquist prototype of [30], [40], which demonstrates radar reception at sub-Nyquist rates. The input signal simulates reflections from arbitrary targets and is corrupted by additive noise and clutter. The radar receiver implements the multichannel topology described in the previous section and samples a signal with Nyquist rate of 30 MHz with a compression factor of 30. Hardware experiments demonstrate the feasibility of detecting targets from the low-rate samples of an analog radar signal using standard radio-frequency (RF) hardware [30], [40]. Typical experiment results are shown in Figure 6, which depicts the input signal, the low-rate samples, and the original and recovered delay-Doppler maps, including close targets, both in terms of range and velocity.

At the heart of the receiver lies the Xampling-based ADC, which performs analog prefiltering of the signal before taking pointwise samples. A multiple bandpass-sampling approach

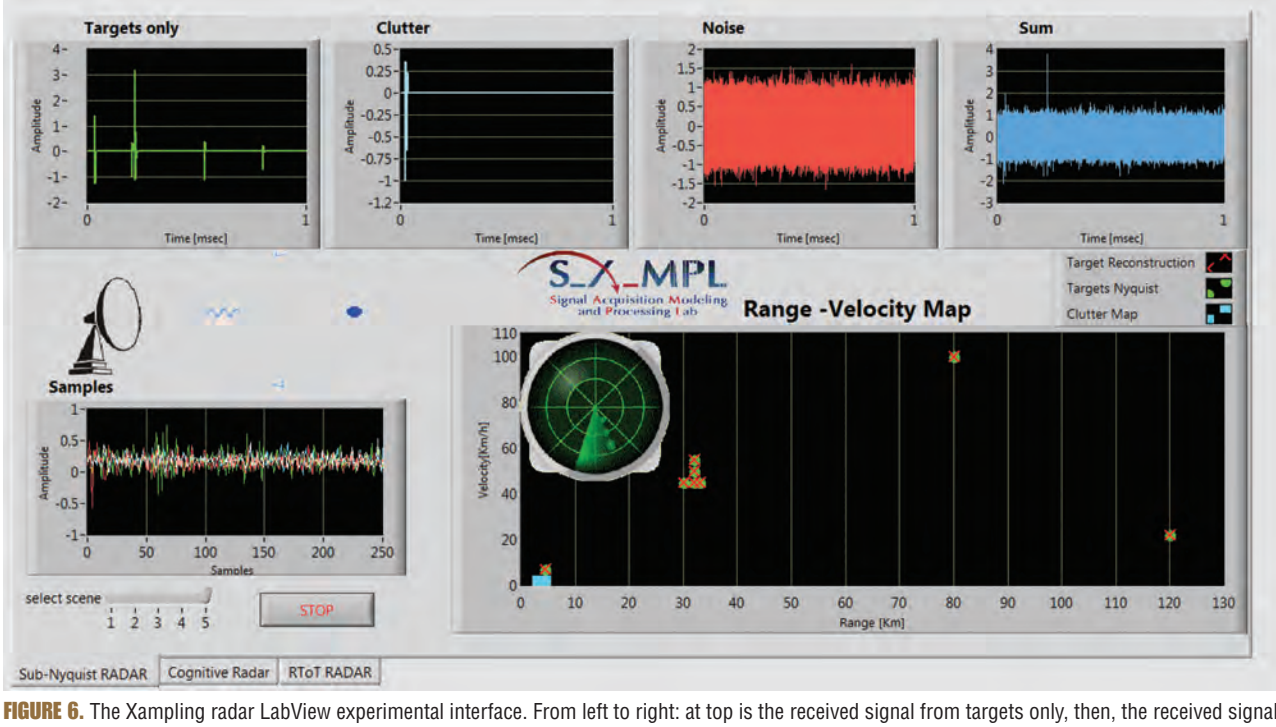


FIGURE 6. The Xampling radar LabView experimental interface. From left to right: at top is the received signal from targets only, then, the received signal from clutter, noise, and overall received signal $x_p(t)$. At the bottom are the sub-Nyquist samples of the four channels at 1/30 of the Nyquist rate, then, the true and recovered delay-Doppler maps. All of the targets (including close targets both in range and in velocity) are correctly detected.

with four channels is adopted. Each channel is composed of a crystal filter with a bandwidth of 80 KHz and extremely narrow transition bands and then sampled at a rate of 250 kHz. The front-end samples four distinct bands of radar–signal spectral content, yielding 320 Fourier coefficients after digital processing with a total sampling rate of 1 MHz. The samples are fed into the chassis controller, and a MATLAB function is launched that computes the 320 Fourier coefficients via FFT, composed of four groups of 80 consecutive Fourier coefficients. These are then used for digital recovery of the delay-Doppler map using the Doppler focusing reconstruction algorithm.

At the heart of the receiver lies the Xampling-based ADC, which performs analog prefiltering of the signal before taking pointwise samples.

The experimental setup is based on National Instrument (NI) PXI-series equipment that is used to synthesize a radar environment and ensure system synchronization. The entire component ensemble wrapped in the NI chassis as well as the analog receiver board are depicted in Figure 7. Additional information regarding the system's configuration and synchronization can be found in [40].

To demonstrate target detection from low-rate samples, the Applied Wave Research (AWR) software is used to simulate the radar scenario, including pulse transmission and accurate power loss due to wave propagation in a realistic medium. AWR software provides a computer-based environment for designing hardware for use with wireless and high-speed digital products. It is used for RF, microwave, and high-frequency analog circuits and system design. A large variety of scenarios, consisting of different targets' parameters, i.e., delays, Doppler frequencies, and amplitudes, are examined in [30] and [40]. An arbitrary waveform generator module produces an analog signal that is amplified and routed to the radar receiver board. The received radar waveform is contaminated with noise and clutter, showing the capabilities of the Xampling receiver to deal with these [30], [40], [54]. The Nyquist rate of the signal is 30 MHz, so that sampling at 1 MHz corresponds to a fast time compression factor of 30.

Slow time compression

Most works on CS radar focus on compression in the fast time domain, reducing the number of samples per pulse below the Nyquist rate. As we have seen, using appropriate CS techniques allows for preserving the range resolution while operating in low-rate regimes by breaking the link between bandwidth and sampling rate. This is illustrated in Figure 5, in which Doppler focusing is shown to achieve the same hit rate as classic processing above a certain SNR, and in Figure 6, where close targets are seen to be correctly recovered despite sampling at 3.3% of the Nyquist rate. We will now see that compression may be similarly performed in the slow time domain, as demonstrated in [65], where the number of transmitted pulses is reduced without decreasing Doppler resolution.

Nonuniform pulse Doppler

The resolution in Doppler frequency in standard processing is governed by the number of transmitted pulses P. More precisely, it is equal to $2\pi/P\tau$. However, a large P leads to large CPI and long time on target. Slow time compression breaks the relation between CPI and time on target. To that end, $M < P$ pulses are sent nonuniformly over the entire CPI $P\tau$, implementing nonuniform time steps between the pulses [65]. This way, the same CPI is kept, but a smaller number of pulses is transmitted, thereby reducing power consumption. In addition, the periods of time in which no pulse is transmitted in a certain direction can be exploited to send pulses in other directions. This allows the radar to scan several directions at the same time and obtain the corresponding delay-Doppler maps in a single CPI. However, at the same time, this reduces SNR because fewer pulses are transmitted in every direction.

Consider a nonuniform pulse-Doppler radar such that the pth pulse is sent at time $m_p\tau$, where $\{m_p\}_{p=0}^{M-1}$ is an ordered set of integers satisfying $m_p \geq p$. In this case, (1) becomes

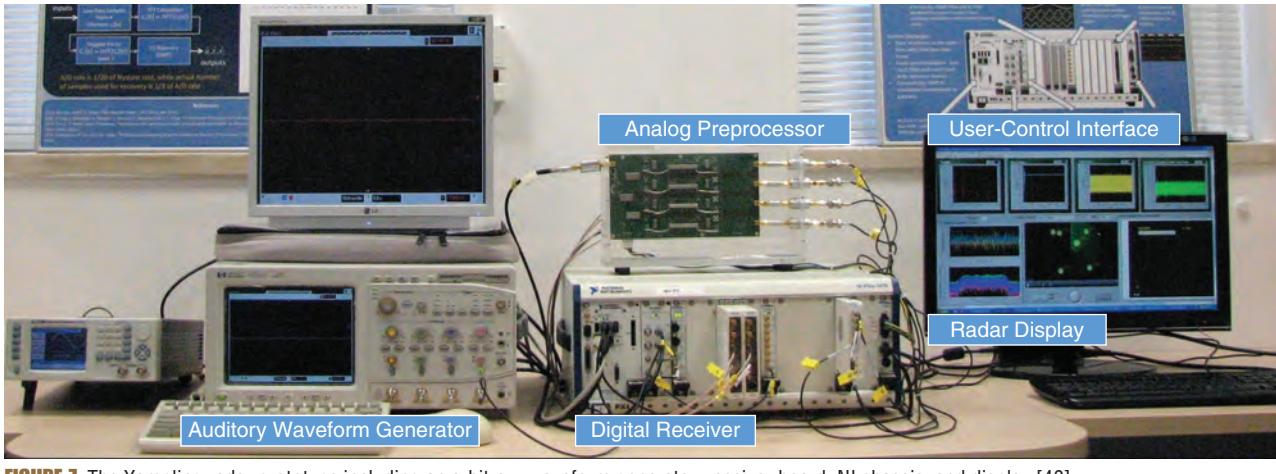


FIGURE 7. The Xampling radar prototype including an arbitrary waveform generator, receiver board, NI chassis, and display [40].

$$x_T(t) = \sum_{p=0}^{M-1} h(t - m_p\tau), \quad 0 \le t \le P\tau, \tag{33}$$

and the received frames (4) are written as

$$x_p(t) = \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l - m_p\tau) e^{-j\nu_l m_p \tau}, \quad 0 \le t \le P\tau. \tag{34}$$

The same Xampling-based method in [30] is used to obtain the Fourier coefficients $c_p[k]$ of the received pulses. Suppose we limit ourselves to the Nyquist grid, as previously mentioned, so that $\tau_l/\tau = s_l/N$, where s_l is an integer satisfying $0 \le s_l \le N-1$, and $\nu_l\tau = 2\pi r_l/M$, where r_l is an integer in the range $0 \le r_l \le M-1$. Similar to the derivations in the previous section, we can write the Fourier coefficients $c_p[k]$in matrix form [65] as

$$\mathbf{X} = \mathbf{H}\mathbf{F}_N^K \mathbf{A} (\mathbf{F}_P^M)^T, \tag{35}$$

where $\mathbf{H}$ is a diagonal matrix that contains the Fourier coefficients $H[k]$. The partial-Fourier matrix $\mathbf{F}_M^P$ contains M rows from the $P \times P$ Fourier matrix, indexed by the values of the transmitted pulses m_p, $1 \le p \le M$; when sampling at the Nyquist rate, $K = N$ and $\mathbf{F}_N^K$ become the standard $N \times N$ Fourier matrix. Similarly, when considering uniformly spaced pulses, $M = P$ and $\mathbf{F}_P^M$ are the standard $P \times P$ matrix. The goal is to recover the sparse matrix $\mathbf{A}$ that contains the values α_l at the L indices $\{s_l, r_l\}$ from the Fourier coefficients matrix $\mathbf{X}$.

CS matrix-recovery algorithms are directly applicable to (35) by extending CS techniques presented in vector form, such as orthogonal matching pursuit or the fast iterative-shrinkage-thresholding algorithm [10], [12] to matrix settings [66]. Alternatively, instead of solving the matrix problem of (35), we can apply the Doppler focusing operation [30] described in "Doppler Focusing." As illustrated in Figure 8, the approximation from (S12) may still be applied in the nonuniform case, where $m_p \ge p$. Therefore, we can rewrite the Fourier coefficients from (27) by replacing p with m_p for the nonuniform case. These may then be approximately expressed in vector form as in (32) and recovered as previously described. It is shown in [65] that the minimal number of nonuniform pulses required to recover the Doppler frequencies of L targets is identical to the uniform case, that is, two L pulses.

Hardware simulation

The transmission of nonuniform pulses has been implemented in the Xampling prototype [40]. Recall that the received signal has a bandwidth of 30 MHz and is sampled at the rate of 1 MHz. To this fast time compression, we now add compression in the slow time domain. In the hardware simulation, $P = 50$ pulses over a CPI of $MP\tau = 2.5$ s are considered. Half of the pulses, i.e., $M = 25$, chosen at random, are sent in one direction, while the other half are sent in a second direction. Two delay-Doppler maps are then simultaneously recovered during a single CPI, as shown in Figure 9. Both of the maps are fully recovered, as previously mentioned, from compressed samples in both the fast and slow time domains.

Range-velocity ambiguity resolution

As presented thus far, targets are traditionally assumed to lie in the radar-unambiguous range-velocity region. For a given PRI τ, the maximum unambiguous range is $r_{\max} = c\tau/2$, and the maximum unambiguous velocity is $\dot{r}_{\max} = \lambda/(4\tau)$, where λ is the radar wavelength. When the target range and velocity intervals of interest are large, traditional pulse-Doppler radar

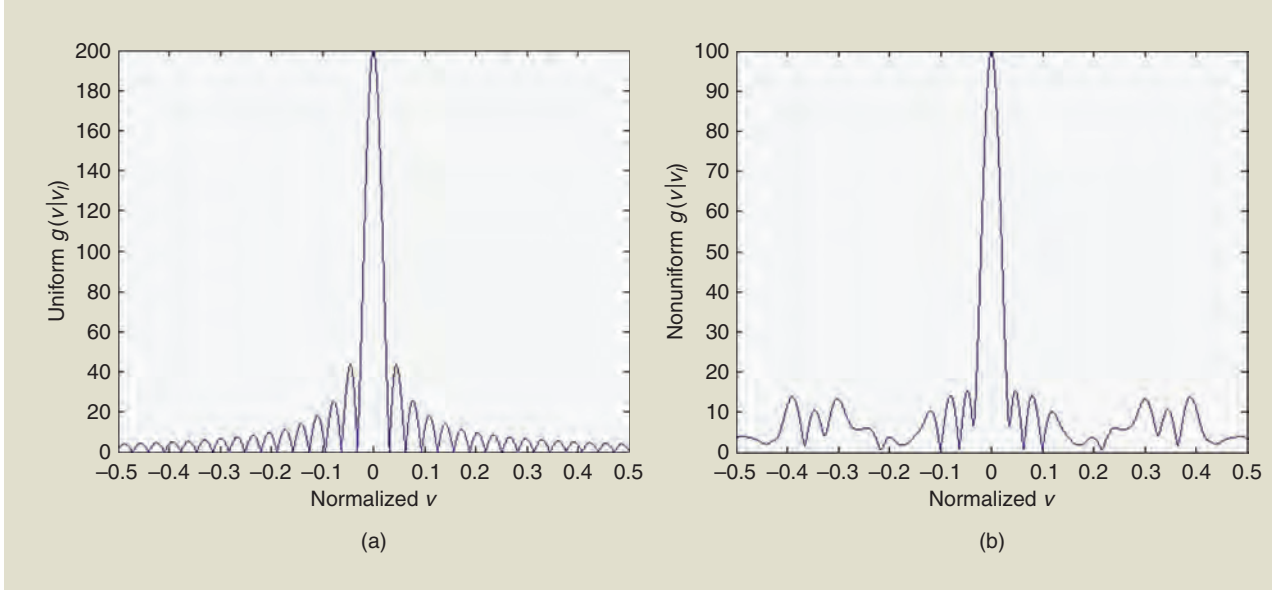


FIGURE 8. The sum of exponents $|g(\nu \mid \nu_l)|$ for $M = 200$, $\tau = 1$s, and $\nu_l = 0$ in the (a) uniform and (b) nonuniform cases. In the nonuniform case, $P = 100$ pulses are chosen uniformly at random [65].

systems suffer from the so-called Doppler dilemma [67], a tradeoff between range and velocity ambiguity, whose product is limited to $r_{\max}\dot{r}_{\max} = c\lambda/8$.

Several techniques have been proposed over the years to mitigate the range-velocity ambiguity by increasing either of these parameters. Two main PRF variation-based methods are staggered PRFs and multiple PRFs (MPRFs). Staggered PRFs are used to raise the first blind speed $\dot{r}_{\max}$ significantly without degrading the unambiguous range [7]. Pulse-to-pulse stagger varies the PRF from one pulse to the next, achieving increased Doppler coverage [68]. The main disadvantage of this approach is that the data correspond to a nonuniformly sampled sequence, making it more difficult to apply coherent Doppler filtering [7]. In addition, clutter cancellation becomes more challenging, and the sensitivity to noise increases [69]; therefore, MPRF techniques are typically preferred. We now review some of the MPRF-based methods and then present a Xampling approach that solves the delay-Doppler ambiguity using phased-coded-transmit pulses.

Ambiguity resolution is typically achieved by searching for coincidence between either unfolded Doppler or delay estimates for each PRF.

MPRF

The MPRF approach transmits several pulse trains, each with a different PRF. Ambiguity resolution is typically achieved by searching for coincidence between either unfolded Doppler or delay estimates for each PRF. A popular approach, adopted in [70], relies on the Chinese remainder theorem [5] and uses two PRFs, such that the numerator and denominator of the ratio between them are prime numbers. The ambiguous velocities are computed for each train i as

$$\hat{\dot{r}}_{i,k} = \hat{\dot{r}}_{i,0} + k\frac{\lambda}{2\tau}, \quad k \in \mathbb{Z}, \tag{36}$$

where $\hat{\dot{r}}_{i,0}$ is the velocity estimate within the unambiguous velocity interval $(-\dot{r}_{\max}, \dot{r}_{\max}]$. Congruence between these are found by an exhaustive search, so that all $\hat{\dot{r}}_{i,k}$ fall within a small, interval, or correlation bin. The resulting velocity estimate is computed by averaging overall $\hat{\dot{r}}_{i,k}$. Assuming $T = 2$ pulse trains with PRFs and ratio $\tau_1/\tau_2 = m/n$, where m and n are relative prime numbers, the expanded velocity interval is of size $m\lambda/2\tau_1 = n\lambda/2\tau_2$. However, in this approach, a small range error on a single PRF can cause a large error in the resolved range with no indication that this has happened [71].

A clustering algorithm proposed in [71] implements the search for a matching interval by computing average distances to cluster centers. The average squared error is defined as

$$C(k) = \sum_{i=1}^{T} |\hat{r}_{i,k} - \bar{r}_k|^2, \quad k = 0, \ldots, r_{\text{amb}}/r_{\max}, \tag{37}$$

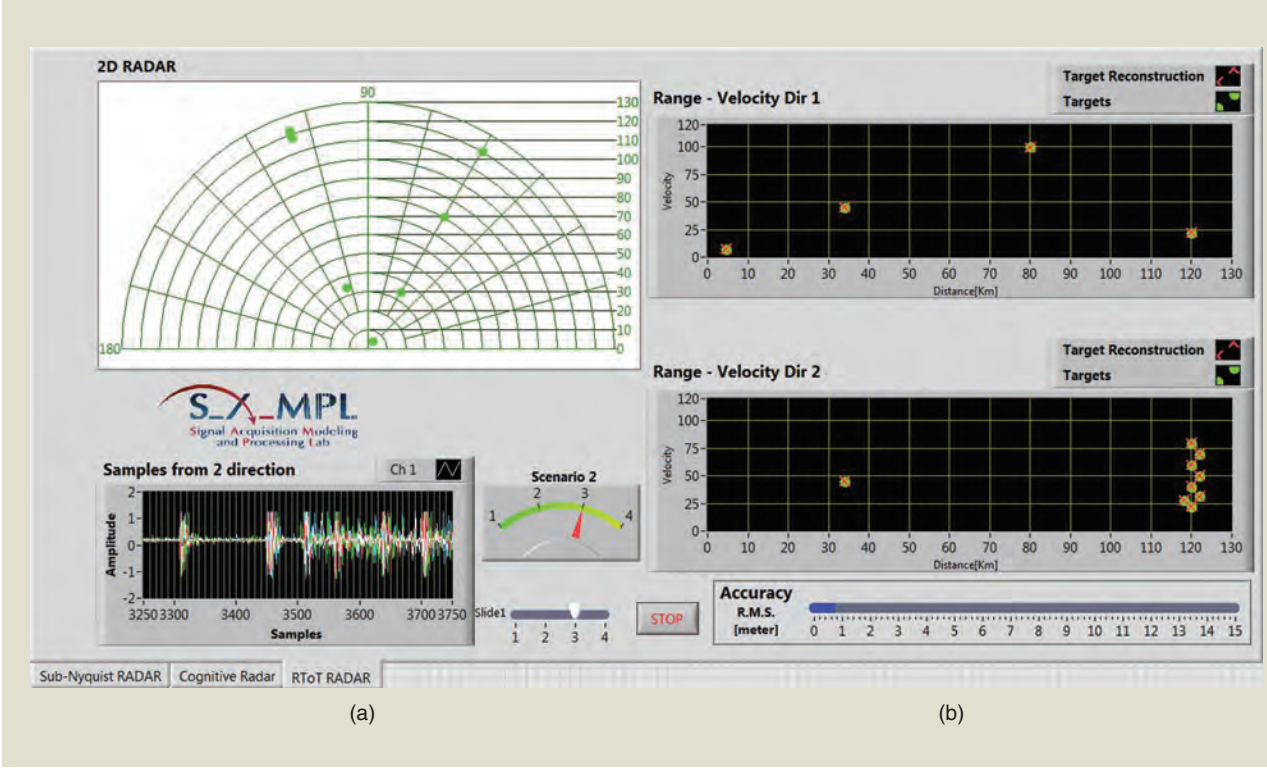


FIGURE 9. The experimental interface of the Xampling radar, with both fast and slow time compression. (a) The true targets' ranges in two directions (top) and superposed low-rate samples from both directions (bottom). (b) The range-velocity map of true and recovered targets in both directions [65].

where $\bar{r}_k$ is the median value of the T ranges with index k and r_{amb} is the maximal ambiguous range. The best cluster occurs at the value of k, where $C(k)$ is minimized. This happens when all of the ambiguous ranges are unfolded correctly, and, hence, all of the range estimates have nearly the same range. This technique still requires an exhaustive search over clusters and does not process the samples jointly, thereby decreasing the SNR.

Random PC has been adopted in polarimetric weather radars, which exploit the inherent random phase between pulses of the popular magnetron transmitters.

Phased-coded pulses

A random-pulse, phase-coding (PC) approach is adopted in [72] to increase the range-unambiguous region, while preserving that of the Doppler frequency and using a single PRF. A similar technique may be used to increase the Doppler-frequency-unambiguous region. Random PC has been adopted in polarimetric weather radars, which exploit the inherent random phase between pulses of the popular magnetron transmitters. In this context, PC mitigates out-of-trip echoes [73]. The approach of [72] introduces a random phase, which differs from pulse to pulse. The joint processing of received signals from all of the trains is the key to range ambiguity resolution.

The pulse-Doppler radar transceiver sequentially transmits one modulated pulse train, consisting of P equally spaced pulses. For $0 \le t \le P\tau$, the transmitted signal is given by

$$x_T(t) = \sum_{p=0}^{P-1} h(t - p\tau) e^{jc[p]}, \tag{38}$$

where $c[p]$ is uniformly distributed in the interval $[0, 2\pi)$ and represents the phase shift of the pth pulse.

As opposed to the common assumption in traditional radar, the targets' time delays $\tilde{\tau}_l$ are not assumed to lie in the unambiguous time region, i.e., less than the PRI τ, but rather in the ambiguous range $\tilde{\tau}_l \in [0, Q\tau)$, where $Q < P$ is the ambiguous factor defined by the targets' maximal range. For convenience, the delay $\tilde{\tau}_l$ is decomposed into its integer part (the ambiguity order) $q_l\tau$ and fractional part (the folded or reduced delay) τ_l such that

$$\tilde{\tau}_l = \tau_l + q_l\tau, \tag{39}$$

where $0 \le q_l \le Q - 1$ is an integer and $0 \le \tau_l < \tau$.

The received signal is then

$$x_R(t) = \sum_{p=0}^{P-1} \sum_{l=0}^{L-1} \alpha_l h(t - \tilde{\tau}_l - p\tau) e^{-j2\pi\nu_l(p+q_l)\tau} e^{jc[p]}, \tag{40}$$

for $0 \le t < (P + Q)\tau$. The main difference with traditional pulse-Doppler radar, aside from the coded phase, is that the PRI index in the Doppler shift term is $p + q_l$, rather than the pulse index p.

The Fourier series of the received signal (40) can be written in matrix form, similarly to (35), and recovered using matrix CS recovery techniques [72]. The minimal number of samples per pulse allowing recovery of $\mathbf{X}$ with high probability is found to be $K > 2L$, and the minimal number of pulses P is $2L + Q + 2$. This method resolves a maximum unambiguous range $r_{\max} = cQ\tau/2$, while preserving the maximum unambiguous velocity $\dot{r}_{\max} = \lambda/(4\tau)$, thereby increasing their product $r_{\max}\dot{r}_{\max}$ by a factor of Q, under the prior mentioned conditions on the number of samples and pulses.

This approach has three main advantages. First, it improves the delay estimation with respect to the MPRF methods since it preserves the resolution of traditional pulse-Doppler radar, i.e., $1/B_h$, while increasing the unambiguous delay region to $Q\tau$. Second, it increases the SNR by jointly processing the samples from all of the pulse trains, rather than matching the estimated parameters from each pulse processed separately. Finally, it provides a systematic delay-Doppler-recovery method that does not involve an exhaustive search. From a practical point of view, this approach does not require the use of several pulse trains with different PRFs, thus simplifying hardware implementation.

Cognitive radar and spectrum sharing

Recently, the concept of cognitive radar (CR) [74], inspired by the echo-location system of a bat, has been presented as a natural next step for traditional radar. The cognition property requires adaptive transmission and reception capabilities, i.e., both the transmitter and receiver are able to dynamically adjust to the environment conditions. Many interpretations of this idea have been proposed. We focus on one aspect of cognition, the dynamic and flexible adaptation to the spectral environment, which allows for spectrum sharing between communication and radar systems [36]–[38]. The interest in these spectrum sharing radars is largely due to electromagnetic spectrums being a scarce resource, with most services having a need for a greater access to it.

The spectrum sharing solution proposed in [39] capitalizes on the cognitive abilities of the radar system. It is shown how compressed radars may be adapted to allow for spectral coexistence between communication and radar signals and flexibility of the radar transmission. This demonstrates that, beyond increasing resolution and realizing compression in the time, frequency, and spatial domains, compressed radars have the potential to enable otherwise challenging technologies.

Spectral adaptive transmission

In previous works that implement fast time compression, e.g., Xampling radar [30], [40], the transmitter broadcasts a wideband signal, which reflects off the targets and propagates back to the receiver. The received signal is then filtered before sampling so that only the content of a few narrow bands is sampled and processed. These works only deal with the reception

side of the radar, providing sampling and processing techniques that can be used with any traditional radar transmitter. However, for broadband frequency occupation and power savings, only the narrow frequency bands that are to be sampled may be transmitted [31], [39]. This will not affect any aspect of the processing since the received signal is preserved in the bands of interest. In fact, since all the signal power is concentrated in the processed bands, the SNR increases, and the detection performance improves [75].

Let $\tilde{H}(f)$ be the CTFT of the new transmitted radar pulse,

$$\tilde{H}(f) = \begin{cases} H(f) & f \in \left[f_r^i - B_r^i/2,\, f_r^i + B_r^i/2\right] \text{ for } 1 \le i \le N_b \\ 0 & \text{else,} \end{cases} \tag{41}$$

where N_b is the number of filtered bands and B_r^i and f_r^i are the bandwidth and center frequency of the ith band, respectively. Obviously, the computation of the relevant Fourier coefficients $c_p[k]$ from (25) will not change. Therefore, the recovery methods presented in the "Fast Time Xampling" section are applicable here as well.

The concept of transmitting only a few subbands that the receiver processes is one way to formulate a frequency-agile CR in terms of its ability to adapt to spectral demands. Complying with CR requirements, the support of the subbands varies with time to allow for dynamic and flexible adaptation. Such a system also enables the radar to disguise the transmitted signal as an electronic countermeasure or to cope with crowded spectrums by using a smaller, interference-free portion, as further discussed in the following section.

Application to spectrum sharing

The unhindered operation of a radar that shares its spectrum with communication systems has captured a great deal of attention within the operational radar community in recent years [36]–[39]. Recent research programs in spectrum sharing radars include the Enhancing Access to the Radio Spectrum project by the National Science Foundation [38] and the Shared Spectrum Access for Radar and Communication (SSPARC) program [37], [76], initiated by the Defense Advanced Research Projects Agency.

A variety of system architectures have been proposed for spectrum sharing radars, and most place an emphasis on optimizing the performance of either radar [77] or communication [78] while ignoring the performance of the other. In nearly all cases, the real-time exchange of information between radar and communication hardware has not yet been integrated into the system architectures. Exceptions to this are automotive solutions in which the same waveform is used for both target detection and communication [79].

In a similar vein, the sub-Nyquist, CR-based approach from [39] incorporates the handshaking of spectral information between the two systems. The CR configuration is key to spectrum sharing since the radar transceiver can adapt its transmission to available bands, thus achieving coexistence with communication signals. Suppose the set of all frequencies of the available common system spectrum is given by $\mathcal{F}$.The communication and radar systems occupy the subsets $\mathcal{F}_C$ and $\mathcal{F}_R$ of $\mathcal{F}$, respectively. The goal is to design the radar waveform and its support $\mathcal{F}_R$, conditional on the fact that the communication occupies frequencies $\mathcal{F}_c$, which are unknown to the radar transceiver [39]. To detect the bands left vacant by the communication signals, spectrum sensing needs to be performed over a large bandwidth. Such a task has recently received tremendous interest in the communication community, which faces a bottleneck in terms of spectrum availability. To increase the efficiency of spectrum managment, a dynamic opportunistic exploitation of temporarily vacant spectral bands by secondary users has been considered, called *cognitive radio* (*CRo*) [80], [81].

To increase the efficiency of spectrum management, a dynamic opportunistic exploitation of temporarily vacant spectral bands by secondary users has been considered, called *cognitive radio*.

A spectrum sharing paradigm using Xampling techniques, the spectral coexistence via Xampling (SpeCX) system [39] is composed of a sub-Nyquist CRo receiver [81] to detect the occupied communication bands so that the radar transmitter may subsequently exploit the spectral holes. In this setting, the received signal at the communication receiver is given by

$$x(t) = x_C(t) + x_R(t), \tag{42}$$

where $x_R(t) = r_{Tx}(t) + r_{Rx}(t)$ is the radar signal sensed by the communication receiver, composed of the transmitted and received radar signals. The goal is therefore to recover the support of $x_C(t)$, given the known support of $x_R(t)$, which is shared by the radar transmitter with the communication receiver. This can be formulated as a sparse recovery with partial-support knowledge, studied under the framework of a modified CS [82].

Once $\mathcal{F}_C$ is identified, the communication receiver provides a spectral map of occupied bands to the radar. Equipped with the detected spectral map and known radio environmental map, the objective of the radar is to identify an appropriate transmit-frequency set $\mathcal{F}_R \subset \mathcal{F} \setminus \mathcal{F}_C$ such that the radar's probability of detection P_d is maximized. For a fixed probability of false alarm P_{fa}, the P_d increases with a higher signal-to-interference plus-noise ratio (SINR) [83]. Hence, the frequency selection process can, alternatively, choose to maximize the SINR or minimize the spectral power in the undesired parts of the spectrum. To find available bands with the least amount of interference, a structured sparsity framework [84] is adopted in [39]. Additional requirements of transmit-power constraints, range-sidelobe levels, and minimum separation between the bands can also be imposed. At the receiver of this spectrum-sharing radar, the sub-Nyquist processing method of [30] recovers the delay-Doppler map from the subset of Fourier coefficients defined by $\mathcal{F}_R$.

This CR system leads to three main advantages. First, the CS reconstruction, as presented in [30] on the transmitted fragmented bands, achieves the same resolution as traditional Nyquist processing over a significantly smaller bandwidth. Second, by concentrating all of the available power in the transmitted narrow bands rather than over a wide bandwidth, the CR increases the SNR. Finally, this technique allows for a dynamic form of the transmitted signal spectrum, in which only a small portion of the whole bandwidth is used at each transmission, thereby enabling spectrum sharing with communication signals, as illustrated in Figure 10(d). There, the coexistence between radar-transmitted bands in red and existing communication bands in white is shown.

SpeCX prototype

The SpeCX prototype, presented in Figure 10, demonstrates radar and communication spectrum sharing. It is composed of a CRo receiver and a CR transceiver. At the heart of the CRo system lies the proprietary modulated wideband converter board [29] that implements a sub-Nyquist analog front-end receiver, which processes signals with Nyquist rates up to 6 GHz. The card first splits the wideband signal into $M = 4$ hardware channels with an expansion factor of $q = 5$, yielding $M_q = 20$ virtual channels after digital expansion (see [85]). In each channel, the signal is mixed with a periodic sequence $p_i(t)$, which are truncated versions of Gold codes [86], generated on a dedicated field-programmable gate array, with a periodic frequency $f_p = 20$ MHz.

Next, the modulated signal passes through an analog antialiasing LPF. Finally, the low-rate analog signal is sampled by an NI ADC operating at $f_s = (q+1)f_p = 120$ MHz (with intended oversampling), leading to a total sampling rate of 480 MHz. The digital receiver is implemented on an NI PXIe-1065 computer with a dc-coupled ADC. Since the digital processing is performed at the low rate of 120 MHz, very low computational load is required to achieve real-time recovery. The prototype is fed with RF signals composed of up to $N_{\text{sig}} = 5$ real communication transmissions, i.e., ten spectral bands with a total bandwidth occupancy of up to 200 MHz and varying support, with a Nyquist rate of 6 GHz.

The input transmissions then go through an RF combiner, resulting in a dynamic multiband input signal that enables fast carrier switching for each of the bands. This input is specially designed to allow the testing of the system's ability to rapidly sense the input spectrum and adapt to changes, as required by modern CRo and shared spectrum standards, e.g., in the SSPARC program. The system's effective sampling rate, equal to 480 MHz, is only 8% of the Nyquist rate. Support recovery is digitally performed on the low-rate samples. The prototype successfully recovers the support of the communication transmitted bands, as demonstrated in Figure 10(b) and (c). Once the support is recovered, the signal itself can be reconstructed from the sub-Nyquist samples. This step is performed in real time, reconstructing the signal bands one sample at a time.

The CR receiver system is identical to the sub-Nyquist sampling prototype of [30], [31] and [40]. In the cognitive case, the transmitter only transmits over $N_b = 4$ bands, which constitute 3.2% of the original wideband signal bandwidth after the spectrum-sensing process has been completed by the communication receiver. Figure 10(d) illustrates the coexistence between the radar-transmitted bands in red and the existing communication bands in white. The gain in power is demonstrated in Figure 10(e); the wideband radar spectrum is shown in blue, the CR in red, and the noise in yellow on a logarithmic scale. The true and recovered range-velocity maps are presented in Figure 10(f). All of the $L = 10$ targets are perfectly recovered, and the clutter, depicted in blue, is discarded. Below the map, the range-recovery accuracy is shown for three scenarios: from

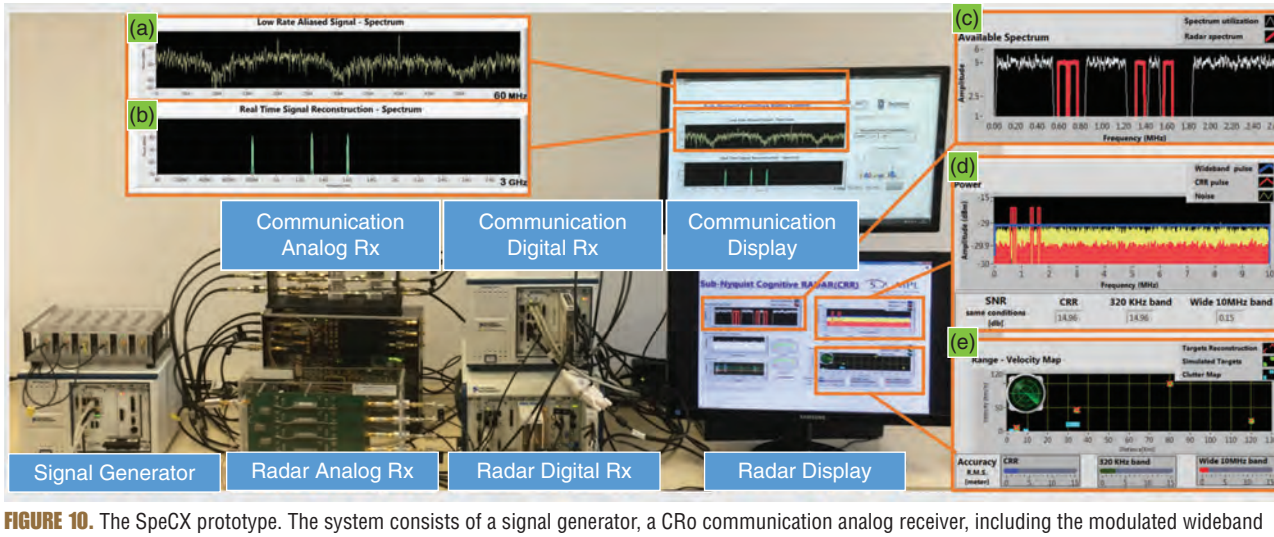


FIGURE 10. The SpeCX prototype. The system consists of a signal generator, a CRo communication analog receiver, including the modulated wideband converter (MWC) analog front-end board, a communication digital receiver, a CR analog, and a receiver. The SpeCX communication system display shows (a) low-rate samples acquired from one MWC channel at a rate of 120 MHz and (b) a digital reconstruction of the entire spectrum from sub-Nyquist samples. The SpeCX radar display shows (c) the coexisting communication and CR, (d) the CR spectrum compared with the full-band radar, and (e) the range-velocity display of both the detected and true locations of the targets [39].

left to right, the CR in blue (2.5 m), the four adjacent bands with the same bandwidth (12.5 m), and the wideband (4 m). The second configuration selects four adjacent frequency bands with the same bandwidth as the CR (with nonadjacent bands) for transmission. Its poor resolution stems from its small aperture. The CR system with nonadjacent bands yields better resolution than traditional wideband transmission, sampling, and processing at the Nyquist rate, due to the increased SNR.

Compressed MIMO radar

Compressed radar methods have recently been extended to MIMO settings, in which their impact may be even greater than for single-antenna configurations. MIMO radar systems belong to the family of array radars, which allow for the simultaneous recovery of the targets' ranges, Dopplers, and azimuths. This 3-D recovery results in high digital processing complexity. One of the main challenges of MIMO radar is therefore coping with complicated systems in terms of cost, high computational load, and hardware implementation. CS has been naturally applied to MIMO to reduce the processing complexity on the digital side as well as allow for spatial compression, in addition to the time compression achieved in single-antenna systems. In MIMO radars, the array aperture, which depends on the number of antennas, dictates the azimuth resolution. Since the aperture is determined by the number of antennas in traditional virtual ULAs, high-azimuth resolution requires a large number of antennas.

Increased resolution

As in single-antenna radar systems, CS has first been exploited to increase the parameter resolution. Here, the MIMO array is composed of T transmitters and R receivers so as to achieve the desired aperture $Z = TR/2$, as shown in Figure 2. The transmitted signal at the mth transmit antenna is given by (12), and each receiver samples the received signal at the Nyquist rate, as in a traditional MIMO. Assuming a sparse target scene, in which the ranges, Dopplers, and azimuths lie on a predefined grid, the work of [15] is extended to MIMO architectures in [18] and [19]. The transmit and receive array manifolds respectively, are given by

$$\mathbf{a}_T(\theta) = \left[e^{j2\pi\xi_1\theta}, e^{j2\pi\xi_2\theta}, \ldots, e^{j2\pi\xi_T\theta}\right]^T, \tag{43}$$

and

$$\mathbf{a}_R(\theta) = \left[e^{j2\pi\zeta_1\theta}, e^{j2\pi\zeta_2\theta}, \ldots, e^{j2\pi\zeta_R\theta}\right]^T, \tag{44}$$

where ξ_m and ζ_q are the relative mth transmit and qth receive antenna spacings. The $R \times N$ received signal matrix from a unit strength target at direction θ, with delay τ and Doppler ν is defined as

$$\mathbf{Z} = \mathbf{a}_R(\theta)\mathbf{a}_T^T(\theta)\mathbf{S}^T(\tau, \nu). \tag{45}$$

Here, $\mathbf{S}^T(\tau,\nu)_{i,m} = s_m(t_i - \tau)e^{j2\pi\nu t_i}$, where t_i are the sampling times and m indexes the transmitted waveforms. In this case, the columns of the dictionary $\mathbf{A}$ are given by $\mathrm{vec}(\mathbf{Z})$ for all possible combinations of θ, ν, and τ on a predefined grid.

The targets' parameters are recovered by matching the received signal with dictionary atoms. To achieve measurement diversity, random waveforms may be used, while the antenna locations are deterministic. ULAs are considered in [18] for both the transmit and receive arrays that do not benefit from the virtual array configuration. Alternatively, deterministic waveforms can be used, e.g., Kerdock codes [19], while the antenna locations are selected uniformly at random over the aperture $Z = TR/2$. Bounds on N, with respect to the number of antennas T and R and the number of samples that ensure targets' parameters recovery, are provided in [18] and [19].

A similar approach extends the framework of [15] to the MIMO setting by adding an azimuth matrix to the time-shift and frequency-modulation matrices $\mathbf{T}$ and $\mathbf{M}$, respectively, as defined in (14). In this case, each target lying on the grid is represented by a time shift, a frequency modulation, and an angle $\mathbf{A}_{q,m} = e^{j\theta(\xi_m + \zeta_q)}$ [87].

In both works, assuming N grid points in each dimension, the number of columns of $\mathbf{A}$ is N^3. The processing efficiency is thus penalized by a very large dictionary that contains every parameter combination. Note that the previously mentioned works focus on increased parameter resolution and do not deal with reduced time/spatial sampling and processing rates.

Reduced processing

Fast time compression is performed in [23]–[25], in which the Nyquist rate samples are compressed in each antenna before being forwarded to the central unit. A circular array is adopted in [23], with transmit and receive nodes uniformly distributed on a disk with a small radius. At each receive antenna, linear projections of the measurement vector are retained so that the resulting samples are compressed in both the slow and fast time domains. Both individual reconstruction at each receiver and joint processing at a fusion center are proposed, using CS recovery methods. The actual sampling is still performed at the Nyquist rate.

