Introduction to the Issue on Information-Theoretic Methods in Data Acquisition, Analysis, and Processing

The field of information theory – dating back to the seminal work by Claude E. Shannon in 1948 – is considered to be one of the landmark intellectual achievements of the 20th century, underpinning advances in compression and communication of data that underpins the information age. In particular, information-theoretic methods have been used to illuminate fundamental limits and gauge the effectiveness of algorithms for various problems in statistical decision theory, data communications, data compression, security, and networking. This has led to a series of technological breakthroughs in areas such as data storage, optical communications and networks, wireless communications and networks, and Internet technology.

Information theory – including its key ideas, methods and tools – has also played an important role in the general area of data science, most notably in information acquisition, information analysis and processing, statistics, probability, and learning. In particular, recent years have witnessed a renaissance in the use of information-theoretic methods to address problems beyond data compression, data communications, and networking, such as, e.g., compressive sensing, dictionary learning, supervised and unsupervised learning, reinforcement learning, graph mining, community detection, privacy, and fairness.

This special issue explores applications of information-theoretic methods to emerging data science problems. In particular, the papers cover a wide range of topics that can broadly be organized into four themes: (1) data acquisition, (2) data analysis and processing, (3) statistics and machine learning, and (4) privacy and fairness.

Overall, this special issue contains 20 papers out of 85 initial submissions, representing an acceptance ratio of just above 17.5%.

I. DATA ACQUISITION

Modern data science problems call for new data acquisition approaches beyond the classical Shannon-Nyquist sampling paradigm.

The paper “A Modulo-Based Architecture for Analog-to-Digital Conversion” by Ordentlich et al. develops a new approach to analog-to-digital conversion that aims at minimizing the number of bits per sample. This approach relies on the reduction of the dynamic range of the analog signal by performing a modulo operation on its amplitude followed by a quantization operation. The reconstruction procedure exploits the signal’s statistical structure in order to unwrap the modulo folding. This method is shown to approach the classical rate-distortion function and to exhibit various advantages over standard sigma-delta modulation.

The next paper, “Zero-Delay Rate Distortion via Filtering for Vector-Valued Gaussian Sources” by Stavrou et al. considers causal data acquisition and quantization of a Gaussian autoregressive source under mean-squared error. After quantization the source must be reconstructed immediately (after zero-delay). This problem is relevant to applications in distributed control where an unstable system must be stabilized over a reliable but rate-limited data channel. In this contribution the authors establish upper and lower bounds on the zero-delay rate distortion function. Lower bounds are established by making connections to the non-anticipatory rate-distortion function which can be characterized through a Kalman-filtering based scheme. Upper bounds are established through entropy coded dithered quantization. Illustrative examples are provided for both stable and unstable sources.

The paper “Performance Analysis of Approximate Message Passing for Distributed Compressed Sensing” by Hannak et al. derives a Bayesian approximate message passing method to recover sets of vectors that have identical (sparse) support but arbitrary correlation. Derivation of the state evolution is provided as well as a replica analysis of the Bayes optimal minimum mean squared error. While some aspects of this model have appeared in the literature before, the generalization to arbitrary correlation is the central contribution.

The following three papers explore another important topic arising in modern data acquisition problems. In particular, in contrast to standard task-agnostic data acquisition approaches where one aims to acquire a signal independently of the desired signal processing task, “Improved Target Acquisition Rates with Feedback Codes” by Lalitha et al., “Maximum entropy low-rank matrix recovery,” by Mak and Xie, and “Near-Optimal Noisy Group Testing via Separate Decoding of Items” by Scarlett and Cevher consider task-specific data acquisition.

Lalitha et al. consider the problem of determining an unknown target location via a sequence of measurements, where each measurement consists of simultaneously probing a group of locations. The proposed target detection approach is based on an equivalence drawn between the target acquisition problem and channel coding over a binary input additive white Gaussian noise (BAWGN) channel with state and feedback. This perspective then leads to a new two-stage adaptive target search strategy based on the sorted Posterior Matching channel coding strategy, that provides various adaptivity gains in different asymptotic regimes of interest.

Mak and Xie consider the problem of data acquisition for low-rank matrix recovery. In particular, the authors exploit the
maximum entropy principle to show that the measurement masks that maximize the entropy of the observations also maximize the information gain of the unknown low-rank matrix.

