## High-Dimensional Data Analysis with Low-Dimensional Models

Connecting theory with practice, this systematic and rigorous introduction covers the fundamental principles, algorithms, and applications of key mathematical models for high-dimensional data analysis. Comprehensive in its approach, it provides unified coverage of many different low-dimensional models and analytical techniques, including sparse and low-rank models, and both convex and nonconvex formulations. Readers will learn how to develop efficient and scalable algorithms for solving real-world problems, supported by numerous examples and exercises throughout, and how to use the computational tools learnt in several application contexts. Applications presented include scientific imaging, communication, face recognition, three-dimensional vision, and deep networks for classification. With code available online, this is an ideal text for graduate students in electrical engineering, computer science, and data science, as well as for those taking courses on sparsity, low-dimensional structures, and high-dimensional data.

**John Wright** is an Associate Professor in the Electrical Engineering Department at Columbia University. He is also affiliated with Columbia's Department of Applied Physics and Applied Mathematics and the Data Science Institute.

**Yi Ma** is a Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. He is a Fellow of the IEEE, ACM, and SIAM.

# High-Dimensional Data Analysis with Low-Dimensional Models

## Principles, Computation, and Applications

JOHN WRIGHT
*Columbia University, New York*

YI MA
*University of California, Berkeley*

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, 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

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108489737

DOI: 10.1017/9781108779302

First published 2022
3rd printing 2022

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

ISBN 978-1-108-48973-7 Hardback

Additional resources for this publication at www.cambridge.org/wright-ma.

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

**To Mary, Isabella, and Mingshu (J.W.)**

**To Henry, Barron, and Diana,**
**and in memory of my father (Y.M.)**

Cambridge University Press & Assessment
978-1-108-48973-7 — High-Dimensional Data Analysis with Low-Dimensional Models
John Wright, Yi Ma
Frontmatter
More Information

# Contents

Cambridge University Press & Assessment
978-1-108-48973-7 — High-Dimensional Data Analysis with Low-Dimensional Models
John Wright, Yi Ma
Frontmatter
More Information

Cambridge University Press & Assessment
978-1-108-48973-7 — High-Dimensional Data Analysis with Low-Dimensional Models
John Wright, Yi Ma
Frontmatter
More Information

# Foreword

I recall a moment, perhaps ten or fifteen years ago, of prodigious scientific activity. To give our reader a sense of this blessed time, consider a series of regular scientific workshops, each involving at most forty participants. Despite the small size and almost intimate nature of these workshops, they brought together an energized and enthusiastic mix of people from an array of disciplines, including mathematics, computer science, engineering, and the life sciences. What a privilege to be in a room with mathematicians such as Terence Tao and Roman Vershynin and learn about high-dimensional geometry; with applied mathematicians and engineers such as David Donoho, Joel Tropp, Thomas Ströhmer, Michael Elad, and Freddy Bruckstein and learn about the power of algorithms; with statistical physicists such as Andrea Montanari and learn about phase transitions in large stochastic systems. What a privilege to learn about fast numerical methods for large-scale optimization from computer scientists such as Stephen Wright and Stanley Osher. What a privilege to learn about compressive optical systems from David Brady, and Richard Baraniuk and Kevin Kelly (of single-pixel camera fame); about compressive analog-to-digital conversion and wideband spectrum sensing from Dennis Healy, Yonina Eldar, and Azita Emami Neyestanak; about breakthroughs in computer vision from Yi Ma, John Wright, and René Vidal; and about dramatically faster scan times in magnetic resonance imaging from Michael Lustig and Leon Axel. Bringing all these people – and others I regretfully cannot name for lack of space – together, with their different perspectives and interests, sparked spirited discussions. Excitement was in the air and progress quickly followed.

Yi Ma and John Wright were frequent participants to these workshops and their book magically captures their spirit and richness. It exposes readers to (1) a variety of real-world applications including medical and scientific imaging, computer vision, wideband spectrum sensing, and so on, (2) the mathematical ideas powering algorithms in use in these fields, and (3) the algorithmic ideas needed to implement them. Let me illustrate with an example. On the one hand, this is a book in which we learn about the principles of magnetic resonance (MR) imaging. There is a chapter in which we learn how an MR scan excites the nucleus of atoms by means of a magnetic field. These nuclei have a magnetic spin, and will respond to this excitation, and it is precisely this response that gets recorded. As for other imaging modalities, such as computed tomography, there is a mathematical transformation that relates the object we wish to infer and the data we collect. In this case, after performing a few approximations, this mathematical transformation is given by the Fourier transform.