The MIMO matrix completion (MIMO-MC) radar [24], [25] employs MC techniques to avoid parameter discretization, which is typically used in CS methods. Two configurations are proposed for azimuth-Doppler recovery in a range bin of interest. In the first scenario, each receiver performs an MF and forwards the maximum of each MF output to the fusion center. The samples from the pth pulse transmitted to the fusion center can then be written in matrix form as

$$\mathbf{X}_p = \mathbf{A}_R \boldsymbol{\Sigma} \mathbf{D}_p \mathbf{A}_T^T, \tag{46}$$

where $\mathbf{X}_p$ is the $R \times T$ matrix that contains the maximum of the MF output for each transmitter and each receiver. For ULA configurations, the lth column of the $T \times L$ transmitter-steering matrix $\mathbf{A}_T$ is given by $(\mathbf{A}_T)_l = \left[1, e^{j\frac{2\pi}{\lambda}d_T\sin(\theta_l)}, \ldots, e^{j\frac{2\pi}{\lambda}(T-1)d_T\sin(\theta_l)}\right]^T$, where d_T is the interelement spacing. The steering matrix $\mathbf{A}_R$ at the receiver is similarly defined. The diagonal matrix $\boldsymbol{\Sigma}$ contains the targets' RCS α_l, and the diagonal matrix $\mathbf{D}$ contains the targets' Dopplers such that

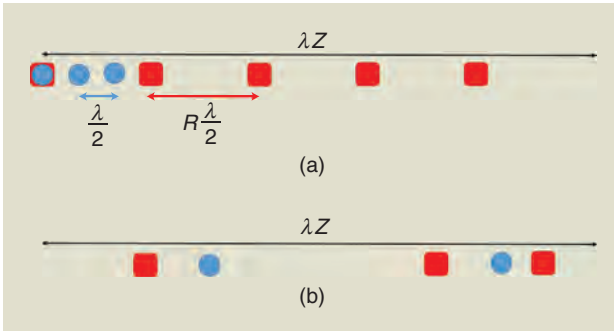


FIGURE 11. An illustration of MIMO arrays: (a) the standard array and (b) the random thinned array [32].

$\mathbf{D}_{p(l,l)} = e^{j\frac{2\pi}{\lambda}2\nu_l(p-1)\tau}$. In this scheme, each receiver transmits the output of a few randomly chosen MFs to the fusion center so that $\mathbf{X}_p$ is only partially known.

In the second scenario, the receivers forward Nyquist samples to the fusion center without performing the MF. In this case, the samples are written as

$$\mathbf{X}_p = \mathbf{A}_R \Sigma \mathbf{D}_p \mathbf{A}_R^T \mathbf{S}, \tag{47}$$

where $\mathbf{S}$ is the $T \times N$ matrix that contains the Nyquist rate samples of each transmitted waveform $s_m(t)$. In this scheme, each receive antenna randomly acquires a subset of the Nyquist samples and transmits these to the fusion center. In both cases, the fusion center performs MC before parametric estimation methods are applied to extract θ_l and ν_l, such as multiple signal classification, also known as *MUSIC* [88]. In these works, sampling and processing rate reduction are not addressed since compression is performed in the digital domain after sampling, and the missing samples are reconstructed before recovering the targets' parameters. Instead, these approaches are aimed at reducing the communication overhead between the receivers and the fusion center.

Spatial compression

Several recent works have considered applying CS to MIMO radar to reduce the number of antennas or the number of samples per receiver without degrading resolution. The problem of azimuth recovery of targets all in the same range-Doppler bin is investigated in [28]. Spatial compressive sampling is performed, in which the number of antennas is reduced while preserving the azimuth resolution. The classic MIMO virtual array configuration requires receivers with maximum spacing $\lambda/2$ and transmitters with spacing $R\lambda/2$ (or vice versa). The product RT thus scales linearly with aperture, which sets the azimuth resolution. Spatial compression is achieved by using a sparse random-array architecture [28], in which a low number of transmit and receive elements are placed at random over the same aperture Z, achieving similar resolution as a filled array, but with significantly fewer elements. The random-array configuration is illustrated in Figure 11. Beamforming is applied on the time-domain samples obtained from the thinned array at the Nyquist rate, and the azimuths are recovered using CS techniques. Recovery guarantees and guidelines concerning the choice of the product RT and the antenna locations are provided. The methods for choosing the antenna locations using deep networks are investigated in [89].

Spatial compression is achieved by using a sparse random-array architecture, in which a low number of transmit and receive elements are placed at random over the same aperture *Z*, achieving similar resolution as a filled array, but with significantly fewer elements.

Time and spatial compression

In all of the previously discussed works, recovery is performed in the time domain on acquired or reconstructed Nyquist rate samples for each antenna. The sub-Nyquist MIMO radar (SUMMeR) system, presented in [32], extends the Xampling concept to MIMO configurations and breaks the link between the aperture and the number of antennas, similar to [28]. The concept of Xampling is applied both in space (antenna deployment) and in time (sampling scheme) to simultaneously reduce the required number of antennas and samples per receiver, while preserving time and spatial resolution. In particular, targets' azimuths, ranges, and Dopplers are recovered from compressed samples in both space and time, while keeping the same resolution induced by Nyquist rate samples obtained from a full virtual array with low computational cost.

The SUMMeR system implements a collocated MIMO radar system with $M < T$ transmit and $Q < R$ receive antennas, whose locations are chosen uniformly at random within the aperture of the virtual array described previously in this section, i.e., $\{\xi_m\}_{m=0}^{M-1} \sim \mathcal{U}[0,Z]$ and $\{\zeta_q\}_{q=0}^{Q-1} \sim \mathcal{U}[0,Z]$, respectively, as shown in Figure 11. Note that, in principle, the antenna locations may be chosen on the ULAs' grid; however, this configuration is less robust than range-azimuth ambiguity and leads to coupling between these parameters in the presence of noise [32]. An FDMA framework is adopted so that spatial compression, which, in particular reduces the number of transmit antennas, removes the corresponding transmitting frequency bands as well. The transmitted signals are illustrated in Figure 12 in the frequency domain. Figure 12(a) and (b) shows a standard FDMA transmission for $T = 5$ and the resulting signal after spatial compression for $M = 3$.

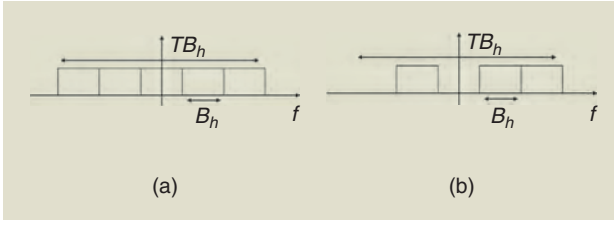


FIGURE 12. The frequency division multiple access transmissions: (a) standard and (b) spatial compression [32].

The transmitted pulses, defined in (12), are reflected by the targets and collected at the receive antennas. Under the assumptions described in "Targets' Assumptions," the received signal $\tilde{x}_q(t)$ at the qth antenna is the sum of time-delayed, scaled replicas of the transmitted signals:

$$\tilde{x}_q(t) = \sum_{m=0}^{T-1} \sum_{l=1}^{L} \alpha_l s_m\left(\frac{c+v_l}{c-v_l}\left(t - \frac{R_{l,mq}}{c+v_l}\right)\right), \tag{48}$$

where $R_{l,mq}$ is the sum of the distances from the mth transmitter and qth receiver to the lth target, which account for the array geometry. After demodulation to the baseband, the received signal can be further simplified to

$$x_q(t) = \sum_{p=0}^{P-1} \sum_{m=0}^{M-1} \sum_{l=1}^{L} \alpha_l h_m(t - p\tau - \tau_l) e^{j2\pi\beta_{mq}\vartheta_l} e^{j2\pi f_l^D p\tau}, \tag{49}$$

where $\beta_{mq} = (\zeta_q + \xi_m)(f_m \frac{\lambda}{c} + 1)$, with f_m the mth-transmission-carrier frequency and λ the signal wavelength. The goal is to estimate the targets' ranges, azimuths, and velocities, i.e., to estimate τ_l, ϑ_l, and f_l^D from low-rate samples of $x_q(t)$, and small numbers m and Q of the antennas.

Similar to the Xampling processing in [30], SUMMeR considers the Fourier coefficients of the received signal $x_q^p(t)$ at the qth antenna. To jointly recover the targets' ranges, azimuths, and Doppler frequencies, the concept of Doppler focusing from [30] (see "Doppler Focusing") is applied to the MIMO setting, and the CS algorithms are extended to simultaneous matrix recovery [32]. The minimal number of channels required for the recovery of L targets' parameters in noiseless settings is $MQ \geq 2L$, with a minimal number of $MK \geq 2L$ samples per receiver and $P \geq 2L$ pulses per transmitter [32]. The SUMMeR system has been implemented in hardware, as described in the following section.

Hardware prototype

The cognitive SUMMeR prototype [41], [42] extends the Doppler focusing, Xampling-based prototype [40] to the MIMO configuration. It simultaneously recovers the targets' delays, Dopplers, and azimuths from sub-Nyquist samples. More specifically, it implements a receiver with a maximum of eight transmit and ten receive antenna elements. The same hardware is used for each receive element and serially feeds the signals of all $R = 10$ receivers to the same prototype.

To avoid the use of an overwhelmingly large number of ADCs and bandpass filters for an 8×10 array, a cognitive transmission is adopted wherein each transmit signal lies in $N_b = 8$ disjoint, narrow slices over a 15-MHz band. Each subband is the width of 375 kHz, leading to a total signal bandwidth of 3 MHz. The transmit subbands, locations were chosen so that all can be subsampled using a single low-rate ADC without aliasing between them [41]. This allows the reduction of the number of samplers. The signal is subsampled at 7.5 MHz, whereas a noncognitive signal would have occupied the entire 15-MHz spectrum requiring a Nyquist sampling rate of 30 MHz. Therefore, the use of cognitive transmission enables spectral sampling reduction by a factor of four for each channel. The effective signal bandwidth is reduced by a factor of five (= 15/3 MHz), respectively, for each channel.

The system may be configured to operate in various array configurations simulating different numbers and locations of the antennas. The hardware switches off the inactive channels and does not sample any data over the corresponding ADCs. This governs the spatial compression by reducing the number of receivers and transmitters. In its baseline configuration, the system uses only half of the antennas with respect to the full virtual array, i.e., $M = 4$ transmitters and $Q = 5$ receivers. Figure 13 shows the sub-Nyquist MIMO prototype, user interface, and

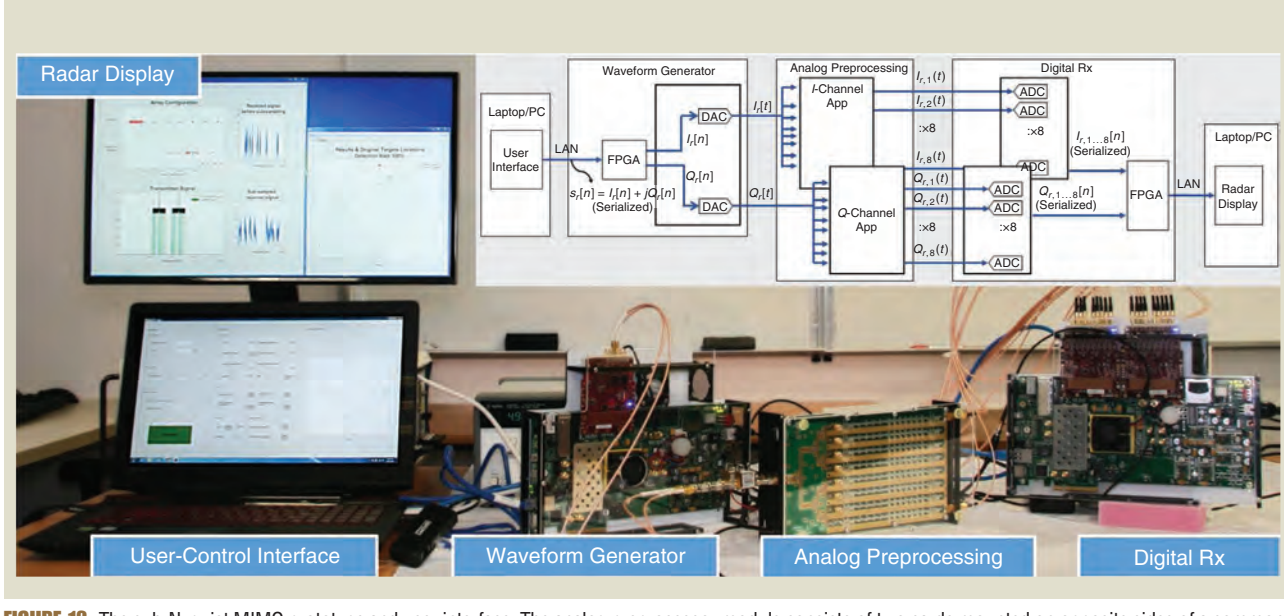


FIGURE 13. The sub-Nyquist MIMO prototype and user interface. The analog preprocessor module consists of two cards mounted on opposite sides of a common chassis. The inset shows the simplified block diagram of the system. The subscript r represents the received signal samples for the rth receiver. Wherever applicable, the second subscript corresponds to a particular transmitter. The square brackets (parentheses) are used for digital (analog) signals [41], [42].

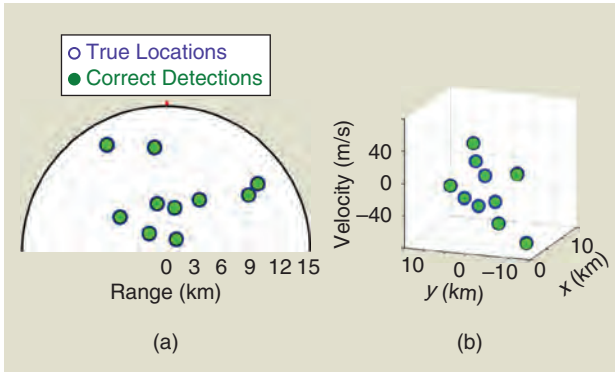


FIGURE 14. The SUMMeR prototype recovery performance: (a) the plan position indicator (PPI) display. The origin is the location of the radar. The red dot indicates the north direction relative to the radar. The positive/negative distances along the horizontal axis correspond to the east/west direction of the radar. Similarly, the positive/negative distances along the vertical axis correspond to the north/south direction of the radar. The estimated targets are plotted over the ground truth. (b) The range-azimuth-Doppler map for the same targets. The lower axes represent the Cartesian coordinates of the polar representation of the PPI plots from (a). The vertical axis represents the Doppler spectrum [32].

radar display. The inset graph depicts the signal flow through a simplified block diagram.

The experimental process consists of the following steps. The simulated radar scenario is stored in a custom-designed waveform generator. The scenario includes pulse-transmission modeling, accurate power loss due to wave propagation in a realistic medium, and interaction of a transmit signal with the target. A large variety of scenarios, consisting of different targets' parameters, i.e., delays, Doppler frequencies, and amplitudes, and array configurations, i.e., the number of transmitters and receivers and antenna locations, may be examined using the prototype. The waveform generator board then produces an analog signal corresponding to the synthesized radar environment, which is amplified and routed to the MIMO radar-receiver board. The prototype samples and processes the signal in real time. The physical array aperture and simulated target response correspond to an X-band ($f_c = 10$ GHz) radar.

The transmit subbands, locations were chosen so that all can be subsampled using a single low-rate ADC without aliasing between them.

Figure 14 presents some recovery results from the prototype. In the experiment, $P = 10$ pulses were transmitted at a uniform PRF of 100 μHz. The received signal corresponding to the echoes from $L = 10$ targets, placed at arbitrary ranges with azimuths and with arbitrary velocities, was injected into the transmit waveform generator. In the experiment, when the angular spacing (in terms of the sine of azimuth) between any two targets was greater than 0.025 and the signal SNR $= -8$ dB, the recovery performance of the compressed configuration in time and in space was equivalent to that of a full array, i.e., with eight transmitters and ten receivers. The figure shows the obtained plan position indicator plot and range-azimuth-Doppler maps for both true and recovered targets. Here, a successful detection (green circle) occurs when the estimated target is within one range cell, one azimuth bin, and one Doppler bin of the ground truth (blue circle). More experiments in [41] and [42] demonstrate that the prototype performance is robust, with SNRs dropping to as low as –10 dB, and the time and spatial resolution are preserved by simulating couples of close targets in range, Doppler, and azimuth.

Conclusions and future challenges

In this article, we reviewed several compressed radar systems that aim to reduce complexity while preserving parameter resolution. Throughout this article, we considered different popular radar systems, including pulse-Doppler and step-frequency radars as well as MIMO configurations. In particular, we showed that temporal, spectral, and spatial compression can be implemented without decreasing Doppler, range, and azimuth resolution. To recover these parameters for L targets, the minimum number of required samples per pulse, the minimum number of pulses, and the minimum number of channels are each equal to $2L$. These are determined by the actual number of DoF of the parameter estimation problem, governed by L, rather than a function of design parameters, such as signal bandwidth, CPI, or aperture. This is essential since the latter determine range, Doppler, and azimuth resolution and are increased for higher performance. By breaking the traditional links between the sampling rate, number of pulses, and antennas on the one hand and parameter estimation on the other hand, increased performance may be achieved without increasing sampling and processing rates.

An advantage of the Xampling system is that traditional radar-processing algorithms can be easily adapted and applied directly to the sub-Nyquist samples. For example, clutter-cancellation techniques have been implemented on the Xampling radar prototypes. These significantly enhance the performance of compressed radars without requiring the reconstruction of Nyquist rate samples. In addition, while CS-based methods traditionally do not perform well in the presence of large noise, since they inherently reduce SNR, Doppler focusing, applied to samples obtained using Xampling, enjoys an SNR improvement that scales linearly with the number of pulses, obtaining good detection at low SNRs.

An essential part of the approach adopted in this survey is the relation between the theoretical algorithms and practical hardware implementation, demonstrating real-time target detection from compressed samples in the fast and slow time domains, as well as in space. The prototypes presented here were built from off-the-shelf components, paving the way to enable commercial, compressed radar systems. To this end, such hardware prototypes should be further extended to implement radar transmit and receive systems and deploy them to be tested on real data. This would permit assessing their performance in real-world conditions, including different types of noise, clutter, and interference.

Authors

Deborah Cohen (deborah.co88@gmail.com) received her B.S. degree in electrical engineering (summa cum laude) in 2010 and her Ph.D. degree in electrical engineering in 2016, both from the Technion–Israel Institute of Technology, Haifa. She has been granted several awards, including the Meyer Foundation Excellence Award, the David and Tova Freud and Ruth Freud-Brendel Memorial Scholarship, the Sandor Szego Award, the Vivian Konigsberg Award, and the Muriel and David Jacknow Award for Excellence in Teaching. Since 2014, she has been an Azrieli fellow. She is currently a research scientist with Google Israel. Her research interests include theoretical aspects of signal processing, compressed sensing, reinforcement learning, and machine learning for dialogues.

Yonina C. Eldar (yonina@ee.technion.ac.il) is a professor in the Department of Electrical Engineering at the Technion–Israel Institute of Technology, Haifa, where she holds the Edwards chair in engineering. She is also an adjunct professor at Duke University, Durham, North Carolina, and a research affiliate with the Research Laboratory of Electronics at the Massachusetts Institute of Technology, Cambridge, and she was a visiting professor at Stanford University, California. She is a member of the Israel Academy of Sciences and Humanities and a Fellow of the IEEE and the European Association for Signal Processing. She has received many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award, the IEEE/Aerospace and Electronic Systems Society Fred Nathanson Memorial Radar Award, the IEEE Kiyo Tomiyasu Award, the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, and the Wolf Foundation Krill Prize for Excellence in Scientific Research. She is the editor-in-chief of *Foundations and Trends in Signal Processing* and serves the IEEE on several technical and award committees.

References

[1] X. P. Masbernat, M. G. Amin, F. Ahmad, and C. Ioana, "An MIMO-MTI approach for through-the-wall radar imaging applications," in *Proc. IEEE Int. Waveform Diversity and Design Conf.*, 2010, pp. 188–192.

[2] D. J. Daniels, *Ground Penetrating Radar*. London, U.K.: Institution of Engineering and Technology, 2004.

[3] K. Schuler, M. Younis, R. Lenz, and W. Wiesbeck, "Array design for automotive digital beamforming radar system," in *Proc. IEEE Int. Radar Conf.*, 2005, pp. 435–440.

[4] V. Bringi and V. Chandrasekar, *Polarimetric Doppler Weather Radar: Principles and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[5] M. Skolnik, *Radar Handbook*. New York: McGraw-Hill, 1970.

[6] P. Z. Peebles, *Radar Principles*. Hoboken, NJ: Wiley, 2007.

[7] M. A. Richards, *Fundamentals of Radar Signal Processing*. New York: McGraw-Hill, 2014.

[8] E. Fishler, A. Haimovich, R. Blum, D. Chizhik, L. Cimini, and R. Valenzuela, "MIMO radar: An idea whose time has come," in *Proc. IEEE Radar Conf.*, 2004, pp. 71–78.

[9] J. Li and P. Stoica, *MIMO Radar Signal Processing*. Piscataway, NJ: IEEE Press, 2009.

[10] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[11] J. C. Curlander and R. N. McDonough, *Synthetic Aperture Radar*. New York: Wiley, 1991.

[12] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[13] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[14] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[15] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2275–2284, 2009.

[16] Y. Chi, R. Calderbank, and A. Pezeshki, "Golay complementary waveforms for sparse delay-Doppler radar imaging," in *Proc. IEEE Int. Workshop Computational Advances Multi-Sensor Adaptive Processing*, 2009, pp. 177–180.

[17] S. Shah, Y. Yu, and A. Petropulu, "Step-frequency radar with compressive sampling (SFR-CS)," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2010, pp. 1686–1689.

[18] T. Strohmer and H. Wang, "Sparse MIMO radar with random sensor arrays and Kerdock codes," in *Proc. IEEE Int. Conf. Sampling Theory and Applications*, 2013, pp. 517–520.

[19] T. Strohmer and B. Friedlander, "Analysis of sparse MIMO radar," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 3, pp. 361–388, 2014.

[20] W. U. Bajwa, K. Gedalyahu, and Y. C. Eldar, "Identification of parametric underspread linear systems and super-resolution radar," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2548–2561, 2011.

[21] O. Teke, A. C. Gurbuz, and O. Arikan, "A robust compressive sensing-based technique for reconstruction of sparse radar scenes," *Digit. Signal Process.*, vol. 27, pp. 23–32, Apr. 2014.

[22] J. H. Ender, "On compressive sensing applied to radar," *Signal Process.*, vol. 90, no. 5, pp. 1402–1414, 2010.

[23] Y. Yu, A. P. Petropulu, and H. V. Poor, "MIMO radar using compressive sampling," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 1, pp. 146–163, 2010.

[24] D. S. Kalogerias and A. P. Petropulu, "Matrix completion in colocated MIMO radar: Recoverability, bounds & theoretical guarantees," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 309–321, 2014.

[25] S. Sun, W. U. Bajwa, and A. P. Petropulu, "MIMO-MC radar: A MIMO radar approach based on matrix completion," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 3, pp. 1839–1852, 2015.

[26] E. Ertin, L. C. Potter, and R. L. Moses, "Sparse target recovery performance of multi-frequency chirp waveforms," in *Proc. IEEE European Signal Processing Conf.*, 2011, pp. 446–450.

[27] C. Liu, F. Xi, S. Chen, Y. D. Zhang, and Z. Liu, "Pulse-Doppler signal processing with quadrature compressive sampling," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 2, pp. 1217–1230, 2015.

[28] M. Rossi, A. M. Haimovich, and Y. C. Eldar, "Spatial compressive sensing for MIMO radar," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 419–430, 2014.

[29] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Devices Syst.*, vol. 5, no. 1, pp. 8–20, 2011.

[30] O. Bar-Ilan and Y. C. Eldar, "Sub-Nyquist radar via Doppler focusing," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1796–1811, 2014.

[31] D. Cohen, A. Dikopoltsev, R. Ifraimov, and Y. C. Eldar, "Towards sub-Nyquist cognitive radar," in *Proc. IEEE Radar Conf.*, 2016. doi: 10.1109/RADAR.2016.7485122.

[32] D. Cohen, D. Cohen, Y. C. Eldar, and A. M. Haimovich, "SUMMeR: Sub-Nyquist MIMO radar," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4315–4330, 2018.

[33] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, "Sparsity and compressed sensing in radar imaging," *Proc. IEEE*, vol. 98, no. 6, pp. 1006–1020, 2010.

[34] M. Cetin, I. Stojanovic, N. O. Onhon, K. R. Varshney, S. Samadi, W. C. Karl, and A. S. Willsky, "Sparsity-driven synthetic aperture radar imaging: Reconstruction, autofocusing, moving targets, and compressed sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 27–40, 2014.

[35] L. Zhao, L. Wang, L. Yang, A. M. Zoubir, and G. Bi, "The race to improve radar imagery: An overview of recent progress in statistical sparsity-based techniques," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 85–102, 2016.

[36] H. Griffiths, L. Cohen, S. Watts, E. Mokole, C. Baker, M. Wicks, and S. Blunt, "Radar spectrum engineering and management: Technical and regulatory issues," *Proc. IEEE*, vol. 103, no. 1, pp. 85–102, 2015.

[37] M. P. Fitz, T. R. Halford, I. Hossain, and S. W. Enserink, "Towards simultaneous radar and spectral sensing," in *Proc. IEEE Int. Symp. Dynamic Spectrum Access Networks*, 2014, pp. 15–19.

[38] J. Bernhard, J. Reed, J. M. Park, A. Clegg, A. Weisshaar, and A. Abouzeid. (2010). Final report of the National Science Foundation workshop on enhancing

access to the radio spectrum (EARS). Nat. Sci. Foundation. Arlington. [Online]. Available: https://www.nsf.gov/mps/ast/nsf_ears_workshop_2010_final_report.pdf

[39] D. Cohen, K. V. Mishra, and Y. C. Eldar, "Spectrum sharing radar: Coexistence via Xampling," *IEEE Trans. Aerosp. Electron. Syst.,* vol. 54, pp. 1279–1296, 2018.

[40] E. Baransky, G. Itzhak, I. Shmuel, N. Wagner, E. Shoshan, and Y. C. Eldar, "Sub-Nyquist radar prototype: Hardware and algorithms," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 1, pp. 809–822, 2014.

[41] K. V. Mishra, E. Shoshan, M. Namer, M. Meltsin, D. Cohen, R. Madmoni, S. Dror, R. Ifraimov, et al., "Cognitive sub-Nyquist hardware prototype of a collocated MIMO radar," in *Proc. Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing (CoSeRa)*, 2016, pp. 56–60.

[42] D. Cohen, K. V. Mishra, D. Cohen, E. Ronen, Y. Grimovich, M. Namer, M. Meltsin, and Y. C. Eldar, "Cognitive sub-Nyquist MIMO radar prototype with Doppler processing," in *IEEE Radar Conf.*, 2016, pp. 1179–1184.

[43] J. Yoo, C. Turnes, E. B. Nakamura, C. K. Le, S. Becker, E. A. Sovero, M. B. Wakin, M. C. Grant, et al., "A compressed sensing parameter extraction platform for radar pulse signal acquisition," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 3, pp. 626–638, 2012.

[44] E. Fishler, A. Haimovich, R. S. Blum, and L. J. Cimini, "Spatial diversity in radars-models and detection performance," *IEEE Trans. Signal Process.*, vol. 54, pp. 823–838, Mar. 2006.

[45] C. Cook, *Radar Signals: An Introduction to Theory and Application*. Amsterdam, The Netherlands: Elsevier, 2012.

[46] J. Li and P. Stoica, "MIMO radar with collocated antennas," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 106–114, 2007.

[47] A. M. Haimovich, R. S. Blum, and L. J. Cimini, "MIMO radar with widely separated antennas," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 116–129, 2008.

[48] D. Bliss and K. Forsythe, "Multiple-input multiple-output (MIMO) radar and imaging: Degrees of freedom and resolution," in *Proc. IEEE Asilomar Conf. Signals, Systems and Computers*, 2003, pp. 54–59.

[49] C.-Y. Chen, "Signal processing algorithms for MIMO radar," Ph.D. dissertation, California Inst. of Technol., Pasadena, 2009.

[50] P. Vaidyanathan, P. Pal, and C.-Y. Chen, "MIMO radar with broadband waveforms: Smearing filter banks and 2D virtual arrays," in *Proc. IEEE Asilomar Conf. Signals, Systems and Computers*, , pp. 188–192.

[51] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. IEEE Asilomar Conf. Signals, Systems and Computers*, 1993, pp. 40–44.

[52] G. M. Davis, S. G. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions with matching pursuit," *SPIE*, vol. 2242, pp. 402–414, Mar. 1994. doi: 10.1117/12.170041.

[53] T. Blumensath and M. Davies, "Gradient pursuits," *IEEE Trans. Signal Process.*, vol. 56, pp. 2370–2382, Jun. 2008.

[54] Y. C. Eldar, R. Levi, and A. Cohen, "Clutter removal in sub-Nyquist radar," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 177–181, 2015.

[55] R. Baraniuk and P. Steeghs, "Compressive radar imaging," in *Proc. IEEE Radar Conf.*, 2007, pp. 128–133.

[56] W. O. Alltop, "Complex sequences with low periodic correlations," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 350–354, 1980.

[57] Z. Yang, C. Zhang, and L. Xie, "Robustly stable signal recovery in compressed sensing with structured matrix perturbation," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4658–4671, 2012.

[58] M. A. Hadi, S. Alshebeili, K. Jamil, and F. E. A. El-Samie, "Compressive sensing applied to radar systems: An overview," *Signal, Image Video Process.*, vol. 9, no. 1, pp. 25–39, 2015.

[59] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, pp. 520–544, Jan. 2010.

[60] J. Yoo, S. Becker, M. Monge, M. Loh, E. Candes, and A. Emami-Neyestanak, "Design and implementation of a fully integrated compressed-sensing signal acquisition system," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 5325–5328.

[61] H. Liu, A. Ghafoor, and P. H. Stockmann, "A new quadrature sampling and processing approach," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 25, no. 5, pp. 733–748, 1989.

[62] X. Song, S. Zhou, and P. Willett, "The role of the ambiguity function in compressed sensing radar," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2010, pp. 2758–2761.

[63] K. Gedalyahu, R. Tur, and Y. C. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1491–1504, 2011.

[64] R. Tur, Y. C. Eldar, and Z. Friedman, "Innovation rate sampling of pulse streams with application to ultrasound imaging," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1827–1842, 2011.

[65] D. Cohen and Y. C. Eldar, "Reduced time-on-target in pulse Doppler radar: Slow time domain compressed sensing," in *Proc. IEEE Radar Conf.*, 2016. doi: 10.1109/RADAR.2016.7485243.

[66] T. Wimalajeewa, Y. C. Eldar, and P. K. Varshney. (2013, 11 Nov.). Recovery of sparse matrices via matrix sketching. arXiv. [Online]. Available: http://arxiv.org/abs/1311.2448

[67] R. J. Doviak and D. S. Zrnic, *Doppler Radar & Weather Observations*. New York: Academic, 2014.

[68] A. Ferrari, G. Alengrin, and C. Theys, "Doppler ambiguity resolution using staggered PRF with a new chirp sweep-rate estimation algorithm," *IEEE Proc. Radar, Sonar Navigation*, vol. 142, no. 4, pp. 191–194, 1995.

[69] V. Venkatesh, L. Li, M. McLinden, G. Heymsfield, and M. Coon, "A frequency diversity pulse-pair algorithm for extending Doppler radar velocity Nyquist range," in *Proc. IEEE Radar Conf.*, 2016, pp. 1–6.

[70] A. Ludloff and M. Minker, "Reliability of velocity measurement by MTD radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-21, no. 4, pp. 522–528, 1985.

[71] G. Trunk and S. Brockett, "Range and velocity ambiguity resolution," in *Proc. IEEE Nat. Radar Conf.*, 1993, pp. 146–149.

[72] X. Liu, D. Cohen, T. Huang, Y. Liu, G. Winerich, L. Shani, and Y. C. Eldar, "Unambiguous delay-Doppler recovery from phased coded pulses," 2016, to be published.

[73] Q. Cao, G. Zhang, R. D. Palmer, and L. Lei, "Detection and mitigation of second-trip echo in polarimetric weather radar employing random phase coding," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1240–1253, 2012.

[74] S. Haykin, "Cognitive radar: A way of the future," *IEEE Signal Process. Mag.*, vol. 23, no. 1, pp. 30–40, 2006.

[75] K. V. Mishra, D. Cohen, S. Tsiper, S. Stein, E. Shoshan, M. Namer, M. Meltsin, R. Madmoni, E. Ronen, Y. Grimovich, and Y. C. Eldar, "Xampling-enabled coexistence in spectrally crowded environments," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2017, pp. 6580–6581.

[76] G. M. Jacyna, B. Fell, and D. McLemore, "A high-level overview of fundamental limits studies for the DARPA SSPARC program," in *Proc. IEEE Radar Conf.*, 2016. doi: 10.1109/RADAR.2016.7485100.

[77] P. Stinco, M. S. Greco, and F. Gini, "Spectrum sensing and sharing for cognitive radars," *IET Radar, Sonar Navigation*, vol. 10, no. 3, pp. 595–602, 2016.

[78] N. Nartasilpa, D. Tuninetti, N. Devroye, and D. Erricolo, "Let's share CommRad: Effect of radar interference on an uncoded data communication system," in *Proc. IEEE Radar Conf.*, 2016. doi: 10.1109/RADAR.2016.7485064.

[79] P. Kumari, N. Gonzalez-Prelcic, and R. W. Heath, "Investigating the IEEE 802.11ad standard for millimeter wave automotive radar," in *Proc. Vehicular Technol. Conf.*, 2015. doi: 10.1109/VTCFall.2015.7390996.

[80] J. Mitola and C. Q. Maguire, Jr., "Cognitive radio: Making software radios more personal," *IEEE Personal Commun.*, vol. 6, no. 4, pp. 13–18, 1999.

[81] D. Cohen, S. Tsiper, and Y. C. Eldar, "Analog-to-digital cognitive radio: Sampling, detection, and hardware," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 137–166, 2018.

[82] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4595–4607, 2010.

[83] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Englewood Cliffs, NJ: Prentice Hall, 1998.

[84] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, pp. 3371–3412, Nov. 2011.

[85] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 375–391, 2010.

[86] R. Gold, "Optimal binary sequences for spread spectrum multiplexing (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 13, no. 4, pp. 619–621, 1967.

[87] C.-Y. Chen and P. Vaidyanathan, "Compressed sensing in MIMO radar," in *Proc. IEEE Asilomar Conf. Signals, Systems and Computers*, 2008, pp. 41–44.

[88] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 4, pp. 276–280, 1986.

[89] A. M. Elbir, K. V. Mishra, and Y. C. Eldar. (2018). Cognitive radar antenna selection via deep learning. arXiv. [Online]. Available: https://arxiv.org/abs/1802.09736

SP

Privacy-Aware Smart Metering

Progress and challenges

©ISTOCKPHOTO.COM/NATALI_MIS

Giulio Giaconi, Deniz Gündüz, and H. Vincent Poor

The next-generation energy network, the so-called smart grid (SG), promises tremendous increases in efficiency, safety, and flexibility in managing the electricity grid as compared to the legacy energy network. This is needed today more than ever, as global energy consumption is growing at an unprecedented rate and renewable energy sources (RESs) must be seamlessly integrated into the grid to assure a sustainable human development.

Smart meters (SMs) are among the crucial enablers of the SG concept. They supply accurate, high-frequency information about users' household energy consumption to a utility provider (UP), which is essential for time-of-use (ToU) pricing, rapid fault detection, and energy theft prevention, while also providing consumers with more flexibility and control over their consumption. However, highly accurate and granular SM

Digital Object Identifier 10.1109/MSP.2018.2841410
Date of publication: 13 November 2018

data also pose a threat to consumer privacy, as nonintrusive load monitoring (NILM) techniques enable a malicious attacker to infer many details of a user's private life.

This article focuses on privacy-enhancing energy management techniques that provide accurate energy consumption information to the grid operator without sacrificing consumer privacy. In particular, we focus on techniques that shape and modify actual user energy consumption by means of physical resources, such as rechargeable batteries (RBs), RESs, and demand shaping. A rigorous mathematical analysis of privacy is presented under various physical constraints on the available physical resources. Finally, open questions and challenges that need to be addressed to pave the way to the effective protection of users' privacy in future SGs are presented.

SMs for an SG

The current energy grid is one of the engineering marvels of the 20th century. However, it has become inadequate for satisfying the steadily growing global electricity demand of the 21st century. In fact, world energy consumption is predicted to increase 48% from 2012 to 2040 [1], driven by factors such as the growth of the global economy, the rise of the gross domestic product per person, the expansion of the planet's population, an increased penetration of electric vehicles, and a broader mobility revolution [2]. Other issues that need to be addressed are the effective integration of RESs and storage capabilities into the grid, the improvement of the grid's environmental sustainability, and the promotion of plug-in hybrid electric vehicles.

To address these challenges, SGs are being engineered. They are intended to substantially improve energy generation, transmission, distribution, consumption, and security, providing enhanced reliability and quality of the electricity supply, quicker detection of energy outages and theft, better matching of the energy supply with demand, and greater environmental sustainability by enabling an easier integration of distributed generation and storage capabilities. The smartness of an SG resides in its advanced metering infrastructure, which enables two-way communication between the utility and its customers and whose pivotal element in the distribution network is the SM, the device that monitors a user's electricity consumption in almost real time.

In contrast to legacy grids, in which billing data are gathered at the end of a use period, SMs send electricity consumption measurements automatically and at a much higher resolution. They enable two-way communication with the UP, the entity that sells energy to the customers, transmitting a great amount of detailed information. SMs collect and send bidirectional readings of active, reactive, and apparent power and energy—i.e., so-called four-quadrant metering—that is purchased from the grid (or sold to it, if the user produces energy, e.g., by means of a photovoltaic panel). In the latter case, the user is referred to as a *prosumer*, i.e., at once a producer and consumer of electricity who can be financially rewarded for the energy sold to the grid.

Table 1. The time resolution of SMs currently in use.

SM Model	Time Resolution
Itron Centron [72]	1 min
REX2 [73]	5 min
Kamstrup Omnipower [74]	5 min
Enel Open Meter [75]	15 min

SMs also keep track of historical consumption data over the previous days, weeks, and months and provide high-resolution consumption data analytics to the customers to enable them to monitor their energy consumption via an in-home display, web portal, or smartphone application in near real time. SMs also send alerts about voltage quality measurements, helping UPs fulfill their obligations toward customers concerning energy, power, and voltage quality, e.g., in accordance with the EN 50160 European standard. Examples of these measurements include the root mean square voltage variations, e.g., voltage dropout, sags and swells, and total harmonic distortion.

Data used for billing, such as the current ToU tariff, balance and debts, credit and prepayment modes, credit alerts, and topping up, are also sent to the UP. SMs can detect if tampering takes place and send relevant data about it, along with the security credentials for enabling the correct functioning of cryptographic protocols, e.g., hashing, digital signatures, and cyclic redundancy checks. Finally, SM firmware information and updates are also communicated.

The increased data resolution is crucial for enabling SG functionalities. Table 1 shows the smallest time resolution of some SMs currently in use, which is on the order of a few minutes. The European Union recommends a time resolution of at least 15 min to allow the new SG functionalities [3]. For example, the current SM specifications in the United Kingdom mandate that an SM send integrated energy readings every 30 min to the UP, while the data sent to a user's in-home display can have a resolution of up to 10 s [4]. It should be noted that, with the increased adoption of renewable energy generation by prosumers, the increased penetration of electric vehicles and energy storage technologies, and the diversification of the energy market, it is expected that SGs will become more volatile, requiring meter readings at a much higher rate in the near future.

SMs provide a wide range of benefits to all of the parties in an SG. Thanks to SMs, UPs can gain better knowledge of their customers' needs while reducing the cost of meter readings. SMs allow UPs to dynamically determine the electricity cost and produce more accurate bills, thus reducing customers' complaints and back-office rebilling. The implementation of ToU pricing can incentivize demand response and control customer behavior, while improved demand forecasts and load-shaping techniques can reduce peak electricity demands. Finally, energy theft can be detected more easily and quickly.

Distribution system operators (DSOs), i.e., the entities that operate the grid, benefit from SMs as well, by being able to better monitor and manage the grid. SMs allow DSOs to reduce operational costs and energy losses and improve grid efficiency and system design, distributed system state estimation, and volt/VAR control. Moreover, DSOs are able to better match distributed resources with the ongoing electricity demand and

the grid's power delivery capability, thus reducing the need to build new power plants.

Consumers themselves take advantage of SMs to monitor their consumption in near real time, leading to better consumption awareness and energy usage management. Moreover, consumers receive accurate and timely billing services, with no more estimated bills, and benefit from ToU pricing by shifting nonurgent loads to off-peak price periods. Microgeneration and energy storage devices can be integrated more easily, and profits from selling the generated excess energy can be collected automatically. Failing or inefficient home appliances, unexpected activity or inactivity, and wasted energy are detected faster and more accurately; in addition, switching between UPs is made easier by requesting on-demand readings, which, in turn, increases the competition among UPs and reduces costs for consumers.

For the aforementioned reasons, the installation of SMs is proceeding rapidly and attracting massive investment globally. The SM market is expected to grow from an estimated US$12.79 billion in 2017 to US$19.98 billion by 2022, registering a compound annual growth rate of 9.34% [5]. Moreover, the global SM data analytics market, which includes demand response analytics and grid optimization tools, is expected to reach US$4.6 billion by 2022 [6], while the global penetration of SMs is expected to climb from approximately 30% at the end of 2016 to 53% by the end of 2025 [7]. These figures show how timely and crucial the research in this field is, and they highlight the need to quickly resolve potential obstructions that can threaten the future benefits from this critical technology.

SM privacy risks

An SM's ability to monitor a user's electricity consumption in almost real time presents serious implications for consumer privacy. In fact, by employing NILM techniques, it is possible to identify the power signatures of specific appliances from aggregated household SM measurements. NILM approaches date back to the 1980s work of George Hart, who first proposed a prototype of an NILM device [9]. Since then, NILM methods have improved in different directions, e.g., by assuming either high- or low-frequency measurements, by considering known or learned signatures [10], and even by using off-the-shelf statistical methods without any a priori knowledge of household activities [11].

An example of a typical power consumption profile, along with some detected appliances, is illustrated in Figure 1. As shown in Figure 2, the UP, a third party that has access to SM data (by, for example, buying it from the UP), or a malicious eavesdropper may gain insights into users' activities and behaviors, and determine such information as a person's presence at home, religious beliefs, disabilities, illnesses, and even the TV channel being watched [12]–[14].

