Generalized Principal Component Analysis

Interdisciplinary Applied Mathematics

Volume 40

Problems in engineering, computational science, and the physical and biological sciences are using increasingly sophisticated mathematical techniques. Thus, the bridge between the mathematical sciences and other disciplines is heavily traveled. The correspondingly increased dialog between the disciplines has led to the establishment of the series: *Interdisciplinary Applied Mathematics.*

The purpose of this series is to meet the current and future needs for the interaction between various science and technology areas on the one hand and mathematics on the other. This is done, firstly, by encouraging the ways that mathematics may be applied in traditional areas, as well as point towards new and innovative areas of applications; and, secondly, by encouraging other scientific disciplines to engage in a dialog with mathematicians outlining their problems to both access new methods and suggest innovative developments within mathematics itself.

The series will consist of monographs and high-level texts from researchersworking on the interplay between mathematics and other fields of science and technology.

More information about this series at http://www.springer.com/series/1390

René Vidal • Yi Ma • S. Shankar Sastry

Generalized Principal Component Analysis

Springer

René Vidal
Center for Imaging Science
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD, USA

Yi Ma
School of Information Science
and Technology
ShanghaiTech University
Shanghai, China

S. Shankar Sastry
Department of Electrical Engineering
and Computer Science
University of California Berkeley
Berkeley, CA, USA

ISSN 0939-6047 ISSN 2196-9973 (electronic)
Interdisciplinary Applied Mathematics
ISBN 978-0-387-87810-2 ISBN 978-0-387-87811-9 (eBook)
DOI 10.1007/978-0-387-87811-9

Library of Congress Control Number: 2015958763

Mathematics Subject Classification (2010): 30C10, 30C40, 62-XX, 62-07, 62-08, 62B10, 62Fxx, 62H12, 62H25, 62H35, 62Jxx, 62J05, 62J07, 14-XX, 14N20, 15-XX

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

To Camille, Margot, and Romeo Vidal (R. V.)
To Diana Xiaoyuan Zhu, Barron, and Henry Ma (Y. M.)
To Claire Tomlin, Samuel, and Lucy Sastry (S. S.)

Preface

We are not very pleased when we are forced to accept a mathematical truth by virtue of a complicated chain of formal conclusions and computations, which we traverse blindly, link by link, feeling our way by touch. We want first an overview of the aim and of the road; we want to understand the idea of the proof, the deeper context.

—Hermann Weyl

Classical theory and methods for the analysis of data were established mainly for engineering and scientific problems that arose five or six decades ago. In these classical settings, engineers or scientists usually had full control of the data acquisition process. As a result, the data to be processed and analyzed were typically *clean* and *complete*: they contained only moderate amounts of noise and were often adequately collected for the specific task or problem of interest. In that regime, many data analysis methods were based on the assumption that most data sets have fewer effective degrees of freedom than the dimension of the ambient space. For example, the number of pixels in an image can be rather large, yet most computer vision models used only a few parameters to describe the appearance, geometry, and dynamics of a scene. This assumption motivated the development of a number of techniques for identifying low-dimensional structures in high-dimensional data, a problem that is important not only for understanding the data, but also for many practical purposes such as data compression and transmission. A popular technique for discovering low-dimensional structure in data is principal component analysis (PCA), which assumes that the data are drawn from a *single* low-dimensional affine subspace of a high-dimensional space (Jolliffe 1986, 2002). PCA is arguably the simplest and most popular dimensionality reduction tool, and it has found widespread applications in many fields such as computer vision (Turk and Pentland 1991).

However, in the past decade or so, there has been a fundamental regime shift in data analysis. Currently, scientists and engineers often must deal with data whose dimension, size, and complexity expand at an explosive rate. Moreover, they have pretty much lost control of the data acquisition process. For instance, in 2012, 350 million photos were uploaded to Facebook every day, and 100 hours

of video were uploaded to YouTube each minute. Moreover, it is estimated that 3.8 trillion photos had been taken by 2012, 10% of them in the last 12 months.[1] This and other forms of massive amounts of data on the Internet and mobile networks are being produced by billions of independent consumers and businesses. How to extract useful information from such massive amounts of data for numerous tasks (such as search, advertisement, scientific analysis) has become one of the biggest engineering endeavors of mankind. Many call it the era of *Big Data*. Obviously, such a regime shift demands a fundamental paradigm shift in data analysis, since classical theory and methods for data analysis were simply not designed to work under such conditions. The website of Theoretical Foundations of Big Data Analysis[2] puts things into perspective:

> *The Big Data phenomenon presents opportunities and perils. On the optimistic side of the coin, massive data may amplify the inferential power of algorithms that have been shown to be successful on modest-sized data sets. The challenge is to develop the theoretical principles needed to scale inference and learning algorithms to massive, even arbitrary, scale. On the pessimistic side of the coin, massive data may amplify the error rates that are part and parcel of any inferential algorithm. The challenge is to control such errors even in the face of the heterogeneity and uncontrolled sampling processes underlying many massive data sets.*