On the other hand, this is a book in which we learn that most of the mass of a high-dimensional sphere is concentrated not just around the equator – this is already sufficiently surprising – but around any equator! Or that the intersection between two identical high-dimensional cubes, one being randomly oriented vis-à-vis the other, is essentially a sphere! These are fascinating subjects, but what is the connection? There is one, of course, and explaining it is the most wonderful strength of the book. In a nutshell, ideas and tools from probability theory, high-dimensional geometry, and convex analysis inform concrete applied problems and explain why algorithms actually work. Returning to our MR imaging problem, we learn how to leverage mathematical models of sparsity to recover exquisite images of body tissues from what appear to be far too few data points. Such a feat allows us to scan patients ten times faster today.

Through three fairly distinct parts – roughly, theory, computations, and applications – the book proposes a scientific vision concerned with the development of insightful mathematics to create models for data, to create processing algorithms, and to ultimately inspire real concrete improvements; for instance, in human health as in the example above.

The first part of the book explores data models around two main themes, namely, sparsity and low-rankedness. Sparsity expresses the idea that most of the entries of an $n$-dimensional signal vanish or nearly vanish so that the information can be effectively summarized using fewer than $n$ data bits. Low-rankedness expresses the idea that the columns of a data matrix ‘live’ near a linear subspace of lower dimension, thereby also suggesting the possibility of an effective summary. We then find out how to use these data models to create data processing algorithms, for instance, to find solutions of underdetermined systems of linear equations. The emphasis is on algorithms formulated as solutions to well-formulated convex optimization problems. That said, we are also introduced to nonconvex methods in Chapter 7 to learn effective empirical representations from data in which signals exhibit enhanced sparsity. All along, the authors use their rich experiences to communicate insights and to explain why some things work while others do not.

The second part reviews effective methods for solving optimization problems – convex or not – at scale; that is, involving possibly millions of decision variables and a possibly equally large number of constraints. This is an area that has seen tremendous progress in the last fifteen years and the book provides readers with a valuable point of entry to the key ideas and vast literature.

The last part is a deep dive into applications. In addition to the imaging challenges I already mentioned, we find a chapter on wireless radio communication, where we see how ideas from sparse signal processing and compressed sensing allow cognitive radios to efficiently identify the available spectrum. We also find three chapters on crucial problems in computer vision, a field in which the authors have brought and developed formidable tools, enabling major advances and opening new perspectives. Exposition starts with a special contribution, which also exploits ideas from compressed sensing, to the crucial problem of face recognition in the presence of occlusions and other nonidealities. (I recall an exciting *Wired* article about this work when it came out.) The book then introduces methods for inferring 3D structure from a series of 2D photographs, and to identify structured textures from a single photograph; solving the latter problem is often the starting point to recover the appearance, pose, and shape of multiple objects in a scene. Finally, at the time of this writing, deep learning (DL) is all the rage. The book contains an epilogue which establishes connections between all the better understood data models reviewed in the book and DL: the one hundred million dollar question is whether they will shed significant insights on deep learning and influence or improve its practice.

Who would enjoy this book? First and foremost, students in mathematics, applied mathematics, statistics, computer science, electrical engineering, and related disciplines. Students will learn a lot from reading this book because it is so much more than a text about a tool being applied with minor variations. They will learn about mathematical reasoning, they will learn about data models and about connecting those to reality, and they will learn about algorithms. The book also contains computer scripts so that we can see ideas in action and carefully crafted exercises making it perfect for upper-level undergraduate or graduate-level instruction. The breadth and depth make this a reference for anyone interested in the mathematical foundations of data science. I also believe that members of the applied mathematical sciences community at large would enjoy this book. They will be reminded of the power of mathematical reasoning and of the all-around positive impact it can have.

Emmanuel Candès
Stanford, California
December 2020

# Preface

"*The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently.*"

– David Donoho, *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality, 2000*

## *The Era of Big Data*

In the past two decades, our world has entered the age of "Big Data." The information technology industry is now facing the challenge, and opportunity, of processing and analyzing massive amounts of data on a daily basis. The size and the dimension of the data have reached an unprecedented scale and are still increasing at an unprecedented rate.