Apart from residential users, SM privacy is particularly critical for businesses, e.g., factories and data centers, as their power consumption profile may reveal sensitive information about the state of their businesses to their competitors. SM privacy has attracted significant public attention and continues to be a topic of heated public and political debate. The issue even stopped the mandatory SM rollout plan in The Netherlands in 2009 after a court decided that the forced installation of SMs would violate consumers' right to privacy and be in breach of the European Convention of Human Rights [15]. Indeed, concerns about consumer privacy threaten the widespread adoption of SMs and can be a major roadblock for this multibillion-dollar industry.

It is worth pointing out that the privacy problem involving SMs is different from the SM data security problem [16]. In the latter, there is a sharp distinction between legitimate users and malicious attackers, whereas in the privacy problem in the SM context, any legitimate receiver of data can also be considered malicious. To benefit from the advantages provided by the SG, users need to share some information about their electricity consumption with the UP and DSO. However, by sharing accurate and high-frequency information about their energy consumption, consumers also expose their private lives and behavior to the UP, which is a fully legitimate user and, at the same time, a potential malevolent party. This renders traditional encryption techniques for data privacy ineffective in achieving privacy against the UP and calls for novel privacy measures and privacy-protection techniques.

Privacy-enabling techniques for SMs

There is a growing literature on SM privacy-preserving methods, which can be classified into two main families. The first, which we call the *SM data manipulation* (*SMDM*) approach, consists of techniques that process the SM data before reporting them to the UP; the practices in the second family, called *user demand shaping* (*UDS*), aim at modifying the user's actual energy consumption. Considered within the first class are methods such as data obfuscation, data aggregation, data anonymization, and downsampling.

Data obfuscation, i.e., the perturbation of metering data by adding noise, is a classical privacy-protection method and has been adapted to SGs in [17] and [18]. In [19], differential

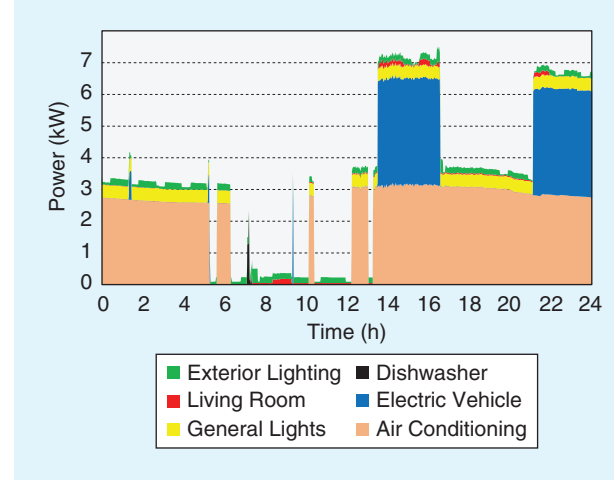


FIGURE 1. An example of a household electricity consumption profile with some appliances highlighted. (Figure courtesy of the Dataport database [8].)

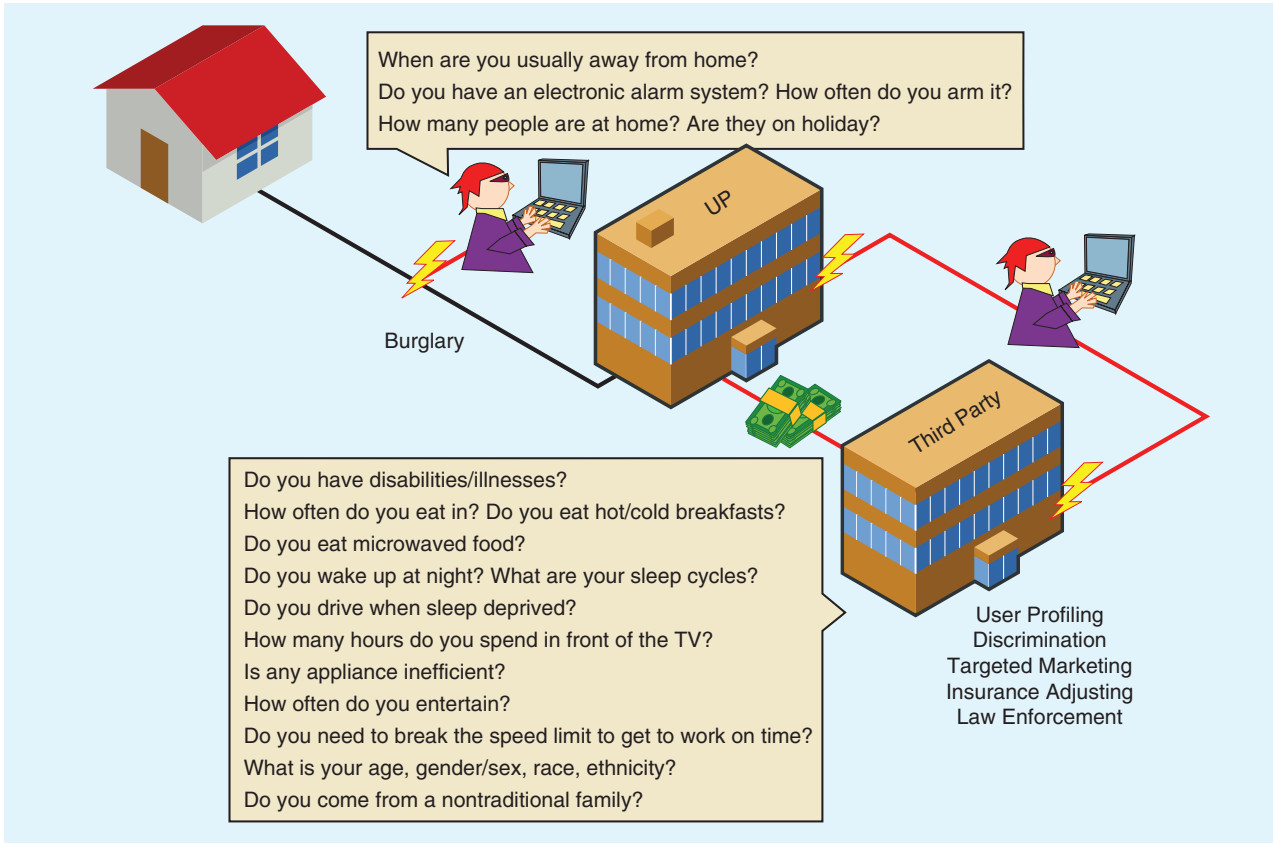


FIGURE 2. Some questions an attacker may be able to answer by having access to SM data.

privacy, a well-established concept in the data-mining literature, is applied to SMs, where noise is added not only to the user's energy consumption, via the RB, but also to the energy used for charging the RB itself to provide differential privacy guarantees. Along these lines, the authors in [20] introduce an information-theoretic framework to study the tradeoff between the privacy obtained by altering the SM data and the utility of data for various SG functionalities. Note that the more noise added to the data, the higher the privacy but the less relevant and useful the data are for monitoring and controlling the grid. In [20], an additive distortion measure is considered to model the utility, which allows the characterization of the optimal privacy-versus-utility tradeoff in an information-theoretic single-letter form.

The data aggregation approach, proposed in [18], [21], and [22], considers sending the aggregate power measurements for a group of households so that the UP is prevented from distinguishing individual consumption patterns. The aggregation can be performed with or without the help of a trusted third party (TTP).

The data anonymization approach, on the other hand, mainly considers utilizing pseudonyms rather than the real identities of consumers [23], [24]. Another method, proposed in [25], reduces the SM sampling rate to a level that does not pose any privacy threat. However, the SMDM family suffers from the following shortcomings.

- Adding noise to the SM readings causes a mismatch between the reported values and the real energy consumption, which prevents DSOs and UPs from accurately monitoring the grid state; rapidly reacting to outages, energy theft, or other problems; and producing accurate and timely billing services. These would significantly limit the SM benefits.
- DSOs, UPs, or, more generally, any eavesdropper can embed additional sensors right outside a household or a business (street-level measurements are already available to DSOs and UPs) to monitor the energy consumption without fully relying on SM readings.
- The anonymization and aggregation techniques that include the presence of a TTP only shift the problem of trust from one entity (the UP) to another (the TTP).

These issues are avoided by the UDS approaches, which directly modify the actual energy consumption profile of the user, called the *user load*, rather than modifying the data sent to the UP. In this family, the SM accurately reports the energy taken from the grid without any modification; however, this is not the energy that is actually consumed by the appliances. This is achieved by filtering the user's actual electricity consumption via a rechargeable energy storage device, i.e., an RB, or by exploiting an RES, which can be used to partially hide the consumer's energy consumption. Examples of RESs include solar panels and micro wind farms.

Another technique is to partially shift a consumer's demand. If we denote the energy received from the grid as the *grid load*, the idea is to physically differentiate the grid load from the user

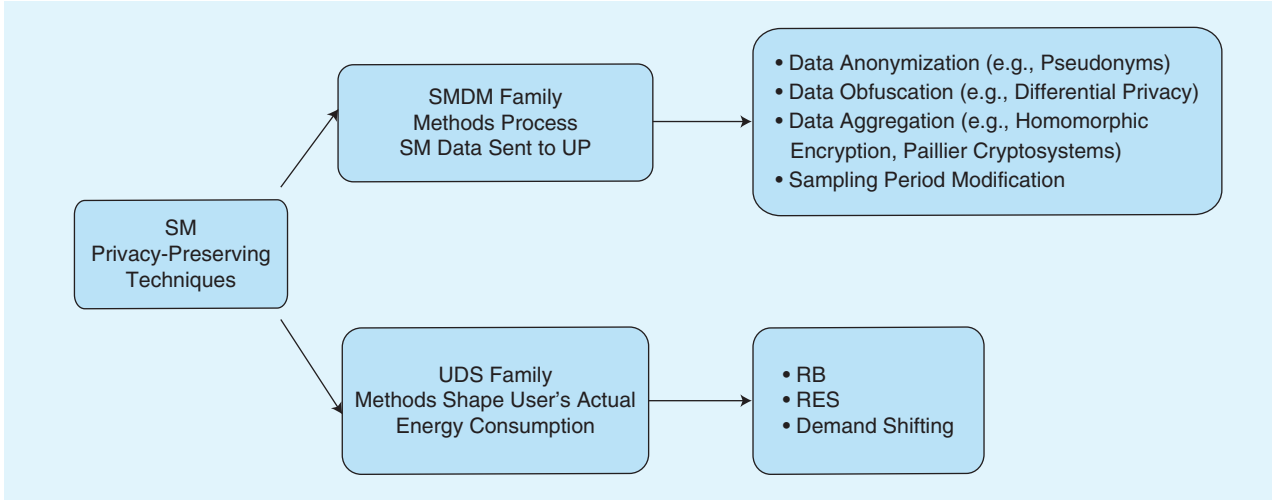


FIGURE 3. An overview of the privacy-enabling approaches for SMs.

load. Note that the effect of using an RB or an RES can also be considered as adding noise to the household consumption, but the noise in this case corresponds to a physical variation in the energy received from the grid. Moreover, different from the approaches in the SMDM family, the SM measurements provided by the UDS methods are exact, and there is no issue of any data mismatch between the SM data and the effective user demand from the grid. Thus, when UDS methods are deployed, the utility of SMs for the SG is not diminished since the users' energy consumption is neither misreported nor distorted. As a result, while the privacy-versus-utility tradeoff is of particular concern for the SMDM techniques, with the UDS methods, SG utility is never diminished. Instead, other tradeoffs are considered, such as privacy versus cost or privacy versus wasted energy. Figure 3 shows an overview of the privacy-enabling approaches.

The focus of this article is on UDS techniques, which have been receiving growing attention from the research community in recent years. The physical resources these techniques rely on, such as RBs or RESs at consumer premises, are already becoming increasingly available, thanks to government incentives and the decreasing cost of solar panels and household and electric vehicle RBs. Moreover, shaping and filtering users' actual energy consumption by means of physical resources render any data misreporting or distortion unnecessary and thus, do not undermine the utility of the SG concept itself.

We present a signal processing perspective on SM privacy by treating the user load as a stochastic time series, which can be filtered and distorted by using an RB, an RES, and/or demand shaping/scheduling. The available energy generated by the RES can also be modeled as a random sequence, whose statistics depend on the energy source (e.g., solar or wind) and the specifications of the renewable energy generator. Additionally, the finite-capacity battery imposes instantaneous limitations on the available energy. We also note that such physical resources can be used as well for cost minimization purposes by the users, e.g., by acquiring and storing energy over low-cost periods and utilizing the stored energy in the RB and the energy generated by an RES over peak-cost periods. Accordingly, we also study the tradeoff between privacy and cost, as well as the minimization of the wasted renewable energy. Next, we describe and summarize the progress made in recent years toward quantifying SM privacy leakage in a rigorous manner, report the most significant results, and highlight a number of future research directions.

Current household batteries, typical energy demands, and renewable energy generation

Table 2 lists the storage capacity and peak power of some of the currently available RBs for residential use. As can be seen, the capacities are in the range of a few kilowatthours. For example, the peak power that batteries with a 4-kWh capacity

Table 2. The specifications of some currently available residential batteries.

Household RB	Capacity (kWh)	RB Charging Peak Power (kW)	RB Discharging Peak Power (kW)
Sunverge SIS-6848 [76]	7.7, 11.6, 15.5, 19.4	6.4	6
SonnenBatterie eco [77]	4–16	3–8	3–8
Tesla Powerwall 2 [78]	13.5	5	5
LG RESU 48V [79]	2.9, 5.9, 8.8	3, 4.2, 5	3, 4.2, 5
Panasonic battery system LJ-SK56A [80]	5.3	2	2
Powervault G200-LI-2/4/6KWH [81]	2, 4, 6	0.8, 1.2	0.7, 1.4
Orison Panel [82]	2.2	1.8	1.8
SimpliPhi PHI 3.4–48V [83]	3.4	1.5	1.5

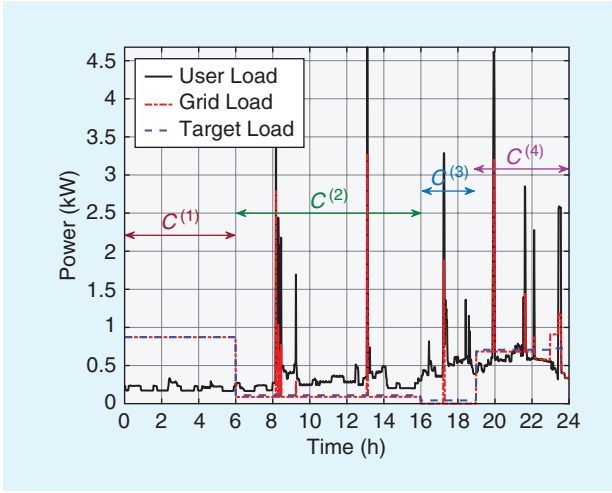


FIGURE 4. Examples of the user load, grid load, and target load over the course of a day when a piecewise target profile is considered [33]. The price periods are highlighted by arrows of different colors. Note that the target load assumes a different constant value for each price period.

can output sustainably is on the order of 1–2 kW. However, as the typical electricity consumption is very spiky (e.g., see Figure 1), current batteries cannot fully hide the consumption spikes, because of the charging/discharging peak power constraints. For example, while a 4-kWh battery can hide a constant consumption of 2 kW over 2 h, it cannot fully conceal spikes in the user load of more than 2 kW. An example of this effect can be noticed in the simulation of Figure 4.

The typical household average power consumption also lies within the range of a few kilowatts, as shown in Table 3, where the distribution of the average user power consumption values over different years obtained from various databases is reported, with various time resolutions. Analyzing the Dataport database [8], we observe that, independent of the period considered, the average user energy demand is fewer than 2 kWh 80–90% of the time. Current batteries charged at full capacity would then be able to satisfy the demand continuously for only a few hours. However, completely covering the consumption over a few hours may come at the expense of revealing the energy consumption fully at future time periods. In fact, once the RB is discharged, it needs to be charged again before being able to hide the user consumption; hence, the use of the RB introduces memory into the system, as decisions taken at a certain time have an impact on the privacy performance at later times.

We should also remark that residential electricity consumption is forecast to increase significantly in the coming years [71], emphasizing the need to intelligently exploit limited-capacity storage devices to hide energy consumption behavior. We also would like to emphasize that the privacy leakage is caused mostly by these spikes, which are typically more informative (e.g., the oven, microwave, and heater) compared to more regular consumption (e.g., the refrigerator). Moreover, because of electricity price variations, users may prefer charging/discharging the battery during certain time periods, which limits the available energy that can be used for privacy. Finally, it is expected that the increasingly wider adoption of electric vehicles and the mass production and adoption of energy-hungry smart devices will inevitably increase the typical household electricity consumption, further limiting RBs' capability to fully hide the user load.

Table 4 shows the average power generated by typical residential solar panels, which are the most common residential RESs. The location, technology, inclination, and size of the solar panel affect the generated power, as shown in Table 5 for one of the databases considered, where *kWp* denotes the kilowatt peak, i.e., the output power achieved by a panel under full solar radiation. As expected, around 50% of the time, i.e., at night, no energy is generated at all, while there are differences in the distribution of the average values for the two databases considered, due to the different locations. Comparing these

Table 3. The distribution of the average household power consumption (resolution refers to the measurement frequency).

Source	Location	Resolution	Time Frame	Number of Houses	$[0,0.5]$ kW	$(0.5,1]$ kW	$(1,2]$ kW	$(2,3]$ kW	$(3,4]$ kW	$(4,+\infty)$ kW
Dataport [8]	Texas, United States	60 min	1 January–31 May 2016	512	38	30	20	7	3	2
			1 January–31 December 2015	703	36	26	20	9	5	4
			1 January–31 December 2014	720	39	25	20	8	4	4
			1 January–31 December 2013	419	35	25	21	9	5	5
			1 January–31 December 2012	182	31	26	24	10	5	5
Intertek [26]	United Kingdom	2 min	1 May 2010–31 July 2011	251	18	24	47	11	0	0
Dred [27]	The Netherlands	1 s	5 July 2015–5 December 2015	1	98	1.8	0.4	0	0	0
Uci [28]	France	1 min	16 December 2006–26 November 2010	1	47	9	28	8	4	2

The values in each column indicate the percentage of time the average consumption falls into the corresponding interval.

Table 4. The distribution of the average power generated by residential photovoltaic systems.

Source	Location	Resolution	Time Frame	Number of Houses	0 kW	(0, 0.5] kW	(0.5, 1] kW	(1, 2] kW	(2, 3] kW	(3, 4] kW	(4, +∞) kW
Dataport [8]	Texas, United States	60 min	1 January 2012–31 May 2016	351	49	17	7	9	7	6	5
Microgen [29]	United Kingdom	30 min	1 January–31 December 2015	100	51.7	36.4	9.8	2	0.1	0	0

The values in each column indicate the percentage of time the average generation falls into the corresponding interval.

Table 5. The specifications of the solar panels studied in the Microgen [29] database.

Solar Panel Area (m^2)					Solar Panel Cell Type		Nominal Installed Capacity (kWp)			
(0, 15]	(15, 20]	(20, 25]	(25, 30]	(30, +∞)	Monocrystalline	Polycrystalline	(0, 2]	(2, 3]	(3, 4]	(4, +∞)
5	35	44	15	1	93	7	4	36	59	1

The values in each column indicate the percentage of solar panels that satisfy the corresponding property.

values with those in Table 2, we note that the battery capacities are large enough to store many hours of average solar energy generated by the solar panels most of the time.

A signal processing perspective on SM privacy

A generic discrete-time SM system model is depicted in Figure 5. In this model, each time slot is normalized to unit time; therefore, the power and energy values within a time slot are used interchangeably. $X_t \in \mathcal{X}$ denotes the total power demanded by the appliances in the household in time slot t, i.e., the user load, where $\mathcal{X}$ is the user load alphabet, i.e., the set of values that X_t can assume. The sequence $\{X_t\}$ represents the user's private information that needs to be protected. $Y_t \in \mathcal{Y}$ is the power received from the grid in time slot t, i.e., the grid load, which is measured and reported to the UP by the SM, while $\mathcal{Y}$ denotes the grid load alphabet. We assume that the user load and grid load power values remain constant within a time slot. In practice, this can be considered as a discrete-time linear approximation of a continuous-load profile. This approximation can be made as accurate as desired by reducing the time slot duration.

In current systems, where no energy manipulation is employed, $Y_t = X_t, \forall t$; that is, the actual energy consumption of the appliances is reported to the UP by the SM. Instead, we will assume that an RB and an RES are available to the user to physically distort the energy consumption, so that what the user receives from the grid, Y_t, does not reveal too much information about the energy used by the appliances, X_t. We remark here that the time slots in our model correspond to time instants when the electricity is actually requested by the user and drawn from the grid, rather than the typically longer sampling interval used for sending SM measurements to the UP. In fact, we assume that the SM measures and records the output power values at each time slot. This is because our aim is to protect consumers' privacy not only from the UP but also from the DSO or any other attacker that may deploy a sensor on the consumer's power line, recording the electricity consumption in almost real time.

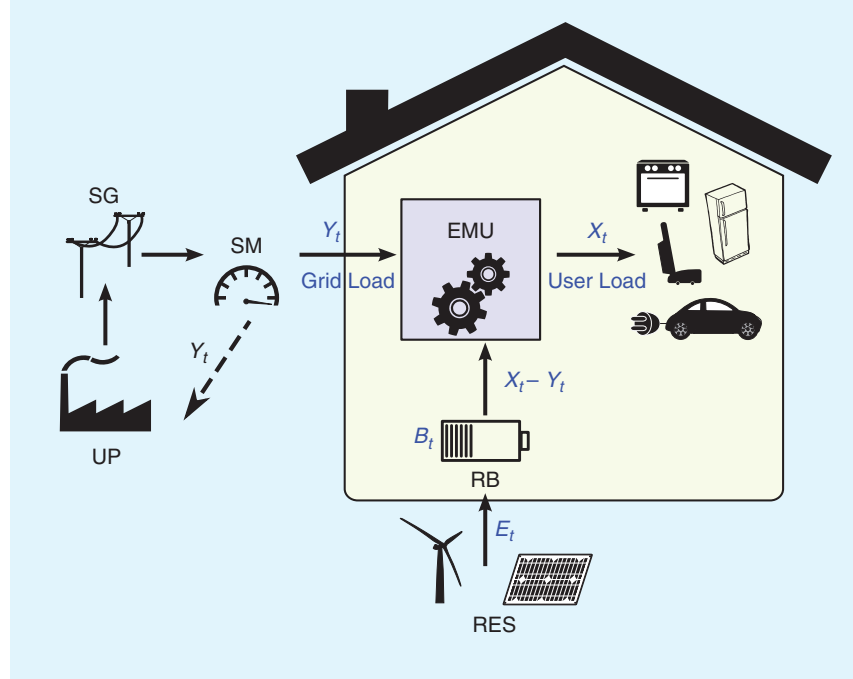


FIGURE 5. The system model. X_t, Y_t, E_t, and B_t denote the consumer's power demand, i.e., the user load; the SM readings, i.e., the grid load; the power produced by the RES; and the battery state of charge at time t, respectively. The dashed line represents the meter readings being reported to the UP. EMU: energy management unit.

The amount of energy stored in the RB at time t is $B_t \in [0, B_{\max}]$, where $B_{\max}$ denotes the maximum battery capacity. $X_t - Y_t$ is the power taken from the RB, and the battery charging and discharging processes are often constrained by the so-called charging and discharging power constraints $\hat{P}_c$ and $\hat{P}_d$, respectively, i.e., $-\hat{P}_c \le X_t - Y_t \le \hat{P}_d, \forall t$. There is also typically a constraint on the average energy that can be retrieved from an RB, imposed by an average power constraint $\bar{P}$, i.e., $\mathbb{E}\left[\frac{1}{n}\Sigma_{t=1}^{n}(X_t - Y_t)\right] \le \bar{P}$. Losses in the battery charging and discharging processes may also be taken into account to model a more realistic energy management system. The renewable energy generated at time t by the RES is represented by $E_t \in \mathcal{E}$, where $\mathcal{E} = [0, E_{\max}]$. RBs and RESs are expensive facilities, and installation and operation costs can be reduced if they are shared by multiple users, e.g., users within the same neighborhood or block of apartments. Moreover, sharing these resources allows the centralized management of the energy system, which also leads to a more efficient use of the available resources. The renewable energy can be stored in the RB or used immediately so that a user can

- increase privacy by not reporting the actual power consumption to the UP
- decrease electricity costs by purchasing and storing electricity from the grid when it is cheaper and using it to satisfy future demand—or even selling it back to the UP when the price increases
- increase energy efficiency by reducing the waste of generated renewable energy when it is not needed and, when it is not profitable, to sell it to the UP.

The random processes X and E are often modeled as Markov processes or as sequences of independent and identically distributed (i.i.d.) random variables. Although the UP typically does not know the instantaneous outcomes of these processes, it may well know their statistics. In some cases, the UP may know the realizations of the renewable energy process E, for example, if it has access to additional information from sensors deployed near the household that measure different parameters, e.g., the solar or the wind power intensity, and if it knows the specifications of the user's renewable energy generator, e.g., the model and size of the solar panel.

Given these definitions, the equation expressing the evolution of the energy level in the battery is

$$B_{t+1} = \min\{B_t + E_t - (X_t - Y_t), B_{\max}\}. \tag{1}$$

Sometimes, the user load does not need to be satisfied immediately in its entirety. In fact, it can be further classified into demand that must be met immediately (e.g., lighting or cooking) and demand that can be satisfied at a later time, the so-called elastic demand (e.g., charging an electric vehicle or running a dishwasher or washing machine). For the latter demand, the user's only concern is that a certain task needs to be finished by a certain deadline (e.g., the electric car must be fully charged by 8 a.m.), and it does not matter exactly when the consumption takes place. This flexibility allows the consumer to employ demand response to increase privacy and lower the energy cost.

The electricity unit cost at time t, denoted by C_t, can be modeled as a random variable or in accordance with a specific ToU tariff. The cost incurred by a user to purchase Y_t units of power over a time interval τ_t at price C_t is thus given by $\tau_t Y_t C_t$. When the presence of an RES is considered, the prosumer may be able to sell part of the energy generated to the grid to further improve privacy and minimize the energy cost. If this occurs, the net metering approach is typically considered, i.e., the utilities purchase consumer-generated electricity at the current retail electricity rate. The battery wear and tear due to charging and discharging the RB can also be taken into account and modeled as an additional cost [30].

The energy management policy

The energy management unit (EMU) is the intelligence of the system, located at the user's premises, where the SM privacy-preservation and cost-optimization algorithms are physically implemented. The energy management policy (EMP), implemented by the EMU, determines at any time t the amount of energy that should be drawn from the grid and the RB, given the previous values of the user load X^t, renewable energy E^t, level of energy in the battery B^t, and grid load Y^{t-1}, i.e.,

$$f_t : \mathcal{X}^t \times \mathcal{E}^t \times \mathcal{B}^t \times \mathcal{Y}^{t-1} \to \mathcal{Y}, \quad \forall t, \tag{2}$$

where $f \in \mathcal{F}$, and $\mathcal{F}$ denotes the set of feasible policies, i.e., policies that produce grid load values that satisfy the RB and RES constraints at any time as well as the battery update equation in (1). The optimal policy is chosen to minimize the long-term information leakage about a consumer's electricity consumption, possibly along with other criteria, such as the minimization of electricity cost or wasted energy. The EMP prevents outages, and typically it is not allowed to draw more energy from the grid to be wasted simply for the sake of increased privacy.

The policy f_t in (2) corresponds to an online EMP, i.e., one in which the action taken by the EMU at any time slot depends only on the information available causally right up to that time. Alternatively, in an offline optimization framework, the policy takes actions based also on future information about the system state, i.e., the user load and RES energy generation, in a noncausal fashion. In the SM privacy literature, both offline and online SM privacy-preserving algorithms have been considered. Online algorithms are more realistic and relevant for real-world applications; however, offline algorithms may lead to interesting intuition or bounds on the performance. Moreover, noncausal knowledge of the electricity price process is a realistic assumption in today's energy networks; and even the noncausal knowledge of power consumption may be valid for certain appliances, such as refrigerators, boilers, heaters, and electric vehicles, whose energy consumption can be accurately predicted over certain finite time frames.

A heuristic privacy measure: Variations in the grid load profile

As in many other problems involving privacy, a wide consensus over the best privacy measures for SMs has not yet been reached. A number of privacy measures have been proposed in the literature, each with its own benefits and limitations. Although it is clear that privacy is achieved when the UP cannot infer a user's behavior on the basis of SM measurements, it is challenging to define a corresponding mathematical measure that is independent of the particular detection technique employed by the attacker.

Grid load variance as a privacy measure

One can argue that privacy in SMs can be ensured by opportunely charging and discharging the RB so that the grid load is always constant. In fact, the differences in consecutive load measurements $Y_t - Y_{t-1}$ are indicative of the appliances' switch-on/off events, the so-called features, and are typically exploited by the existing NILM algorithms. Ideally, a completely flat grid load profile would not reveal any feature and would only leak a user's long-term average power consumption. However, this would require a very large battery capacity and/or a powerful RES. Alternatively, the level of privacy can be measured by what we might call the *distance* of the grid load from a completely flat target load profile, based on the intuition that the smaller the distance, the higher the level of privacy achieved [31]. Accordingly, privacy can be defined as the grid load variance around a prefixed target load profile W,i.e.,

$$\mathcal{V}_n \triangleq \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(Y_t - W)^2], \tag{3}$$

where the expectation is over X_t and Y_t, and typically $W = \mathbb{E}[X]$.

Another important concern for consumers is their energy cost. With the integration of unreliable RESs into the grid, it is expected that the unit cost of energy from different UPs will fluctuate over time. RBs for residential use provide flexibility to consumers, as they can buy and store energy during low-cost periods to be used during peak-price periods. The impact of RBs in reducing the cost of energy to consumers has been extensively studied in the literature [32]. Note, however, that the operation of the EMU to minimize the energy cost does not necessarily align with the goal of minimizing privacy leakage. Therefore, it is essential to jointly optimize the electricity cost and user privacy. If the cost of energy and battery wear and tear are considered, the overall optimization problem becomes

$$\min \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[C_t Y_t + \mathbb{1}_B(t) C_B + \alpha(Y_t - W)^2], \tag{4}$$

where $\mathbb{1}_B(t) = 1$ if the battery is charging/discharging at time t and equals 0; otherwise, C_B is the battery operating cost due to the battery deterioration caused by charging and discharging the RB, and α strikes the tradeoff between privacy and cost. The expectation in (4) is over the probability distributions of all of the involved random variables, i.e., X_t, Y_t, and C_t.

If $W_t = \mathbb{E}[X], \forall t$, the EMU tries to achieve a flat grid load profile around the average user energy consumption with as few deviations as possible. This scenario is illustrated in Figure 6, where the straight blue line is the fixed target consumption profile W_t and the red line indicates the achieved grid load profile Y_t. For i.i.d. X and C processes, an online EMP can be obtained using Lyapunov optimization [30]. The online control algorithm can be formulated as a Lyapunov function with a perturbed weight, and the drift-plus-penalty framework is adopted, which is typically used for stabilizing a queuing network, by minimizing the so-called drift while at the same time minimizing a penalty function. Here, the penalty is represented by the optimization target, while the Lyapunov drift is defined as the difference of the level of energy in the RB at successive time instants. The authors in [30] show that this approach leads to a mixed-integer nonlinear program, which they solve by decomposing it into multiple cases and finding a closed-form solution to each of them.

This problem can also be studied in an offline framework by assuming that the future user demand profile can be accurately estimated for a certain time horizon and that the energy cost is known in advance. When privacy and cost of energy are jointly optimized over a certain time horizon, one can characterize the points on the Pareto boundary of the convex region formed by all of the cost and privacy leakage pairs by solving the following convex optimization problem [31]:

$$\min_{Y_t \geq 0} \sum_{t=1}^{n}[(1-\alpha) Y_t C_t + \alpha(Y_t - W)^2]. \tag{5}$$

It is shown in [31] that the optimal offline solution has a water-filling interpretation. However, unlike the classical water-filling algorithm, which appears as the solution of the power allocation problem across parallel Gaussian channels under a total power constraint, here the water level is not constant but changes across time because of the instantaneous power constraints.

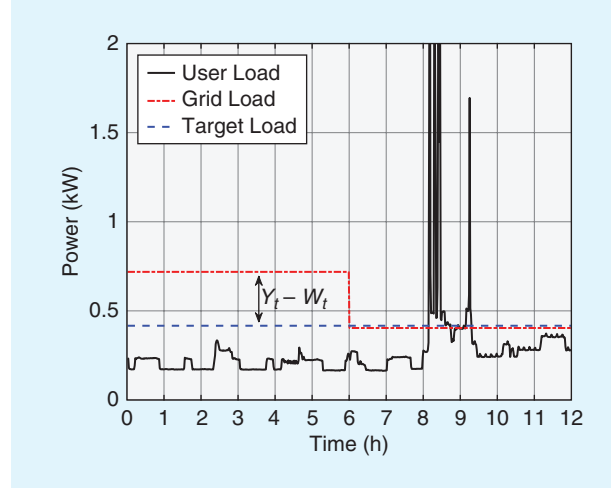


FIGURE 6. An example of the user load, grid load, and constant target load profiles, where the distance $Y_t - W_t$ is highlighted. The aim of the algorithms presented in the "Grid Load Variance as a Privacy Measure" section is to minimize the average squared distance.

A completely flat consumption profile may not be feasible or even desirable—e.g., if the cost varies greatly during the system operation because of ToU tariffs. Thus, it is reasonable to assume that a user requests more energy during off-peak price periods as compared to peak price periods and hence allows a piecewise constant target load [33]. An example of this strategy is shown in Figure 4, where it is applied to real power consumption data from the UK-DALE data set [34]. The optimization problem (5) becomes

$$\min_{Y_t, W^{(i)}} \frac{1}{N} \sum_{i=1}^{M} \left[\sum_{t=t_{c^{(i-1)}}}^{t_{c^{(i)}}} (1-\alpha) Y_t C^{(i)} + \alpha (Y_t - W^{(i)})^2 \right], \tag{6}$$

where $C^{(i)}$ and $W^{(i)}$ are the cost of the energy purchased from the UP and the target profile during the ith price period, respectively, where $1 \le i \le M$; M is the total number of price periods during time T; and the ith price period spans from time slot $t_{c^{(i-1)}}$ to $t_{c^{(i)}}$. Figure 7 depicts the timing convention considered in this scenario. Energy can be sold to the UP to further improve the privacy-versus-cost tradeoff, as assumed in [33]. Considering a piecewise target profile improves the overall privacy-versus-cost tradeoff compared to a constant target profile, as shown in Figure 8 for a Powervault G200-LI-4KWH RB when using power consumption data from [34].

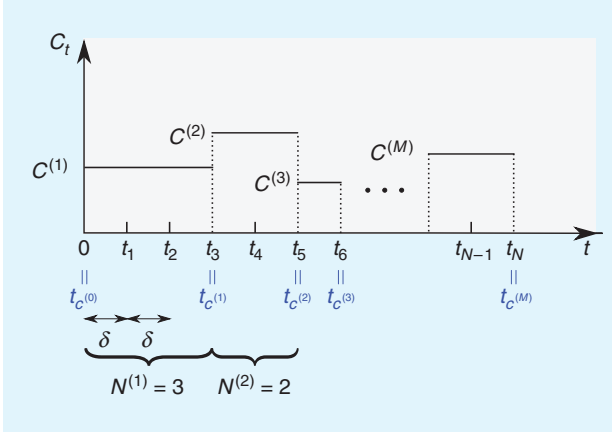


FIGURE 7. The ToU tariff and timing conventions used for a piecewise target profile [33]. The time instants at which the price of energy changes are represented by $t_{c^{(i)}}$, for $i = 1, \ldots, M$, and $t_{c^{(0)}} = 0$.

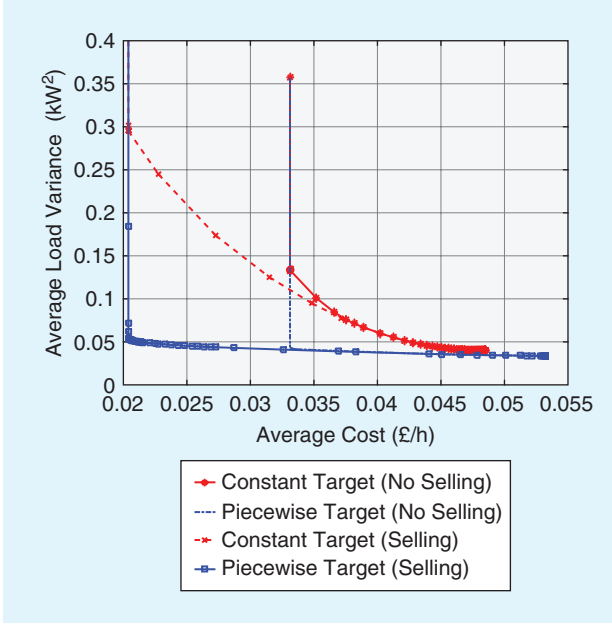


FIGURE 8. The privacy-versus-cost tradeoff when using a Powervault G200-LI-4KWH battery for the strategies in [31] and [33].

A possible extension of the latter work is to consider the multiuser scenario, where, in principle, each user can fix its own target profile. As long as the target profile does not depend on the user's energy consumption profile, the UP does not receive much information about the consumer's activities. On the other hand, the UP can implicitly incentivize users to choose different target profiles by setting different ToU prices for different consumers. Since consumers will tend to buy more energy when it is cheaper, each of the users in the neighborhood will shift the load to a different time slot, also balancing the total load on the grid.

Markov decision process formulation

In the online optimization framework, where the user load and the energy generated by the RES can be modeled as Markov processes (or as i.i.d. sequences as a special case), the SM privacy problem can be cast as a Markov decision process (MDP). An MDP is a discrete-time state-transition system that is formally characterized by the following:

- a state space
- an action space, which includes the possible actions that can be taken by the decision maker at each state
- the transition probabilities from the current state to the next state, which describe the dynamics of the system
- the reward (or inversely the cost) process, which indicates the reward received (or cost incurred) by the decision maker by taking a particular action in a particular state.

The goal of an MDP is to find the optimal policy that minimizes the average (or discounted) cost either by a specified time in the future, i.e., by considering the so-called finite-horizon setting, or over an indefinite time period, by considering the infinite-horizon setting. To solve the corresponding MDP, the Bellman optimality equations should be formulated [35], which can be solved to obtain the optimal policy at each state and time instant. The problem can be solved numerically for the finite-horizon setting, while the value iteration algorithm can be employed to obtain the optimal stationary policy in the infinite-horizon scenario.

In the SM problem, the state at any time t is typically represented by a combination of the current level of energy in the battery B_t, user demand X_t, and renewable energy E_t. The action, performed by the EMU, is represented by the current grid load and the energy used from the RB and RES. State transitions are modeled by the battery update equation, which is typically assumed to be deterministic, and by transitions in the user demand and renewable generation states, which typically do not depend on users' actions. The cost function is the privacy

loss that is experienced when moving from one level of energy in the battery to another by following a certain action.

However, to consider privacy as the cost function in an MDP, it is necessary to formulate the privacy leakage in an additive form across time, so that the total loss of privacy over multiple time slots is given by the summation of the privacy leakage at different time slots. This may be challenging, depending on the privacy measure employed. For example, measuring privacy via the squared distance of the grid load from a constant target profile has a straightforward additive formulation, while the same does not hold when privacy is measured by the mutual information (MI) between the user and grid load sequences. This is because the MI takes into account the dependence between the realization of the user load at time t, X_t and the current, past, and future realizations of $Y, Y_1, \ldots, Y_t, \ldots$.

When the state and action spaces are continuous, it is necessary to discretize them to solve the problem numerically. The accuracy of the numerical solution can be improved by decreasing the discretization step size, though at the expense of significantly higher computational complexity. When the dimensions of the state and action spaces render numerical evaluation of the optimal policy unfeasible, one can resort to suboptimal solutions that are easier to optimize and compute numerically yet may provide near-optimal performance or interesting intuition. Also, when the information-theoretic privacy measures are used, it may be possible to simplify the infinite-horizon optimization problem and write it in a single-letter form. We will provide further insight into this next.

The SM problem is cast as an MDP in [36], where the loss of privacy is measured by the fluctuations of the grid load around a constant target load, and the joint optimization of privacy and cost is studied. The optimal privacy-preserving policies are characterized by minimizing the expected total cost. Denote by u_t the action at time t. To solve the MDP, the transition probabilities $p(X_t \mid X_{t-1})$ and $p(B_t \mid B_{t-1}, u_t)$ need to be known; however, this is normally not the case, as the user load and the energy storage usage are typically nonstationary. The authors in [36] overcome this issue by adopting the Q-learning algorithm [37], which is an iterative algorithm used for characterizing the expected cost for each state–action pair by alternating the exploitation and exploration phases. The corresponding offline optimization policy is also characterized in [36] to be considered as a benchmark for the online algorithm. The paper characterizes the privacy–cost tradeoff curves and also evaluates the performance of the proposed algorithm by means of the empirical mutual information.

Temporal and spatial similarities in the grid load as a privacy measure

Variations in the grid load profile can be captured by considering the power traces of single appliances and computing differences in power consumption both in the time domain, i.e., the consumption deviation over time of a specific appliance, and in the space domain, i.e., the consumption profiles of different appliances. As these variations are computed over a certain time horizon, when an online algorithm is considered, future user electricity consumption is estimated by forecasting the future electricity prices and running Monte Carlo simulations. The optimal decision at any time is characterized by considering both the current inputs and the forecasts through a rolling online stochastic optimization process.

Load shifting, i.e., the scheduling of the user's flexible electricity demand in accordance with privacy as well as cost concerns, can also be considered. Load shifting is analyzed in [38] and [39], where privacy, cost of energy, and battery wear and tear are jointly optimized and an online algorithm is formulated. The objective is to minimize the sum of the current and expected electricity and charging/discharging costs together with the weighted power profile differences measured through the similarity parameters for an entire day. In [39], the effectiveness of three similarity measures are examined separately and jointly, considering only four typical appliances—an oven, a clothes dryer, a dishwasher, and an electric vehicle—for the sake of simplicity.

Heuristic algorithms

While the grid load can be flattened by minimizing its variation around a constant consumption target, several works in the literature propose heuristic battery charging and discharging algorithms that keep the grid load variations limited. An intuitive approach is to try to keep the grid load equal to its most recent value by discharging (or charging) the RB when the current user load is larger (or smaller) than the previous one. This approach, called the *best-effort* (*BE*) *algorithm* in [40], tends to eliminate the higher-frequency components of the user load while still revealing the lower-frequency components.