Since the data acquisition process is no longer under the data gatherer's control, the structure of the data to be processed or analyzed can no longer be assumed to be relatively simple or clean: very often, the data contain significant amounts of noise, corrupted entries, and outliers; or the data could be incomplete or inadequate for a task that arises only after the data have been collected; or the data could even have some degree of unknown nonlinearity due to lack of calibration in the data acquisition. In the past decade, these challenges have led to many revolutionary discoveries and much progress in which many of the classical models and methods for data analysis have been systematically generalized or improved to make them robust to such bad nuisances in the data. In the context of identifying low-dimensional structures in the data, classical PCA is *generalized* so that it can robustly find the correct subspace structure of the data despite such nuisances. The forms of progress include entirely new methods for low-rank matrix completion, robust PCA, kernel PCA, and manifold learning.

Another challenge that arises in the new regime is that we can no longer assume that the data lie on a single low-dimensional subspace or submanifold. This is because many modern data sets are not collected for any specific task. Instead, the data may have already been collected, and the task emerges only afterward. Hence a data set can be mixed with multiple classes of data of different natures, and the intrinsic structure of the data set may be *inhomogeneous* or *hybrid*. In this case, the data set may be better represented or approximated by not one, but multiple low-dimensional subspaces or manifolds. Figure 1 gives an example of face images

[1] http://www.buzzfeed.com/hunterschwarz/how-many-photos-have-been-taken-ever-6zgv.

[2] http://simons.berkeley.edu/programs/bigdata2013.

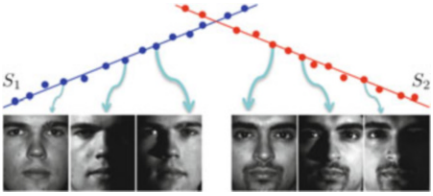

Fig. 1 Face images from multiple individuals can be well approximated by multiple low-dimensional subspaces.

under varying illumination conditions, where each affine subspace corresponds to face images of a different individual. This leads to a general problem: Given a set of data points from a *mixture* of affine subspaces, how does one automatically learn or infer those subspaces from the data? A solution to this problem requires one to *cluster* or *segment* the data into multiple groups, each belonging to one subspace, and then *identify* the parameters of each subspace. To model data with such mixed subspace structures, the classical PCA method needs to be *generalized* so that it can simultaneously identify multiple subspaces from the data. This leads to the so-called *subspace clustering* problem, which has received great attention in the last decade and has found widespread applications in computer vision, image processing, pattern recognition, and system identification.

Purpose of This Book

The purpose of this book is to provide a comprehensive introduction to the latest advances in the mathematical theory and computational tools for modeling high-dimensional data drawn from one or more low-dimensional subspaces (or manifolds) and corrupted by noise, missing entries, corrupted entries, and outliers. This will require the development of new algebraic, geometric, statistical, and computational theory and methods for efficient and robust estimation of one or more subspaces. To distinguish this theory and these methods from classical PCA, we call all such advanced approaches as *generalized principal component analysis* or GPCA for short.[3]

As we will see in this book, in order to generalize classical PCA to the case of corrupted and mixed data, we need to resort to a body of more advanced mathematical tools from estimation theory, algebraic geometry, high-dimensional

[3]In the literature, the word *generalized* is sometimes used to indicate any particular extension to classical PCA (Jolliffe 1986, 2002). In our opinion, each of these extensions is a particular generalization rather than the more systematic generalization that we present in this book. In addition, for the case in which we want PCA to handle large amounts of corruptions or outliers, we may use the special name *robust PCA* (RPCA); for the *nonlinear* case in which each component is an algebraic variety of higher degree such as a quadratic surface or a more complicated manifold, we may use the name *nonlinear PCA* or *manifold learning*; for the case of multiple subspaces or manifolds, names such as *mixtures of probabilistic PCA* (MPPCA), *subspace clustering* (SC) and *hybrid component analysis* (HCA) have been suggested and would also be appropriate.

statistics, and convex optimization. In particular, in this book and its appendices, we will give a systematic introduction to effective and scalable optimization techniques tailored to estimating low-dimensional subspace structures from high-dimensional data (see Appendix A), all the related statistical theory and methods for robust estimation of mixture models (see Appendix B), as well as a complete characterization of the algebraic properties of a union of multiple subspaces as an algebraic set (see Appendix C). As we will see throughout this book, the statistical, algebraic-geometric, and computational aspects of GPCA are highly complementary to each other. Each of them leads to solutions and algorithms of their own that hold certain conceptual or computational advantages against other approaches under certain assumptions about the data and/or the subspaces.