For instance, on the technological side, the resolution of consumer digital cameras has increased nearly ten-fold in the past decade or so. Each day, over 300 million photos are uploaded to Facebook;[1] 300 hours of videos are posted on YouTube every minute; and over 20 million entertaining short videos are produced and posted to Douyin (also known as TikTok) of China.

On the business side, on a single busy day, Alibaba.com needs to take in over 800 million purchase orders for over 15 million products, handle over a billion payments, and deliver more than 30 million packages. Amazon.com also operates at a similar scale, if not even larger. Those numbers are still growing and growing fast!

On the scientific front, super-resolution microscopy imaging technologies have undergone tremendous advances in the past decades,[2] and some are now capable of producing massive quantities of images with subatomic resolution. High-throughput gene sequencing technologies are capable of sequencing hundreds of millions of DNA molecule fragments at a time,[3] and can sequence in just a few hours an entire human genome that has a length of over 3 billion base pairs and contains 20 000 protein-encoding genes!

[1] Almost all of them are passing through several processing pipelines for face detection, face recognition, and general object classification for content screening, etc.

[2] For example, in 2014, Eric Betzig, Stefan W. Hell, and William E. Moerner were awarded the Nobel Prize in Chemistry for the development of super-resolution fluorescence microscopy that bypasses the limit of 0.2 micrometers of traditional optical microscopy.

[3] In 2002, Sydney Brenner, John Sulston, and Robert Horvitz were awarded the Nobel Prize in Physiology or Medicine for their pioneering work and contributions to the Human Genome project.

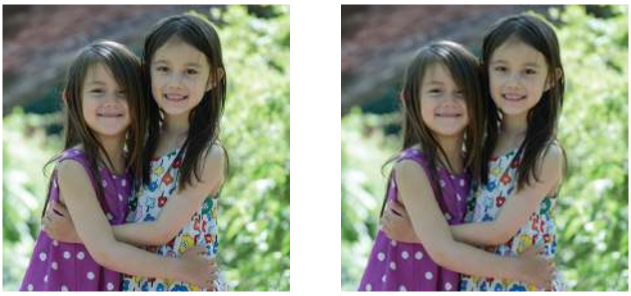


**Figure 0.1** Images of Mary and Isabella: the resolution of the image on the left is 2500 × 2500, whereas the image on the right is down-sampled to 250 × 250, with only 1/100-th fraction of pixels of the original one.

### *Paradigm Shift in Information Acquisition, Processing, and Analysis*

In the past, scientists or engineers have sought to carefully control the data acquisition apparatus and process. Since the apparatus was expensive and the process time-consuming, typically only necessary data (or measurements) were collected for a specific given task. The data or signals collected were mostly informative for the task and did not contain much redundant or irrelevant information, except for some uncontrollable noise. Hence, classical signal processing or data analysis typically operated under the following

$$\text{Classical Premise:} \qquad \textbf{Data} \;\approx\; \textbf{Information}.$$

In this classical paradigm, practitioners mostly needed to deal with problems such as removing noise or compressing the data for storage or transport.

As mentioned above, technologies such as the Internet, smart phones, high-throughput imaging, and gene sequencing have fundamentally changed the nature of data acquisition and analysis. We are moving from a “data-poor” era to a “data-rich” era. As pointed out by Jim Gray (a Turing Award winner), “increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.” This is now heralded as *the Fourth Paradigm* of scientific discovery [HTT09].

Nevertheless, data-rich does not necessarily imply “information-rich,” at least not for free. Massive amounts of data are being collected, sometimes without any specific purpose in advance. Scientists or engineers often do not have direct control of the data acquisition process anymore, neither in the quantity nor in the quality of the acquired data. Therefore, any given new task could be inundated with massive amounts of irrelevant or redundant data.

To see intuitively why this is the case, let us first consider the problem of *face recognition*. Figure 0.1 shows two images of two sisters. It is arguably the case that, to human eyes, both images convey the identity of the persons equally well, even though pixels of the second image are merely 1/100-th of the first one. In other words, if we view both images as vectors with their pixel values as coordinates, then the dimension of the low-resolution image vector is merely 1/100-th of the original one.