In [40], the similarity between the two probability distributions of the user and grid loads is quantified via the empirical relative entropy, i.e., the Kullback–Leibler (KL) divergence [41]. In the same work, the authors also consider cluster classification, whereby data are clustered according to power levels and cross-correlation and regression procedures, according to which the grid load is shifted in time at the point of maximum cross-correlation with the user load, and regression methods are then used to compare the two aligned signals.

The study in [42] considers a slightly more sophisticated approach, called the *nonintrusive load-leveling* (*NILL*) *algorithm*, in which more than one grid load target value, namely, a steady-state target and low and high recovery state targets, are allowed, and where the EMU tries to maintain the grid load at one of these values across time. If the steady-state load cannot be maintained, the EMU switches to a high (or low) recovery state in case of persistent light (or heavy) user demand. When one of the recovery states is reached, the target load is adapted accordingly to permit the battery to charge or discharge, similar to the empirical strategies outlined in [43]. The value of the steady-state target load can be updated whenever a recovery state is reached, to reduce the occurrences of recovery states, which is achieved by using an exponential weighted moving average of the demand. To assess their proposed approach, the authors in [42] count the number of features, i.e., the number of times a device is recognized as being on or off based on the grid load as compared to the user load.

As also pointed out in [44], these heuristic algorithms suffer from precise load change recovery attacks that can identify peaks of user demand. We note that the NILL algorithm essentially quantizes the input load to three values with the help of the RB. This idea is generalized in [44] by considering an arbitrary number of quantization levels. Since quantization is a many-to-few mapping, converting the grid load to a step function is inherently a nonlinear and irreversible process, which can be used to provide privacy by maximizing the quantization error under battery limitations. More specifically, the grid load is forced to be a multiple of a quantity β, i.e., $Y_t = h_t\beta$, where h_t is an integer value and β is the largest value that satisfies the battery's maximum capacity and power constraints. At any time slot, given the user load, the grid load is chosen between the two levels adjacent to the user load, namely, $\left\lceil \frac{X_t}{\beta} \right\rceil$ and $\left\lfloor \frac{X_t}{\beta} \right\rfloor$, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor functions, respectively.

The study in [44] proposed three stepping algorithms that have different quantization levels: 1) the lazy stepping algorithm, which tries to maintain the external load constant for as long as possible; 2) the lazy charging algorithm, which keeps charging (or discharging) the battery until it is full (or empty); and 3) the random charging algorithm, which chooses its actions at random. While the simulation results show that these algorithms outperform the BE and NILL algorithms, with the lazy stepping algorithm typically performing the best, it is hard to make general claims because of the heuristic nature of these algorithms. In fact, these approaches do not provide theoretical guarantees on the level of privacy achieved. Thus, they are not able to make any general claim about the strength of the proposed privacy-preserving approaches and their absolute performance. This is an important limitation, as consumers would like to know the level of privacy they can achieve, even if it is in statistical terms. Also, because such heuristics are often based on deterministic schemes, they are prone to be easily reverse-engineered.

Theoretical guarantees on SM privacy

One of the challenges in SM privacy is to provide theoretical assurances and fundamental limits on the information leaked by an SM system, independently of any assumption on the capability of an attacker or of the particular NILM algorithm employed. This is essential in privacy research, as privacy-preserving techniques may perform extremely well against some NILM algorithms and very poorly against others. Moreover, the privacy assurances should not be based on the complexity limitations of a potential attacker, as techniques that are currently thought to be not feasible may become available to attackers in the future if computational capabilities improve or if new methods are developed.

Last but not least, establishing a coherent mathematical framework would allow us to rigorously compare various SM scenarios and the use of different physical resources, e.g., RBs of various capacities, RESs of various kinds, and so forth. Accordingly, signal processing and information-theoretic tools have been employed in the literature to provide theoretical privacy assurances. We will overview various statistical measures for privacy, particularly conditional entropy [43], Fisher information (FI) [45], and type II error probability for detecting user activity [46].

In this statistical framework, it is commonly assumed that the statistics of the user load and the RES are stationary over the period of interest and known to the EMU. This assumption is reasonable, especially if the period of stationarity is sufficiently long for the EMU to observe and learn these statistics [47]–[49]. On the other hand, an online learning theoretic framework can also be considered to account for the convergence time of the learning algorithm. Alternatively, most of the works in the literature that carry out a theoretical analysis also propose suboptimal policies that can be applied on real power traces, thus allowing the reader to gain an idea of the practical application and performance of these theoretically motivated techniques. We take a worst-case approach and assume that the statistics governing the involved random processes are known by the attacker. Note that this can only empower the attacker and strengthen the stated privacy guarantees.

The significance of single-letter expressions

It is expected that a meaningful privacy measure should consider the leakage of a user's information over a certain time period of reasonable length, because of the memory effects introduced by the RB and the RES. The energy consumption over a short period of time can be easily covered by satisfying all of the demand from the RB or the RES over this period, but this may come at the expense of fully revealing the energy consumption at future time periods. Therefore, the information-theoretic analysis typically considers an average information rate measured over a given finite time period and often studies its infinite-horizon asymptotics as well.

However, increasing the time horizon also increases the problem complexity, and one of the challenges of the information-theoretic analysis is to obtain a so-called single-letter expression for the optimal solution, which would significantly reduce the problem complexity, particularly when the involved random variables are defined over finite alphabets. Unfortunately, to date, closed-form or single-letter expressions for the information leaked in an SM system have been characterized only for specific settings under various simplifications, e.g., considering an i.i.d. or Markov user load or RES generation.

MI as a privacy measure

The entropy of a random variable X, $H(X)$ is a measure of the uncertainty of its realization. The MI between random variables X and Y, $I(X;Y)$, measures the amount of information shared between the two random variables [41]. The MI can also be considered as a measure of dependence between the random variables X and Y, and it is equal to zero if and only if they are independent. Rewriting the MI as $I(X;Y) = H(X) - H(X \mid Y)$, where $H(X \mid Y)$ is the conditional entropy, we can also interpret MI as the average reduction in the uncertainty of X from the knowledge of Y. Therefore, we can measure the privacy leakage about the input load sequence X^n through the SM readings Y^n by the MI between the two sequences $I(X^n; Y^n)$. This will measure the

reduction in the uncertainty of the UP about the real energy consumption of the appliances X^n after receiving the SM measurements Y^n. For an SM system with only an RB (but no RES) and a given EMP f in (2) running over n time slots, the average information leakage rate $\mathcal{I}_f^n(B_{\max}, \hat{P}_d)$ is defined as

$$\mathcal{I}_f^n(B_{\max}, \hat{P}_d) \triangleq \frac{1}{n} I(X^n; Y^n) = \frac{1}{n}[H(X^n) - H(X^n \mid Y^n)], \quad (7)$$

where $0 \le X_t - Y_t \le \hat{P}_d$. The parameters $B_{\max}$ and $\hat{P}_d$ emphasize the dependence of the EMP and, therefore, of the achievable information leakage rate on the battery capacity and the discharging peak power constraint. The optimal EMP and the corresponding minimum information leakage rate are obtained by minimizing (7) over all of the feasible policies $f \in \mathcal{F}$ to obtain $\mathcal{I}^n(B_{\max}, \hat{P}_d)$.

Privacy with an RES

Alternatively, one can also consider the SM system of Figure 5 with an RES but no RB. Assume that the renewable energy that can be used over the operation period is constrained by an average and a peak power constraint. We do not allow selling the generated renewable energy to the UP, as our goal is to understand the impact of the RES on providing privacy to the user. The minimum information leakage rate achieved under these assumptions and for an i.i.d. user load can be characterized by the so-called privacy-power function $\mathcal{I}(\bar{P}, \hat{P}_d)$ and formulated in the following single-letter form:

$$\mathcal{I}(\bar{P}, \hat{P}) = \inf_{p_{Y|X} \in \mathcal{P}} I(X; Y), \quad (8)$$

where $\mathcal{P} \triangleq \{p_{Y|X} : y \in \mathcal{Y}, \mathbb{E}[(X - Y)] \le \bar{P}, 0 \le X - Y \le \hat{P}\}$. This formulation is presented in [50] for a discrete user load alphabet (i.e., X can assume only values that are multiples of a fixed quantum) and in [51] for a continuous user load alphabet (i.e., X can assume any real value within the limits specified by the peak power constraints of the appliances). The optimal EMP that minimizes (8) is stochastic and memoryless; that is, the optimal grid load at each time slot is generated randomly via the optimal conditional probability that minimizes (8) by only considering the current user load. Another interesting observation is that (8) is in a form similar to the well-known rate-distortion function in information theory, which characterizes the minimum compression rate R of data, in bits per sample, that is required for the receiver to reconstruct the source sequence within a specified average distortion level D [41]. Formally, the rate-distortion function $R(D)$ for an i.i.d. source $X \in \mathcal{X}$ with distribution p_X, reconstruction alphabet $\hat{\mathcal{X}}$, and distortion function $d(\hat{x}, x)$, where the distortion between sequences X^n and $\hat{X}^n$ is given by $\frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$, characterizes the minimum rate with which an average distortion of D is achievable. The compression rate specifies the size of the codebook 2^{nR} required to compress the source sequence of length n, X^n. Shannon showed that the rate-distortion function can be obtained in the following single-letter form:

$$R(D) = \min_{p_{\hat{X}|X}: \sum_{(x,\hat{x})} p_X p_{\hat{X}|X} d(x,\hat{x}) \le D} I(\hat{X}; X). \quad (9)$$

The analogy between (8) and (9) becomes clear considering the following distortion measure:

$$d(x, y) = \begin{cases} x - y, & \text{if } 0 \le x - y \le \hat{P}, \\ \infty, & \text{otherwise}, \end{cases} \quad (10)$$

and such an analogy enables the use of tools from rate-distortion theory to evaluate the privacy-power function for an SM system. However, it is important to highlight that, despite the functional similarity, there are major conceptual differences between the two problems: 1) in the SM privacy problem, Y^n is the direct output of the encoder rather than the reconstruction at the decoder side, and 2) unlike the lossy source encoder, the EMU does not operate over blocks of user load realizations; instead, it operates symbol by symbol, acting instantaneously after receiving the appliance load at each time slot.

For discrete user load alphabets, the grid load alphabet can be constrained to the user load alphabet without loss of optimality [52], and, since MI is a convex function of the conditional probability $p_{Y|X} \in \mathcal{P}$, the privacy-power function can be written as a convex optimization problem with linear constraints. Algorithms such as the Blahut–Arimoto (BA) algorithm can be used to numerically compute the optimal conditional distribution [41]. For continuous user load distributions, the Shannon lower bound is derived in [52], which is a computable lower bound on the rate-distortion function widely used in the literature and is shown to be tight for exponential user load distributions.

These results can be generalized to a multiuser scenario in which N users, each equipped with a single SM, share the same RES [52]. This scenario is represented in Figure 9, where the objective is to minimize the total privacy loss of N consumers (or devices) considered jointly, rather than minimizing the privacy loss for each of them separately. This requires the EMU to allocate the shared RES among all of the users in the most effective manner. The average information leakage rate can still be written as in (7), by replacing X_t and Y_t with $\mathbf{X}_t = [X_{1,t}, \ldots, X_{N,t}]$ and $\mathbf{Y}_t = [Y_{1,t}, \ldots, Y_{N,t}]$, where the boldface characters denote the vectors representing the N power measurements. The privacy-power function has the same expression as in (8), and, for the case of independent but not

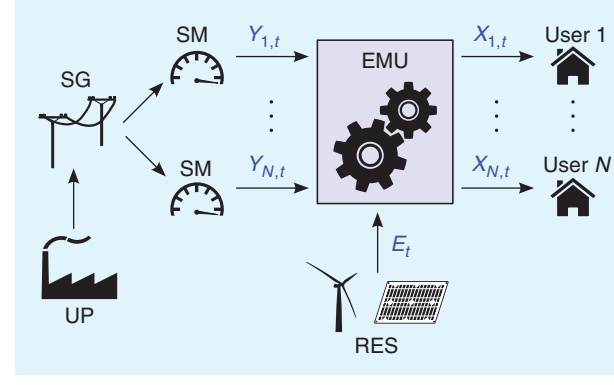


FIGURE 9. A single EMU and RES are shared among N users, each equipped with an SM. The EMU decides how much energy each user can retrieve from the RES and from the grid.

necessarily identically distributed user loads, the optimization problem (ignoring the peak power constraint) can be cast as

$$\mathcal{I}(\bar{P}) = \inf_{\sum_{i=1}^{N} P_i \le \bar{P}} \sum_{i=1}^{N} \mathcal{I}_{X_i}(P_i), \tag{11}$$

where $\mathcal{I}_{X_i}(\cdot)$ denotes the privacy-power function for the ith user having user load distribution $p_{X_i}(x_i)$. For continuous and exponential user loads, the optimal allocation of the energy generated by an RES can be obtained by the reverse water-filling algorithm, according to which energy from the RES is used only to satisfy the users with a low average load, while users with higher average load need to request energy from the grid as well.

Privacy with an RB

We can also consider the presence of just an RB in the system, which is thus charged only via the grid (no RES is available to the EMU). Including an RB complicates the problem significantly, and the level of energy in the battery B_t plays an important role when designing a feasible EMP.

This problem can be solved by putting it in the form of an MDP and finding a suitable additive formulation for the privacy cost function [53]. The optimization problem is formulated as

$$L^* \triangleq \min_{f} \frac{1}{n} I(B_1, X^n; Y^n), \tag{12}$$

where f can be any feasible policy, as specified in (2) (without including the renewable energy process). The approach in [53] casts (12) in an additive formulation by noting that there is no loss of optimality in restricting the focus to charging strategies f' that decide on the grid load, based only on the current values of the user load X_t and level of energy in the battery B_t and on the past values of the grid load Y^{t-1}. That is, the general strategy f in (2) is specified as $f_t' : \mathcal{X} \times \mathcal{B} \times \mathcal{Y}^{t-1} \to \mathcal{Y}, \forall t$, because of the following inequality:

$$\frac{1}{n} I(X^n, B_1; Y^n) \ge \frac{1}{n} \sum_{t=1}^{n} I(X_t, B_t; Y_t \mid Y^{t-1}). \tag{13}$$

The conditional distributions in (13) grow exponentially with time because of the term Y^{t-1}, and the problem becomes computationally infeasible very quickly. To overcome this issue, the knowledge of Y^{t-1} is summarized into a belief state, defined as $p(X_t, B_t \mid Y^{t-1})$, which can be computed recursively and interpreted as the belief that the UP has about (X_t, B_t) at time t, given its past observations Y^{t-1}. This way, the Bellman equations can be formulated, and the optimal policy can be identified numerically (with a discretization of the belief state).

For an i.i.d. user load, the single-letter characterization of the minimum information leakage rate is given by [53] as

$$J^* \triangleq \min_{\theta \in \mathcal{P}_{\mathcal{B}}} I(B - X; X), \tag{14}$$

where θ is the probability distribution over B, given the past output and actions, i.e., $\theta_t \triangleq p\left(b_t \middle| y^{t-1}, a^{t-1}\right)$, and the action a_t is defined as the transition probability from the current belief, user load, and level of energy in the battery to the current grid load. This result is obtained by considering a belief on $W_t \triangleq B_t - X_t$, rather than (B_t, X_t), and by further restricting to policies of the type $f_t'' : \mathcal{W} \times \mathcal{Y}^{t-1} \to \mathcal{Y}, \forall t$. Since (14) is convex in θ, the optimal θ^* may be obtained by using the BA algorithm. The optimal grid load turns out to be i.i.d. and indistinguishable from the demand, while the optimal policy is memoryless, and the distribution of Y_t depends only on W_t. Such a characterization is provided in [54] for a binary i.i.d. user load, while the authors extend it to an i.i.d. user load of generic alphabet size in [53], [55], and [56].

Another approach is to model the level of energy in the RB as a trapdoor channel [57]. In a trapdoor channel, a certain number of red or blue balls are within the channel, and a new ball of either color is inserted into it as the channel input at each time step. After the new ball is inserted, one of the balls present in the channel is randomly selected and removed from the channel. In an SM setting, the finite-capacity RB can be viewed as a trapdoor channel, whereby inserting or extracting a ball from the channel represents charging or discharging the RB, respectively. An upper bound on the information leakage rate is characterized in [58] through this model by minimizing the information leakage rate over the set of stable output balls, i.e., the set of feasible output sequences Y^n that can be extracted from the channel, given a certain initial state and an input sequence X^n and by taking inspiration from codebook construction strategies in [59]. Such an upper bound is characterized in [58] as

$$\frac{1}{n} I(X^n; Y^n) \le \frac{1}{\left\lfloor (B_{\max} + 1)/X_{\max} \right\rfloor}, \tag{15}$$

where $X_{\max}$ is the largest value X can assume. It is also shown in [58] that the average user energy consumption determines the level of achievable privacy.

Apart from only maximizing privacy, it is of interest to also minimize the cost. Different from privacy, the cost of energy has an immediate additive formulation and can be easily incorporated into the MDP construction. Considering the random price vector $C^t = (C_1, \ldots, C_t)$, where C_t denotes the unit cost of energy at time slot t, privacy can be defined in the long time horizon as

$$\mathcal{P} \triangleq \lim_{t \to \infty} \frac{H(X^t \mid Y^t, C^t)}{t}. \tag{16}$$

This formulation is presented in [43], where the corresponding MDP is constructed and two suboptimal algorithms are proposed. The first is a greedy algorithm, which maximizes at any time the current instantaneous reward, while the second is a battery-centering approach that is aimed at keeping the battery at a medium level of charge so that the EMU is less constrained by the battery or the demand in determining the grid load. In the latter approach, if the grid load depends not on the current user load or the battery level but only on the current electricity price, the system is said to be in a hidden state, while it is said to be in a revealing state otherwise. The latter strategy is analyzed for an i.i.d. user load by considering the system as a recurrent Markov chain and adopting random walk theory.

Privacy with both an RES and an RB

When both an RES and an RB are present, the information-theoretic privacy analysis becomes more challenging. As an initial step, we can consider infinite and zero battery capacities, which represent, respectively, lower and upper bounds on the privacy leakage achievable for a practical SM system with a finite-capacity battery [60], [61]. When $B_{\max} = \infty$, the problem can be shown to be equivalent to the average and peak power-constrained scenario, and, interestingly, the privacy performance does not deteriorate, even if the UP knows the exact amount of renewable energy generated. This shows that keeping the renewable energy generation process private is more critical when the RB has a limited capacity.

Two different EMPs are shown to achieve the lower bound in [61]. In the BE policy, at any time slot, the optimal EMP derived from (11) is employed independently of the level of energy in the RB if there is sufficient energy in the RB, while otherwise all of the energy request is satisfied from the grid. The latter approach leads to full leakage of user consumption, but it can be shown that these events are rare enough that the information leakage rate does not increase. In the alternative store-and-hide policy, an initial storage phase is employed, during which all of the user's energy requests are satisfied from the grid while all of the generated renewable energy is stored in the battery. In the following hiding phase, the EMU deploys the optimal policy designed under average and peak power constraints.

On the other extreme, when $B_{\max} = 0$, the renewable energy that can be used at any time slot is limited by the amount of energy generated within that period. As expected, assuming that the UP has knowledge of the renewable energy process significantly degrades the privacy performance for this scenario. Figure 10 compares the minimum information leakage rate with respect to the renewable energy generation rate p_e for $|\mathcal{X}| = |\mathcal{E}| = |\mathcal{Y}| = 5$ when $B_{\max} = \{0, 1, 2, \infty\}$. In this figure, the curves for a finite battery capacity of $B_{\max} = 1$ and $B_{\max} = 2$ are obtained numerically by considering a suboptimal EMP [61].

The presence of a finite-capacity battery increases the problem complexity dramatically because of the memory effects induced by the finite battery, and single-letter expressions are still lacking for this scenario. A possible approach to find a theoretical solution to this problem is by extending the MDP formulation, as investigated in [62].

Detection error probability as a privacy measure

So far, we have considered approaches that try to hide the complete user energy demand from the UP. However, rather than hiding the entire energy consumption profile, in some cases it may be more meaningful to keep specific user activities private, such as whether there is anyone at home, whether the alarm has been activated, or whether someone is eating microwaved food. To keep the answer to such details private, the goal of the EMU is to maximize the attacker's probability of making errors when attempting to discern them.

Let the consumer's behavior that needs to be kept private belong to a set of M possible activities. Thus, we can treat the attacker's decision and the user's action as an M-ary hypothesis test, i.e., $H \in \mathcal{H} = \{h_0, h_1, \ldots h_{M-1}\}$. When $M = 2$, the hypothesis test is said to be binary, and, by convention, the hypothesis h_0, called the *null hypothesis*, represents the absence of some factor or condition, while the hypothesis h_1, called the *alternative hypothesis*, is the complementary condition. For example, answering the question, "Is somebody at home?" corresponds to a binary hypothesis test, where h_0 is the hypothesis "somebody is not at home" and h_1 is the hypothesis "somebody is at home." It is reasonable to assume that the input load will have different statistics under these two hypotheses; accordingly, we assume that under hypothesis h_0 (h_1), the energy demand at time slot t is i.i.d. with $p_{X|h_0}$ $(p_{X|h_1})$. Based on the SM readings, the attacker aims at determining the best decision rule $\hat{H}(\cdot)$, i.e., the optimal map between the SM readings and the underlying hypothesis. In other words, the space of all possible SM readings $\mathcal{Y}^n$ is partitioned into two disjoint decision regions $\mathcal{A}_0$ and $\mathcal{A}_1$, defined as follows:

$$\mathcal{A}_0 \triangleq \left\{ y^n \middle| \hat{H}(y^n) = h_0 \right\}, \tag{17}$$

$$\mathcal{A}_1 \triangleq \left\{ y^n \middle| \hat{H}(y^n) = h_1 \right\}, \tag{18}$$

which correspond to the subsets of the SM readings for which the UP decides for one of the two hypotheses. The attacker's binary hypothesis test can incur two types of errors:

- type 1 error, in which a decision h_1 is made when h_0 is the true hypothesis (a false positive or false alarm); the type 1 error probability is $p_{\mathrm{I}} = p_{Y^n|h_1}(\mathcal{A}_0)$
- type 2 error, in which a decision h_0 is made when h_1 is the true hypothesis (a false negative or miss); the type 2 error probability is $p_{\mathrm{II}} = p_{Y^n|h_0}(\mathcal{A}_1)$.

The Neyman–Pearson test minimizes the type 2 error probability for a fixed maximum type 1 error probability and makes decisions by thresholding the likelihood ratio

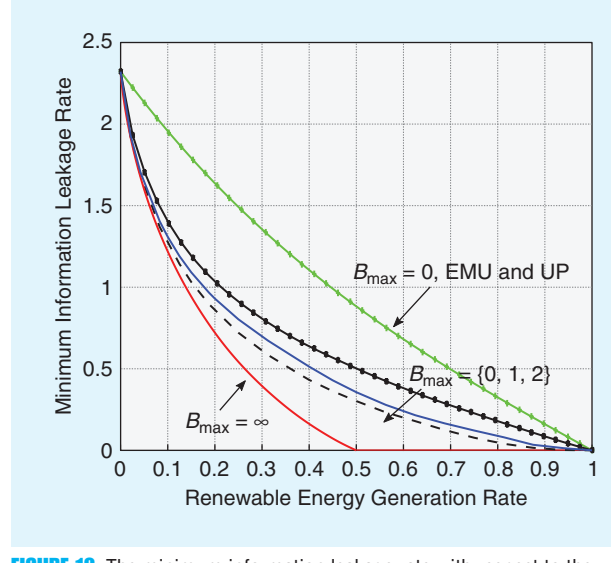


FIGURE 10. The minimum information leakage rate with respect to the renewable energy generation rate p_e with $\mathcal{X} = \mathcal{E} = \mathcal{Y} = \{0, 1, 2, 3, 4\}$. The leakage for $B_{\max} = \infty$ has been found by setting $\hat{P} = 4$ [61].

$(p_{Y^n|h_0}(y^n|h_0))/(p_{Y^n|h_1}(y^n|h_1))$. Consider the worst case of an all-powerful attacker that has perfect knowledge of the EMP employed in the asymptotic regime $n \to \infty$ and denote by $p_{\text{II}}^{\min}$ the minimal type 2 probability of error subject to a constraint on the type 1 error probability. Assuming that a memoryless EMP is employed by the EMU—i.e., the grid load at any time slot t depends only on the input load at the same time slot—then the attacker runs a Neyman–Pearson detection test on the grid load. We note that the memoryless EMP assumption is not without loss of optimality. However, it is justified on the grounds that characterizing the more general optimal policy with memory seems to be significantly more challenging and is unlikely to lend itself to a single-letter expression. The Chernoff–Stein lemma [41] links the minimal type 2 error probability $p_{\text{II}}^{\min}$ to the KL divergence $D(\cdot\|\cdot)$ between the grid load distributions conditioned on the two hypotheses in the limit of the number of observations going to infinity:

$$\lim_{n\to\infty} -\frac{\log p_{\text{II}}^{\min}}{n} = D(p_{Y|h_0}\|p_{Y|h_1}), \tag{19}$$

where the KL divergence between two probability distribution functions on X, p_X and q_X, is defined as [41]

$$D(p_X\|q_X) \triangleq \sum_{x\in\mathcal{X}} p_X(x)\log\frac{p_X(x)}{q_X(x)}. \tag{20}$$

Not surprisingly, to maximize privacy, the goal of the EMU is to find the optimal grid load distribution, which, given the user load X and the true hypothesis H, minimizes the KL divergence in (19) or, equivalently, minimizes the asymptotic exponential decay rate of $p_{\text{II}}^{\min}$. However, the EMU is constrained by the available resources in making the two input load distributions produce similar grid load distributions. In particular, we impose a constraint on the average RES it can use. Thus, the objective is to solve the following minimization problem:

$$\min_{p_{Y|H}\in\mathcal{P}_{Y|H}} D(p_{Y|h_0}\|p_{Y|h_1}), \tag{21}$$

where $\mathcal{P}_{Y|H}$ is the set of feasible EMPs, i.e., those that satisfy the average RES generation rate $\bar{P}$, so that $(1/n)\mathbb{E}[\Sigma_{i=1}^{n} X_i - Y_i|h_j] \le \bar{P}$, with $j = 0, 1$. This setting is studied in [63], where the asymptotic single-letter expressions of two privacy-preserving EMPs in the worst-case scenario are considered, i.e., when the probability of a type 1 error is close to 1. The first policy is a memoryless hypothesis-aware policy that decides on Y_t based only on the current X_t and H, while the second policy is unaware of the correct hypothesis H but takes into account all of the previous realizations of X and Y.

It is noteworthy that, even if the hypothesis-unaware policy with memory does not have access to the current hypothesis, it performs at least as well as the memoryless hypothesis-aware policy. This is because the hypothesis-unaware policy is able to learn the hypothesis with negligible error probability after observing the energy demand process for a sufficiently long period. Additionally, the energy supply alphabet can be constrained to the energy demand alphabet without loss of optimality, which greatly simplifies the numerical solution to the problem.

FI as a privacy measure

FI is another statistical measure that can be employed as a measure of SM privacy [45]. Let some sample data x be drawn according to a distribution depending on an underlying parameter. Then, FI is a measure of the amount of information that x contains about the parameter. In the SM setting, Y^n is the sample data available to the attacker, while X^n is the parameter underlying the sample data that is to be estimated by the UP. Let $\hat{X}^n$ denote the estimate of the UP. The FI can be generalized to the multivariate case by the FI matrix, defined as

$$\mathcal{FI}(x^n) = \int_{y^n\in\mathcal{Y}^n} p\left(y^n|x^n\right)\left[\frac{\partial\log\left(p\left(y^n|x^n\right)\right)}{\partial x^n}\right]\left[\frac{\partial\log\left(p\left(y^n|x^n\right)\right)}{\partial x^n}\right]^T \mathrm{d}y^n. \tag{22}$$

Assuming an unbiased estimator at the attacker, i.e., the difference between the estimator's expected value and the true average value of the parameter being estimated is zero, the variance of the estimation error can be bounded via the Cramér–Rao bound as follows:

$$\mathbb{E}\left[\left\|x^n - \hat{x}^n(y^n)\right\|_2^2\right] \ge \mathrm{Tr}(\mathcal{FI}(x^n)^{-1}), \tag{23}$$

where $\left\|x^n - \hat{x}^n(y^n)\right\|_2^2$ denotes the squared Euclidean norm and $\mathrm{Tr}(A)$ represents the trace of the matrix A. To maximize the privacy, it is then necessary to maximize the trace of the inverse of the FI matrix. In [45], two SM settings with RBs are studied, specifically when the battery charging policy is independent of the user load and when it is dependent noncausally on the entire user load sequence. For both cases, single-letter expressions are obtained for the maximum privacy. Moreover, the case of biased estimators, wear and tear on the batteries, and peak power charging and discharging constraints are also briefly analyzed in [45].

Empirical MI as a privacy measure

Approaches aimed at determining theoretical privacy limits provide important insights and intuition for the optimal EMP to limit privacy leakage. However, they are often difficult to optimize or even evaluate numerically, and the relatively simplified formulations obtained in various special cases rely on restrictive assumptions, e.g., i.i.d. user load and infinite RB capacity. An alternative is to follow a suboptimal or heuristic EMP. Although such a policy does not provide theoretical privacy guarantees, one can evaluate the corresponding privacy leakage numerically using empirical MI.

One way to compute the empirical MI is by simulating a discrete time system for a large enough time interval and sampling

the resulting X^n and Y^n sequences [64]. The MI between two observed sequences x^n and y^n can be approximated as

$$I(X;Y) \approx -\frac{1}{n}\log p(y^n) - \frac{1}{n}\log p(x^n) + \frac{1}{n}\log p(x^n, y^n), \quad (24)$$

where $p(y^n)$, $p(x^n)$ and $p(x^n, y^n)$ are calculated recursively through a sum–product computation. When using this method, the RB is modeled as a finite state machine (FSM), and the level of energy in the RB evolves in time through a Markov chain with transition probabilities depending on the specific policy implemented. An example of an FSM is illustrated in Figure 11, where all the processes are considered to be binary and Bernoulli distributed, and the parameters are $q_x = \Pr\{X = 1\}$, $p_e = \Pr\{E = 1\}$ and p_v, the latter being the probability of using energy from the battery, provided there is available energy. The support space for the parameters is discretized, and the optimal combination of parameters is found, which minimizes the empirical MI.

This approach is followed in [65], where only a binary RB is present and for an i.i.d. Bernoulli-distributed user demand, and in [66] where an RES is also considered. The latter work also analyzes the wasted energy and characterizes the privacy-versus-energy efficiency tradeoff for the binary scenario and the equiprobable user load and renewable energy generation processes. For larger battery capacities and for an equiprobable user load, the authors note that there is a symmetry and complementarity in the optimal transition probabilities in the FSM model, which simplifies the numerical analysis. This model is also employed in [60] and [61] by considering an RES and designing a suboptimal policy, which, at each time instant, decides among using all of the available energy, half of it, or no energy at all, according to a probability chosen to minimize the overall information leakage.

Another technique for approximating MI is to assume X and Y to be i.i.d. over a time interval and approximate the MI via the relative frequency of events (X_t, Y_t) during the same time window. In [67], this approach is enriched by additive smoothing, i.e., avoiding zero probability estimates by adding a positive scalar, and it is employed together with a model-distribution predictive controller, such that, at each time slot t, the EMU chooses its actions for a prediction horizon of length T, i.e., up to time $t + T$. Privacy and cost are jointly optimized by considering noncausal knowledge of the renewable energy generation process, user load, and energy prices, while the EMU's actions, i.e., the energy that is requested from the grid and the battery, are forecast over the prediction horizon. The user and grid load processes are assumed to be i.i.d. within a time window $N \gg T$, which also includes the prediction horizon T, and the finite alphabets $\mathcal{X}$ and $\mathcal{Y}$ are considered. As $N \gg T$, first-order Taylor approximation of the logarithm function is used, and the corresponding mixed-integer quadratic program is formulated, which is of manageable size and can be solved recursively whenever new SM readings are available.

Results show that considering a relatively small prediction horizon T prevents the EMU from fully utilizing the RB capacity, as the user load that is considered by the algorithm is generally smaller than the RB capacity. Allowing a longer prediction interval dramatically improves the performance in terms of both privacy and cost, at the expense of a much higher computational complexity. The work also shows that by increasing the alphabet sizes of $\mathcal{X}$ and $\mathcal{Y}$, better privacy performance can be achieved.

Empirical MI normalized by the empirical entropy of the user load is considered in [68], where an RB is used to minimize the energy cost subject to privacy constraints. Here, two cost tariffs are considered, a low price and a high price, and a dynamic programming approach is developed to maximize the energy stored in the battery at the end of the low-price period and minimize it at the end of the high-price period. At every time slot, the optimal probability distribution of the grid load is computed, which is forced to be independent of the user load distribution.

Concluding remarks and future challenges

Privacy and the so-called right to be let alone are considered to be an individual's inalienable, fundamental rights, which are safeguarded in many national constitutions worldwide. In Europe, the General Data Protection Regulation (GDPR) [69],

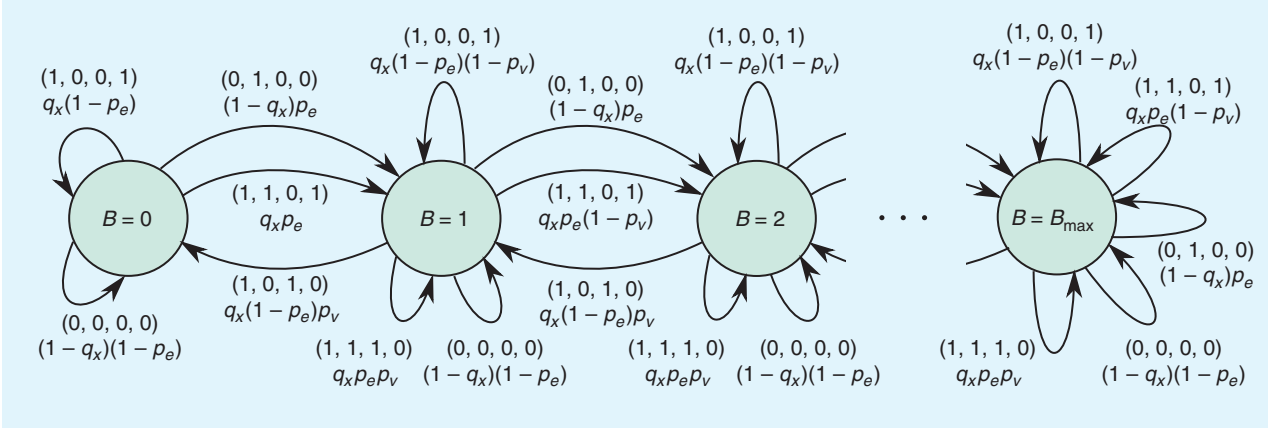


FIGURE 11. An example of RB evolution modeled as a finite state machine, with $\mathcal{B} = \{0, 1, \ldots B_{\max}\}$ and $\mathcal{X} = \mathcal{E} = \mathcal{Y} = \{0, 1\}$. The (x, e, v, y) represent, for every time t, the values of the user load, the renewable energy produced, the energy taken out of the battery by the EMU, and the grid load, respectively [61].

which went into effect on 25 May 2018, sets even more stringent requirements for every technology or device that collects and processes customer data, including SMs. For these reasons, addressing the SM privacy problem is crucial for the adoption of the SG concept. In fact, considering consumers' growing privacy concerns over SMs and many other emerging technologies [70], a critical growth in the adoption of SMs and other SG technologies will take place only when consumers are given full control of their privacy and feel they have clear and honest information on how their data are being used. Only then can consumer resistance be overcome and users' trust be assured, thus paving the way to a more fertile and fair ground for new products and increased innovation in this domain.

UPs and their partners, including governments, may be too keen on collecting users' data indiscriminately and not well incentivized to develop privacy-enhancing technologies. Therefore, legislators, public commissions, consumer advocacy groups, and researchers have important roles to play in tackling the SM privacy problem and preventing SM data from being gathered haphazardly and sold to third parties without explicit user consent or even passed to government intelligence agencies for mass surveillance. The GDPR is a good example of the initiatives that are needed.

However, given that such a legal framework is still lacking and not yet fully developed globally, it becomes imperative to push forward the concept of privacy by design, according to which privacy should be designed in to new products and services rather than considered only after user complaints and regulatory impositions. This is because more options are available during the design stage as compared to the completion stage, when the product has to be modified following a privacy incident or a user complaint. Achieving privacy by design is the ultimate goal of the techniques analyzed in this article.

In this article, we have focused exclusively on techniques that adopt physical resources, such as RESs and RBs, to provide privacy to users. The main motivation and benefits of these techniques is that they do not undermine the value of the SG concept. Each of the outlined techniques has its unique advantages and disadvantages and focuses on a particular aspect of privacy. However, despite the considerable efforts put into developing SM privacy-preserving techniques, the full extent of the SM privacy problem is far from completely understood, and a unified and coherent vision for SM privacy (just as in many other domains) is still elusive.

In the context of SM privacy, UDS-based methods manipulate a physical quantity, energy, to ensure privacy for users. This entails that physical constraints, such as those related to an RB or an RES, play a crucial role in finding the optimal privacy-preserving strategy. We expect that the techniques developed for enhancing SM privacy can prove useful in other privacy-sensitive settings in which physical quantities are involved, such as gas and water meters and location privacy.

Research challenges

Various challenges must be addressed before privacy by design can become a reality in SM systems. First, a generic privacy measure or a combination of different measures must be determined and adopted to formally quantify loss of privacy, in the same way a user's electricity bill is computed. Such a measure should be device independent and should enable the comparison of various privacy-preserving strategies. It is also necessary to understand the implications of the various privacy measures on the grid load. From this point of view, theoretical measures may be preferable because of their abstract and fundamental nature, i.e., they are independent of any assumptions about the attacker's algorithms. However, their relevance in real-world scenarios must be assessed further, and, if necessary, valid suboptimal privacy measures or algorithms should be put forward and standardized as a proxy for more rigorous privacy assurances.

Another important goal is to give consumers as much flexibility as possible in setting their desired level of privacy, trading off privacy with the cost of electricity or other services. It is also essential to allow consumers the possibility of setting different privacy requirements for different devices, as users may consider the information about the usage of a certain device as more sensitive compared to others. This may happen because certain devices are naturally more correlated to the user's activities or presence at home, such as the use of a teakettle, a microwave, or an oven, or because a user may decide to hide the usage of a certain appliance for personal reasons.

In the near future, a wider use of electric vehicles will also bring additional complications to the SM privacy problem, as mobility patterns may be inferred by analyzing the charging and discharging events. This problem can be tackled by load shifting, which is expected to play an important role in jointly optimizing electricity cost and privacy. Load shifting, as well as other privacy-preserving techniques introduced here, will be more accurate and relevant thanks to the development of reliable prediction techniques for future electricity consumption, e.g., by using machine-learning techniques. The proliferation of various energy-hungry smart devices will complicate the problem further and overburden RBs even more.

Finally, the use of shared physical resources should also be investigated in more depth, as cities are becoming more and more densely populated and consumers may want to team up to install storage devices or energy generators that are still rather costly. In cities, solar panels or mini wind turbines may be installed on the roofs of blocks of apartments, and RBs may be put in communal areas; these resources can be used jointly by all of the users in a building. Such resource-sharing models make the privacy problem even more complicated and challenging and might call for a game-theoretic formulation of the problem.

Overall, we hope that presenting this overview of the SM privacy problem and current solutions will further encourage research and development in this area, so that remaining open issues will be solved and SMs' full potential will be realized.

Acknowledgments

Giulio Giaconi gratefully acknowledges the Engineering and Physical Sciences Research Council (EPSRC) of the United Kingdom for funding his Ph.D. studies (award reference #1507704). This work was also supported in part by the

EPSRC through the project COPES (#173605884), and by the U.S. National Science Foundation under grants CNS-1702808 and ECCS-1549881.

Authors

Giulio Giaconi (g.giaconi@imperial.ac.uk) received his B.Sc. and M.Sc. degrees (honors) in communications engineering from Sapienza University of Rome, Italy, in 2011 and 2013, respectively, and his Ph.D. degree from the Department of Electrical and Electronic Engineering, Imperial College London, in 2018. He is currently a research scientist with BT Applied Research, Security Futures Practice. In 2013, he was a visiting student with Imperial College London, working on indoor localization via visible light communications. His current research interests include cybersecurity, data privacy, information and communication theory, signal processing, and machine learning. In 2014, he received the Excellent Graduate Student Award from Sapienza University of Rome.

Deniz Gündüz (d.gunduz@imperial.ac.uk) received his B.S. degree from Middle East Technical University and his M.S. and Ph.D. degrees from the New York University Polytechnic School of Engineering in 2004 and 2007, respectively. He is a reader in information theory and communications in the Department of Electrical and Electronic Engineering, Imperial College London. He is an editor of *IEEE Transactions on Communications* and *IEEE Transactions on Green Communications and Networking.* He was the recipient of the IEEE Communications Society Communication Theory Technical Committee Early Achievement Award in 2017; a starting grant of the European Research Council in 2016; the IEEE Communications Society Best Young Researcher Award for the Europe, Middle East, and Africa Region in 2014; and the Best Paper Award at the 2016 IEEE Wireless Communications and Networking Conference.

H. Vincent Poor (poor@princeton.edu) is the Michael Henry Strater University Professor of Electrical Engineering at Princeton University, New Jersey. His interests include information theory and signal processing, with applications in wireless networks, energy systems, and related fields. He is an IEEE Fellow, a member of the National Academy of Engineering and National Academy of Sciences, and a foreign member of the Chinese Academy of Sciences and the Royal Society. He received the IEEE Signal Processing Society Technical Achievement and Society Awards in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Sc. *honoris causa* from Syracuse University, also in 2017.

References

[1] U.S. Energy Information Administration. (2016). International energy outlook 2016. U.S. Energy Inform. Admin. Washington, D.C. Tech. Rep. DOE/EIA-0484(2016). [Online]. Available: https://www.eia.gov/outlooks/ieo/pdf/0484(2016).pdf

[2] BP Global. (2016). BP energy outlook, 2016 edition. BP p.l.c. London. [Online]. Available: https://biomasspower.gov.in/document/Reports/bp-energy-outlook-2016.pdf

[3] ERGEG. (2011). Final guidelines of good practice on regulatory aspects of smart metering for electricity and gas. European Regulators Group for Electricity and Gas. Brussels, Belgium. Tech. Rep. E10-RMF-29-05. [Online]. Available: http://www.smartgrids-cre.fr/media/documents/ERGEG_Guidelines_of_good_practice.pdf

[4] Department of Energy and Climate Change. (2014). Smart metering equipment technical specifications: Version 2. Dpt. Energy and Climate Change, London. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381535/SMIP_E2E_SMETS2.pdf

[5] Markets and Markets, "Smart meters market by type (electric, water, and gas), application (commercial, residential, and industrial), technology (automatic meter reading and advanced metering infrastructure), and by region: Global forecasts to 2022," Rep. 2477, 2017.