There are several reasons why we feel that the time is now ripe to write a book about GPCA:

1. The limitations of classical PCA have been well known to engineers and practitioners of modern data analysis. However, PCA remains the method of choice by many field engineers simply because they do not have a systematic body of theory and methods for handling different types of nuisances in the data. In the past few years, with advances in algebraic geometry, high-dimensional statistics, and convex optimization, our understanding of the problem of estimating a low-dimensional subspace has gone well beyond classical settings: we have not only a better understanding of the geometric, statistical, and probabilistic nature of PCA, but also computationally efficient algorithms for PCA with missing and corrupted data that give provably correct solutions under broad conditions. In addition, the field of estimating mixture models, in particular a mixture of subspaces, has also gone through revolutionary developments in the past few years. The statistical, algebraic, geometric, and computational properties of this class of models have been reasonably well understood. As result, many effective and efficient algorithms have been developed for this problem.
2. These new developments obviously come at a very good time, since both science and engineering are entering the era of *Big Data*. Many of the new algorithms and techniques have already demonstrated great success and potential in many important practical problems of image processing and pattern analysis, as we will demonstrate with some concrete applications and examples in this book. We anticipate that these new theoretical results and the associated computational methods will provide scientists and engineers with a new set of models, principles, and tools that can be readily applied to a broad range of practical problems and real-world data, far beyond the applications and data illustrated in this book.

Intended Audience of This Book

We have written this book with the idea that it will have both research and pedagogical value. From a research perspective, the topics covered in this book are of great relevance and importance to both theoreticians and practitioners in such areas as data science, machine learning, pattern recognition, computer vision, signal and image processing, and system identification. The motivating examples

and applications given in this book are purposely biased by our own research interests in image processing and computer vision, because we believe that from a pedagogical perspective, visual data and examples can best illustrate some of the abstract models and properties introduced. Nevertheless, the basic theory and algorithms are established in fairly general terms, and are obviously applicable to many practical engineering and scientific problems well beyond those described in this book.

We believe that the material of the book is ideal for an introductory graduate course for students in data science, machine learning, and signal processing, or an advanced course for students in computer vision, estimation theory, and systems theory. Through arguably the simplest class of models, the low-dimensional linear models, the book introduces to students some of the most fundamental principles in data modeling, statistical inference, optimization, and computation. Knowledge about these basic models and their properties is absolutely necessary for anyone who strives to study more sophisticated classes of models in which low-dimensional linear models are the key building blocks, such as the sparse models in compressive sensing and the deep neural networks in machine learning (see Chapter 13 for further discussion).

The book is written to be friendly to beginning graduate students and instructors. At the end of each chapter, we have provided many basic exercises and programs from which students may gain hands-on experience with the material covered in the chapters as well as an extensive survey of related literature for research purposes. Additional information, resources, and sample code for most of the examples, algorithms, and applications featured in this book will be made available at the book's website: http://www.vision.jhu.edu/gpca.

We have used material from this book many times to teach a one-semester graduate course at the Johns Hopkins University, the University of Illinois at Urbana-Champaign, the University of California at Berkeley, and the ShanghaiTech University in China. As the reader will see, GPCA is a very unique subject that touches on many fundamental concepts, facts, and principles across engineering, computation, statistics, and mathematics. Therefore, this is a great topic that can shepherd researchers and students to systematically establish some of the most fundamental and useful knowledge for modern data science and machine learning. We also believe that the reader will learn to appreciate the complementary nature of different perspectives and approaches presented in this book, and in the end develop a deep and comprehensive understanding of the subject.

Organization of This Book

Chapter 1 gives a nontechnical introduction to the basic problems, ideas, and principles studied in this book. The remainder of the book is organized into four *parts*:

Part I covers classical and modern theory and methods for modeling data with a single low-dimensional linear or affine subspace (or a nonlinear submanifold). More specifically, **Chapter 2** gives a review of classical PCA theory and methods for subspace estimation, including its statistical, geometric, and rank minimization

interpretations. The chapter also covers a simple generative model for PCA, called probabilistic PCA, as well as model selection issues for PCA. **Chapter 3** shows how to estimate a subspace when the data are incomplete or corrupted. The chapter discusses statistical and alternating minimization methods for robust PCA, as well as some advanced tools from compressive sensing for sparse and low-rank recovery. Since complete proofs for these results are beyond the scope of this book, we will simply discuss their implications and show how to use them to develop effective algorithms for robust PCA. **Chapter 4** shows how to extend the methods for learning linear subspaces to nonlinear submanifolds. In particular, the chapter introduces both parametric and nonparametric methods for manifold learning, including nonlinear PCA, kernel PCA, locally linear embedding, and Laplacian eigenmaps. The chapter also introduces the basic K-means algorithm for clustering data distributed around a few cluster centers, as well as the more advanced spectral clustering algorithm, which combines manifold learning methods with K-means to cluster mixed data that have more complex nonlinear structures.