**Figure 0.2** Detecting and recognizing faces in a large group photo, from the BIRS workshop on "*Applied Harmonic Analysis, Massive Data Sets, Machine Learning, and Signal Processing*," held at Casa Matemática Oaxaca (CMO) in Mexico, 2016.

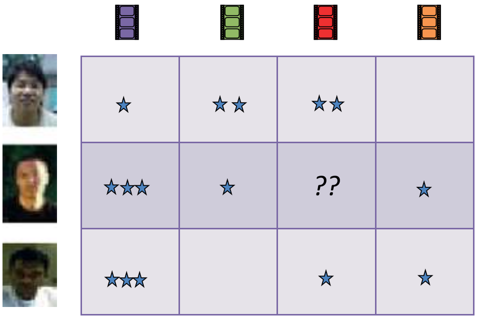
Clearly, the information about the identity of a person relies on statistics of much lower dimension than the original high-resolution image.[4] Hence, in such scenarios, we have the following

$$\text{New Premise I:} \quad \textbf{Data} \gg \textbf{Information}.$$

For *object detection* tasks such as face detection in images or pedestrian detection in surveillance videos, the issue is no longer with redundancy. Instead, the difficulty is to find any relevant information at all in an ocean of irrelevant data. For example, to detect and recognize familiar people from a group photo shown in Figure 0.2, image pixels associated with human faces only occupy a very tiny portion of the image pixels (10 millions in this case) whereas the mass majority of the pixels belong to completely irrelevant objects in the surroundings. In addition, the subjects of interest, say the two authors, are only two among many human faces. Now imagine scaling this problem to billions of images or millions of videos captured with mobile phones or surveillance cameras. Similar "detection" and "recognition" tasks also arise in studying genetics: out of the nearly 20 000 genes and millions of proteins they encode, scientists need to identify which one (or handful of ones) is responsible for certain genetic diseases. In scenarios like these, we have

$$\text{New Premise II:} \quad \textbf{Data} = \textbf{Information} + \textbf{Irrelevant Data}.$$

[4] In fact, one can continue to argue that even such a low-resolution image is still highly redundant. Studies have shown that humans can recognize familiar faces from images with a resolution as low as around $7 \times 10$ pixels [SBOR06]. Recent studies in neuroscience [CT17] reveal that it is possible for the brain to encode and decode any human face using just 200 cells in the inferotemporal (IT) cortex. Modern face recognition algorithms extract merely a few hundred features for reliable face verification.



| | | | |
|---|---|---|---|
| ★ | ★★ | ★★ | |
| ★★★ | ★ | ?? | ★ |
| ★★★ | | ★ | ★ |

**Figure 0.3** An example of collaborative filtering of user preferences: how to guess a customer's rating for a movie even if he or she has not seen it yet?

The explosive growth of e-commerce, online shopping, and social networks has created tremendous datasets of user preferences. Major internet companies typically have records of billions of people's preferences, across millions of commercial products, media contents, and more. By nature, such datasets of user preferences, however massive, are far from complete. For instance, in the case of a dataset of movie ratings as shown in Figure 0.3, no one could have seen all the movies and no movie would have been seen by all people. Nevertheless, companies like Netflix need to guess from such incomplete datasets a customer's preferences so that they could send the most relevant recommendations or advertisements to the customer. This problem in information retrieval literature is known as *collaborative filtering*, and most internet companies' business[5] relies on solving problems such as this one effectively and efficiently. The most fundamental reason why complete information can be derived from such a highly incomplete dataset is that user preferences are not random and the data have structure. For instance, many people have similar tastes in movies and many movies are similar in style. Rows and columns of the user preference table would be strongly correlated, hence the intrinsic dimension (or rank) of the complete table is in fact extremely low compared to its size. Hence, for large (incomplete) datasets drawn from low-dimensional structures, we have

New Premise III: **Incomplete Data** $\approx$ **Complete Information**.

As the above examples suggest, in the modern era of big data, we often face problems of recovering specific information that is buried in highly redundant, irrelevant, seemingly incomplete, or even corrupted[6] datasets. Such information without exception is encoded as certain low-dimensional structures underlying the data, and may only depend on a small (or sparse) subset of the (massive) dataset. This is very different from the classical settings and is precisely the reason why modern data science and engineering are undergoing a fundamental shift in their mathematical and computational paradigms. At its foundation, we need to develop a new mathematical

[5] Most internet companies make money from advertisements, including but not limited to Google, Baidu, Facebook, Bytedance, Amazon, Alibaba, Netflix, etc.

