Information-Theoretic Methods in Data Science

Learn about the state-of-the-art at the interface between information theory and data science with this first unified treatment of the subject. Written by leading experts in a clear, tutorial style, and using consistent notation and definitions throughout, it shows how information-theoretic methods are being used in data acquisition, data representation, data analysis, and statistics and machine learning.

Coverage is broad, with chapters on signal data acquisition, data compression, compressive sensing, data communication, representation learning, emerging topics in statistics, and much more. Each chapter includes a topic overview, definition of the key problems, emerging and open problems, and an extensive reference list, allowing readers to develop in-depth knowledge and understanding.

Providing a thorough survey of the current research area and cutting-edge trends, this is essential reading for graduate students and researchers working in information theory, signal processing, machine learning, and statistics.

Miguel R. D. Rodrigues is Professor of Information Theory and Processing in the Department of Electronic and Electrical Engineering, University College London, and a Turing Fellow at the Alan Turing Institute, London.

Yonina C. Eldar is a Professor in the Faculty of Mathematics and Computer Science at the Weizmann Institute of Science, a Fellow of the IEEE and Eurasip, and a member of the Israel Academy of Sciences and Humanities. She is the author of *Sampling Theory* (Cambridge, 2015), and co-editor of *Compressed Sensing in Radar Signal Processing* (Cambridge, 2019), *Compressed Sensing* (Cambridge, 2012), and *Convex Optimization in Signal Processing and Communications* (Cambridge, 2009).

"How should we model the opportunities created by access to data and to computing, and how robust and resilient are these models? What information is needed for analysis, how is it obtained and applied? If you are looking for answers to these questions, then there is no better place to start than this book."

Robert Calderbank, *Duke University*

"This pioneering book elucidates the emerging cross-disciplinary area of information-theoretic data science. The book's contributors are leading information theorists who compellingly explain the close connections between the fields of information theory and data science. Readers will learn about theoretical foundations for designing data collection and analysis systems, from data acquisition to data compression, computation, and learning."

Al Hero, *University of Michigan*

Information-Theoretic Methods in Data Science

Edited by
MIGUEL R. D. RODRIGUES
University College London

YONINA C. ELDAR
Weizmann Institute of Science

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108427135
DOI: 10.1017/9781108616799

First published 2021

Printed in the United Kingdom by TJ International Ltd, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-42713-5 Hardback

To my husband Shalomi and children Yonatan, Moriah, Tal, Noa, and Roei for their boundless love and for filling my life with endless happiness

YE

To my wife Eduarda and children Isabel and Diana for their unconditional love, encouragement, and support

MR

CAMBRIDGE

Cambridge University Press
978-1-108-42713-5 — Information-Theoretic Methods in Data Science
Edited by Miguel R. D. Rodrigues , Yonina C. Eldar
Frontmatter
More Information

Contents

Preface

Since its introduction in 1948, the field of information theory has proved instrumental in the analysis of problems pertaining to compressing, storing, and transmitting data. For example, information theory has allowed analysis of the fundamental limits of data communication and compression, and has shed light on practical communication system design for decades. 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 compressive sensing, data acquisition, data analysis, machine learning, graph mining, community detection, privacy, and fairness. In this book, we explore a broad set of problems on the interface of signal processing, machine learning, learning theory, and statistics where tools and methodologies originating from information theory can provide similar benefits. The role of information theory at this interface has indeed been recognized for decades. A prominent example is the use of information-theoretic quantities such as mutual information, metric entropy and capacity in establishing minimax rates of estimation back in the 1980s. Here we intend to explore modern applications at this interface that are shaping data science in the twenty-first century.

There are of course some notable differences between standard information-theoretic tools and signal-processing or data analysis methods. Globally speaking, information theory tends to focus on asymptotic limits, using large blocklengths, and assumes the data is represented by a finite number of bits and viewed through a noisy channel. The standard results are not concerned with complexity but focus more on fundamental limits characterized via achievability and converse results. On the other hand, some signal-processing techniques, such as sampling theory, are focused on discrete-time representations but do not necessarily assume the data is quantized or that there is noise in the system. Signal processing is often concerned with concrete methods that are optimal, namely, achieve the developed limits, and have bounded complexity. It is natural therefore to combine these tools to address a broader set of problems and analysis which allows for quantization, noise, finite samples, and complexity analysis.