Part II covers three complementary approaches and methods for modeling data with a mixture of multiple subspaces. More specifically, **Chapter 5** studies the algebraic-geometric properties of a mixture of subspaces, also known in modern algebra as a subspace arrangement. The chapter introduces a basic noniterative algebraic method for estimating multiple subspaces, which works effectively and efficiently when the data are relatively clean and the ambient dimension is low. **Chapter 6** introduces several statistical methods for estimating mixture subspace models. They are based on different but related statistical principles, including the minimax principle (the K-subspaces method), the maximum likelihood principle (the EM algorithm), and the minimum description/coding length principle (the compression-based agglomerative clustering method). **Chapter 7** explores the nonparametric spectral clustering method for subspace clustering and introduces many different ways to establish affinity matrices for data points in a mixture of subspaces, based on local, semilocal, and global geometric information. **Chapter 8** develops principled ways to establish affinity matrices for subspace clustering via self-expressive low-rank or sparse representations. It introduces modern convex optimization techniques to find such representations. It also studies under what conditions this approach gives provably correct solutions.

Part III demonstrates a few representative applications of the methods and algorithms introduced in earlier chapters. More specifically, **Chapter 9** shows how to cluster image patches into multiple subspaces and learn a hybrid linear model from them for the purpose of building highly compact and sparse representations of natural images. **Chapter 10** shows how to segment natural images into multiple regions corresponding to different colors and textures based on data compression and subspace clustering techniques introduced in this book. **Chapter 11** shows how to segment multiple moving objects in an image sequence using many of the subspace clustering algorithms presented in this book. The chapter also provides an empirical comparison of these methods on motion segmentation data, and discusses their strengths and weaknesses. The chapter also shows how to extend subspace clustering algorithms to a special class of nonlinear manifolds

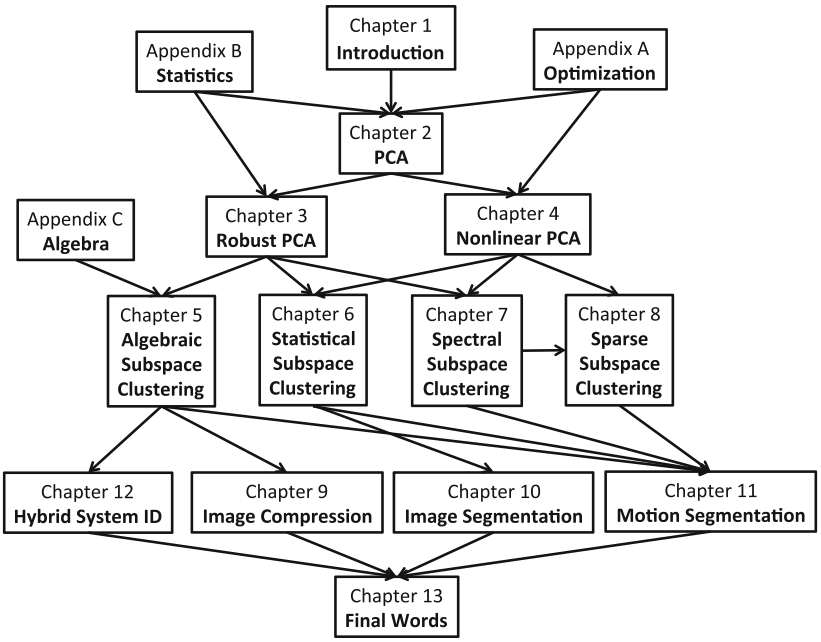

Fig. 2 Organization of the Book—logical dependency among all the chapters and the appendices.

arising in the motion segmentation problem. The chapter also shows how subspace clustering algorithms can be used to segment video and time series into multiple events or actions. **Chapter 12** studies the temporal segmentation problem more systematically. The algebraic subspace clustering method is modified and extended to segment observations that are generated by a hybrid linear dynamical system and to subsequently identify all the underlying dynamical models.

Part IV covers relevant concepts and results in optimization, mathematical statistics, and algebraic geometry in order to make the book self-contained. More specifically, **Appendix A** covers basics notions from optimization, such as first- and second-order conditions for optimality, convexity, gradient descent methods, alternating minimization methods, constrained optimization, duality, Lagrange methods, augmented Lagrange methods, and the alternating direction method of multipliers. **Appendix B** covers basic notions from statistics, such as sufficient statistics, unbiased estimators, maximum likelihood estimation, expectation maximization, mixture models, model selection, and robust statistics. **Appendix C** covers basic notions from algebraic geometry, including polynomial rings, ideals, algebraic sets, subspace arrangements, ideals of subspace arrangements, and Hilbert functions of subspace arrangements. All these concepts and results may come in handy for readers who are not so familiar with certain mathematical facts used in the book, especially for the early chapters.

Last but not least, **Chapter 13** discusses some of the related open research topics and future directions that are not covered by this book.

We have taught the material of Chapters 1–8 several times in a one-semester course, and have covered the entire book with some of the additional proofs for the material in Chapters 3–8 and applications in Part III in a two-semester sequence. We invite instructors to experiment with alternative ways of covering this material. To help instructors design their courses, we have outlined in Figure 2 the overall book organization and logical dependency among all the chapters and appendices. We would be delighted to hear of your experiences in this regard.

Baltimore, MD, USA — René Vidal
Shanghai, China — Yi Ma
Berkeley, CA, USA — S. Shankar Sastry
August 2015

Acknowledgments

As we express our gratitude, we must never forget that the highest appreciation is not to utter words, but to live by them.