[6] Say due to negligence, misinformation, rumors, or malicious tampering.

framework that characterizes precise conditions under which such low-dimensional information can be correctly and effectively acquired and retained. Equally importantly, we need to develop efficient algorithms that are capable of retrieving such information from massive high-dimensional datasets, at unprecedented speed, at arbitrary scale, and with guaranteed accuracy.

### *Purposes of This Book*

Over the past two decades, there have been explosive developments in the study of low-dimensional structures in high-dimensional spaces. To a large extent, the geometric and statistical properties of representative low-dimensional models (such as sparse and low-rank, and their variants and extensions) are now well understood. Conditions under which such models can be effectively and efficiently recovered from (a minimal amount of sampled) data have been clearly characterized. Many highly efficient and scalable algorithms have been developed for recovering such low-dimensional models from high-dimensional data. The working conditions, and data and computational complexities of these algorithms, have also been thoroughly and precisely characterized. These new theoretical results and algorithms have revolutionized the practice of data science and signal processing, and have had significant impacts on sensing, imaging, and information processing. They have significantly advanced the state of the art for many applications in areas such as scientific imaging,[7] image processing,[8] computer vision,[9] bioinformatics,[10] information retrieval,[11] and machine learning.[12] As we will see from applications featured in this book, some of these developments seem to defy conventional wisdom.

As witnesses to such historical advancements, we believe that the time is now ripe to give a comprehensive survey of this new body of knowledge and to organize these rich results under a unified theoretical and computational paradigm. There are a number of excellent existing books on this topic that already focus on the mathematical/statistical principles of compressive sensing and sparse/low-dimensional modeling [FR13, HTW15, Van16, Wai19, FLZZ20]. Nevertheless, the goal of this book is to bridge, through truly tractable and scalable computation, the gap between principles and applications of low-dimensional models for high-dimensional data analysis with

$$\text{A New Paradigm:}\quad \textbf{Principles} \xleftrightarrow{\textbf{Computation}} \textbf{Applications}.$$

Hence, not only does this book establish mathematical principles for modeling low-dimensional structures and understanding the limits on when they can be recovered, but it also shows how to systematically develop provably efficient and scalable algorithms for solving the recovery problems, leveraging both classical and recent developments in optimization.

[7] Compressive sampling and recovery of medical and microscopic images, etc.
[8] Denoising, super-resolution, inpainting of natural images, etc.
[9] Regular texture synthesis, camera calibration, and 3D reconstruction, etc.
[10] Microarray data analysis for gene–protein relations, etc.
[11] Collaborative filtering of user preferences, documents, and multimedia data, etc.
[12] Especially for interpreting, understanding, and improving deep networks.

Furthermore, through a rich collection of exemplar applications in science and technology, the book aims to further coach readers and students on how to incorporate additional domain knowledge or other nonideal factors (e.g., nonlinearity) in order to correctly apply these new principles and methods to model real-world data and solve real-world problems successfully.

Although the applications featured in this book are inevitably biased by the authors' own expertise and experiences in practicing these general principles and methods, they are carefully chosen to convey diverse and complementary lessons we have learned (often in a hard way). We believe these lessons have value for both theoreticians and practitioners.

### *Intended Audience*

In many ways, the body of knowledge covered in this book has great pedagogical value to young researchers and students in the area of data science. Through rigorous mathematical development, we hope our readers are able to gain new knowledge and insights about high-dimensional geometry and statistics, far beyond what has been established in classical signal processing and data analysis. Such insights are generalizable to a wide range of useful low-dimensional structures and models, including modern deep networks, and can lead to entirely new methods and algorithms for important scientific and engineering problems.

Therefore, this book is intended to be a textbook for a course that introduces basic mathematical and computational principles for sensing, processing, analyzing, and learning low-dimensional structures from high-dimensional data. The *targeted core audience* of this book are entry-level graduate students in electrical engineering and computer science (EECS), especially in the areas of *data science, signal processing, optimization, machine learning*, and *applications*. This book equips students with systematic and rigorous training in concepts and methods of high-dimensional geometry, statistics, and optimization. Through a very diverse and rich set of applications and (programming) exercises, the book also coaches students how to correctly use such concepts and methods to model real-world data and solve real-world engineering and scientific problems.