This book is aimed at providing a survey of recent applications of information-theoretic methods to emerging data-science problems. The potential reader of this book could be a researcher in the areas of information theory, signal processing, machine learning, statistics, applied mathematics, computer science or a related research area, or a graduate student seeking to learn about information theory and data science and to scope out open problems at this interface. The particular design of this volume ensures that it can serve as both a state-of-the-art reference for researchers and a textbook for students.

The book contains 16 diverse chapters written by recognized leading experts worldwide, covering a large variety of topics that lie on the interface of signal processing, data science, and information theory. The book begins with an introduction to information theory which serves as a background for the remaining chapters, and also sets the notation to be used throughout the book. The following chapters are then organized into four categories: data acquisition (Chapters 2–4), data representation and analysis (Chapters 5–9), information theory and machine learning (Chapters 10 and 11), and information theory, statistics, and compression (Chapters 12–15). The last chapter, Chapter 16, connects several of the book's themes via a survey of Fano's inequality in a diverse range of data-science problems. The chapters are self-contained, covering the most recent research results in the respective topics, and can all be treated independently of each other. A brief summary of each chapter is given next.

Chapter 1 by Rodrigues, Draper, Bajwa, and Eldar provides an introduction to information theory concepts and serves two purposes: It provides background on classical information theory, and presents a taster of modern information theory applied to emerging data-science problems.

Chapter 2 by Kipnis, Eldar, and Goldsmith extends the notion of rate-distortion theory to continuous-time inputs deriving bounds that characterize the minimal distortion that can be achieved in representing a continuous-time signal by a series of bits when the sampler is constrained to a given sampling rate. For an arbitrary stochastic input and given a total bitrate budget, the authors consider the lowest sampling rate required to sample the signal such that reconstruction of the signal from a bit-constrained representation of its samples results in minimal distortion. It turns out that often the signal can be sampled at sub-Nyquist rates without increasing the distortion.

Chapter 3 by Jalali and Poor discusses the interplay between compressed sensing and compression codes. In particular, the authors consider the use of compression codes to design compressed sensing recovery algorithms. This allows the expansion of the class of structures used by compressed sensing algorithms to those used by data compression codes, which is a much richer class of inputs and relies on decades of developments in the field of compression.

Chapter 4 by Pilanci develops information-theoretical lower bounds on sketching for solving large statistical estimation and optimization problems. The term sketching is used for randomized methods that aim to reduce data dimensionality in computationally intensive tasks for gains in space, time, and communication complexity. These bounds allow one to obtain interesting trade-offs between computation and accuracy and shed light on a variety of existing methods.

Chapter 5 by Shakeri, Sarwate, and Bajwa treats the problem of dictionary learning, which is a powerful signal-processing approach for data-driven extraction of features from data. The chapter summarizes theoretical aspects of dictionary learning for vector- and tensor-valued data and explores lower and upper bounds on the sample complexity of dictionary learning which are derived using information-theoretic tools. The dependence of sample complexity on various parameters of the dictionary learning problem is highlighted along with the potential advantages of taking the structure of tensor data into consideration in representation learning.

Chapter 6 by Riegler and Bölcskei presents an overview of uncertainty relations for sparse signal recovery starting from the work of Donoho and Stark. These relations are then extended to richer data structures and bases, which leads to the recently discovered set-theoretic uncertainty relations in terms of Minkowski dimension. The chapter also explores the connection between uncertainty relations and the "large sieve," a family of inequalities developed in analytic number theory. It is finally shown how uncertainty relations allow one to establish fundamental limits of practical signal recovery problems such as inpainting, declipping, super-resolution, and denoising of signals.

Chapter 7 by Reeves and Pfister examines high-dimensional inference problems through the lens of information theory. The chapter focuses on the standard linear model for which the performance of optimal inference is studied using the replica method from statistical physics. The chapter presents a tutorial of these techniques and presents a new proof demonstrating their optimality in certain settings.

Chapter 8 by Shah discusses the question of learning distributions over permutations of a given set of choices based on partial observations. This is central to capturing choice in a variety of contexts such as understanding preferences of consumers over a collection of products based on purchasing and browsing data in the setting of retail and e-commerce. The chapter focuses on the learning task from marginal distributions of two types, namely, first-order marginals and pair-wise comparisons, and provides a comprehensive review of results in this area.

Chapter 9 by Raman and Varshney studies universal clustering, namely, clustering without prior access to the statistical properties of the data. The chapter formalizes the problem in information theory terms, focusing on two main subclasses of clustering that are based on distance and dependence. A review of well-established clustering algorithms, their statistical consistency, and their computational and sample complexities is provided using fundamental information-theoretic principles.