—John F. Kennedy

Our initial motivation for trying to generalize principal component analysis to multiple subspaces can be traced back to early 2001, when René, Shankar, and colleagues were developing methods for having a team of robots pursue another team of robots using visual information. For this purpose, we needed to develop methods for estimating the pose of multiple moving objects in a video taken by a moving camera. At the time, methods for estimating the pose of an object relative to a moving camera were well understood, including many methods developed by Yi in his PhD thesis. However, the problem of estimating the pose of multiple moving objects in a video was not as well understood. In particular, the main challenge was that we often do not know which pieces of the video correspond to the same moving object; hence we needed both to segment the video and to estimate the pose of each object, i.e., we needed to solve the *motion segmentation problem*.

To address these issues, René and Yi began to work on a polynomial-based method for solving the motion segmentation problem. The approach was based on fitting a high-order polynomial to the image data and factorizing it into multiple bilinear factors, each one encoding the pose of each one of the moving objects. Interestingly, we observed that the bilinear factorization problem could be reduced to the problem of factorizing a polynomial into a product of linear factors, which in turn provided a solution to the problem of clustering data drawn from a union of planes in three-dimensional space, i.e., the *plane clustering problem*. We soon realized that this polynomial-based method for solving the plane clustering problem could be extended to subspaces of arbitrary dimensions, a problem that was common and fundamental to many data modeling, clustering, and classification problems in pattern recognition, computer vision, signal/image processing, and systems theory. However, at the time there was a serious lack of systematic study and understanding of this very important class of models and problems, and many algorithms at the time were heuristic or ad hoc. This inspired us to work very actively during the next

few years to develop a more complete theoretical and algorithmic foundation for this class of models. As of August 2003, René had summarized much of the algebraic-geometric method in his PhD dissertation at Berkeley (see Chapter 5), including some initial applications to motion segmentation.

In early 2004, on the day after Yi's wedding reception, we decided to formalize our findings with a manuscript and sketched an early outline of this book at Café Kopi in downtown Champaign, Illinois. Our initial plan was to extend the algebraic-geometric approach by developing more efficient and robust techniques for estimating low-dimensional subspace structures from imperfect mixed high-dimensional data, and to apply these techniques to a broad class of engineering and scientific problems. Following René's PhD dissertation and some of our earlier papers, many of our graduate students, postdocs, and colleagues enriched and extended the theory, algorithms, and applications of the algebraic-geometric approach to many new settings and problems in computer vision, image processing, and system identification. We especially thank our former students Laurent Bako, Yasmin Hashambhoy, Wei Hong, Kun Huang, Jacopo Piazzi, Dheeraj Singaraju, Roberto Tron, Allen Yang, and John Wright for their development of robust algebraic-geometric approaches to subspace clustering and their applications to image compression, image segmentation, motion segmentation and hybrid system identification, which are featured in Chapters 6, 9, 10, 11, and 12. We are also greatly indebted to Professor Robert Fossum and Professor Harm Derksen. They painstakingly taught Yi algebraic geometry and have helped develop a rather complete characterization of the algebraic properties of subspace arrangements. Their work has helped to provide a rigorous mathematical foundation for the algebraic subspace clustering algorithms developed in this book (Chapter 5 and Appendix C). Professor Harm Derksen was also the first to suggest to Yi the use of compression for subspace clustering. He and John Wright helped develop the compression-based subspace clustering work (featured in Chapter 6), which is very complementary to other existing methods. We also thank Professor Richard Hartley and Professor Brian Anderson for their contributions to the applications of the algebraic-geometric methods to motion segmentation (featured in Chapter 11) and hybrid system identification (featured in Chapter 12), respectively.

As it turned out, around 2007 we realized that our initial plan based on extending the algebraic-geometric approach to subspace clustering was a little premature: We did not fully realize at the time how much and how fast this topic was to evolve in years to come. In fact, during the next few years, even classical PCA for learning a single subspace went through unprecedented development with many new algorithms based on more solid statistical and mathematical principles that are much more effective and robust than classical techniques. We especially thank our former students Arvind Ganesh, Hossein Mobahi, Shankar Rao, Allen Yang, Andrew Wagner, and John Wright of UIUC for their development of robust approaches to classical PCA based on sparse and low-rank minimization, and their applications to face recognition. We are also extremely grateful to Professor Emmanuel Candès for his pioneering and inspiring work in this area. His collaboration with Yi on robust

PCA is featured in Chapter 3 together with his elegant treatment of the low-rank matrix completion problem.

Such exciting developments in the theory of classical PCA led us to revisit the subspace clustering problem with much more advanced mathematical tools and entirely new perspectives. In particular, we are greatly indebted to our former students Ehsan Elhamifar, Alvina Goh, Shankar Rao, Guangcan Liu, and Roberto Tron for the development of manifold learning and spectral clustering approaches to subspace clustering, which we have featured in Chapter 7 and Chapter 8. Special thanks go to Ehsan Elhamifar, who was a key contributor to the development of subspace clustering methods based on sparse representation, and to Guangcan Liu and Professor Paolo Favaro for the development of subspace clustering methods based on low-rank representation, both featured in Chapter 8. Their work extended not only the theory of subspace clustering, but also their applicability to midsize data sets, including face and digit clustering. We also thank Mahdi Soltanolkotabi and Professor Emmanuel Candès for their recent elegant theoretical analysis of the sparse subspace clustering algorithm, which is also featured in Chapter 8.