Scarlett and Cevher conduct an information-theoretic performance analysis of an algorithm by Mal'yutov and Mateev from 1980, for the noisy group testing problem. It is shown, under certain assumptions on the noise statistics, that the number of tests required for vanishing error probability is close to the information-theoretic limit provided that the number of defective items is small enough (sparse regime). A matching converse for separate decoding and independent and identically distributed randomized testing complements the achievability result.

II. DATA ANALYSIS AND PROCESSING

Modern data science problems also call for new approaches to data analysis and processing in view of the fact that one often desires to perform non-conventional tasks on increasingly complex data drawn from multi-dimensional manifolds, graphs, or more.

The paper “On the Fundamental Limit of Multipath Matching Pursuit” by Li et al. considers a new approach to the recovery of sparse signals based on the multipath matching pursuit (MMP) algorithm, representing an extension of the ubiquitous orthogonal matching pursuit (OMP) algorithm commonly used in compressed sensing. The authors derive optimal bounds on the restricted isometry constant for the sensing matrix so that MMP can successfully recover the underlying sparse signal from incomplete or noisy observations.

The paper “Universal Joint Image Clustering and Registration using Multivariate Information Measures” by Raman et al. explores the joint design of image clustering and image registration. The objective of clustering is to sort images into groups of similar images. The objective of registration is to align images. While well studied individually, the authors show that joint clustering and registration (within each cluster) of images is quite beneficial. Although classic approaches such as the maximum mutual information method suffice for pairwise registration, the authors introduce the max multi-information functional for contexts wherein many images must be clustered and aligned. The authors show that their method is asymptotically universal and asymptotically optimal.

The paper “Community Detection with Side Information: Exact Recovery under the Stochastic Block Model” by Saad and Nosratinia considers a community detection problem within the so-called stochastic block model framework where one aims at making inferences about node labels on a graph given a certain realization of the edges of the graph. The authors imbue the work with an interesting twist by assuming that one has access to additional, non-graphical side information that manifests itself in the form of noisy labels with error probability \( \alpha \). Fundamental limits on the scaling of \( \alpha \) with the problem size are established so that the phase transition of recovery of the node labels can be improved with respect to problems where no side information is available.

The paper “Hypergraph Spectral Clustering in the Weighted Stochastic Block Model” by Ahn et al. considers generalizations of spectral clustering algorithms from the standard graph clustering setting to the hypergraph one. In particular, by considering a weighted stochastic block model, where nodes represent objects and hyper-edges weights represent multi-way measures between objects, the authors propose two poly-time algorithms, namely Hypergraph Spectral Clustering and Hypergraph Spectral Clustering with Local Refinement, to recover the hidden partition from the weighted hypergraph. The authors also analyze the performance of these algorithms including their consistency, showing that certain algorithms achieve information-theoretic limits for a specific practically-relevant model.

Two more papers in this section focus on ranking problems: “Lower Bounds on the Bayes Risk of the Bayesian BTL Model with Applications to Comparison Graphs” by Alsan et al., considers the problem of aggregating pairwise comparisons to obtain a consensus ranking order over a collection of objects, where the pairwise comparisons are described using the Bradley-Terry-Luce (BTL) model. The authors derive information-theoretic lower bounds on the Bayes risk of estimators for norm-based distortion functions. Moreover, in the special case of the mean-squared error, the information-theoretic bounds are compared to the well-known Bayesian Cramér-Rao lower bound. Finally, the results are related to graph theory by analyzing the effect of various comparison graph structures on the obtained lower bounds.

In “Top-K Rank Aggregation from M-wise Comparisons” by Jang et al., the authors consider a generalization of the top-K ranking problem where one is asked to rank a small subset of objects \( M \) out of the total number of objects. By capitalizing on the so-called Plackett-Luce model involving comparisons among \( M \) objects, the authors propose efficient algorithms and establish fundamental limits on the recovery performance.

III. STATISTICS AND MACHINE LEARNING

Several papers in this special issue are devoted to exploring the use of information-theoretic methods in emerging problems in statistics and machine learning, including (1) hypothesis testing, (2) classification, (3) causal inference, and (4) a variety of learning problems, such as dictionary learning and multi-task learning.