The book is written to be friendly to both instructors and students. It provides ample illustrations, examples, exercises, and programs from which students may gain hands-on experience with the concepts and methods covered in the book. Materials in this book were developed from several one-semester graduate courses or summer courses offered at the University of Illinois at Urbana-Champaign, Columbia University, ShanghaiTech University, Tsinghua University, and the University of California at Berkeley in the past ten years. The main prerequisites for such a course are college-level linear algebra, optimization, and probability. To make this book accessible to a broader audience, we have tried to make the book as self-contained as possible: we give a crisp summary of facts used in this book from linear algebra, optimization, and statistics in the Appendices. For EECS students, preliminary courses on signal processing, matrix analysis, optimization, or machine learning will improve their appreciation. From our experiences, besides beginning graduate students, many senior undergraduate students at these institutes were able to take the course and read the book without serious difficulty.

### *Organization of This Book*

The main body of this book consists of three inter-related parts: *Principles, Computation*, and *Applications*. The book also contains five *Appendices* on related background knowledge.

- *Part I: Principles (Chapters 2–7)* develops the fundamental properties and theoretical results for sparse, low-rank, and general low-dimensional models. It characterizes the conditions, in terms of sample/data complexity, under which the inverse problems of recovering such low-dimensional structures become tractable and can be solved efficiently, with guaranteed correctness or accuracy.
- *Part II: Computation (Chapters 8–9)* introduces methods from convex and nonconvex optimization to develop practical algorithms that are tailored for recovering the low-dimensional models. These methods show powerful ideas how to systematically improve algorithm efficiency and reduce overall computational complexity so that the resulting algorithms are fast and scalable to large-size and high-dimensional data.
- *Part III: Applications (Chapters 10–16)* demonstrates how principles and computational methods in the first two parts could significantly improve the solutions to a variety of real-world problems and practices. These applications also coach how the idealistic models and algorithms introduced in this book should be properly customized and extended to incorporate additional domain-specific knowledge (priors or constraints) about the applications.
- *Appendices A–E* at the end of the book are meant to make the book largely self-contained. The appendices cover basic mathematical concepts and results from linear algebra, optimization, and high-dimensional statistics that are used in the main body of the book.

The overall organization of these chapters (and appendices) as well as their logical dependence is illustrated in Figure 0.4.

### *How to Use This Book to Teach or to Learn*

The book contains enough material for a two-semester course series. We have purposely organized the material in the book in a modular fashion so that the chapters and even sections can be easily selected and organized to support different types of courses. Here are some examples:

- *A One-Quarter Course on Sparse Models and Methods* for Graduate or Upper Division Undergraduate Students: the introduction Chapter 1 and two theoretical Chapters 2 and 3; the convex optimization Chapter 8, and two to three applications from Chapters 10, 11, and 13, plus some appendices will be ideal for an eight- to ten-week summer or quarter course for senior undergraduate students and early-year graduate students. That is essentially the red route highlighted in Figure 0.4.