While we were witnessing such exciting developments, our book plan had been delayed repeatedly. Only very recently did we all become convinced that this topic had become stable and mature enough for us to fulfill our ten-year-old commitment. The final version of this book would have not been completed without the help of René's students Chong You and Manolis Tsakiris. Chong was kind enough to help generate most of the wonderful running examples for the algorithms presented this book, especially those on face images and motion capture, while Manolis generated many of the synthetic examples, especially those based on algebraic methods. In return, this book has inspired their own PhD work: Chong is currently developing theory and algorithms to scale up sparse and low-rank subspace clustering methods to the big data domain, while Manolis is revisiting the algebraic-geometric approach to make it robust. Overall, it looks like this book has come full circle, and we may have robust algebraic-geometric algorithms in the near future thanks to Manolis's work. We also thank Chong, Manolis, and Ben Haeffele for proofreading the final version of this book and giving us fantastic comments on how to improve the presentation of the material.

We thank Professor Martin Vetterli of École Polytechnique Fédérale de Lausanne, Professor David Donoho of Stanford University, and Professor Guillermo Sapiro of Duke University for their encouragement for us to apply generalized PCA models to sparse image representation. Yi would like to thank Professor David Donoho in particular for making the early suggestion about the strong connection between subspace arrangements and sparse representation. One could say that the story of GPCA would never be so profound and complete without the advanced theory and computational tools from compressive sensing and sparse representation.

We thank many of our colleagues for valuable collaborations, discussions, suggestions, and moral support. They are Professor Stefano Soatto of the University of California at Los Angeles, Professor Jana Kosecka of George Mason University, Professor Richard Hartley of the Australian National University, Professors Jitendra Malik and Ruzena Bajcsy of the University of California at Berkeley, Dr. Harry Shum, Dr. Yasuyuki Matsushita, and Dr. David Wipf of Microsoft, Professor Zhouchen Lin of Peking University (or Microsoft Research at the time of collaboration), Professor Shuicheng Yan of National University of Singapore, Professors Robert Fossum, Minh Do, Thomas Huang, Narendra Ahuja, Daniel Liberzon, and Yizhou Yu of the University of Illinois at Urbana-Champaign, Professor Rama Chellapa from the University of Maryland at College Park, and Professors Don Geman, Gregory Hager, Michael Miller, Daniel Robinson, Laurent Younes, of The Johns Hopkins University. In particular, Professor Robert Fossum helped with proofreading an early version of the manuscript (containing essentially Chapters 1–6), which Yi used to teach an earlier course on GPCA at UIUC in 2006. We also thank Professors Alvaro Soto and Domingo Mery from the Catholic University of Chile, Professors Jean Ponce and Francis Bach from INRIA, and Professor Emmanuel Candès from Stanford University for hosting René during his sabbatical in 2012, when many chapters of this book were written.

We are obviously grateful for all the funding agencies and institutes that have supported us through all these years. In particular, we would like to thank our funders and program managers, Dr. Daniel DeMenthon, Dr. Helen Gill, Dr. Haesun Park, Dr. John Cozzens, and Dr. Jie Yang of the National Science Foundation, and Dr. Behzad Kamgar-Parsi of the Office of Naval Research. They have generously supported our research in this direction even when much of the theory and results were still in their infancy. Without their vision and trust, the area of GPCA let alone this book, would not have been possible. We would like to acknowledge the research funding of NSF under grants IIS-0347456, CAREER-IIS-0447739, CNS-EHS-0509101, CRS-EHS-0509151, CCF-TF-0514955, ECCS-0701676, IIS-0703756, CNS-0834470, CCF-0964215, ECCS-0941463, CSN-0931805, OIA-0941362, IIS-0964416, IIS-1116012, IIS-1218709, IIS-1335035, and IIS-1447822, ONR under grants N00014-00-10621, N00014-05-10633, N00014-05-10836, N00014-09-10084, N00014-09-10230, N00014-09-10839, and N00014-13-10116, and DARPA under grants F33615-98-C-3614 and KECoM 10036-100471 for their support of our work. René would like to thank the Sloan Research Fellowship for partially supporting his sabbatical. Yi would like to pay special thanks to the generous startup support from ShanghaiTech and moral support from Dean Cher Wang and President Mianheng Jiang. Their vision and determination to reform Chinese higher education and research has encouraged Yi to focus on writing this book till its completion over the past two years at ShanghaiTech. GPCA has now become part of the regular curricula for the areas of data science, signal processing, and machine learning for the School of Information Science and Technology of ShanghaiTech.