[6] "Smart grid data analytics market—Global industry analysis, size, share, growth, trends and forecast 2015–2022," Transparency Market Res., Albany, NY, Tech. Rep. TMRGI 3966, 2015.

[7] Navigant Research. (2016). Market data: Smart meters. Navigant Research, Washington, D.C. [Online]. Available: https://www.navigantresearch.com/reports/market-data-smart-meters

[8] Pecan Street Inc. Dataport. [Online]. Available: https://dataport.cloud/

[9] G. Hart, "Prototype nonintrusive appliance load monitor," MIT Energy Lab. Tech. Rep. and Electric Power Res. Inst., Cambridge, MA, Tech. Rep. RP 2568-2, 1985.

[10] G. Hart, "Nonintrusive appliance load monitoring," *Proc. IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[11] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter," in *Proc. ACM Workshop Embedded Sensing Systems Energy Efficiency Building*, Zurich, 2010, pp. 61–66.

[12] A. Prudenzi, "A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recordings at meter panel," in *Proc. IEEE Power Engineering Society Winter Meeting*, New York, 2002, pp. 941–946.

[13] E. Quinn, "Privacy and the new energy infrastructure," *Social Sci. Res. Network*, 2009. doi: 10.2139/ssrn.1370731.

[14] I. Rouf, H. Mustafa, M. Xu, W. Xu, R. Miller, and M. Gruteser, "Neighborhood watch: Security and privacy analysis of automatic meter reading systems," in *Proc. ACM Conf. Computer and Communications Security*, Raleigh, NC, 2012, pp. 462–473.

[15] C. Cuijpers and B.-J. Koops, "Smart metering and privacy in Europe: Lessons from the Dutch case," in *European Data Protection: Coming of Age*, S. Gutwirth, R. Leenes, P. de Hert, and Y. Poullet, Eds. Dordrecht, The Netherlands: Springer-Verlag, 2012, pp. 269–293.

[16] J. Loeb, "Smart meters: What would it take to stop the national rollout juggernaut?" *IET Eng. Technol.*, vol. 12, no. 5, pp. 32–34, 2017.

[17] Y. Kim, E. Ngai, and M. Srivastava, "Cooperative state estimation for preserving privacy of user behaviors in smart grid," in *Proc. IEEE Int. Conf. Smart Grid Communications.* Brussels, Belgium, 2011, pp. 178–183.

[18] J.-M. Bohli, C. Sorge, and O. Ugus, "A privacy model for smart metering," in *Proc. IEEE Int. Conf. Communications*. Cape Town, South Africa, 2010, pp. 1–5.

[19] M. Backes and S. Meiser, "Differentially private smart metering with battery recharging," in *Proc. Int. Workshop Data Privacy Management and Autonomous Spontaneous Security.* Egham, U.K., 2014, pp. 194–212.

[20] L. Sankar, S. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart meter privacy: A theoretical framework," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 837–846, 2013.

[21] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Proc. Int. Conf. Security and Trust Management.* Athens, Greece, 2010, pp. 226–238.

[22] F. Li, B. Luo, and P. Liu, "Secure and privacy-preserving information aggregation for smart grids," *Int. J. Security Netw.*, vol. 6, no. 1, pp. 28–39, 2011.

[23] R. Petrlic, "A privacy-preserving concept for smart grids," in *Proc. Sicherheit in vernetzten Systemen:18. DFN Workshop*, 2010, pp. B1–B14.

[24] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *Proc. IEEE Int. Conf. Smart Grid Communications*. Gaithersburg, MD, 2010, pp. 238–243.

[25] A. Cárdenas, S. Amin, and G. A. Schwartz, "Privacy-aware sampling for residential demand response programs," in *Proc. ACM Int. Conf. High Confidence Networked Systems*, Beijing, China, 2012.

[26] J.-P. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, and C. Evans, "Household electricity survey: A study of domestic electrical product usage," Intertek Testing and Certification Ltd., Milton Keynes, U.K., Tech. Rep. R66141, 2012.

[27] A. S. N. U. Nambi, A. R. Lua, and R. V. Prasad, "LocED: Location-aware energy disaggregation framework," in *Proc. ACM Int. Conf. Embedded Systems Energy-Efficient Built Environments*, 2015, pp. 45–54.

[28] D. Dheeru and E. Karra Taniskidou. (2017). UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml

[29] University of Sheffield, Sheffield Solar, Microgen database. [Online]. Available: https://microgen-database.sheffield.ac.uk/

[30] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Cost-effective and privacy-preserving energy management for smart meters," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 486–495, 2015.

[31] O. Tan, J. Gómez-Vilardebó, and D. Gündüz, "Privacy-cost trade-offs in demand-side management with storage," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 6, pp. 1458–1469, 2017.

[32] A. Lever, D. Sanders, N. Lehmann, M. Ravishankar, M. Ashcroft, G. Strbac, M. Aunedi, F. Teng, and D. Pudijanto. (2016). Can storage help reduce the cost of a future UK electricity system? Carbon Trust and Imperial Coll. London. [Online]. Available: https://www.carbontrust.com/media/672486/energy-storage-report.pdf

[33] G. Giaconi, D. Gündüz, and H. V. Poor, "Optimal demand-side management for joint privacy-cost optimization with energy storage," in *Proc. IEEE Int. Conf. Smart Grid Communications*, Dresden, Germany, 2017, pp. 265–270.

[34] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, Mar. 2015. doi: 10.1038/sdata.2015.7.

[35] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 2, 3rd ed. Belmont, MA: Athena Scientific, 2007.

[36] Y. Sun, L. Lampe, and V. W. S. Wong, "Smart meter privacy: Exploiting the potential of household energy storage units," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 69–78, 2018.

[37] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[38] J. Wu, J. Liu, X. S. Hu, and Y. Shi, "Privacy protection via appliance scheduling in smart homes," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, Austin, TX, 2016, pp. 1–6.

[39] Z. Chen and L. Wu, "Residential appliance DR energy management with electric privacy protection by online stochastic optimization," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 1861–1869, 2013.

[40] G. Kalogridis, C. Efthymiou, S. Denic, T. Lewis, and R. Cepeda, "Privacy for smart meters: Towards undetectable appliance load signatures," in *Proc. IEEE Int. Conf. Smart Grid Communications*, Gaithersburg, MD, 2010, pp. 232–237.

[41] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley-Interscience, 1991.

[42] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring," in *Proc. ACM Conf. Computer and Communications Security*, Chicago, 2011, pp. 87–98.

[43] J. Yao and P. Venkitasubramaniam, "The privacy analysis of battery control mechanisms in demand response: Revealing state approach and rate distortion bounds," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2417–2425, 2015.

[44] W. Yang, N. Li, Y. Qi, W. Qardaji, S. McLaughlin, and P. McDaniel, "Minimizing private data disclosures in the smart grid," in *Proc. ACM Conf. Computer and Communications Security*, Raleigh, NC, 2012, pp. 415–427.

[45] F. Farokhi and H. Sandberg, "Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries," *IEEE Trans. Smart Grid*, 2017. doi: 10.1109/TSG.2017.2667702.

[46] Z. Li and T. J. Oechetering, "Privacy on hypothesis testing in smart grids," in *Proc. IEEE Information Theory Workshop*, Jerusalem, Israel, 2015, pp. 337–341.

[47] K. Qian, C. Zhou, M. Allan, and Y. Yuan, "Modeling of load demand due to EV battery charging in distribution systems," *IEEE Trans. Power Syst.*, vol. 26, no. 2, pp. 802–810, 2011.

[48] P. A. Leicester, C. I. Goodier, and P. N. Rowley, "Probabilistic analysis of solar photovoltaic self-consumption using bayesian network models," *IET Renew. Power Gen.*, vol. 10, no. 4, pp. 448–455, 2016.

[49] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and Markov models," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1561–1569, 2013.

[50] D. Gündüz and J. Gómez-Vilardebó, "Smart meter privacy in the presence of an alternative energy source," in *Proc. IEEE Int. Conf. Communications*, Budapest, Hungary, 2013, pp. 2027–2031.

[51] J. Gómez-Vilardebó and D. Gündüz, "Privacy of smart meter systems with an alternative energy source," in *Proc. IEEE Int. Symp. Information Theory*, Istanbul, Turkey, 2013, pp. 2572–2576.

[52] J. Gómez-Vilardebó and D. Gündüz, "Smart meter privacy for multiple users in the presence of an alternative energy source," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 1, pp. 132–141, 2015.

[53] S. Li, A. Khisti, and A. Mahajan, "Information-theoretic privacy for smart metering systems with a rechargeable battery," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3679–3695, 2018.

[54] S. Li, A. Khisti, and A. Mahajan, "Structure of optimal privacy-preserving policies in smart-metered systems with a rechargeable battery," in *Proc. IEEE Int. Workshop Signal Processing Advances Wireless Communication*, 2015, pp. 375–379.

[55] S. Li, A. Khisti, and A. Mahajan, "Privacy preserving rechargeable battery policies for smart metering systems," in *Proc. Int. Zurich Seminar Communicatioins*, Zurich, 2016, pp. 121–124.

[56] S. Li, A. Khisti, and A. Mahajan, "Privacy-optimal strategies for smart metering systems with a rechargeable battery," in *Proc. American Control Conf.*, Boston, 2016, pp. 2080–2085.

[57] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3150–3165, 2008.

[58] M. Arrieta and I. Esnaola, "Smart meter privacy via the trapdoor channel," in *Proc. IEEE Int. Conf. Smart Grid Communications*, Dresden, Germany, 2017, pp. 277–282.

[59] R. Ahlswede and A. Kaspi, "Optimal coding strategies for certain permuting channels," *IEEE Trans. Inf. Theory*, vol. 33, no. 3, pp. 310–314, 1987.

[60] G. Giaconi, D. Gündüz, and H. V. Poor, "Smart meter privacy with an energy harvesting device and instantaneous power constraints," in *Proc. IEEE Int. Conf. Communications*, London, 2015, pp. 7216–7221.

[61] G. Giaconi, D. Gündüz, and H. V. Poor, "Smart meter privacy with renewable energy and an energy storage device," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 129–142, 2018.

[62] G. Giaconi and D. Gündüz, "Smart meter privacy with renewable energy and a finite capacity battery," in *Proc. IEEE Int. Workshop Signal Processing Advances Wireless Communication*, Edinburgh, U.K., 2016, pp. 1–5.

[63] Z. Li, T. J. Oechtering, and D. Gündüz, "Smart meter privacy based on adversarial hypothesis testing," in *Proc. IEEE Int. Symp. Information Theory*, Aachen, Germany, 2017, pp. 774–778.

[64] D.-M. Arnold, H.-A. Loeliger, P. Vontobel, A. Kavcic, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3498–3508, 2006.

[65] D. Varodayan and A. Khisti, "Smart meter privacy using a rechargeable battery: Minimizing the rate of information leakage," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Prague, Czech Republic, 2011, pp. 1932–1935.

[66] O. Tan, D. Gündüz, and H. V. Poor, "Increasing smart meter privacy through energy harvesting and storage devices," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1331–1341, 2013.

[67] J. X. Chin, T. T. D. Rubira, and G. Hug, "Privacy-protecting energy management unit through model-distribution predictive control," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 3084–3093, 2017.

[68] J. Koo, X. Lin, and S. Bagchi, "Privatus: Wallet-friendly privacy protection for smart meters," in *Proc. European Symp. Research Computer Security*, Pisa, Italy, 2012, pp. 343–360.

[69] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," *Official J. European Union*, pp. L 119/1–L 119/88, May 2016.

[70] "Consumer intelligence series: Protect.me (an in-depth look at what consumers want, what worries them, and how companies can earn their trust—and their business)," PwC, White Paper, Sept. 2017.

[71] European Environment Agency. (2015, Sept. 4). Total electricity consumption: Ooutlook from IEA. [Online]. Available: https://www.eea.europa.eu/data-and-maps/indicators/total-electricity-consumption-outlook-from-iea/total-electricity-consumption-outlook-from-1

[72] Itron. Itron Centron. [Online]. Available: https://www.itron.com/na/technology/product-services-catalog/products/0/7/5/centron

[73] Honeywell. REX2 meter. [Online]. Available: https://www.elstersolutions.com/en/product-details-na/826/en/REX2_meter

[74] Kamstrup. Omnipower electricity meters. [Online]. Available: https://www.kamstrup.com/en-us/products-and-solutions/smart-grid/electricity-meters

[75] Enel SpA. (2016, June 27). Enel presents Enel Open Meter, the new electronic meter. [Online]. Available: https://www.enel.com/content/dam/enel-com/pressrelease/porting_pressrelease/1666038-1_PDF-1.pdf

[76] Sunverg Energy Inc. Maximize the value of your solar power. [Online]. Available: http://www.sunverge.com/energy-management/

[77] Sonnen. The sonnenBatterie. [Online]. Available: https://sonnen-batterie.com/en-us/sonnenbatterie

[78] Tesla. Powerwall. [Online]. Available: https://www.tesla.com/powerwall

[79] LG Chem. ESS battery. [Online]. Available: http://www.lgchem.com/global/ess/ess/product-detail-PDEC0001

[80] Panasonic. LJ-SK56A residential storage battery system. [Online]. Available: http://www.panasonic.com/au/consumer/energy-solutions/residential-storage-battery-system/lj-sk56a.html

[81] Powervault. [Online]. Available: http://www.powervault.co.uk/technical/technical-specifications/

[82] Orison. Meet Orison: The first all-in-one, self-installable home battery system. [Online]. Available: http://orison.energy/

[83] SimpliPhi Power. Phi 3.5 battery. [Online]. Available: http://simpliphipower.com/product/phi3-4-smart-tech-battery/

SP

Rafael C. Gonzalez

Deep Convolutional Neural Networks

Neural networks are a subset of the field of artificial intelligence (AI). The predominant types of neural networks used for multidimensional signal processing are *deep convolutional neural networks* (CNNs). The term *deep* refers generically to networks having from a "few" to several dozen or more convolution layers, and *deep learning* refers to methodologies for training these systems to automatically learn their functional parameters using data representative of a specific problem domain of interest. CNNs are currently being used in a broad spectrum of application areas, all of which share the common objective of being able to automatically learn features from (typically massive) data bases and to generalize their responses to circumstances not encountered during the learning phase. Ultimately, the learned features can be used for tasks such as classifying the types of signals the CNN is expected to process. The purpose of this "Lecture Notes" article is twofold: 1) to introduce the fundamental architecture of CNNs and 2) to illustrate, via a computational example, how CNNs are trained and used in practice to solve a specific class of problems.

Relevance

After decades of languishing in research laboratories, AI has recently experienced an explosion in worldwide interest as a strategic tool in industry, government, and research institutions. This interest is based on the fact that AI makes it possible for computers to learn from experience, generalize their behavior, and perform tasks that one normally associates with human intelligence. Some applications of AI are well known to the general public, such as computers that beat grand masters at chess, recognize fingerprints, and interpret verbal commands. Other applications are less well known, such as fraud detection, searching for patterns in large amounts of data, and controlling complex industrial processes. As varied as they are, however, all of these applications are based on the same concepts from deep learning. Of particular interest in two-dimensional (2-D) signal processing is automatic recognition of the contents of digital images using deep learning, which is currently being applied with unprecedented success in fields ranging from biometrics, such as face and retinal identification, to visual quality inspection, medical diagnoses, and autonomous vehicle navigation.

The purpose of this "Lecture Notes" article is twofold: 1) to introduce the fundamental architecture of CNNs and 2) to illustrate, via a computational example, how CNNs are trained and used in practice to solve a specific class of problems.

Prerequisites

The only prerequisites for understanding this article are calculus (in particular, differentiation and the chain rule) and linear algebra, both at the undergraduate level.

Background and problem statement

Interest in using computers to perform automated image recognition tasks dates back more than half a century. During the mid 1950s and early 1960s, a class of so-called learning machines [1] caused a great deal of excitement in the field of machine learning. The reason was the development of mathematical proofs showing that basic computing units, called *perceptrons*, when trained with linearly separable data sets, would converge to a solution in a finite number of iterative steps. The solution took the form of coefficients of hyperplanes that were capable of correctly separating these data classes in feature hyperspace. Unfortunately, the basic perceptron was inadequate for tasks of practical significance. Subsequent attempts to extend the power of perceptrons by assembling multiple layers of these devices lacked effective training algorithms, such as those that had created interest in the perceptron itself [2]. This discouraging state of the art changed with the development in 1986 of *backpropagation*, a method for training neural networks

Digital Object Identifier 10.1109/MSP.2018.2842646
Date of publication: 13 November 2018

composed of layers of perceptron-like units [3]. Backpropagation was first applied to 2-D signals in 1989 in the context of what we now refer to as *deep CNNs* [4]. Similar efforts followed at a relatively low level for the next two decades, but it was not until 2012, when publication of the results of the 2012 ImageNet Challenge demonstrated the power of deep CNNs, that these neural nets began to be used widely in image pattern recognition and other imaging applications [5], [6]. Today, CNNs are the approach of choice for addressing complex image recognition tasks and other important fields, which will be mentioned shortly.

Pattern recognition by machine involves the following four basic stages:

1) acquisition
2) preprocessing
3) feature extraction
4) classification.

Acquisition generates the raw input patterns (e.g., digital images); preprocessing deals with tasks such as noise reduction and geometric corrections; feature extraction deals with computing attributes that are fundamental in differentiating one class of patterns from another; and classification is the process that assigns a given input pattern to one of several predefined classes. Feature extraction usually is the most difficult problem to solve, with extensive engineering often being required to define and test a suitable set of features for a given application. CNNs offer an alternative approach that automates the learning of features by utilizing large databases of samples, called *training sets*, that are representative of an application domain of interest.

The problem addressed in this tutorial is to define a CNN-based strategy for extracting features automatically from a large training database and to use those features for accurately recognizing images from both the training database and also from an independent set of test images. This type of problem is by far the predominant application of CNNs, but it is not their only use. CNNs are currently being applied successfully in a number of other areas that include speech recognition, semantic image segmentation, and natural language processing [8]. In each case, the specifics of how CNNs are structured may vary, but their principles of operation are the same as those discussed in this article.

Solution

We approach the solution to the problem stated in the previous section by using a deep modular CNN architecture consisting of layers of convolution, activation, and pooling. The output of the CNN is then fed into a deep, *fully connected neural network* (FCN), whose purpose is to map a set of 2-D features into a class label for each input image. Central to this approach is the ability to use sample training data to learn the operational parameters of each network layer. For this, we use backpropagation as a tool for iteratively adjusting the network weights (also referred to as *coefficients*, *parameters*, and *hyperparameteres*) based on cycling through the training data. Finally, we demonstrate the effectiveness of the solution by training the CNN/FCN system using a large database of handwritten numeric characters and then testing it with a set of images not used in the training phase. As we show in the "A Computational Example" section, the recognition accuracy achieved by the system on the images of both data sets exceeded 99%.

The fundamental operation performed in each stage of a CNN is convolution, from which these neural nets derive their name.

Deep CNNs

Figure 1 shows the basic components of one stage of a CNN. In practice, a CNN can have tens of such stages, interconnected in series. In addition to the number of stages, CNN architectures differ in how the elements of each stage are defined and used, but the basic structure in Figure 1 is fundamental to all of them.

As the figure shows, one stage of a CNN is composed in general of three volumes, consisting, respectively, of *input maps*, *feature maps*, and *pooled feature maps* (or *pooled maps*, for short). Pooled maps are not always used in every stage and, in some applications, not at all. All maps are 2-D arrays whose size generally varies from volume to volume, but all maps within a volume are of the same size. If the input to the CNN is an RGB color image, the input volume will consist of three maps—the red, green, and blue component images, or channels, of the RGB image. The term *input maps volume* comes from the fact that the inputs have height and width (the spatial dimensions of each map) as well as depth, equal to the number of maps in a volume. In the context of our discussion, the input volume to the first stage consists in general of the channels of multispectral images; the input volumes to all other stages are the pooled maps (or feature maps for stages with no pooling) from the previous stage. When present, the number of pooled maps in a stage is equal to the number of feature maps.

The fundamental operation performed in each stage of a CNN is convolution, from which these neural nets derive their name. Although convolution is a ubiquitous operation in signal processing, it is not always explicitly stated that the type of convolution performed in CNNs is, in general, volume convolution, with the restriction that there is no displacement of the convolution kernel volume (also referred to as a *filter*) in the depth dimension. Figure 1 illustrates this concept, in which a kernel volume, shown in yellow, consists of three individual 2-D kernels. It is evident from this figure that the depth of each kernel volume in any stage is always equal to the depth of the input volume to that stage. Convolution is performed between a different 2-D kernel and its corresponding 2-D input map. Because there is no displacement in the depth dimension, a volume convolution in this case is simply the sum of the individual 2-D convolutions. To understand how a CNN works, it helps to focus attention on the result of volume convolution at one pair of spatial coordinates, (x, y).

Let $w_{m,n,k}$ denote the weights of a 2-D kernel associated with the kth map in the input volume, where m and n are variables that index over the kernel height and width. The convolution between this kernel and the kth map, at any specific spatial location, (x, y), of

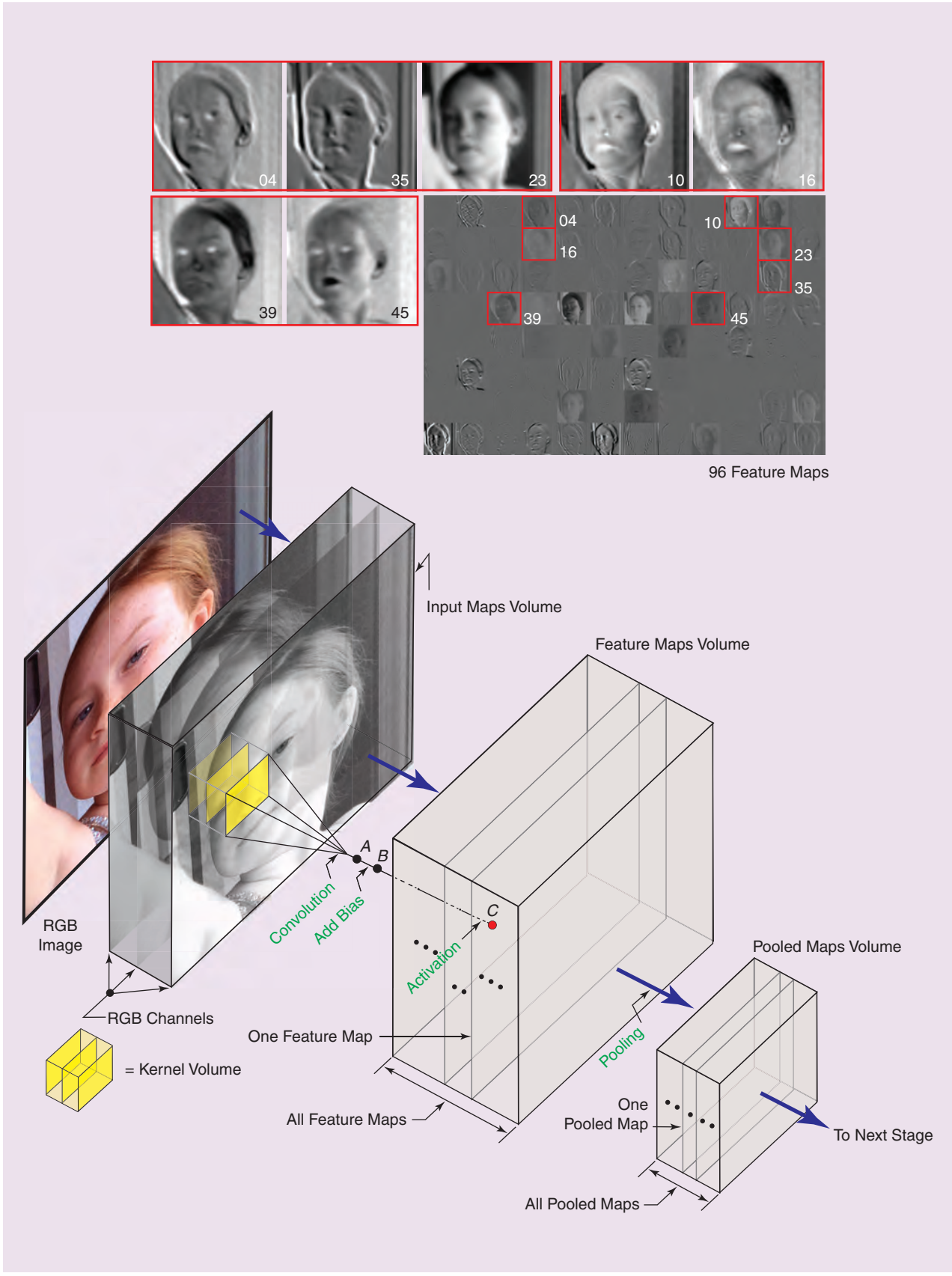


FIGURE 1. The components of one stage of a CNN, consisting of an input maps volume, a feature maps volume, and an optional pooled maps volume. The maps in the input volume correspond to the three channels of the RGB image shown. The stage has 96 feature maps and 96 pooled maps. The highlighted feature maps, displayed as images and identified numerically, illustrate the types of features that a CNN is capable of extracting from an input image.

the map, is the sum of products of the weights of the kernel and the elements of the map that are spatially coincident with the kernel. To obtain a volume convolution, the sum of products operation is performed between each corresponding 2-D kernel and its map at that same spatial location. Each sum of products is a scalar, and the volume convolution at that point is the sum of the K resulting scalars, where K is the depth of the input volume. To write this in equation form would require K 2-D summations. However, for reasons that will be explained in the next section, we can redefine the indices and write the K summations as one:

$$\text{conv}_{x,y} = \sum_i w_i v_i, \tag{1}$$

where the ws are kernel weights, the vs are values of the spatially corresponding elements in the input maps, and $\text{conv}_{x,y}$ is the result of volume convolution at the same spatial coordinates, (x,y), for all maps of the input volume. Equation (1) gives the result at point A in Figure 1. The result at point B is obtained by adding a scalar bias, b, to (1)

$$z_{x,y} = \sum_i w_i v_i + b. \tag{2}$$

We discuss the nature of this bias in the next section.

The result at point C is obtained by passing scalar $z_{x,y}$ through a nonlinearity called an *activation function*, h

$$a_{x,y} = h(z_{x,y}). \tag{3}$$

Activation functions used in practice include sigmoids $h(z) = 1/(1+\exp(-z))$, hyperbolic tangents $h(z) = \tanh(z)$, and so-called rectified linear units (ReLUs) $h(z) = \max(0,z)$. The resulting $a_{x,y}$, called an *activation value,* becomes the value of the feature map at location (x,y), as illustrated by the point labeled C in Figure 1. A complete feature map, also referred to as an *activation map*, is generated by performing the three operations just explained at all spatial locations of the input maps. Each feature map has one kernel volume and one bias associated with it. The objective is to use training data to learn the weights of the kernel volume and bias of each feature map. We explain in the following two sections how these coefficients are learned, and give a detailed computational example of a CNN application.

Figure 1 also illustrates the types of features that volume convolution is able to extract. The input to the CNN stage in Figure 1 was an RGB image of size 277×277 pixels, which resulted in an input volume of depth three, corresponding to the red, green, and blue channels of the RGB image. We used the image of a human subject as the input so that the resulting feature maps would be easier to interpret visually. The feature maps volume in this case was specified to have 96 feature maps, each obtained by filtering the maps of the input volume with a different kernel volume of size $11 \times 11 \times 3$. Thus, there are 96 kernel volumes of depth three, for a total of $3 \times 96 = 288$ 2-D convolution kernels of size 11×11 in this CNN stage. The 96 feature maps resulting from the input image are shown as images in the upper right of Figure 1 as an 8×12 montage. The feature maps shown in enlarged detail are numbered and grouped to illustrate the variety of complementary features that can result from volume convolution. The first group shows three feature maps. Two of them (4 and 35) emphasize edge content, and the third (23) is a blurred version of the input. The second group has two maps (10 and 16) that capture complementary shades of gray (note the difference in the hair intensity, for example). In the third group, feature map 39 emphasizes the subject's eyes and dress, both of which are blue in the input RGB image. Map 45 also emphasizes blue, but it also emphasizes areas that correspond to red tones in the RGB image, such as the subject's lips, hair, and skin. These two feature maps are more sensitive to color content than the maps in the other two groups. Subsequent stages would operate on these feature maps to extract further abstractions from the data, as we illustrate later in the "A Computational Example" section. The weights of the convolution kernel volumes used to generate the 96 feature maps came from AlexNet, a CNN trained using more than 1 million images belonging to 1,000 object categories [5]. The system had never "seen" the image we used in Figure 1.

The pooling, or subsampling, shown in Figure 1 is motivated by studies that suggest that the brains of mammals perform an analogous operation during visual cognition. A pooled map is simply a feature map of lower resolution. A typical pooling method is to replace the values of every neighborhood of size, say, 2×2, in the feature maps by the average of the values in the neighborhood. Using a neighborhood of size 2×2 results in pooled maps of size one-half in each spatial dimension of the size of the feature maps. Thus, a consequence of pooling is significant data reduction, which helps speed up processing. However, a major disadvantage is that map size also decreases significantly every time pooling is performed. Even with neighborhoods of size 2×2 the reduction by half in each spatial dimension quickly becomes an issue when the number of layers is large with respect to the size of the input images. This is one of the reasons why pooling is used only sporadically in large CNN systems. As with activation functions, the type of pooling used also plays a role in defining the architecture of a CNN. In addition to *neighborhood averaging*, two additional pooling methods used in practice are *max pooling*, which replaces the values in a neighborhood by the maximum value of its elements, and *L2 pooling*, in which the pooled value in a neighborhood is the square root of the sum of their values squared. Max pooling has been demonstrated to be particularly effective in classifying large image databases, and it has the added advantage of simplicity and speed. As noted previously, when pooling is used in a layer, each pooled map is generated from only one feature map, so the number of feature and pooled maps is the same.

The basic architecture of each stage of a CNN is defined by specifying the number of feature maps and by whether or not pooling is used in that stage. Also specified are kernel and pooling sizes, and the *convolution stride*, defined as the number of increments of displacement of the kernel between convolution operations. For example, a stride of two means that convolution is performed at every other spatial location in the input

maps. The number of 2-D convolution kernels needed in each stage is equal to the depth of the input volume multiplied by the number of feature maps. The spatial dimensions of all kernels in a stage are the same and are specified as part of the definition of a CNN stage. Generally, the same type of activation is used in all stages of a CNN. This is true also of the size and type of pooling method used when pooling is defined for one or more stages of the network.

There are two major ways in which CNNs are structured: A fully convolutional network (not to be confused with a fully connected network) consists exclusively of stages of the form described in Figure 1, connected in series. The major application of fully convolutional architectures is image segmentation in which the objective is labeling each individual pixel in an input image. Because map size decreases as the number of stages increases, additional processing, such as upsampling, is used so that the output maps are of the same size as the input images. In fact, fully convolutional nets can be connected "end to end" so that map size is first allowed to decrease as a result of convolution and then are run in a reverse process through an identical network whose maps increase from stage to stage using "backward" convolution. The final output is an image of the same size as the input, but in which pixels have been labeled and grouped into regions [8].

The second major way in which CNNs are used is for image classification which, as noted previously, is by far the widest use of CNNs. In this application, the output maps in the last stage of a CNN are fed into an FCN whose function is to classify its input into one of a predetermined number of classes. Because the output volume of a CNN consists of 2-D maps and, as we will show in the next section, the inputs to FCNs are vectors, the interface between a CNN and an FCN is a simple stage that converts 2-D arrays to vectors. A discussion of how all of this is accomplished and applied to solve a specific problem is the subject of the section "A Computational Example."

Deep FCNs

A single perceptron is a computational unit that performs a sum-of-products operation, $z = \Sigma_{i=1}^{n} w_i x_i + w_{n+1}$, between a set of weights, $w_1, w_2, \ldots, w_n, w_{n+1}$, and a set of input scalar pattern features, $x_1, x_2, \ldots x_n$. A vector formed from these features is referred to as a *pattern* (or *feature*) vector. Setting $z = 0$ gives the equation of an n-dimensional hyperplane, where coefficient w_{n+1} is a bias that offsets the hyperplane from the origin of the corresponding n-dimensional Euclidean space. In the "classic" perceptron, the output of the sum-of-products computation is fed into a hard threshold, h, to produce an activation value, $a = h(z)$, with a binary output denoted typically by $[+1, -1]$. Then, if $a = 1$, an input pattern is assigned by the single perceptron to one class, and, if $a = -1$, the pattern is assigned to another. Neural networks are composed of perceptrons in which the activation function is changed from a hard threshold to a smoother function, such as a sigmoid, hyperbolic tangent, or ReLU function, as defined in the previous section. The resulting unit is referred to as an *artificial neuron* because of postulated similarities between its response and the way neurons in the brains of mammals are believed to function.

The objective of training a CNN/FCN network is to determine the weights and biases of convolution volumes in the former, and of the neuron weights and biases in the latter, that solve a given problem.

Figure 2 is a schematic of a deep FCN consisting of layers of artificial neurons in which the output of every neuron in a layer is connected to the input of every neuron in the next layer, hence the term *fully connected*. The input layer is formed from the components of a pattern vector, $x_1, x_2, \ldots, x_n$, and the number of neurons in the output layer is equal to the number of pattern classes in a given application. The input and output layers are visible because we can observe the values of their outputs. All other layers in a neural net are *hidden layers*. Note that CNNs are not fully connected, in the sense that each element of a map in one layer is not connected to every element of maps in the following layer.

The objective of training a CNN/FCN network is to determine the weights and biases of convolution volumes in the former, and of the neuron weights and biases in the latter, that solve a given problem. As noted in the "Background and Problem Statement" section, these parameters are estimated using backpropagation, a methodology for iteratively adjusting the coefficients based on values of the error observed at the output neurons of the FCN.

The computation performed by the zoomed neuron in Figure 2 is

$$z_i(\ell) = \sum_{j=1}^{n_{\ell-1}} w_{ij}(\ell) a_j(\ell - 1) + b_i(\ell), \quad (4)$$

where $w_{ij}(\ell)$ is the weight of the ith neuron in layer ℓ that associates that neuron with the output of the jth neuron in layer $\ell - 1$; $a_j(\ell - 1)$ is the output of the jth neuron in layer $\ell - 1$; $b_i(\ell)$ is the bias of the ith neuron in layer ℓ; and $n_{\ell-1}$ is the number of neurons in layer $\ell - 1$. The output of the ith neuron is obtained by passing $z_i(\ell)$ through a nonlinearity, h, of the form discussed in the previous section:

$$a_i(\ell) = h(z_i(\ell)). \quad (5)$$

These two simple expressions completely characterize the behavior of a neuron in any layer of an FCN. Basically, these equations indicate that the inputs to a neuron in any layer of an FCN are the outputs of all neurons in the previous layer and that the output of that neuron is the sum of products of the neuron weights and its inputs, to which we add a scalar value, and then pass the total sum through a nonlinearity. The important thing to note in (4) and (5) is that they are identical in form to (2) and (3), indicating that CNNs and FCNs perform the same types of neural computations. The ultimate result of this similarity is that training a CNN and an FCN follows the same computational rules, with allowances being made for the fact that CNNs operate on volumes, while FCNs work with vectors.

Training of an FCN begins by assigning small random values to all weights and

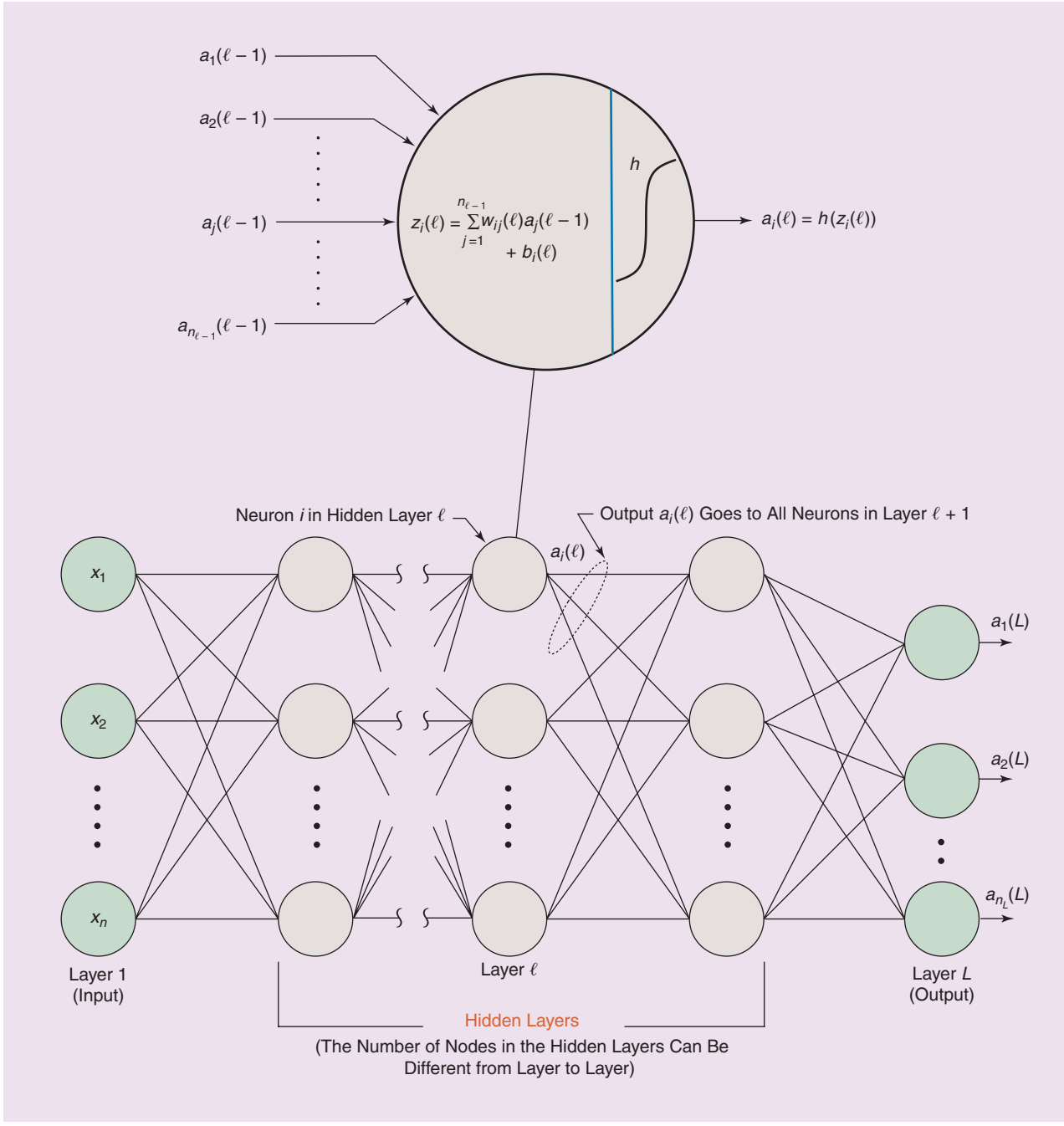


FIGURE 2. A schematic of a fully connected neural network. The zoomed section shows the computations performed by each neuron in the network. The activation function, h, shown is in the shape of a sigmoid.

biases. Because we know that $a_j(1) = x_j$, we can use (4) and (5) to compute $z_j(\ell)$ and $a_j(\ell)$ for all layers in the network, past the first. Although it is not shown in the diagram, we also compute $h'(z_i(\ell))$ for use later in backpropagation. Propagating a pattern vector through a neural net to its output is called *feedforward*, and training consists of feedforward and backpropagation passes through the network, periodically adjusting the weights and biases between such passes.

Measuring performance during training requires an error, or cost, function. The function used most frequently for this purpose is the mean-squared error (MSE) between actual and desired outputs:

$$E = \frac{1}{2}\sum_{j=1}^{n_L}(r_j - a_j(L))^2, \qquad (6)$$

where $a_j(L)$ is the activation value of the jth neuron in the output layer of the FCN. During training, we let $r_j = 1$ if the pattern being processed belongs to the jth class and $r_j = 0$ if it does not. Thus, if a pattern belongs to the kth class, we want the response of the kth output neuron, $a_k(L)$, to be 1 and the response of all other output neurons to be 0. When this occurs, the error is zero and no adjustments

are made to the weights because the input vector was classified correctly.

The objective of training is to adjust all the weights and biases in the network when a classification mistake is made, so that the error at the output is minimized. This is done using gradient descent for the weights and biases

$$w_{ij}(\ell) = w_{ij}(\ell) - \alpha \frac{\partial E}{\partial w_{ij}(\ell)} \tag{7}$$

and

$$b_i(\ell) = b_i(\ell) - \alpha \frac{\partial E}{\partial b_i(\ell)}, \tag{8}$$

where α is a scalar correction increment called the *learning rate constant*. Unfortunately, the change in the output error with respect to changes in the weights and biases in the hidden layers is not known. In a nutshell, backpropagation is a scheme that 1) propagates the error in the output, which is known, backward through all the hidden layers of the network and 2) uses the backpropagated error to express the two partials in (7) and (8) in terms of the activation function, the output error, and the current values of the weights and biases, all of which are known quantities at every layer in the network during training. A derivation of this important result is outside the scope of our discussion, but a sketch of the fundamental equations of backpropagation will help demonstrate the surprising simplicity of this method. The original derivation is given in [3], and is further illustrated and formulated in a more computationally effective matrix form, in [7].

Backpropagation is based on the following four results:

$$\frac{\partial E}{\partial w_{ij}(\ell)} = a_j(\ell - 1)\Delta_i(\ell) \tag{9}$$

and

$$\frac{\partial E}{\partial b_i(\ell)} = \Delta_i(\ell), \tag{10}$$

where

$$\Delta_j(\ell) = h'(z_j(\ell)) \sum_i w_{ij}(\ell + 1)\Delta_i(\ell + 1) \tag{11}$$

and

$$\Delta_j(L) = h'(z_j(L))[a_j(L) - r_j]. \tag{12}$$

Equations (9) and (10) are used to compute the gradients in (7) and (8), based on known or computable quantities. The fact that the quantities in (9) and (10) are known is established by (11) and (12). In the latter equation, $h'(z_j(L))$ and $a_j(L)$ are computed during feedforward, and r_j is given during training, so $\Delta_j(L)$ can be computed. But if we know this quantity, we can compute $\Delta_j(L-1)$ using (11) because all of its terms are known also during any training iteration. Another application of this equation gives $\Delta_j(L-2)$, and so on for all values of $\ell = L-1,\ L-2, \ldots, 2$. In other words, at any iterative step in training, we are able to compute all the quantities necessary to implement the gradient descent formulation given in (7) and (8), which seeks a minimum of the MSE in (6). Observe that we compute the terms necessary for gradient descent by proceeding backward from the output, hence the use of the term *backpropagation* to describe this method.