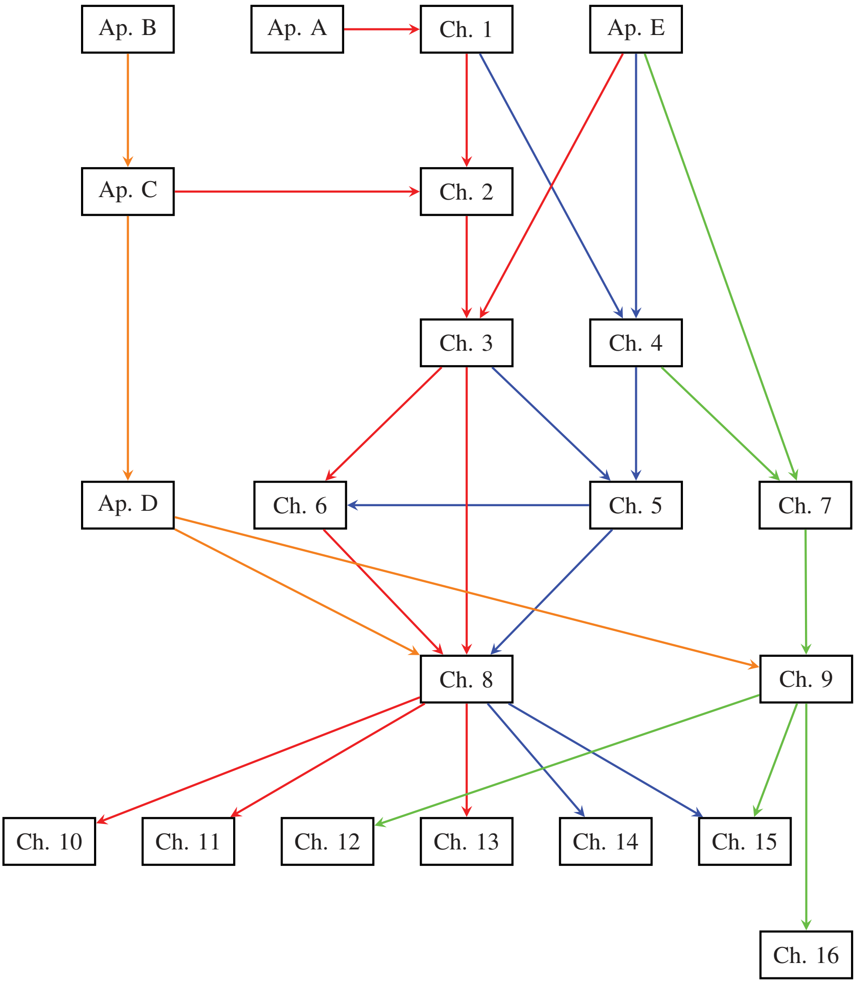


**Figure 0.4** **Organization chart of the book:** dependence among chapters and appendices. Red route: sparse recovery via convex optimization. Blue route: low-rank recovery via convex optimization. Green route: nonconvex approach to low-dimensional models. Orange route: development of optimization algorithms.

- *A One-Semester Course on Low-Dimensional Models* for early-year Graduate Students: the introduction Chapter 1 and the four theoretical Chapters 2–5; the convex optimization Chapter 8, and the several application Chapters 10, 11, 13–15, plus the appendices will be adequate for a one-semester course on low-dimensional models for graduate students. That is essentially both the red and the blue routes highlighted in Figure 0.4.
- *An Advanced-Topic Course on High-Dimensional Data Analysis* for senior Graduate Students who conduct research in related areas: with the previous course as prerequisite, a more in-depth exposition of the mathematical principles, including Chapter 6 on convex methods for general low-dimensional models and Chapter 7

on nonconvex methods. One then can give a more in-depth account of the associated convex and nonconvex optimization methods in Chapters 8 and 9, and several application Chapters 12, 15, and 16 for nonlinear and nonconvex problems. Those are essentially the green and the orange routes highlighted in Figure 0.4. In addition, the instructor may choose to cover new developments in the latest literature, such as broader families of low-dimensional models, more advanced optimization methods, and extensions to deep networks (for low-dimensional submanifolds), say along open directions suggested in the epilogue of Chapter 16.

Certainly, this book can be used as a supplementary textbook for existing (graduate-level) courses on *signal processing* or *image processing*, since it offers more advanced new models, methods, and applications. It can also be used as a complementary textbook for more traditional courses on *optimization* as Chapters 8 and 9 give a rather complete and modern coverage of the first-order (hence more scalable) methods. For a conventional *machine learning* or *statistical data analysis* course, this book may serve as an additional reference for deeper and broader extensions to classic regression analysis, principal component analysis, and deep learning. For a more theoretical course on *high-dimensional statistics and probability*, this book can be used as a secondary text and provides ample motivating and practical examples.