Finally and most importantly, our families, including the six little ones who were born during the gestation of this book, have provided us with a huge amount of love, encouragement, and support in the writing of this book.

Baltimore, MD, USA — René Vidal
Shanghai, China — Yi Ma
Berkeley, CA, USA — S. Shankar Sastry
August 2015

Contents

Glossary of Notation

Frequently used mathematical symbols are defined and listed according to the following categories:

0. Set theory and logic symbols
1. Sets and linear spaces
2. Transformation groups
3. Vector and matrix operations
4. Geometric primitives in space
5. Probability and statistics
6. Graph theory
7. Image formation

Throughout the book, **every vector is a column vector unless stated otherwise!**

0. Set theory and logic symbols

$\cap$	$S_1 \cap S_2$ is the intersection of two sets
$\cup$	$S_1 \cup S_2$ is the union of two sets
$\doteq$	Definition of a symbol
$\exists$	$\exists s \in S, P(s)$ means there exists an element s of set S such that proposition $P(s)$ is true
$\forall$	$\forall s \in S, P(s)$ means for every element s of set S, proposition $P(s)$ is true
$\in$	$s \in S$ means s is an element of set S
$\|S\|$	The number of elements in set S
$\setminus$	$S_1 \setminus S_2$ is the difference of set S_1 minus set S_2
$\subset$	$S_1 \subset S_2$ means S_1 is a proper subset of S_2
$\{s\}$	A set consists of elements like s
$\rightarrow$	$f : D \rightarrow R$ means a map f from domain D to range R
$\mapsto$	$f : x \mapsto y$ means f maps an element x in the domain to an element y in the range
$\circ$	$f \circ g$ means composition of map f with map g
$\vee$	$\mathcal{P} \vee \mathcal{Q}$ is true if either proposition $\mathcal{P}$ or proposition $\mathcal{Q}$ is true
$\wedge$	$\mathcal{P} \wedge \mathcal{Q}$ is true if both proposition $\mathcal{P}$ and proposition $\mathcal{Q}$ are true
$\implies$	$\mathcal{P} \implies \mathcal{Q}$ means proposition $\mathcal{P}$ implies proposition $\mathcal{Q}$
$\iff$	$\mathcal{P} \iff \mathcal{Q}$ means propositions $\mathcal{P}$ and $\mathcal{Q}$ imply each other
$\mid$	$\mathcal{P} \mid \mathcal{Q}$ means proposition $\mathcal{P}$ holds given the condition $\mathcal{Q}$

1. Sets and linear spaces

$\mathbb{C}$	The set of all complex numbers

$\mathbb{C}^n$	The n-dimensional complex linear space
$\mathbb{P}^n = \mathbb{RP}^n$	The n-dimensional real projective space
$\mathbb{R}$	The set of all real numbers
$\mathbb{R}^n$	The n-dimensional real linear space
$\mathbb{R}_+$	The set of all nonnegative real numbers
$\mathbb{Z}$	The set of all integers
$\mathbb{Z}_+$	The set of all nonnegative integers
L	A generic 1-D line in space
S	Typically represents a generic linear or affine subspace
P	A generic 2-D plane in space

2. Geometric primitives in space

$x \in \mathbb{R}$	A lower-case letter normally represents a scalar
$\boldsymbol{x} \in \mathbb{R}^D$	A bold lower-case letter represents a vector or a random vector
$\boldsymbol{x}_j \in \mathbb{R}^D$	The jth sample vector in a data set
$\mathcal{X} \subset \mathbb{R}^D$	Represents a set of data points: $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$
$X \in \mathbb{R}^{D \times N}$	A capital letter represents a matrix, very often representing the data matrix with the data points as its columns: $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$
$\mathcal{X}_i \subset \mathcal{X}$	The ith subset or cluster of the dataset $\mathcal{X}$
X_i	The submatrix of X associated with the ith cluster $\mathcal{X}_i$

3. Vector and matrix operations

$\|\|\boldsymbol{x}\|\|_2$	The 2-norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$: $\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$
$\|\|\boldsymbol{x}\|\|_1$	The 1-norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$: $\|x_1\| + \|x_2\| + \cdots + \|x_n\|$
$\|\|\boldsymbol{x}\|\|_0$	The 0-norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$: the number of nonzero values
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \in \mathbb{R}$	The inner product of two vectors: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}$
$\boldsymbol{x} \sim \boldsymbol{y}$	Homogeneous equality: two vectors or matrices $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal up to a nonzero scalar factor
$\boldsymbol{x} \times \boldsymbol{y} \in \mathbb{R}^3$	The cross product of two 3-D vectors: $\boldsymbol{x} \times \boldsymbol{y} = \widehat{\boldsymbol{x}}\boldsymbol{y}$
$\boldsymbol{x} \otimes \boldsymbol{y}$	The Kronecker (tensor) product of $\boldsymbol{x}$ and $\boldsymbol{y}$
$\text{span}(M)$	The range or subspace spanned by the columns of a matrix M
$\text{rank}(M)$	The rank of a matrix M
$\text{null}(M)$	The null space or kernel of a matrix M
$\det(M)$	The determinant of a square matrix M
$M^\top \in \mathbb{R}^{n \times m}$	Transpose of a matrix $M \in \mathbb{R}^{m \times n}$ (or a vector)
$\text{trace}(M)$	The trace of a square matrix M, i.e., the sum of all its diagonal entries, sometimes shorthand as $\text{tr}(M)$
$M = U \Sigma V^\top$	The singular value decomposition of a matrix M
$\|\|M\|\|_*$	The nuclear norm of a matrix M: the sum of all its singular values
$\|\|M\|\|_0$	The 0-norm of a matrix M: the number of nonzero values