Training a CNN for image classification is performed in conjunction with training its attached FCN.

Using the preceding relatively simple equations, the procedure for training an FCN can be summarized as follows:

1) Initialize all weights and biases to small random values.
2) Using a pattern vector from the training set, perform a forward pass through the network and compute all values of $a_j(\ell)$ and $h'(z_j(\ell))$.
3) Compute the MSE using (6).
4) Compute $\Delta_j(L)$ using (12) and propagate it back through the network, using (11) to compute $\Delta_j(\ell)$ for $\ell = L-1, L-2, \ldots, 2$.
5) Update the weights and biases using (7)–(10).
6) Repeat steps 2–5 for all patterns of the training set. One pass through all training patterns constitutes one epoch of training. This procedure is repeated for a specified number of epochs, or until the MSE stabilizes to within a predefined range of acceptable variation.

Training a CNN for image classification is performed in conjunction with training its attached FCN. During feedforward, an image propagates through the CNN, resulting in a set of output maps in the last stage, as explained in Figure 1. The elements of these maps are vectorized and input into the FCN so that they propagate to the output of the fully connected net, at which point the MSE is computed, as described previously. The error delta, $\Delta_j(L)$, is backpropagated all the way to the input of the FCN. The vectorization applied on feedforward is then reversed into the 2-D format of the output maps. The reformatted quantities are the "deltas" of the CNN, which are then backpropagated to its input stage. The error deltas at each layer are computed during backpropagation through both networks, and these are then used to update the weights and biases of the CNN and FCN, using (7) and (8) for the latter, and their equivalents for the CNN [7]. Given the similarities between the computations performed by a CNN [(2) and (3)], and those performed by an FCN [(4) and (5)], the reader should not be surprised that the equations of backpropagation for the two networks are also similar. The fundamental difference between the equations for the two neural networks is that FCNs, which work with vectors, use multiplications, while CNNs, which work with 2-D arrays, use convolution.

As noted previously, the feedforward/backpropagation training procedure just explained is repeated for a specified number of epochs or until changes in the MSE stabilize to within a specified range of acceptable variation. After training, the CNN and FCN are completely specified by the learned weights and biases. When deployed for autonomous operation, the system classifies an unknown image into one of the classes on which the system was trained, by performing a feedforward pass and detecting which neuron at the output of the FCN yields the largest value.

A computational example

In this section, we illustrate how to train and test a CNN/FCN for image classification, using an image database that contains a training set of 60,000 grayscale images of handwritten numerals. The database also contains a set of 10,000 test images. Figure 3 shows the CNN and

FCN architectures we used. The layout is more detailed than in Figure 1 to simplify explanations. This network, which we explain below, was trained for 200 epochs using all 60,000 training images. The performance of the resulting trained system on the images of the training set was 99.4% correct classification. When subjected to the 10,000 test images, which the system had never "seen" before, the performance was 99.1%. These are impressive results, considering the simplicity of the architecture in Figure 3, and the fact that the inputs are handwritten characters that exhibit significant variability.

The input grayscale images are of size 28×28 pixels. The first stage of the CNN has six feature maps, and the second has 12. Both stages use pooling with 2×2 neighborhoods. The convolution kernels are of size 5×5 in both stages. The FCN has no hidden layers, consisting instead of only an input and an output layer. This means that the FCN is a linear classifier that implements hyperplane boundaries, as we noted previously in the discussion of perceptrons.

Because the inputs are grayscale images, the depth of the input volume to the first stage of the CNN is one, indicating that six 2-D kernels, one for each of the six feature maps, are needed in the first stage. The depth of the input volume to the second stage is six because there are six pooled maps at the output of the first stage. This means that 12 kernel volumes, each consisting of six 2-D kernels, are required to generate the 12 feature maps in the second stage, for a total of $6 \times 12 = 72$, 2-D convolution kernels in that stage. There is one bias per feature map, for a total of six biases in the first stage and 12 in the second.

For 2-D convolution without padding, we require that the 2-D kernels be completely contained in their respective maps during spatial translation. Because the input images are of size 28×28 pixels and the kernels are of size 5×5, this means that the feature maps in the first stage are of size 24×24 elements. Pooling reduces the size of these maps to 12×12 elements. These are the input maps to the second stage which, when convolved with kernels of size 5×5, result in feature maps of size 8×8. The

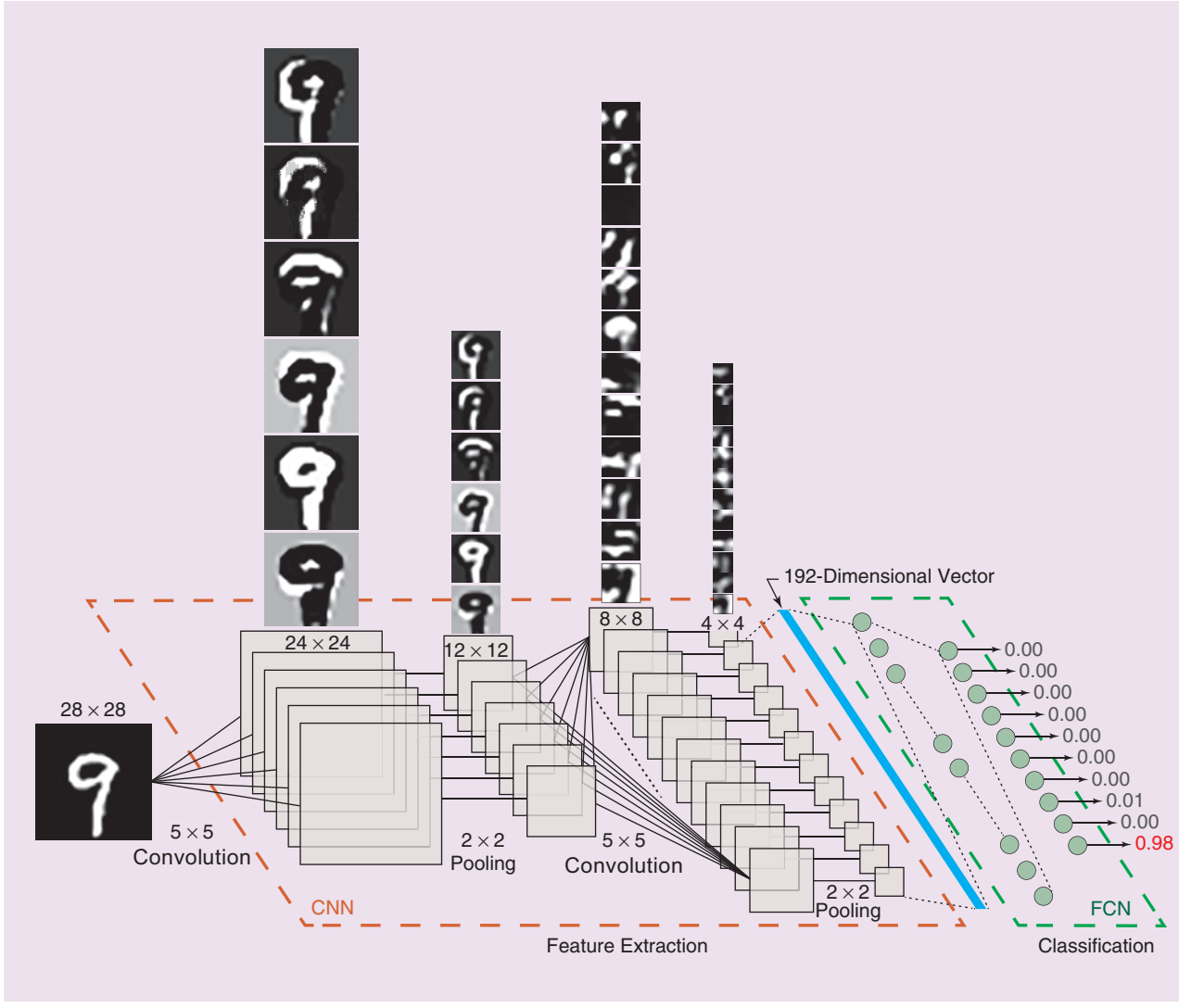


FIGURE 3. A CNN trained to extract features that are then used by an FCN to classify handwritten numerals. The input image shown is from the National Institute of Standards and Technology database. (A formatted version of this database is available for experimental work at yann.lecun.com/exdb/mnist.)

output maps in the second stage are obtained by pooling the feature maps in that stage, which results in 12 maps of size 4×4. These maps are then converted to vectors by linear indexing, which concatenates the elements of all the 2-D maps, column by column, into a one-dimensional string. When vectorized, these maps result in input vectors to the FCN that have $4 \times 4 \times 12 = 192$ elements. There are ten numeric classes, so the number of neurons in the output layer of the FCN is ten.

We illustrate the operations performed by our CNN/FCN neural net by following the flow of the image in Figure 3 from the input to the CNN to the output of the FCN. The weights and biases used in this example were obtained by training the CNN/FCN with the 60,000 images mentioned previously. Each feature map in the first stage of the CNN was generated by convolving a different 5×5 kernel with the input image. The resulting feature maps are shown as images above the feature maps volume in the first CNN stage. The feature maps in the first stage are of size 24×24 pixels, which we enlarged using bicubic interpolation to a size of 300×300 pixels, to make it easier to interpret them visually. These maps illustrate that each kernel was capable of detecting different features in the input image. For example, the first feature map at the top of the figure exhibits strong vertical components on the left of the character. The second feature map shows strong components in the northwest area of the top of the character and the left vertical lower area. The third feature map shows strong horizontal components in the top of the character. Similarly, each of the other three feature maps exhibits features distinct from the others.

As Figure 3 shows, the pooled maps are lower-resolution versions of the feature maps, but the former retain the basic characteristics of the latter. The volume containing these six maps is the input to the second stage. Each feature map in the second stage was generated by convolving a different kernel volume with the input volume to that stage, as explained in Figure 1. The feature maps resulting from these operations are of size 8×8; they are shown as enlarged images above the second CNN stage in Figure 3. These are not as easy to interpret visually as the feature maps in the first stage, other than to say that each exhibits a different response. Based on the accuracy of the training and test results, we know that these responses do a good job of characterizing all ten numeral classes over the entire database.

Each 192-dimensional vector resulting from vectorizing the output maps of the second stage of the CNN was fed into a fully connected net. This vector then propagated through the FCN, as explained previously. The values of the output neurons corresponding to the input image are zero or nearly zero, with the exception of the tenth neuron, whose output was 0.98. This indicates that the system correctly recognized the input image as being from the tenth class, which is the class of nines. These values of the output neurons resulted in a value for the MSE in (6) that is close to zero.

As mentioned previously in this example, training was carried out for 200 epochs. We trained the system using minibatches of 50 images between weight updates. The patterns were ordered randomly after each epoch of training, and the learning rate increment we used was $\alpha = 1.0$. This "standard" approach to training yielded excellent results in our example, but it can be refined further in more complex situations. For instance, experimental evidence suggests that large databases of RGB images containing 1,000 or more object classes require significantly deeper architectures and more complex training methodology. A good example is the deep learning neural network, *AlexNet*, that won the 2012 *ImageNet Challenge* [5].

What we have learned

After giving a brief historical account of how adaptive learning systems evolved, we introduced the basic concepts underlying the architecture and operation of deep CNNs and FCNs. The usefulness of these networks, working together to address complex image processing applications, is made possible by training the complete CNN/FCN system using backpropagation. We presented the underpinnings of backpropagation and discussed the basic equations used to implement this deep-learning scheme. The effectiveness of combining CNNs and FCNs for image pattern recognition was illustrated by training and testing a system capable of recognizing with high accuracy a large database of handwritten numeric characters.

Author

Rafael C. Gonzalez (rcg@utk.edu) received a B.S.E.E. degree (1965) from the University of Miami, FL, and M.S. (1967) and Ph.D. (1970) degrees from the University of Florida, Gainesville, all in electrical engineering. He is a distinguished service professor, emeritus in the Electrical Engineering and Computer Science Department at the University of Tennessee, Knoxville. He is a pioneer in the fields of image processing and pattern recognition and is the author or coauthor of four books, several edited books, and more than 100 publications in these fields. His books are used in more than 1,000 universities and research institutions throughout the world, and his work spans highly successful academic and industrial careers. He is a Life Fellow of the IEEE.

References

[1] F. Rosenblatt, "Two theorems of statistical separability in the perceptron," in *Proc. Symp. No. 10 Mechanisation Thought Processes,* London, 1959, vol. 1, pp. 421–456.

[2] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, D.C.: Spartan, 1962.

[3] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. 1, D. E. Rumelhart et al., Eds. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[4] Y. LeCun, B. Boser, J. S. Denker D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[6] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May, 2015.

[7] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York: Pearson-Prentice Hall, 2018.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

Zafar Rafii

Sliding Discrete Fourier Transform with Kernel Windowing

The sliding discrete Fourier transform (SDFT) is an efficient method for computing the N-point DFT of a given signal starting at a given sample from the N-point DFT of the same signal starting at the previous sample [1]. However, the SDFT does not allow the use of a window function, generally incorporated in the computation of the DFT to reduce spectral leakage, as it would break its sliding property. This article will show how windowing can be included in the SDFT by using a kernel derived from the window function, while keeping the process computationally efficient. In addition, this approach allows for turning other transforms, such as the modified discrete cosine transform (MDCT), into efficient sliding versions of themselves.

Relevance

The SDFT can be used to perform spectral analysis on successive samples in a signal without having to compute a new DFT from scratch every time, provided that windowing can be incorporated into the computation of the DFTs without harming the efficiency of the method. A notable application of the SDFT with windowing can then be framing detection in audio signals that have undergone lossy compression in the context of audio compression identification [2]. A lossy compression algorithm will typically introduce traces of compression in the signal being encoded, which can become visible in the time-frequency representation when using the same parameters and framing that were used for the encoding. Therefore, the parameters and framing can be recovered by computing time-frequency representations at successive samples in the signal and identifying when traces of compression become visible. This demanding process can be translated into an efficient one by using the SDFT with kernel windowing.

Prerequisites

Basic knowledge of digital signal processing is required to understand this article, particularly concepts such as the DFT, windowing, and general spectral analysis. More details about the SDFT and lossy audio compression identification can also be found in [1] and [2], respectively.

Problem statement and solution

Problem statement

The SDFT allows for the computation of the N-point DFT of a signal from the N-point DFT of the same signal starting one sample earlier, in a sense by sliding a rectangular window of length N one sample forward. The SDFT essentially relies on the shift theorem, which states that multiplying a signal by a linear phase is equivalent to a circular shift in the corresponding DFT.

Equation (1), shown at the bottom of the page, shows the derivation of $X^{(i)}$, the N-point DFT of signal x starting at sample i, from $X^{(i-1)}$, the N-point DFT of x starting at $i-1$, a process hence known as SDFT.

The SDFT thus only requires two N additions and N multiplications, leading to a linear time complexity of $O(N)$, while the full and direct computation of the DFT and the fast Fourier transform (FFT) are $O(N^2)$ and $O(N \log N)$, respectively.

Transforms such as the DFT typically use a window function in their computation to reduce spectral leakage and enhance spectral analysis. However, the SDFT does not allow the incorporation of a window function as it will break the process shown in (1). One solution would be to perform the windowing in the frequency domain, i.e., on the derived DFT through convolution. A practical window function for that matter could be the Hanning window, as the corresponding windowing in the frequency domain equals a simple three-point convolution [1]. Other window functions, however, may not be as practical, as the corresponding convolutions may involve many more operations, which will ultimately hurt the computational efficiency of the SDFT. Therefore, the problem is to incorporate any window function into the computation of the DFTs in an efficient manner without breaking the SDFT process.

Solution: Kernel windowing

The idea of performing the windowing in the frequency domain can still be exploited by reformulating the convolution

$$
\begin{aligned}
X^{(i)}_{k}_{0 \le k < N} &= \sum_{n=0}^{N-1} x_{i+n} e^{\frac{-j2\pi nk}{N}} \\
&= \sum_{n=0}^{N-1} x_{i+n} e^{\frac{-j2\pi (n+1)k}{N}} e^{\frac{j2\pi k}{N}} \\
&= \sum_{n=1}^{N} x_{i-1+n} e^{\frac{-j2\pi nk}{N}} e^{\frac{j2\pi k}{N}} \\
&= \left(\sum_{n=0}^{N-1} x_{i-1+n} e^{\frac{-j2\pi nk}{N}} - x_{i-1} + x_{i+N-1} \right) e^{\frac{j2\pi k}{N}} \\
&= (X^{(i-1)}_{k} - x_{i-1} + x_{i+N-1}) e^{\frac{j2\pi k}{N}}. \qquad (1)
\end{aligned}
$$

Digital Object Identifier 10.1109/MSP.2018.2855727
Date of publication: 13 November 2018

 1053-5888/18©2018IEEE

as a multiplication by a kernel that can be derived from any window function. Such a kernel will be independent from the signal to be processed and only need to be computed once. It will typically have a very small number of values that would be significant, which means that most of the values can then be ignored. This would lead to a very sparse kernel, which can then be applied to the DFT of the signal, producing results virtually equivalent to the DFT of the same signal modified by the corresponding window function while preserving the computational efficiency of the SDFT.

The constant-Q transform (CQT) is a transform with a logarithmic frequency resolution that was proposed as a more adapted alternative to the FT for analyzing music signals [3]. A fast algorithm was proposed soon after, which translated the slow computation of the CQT into the multiplication of a DFT, which can be efficiently computed using the FFT, and a kernel, which is computed once beforehand and typically very sparse [4]. The idea was to use Parseval's theorem to turn the direct computation in the time domain into a multiplication between a DFT and a kernel in the frequency domain, essentially demonstrating the property of energy conversation between the time and the frequency domains [5].

Parseval's theorem is recalled in (2). The value X is the N-point DFT of x and $\bar{x}$ represents the complex conjugate of x

$$\sum_{n=0}^{N-1} x_n \overline{y_n} = \frac{1}{N} \sum_{k=0}^{N-1} X_k \overline{Y_k}. \tag{2}$$

Following a similar idea, we propose the use of Parseval's theorem to translate the DFT of a windowed signal into the DFT of the signal, multiplied by a kernel that is derived from the corresponding window function: a multiplication that will happen after the SDFT process. Unlike in the fast CQT case, the purpose here is not to speed up the computation of the transform by taking advantage of the efficiency of the FFT algorithm in conjunction with the use of sparse kernel but to extract the windowing operation from the DFT computation so that the SDFT process shown in (1) still holds.

Equation (3), shown at the bottom of the page, illustrates the computation of $\mathcal{X}^{(i)}$, the N-point DFT of the signal x starting at sample i and modified by the window function w, from $X^{(i-1)}$, the N-point DFT of x starting at $i-1$ without windowing, multiplied by the kernel K, which is derived from w.

As we can see in (3), the kernel is completely independent from the signal; therefore, it only needs to be computed once, before the SDFT process. Furthermore, given the nature of such kernel, typically only a very small number of its values will be significant, which means that most of the values can then be zeroed, given some threshold, leading to a very sparse kernel. The multiplication of the derived DFT by such kernel will thus only involve few more operations, keeping the whole process computationally efficient.

A notable application of the SDFT with windowing can then be framing detection in audio signals that have undergone lossy compression in the context of audio compression identification.

Figure 1 shows the kernels derived from some common window functions, i.e., Hanning, Blackman, triangular, Gaussian, Parzen, and Kaiser windows. As we can see, the Hanning window kernel shows only three nonzero values per row, confirming that the corresponding windowing in the frequency domain equals a simple three-point convolution, while the Blackman window kernel shows five nonzero values per row. Both those windows are actually special cases of the generalized cosine window whose corresponding windowing in the frequency domain equals convolutions with typically only few points. Unlike the Hanning and Blackman window kernels, the triangular, Parzen, Gaussian, and Kaiser window kernels show additional nonzero values around their main diagonal, suggesting that the corresponding windowings in the frequency domain equal convolutions with many more points. However, most of those nonzero values have very small magnitudes ($\ll 0.01$) and could then be ignored without significantly affecting the actual windowing process. By using an appropriate threshold, those kernels can therefore be made very sparse with only a few meaningful values per row in the same manner as in the fast CQT case [4].

As proposed in [4], we computed for each of those kernels the error in keeping the values greater than a chosen threshold by dividing the sum of the magnitudes of the values after thresholding by the sum of the magnitudes of all the values before thresholding. A threshold of 0.01 will thus give very small errors of 0.049, 0.009, 0.020, and 0.015, for the triangular, Parzen, Gaussian, and Kaiser window kernels, respectively, when derived for an N-point DFT with $N = 2{,}048$. With such a threshold, the first three kernels will then only have

$$\begin{aligned}
\underset{0\le k<N}{\mathcal{X}_k^{(i)}} &= \sum_{n=0}^{N-1} x_{i+n} \underbrace{w_n e^{\frac{-j2\pi nk}{N}}}_{y_n} \\
&= \sum_{k'=0}^{N-1} X_{k'}^{(i)} \underbrace{K_{k,k'}}_{\frac{1}{N}\overline{Y_{k'}}} \\
&= \sum_{k'=0}^{N-1} \left[(X_{k'}^{(i-1)} - x_{i-1} + x_{i+N-1}) e^{\frac{j2\pi k'}{N}} \right] K_{k,k'} \\
\underset{\substack{0\le k<N \\ 0\le k'<N}}{K_{k,k'}} &= \frac{1}{N}\overline{Y_{k'}} = \frac{1}{N}\overline{\sum_{n=0}^{N-1} y_n e^{\frac{-j2\pi nk'}{N}}} \\
&= \frac{1}{N} \sum_{n=0}^{N-1} \overline{w_n e^{\frac{-j2\pi nk}{N}}} e^{\frac{-j2\pi nk'}{N}} \\
&= \frac{1}{N} \sum_{n=0}^{N-1} w_n e^{\frac{j2\pi n(k'-k)}{N}}.
\end{aligned} \tag{3}$$

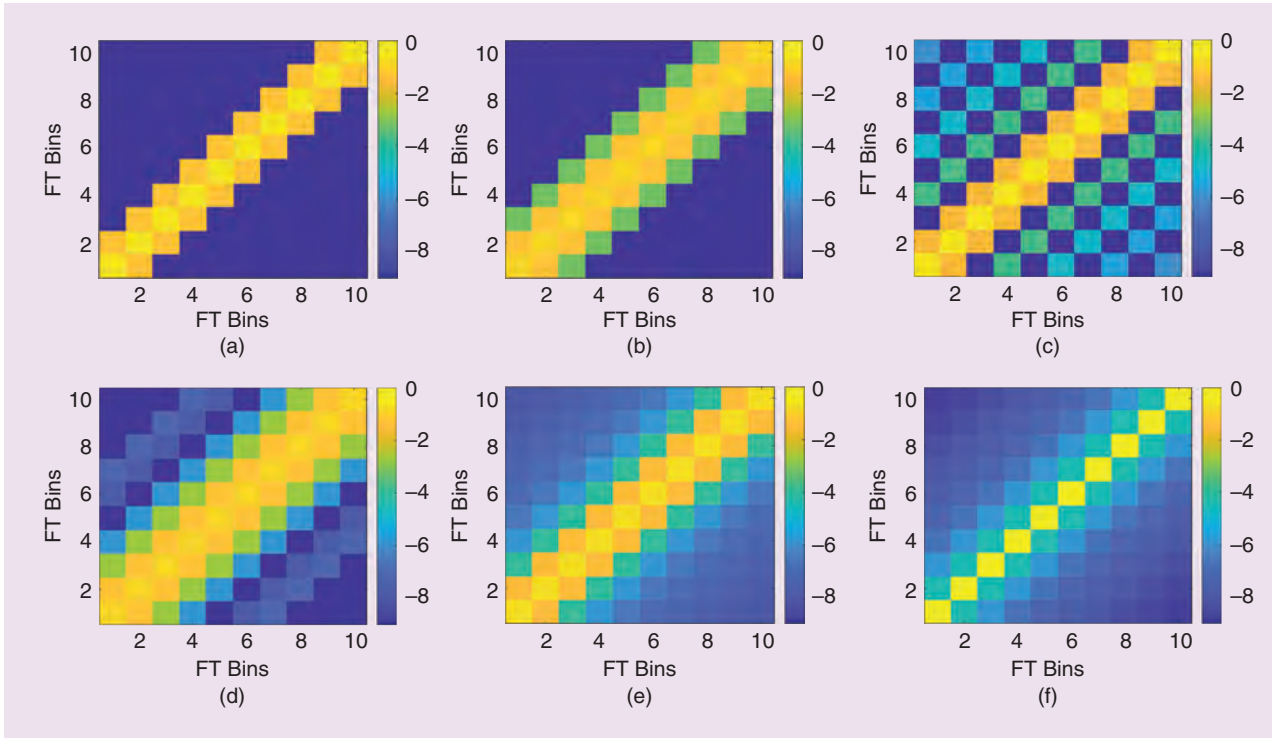


FIGURE 1. The kernels derived from the (a) Hanning, (b) Blackman, (c) triangular, (d) Parzen, (e) Gaussian (with $\alpha = 2.5$), and (f) Kaiser (with $\beta = 0.5$) windows. The kernels were derived for an N-point DFT where $N = 2{,}048$ samples. Only the first 100 coefficients at the bottom-left corner of the N-by-N kernels are shown. The values are displayed in log of amplitude.

five nonzero values per row, while the latter one will have three nonzero values per row. This shows that only a very small number of values is actually significant in such kernels. The multiplication of the DFT by those very sparse kernels will then only involve KN multiplications and KN additions, with $K = 3$ or 5, barely affecting the computational efficiency of the SDFT, still maintaining a linear complexity of $O(N)$, and producing results virtually equivalent to taking the DFT of the signal modified by the corresponding window functions.

Computational examples

Framing detection and lossy audio coding

The SDFT with kernel windowing can be particularly useful for fast framing detection in the context of audio compression identification. Audio compression identification is the recovery of information regarding the data compression that an audio signal has undergone. In particular, the recovery of the parameters and framing used at the time-frequency decomposition stage of the encoding could allow for identifying the coding format or detecting alterations in audio signals that have undergone lossy compression [2], [6]–[9]. Lossy compression algorithms typically introduce traces of compression in the audio signal being encoded in the form of time-frequency coefficients quantized to zeros, which can become visible when using the same parameters and framing that were used for the encoding. One approach to identify if and when lossy compression was used would then be to compute the time-frequency representation at successive samples in the audio signal and search for traces of compression every time, given a set of parameters associated with a known coding format, such as time-frequency transform, window length, and window function, a process also known as *framing detection*.

Lossy audio coding formats, perhaps the most popular ones being MP3, Advanced Audio Coding (AAC), AC-3, Vorbis, and Windows Media Audio (WMA), are widely used for storage (e.g., in music and video files) or transmission (e.g., in radio and television broadcasting). Compression algorithms that can encode to such formats first transform the audio signal into a time-frequency representation, derive a psychoacoustic model to locate regions of perceptually less significance, then quantize the data given the psychoacoustic model, and, finally, convert it into a bitstream. The transform used at the time-frequency decomposition stage is typically based on the MDCT, and a variety of window lengths and window functions can be used depending on the coding format. In particular, specialized window functions such as the sine, slope, and Kaiser–Bessel-derived (KBD) windows, are generally required for the MDCT to be invertible. For more information about lossy audio coding, see [10]. The computation of the MDCT without windowing is

$$Y_{k\ 0\le k<\frac{N}{2}} = \sum_{n=0}^{N-1} x_n \cos\Big[\frac{2\pi}{N}\Big(n+\frac{1}{2}+\frac{N}{4}\Big)\Big(k+\frac{1}{2}\Big)\Big]. \tag{4}$$

Sliding MDCT with kernel windowing

In this context, performing framing detection for audio compression identification would involve computing an MDCT for every set of window length and window

function associated with a known lossy coding format at successive samples in the audio signal and searching for time-frequency coefficients quantized to zero until one of the sets shows visible traces of compression for a specific framing of the signal. We can see that such a process will be computationally demanding, as a full transform would have to be computed every time. The direct computation of the MDCT, including the windowing using one of the specialized window functions presented earlier, can actually be translated into an SDFT with a kernel windowing by incorporating the computation of the window function and a part of the MDCT into a kernel, which will still happen to be very sparse, thus making the process computationally efficient.

Equation (5), shown at the bottom of the page, shows the computation of $\mathcal{Y}^{(i)}$, the N-point MDCT of signal x starting at sample i and modified by the window function w, from $X^{(i-1)}$, the N-point DFT of x starting at $i-1$ without windowing, multiplied by the kernel K, which is derived from w.

Figure 2 shows the kernels derived for an N-point MDCT, from the sine window where $N = 1,152$ samples, as in MP3, from the slope window where $N = 2,048$ samples, as in Vorbis, and from the KBD window where $N = 512$ samples, as in AC-3. As we can see, most of the values in those kernels appear to have negligible magnitudes, while the very few values with significant magnitudes appear to be concentrated around two diagonals, one going from the bottom-left to the top-center and one going from the top-center to the bottom-right. As in [4], we computed for each of those kernels the error in keeping the values greater than 0.01 and obtained very small errors of 0.000, 0.022, and 0.013, for the sine, slope, and KBD window kernel, respectively. With such a threshold, the sine kernel will only have around two nonzero values per row and the slope and KBD kernels around six nonzero values per row. Therefore, these very sparse kernels will barely affect the computational efficiency of the SDFT while still producing results equivalent to taking the MDCT of the signal modified by the corresponding window functions.

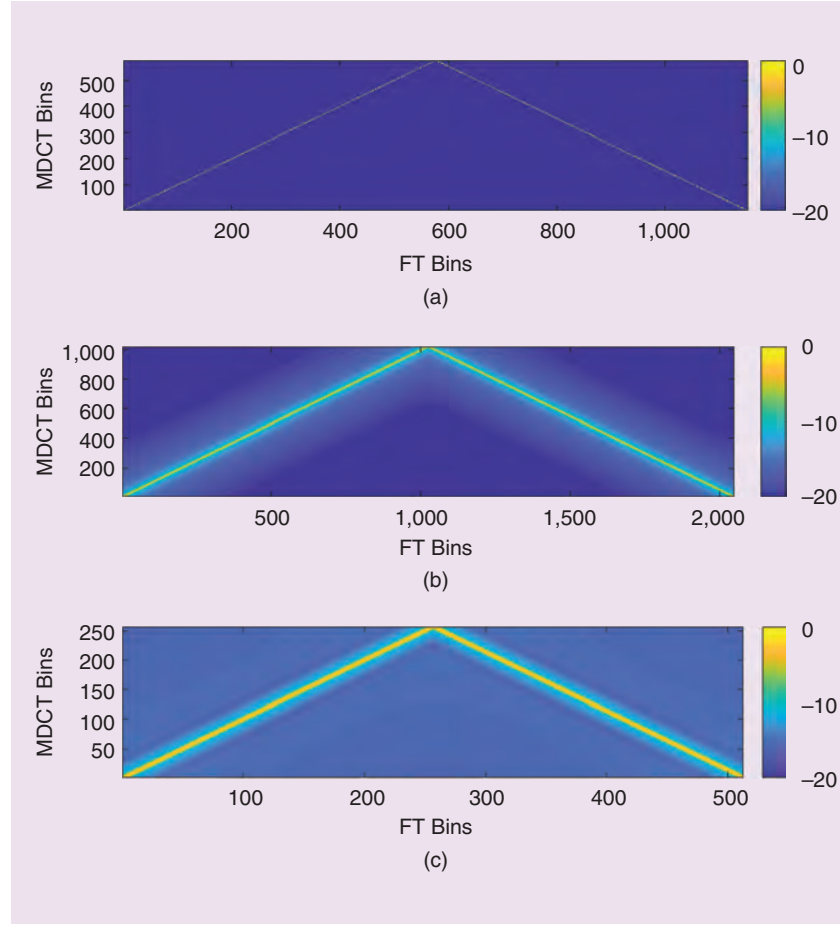


FIGURE 2. The kernels derived for an N-point MDCT, from (a) the sine window where $N = 1,152$ samples, (b) the slope window where $N = 2,048$ samples, and (c) the KBD window where (c) $N = 512$ samples. The values are displayed in log of amplitude.

$$
\begin{aligned}
\underset{0\le k<\frac{N}{2}}{\mathcal{Y}_k^{(i)}} &= \sum_{n=0}^{N-1} x_{i+n} \underbrace{w_n \cos\Big[\frac{2\pi}{N}\Big(n+\frac{1}{2}+\frac{N}{4}\Big)\Big(k+\frac{1}{2}\Big)\Big]}_{y_n} \\
&= \sum_{k'=0}^{N-1} X_{k'}^{(i)} \underbrace{K_{k,k'}}_{\frac{1}{N}\overline{Y_{k'}}} \\
&= \sum_{k'=0}^{N-1} \Big[(X_{k'}^{(i-1)} - x_{i-1} + x_{i+N-1}) e^{\frac{j2\pi k'}{N}}\Big] K_{k,k'} \\
\underset{\substack{0\le k<\frac{N}{2}\\ 0\le k'<N}}{K_{k,k'}} &= \frac{1}{N}\overline{Y_{k'}} = \frac{1}{N}\overline{\sum_{n=0}^{N-1} y_n e^{\frac{-j2\pi nk'}{N}}} \\
&= \frac{1}{N}\sum_{n=0}^{N-1} \overline{w_n \cos\Big[\frac{2\pi}{N}\Big(n+\frac{1}{2}+\frac{N}{4}\Big)\Big(k+\frac{1}{2}\Big)\Big]} e^{\frac{-j2\pi nk'}{N}} \\
&= \frac{1}{N}\sum_{n=0}^{N-1} w_n \cos\Big[\frac{2\pi}{N}\Big(n+\frac{1}{2}+\frac{N}{4}\Big)\Big(k+\frac{1}{2}\Big)\Big] e^{\frac{j2\pi nk'}{N}}. \qquad (5)
\end{aligned}
$$

What we have learned

We have shown that the SDFT can incorporate windowing in its computation by using a kernel that can be derived from any window function and can be made very sparse. This SDFT with kernel windowing will produce results equivalent to the DFT of the signal modified by the

corresponding window function, while keeping the process computationally efficient. This approach may be applied in audio compression identification, in particular by making the process of framing detection much more efficient, allowing for the translation of a transform, such as the MDCT, into an efficient sliding version of itself.

Author information

Zafar Rafii (zafar.rafii@nielsen.com) received his M.S. degree in electrical engineering from the Ecole Nationale Superieure de l'Electronique et de ses Applications, Cergy, France, and another M.S. degree in the same field from the Illinois Institute of Technology, Chicago, in 2006. He received his Ph.D. degree in electrical engineering and computer science from Northwestern University, Evanston, Illinois, in 2014. He is currently a senior research engineer at Gracenote. He also previously worked as a research engineer at Audionamix in France. His research interests focus on audio analysis, somewhere between signal processing, machine learning, and cognitive science, with a predilection for source separation and audio identification in music.

References

[1] E. Jacobsen and R. Lyons, "The sliding DFT," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 74–80, Mar. 2003.

[2] B. Kim and Z. Rafii, "Lossy audio compression identification," in *Proc. 26th European Signal Processing Conf.*, Rome, Italy, Sept. 2018.

[3] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[4] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, Nov. 1992.

[5] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1975.

[6] J. Herre and M. Schug, "Analysis of decompressed audio—The 'inverse decoder,'" in *Proc. 109th Audio Engineering Society Conv.*, Los Angeles, CA, Sept. 2000, p. 5256.

[7] S. Moehrs, J. Herre, and R. Geiger, "Analysing decompressed audio with the 'inverse decoder'—Towards an operative algorithm," in *Proc. 112th Audio Engineering Society Conv.*, Munich, Germany, Apr. 2002, p. 5576.

[8] R. Yang, Z. Qu, and J. Huang, "Detecting digital audio forgeries by checking frame offsets," in *Proc. 10th ACM Workshop on Multimedia and Security*, Oxford, U.K., Sept. 2008, pp. 21–26.

[9] D. Gärtner, C. Dittmar, P. Aichroth, L. Cuccovillo, S. Mann, and G. Schuller, "Efficient cross-codec framing grid analysis for audio tampering detection," in *Proc. 136th Audio Engineering Society Conv.*, Berlin, Apr. 2014, pp. 306–316.

[10] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. New York: Springer, 2003.

SP

Digital Object Identifier 10.1109/MSP.2018.2876581

Alexander Bertrand

Utility Metrics for Assessment and Subset Selection of Input Variables for Linear Estimation

This tutorial article introduces the utility metric and its generalizations, which allow for a quick-and-dirty quantitative assessment of the relative importance of the different input variables in a linear estimation model. In particular, we show how these metrics can be cheaply calculated, thereby making them very attractive for model interpretation, online signal quality assessment, or greedy variable selection. The main goal of this article is to provide a transparent and consistent framework that consolidates, unifies, and extends the existing results in this area. In particular, we 1) introduce the basic utility metric and show how it can be calculated at virtually no cost, 2) generalize it toward group-utility and noise-impact metrics, and 3) further extend it to cope with linearly dependent inputs and minimum norm requirements.

Introduction of the utility metric

When solving a regression problem, one often wants to have some quantitative insights into the relevance of each input variable, i.e., how much it contributes to the reduction of a loss function. Such information can be used to interpret the model, assess the predictive value of specific input variables or signals, or perform a greedy variable subset selection [1]–[4]. The latter allows reducing the dimensionality of a model, e.g., to avoid overfitting [5], to make the model more interpretable or to reduce computational complexity, data storage, data transmission, or sensor costs [3], [6], [7].

In this tutorial, we focus on linear least squares (LS) estimation, although many results can also be extended to other linear estimation frameworks [8]. A naive heuristic that is remarkably commonly used for variable assessment is the magnitude of the weights of the LS solution, thereby (incorrectly) assuming that important input variables will also receive a large weight in the LS solution. However, it is not difficult to see that this reasoning is flawed. For example, if the observations of one of the input variables would all be scaled with a factor α, then the corresponding weight in the LS solution will be scaled with α^{-1}, whereas the information content of that input variable obviously remains unchanged.

When solving a regression problem, one often wants to have some quantitative insights into the relevance of each input variable.

A more relevant metric would consist of calculating the effective loss, i.e., the increase in LS cost, if an input variable would be removed and if the model would be reoptimized. We refer to this resulting metric as the *utility* of that input variable. Utility is a powerful heuristic for input variable assessment [1], [3], [4], [6] and can even be shown to have some optimality properties when used for greedy variable subset selection [1], which can compete with well-known sparse regression techniques, such as the least absolute shrinkage and selection operator (LASSO) [9].

However, computing the utility of M input variables by definition requires solving M different LS problems, i.e., one for each removal of an input variable [1], [3]. As a result, the metric scales poorly with the dimensionality of the model, which can be problematic in real-time applications and can make a greedy variable subset selection in very high-dimensional problems even infeasible.

In this tutorial article, we show how some simple tricks from standard linear algebra allow computing the utility metric at virtually no cost, thereby making it a highly attractive metric for model interpretation, signal quality assessment, greedy variable selection, and so forth, in particular in real-time or large-scale applications. We also address several generalizations and extensions of this utility metric toward

- a group-utility metric, allowing evaluation of the joint utility of a group of input variables
- a noise-impact metric, allowing evaluation of the impact of additive errors in the input variables, e.g., to predict the effect of quantization or measurement noise, and which contains the original utility metric as a special case
- a minimum-norm utility metric for ill-conditioned cases, in which linear dependence relationships exist between the input variables.

Digital Object Identifier 10.1109/MSP.2018.2856632
Date of publication: 13 November 2018

The main goals of this article are to 1) provide an accessible overview and unification of existing results in this context with pointers to the original publications and 2) introduce novel extensions and generalizations presented in a unified framework.

Utility: Definition and core equation

Definition

Consider the LS problem with N measurements of M input variables:

$$J(Y) \triangleq \min_{\mathbf{x}} \frac{1}{N} \| Y\mathbf{x} - \mathbf{d} \|^2, \qquad (1)$$

where $Y \in \mathbb{R}^{N \times M}$ is the regressor matrix, $\mathbf{d} \in \mathbb{R}^N$ is the desired response vector, and $\mathbf{x} \in \mathbb{R}^M$ is the vector with optimization variables (we consider the real-valued case for simplicity, yet all results in this article can be easily generalized to the complex-valued case). Note that $J(Y)$ is defined as an operator that evaluates the LS cost for the case where the information in Y is available, which includes an implicit optimization of $\mathbf{x}$ [the reason for making the dependency on Y explicit will become clear later when defining the utility metric in (3)]. We will refer to the columns of Y as the input variables of the model. Depending on the context, these input variables could represent, e.g., different sensors or channels (in a sensor array), time lags (in a temporal filter), observations of independent variables (in a regression model), and so forth. Assuming Y has full rank, the LS solution $\hat{\mathbf{x}}$ that minimizes (1) is given by

$$\hat{\mathbf{x}} = R^{-1}\mathbf{r}, \qquad (2)$$

with $R = (1/N)Y^TY$ and $\mathbf{r} = (1/N)Y^T\mathbf{d}$.

To quantify the relevance of each input variable, we define the utility metric [3], which will be the focus of this tutorial. The utility of the kth input variable is defined as the increase in LS cost if the kth input variable would be removed and if the LS problem would be reoptimized, i.e.,

$$U_k \triangleq J(Y_{-k}) - J(Y), \qquad (3)$$

where Y_{-k} denotes the matrix Y with the kth column removed. Note that a naive computation of U_k would, in principle, require solving a second LS problem based on Y_{-k}, of which the computational complexity scales cubically with the number of input variables, i.e., $O(M^3)$. Calculating the utility of all M input variables would then have a complexity of $O(M^4)$, which can be unacceptably high for, e.g., real-time systems or for large-scale problems with hundreds or thousands of input variables. In the sequel, we will show how some simple linear algebra tricks allow deriving an efficient and elegant equation to calculate (3) for all input variables, with a total complexity of merely $O(M)$.

In some applications, the input variables are naturally clustered in specific predefined groups.

Core equation

Once the full LS solution (2) has been calculated, we show that calculating the utility (3) does not require solving an extra LS problem to evaluate $J(Y_{-k})$, i.e., it can be calculated as

$$U_k = \frac{|x_k|^2}{q_k}, \qquad (4)$$

where q_k is the kth diagonal element of R^{-1} and where x_k is the kth element of $\hat{\mathbf{x}}$ in (2). This has originally been proven in [3], but we will derive a more general form of (4) in the "Group Utility" section, which will then also prove (4) as a special case. From (4), it follows that the vector $\mathbf{u} = [U_1, \ldots, U_M]^T$ containing the utilities of all input variables can be calculated as

$$\mathbf{u} = \Lambda^{-1} |\hat{\mathbf{x}}|^2, \qquad (5)$$

where $|\cdot|^2$ represents an elementwise squaring and $\Lambda = \mathcal{D}(R^{-1})$ with $\mathcal{D}(\cdot)$ the operator that creates a diagonal matrix by setting all of the off-diagonal elements of the matrix in its argument to zero.