In the future, we would very much like to hear from experienced instructors and seasoned researchers about other good ways to teach or study material in this book. We will share those experiences, suggestions, and even new contributions (examples, exercises, illustrations, etc.) at the book's website, which also contains demos, source code, and other supplementary materials:

https://book-wright-ma.github.io

Further information on the book can also be found on the Cambridge University Press website:
www.cambridge.org/9781108489737 (DOI: 10.1017/9781108779302)

# Acknowledgements

Yi was first introduced to the subject of sparse representation by Professor David Donoho of Stanford when David visited the University of Illinois in 2005. During the dinner, David and Yi discussed Yi's research interests at that time: in particular *generalized principal component analysis* (GPCA) [VMS16], a subject that aims to learn an arbitrary mixture of low-dimensional subspaces from high-dimensional mixed data. David commented that GPCA in its most general setting would be an extremely challenging problem. He suggested starting with the simpler sparse model for which data are assumed to lie on a special family of subspaces.[13] Soon after that, Yi studied the subject of sparse representation systematically, especially during his sabbatical leave at Microsoft Research Asia in Dr. Harry Shum's group in 2006 and later at Berkeley in Professor Shankar Sastry's group in 2007. He was profoundly influenced and inspired by a series of seminal work at the time from Emmanuel Candès and Terence Tao on compressive sensing, error correction, and low-rank matrix recovery.

Since then, we have had the greatest fortune to work closely with many wonderful colleagues in this exciting new field. They are: Emmanuel Candès, Michael Elad, Guillermo Sapiro, Mario Figueiredo, René Vidal, Robert Fossum, Harm Derksen, Thomas Huang, Xiaodong Li, Shankar Sastry, Jitendra Malik, Carlos Fernandez, Julien Mairal, Yuxin Chen, Zhihui Zhu, Daniel Spielman, Peter Kinget, Abhay Pasupathy, Daniel Esposito, Szabolcs Marka, and Zsuzsa Marka. We would also like to thank many of our former colleagues from when we visited or worked at Microsoft Research and other places: Harry Shum, Baining Guo, Weiying Ma, Zhouchen Lin, Yasuyuki Matsushita, Zuowen Tu, David Wipf, Jian Sun, Kaiming He, Shuicheng Yan, Lei Zhang, Liangshen Zhuang, Weisheng Dong, Xiaojie Guo, Xiaoqin Zhang, Kui Jia, Tsung-Han Chan, Zinan Zeng, Guangcan Liu, Jingyi Yu, Shenghua Gao, and Xiaojun Yuan. These collaborations have broadened our knowledge and enriched our experience in this field. Many results featured in this book are conveniently borrowed from these years of fruitful collaborations.

We would like to send special thanks to our former students Allen Yang, Chaobing Song, Qing Qu, and Yuqian Zhang for directly helping with content in some of the chapters. Allen has been of great help during early germination of the book project,

[13] In the last chapter of this book, Chapter 16, we will see a rather unexpected connection between sparse models and GPCA, through an unexpected third party: *deep learning*. Concepts developed for GPCA such as lossy coding rates for clustering subspaces, in Chapter 6 of [VMS16], will play a crucial role in understanding deep networks.

back to early 2013. He has helped draft early versions of the application chapters on magnetic resonance imaging and robust face recognition. Chaobing has helped transform the optimization chapters with a unified parsimonious approach to optimization algorithm design and brought this classic topic to the modern context of scalable computation. We would also like to thank some of our colleagues who have generously shared some material for this book: Bruno Olshausen, Michael Lustig, Julien Mairal, Yuxin Chen, Sam Buchanan, and Tingran Wang.

We would like to thank many of our former and current students. Their research has contributed to many of the results featured in this book. Many of them have also kindly helped with proofreading drafts of the book during different stages or helped in developing exercises as they were taking or teaching assistants for the courses based on early drafts of this book. They are Allen Yang, Arvind Ganesh, Andrew Wagner, Shankar Rao, Zihan Zhou, Hossein Mobahi, Jianchao Yang, Kerui Min, Zhengdong Zhang, Yigang Peng, Xiao Liang, Xin Zhang, Yuexiang Zhai, Haozhi Qi, Yaodong Yu, Christina Baek, Zhengyuan Zhou, Chaobing Song, Chong You, Yuqian Zhang, Qing Qu, Han-Wen Kuo, Yenson Lau, Robert Colgan, Dar Gilboa, Sam Buchanan, Tingran Wang, Jingkai Yan, and Mariam Avagyan.

Last but not the least, we are grateful for generous financial support through all these years from the National Science Foundation, Office of Naval Research, Tsinghua Berkeley Shenzhen Institute, Simons Foundation, Sony Research, HTC, and VIA Technologies Inc.