Chapter 10 by Raginsky, Rakhlin, and Xu introduces information-theoretic measures of algorithmic stability and uses them to upper-bound the generalization bias of learning algorithms. The notion of stability implies that its output does not depend too much on any individual training example and therefore these results shed light on the generalization ability of modern learning techniques.

Chapter 11 by Piantanida and Vega introduces the information bottleneck principle and explores its use in representation learning, namely, in the development of computational algorithms that learn the different explanatory factors of variation behind high-dimensional data. Using these tools, the authors obtain an upper bound on the generalization gap corresponding to the cross-entropy risk. This result provides an interesting connection between mutual information and generalization, and helps to explain why noise injection during training can improve the generalization ability of encoder models.

Chapter 12 by Ding, Yang, and Tarokh discusses fundamental limits of inference and prediction based on model selection principles from modern data analysis. Using information-theoretic tools the authors analyze several state-of-the-art model selection techniques and introduce two recent advances in model selection approaches, one concerning a new information criterion and the other concerning modeling-procedure selection.

Chapter 13 by Wu and Xu provides an exposition on some of the methods for determining the information-theoretical as well as computational limits for high-dimensional statistical problems with a planted structure. Planted structures refer to a ground truth structure (often of a combinatorial nature) which one is trying to discover in the presence of random noise. In particular, the authors discuss first- and second-moment methods for analyzing the maximum likelihood estimator, information-theoretic methods for proving impossibility results using mutual information and rate-distortion theory, and techniques originating from statistical physics. To investigate computational limits, they describe randomized polynomial-time reduction schemes that approximately map planted-clique problems to the problem of interest in total variation distance.

Chapter 14 by Zhao and Lai considers information-theoretic models for distributed statistical inference problems with compressed data. The authors review several research directions and challenges related to applying these models to various statistical learning problems. In these applications, data are distributed in multiple terminals, which can communicate with each other via limited-capacity channels. Information-theoretic tools are used to characterize the fundamental limits of the classical statistical inference problems using compressed data directly.

Chapter 15 by Feizi and Médard treats different aspects of the network functional compression problem. The goal is to compress a source of random variables for the purpose of computing a deterministic function at the receiver where the sources and receivers are nodes in a network. Traditional data compression schemes are special cases of functional compression, in which the desired function is the identity function. It is shown that for certain classes of functions considerable compression is possible in this setting.

Chapter 16 by Scarlett and Cevher provides a survey of Fano's inequality and its use in various statistical estimation problems. In particular, the chapter overviews the use of Fano's inequality for establishing impossibility results, namely, conditions under which a certain goal cannot be achieved by any estimation algorithm. The authors present several general-purpose tools and analysis techniques, and provide representative examples covering group testing, graphical model selection, sparse linear regression, density estimation, and convex optimization.

Within the chapters, the authors point to various open research directions at the interface of information theory, data acquisition, data analysis, machine learning, and statistics that will certainly see increasing attention in the years to come.

We would like to end by thanking all the authors for their contributions to this book and for their hard work in presenting the material in a unified and accessible fashion.