In “Nonparametric Composite Hypothesis Testing in an Asymptotic Regime,” Li et al. study nonparametric, composite hypothesis testing problems for arbitrary unknown distributions in various asymptotic regimes, including the case where the number of hypotheses approaches infinity. In particular, the notion of discrimination capacity is introduced, capturing the largest exponential growth rate of the number of hypotheses relative to the sample size so that there exists a test of asymptotically zero probability of error. Furthermore, the achievable discrimination rates of various tests including maximum mean discrepancy (MMD) and Kolmogorov-Smirnov (KS) distances based tests, are characterized, and an upper bound on the discrimination capacity based on Fano’s inequality is provided.

In “Classification and Representation via Separable Subspaces: Performance Limits and Algorithms,” Jindal and Nokleby study classification problems for Kronecker-structured subspace models in asymptotic regimes. Various notions are introduced such as the classification capacity; in particular, the classification capacity defines the maximum rate at which the number of classes can grow as the signal dimension goes to infinity. Finally, a fast algorithm for Kronecker-structured learning of discriminative dictionaries is proposed, leading to both compact representations and good classification performance.

The paper “Bayesian Nonparametric Causal Inference: Information Rates and Learning Algorithms” by Alaa and van der Schaar, proposes an information-theoretic approach to the problem of estimating the causal effect of a treatment on
individual subjects from observational data. The Kullback-Leibler distance between the estimated and the true distributions is employed as a measure of accuracy with Fano’s method allowing to establish a fundamental limit on the information rate achievable by Bayesian estimators. The paper goes on to devise an information-optimal Bayesian causal inference algorithm, which embeds the potential outcomes in a RKHS.

The paper “Identifiability of Kronecker-structured Dictionaries for Tensor Data” by Shakeri et al. derives sufficient conditions for local recovery of coordinate dictionaries comprising a Kronecker-structured dictionary that is used for representing $K^{th}$-order tensor data. Information-theoretic methods are used to establish sample complexity results, guaranteeing dictionary recovery is successful in the sense that the estimation error does not exceed a small threshold $\varepsilon$.

In “Compression-Based Regularization with an Application to Multi-Task Learning” by Vera et al., an information-theoretic perspective of (supervised) multi-task learning (MTL) is provided, by drawing connections between the MTL problem and the information bottleneck with side information problem. Furthermore, an iterative Arimoto-Blahut like algorithm for solving the non-convex information bottleneck with side information problem is proposed and analyzed; this algorithm leads to an optimal trade-off between the log-loss (average risk) and the information rate of the features (statistical model complexity). The paper also showcases applications of the approach to hierarchical text categorization and distributional word clusters.

Finally, motivated by applications in power monitoring systems, in the paper “Optimal detection and error exponents for hidden semi-Markov models” Bajovic et al. consider a binary hypothesis testing problem. The null hypothesis corresponds to pure noise. The second hypothesis corresponds to a signal in additive noise that alternates between two states of stochastically varying duration. The authors devise a tractable recursive expression for the likelihood function. From this a lower bound on the error exponent is derived. Simulation results demonstrate that the lower bound can be tight.

IV. PRIVACY AND FAIRNESS

In recent years, privacy and fairness issues have come to the forefront in several data science problems. In particular, it is becoming increasingly crucial to develop new approaches to data acquisition, analysis, and processing that preserve privacy and maintain fairness amongst individuals releasing the data for further analyses.

This special issue contains two papers that explore these emerging topics. In “Efficiently Computing Messages that Reveal Selected Inferences While Protecting Others” by Johnson et al., a study of mechanisms for data release allowing a statistician to infer some allowed attributes but at the same time preventing the statistician from inferring private ones are conducted. In particular, by relying on the privacy funnel formulation, the authors consider Pareto-optimal probabilistic mechanisms striking a balance between utility and privacy.

In “Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis” by Calmon et al., a study of data release mechanisms for discrimination prevention is also carried out. The authors rely on an optimization formulation relying on information-theoretic quantities to define such data release mechanisms, showing discrimination prevention in real-world datasets such as the prison recidivism dataset and the UCI Adult dataset.

We would like to end by thanking all those that helped make this special issue possible, in particular Shri Narayanan, whose support, encouragement, and guidance has been instrumental. We are grateful for the support of Allison Fleisher, the journal coordinator, and for her effective assistance throughout the review process. Last but not least, we thank all the reviewers for their volunteer work and valuable comments.