Note that (4) and (5) are remarkably simple and elegant. Because R^{-1} is readily available from the computation of $\hat{\mathbf{x}}$ in (2), the utility can be calculated with a complexity of merely $O(1)$ for a single variable, or $O(M)$ for all variables. This should be contrasted to a naive computation of U_k or $\mathbf{u}$ based on the original utility definition (3), resulting in a complexity of $O(M^3)$ and $O(M^4)$, respectively.

Group utility

In some applications, the input variables are naturally clustered in specific predefined groups, in which case it could make more sense to investigate the utility of groups of variables, rather than of individual variables. For example, in a multichannel filter, the utility of a channel is the joint utility of all of the filter taps in that channel's delay line. Comparably, in a sensor network with multisensor nodes, the utility of a node is the joint utility of all of the sensor signals within that node [4].

Similar to (3), the group utility of a predefined group of G input variables, denoted by the set $\mathcal{G}$, is defined as [4]

$$U_{\mathcal{G}} \triangleq J(Y_{-\mathcal{G}}) - J(Y), \qquad (6)$$

where $Y_{-\mathcal{G}}$ is the matrix Y with all columns corresponding to the input variables in $\mathcal{G}$ removed. From now on, we assume that $\mathcal{G}$ consists of the last G columns of Y, which is without loss of generality (w.l.o.g.), because the order of the inputs can be arbitrarily rearranged. Define the following block partitioning of the (known) inverse of R

$$R^{-1} = \begin{bmatrix} A & B \\ B^T & Q \end{bmatrix}, \qquad (7)$$

where Q is the $G \times G$ matrix capturing the rows and columns with indices corresponding to the variables in $\mathcal{G}$ (here at the bottom right w.l.o.g.). As shown in "Derivation of the Group-Utility Core Equation (8)," the group-utility $U_{\mathcal{G}}$ can efficiently be calculated as

$$U_{\mathcal{G}} = \mathbf{x}_{\mathcal{G}}^T Q^{-1} \mathbf{x}_{\mathcal{G}}, \qquad (8)$$

where $\mathbf{x}_{\mathcal{G}}$ contains the last G entries of $\hat{\mathbf{x}}$ (we do not add a hat to $\mathbf{x}_{\mathcal{G}}$ because it is not an optimal LS solution in itself).

Derivation of the Group-Utility Core Equation (8)

Note that evaluating $J(Y_{-\mathcal{G}})$ in (6) requires solving the reduced least squares (LS) solution

$$\hat{\mathbf{x}}_{-\mathcal{G}} = R_{-\mathcal{G}}^{-1}\mathbf{r}_{-\mathcal{G}} \tag{S1}$$

with $R_{-\mathcal{G}} = (1/N)Y_{-\mathcal{G}}^T Y_{-\mathcal{G}}$ and $\mathbf{r}_{-\mathcal{G}} = (1/N)Y_{-\mathcal{G}}^T\mathbf{d}$. The first step in our derivation is to find a more efficient way to calculate $R_{-\mathcal{G}}^{-1}$, based on a subresult of the blockwise matrix inversion theorem [10].

Lemma

Consider the block partitioning of a matrix V and its inverse V^{-1} as follows:

$$V = \begin{bmatrix} A & B \\ C & D \end{bmatrix},\quad V^{-1} = \begin{bmatrix} E & * \\ * & * \end{bmatrix}$$

with A and E square matrices of equal size. If D and E are invertible, then $E^{-1} = A - BD^{-1}C$.

By setting $V = R^{-1}$ (consequently $E = R_{-\mathcal{G}}$), and using the notation in (7), the lemma immediately yields the following important result:

$$R_{-\mathcal{G}}^{-1} = A - BQ^{-1}B^T. \tag{S2}$$

By plugging (S2) in (S1), the reduced LS solution is given by

$$\hat{\mathbf{x}}_{-\mathcal{G}} = (A - BQ^{-1}B^T)\mathbf{r}_{-\mathcal{G}}. \tag{S3}$$

Using the partitioning in (7), we can define the following partitioning of the LS solution (2):

$$\hat{\mathbf{x}} = \begin{bmatrix} \mathbf{x}_{-\mathcal{G}} \\ \mathbf{x}_{\mathcal{G}} \end{bmatrix} = \begin{bmatrix} A\,\mathbf{r}_{-\mathcal{G}} + B\,\mathbf{r}_{\mathcal{G}} \\ B^T\mathbf{r}_{-\mathcal{G}} + Q\,\mathbf{r}_{\mathcal{G}} \end{bmatrix}, \tag{S4}$$

where $\mathbf{r}_{-\mathcal{G}}$ and $\mathbf{r}_{\mathcal{G}}$ denote subvectors of $\mathbf{r}$ containing the first $M - G$ and last G entries, respectively. From (S3) and (S4), it can be easily verified that

$$\hat{\mathbf{x}}_{-\mathcal{G}} = \mathbf{x}_{-\mathcal{G}} - BQ^{-1}\mathbf{x}_{\mathcal{G}}, \tag{S5}$$

which allows efficiently updating the LS solution. By expanding the LS cost functions in (6) in their quadratic terms and plugging in the corresponding LS solutions (2) and (S1) for $\hat{\mathbf{x}}_{-\mathcal{G}}$ and $\hat{\mathbf{x}}$, respectively, it can be straightforwardly found that

$$U_{\mathcal{G}} = \mathbf{r}^T\hat{\mathbf{x}} - \mathbf{r}_{-\mathcal{G}}^T\hat{\mathbf{x}}_{-\mathcal{G}}. \tag{S6}$$

By plugging in (S5), and by partitioning $\hat{\mathbf{x}}$ in $\mathbf{x}_{-\mathcal{G}}$ and $\mathbf{x}_{\mathcal{G}}$, we immediately find that

$$U_{\mathcal{G}} = \mathbf{r}_{\mathcal{G}}^T\mathbf{x}_{\mathcal{G}} + (\mathbf{r}_{-\mathcal{G}}^T B)\,Q^{-1}\mathbf{x}_{\mathcal{G}}. \tag{S7}$$

From the bottom half of (S4), it follows that $\mathbf{r}_{-\mathcal{G}}^T B = \mathbf{x}_{\mathcal{G}}^T - \mathbf{r}_{\mathcal{G}}^T Q$, such that (S7) eventually yields (8). ■

Note that this group-utility equation reduces to the original utility equation (4) if $G = 1$. Obviously, if $G \ll M$, computing (8) is much cheaper than evaluating $J(Y_{-\mathcal{G}})$ in (6) by explicitly computing the reduced LS solution.

Although the derivation of (8) is not necessary to follow the rest of this tutorial, we include it in "Derivation of the Group-Utility Core Equation (8)" for completeness and because it also reveals two interesting by-products, i.e., two equations, (S2) and (S5), that allow recursively updating 1) the inverse autocorrelation matrix R^{-1} and 2) the LS solution $\hat{\mathbf{x}}$ after the removal of G input variables. This is interesting if the (group-)utility metric would be used for greedy variable selection, where (groups of) input variables are deleted one by one (see the "Computational Benefits and Implications for Variable Subset Selection" section).

Generalization toward noise impact

The utility metric as defined in (3) measures the increase in the LS cost when the kth column of Y is removed. Another relevant metric would be to measure the increase in the LS cost when adding some random noise in the kth column of Y, rather than fully removing that column. This is interesting in situations where one has some freedom in controlling the accuracy of each individual input variable. For example, in quantization or lossy compression, one can often modify the bit depth or the compression rate of each individual signal to reduce the resources required to store or transmit it, while increasing its noise level. Similar tradeoffs appear when deciding between cheap or accurate sensors, in applications or experiments where the accuracy depends on the measurement time, and so forth. In all of these cases, it is important to be able to efficiently assess and quantify how additive noise on each particular input variable would affect the estimation performance, in particular when used in greedy or adaptive resource allocation schemes.

It is important to be able to efficiently assess and quantify how additive noise on each particular input variable would affect the estimation performance.

To quantify the effect of additive noise, the noise-impact metric was originally defined in [11] with the purpose of performing a greedy signal quantization. For the sake of completeness and unification, we generalize this result to a group-impact metric in this article, which has the result

Derivation of the Noise-Impact Core Equation (10)

As the added noise is zero mean and uncorrelated to Y and $\mathbf{d}$, the least squares solution of the first term is equal to

$$\hat{\mathbf{x}}_{\mathcal{G},\Sigma} = R_F^{-1}\mathbf{r} = (R + FF^T)^{-1}\mathbf{r},$$

where $F = [O \;\; \Sigma^{1/2}]^T$ with O the all-zero matrix. Note that $\mathbf{r}$ is unaffected by the noise due to the expected value operator and the fact that the noise is uncorrelated to $\mathbf{d}$. Applying the Sherman–Morrison–Woodbury identity [10] to R_F yields

$$R_F^{-1} = R^{-1} - R^{-1}F(I + F^T R^{-1} F)^{-1} F^T R^{-1}.$$

With the partitioning of R^{-1} in (7), this becomes

$$R_F^{-1} = R^{-1} - \begin{bmatrix} B \\ Q \end{bmatrix} \Sigma^{\frac{1}{2}} (I + \Sigma^{\frac{1}{2}} Q \Sigma^{\frac{1}{2}})^{-1} \Sigma^{\frac{1}{2}} [B^T \;\; Q] \quad \text{(S8)}$$

$$= R^{-1} - \begin{bmatrix} B \\ Q \end{bmatrix} (\Sigma^{-1} + Q)^{-1} [B^T \;\; Q]. \quad \text{(S9)}$$

Similar to (S6), it can easily be verified that the noise impact (9) is equal to

$$I_{\mathcal{G}}(\Sigma) = \mathbf{r}^T\hat{\mathbf{x}} - \mathbf{r}^T\hat{\mathbf{x}}_{\mathcal{G},\Sigma} = \mathbf{r}^T(R^{-1} - R_F^{-1})\mathbf{r}.$$

Plugging (S8) into this equation, and using the fact that $\mathbf{x}_{\mathcal{G}} = [B^T \; Q]\mathbf{r}$ [see (S4)], we eventually find (10). ■

of [11] as a special case and generalizes the group-utility metric (8). Let $Y_{\mathcal{G},\Sigma}$ denote the matrix Y in which zero-mean random noise is added to the input variables in $\mathcal{G}$, with a positive definite noise covariance matrix $\Sigma \in \mathbb{R}^{G\times G}$. In most cases, the noise will be uncorrelated across the input variables, in which case Σ is a diagonal matrix.

In line with (6), we define the noise impact on the group $\mathcal{G}$ as

$$I_{\mathcal{G}}(\Sigma) \triangleq \min_{\mathbf{x}} \frac{1}{N}\mathcal{E}\left\{\left\| Y_{\mathcal{G},\Sigma}\,\mathbf{x} - \mathbf{d}\right\|^2\right\} - J(Y), \quad (9)$$

where $\mathcal{E}\{\cdot\}$ denotes the expected value operator, which is introduced due to the stochastic nature of the first term. In "Derivation of the Noise-Impact Core Equation (10)," we show that (9) can be efficiently calculated as

$$I_{\mathcal{G}}(\Sigma) = \mathbf{x}_{\mathcal{G}}^T(\Sigma^{-1} + Q)^{-1}\mathbf{x}_{\mathcal{G}}, \quad (10)$$

where we again assumed that $\mathcal{G}$ contains the last G columns of Y w.l.o.g. This equation allows computing the noise impact of the input variables in $\mathcal{G}$ using the original LS solution [remember that $\mathbf{x}_{\mathcal{G}}$ consists of the last G entries of $\hat{\mathbf{x}}$ as in (8)]. For the case where $G = 1$, i.e., when evaluating the impact of noise with variance σ^2 on a single input variable, (10) reduces to the elegant equation [compare with (4)]

$$I_k(\sigma^2) = \frac{|x_k|^2}{(1/\sigma^2) + q_k}. \quad (11)$$

This noise-impact metric I_k can be viewed as a generalization of the utility metric U_k, as a comparison between (4) and (11) shows that $I_k \to U_k$ if $\sigma^2 \to \infty$. This should not come as a surprise, as adding infinitely large noise to the observations of the kth input variable essentially results in the same loss of information as when the kth input variable would be removed. Similarly, the group-impact equation (10) reduces asymptotically to the group-utility equation (8) if the diagonal entries in Σ grow to infinity.

Redundant input variables

If there is redundancy in the set of input variables, i.e., there is a linear dependency or almost perfect correlation between some of the columns in Y, then the solution of (1) becomes nonunique or ill conditioned. A common strategy is then to compute the LS solution with the smallest $\ell 2$ norm, which is advantageous against overfitting [5], [12]. In the sequel, we denote $\mathcal{R}$ as the set containing all input variables that are redundant, i.e., all columns of Y that consist of a linear combination of the other columns of Y. Note that, by definition, $U_k = 0$ for $k \in \mathcal{R}$, as the removal of a redundant variable does not impact the LS cost.

If $\mathcal{R}$ is nonempty, then R^{-1} does not exist, and the LS solution of (1) is not unique, in which case the LS solution with minimal $\ell 2$ norm is given by

$$\hat{\mathbf{x}} = Y^+\mathbf{d} = R^+\mathbf{r}, \quad (12)$$

where R^+ denotes the Moore–Penrose pseudoinverse of R and the second equality follows from the identity $X^+ = (X^TX)^+X^T$, which holds for the pseudoinverse of any matrix X [10], [13]. As R^{-1} simply has to be replaced with R^+ in (2), it is then tempting to also compute the utility U_k by setting q_k in (4) equal to the kth diagonal element of R^+ instead of R^{-1}. Although it can be shown that this yields the correct utility values U_k for the nonredundant variables $k \notin \mathcal{R}$, it will result in incorrect (nonzero) utility values for the redundant variables $k \in \mathcal{R}$. The proof of this statement is omitted for conciseness but follows relatively straightforwardly from some subresults in [14].

To fix this issue, we have to modify (4) to enforce that U_k is small (near zero) for $k \in \mathcal{R}$, whereas nonredundant variables $k \notin \mathcal{R}$ should receive a nonzero U_k that approximates (3). Furthermore, although the removal of a redundant variable will not affect the LS cost, it will increase the $\ell 2$ norm, i.e., $\|\hat{\mathbf{x}}_{-k}\| \geq \|\hat{\mathbf{x}}\|$

for $k \in \mathcal{R}$, To maximally avoid overfitting, we would like the modified utility measure to also reflect this change in norm, such that removing the redundant input value with the lowest modified utility also induces the least increase of the $\ell 2$ norm. We will show that both of these goals can be achieved if we generalize the utility definition to a standard ridge regression framework.

In ridge regression, an $\ell 2$-norm penalty is added to the LS cost function [10], i.e., (1) becomes

$$\min_{\mathbf{x}} \frac{1}{N}\left(\| Y\mathbf{x} - \mathbf{d}\|^2 + \lambda \|\mathbf{x}\|^2\right), \quad (13)$$

where λ is a user-defined regularization parameter which has

$$\hat{\mathbf{x}} = R_\lambda^{-1}\mathbf{r} = (R + \lambda I)^{-1}\mathbf{r} \quad (14)$$

as a minimizer [10]. Let us now define utility as we did before in (3) but this time based on the regularized cost function (13) instead, i.e.,

$$U_k(\lambda) \triangleq \frac{1}{N}\left(\| Y_{-k}\hat{\mathbf{x}}_{-k} - \mathbf{d}\|^2 - \| Y\hat{\mathbf{x}} - \mathbf{d}\|^2\right) + \frac{\lambda}{N}\left(\|\hat{\mathbf{x}}_{-k}\|^2 - \|\hat{\mathbf{x}}\|^2\right). \quad (15)$$

Note that we have not used the min operators this time, but instead we plugged in the minimizers to explicitly separate the increase in LS cost in the first term and the increase in $\ell 2$ norm in the second term.

The following three theoretical results, which are proven in "Proofs of the Three Results on the Extended Utility Metric," demonstrate that this new utility definition (15) indeed resolves the aforementioned issues and can still be calculated efficiently (these results can also easily be extended to the group-utility framework):

- *Result 1 (efficient calculation)*: The modified utility $U_k(\lambda)$ defined in (15) can be calculated using the efficient formula (4) where q_k is now set to the kth diagonal element of R_λ^{-1} instead of R^{-1}.
- *Result 2 (consistency)*: If $0 < \lambda \ll \epsilon$, with ϵ the smallest nonzero eigenvalue of R, then $U_k(\lambda) \approx 0$ if $k \in \mathcal{R}$, and $U_k(\lambda) \approx U_k$ if $k \notin \mathcal{R}$, where the approximations become asymptotically exact for an arbitrarily small λ.
- *Result 3 (minimum norm revealing)*: If λ is sufficiently small, $U_k(\lambda)$ will be smallest for the $k \in \mathcal{R}$ that results in the smallest $\ell 2$-norm $\|\hat{\mathbf{x}}_{-k}\|$ after removal of input k. More specifically,

Proofs of the Three Results on the Extended Utility Metric

Here we provide the outline of the proofs for the three results listed in the "Redundant Input Variables" section.

Result 1: It can be easily verified that the derivation in "Derivation of the Group-Utility Core Equation (8)" is also valid for $U_k(\lambda)$ if R is replaced with R_λ everywhere.

Result 2: If $\lambda \to 0$, the second term in (15) will vanish, so it can be ignored. Furthermore, it is a known fact that

$$\lim_{\lambda \to 0} R_\lambda^{-1}\mathbf{r} = R^+\mathbf{r}, \quad (S10)$$

which holds even if R^{-1} does not exist (see, e.g., [10, p. 263]). This means that $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_{-k}$ get asymptotically close to the minimum-norm least squares solution of the nonregularized cost (1) when $\lambda \to 0$. As a result, the first term gets asymptotically close to U_k according to the original definition of the nonregularized utility in (3), which is by definition equal to zero if $k \in \mathcal{R}$.

Result 3: If $k \in \mathcal{R}$, then both terms in (12) will vanish if $\lambda \to 0$, i.e., $U_k(\lambda) \to 0$ (see Result 2). Note that the second term vanishes linearly with λ. Therefore, to prove Result 3, we have to show that the first term vanishes superlinearly with λ, $\forall\, k \in \mathcal{R}$, such that the second term dominates over the first term if λ becomes small. To this end, we study $\lim_{\lambda \to 0} U_k(\lambda)/\lambda$ instead. Based on l'Hôpital's rule, we find that

$$\forall\, k \in \mathcal{R}\colon \lim_{\lambda \to 0} \frac{U_k(\lambda)}{\lambda} = \lim_{\lambda \to 0} \frac{dU_k(\lambda)}{d\lambda}, \quad (S11)$$

i.e., the limit is obtained by taking the derivative of $U_k(\lambda)$. By plugging (14) into (15) and expanding it in its quadratic terms, we find that the derivative of (15) can be calculated as

$$\frac{dU_k(\lambda)}{d\lambda} = \frac{d}{d\lambda}(\mathbf{r}^T R_\lambda^{-1}\mathbf{r} - \mathbf{r}_{-k}^T R_{\lambda,-k}^{-1}\mathbf{r}_{-k}), \quad (S12)$$

where $R_{\lambda,-k}$ denotes the matrix R_λ with the kth row and column removed, and $\mathbf{r}_{-k}$ denotes $\mathbf{r}$ with the kth entry removed. The following basic identity is found in [13] and [18] for an invertible matrix A:

$$\frac{dA^{-1}}{dx} = -A^{-1}\frac{dA}{dx}A^{-1}.$$

Using this identity, it can be straightforwardly found that (S12) reduces to

$$\frac{dU_k(\lambda)}{d\lambda} = (\mathbf{r}_{-k}^T R_{\lambda,-k}^{-2}\mathbf{r}_{-k} - \mathbf{r}^T R_\lambda^{-2}\mathbf{r}).$$

Taking the limit for $\lambda \to 0$ and using (S10) and (S11), this reduces to

$$\forall\, k \in \mathcal{R}\colon \lim_{\lambda \to 0} \frac{U_k(\lambda)}{\lambda} = \| R_{-k}^+\mathbf{r}_{-k}\|^2 - \| R^+\mathbf{r}\|^2 = \|\hat{\mathbf{x}}_{-k}\|^2 - \|\hat{\mathbf{x}}\|^2,$$

where the last equality immediately follows from (12). This shows that selecting the redundant variable $k \in \mathcal{R}$ with the lowest utility will induce the lowest increase in $\ell 2$ norm, which proves Result 3.

■

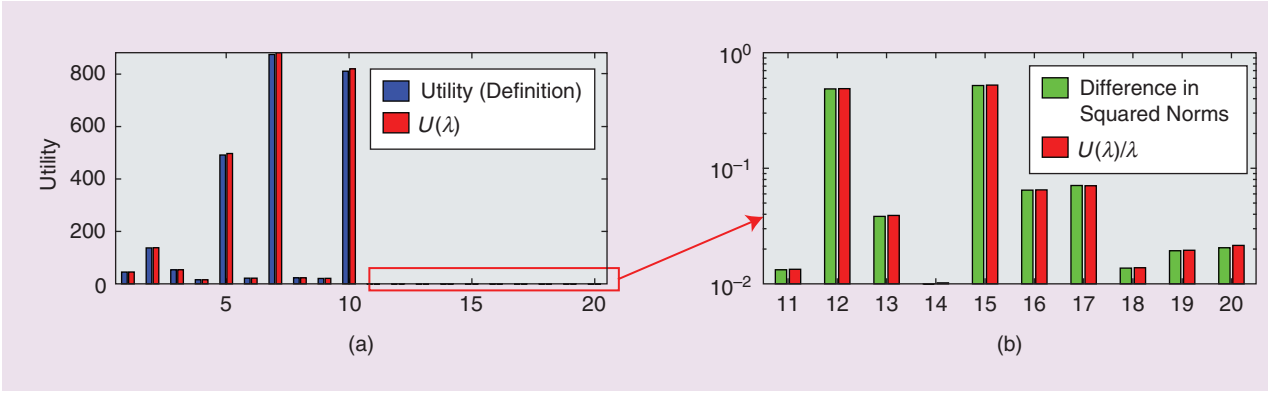


FIGURE 1. (a) The validation of the extended utility metric for the case where the last ten input variables are redundant and (b) the zoom of the last ten entries.

$$\forall\, k \in \mathcal{R}\colon \frac{U_k(\lambda)}{\lambda} \approx \left\|\hat{\mathbf{x}}_{-k}\right\|^2 - \left\|\hat{\mathbf{x}}\right\|^2,$$

where the approximation becomes asymptotically exact for an arbitrarily small λ.

To validate these results, we have calculated the modified utility metric (15) on a toy example with random data with $M = 20$ input variables. The last five columns of Y are generated as random linear combinations of columns 11–15, such that $\mathcal{R} = \{11,\ldots,20\}$. We set $\lambda = \epsilon/100$, where ϵ denotes the smallest nonzero eigenvalue of R. Figure 1(a) shows the values $U_k(\lambda)$ for $k = 1,\ldots,20$ in blue for the naive calculation based on the definition (15) using (12) to find $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_{-k}$, and in red when calculated using the efficient equation (5) where R is replaced with R_λ (see Result 1). It can be observed that both calculation methods result in the same utility value. Note that the ten redundant variables have an almost-zero utility, which is consistent with Result 2. To validate Result 3, we zoom in on the ten redundant variables [Figure 1(b)] and plot the difference $\left\|\hat{\mathbf{x}}_{-k}\right\|^2 - \left\|\hat{\mathbf{x}}\right\|^2$ (in green) versus the value $\left(U_k(\lambda)\right)/\lambda$, where we observe that both result in the same value. This allows selecting the redundant input variable that will yield the smallest increase in $\ell 2$ norm when removed (in this case, variable 14).

Computational benefits and implications for variable subset selection

To demonstrate the impressive reduction in computation time achieved by the core equations (4) and (5) and their generalizations/extensions, we measured the calculation times on a standard laptop running MATLAB. In Figure 2, we compare the time to compute the complete utility vector $\mathbf{u} = [U_1,\ldots,U_M]^T$ when using the efficient equation (5) and using a naive calculation based on the definition (3), as a function of the number of input variables M. We performed the naive calculation two times: once with and once without redundant input variables, where (12) is used to find the minimum-norm LS solution in the former case. In both cases, the computation time is several orders of magnitude lower when using (5).

These strong computational simplifications facilitate the use of (group-)utility metrics for a backward greedy variable selection procedure—even in large-scale problems—in which the input variables with lowest utility are recursively removed one by one, until a sufficiently small set is obtained or until any removal would result in a too-large increase in LS cost [1]–[4], [6]. This can be viewed as an alternative for the well-known (group-)LASSO algorithm [9], [15]. The backward greedy algorithm can even be shown to be optimal if an exact or almost-exact sparse solution exists [1]. In a similar fashion, the noise-impact metrics (10) and (11) can be used for a greedy adaptive quantization, e.g., where in each iteration a certain amount of quantization noise is added to the input with the lowest

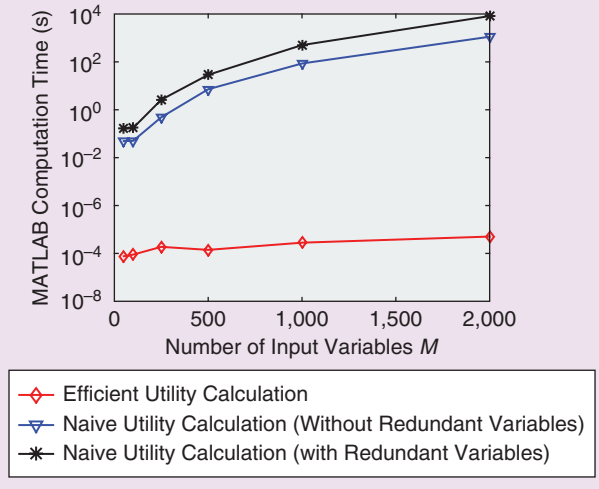


FIGURE 2. The computing time in MATLAB to calculate the utility of M input variables using the efficient equation (5) and using a naive calculation based on the definition (3).

noise impact [11], [16]. Finally, a utility-based greedy variable selection based on (15) yields a combination of variable selection with $\ell 2$-norm minimization, which is akin to the so-called elastic net procedure [17].

Overall, utility-based greedy versions have some useful properties, i.e., they bypass the tedious tuning of sparsity-inducing regularization parameters, they are cheap to compute, and they are easy to implement. Furthermore, the low computational complexity is particularly attractive for online utility tracking, e.g., in recursive LS adaptive filters to (temporarily) eliminate signals of which the utility goes under a predefined threshold [3] or to guarantee that the overall loss does not exceed a predefined threshold. Note that every time a (group of) input variable(s) $\mathcal{G}$ is removed, the LS solution and inverse autocorrelation matrix have to be updated according to the remaining set of variables. These updates can be efficiently calculated using (S2) and (S5) in "Derivation of the Group-Utility Core Equation (8)." Indeed, after the removal of the input variables in $\mathcal{G}$, these equations allow recursively updating the inverse of the reduced autocorrelation matrix $R_{\mathcal{G}}^{-1}$ and the corresponding LS solution at a low cost, based on the original R^{-1} and $\hat{\mathbf{x}}$. For the nongrouped case, i.e., $G = 1$, the matrix inversion of Q in (S2) and (S5) reduces to a simple scalar inversion.

If one also wants to monitor the utility of input variables that are to be added to the model, e.g., for a forward greedy variable selection instead of backward deletion, there also exist additive versions of the utility metric with efficient calculation schemes [3], [4]. However, it should be noted that these metrics are less elegant and computationally less attractive than the deletion-based utility metrics that were introduced previously, yet they are still more efficient than a naive brute-force computation.

Utility is a powerful heuristic for input variable assessment.

Conclusions

In this article, the core equations for the efficient calculation of utility metrics have been reviewed and unified. These have been extended toward group metrics and a minimum-norm revealing utility metric. All of these metrics can be elegantly and inexpensively calculated, thereby making them attractive as a quick-and-dirty tool for model interpretation, online signal quality assessment, or greedy variable selection.

Acknowledgments

I would like to acknowledge the financial support of the KU Leuven Research Council for project C14/16/057; FWO (Research Foundation Flanders) for projects G.0031.14, 1.5.123.16N, G.0D75.16N, and G.0A49.18N; and the EU Horizon 2020 research and innovation program under grant agreement number 766456 (project AMPHORA).

Author

Alexander Bertrand (alexander.bertrand@esat.kuleuven.be) is an assistant professor with the Department of Electrical Engineering, KU Leuven, Belgium. His research focuses on signal processing for (distributed) sensor arrays with a main interest in biomedical applications. He was a visiting researcher at the University of California, Los Angeles, in 2010 and at the University of California, Berkeley, in 2013. Since 2017, he has served as an associate editor of *IEEE Transactions on Signal Processing* and, since 2016, for the IEEE International Engineering in Medicine and Biology Conference. He is a member of the Sensor Array and Multichannel Signal Processing Technical Committee of the IEEE Signal Processing Society (SPS) and a board member of the IEEE SPS Benelux Chapter. He is a Senior Member of the IEEE.

References

[1] C. Couvreur and Y. Bresler, "On the optimality of the backward greedy algorithm for the subset selection problem," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 3, pp. 797–808, 2000.

[2] L. Scott and B. Mulgrew, "Sparse LCMV beamformer design for suppression of ground clutter in airborne radar," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2843–2851, 1995.

[3] A. Bertrand and M. Moonen, "Efficient sensor subset selection and link failure response for linear MMSE signal estimation in wireless sensor networks," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, 2010, pp. 1092–1096.

[4] J. Szurley, A. Bertrand, P. Ruckebusch, I. Moerman, and M. Moonen, "Greedy distributed node selection for node-specific signal estimation in wireless sensor networks," *Signal Process.*, vol. 94, pp. 57–73, Jan. 2014.

[5] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[6] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone subset selection for MVDR beamformer based noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 550–563, 2018.

[7] S. P. Chepuri and G. Leus, "Sensor selection for estimation, filtering, and detection," in *Proc. 2014 Int. Conf. Signal Processing and Communications (SPCOM)*, pp. 1–5.

[8] A. Bertrand, J. Szurley, P. Ruckebusch, I. Moerman, and M. Moonen, "Efficient calculation of sensor utility and sensor removal in wireless sensor networks for adaptive signal estimation and beamforming," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5857–5869, 2012.

[9] R. Tibshirani, "Regression shrinkage and selection via the LASSO: A retrospective," *J. Roy. Statist. Soc.: Series B (Statist. Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[10] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[11] F. de la Hucha Arce, F. Rosas, M. Moonen, M. Verhelst, and A. Bertrand, "Generalized signal utility for LMMSE signal estimation with application to greedy quantization in wireless sensor networks," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1202–1206, 2016.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. (2017). Understanding deep learning requires rethinking generalization. [Online]. Available: https://arxiv.org/abs/1611.03530

[13] K. B. Petersen and M. S. Pedersen. (2012). The matrix cookbook. [Online]. Available: https://archive.org/details/imm3274

[14] C. A. Rohde, "Generalized inverses of partitioned matrices," *J. Soc. Ind. Appl. Math.*, vol. 13, no. 4, pp. 1033–1035, 1965.

[15] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc.: Series B (Statist. Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[16] F. de la Hucha Arce, M. Moonen, M. Verhelst, and A. Bertrand. (2017). Adaptive quantization for multichannel Wiener filter-based speech enhancement in wireless acoustic sensor networks. *Wireless Commun. Mobile Comput.* [Online]. Available: https://www.hindawi.com/journals/wcmc/2017/3173196/

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc.: Series B (Statist. Methodology)*, vol. 67, pp. 301–320, Mar. 2005.

[18] S. M. Selby, *Standard Mathematical Tables*. Boca Raton, FL: CRC, 1974.

Zsolt Kollár, Ferenc Plesznik, and Simon Trumpf

Observer-Based Recursive Sliding Discrete Fourier Transform

In the field of digital signal analysis and processing, the ubiquitous domain transformation is the discrete Fourier transform (DFT), which converts the signal of interest within a limited time window from discrete time to the discrete frequency domain. The active use in real-time or quasi-real-time applications has been made possible by a family of fast implementations of the DFT, called *fast Fourier transform* (*FFT*) algorithms.

Although highly optimized and efficient FFT algorithms are available, their operation remains block oriented with nonrecursive operations.

Although highly optimized and efficient FFT algorithms are available, their operation remains block oriented with nonrecursive operations. An alternative approach to this technique is the sliding DFT (SDFT), where the calculations are performed for a fixed-size sliding window.

The basic idea behind the SDFT algorithm is to recursively calculate the DFT spectrum of the input stream [1], [2]. It is based on a Lagrange structure, built up on a comb filter and complex resonators for the various frequency bins. The biggest disadvantage of this algorithm is that it suffers from stability problems caused by numerical imperfections. Various solutions have been proposed to counteract this effect, keeping the original functionality. The modulated SDFT (mSDFT) [3] addresses the problem with a modified structure moving the complex multiplication factor out of the resonator. Another SDFT variant is the hopping SDFT (hSDFT) [4], which is optimized for the calculation of the SDFT with larger steps (L) than a single sample but smaller than the observation window: $L = 2^a < N$.

Digital Object Identifier 10.1109/MSP.2018.2853196
Date of publication: 13 November 2018

In this article, we investigate the observer-based SDFT (oSDFT), a lesser-known alternative solution for the recursive calculation of the DFT that is based on the observer theory. It was originally developed by Hostetter [5] and generalized by Péceli [6]. Software implementation issues of the structure were recently presented in [7]. The structure is proved to be stable, with a small sensitivity to numerical imperfections. Throughout this article we will compare it to the SDFT and mSDFT structures.

SDFT

The formula for calculating the DFT coefficient in the kth frequency position over the N samples block of $x[n]$ is given as

$$X_k = \sum_{n=0}^{N-1} x[n] W_N^{-kn}, \quad k = 0 \ldots N-1, \tag{1}$$

where $W_N = e^{j(2\pi/N)}$ with j being the imaginary unit. The calculation of (1) in a sliding manner, the DFT component can be expressed as

$$X_k^C[n+1] = \sum_{m=0}^{N-1} x[q+m] W_N^{-km}, \tag{2}$$

where $k = 0 \ldots N-1$ and $q = n - N + 1$. Through this operation we obtain a rotating DFT coefficient, a complex DFT component $X_k^C[n]$, since $x[n]$ slides while W_N^{-km} stands still relative to the sampling window. The upper index C in $X_k^C[n]$ refers to the component nature of the DFT value to distinguish it from the DFT coefficient $X_k[n]$. Given a periodic signal with periodicity of N, the DFT component equals the DFT coefficient at every Nth step

$$X_k^C[n] = X_k, \quad n = 0, N, 2N \ldots. \tag{3}$$

The recursive equivalent of (2) can be expressed based on the previous DFT component $X_k^C[n]$, the current signal sample $x[n]$ and the former signal sample $x[n-N]$ as

$$X_k^C[n+1] = W_N^k (X_k^C[n] + (x[n] - x[n-N])). \tag{4}$$

Figure 1 shows how (4) can be implemented as a comb filter followed by a resonator stage. The resonator stage is an integrator containing a complex multiplication factor, which is an infinite impulse response (IIR) filter. The transfer function of the SDFT structure can be expressed as

$$H_{k,\mathrm{SDFT}}(z) = \frac{X_k^C(z)}{X(z)} = (1 - z^{-N}) H_k(z), \tag{5}$$

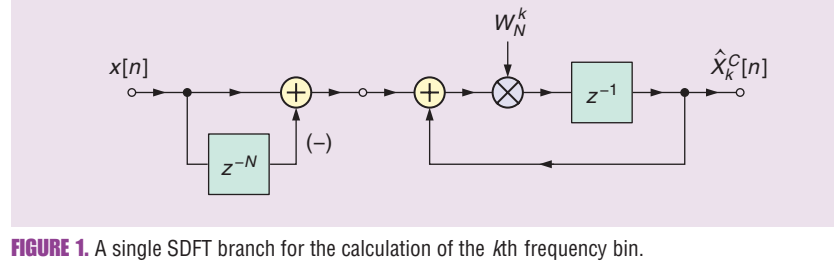


FIGURE 1. A single SDFT branch for the calculation of the kth frequency bin.

where the transfer function of the resonator $H_k(z)$ can be determined as

$$H_k(z) = \frac{W_N^k z^{-1}}{1 - W_N^k z^{-1}}. \tag{6}$$

This structure is considered to be only marginally stable in practice [1] as the W_N^k poles, in the presence of numerical imperfections, may be located inside or outside the unit circle. To avoid a potential divergence in the results, without altering the structure, a straightforward method is given by a compensated SDFT [1] enforcing the poles inside the unit circle by applying a constant multiplication factor r, slightly smaller than one, to all W_N^k factors. As a drawback, it leads to a modified DFT calculation, thus it gives inaccurate results [3].

mSDFT

A slightly modified structure of the SDFT is the mSDFT [3], which aims to solve the aforementioned stability issue without sacrificing accuracy through utilizing the DFT's frequency shift theorem property. The mSDFT-based structure calculating the kth frequency bin is shown in Figure 2.

First, it transforms the kth frequency bin to dc ($k = 0$) by a complex multiplication with the sequence W_N^{-kn}, then the calculations of (2) is applied for $k = 0$. Finally, it transforms the result back by up conversion with a multiplication of the sequence W_N^{kn}. With this described technique, the resonators became stable integrators performing simple averaging.

Via down conversion, the mSDFT calculates the DFT coefficients, recursively as

$$\hat{X}_k[n+1] = \left(\hat{X}_k[n] + W_N^{-kn}(x[n] - x[n-N])\right). \tag{7}$$

To get the same output as the SDFT in (2), namely the DFT component, an upconversion sequence has to be applied by multiplying the DFT coefficient X_k with W_N^{kn},

$$\hat{X}_k^C[n] = \hat{X}_k[n]\, W_N^{kn}. \tag{8}$$

As a result, the transfer function of an mSDFT branch is theoretically identical with the transfer function of an SDFT branch presented in (5).

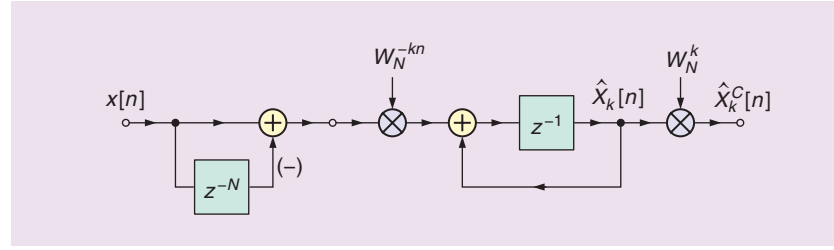


FIGURE 2. A single mSDFT branch for the calculation of the kth frequency bin.

oSDFT

In this section, we introduce a lesser-known alternative approach to the SDFT problem: the oSDFT. The main idea behind the oSDFT, the application of the state observer, is widely used in system control theory [8] and also can be successfully adapted for digital signal processing purposes [5], [6].

The observer theory model

The observer theory supposes the system model that the measured signal ($x[n]$) is a linear combination of the elements of a given basis system

$$x[n] = \sum_{k=0}^{N-1} X_k c_k[n], \tag{9}$$

where N is the rank of the basis system, $c_k[n]$ is the kth basis vector, and X_k its matching weighting factor.

This system model is considered for the signal construction and can be seen as the generator of the signal $x[n]$ on the left side of Figure 3, wherein weighting factors are stored in discrete integrators as initial values.

The observer, which can be seen on the right side of Figure 3, by mirroring the system model's structure, estimates the $x[n]$ input signal's X_k weighting factors in its internal state variables $\hat{X}_k$ through signal decomposition. For the refinement of this estimation, a negative feedback is created with a reconstructed signal $y[n]$ from the estimated $\hat{X}_k$ weighting factors. This negative feedback also acts as a stabilizing control loop for our state observer [6], [9].

The kth state variable can be expressed as

$$\hat{X}_k[n+1] = \hat{X}_k[n] + g_k[n] \times (x[n] - y[n]), \tag{10}$$

where $y[n]$ can be expressed as

$$y[n] = \frac{1}{N}\sum_{k=0}^{N-1} c_k[n]\hat{X}_k[n] = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}_k^C[n]. \tag{11}$$

Péceli proves the following four statements in [6], which are crucial from the SDFT aspect:

1) The observer is convergent, if $c_k[n]$ and $g_k[n]$ are basis-reciprocal basis systems for $n = 0 \ldots N-1$ with a normalization factor of $1/N$:

$$\frac{1}{N}\sum_{n=0}^{N-1} c_k[n]\, g_k[n] = 1, \forall k. \tag{12}$$

   Moreover, in this scenario the system is deadbeat in N step (i.e., after N steps $\hat{X}_k[n] = X_k$).
2) The state variables $\hat{X}_k$ of the observer are the DFT coefficients according to (9), if $g_k[n] = W_N^{-kn}$ and $c_k[n] = W_N^{kn}$. The modulated state variables $\hat{X}_k^C[n+1]$ are the sliding DFT components of the input signal $x[n]$ as presented in (2).
3) Based on the fact that the oSDFT structure is a control loop with a negative feedback, the transfer function of the kth branch of the oSDFT can be expressed as

$$H_{k,\,\mathrm{oSDFT}}(z) = \frac{X_k^C(z)}{X(z)} = \frac{H_k(z)}{1 + \frac{1}{N}\sum_{k=0}^{N-1} H_k(z)}, \tag{13}$$

   where $H_k(z)$ is given in (6).
4) The oSDFT structure is equivalent to the SDFT structure presented in Figure 1 in such a way that their transfer functions of the kth branches are

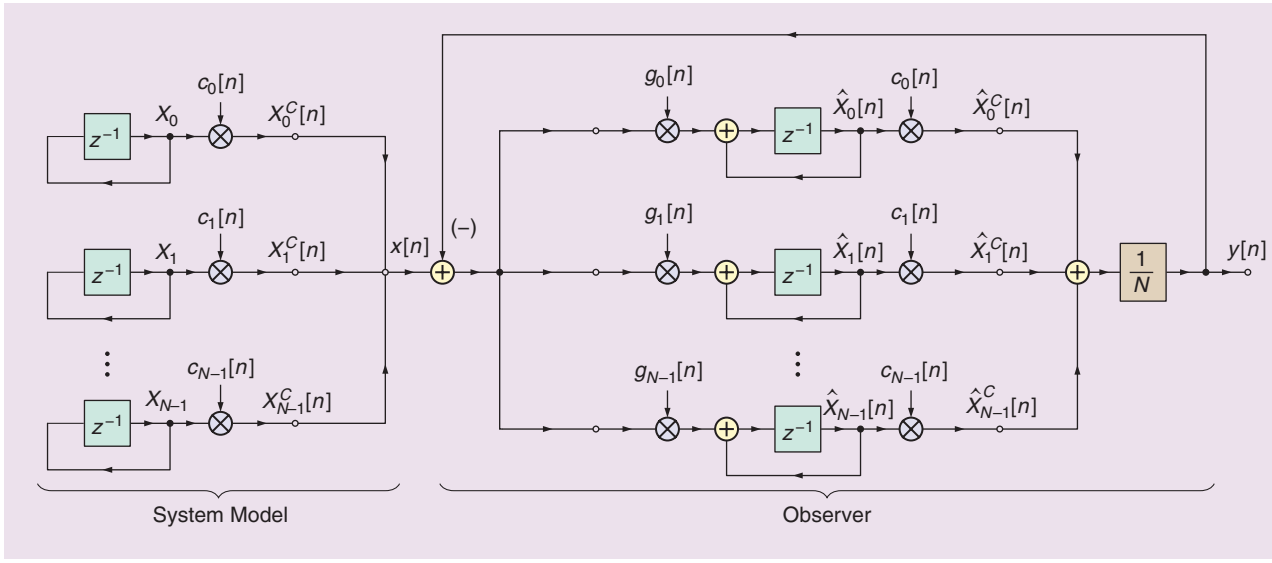


FIGURE 3. The observer theory model: system model and observer.

equal. The proof of the theoretical equivalence can be found in "Proof of the Equivalence of the SDFT and oSDFT."