Notation

z	scalar (or value of random variable Z)
Z	random variable
$\mathbf{z}$	vector (or value of random vector $\mathbf{Z}$)
$\mathbf{Z}$	matrix (or random vector)
$\mathbf{z}_i$	ith entry of vector $\mathbf{z}$
$\mathbf{Z}_{i,j}$	(i, j)th entry of matrix $\mathbf{Z}$
$Z^n = (Z_1, \ldots, Z_n)$	sequence of n random variables
$z^n = (z_1, \ldots, z_n)$	value of sequence of n random variables Z^n
$Z_i^j = \left(Z_i, \ldots, Z_j\right)$	sequence of $j - i + 1$ random variables
$z_i^j = \left(z_i, \ldots, z_j\right)$	value of sequence of $j - i + 1$ random variables Z_i^j
$\lVert \cdot \rVert_p$	p-norm
$(\cdot)^{\mathrm{T}}$	transpose operator
$(\cdot)^*$	conjugate Hermitian operator
$(\cdot)^\dagger$	pseudo-inverse of the matrix argument
$\mathrm{tr}(\cdot)$	trace of the square matrix argument
$\det(\cdot)$	determinant of the square matrix argument
$\mathrm{rank}(\cdot)$	rank of the matrix argument
$\mathrm{range}(\cdot)$	range span of the column vectors of the matrix argument
$\lambda_{\max}(\cdot)$	maximum eigenvalue of the square matrix argument
$\lambda_{\min}(\cdot)$	minimum eigenvalue of the square matrix argument
$\lambda_i(\cdot)$	ith largest eigenvalue of the square matrix argument
$\mathbf{I}$	identity matrix (its size is determined from the context)
$\mathbf{0}$	matrix with zero entries (its size is determined from the context)
$\mathcal{T}$	standard notation for sets
$\lvert\mathcal{T}\rvert$	cardinality of set $\mathcal{T}$
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{R}^n$	set of n-dimensional vectors of real numbers
$\mathbb{C}^n$	set of n-dimensional vectors of complex numbers
j	imaginary unit
$\mathrm{Re}(x)$	real part of the complex number x
$\mathrm{Im}(x)$	imaginary part of the complex number x
$\lvert x\rvert$	modulus of the complex number x
$\arg(x)$	argument of the complex number x
$\mathbb{E}[\cdot]$	statistical expectation
$\mathbb{P}[\cdot]$	probability measure

$H(\cdot)$	entropy
$H(\cdot\|\cdot)$	conditional entropy
$h(\cdot)$	differential entropy
$h(\cdot\|\cdot)$	conditional differential entropy
$D(\cdot\|\|\cdot)$	relative entropy
$I(\cdot;\cdot)$	mutual information
$I(\cdot;\cdot\|\cdot)$	conditional mutual information
$\mathcal{N}\left(\mu,\sigma^2\right)$	scalar Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}(\mu,\mathbf{\Sigma})$	multivariate Gaussian distribution with mean μ and covariance matrix $\mathbf{\Sigma}$

Contributors

Waheed U. Bajwa
Department of Electrical and Computer Engineering
Rutgers University

Helmut Bölcskei
Department of Information Technology and Electrical Engineering
Department of Mathematics
ETH Zürich

Volkan Cevher
Laboratory for Information and Inference Systems
Institute of Electrical Engineering
School of Engineering, EPFL

Jie Ding
School of Statistics
University of Minnesota

Stark C. Draper
Department of Electrical and Computer Engineering
University of Toronto

Yonina C. Eldar
Faculty of Mathematics and Computer Science
Weizmann Institute of Science

Soheil Feizi
Department of Computer Science
University of Maryland, College Park

Andrea J. Goldsmith
Department of Electrical Engineering
Stanford University

Shirin Jalali
Nokia Bell Labs

Alon Kipnis
Department of Statistics
Stanford University

Lifeng Lai
Department of Electrical and Computer Engineering
University of California, Davis

Muriel Médard
Department of Electrical Engineering and Computer Science
Massachussets Institute of Technology

Henry D. Pfister
Department of Electrical and Computer Engineering
Duke University

Pablo Piantanida
Laboratoire des Signaux et Systèmes
Université Paris Saclay
CNRS-CentraleSupélec

Montreal Institute for Learning Algorithms (Mila), Université de Montréal

Mert Pilanci
Department of Electrical Engineering
Stanford University

H. Vincent Poor
Department of Electrical Engineering
Princeton University

Maxim Raginsky
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Alexander Rakhlin
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology

Ravi Kiran Raman
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Galen Reeves
Department of Electrical and Computer Engineering
Department of Statistical Science
Duke University

Leonardo Rey Vega
Department of Electrical Engineering
School of Engineering
Universidad de Buenos Aires
CSC-CONICET

Erwin Riegler
Department of Information Technology and Electrical Engineering
ETH Zürich

Miguel R. D. Rodrigues
Department of Electronic and Electrical Engineering
University College London

Anand D. Sarwate
Department of Electrical and Computer Engineering
Rutgers University

Jonathan Scarlett
Department of Computer Science and Department of Mathematics
National University of Singapore

Devavrat Shah
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Zahra Shakeri
Electronic Arts

Vahid Tarokh
Department of Electrical and Computer Engineering
Duke University

Lav R. Varshney
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Yihong Wu
Department of Statistics and Data Science
Yale University

Aolin Xu
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Jiaming Xu
Fuqua School of Business
Duke University

Yuhong Yang
School of Statistics
University of Minnesota

Wenwen Zhao
Department of Electrical and Computer Engineering
University of California, Davis