We hope that you will find the articles in this special issue interesting, stimulating, and useful for your own research, and that this special issue will continue to inspire further work on the interface of information theory, signal processing, and data science.

M. Rodrigues, Lead Guest Editor
University College London
London WC1E 7JE, U.K.

H. Bolcskei, Guest Editor
ETH Zurich
Zürich CH-8092, Switzerland

S. Draper, Guest Editor
University of Toronto
Toronto, ON M5S 3G4, Canada

Y. Eldar, Guest Editor
Technion—Israel Institute of Technology
Haifa 32000, Israel

V. Tan, Guest Editor
National University of Singapore
Singapore 117576

Miguel Rodrigues is a Reader in information theory and processing with the Department of Electronic and Electrical Engineering, University College London, London, U.K., and a Faculty Fellow with the Turing Institute, London, U.K. He was previously with the Department of Computer Science, University of Porto, Portugal, and the ranks from Assistant to Associate Professor. He also held research positions with Princeton University, Cambridge University, and Duke University. His research interests—which lie in the general areas of information theory and processing—have led to nearly 200 publications in leading journals and conferences in the field, including the prestigious IEEE Communications and Information Theory Societies Joint Paper Award 2011. He is a coauthor of a book on Information-Theoretic Methods in Data Science (Cambridge University Press).
Helmut Bolcskei is a Professor in electrical engineering and mathematics with ETH Zurich, Zurich, Switzerland. His research interests include information theory, mathematical signal processing, machine learning, and statistics. He was a recipient of the 2001 IEEE Signal Processing Society Young Author Best Paper Award, the 2006 IEEE Communications Society Leonard G. Abraham Best Paper Award, the 2010 Vodafone Innovations Award, the ETH “Golden Owl” Teaching Award. He is a 2011 EURASIP Fellow. He was a Distinguished Lecturer (2013–2014) of the IEEE Information Theory Society, an Erwin Schrödinger Fellow (1999–2001) of the Austrian National Science Foundation, was included in the 2014 Thomson Reuters List of Highly Cited Researchers in Computer Science, and is the 2016 Padovani Lecturer of the IEEE Information Theory Society. He was an Editor-in-Chief for the IEEE TRANSACTIONS ON INFORMATION THEORY during the period 2010–2013, the TPC Co-Chair of the 2008 IEEE International Symposium on Information Theory and the 2016 IEEE Information Theory Workshop, and served on the Board of Governors of the IEEE Information Theory Society. He has been a delegate of the president of ETH Zurich for faculty appointments since 2008.

Stark Draper received the B.S. and B.A. degrees in electrical engineering and history from Stanford University, Stanford, CA, USA, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA. He is a Professor in electrical and computer engineering with the University of Toronto, Toronto, ON, Canada. As a research scientist, he has worked with the Mitsubishi Electric Research Labs, Disney’s Boston Research Lab, Arraycomm, Inc., the C. S. Draper Laboratory, and Ktaadn, Inc. His research interests include information theory, optimization, machine learning, coding theory, and application of these fields to cloud computing and computer architecture. He was a recipient of the NSF CAREER Award, the MERL President’s Award, Stanford’s Frederick E. Terman Engineering Scholastic Award, and a U.S. State Department Fulbright Fellowship. He currently chairs the “Machine Intelligence” major at the University of Toronto and serves on the IEEE Information Theory Society Board of Governors.

Yonina Eldar is a Professor with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel, where she holds the Edwards Chair in engineering. She is also an Adjunct Professor with Duke University, Durham, NC, USA, a Research Affiliate with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, and was Visiting Professor with Stanford University, Stanford, CA, USA. She is a member of the Israel Academy of Sciences and Humanities and a Fellow of 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 for the Foundations and Trends in Signal Processing and serves the IEEE on several technical and award committees.

Vincent Tan received the B.A. and M.Eng. degrees in electrical and information sciences from Cambridge University, Cambridge, U.K., in 2005, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2011. He is currently an Associate Professor with the Department of Electrical and Computer Engineering and the Department of Mathematics, National University of Singapore, Singapore. His research interests include information theory, machine learning, and statistical signal processing. He is a currently the IEEE Information Theory Society Distinguished Lecturer and an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.