$\|M\|_F$	The Frobenius norm of a matrix M: the square root of the sum of the square of its entries
$S_1 \oplus S_2$	The direct sum of two linear subspaces S_1 and S_2
$S^\perp$	The orthogonal complement of a subspace S
$P_S(\boldsymbol{x})$	Projecting a vector $\boldsymbol{x}$ onto the subspace S

4. Transformation groups

$GL(n) = GL(n, \mathbb{R})$	The real general linear group on $\mathbb{R}^n$; it can be identified as the set of $n \times n$ invertible real matrices
$SL(n) = SL(n, \mathbb{R})$	The real special linear group on $\mathbb{R}^n$; it can be identified as the set of $n \times n$ real matrices of determinant 1
$A(n) = A(n, \mathbb{R})$	The real affine group on $\mathbb{R}^n$; an element in $A(n)$ is a pair $(A, \boldsymbol{b})$ with $A \in GL(n)$ and $\boldsymbol{b} \in \mathbb{R}^n$ and it acts on a point $\boldsymbol{x} \in \mathbb{R}^n$ as $A\boldsymbol{x} + \boldsymbol{b}$
$O(n) = O(n, \mathbb{R})$	The real orthogonal group on $\mathbb{R}^n$; if $U \in O(n)$, then $U^\top U = I$
$SO(n) = SO(n, \mathbb{R})$	The real special orthogonal group on $\mathbb{R}^n$; if $R \in SO(n)$, then $R^\top R = I$ and $\det(R) = 1$
$SE(n) = SE(n, \mathbb{R})$	The real special Euclidean group on $\mathbb{R}^n$; an element in $SE(n)$ is a pair $(R, \boldsymbol{t})$ with $R \in SO(n)$ and $\boldsymbol{t} \in \mathbb{R}^n$ and it acts on a point $\boldsymbol{x} \in \mathbb{R}^n$ as $R\boldsymbol{x} + \boldsymbol{t}$

5. Probability and statistics

$p_\theta(\boldsymbol{x})$	The probability density function of the random variable or vector $\boldsymbol{x}$ with θ as parameters of the distribution, sometimes also written as $p(\boldsymbol{x}, \theta)$
$p(\boldsymbol{y} \mid \boldsymbol{x})$	The conditional probability density function of the random variable $\boldsymbol{y}$ given $\boldsymbol{x}$
$P(\cdot)$	The probability of a random event
$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$	The expectation (or mean) of a random variable or vector $\boldsymbol{x}$
$\Sigma_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$	The covariance matrix of a random vector $\boldsymbol{x}$
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	The normal (Gaussian) distribution with mean $\boldsymbol{\mu}$ and covariance Σ

6. Graph theory

$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$	An (undirected) graph consisting of a set of vertices $\mathcal{V}$ and (weighted) edges $\mathcal{E}$
$\mathcal{V} = \{1, \ldots, N\}$	The set of N vertices of a graph $\mathcal{G}$, where in this book a vertex typically represents one data point
$\mathcal{E} = \{(i, j)\}$	The set of (weighted) edges of a graph $\mathcal{G}$, where in this book an edge typically represents two data points belonging to the same cluster
$w_{ij} \in \mathbb{R}_+$	A weight associated with the edge $(i, j) \in \mathcal{E}$, where in this book the weight value represents the affinity between two data points

W	Weight matrix of a graph $\mathcal{G}$, with w_{ij} as its entries
$\mathcal{D}$	Degree matrix of a graph $\mathcal{G}$, a diagonal matrix whose diagonal entries are the degree $d_{ii} = \sum_j w_{ij}$ of each vertex $i \in \mathcal{V}$
$\mathcal{L}$	Laplacian matrix of a graph $\mathcal{G}$, defined as $\mathcal{L} = \mathcal{D} - W$

7. Image formation

$(R_i, \boldsymbol{T}_i)$	Relative motion (rotation and translation) from the ith camera frame to the (default) first camera frame: $X_i = R_i X + \boldsymbol{T}_i$
$(R_{ij}, \boldsymbol{T}_{ij})$	Relative motion (rotation and translation) from the ith camera frame to the jth camera frame: $X_i = R_{ij} X_j + \boldsymbol{T}_{ij}$
$H \in \mathbb{R}^{3\times 3}$	The homography matrix, and it usually represents an element in the general linear group $GL(3)$