Resonator-based oSDFT

An alternative version of the oSDFT structure is depicted in Figure 4, which is based solely on resonators, which are IIR filters, without down- and up-converters, similar to the SDFT structure. The proof of the equivalence of the two oSDFT structures is provided in "Proof of the Equivalence of Two oSDFT Structures."

Complexity analysis

In this section we analyze the computational complexity and memory requirements for the various SDFT structures when calculating all N DFT components. The comparison will be performed based on the calculation of a single input sample. All elements are considered to be complex valued. The requirements are summarized in Table 1.

Independent from the chosen algorithms, N registers are required for storing the state variables of the resonators or the integrators. The SDFT and mSDFT algorithms used N additional registers for the comb filter's N-step delay line. Furthermore, the SDFT and the resonator-based oSDFT structure require N memories to store the multiplication factors (W_k^N) for all branches. The mSDFT and the oSDFT structures can obtain the values of the modulator and demodulator signals (W_N^{nk} and W_N^{-nk}) from a look-up-table (LUT). The LUT stores N samples for each branch, as the values are periodic to N.

Resonator-based implementations (i.e., SDFT and resonator-based oSDFT) require N multipliers, whereas the demodulation and modulation approaches (i.e., mSDFT and oSDFT) use $2N$ multipliers.

For all algorithms, each branch requires one two-input adder. In case of the oSDFT structures, they both apply an N-input adder to calculate the feedback signal $y[n]$ and a two-input adder is used to calculate the difference of the input and the feedback signal as shown in (10).

The basic idea behind the SDFT algorithm is to recursively calculate the DFT spectrum of the input stream.

The biggest advantage of these structures compared to the FFT-based block-wise calculation is that the operational load can be distributed between the incoming samples, as the SDFT structure can operate continuously. As soon as the Nth sample of a block has arrived, the calculation with the last input sample can be executed in a single step with parallel calculations. The spectral components will be available faster compared to the block-wise operational FFT where this can be performed in $\log_2 N$ steps.

Simulations

Floating-point implementation

A simulation environment for the comparison of the aforementioned sliding DFT algorithms (i.e., mSDFT, oSDFT, and resonator-based oSDFT) was developed in MATLAB2017a (x64 PC). For the algorithms we applied 32-bit, single-precision, floating-point arithmetic and compared the numerical imperfections of the various methods to the results of a 64-bit, double-precision arithmetic sliding FFT, utilizing the built-in *fft* function. We applied the following simulation scenario: within an $N = 64$ frequency bin setup, an aperiodic white gaussian noise was used, where the noise signal was generated using the built-in *randn* function operating with default seed option and a unit variance as:

```
% setting the seed
rng('default');
% variance of the noise signal
var = 1;
% noise signal with single precision
x = var * randn(1,32000,'single');
```

The usage of white noise as excitation signal ensures that all the branches system-wide are statistically equally

Proof of the Equivalence of the SDFT and oSDFT

To prove the equivalence of the SDFT and oSDFT structures, we will show that the transfer functions for each branch, $H_{k,\mathrm{SDFT}}(z)$ and $H_{k,\mathrm{oSDFT}}(z)$, are equal. The transfer function of the SDFT and the oSDFT structures are expressed according to (5) and (13) as

$$H_{k,\mathrm{SDFT}}(z) = (1 - z^{-N}) \cdot H_k(z), \tag{S1}$$

$$H_{k,\mathrm{oSDFT}}(z) = \frac{H_k(z)}{1 + H_0(z)} = \frac{H_k(z)}{1 + \frac{1}{N}\sum_{k=0}^{N-1} H_k(z)}, \tag{S2}$$

where $H_k(z)$ is the transfer function of the kth resonator and $H_0(z)$ is the transfer function of the open loop in the oSDFT structure. The transfer function $H_k(z)$ is determined as

$$H_k(z) = \frac{W_N^k z^{-1}}{1 - W_N^k z^{-1}}. \tag{S3}$$

First, we will prove that

$$\sum_{k=0}^{N-1} H_k(z) = \sum_{k=0}^{N-1} \frac{W_N^k z^{-1}}{1 - W_N^k z^{-1}} = N\frac{z^{-N}}{1 - z^{-N}}. \tag{S4}$$

As we unfold and rearrange the first part of (S4) using the formula for the sum of a geometric series, we obtain

$$\sum_{k=0}^{N-1} \frac{W_N^k z^{-1}}{1 - W_N^k z^{-1}} = \sum_{k=0}^{N-1} [(W_N^k z^{-1})^1 + (W_N^k z^{-1})^2 + \cdots + (W_N^k z^{-1})^N + \cdots] = \sum_{p=1}^{\infty} \left[(z^{-1})^p \cdot \sum_{k=0}^{N-1} W_N^{kp} \right]. \tag{S5}$$

Emphasizing the fact that for the sum of the powers of a unit root, the following expression is valid:

$$\sum_{k=0}^{N-1} W_N^{kp} = \begin{cases} N, & \text{if } p = 0, N, 2N, \ldots \\ 0, & \text{otherwise.} \end{cases} \tag{S6}$$

We can simplify (S5) using the formula for the sum of a geometric series to

$$\sum_{k=0}^{N-1} \frac{W_N^k z^{-1}}{1 - W_N^k z^{-1}} = [z^{-N} \cdot N + z^{-2N} \cdot N + \cdots + z^{-NN} \cdot N + \cdots] = N\frac{z^{-N}}{1 - z^{-N}}. \tag{S7}$$

This is what we wanted to prove.

Now, if we substitute (S4) into (S2) we get the following simplified equation:

$$H_{k,\mathrm{oSDFT}}(z) = \frac{H_k(z)}{1 + \frac{1}{N} N \frac{z^{-N}}{1 - z^{-N}}} = \frac{(1 - z^{-N}) H_k(z)}{(1 - z^{-N}) + z^{-N}} = (1 - z^{-N}) H_k(z). \tag{S8}$$

As a result, we have proved that the transfer function of the two structures according to (S1) and (S2) are equivalent. ■

excited, so the behavior of each structure can be better characterized and evaluated as a dynamic system.

The results of the various SDFT methods were compared through double-precision arithmetic to the results of the sliding FFT, and the average error signal over the branches was formulated as

$$\varepsilon[n] = \frac{1}{N}\sum_{k=0}^{N-1} \left| \hat{X}_{k,\mathrm{xSDFT}}^C[n] - \hat{X}_{k,\mathrm{FFT}}^C[n] \right|, \tag{14}$$

where xSDFT stands for the mSDFT and oSDFT algorithms.

In Figure 5, the error progress in the function of the time index is compared in the case of mSDFT and oSDFT. The error of the mSDFT is not stable and slowly drifts over the time samples. On the contrary, the oSDFT algorithm is stable, but it is noisy as well due to the numerical errors.

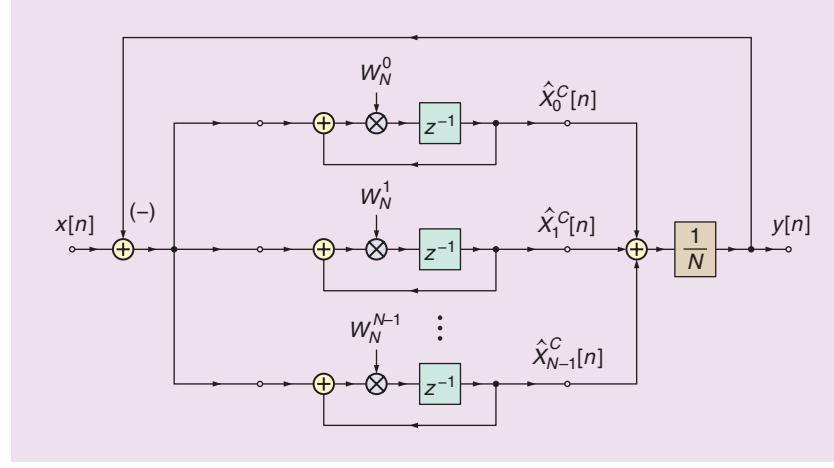


FIGURE 4. A resonator-based oSDFT.

In Figure 6 the error progress in function of the time index is shown for the oSDFT and the resonator-based oSDFT algorithms. Both algorithms produce a stable-but-noisy error over the discrete time samples. Additionally, the oSDFT outperforms the resonator-based oSDFT.

Proof of the Equivalence of Two oSDFT Structures

Both oSDFT algorithms are built upon either one of the two main substructures, namely the down conversion–integrator–up conversion or the resonator scheme as shown in Figure S1. Here, we intend to show the theoretical equivalence of these two substructures. We present this statement through an alternative graphical method, while a mathematical approach can be found in [3] and [12].

Starting from Step (a) in Figure S2:

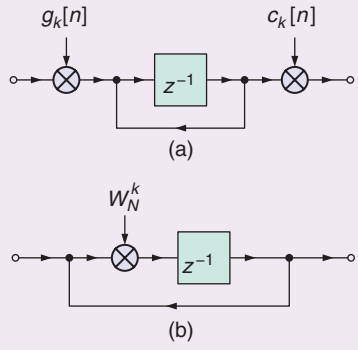


FIGURE S1. Substructures of the oSDFT: the (a) down conversion–integrator–up conversion and (b) resonator scheme.

(b) Push the up-converting sequence into the loop, before the feedback exit point. To ensure the same functionality, we have to compensate for the effect of the newly introduced in-loop multiplication into the feedback path as well.

(c) Move the up-converting sequence even further, through the delay element. Due to the delay element, only the time indexing has to be modified.

(d) Push the compensating term, introduced in Step (b), further down the feedback-loop, until it stands after the feedback entry point. Additionally, to counterpreserve the functionality, we have to also divide the input (i.e., signal) with the compensating term. Finally, as they are in the same position, we can contract the modulating and demodulating sequences into one term. As a result, we have reached the same structure as presented in Figure S1(b) based on the following equivalences:

$$g_k[n] \cdot c_k[n] = W_N^{-nk} \cdot W_N^{nk} = 1, \tag{S9}$$

$$\frac{c_k[n+1]}{c_k[n]} = c_k[1] = W_N^k. \tag{S10}$$

■

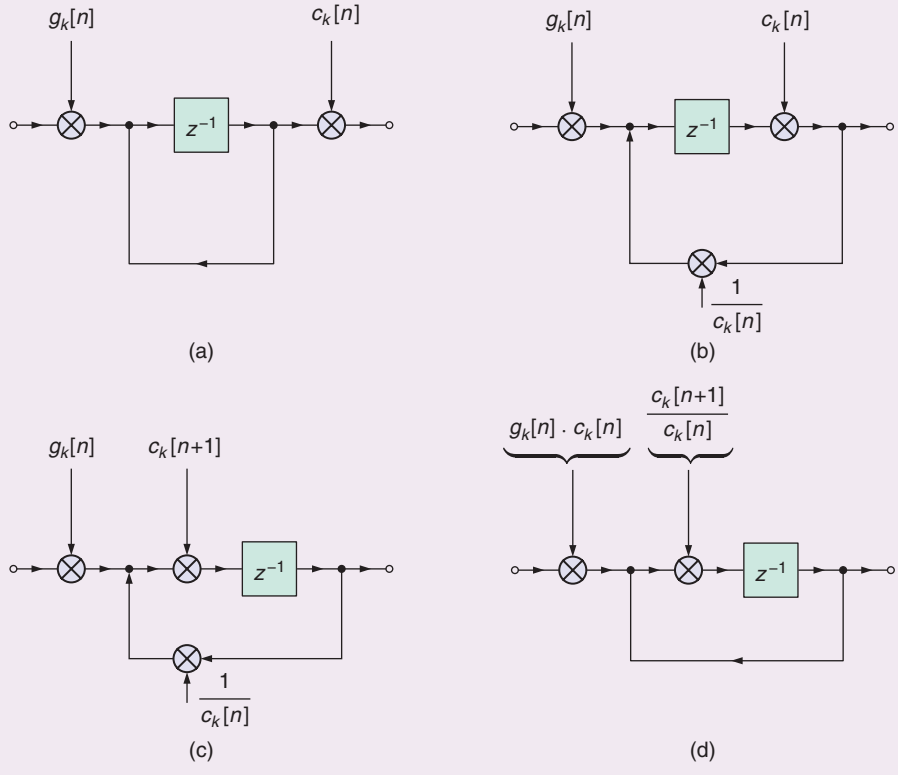


FIGURE S2. (a)–(d) The steps for proving the equivalence of the two oSDFT substructures.

The observed performance difference between the two oSDFT structures has two attributes with a common root cause: multiplication within the resonators with constant W_N^k, at every time step. The two distinct differences experienced in Figure 6 are an offset and a higher noise variance.

The offset is caused by the fact that for every step of n the W_N^{kn} modulator and demodulator values for the oSDFT are taken periodically from a precomputed sequence stored in a LUT, and within this LUT the error introduced by rounding (i.e., finite precision storage) is averaged out over a sequence period. This way the oSDFT's modulation and demodulation process will be more precise regarding the average frequency accuracy over a sequence period than the resonator-based oSDFT, where the numerical error in the constant W_N^k pole can't be averaged out over the same period, thus leading to a constant frequency offset in the center frequency of the resonators.

The higher variance of the error signal comes from the fact that the finite precision multiplication by W_N^k within the resonator's loop is an additional noise source which will dominate the variance due to the structure, thus it will lead to slightly misplaced W_N^k poles in a random manner over the complex plain.

Fixed-point implementation

As to further investigate and cover wider use-case scenarios, the results for fixed-point implementations are also presented.

Table 1. The complexity comparison of the various SDFT structures.

	Memory				Adders	
Type	Read-Only Memory	Random-Access Memory	LUT_N	Multipliers	Two Input	*N* Input
SDFT	*N*	*N* + *N*	0	*N*	*N* + 1	0
mSDFT	0	*N* + *N*	1	2*N*	*N* + 1	0
oSDFT	0	*N*	1	2*N*	*N* + 1	1
oSDFT (resonator)	*N*	*N*	0	*N*	*N* + 1	1

During the comparison simulations with the 32-bit, single-precision, floating-point variants, the signed fixed-point calculations were implemented with a word length of 32 bits, from which 31 bits were used for the fractional part, and a rounding toward zero method was applied to maintain the stability of the feedback structures. Otherwise, the simulation environment, the test signals, and the error term definition were the same as with the floating-point scenario presented earlier.

With this described technique, the resonators became stable integrators performing simple averaging.

The error progress of the various SDFT structures for signed Q0.31 format fixed-point implementation can be seen in Figure 7. The results are similar to the case where the structures are implemented using single-precision, although the overall errors are slightly smaller for each method. The reason for this is that, the IEEE 754-2008 single-precision standard, used by MATLAB, the fractional part is defined as only 23 bits, thus within the same range, it offers lower resolution, resulting in more imprecise W_N^{kn} modulator and W_N^k pole values.

Furthermore, the averaged error $\varepsilon[n]$ over the samples n for the fixed-point implementation in function of the fraction part is shown in Figure 8. For both methods with an enlarged fraction part, the error is exponentially decreasing.

Summary

In this article, an alternative structure for the calculation of the SDFT, the oSDFT, was presented, which is based on the observer theory, a method taken from control theory. The core structure of the oSDFT is similar to the other SDFT

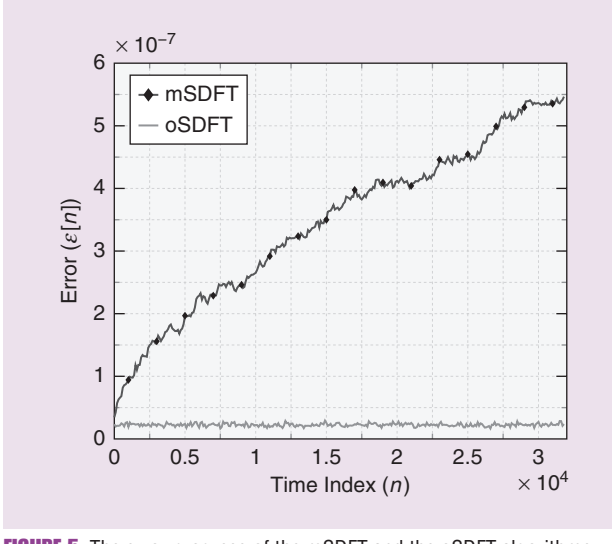


FIGURE 5. The error progress of the mSDFT and the oSDFT algorithms.

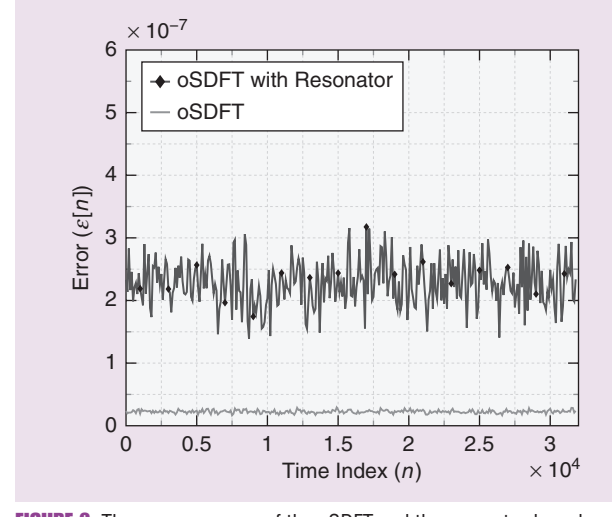


FIGURE 6. The error progress of the oSDFT and the resonator-based oSDFT algorithms.

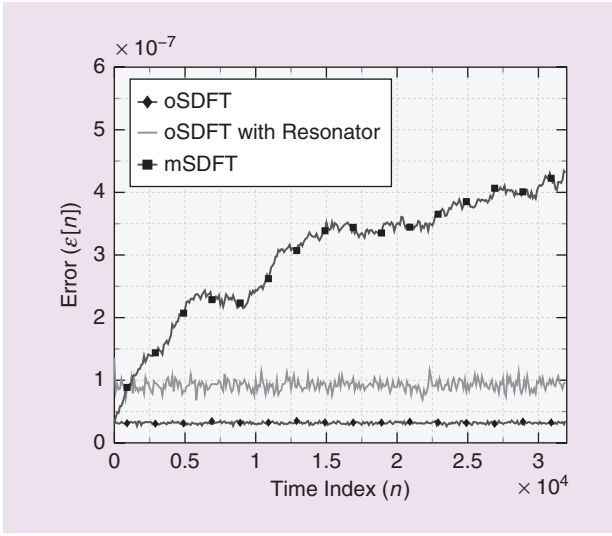


FIGURE 7. The error progress of the oSDFT, the resonator-based oSDFT, and the mSDFT algorithms with fixed-point implementation in signed Q0.31 format using a 32- or 31-bit fractional part.

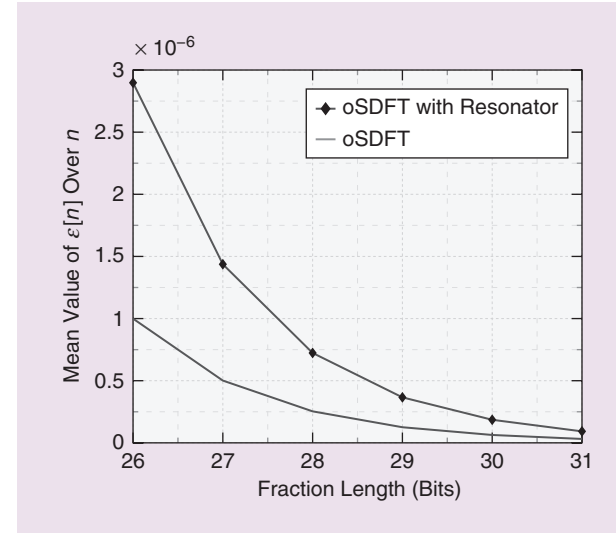


FIGURE 8. The mean error of the oSDFT and the resonator-based oSDFT in function of the fraction length of the signed 32-bit fixed-point implementation.

methods, but it applies a recursive overall feedback branch, which allows the elimination of the feedforward comb filter and its N-tap delay line, achieving long-term stability in contrast to other well-known SDFT methods. The various SDFT structures were also compared based on their memory and arithmetical requirements.

It was also shown that the oSDFT structure has a lower sensitivity to numerical imperfections compared to other SDFT structures. The oSDFT is stable for input signals containing aperiodic white noise as well, due to the control-loop feedback structure, and keeps its stability and behavior with fixed-point implementations as well.

The application of the oSDFT structure can be especially advantageous not only for the long-term stability but because a large percentage of the N DFT components are required to be calculated in a sliding manner. From a practical aspect, the oSDFT structure can be advantageously used as a tunable filter [10] or as a nonlinear adaptive frequency estimator [11].

Independent from the chosen algorithms, *N* registers are required for storing the state variables of the resonators or the integrators.

Acknowledgments

We are thankful to Prof. Péceli Gábor, for his helpful comments and suggestions. This work was supported by the János Bolyai Research Fellowship of the Hungarian Academy of Sciences.

Authors

Zsolt Kollár (kollar@hvt.bme.hu) received his diploma and Ph.D. degree in electric engineering from the Budapest University of Technology and Economics, Hungary, in 2008 and 2013, respectively. He is an associate professor in the Department of Broadband Infocommunications and Electromagnetic Theory at the Budapest University of Technology and Economics, Hungary, where he is the head of the MATLAB laboratory. His research interests are digital signal processing, wireless communication, and quantization issues.

Ferenc Plesznik (plesznik@yahoo.co.uk) received his B.Sc. and M.Sc. degrees in electrical engineering in 2011 and 2016, respectively, from the Budapest University of Technology and Economics, Hungary. His main focus is on digital signal processing and wireless communication systems.

Simon Trumpf (simon.trumpf@student.kit.edu) received his B.Sc. degree in electrical engineering in 2018 from Karlsruhe Institute of Technology, Germany, where he is a currently an M.Sc. degree student in the field of electrical engineering and information technology.

References

[1] E. Jacobsen and R. Lyons, "The sliding DFT," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 74–80, Mar. 2003.

[2] E. Jacobsen and R. Lyons, "An update to the sliding DFT," *IEEE Signal Process. Mag.*, vol. 21, no. 1, pp. 110–111, Jan. 2004.

[3] K. Duda, "Accurate, guaranteed stable, sliding discrete Fourier transform," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 124–127, Nov. 2010.

[4] C. S. Park and S. J. Ko, "The hopping discrete Fourier transform," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 135–139, Mar. 2014.

[5] G. Hostetter, "Recursive discrete Fourier transformation," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 28, no. 2, pp. 184–190, Apr. 1980.

[6] G. Péceli, "A common structure for recursive discrete transforms," *IEEE Trans. Circuits Syst.*, vol. 33, no. 10, pp. 1035–1036, Oct. 1986.

[7] M. Kovács and Z. Kollár, "Software implementation of the recursive discrete Fourier transform," in *Proc. 2017 27th Int. Conf. Radioelektronika*, April 2017, pp. 1–5.

[8] A. V. Oppenheim and G. C. Verghese, *Signals, Systems, and Inference*. Upper Saddle River, NJ; Pearson Education Limited, 2016.

[9] G. Péceli and G. Simon, "Generalization of the frequency sampling method," in *Proc. IEEE Instrumentation and Measurement Technology Conf. IMEKO Tec*, vol. 1, 1996, pp. 339–343.

[10] G. Péceli, "Resonator-based digital filters," *IEEE Trans. Circuits Syst.*, vol. 36, no. 1, pp. 156–159, Jan. 1989.

[11] G. Simon and G. Péceli, "Convergence properties of an adaptive Fourier analyzer," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 46, no. 2, pp. 223–227, Feb. 1999.

[12] C. S. Park, "Fast, accurate, and guaranteed stable sliding discrete Fourier transform," *IEEE Signal Process. Mag.*, vol. 32, no. 4, pp. 145–156, July 2015.

SP

Please send calendar submissions to:
Dates Ahead, Att: Samantha Walter, E-mail: walter.samantha@ieee.org

2018

OCTOBER

25th IEEE International Conference on Image Processing (ICIP)
7–10 October, Athens, Greece.
General Chairs: Christophoros Nikou and Kostas Plataniotis
URL: https://2018.ieeeicip.org

IEEE Workshop on Signal Processing Systems (SiPS)
21–24 October, Cape Town, South Africa.
General Chair: Tokunbo Ogunfunmi
URL: http://www.sips2018.org/

Asilomar Conference on Signals, Systems, and Computers (ACSSC)
28–31 October, Pacific Grove, California, United States.
General Chair: Visa Koivunen
URL: http://www.asilomarsscconf.org/

NOVEMBER

Tenth Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2018)
12–15 November, Honolulu, Hawaii, United States.
General Chairs: Yih-Fang Huang, Anthony Kuh, and Susanto Rahardja
URL: https://apsipa2018.org

Sixth IEEE Global Conference on Signal and Information Processing (GlobalSIP)
26–28 November, Anaheim, California, United States.
General Chairs: Shuguang Cui and Hamid Jafarkhani
URL: http://2018.ieeeglobalsip.org/

15th IEEE International Conference on Advanced Video and Signals-Based Surveillance (AVSS)
27–30 November, Auckland, New Zealand.
General Chairs: Reinhard Klette and Mohan Kankanhalli
URL: https://avss2018.org

Digital Object Identifier 10.1109/MSP.2018.2867121
Date of publication: 13 November 2018

The 44th IEEE International Conference on Acoustics, Speech, and Signal Processing will be held 12–17 May 2019 in Brighton, United Kingdom.

DECEMBER

IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)
6–8 December, Louisville, Kentucky, United States.
General Cochairs: Esam Abdel-Raheem and Myrian Tebald
URL: http://www.isspit.org/isspit/2018/index.html

2018 IEEE International Workshop on Information Forensics and Security (WIFS)
11–13 December, Hong Kong.
General Chair: Ajay Kumar
URL: https://wifs2018.comp.polyu.edu.hk/

2018 IEEE Spoken Language Technology Workshop (SLT)
18–21 December, Athens, Greece.
Cochairs: Vangelis Karkaletsis, Yannis Stylianou, and Srinivas Bangalore
URL: http://www.slt2018.org

2019

MARCH

The Data Compression Conference (DCC)
26–29 March, Snowbird, Utah, United States.
General Chairs: Michael W. Marcellin and James A. Storer
URL: http://www.cs.brandeis.edu/~dcc/index.html

APRIL

IEEE International Symposium on Biomedical Imaging (ISBI)
8–11 April, Venice, Italy.
General Chairs: Marius George Linguraru and Enrico Grisan
URL: https://biomedicalimaging.org/2019/

MAY

44th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
12–17 May, Brighton, United Kingdom.
General Chairs: Saeid Sanei and Lajos Hanzo
URL: http://icassp2019.com

JULY

IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)
2–5 July, Cannes, France.
General Chair: David Gesbert
URL: http://www.spawc2019.org/

SEPTEMBER

27th European Signal Processing Conference (EUSIPCO)
2–6 September, A Coruña, Spain.
General Cochairs: Mónica F. Bugallo and Luis Castedo
URL: http://eusipco2019.org

SP

2018 Index
IEEE Signal Processing Magazine
Vol. 35

This index covers all technical items — papers, correspondence, reviews, etc. — that appeared in this periodical during 2018, and items from previous years that were commented upon or corrected in 2018. Departments and other items may also be covered if they have been judged to have archival value.

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of the paper or other item, and its location, specified by the publication abbreviation, year, month, and inclusive pagination. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication abbreviation, month, and year, and inclusive pages. Note that the item title is found only under the primary entry in the Author Index.

AUTHOR INDEX

Digital Object Identifier 10.1109/MSP.2018.2876858
Date of publication: 13 November 2018

SUBJECT INDEX

Numeric

A

B

C

+ Check author entry for coauthors

+ Check author entry for coauthors

+ Check author entry for coauthors

F

G

H

+ Check author entry for coauthors

+ Check author entry for coauthors

M

+ Check author entry for coauthors

+ Check author entry for coauthors

+ Check author entry for coauthors

S

+ Check author entry for coauthors

T

+ Check author entry for coauthors

U

V

W

+ Check author entry for coauthors

IEEE Access®

Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Best New Journal in STM 2015

Become a published author in 4 to 6 weeks.

Get on the fast track to publication with the multidisciplinary open access journal worthy of the IEEE.

IEEE journals are trusted, respected, and rank among the most highly cited publications in the industry. IEEE *Access* is no exception; the journal is included in Scopus, Web of Science, and has an Impact Factor.

Published online only, IEEE *Access* is ideal for authors who want to quickly announce recent developments, methods, or new products to a global audience.

Publishing in IEEE *Access* allows you to:

- Submit multidisciplinary articles that do not fit neatly in traditional journals
- Reach millions of global users through the IEEE *Xplore*® digital library with free access to all
- Establish yourself as an industry pioneer by contributing to trending, interdisciplinary topics in one of the Special Sections
- Integrate multimedia and track usage and citation data on each published article
- Connect with your readers through commenting
- Publish without a page limit for **only $1,750** per article

Learn more at:
ieeeaccess.ieee.org

IEEE

17-PUB-013 3/17

 (continued from page 128)

of so-called grand challenges, which aim to objectively benchmark the state of the art and promote the open-source publication of standardized data sets, evaluation metrics, algorithms, and software tools. Various successful challenges have already been organized by BISP-TC members and colleagues in recent years, especially at ISBI and MICCAI, but much work remains in keeping them up to date and expanding and harmonizing them. This goes hand in hand with activities to integrate the best-performing data processing methods in user-friendly software tools and to catalog their design criteria, modes of operation, boundary conditions, and optimal parameter settings for the benefit of both researchers and practitioners.

With these and many other ongoing developments, the future is bright for bioimaging and signal processing. Everyone with an interest in the interdisciplinary scope and activities of the BISP-TC is cordially invited to register via the website as an affiliate member. Professionals with expertise in technologies complementary to those covered by the current BISP-TC members are especially encouraged to sign up. The same applies to professionals in industries working within the TC's scope or on closely related topics. Elections of SPS members interested in becoming an associate member or full member are held yearly in the autumn. Membership in any form offers a great opportunity to get actively involved in groundbreaking activities and discussions to help shape the field of bioimaging and signal processing. We look forward to welcoming many new members soon.

Authors

Erik Meijering (meijering@imagescience.org) received his M.Sc. degree in electrical engineering from Delft University of Technology, The Netherlands, and his Ph.D. degree in medical image analysis from Utrecht University, The Netherlands, in 1996 and 2000, respectively. After his postdoctoral work at the Swiss Federal Institute of Technology, Lausanne, Switzerland, he returned to The Netherlands, where he is an associate professor of biomedical image computing at the Erasmus University Medical Center, Rotterdam. He currently serves as the chair of the Bio-Imaging and Signal Processing Technical Committee.

Arrate Muñoz-Barrutia (mamunozb@ing.uc3m.es) received her M.Sc. degree in telecommunication engineering from the Public University of Navarra, Pamplona, Spain, in 1997 and her Ph.D. degree in technical sciences from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. Currently, she is an associate professor in the Department of Bioengineering and Aerospace Engineering at the Universidad Carlos III de Madrid, Spain, where she works in biomedical image processing. She is also a senior researcher at the Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain. In 2016–2017, she was the chair of the Bio-Imaging and Signal Processing Technical Committee and currently serves as its departing vice chair.

Michiel Bacchiani and Eric Fosler-Lussier

An Overview of the IEEE SPS Speech and Language Technical Committee

As part of the IEEE Signal Processing Society (SPS), the Speech and Language Technical Committee (SLTC) promotes research and development activities for technologies that are used to process speech and natural language. Much of the SLTC's efforts are devoted to the annual IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), where the SLTC manages the review of papers covering speech and language and organizes conference sessions, special sessions, and tutorials. In addition, it promotes and supports various workshops, most prominently the Automatic Speech Recognition and Understanding (ASRU) and the Spoken Language Technology (SLT) workshops.

The ASRU and SLT workshops are both held biannually in alternating years. This year, the SLT workshop will be held in Athens, Greece [1]. The SLTC recently conducted a search for the venue

Digital Object Identifier 10.1109/MSP.2018.2863099
Date of publication: 13 November 2018

for ASRU 2019. A four-member workshop subcommittee disseminated calls for proposals as well as used personal connections to reach out and find strong groups that could potentially organize the workshops in diverse geographic locations. The subcommittee supported the prospective proposers in ensuring they constructed a viable, well-planned proposal. The SLTC successfully gathered very strong collaborative proposals from three lead institutions: the Qatar Computing Research (for Doha, Qatar), the National University of Singapore (for Sentosa Island, Singapore), and Friedrich-Alexander-Universität (for Cartagena, Colombia). The SLTC selected Singapore for ASRU 2019 in a close vote (all proposals were within four percentage points), demonstrating the significant influence of the proposers as well as the process to recruit and vet the workshops. We look forward to a successful ASRU 2019 in Singapore!

Our speech and language community is strong and growing. Our 2017 annual election had 54 candidates for 19 positions, with 12 first-time members being elected. Our vice-chair election also had four candidates. ICASSP 2018 submissions in our speech and language area were up as well, with a 40% increase in submissions (to 634) during 2017, representing a 22% share of papers presented at the conference. Our work also received significant appreciation as reflected by IEEE SPS awards including the Society Award earned by Alex Acero and the Meritorious Service Award earned by Mari Ostendorf.

The most recent decade has witnessed language technologies receiving wide acceptance by the general public. Speech recognition interfaces to smartphones and smart speakers that provide question-and-answer or dialog technologies are becoming unremarkable in the eyes of consumers. As a result, the amount of available data from those interactions is growing very rapidly. This increase of data leads to a virtuous cycle where more data allows for larger and/or better-trained state-of-the-art neural network models. The systems' improved performance, in turn, gives an incentive for users to engage more with this technology. As a result, our community has become increasingly entangled with the more fundamental research in machine learning, and the recent maturing of SLT presents itself as a means to widen our community scope. Some exciting directions in that area include leveraging neural machine learning with multiple objectives to transfer systems that work well in one condition to another space (e.g., robustness to unique noise conditions). We are also starting to see work that ties modalities together, such as learning techniques that map speech and language utterances to visual inputs (pictures, movies). This cross-modal integration will be an important direction for the future of our technical committee and a point of contact across technical committees.

ICASSP 2018 submissions in our speech and language area were up as well, with a 40% increase in submissions (to 634) during 2017, representing a 22% share of papers presented at the conference.

For those interested in participating in our community, various opportunities exist. SLTC members are elected every November for a three-year term, with nominations proposed well in advance (as early as summer). For details of our last election nomination, see [2]. As space on the technical committee is limited (we typically have a 3:1 candidate to electee ratio), we also offer an affiliate membership in our committee. Prospective applicants should refer to [3] for details on how to apply for such a position.

Authors

Michiel Bacchiani (michiel@google.com) received his engineer's degree from the Technical University of Eindhoven, The Netherlands, and his Ph.D. degree from Boston University, Massachusetts, both in electrical engineering. He is a senior staff research scientist at Google working on novel machine-learning algorithms for speech recognition and natural-language processing. He is a member of the IEEE Signal Processing Society and currently serves as chair of the IEEE Speech and Language Technical Committee. He is a Senior Member of the IEEE.

Eric Fosler-Lussier (fosler@cse.ohio-state.edu) received his B.A.S. degree in computer and cognitive science and his B.A. degree in linguistics from the University of Pennsylvania, Philadelphia, and his Ph.D. degree in computer science from the University of California, Berkeley. He is a professor of computer science and engineering at The Ohio State University, Columbus, with research interests in robust speech recognition and natural-language semantics, particularly for clinical informatics applications. He is a member of the IEEE Signal Processing Society and currently serves as vice chair of the IEEE Speech and Language Technical Committee. He is a Senior Member of the IEEE.

References

[1] IEEE Spoken Language Technology. (2018). [Online]. Available: http://www.slt2018.org/

[2] IEEE Signal Processing Society. (2017, Oct.). Nomination for new SLTC Members (2018–2020). [Online]. Available: https://signalprocessingsociety.org/get-involved/speech-and-language-processing/newsletter/nomination-new-sltc-members-2018-2020

[3] IEEE Signal Processing Society. (2018). Speech and Language Processing Technical Committee affiliate members. [Online]. Available: https://signalprocessingsociety.org/get-involved/speech-and-language-processing/affiliate-members

SP

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*.

IEEE SIGNAL PROCESSING MAGAZINE REPRESENTATIVE

Mark David, Director, Business Development — Media & Advertising, Phone: +1 732 465 6473, Fax: +1 732 981 1855, m.david@ieee.org

COMPANY	PAGE NUMBER	WEBSITE	PHONE
IEEE Information Theory Society	4	informationinsmallbits.com	
MathWorks	CVR 4	www.mathworks.com/wireless	
Southern University of Science and Technology	7	http://talent.sustc.edu.cn/en	+86 755 88018558
University of Vermont	6	http://www.uvmjobs.com/postings/31457	

Digital Object Identifier 10.1109/MSP.2017.2770717

Conference Planning for Professors

by Nuria González-Prelcic and Robert W. Heath, Jr.

Digital Object Identifier 10.1109/MSP.2018.2877374
Date of publication: 13 November 2018

SP

Erik Meijering and Arrate Muñoz-Barrutia

Spotlight on Bioimaging and Signal Processing

The Bio-Imaging and Signal Processing Technical Committee (BISP-TC) of the IEEE Signal Processing Society (SPS) promotes activities in the broad technical areas of computerized image and signal processing with a clear focus on applications in biology and medicine. Specific topics of interest include image reconstruction, compressed sensing, superresolution, image restoration, registration and segmentation, pattern recognition, object detection, localization, tracking, quantification and classification, machine learning, multimodal image and signal fusion, analytics, visualization, and statistical modeling. Application areas covered by the TC include biomedical imaging from nano to macroscale, encompassing all modalities of molecular imaging and microscopy, anatomical imaging, and functional imaging, as well as genomic signal processing, computational biology, and bioinformatics, with the ultimate overarching aim of enabling precision medicine.

Since its creation in 2004, the TC has served as the expert review and organization panel for the IEEE International Symposium on Biomedical Imaging (ISBI) as well as the bioimaging and signal processing tracks of the IEEE International Conference on Acoustics, Speech, and Signal Processing and the IEEE International Conference on Image Processing. Over the years, members of the TC have played leading roles in these flagship SPS meetings and organized numerous workshops, special sessions, and tutorials to deepen the understanding of theoretical concepts, broaden their range of applications, and highlight emerging hot topics in the field.

The TC maintains strong ties with other communities within SPS, the IEEE, and beyond. For example, multiple past and present members are active in the SPS Computational Imaging Special Interest Group, the Engineering in Medicine and Biology Society Biomedical Imaging and Image Processing TC, the cross-Society IEEE Life Sciences Technical Community, the Medical Image Computing and Computer-Assisted Intervention Society (MICCAI), the International Society for Optical Engineering, and the Society for Industrial and Applied Mathematics. Many TC members also serve on the editorial boards of SPS publications, such as *IEEE Transactions on Medical Imaging*, *IEEE Transactions on Computational Imaging*, *IEEE Transactions on Image Processing*, and *IEEE Transactions on Signal Processing*.

Computerized image and signal processing technologies have been key to biological research and medical diagnostics for at least a half-century. Revolutionary inventions such as magnetic resonance imaging (2003 Nobel Prize), superresolution microscopy (2014 Nobel Prize), cryo-electron microscopy (2017 Nobel Prize), and, in a sense, even the sequencing of the human genome (which will inevitably be awarded a Nobel Prize) all relied on or spurred the development of image and signal processing. As a consequence, we now live in an era that is flooded with data produced by imaging and sequencing devices and craves powerful solutions to extract maximum knowledge from them. Undoubtedly, machine-learning approaches, in particular, deep learning using artificial neural networks, will be an essential ingredient of these solutions and are already pervading the field, outperforming classical image and signal processing approaches in many tasks. But many open questions remain as to how they can be optimally designed and trained in a semi- or weakly supervised manner to get around the need for excessive human input in data annotation and to improve their transferability between applications.

Another question is how to improve the interpretability of the decisions made by deep neural networks, which is crucial, especially in areas such as medical diagnostics, where accountability is important and failures may have serious legal consequences. Notwithstanding open questions, deep learning seems perfectly suited for integrative data processing in emerging fields, e.g., imaging genomics, in which multiscale and multimodal imaging and genomic information are harnessed for the comprehensive and systematic diagnoses of complex diseases, such as cancer and dementia.

One activity that has stimulated the development of innovative solutions and has fostered collaborative efforts perhaps more than any is the organization

(continued on page 125)

Digital Object Identifier 10.1109/MSP.2018.2864772
Date of publication: 13 November 2018

420,000+ members in 160 countries.
Embrace the largest, global, technical community.

People Driving Technological Innovation.

ieee.org/membership **#IEEEmember**

IEEE

KNOWLEDGE COMMUNITY PROFESSIONAL DEVELOPMENT CAREER ADVANCEMENT

MATLAB SPEAKS WIRELESS DESIGN

You can simulate, prototype, and verify wireless systems right in MATLAB. Learn how today's MATLAB supports RF, LTE, WLAN and 5G development and SDR hardware.

mathworks.com/wireless

MathWorks®