Rethinking PCA for Modern Data Sets: Theory, Algorithms, and Applications

By NAMRATA VASWANI
Guest Editor

YUEJIE CHI, ***Senior Member IEEE***
Guest Editor

THIERRY BOUWMANS
Guest Editor

I. INTRODUCTION

In today's big and messy data age, there is a lot of data generated everywhere around us. Examples include texts, tweets, network traffic, changing Facebook connections, or video surveillance feeds coming in from one or multiple cameras. Dimension reduction and noise/outlier removal are usually important preprocessing steps before any high-dimensional (big) data set can be used for inference. A common way to do this is via solving the principal component analysis (PCA) problem or its robust extensions. The basic PCA problem has been studied for over a century since the early work by Pearson in 1901 and Hotelling in 1933. The aim of PCA is to reduce the dimensionality of multivariate data while preserving as much of the relevant information as possible. It is often the first step in various types of exploratory data analysis, predictive modeling, and classification and clustering tasks, and finds applications in biomedical imaging, computer vision, process fault detection, recommendation systems' design, and many more domains.

The papers in this special issue introduce the reader to the theory, algorithms, and applications of principal component analysis and its many extensions.

"PCA" refers to the following problem. Given a data set (a set of data vectors, or, more generally a set of data "tensors") and a dimension k, find the k-dimensional subspace that "best" approximates the given data set. There are various notions of "best"; the traditional one used for classical PCA is either minimum Frobenius norm or minimum spectral norm of the approximation error of the data matrix. PCA, without constraints, and for clean data, is a solved problem. By the Eckart–Young–Mirsky theorem, computing the top k left singular vectors of the data matrix returns the PCA solution. On the other hand, robust PCA, which refers to the problem of PCA in the presence of outliers, is a much harder problem and one for which provably correct solutions have started appearing only recently. The same is true for dynamic PCA (subspace tracking or streaming PCA), dynamic or recursive robust PCA (robust subspace tracking), PCA and subspace tracking with missing data, and the related low-rank matrix completion problem, as well as for sparse PCA. Sparse PCA refers to the PCA problem when the principal components are assumed to be sparse. In fact, even the classical PCA problem with speed or memory constraints is not well understood.

The above issues have become particularly important for modern data sets because 1) the data matrix is often so large that it cannot be directly stored in the computer's memory (need for streaming solutions); 2) a lot of data consist of missing entries and/or outlier-corrupted entries (need for matrix completion and robust versions of PCA and subspace recovery);

Digital Object Identifier 10.1109/JPROC.2018.2853498

0018-9219 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

3) a lot of data arrive sequentially, the data subspace itself may change over time, the entire data set cannot be stored but short batches can be, and decisions are often needed in real time or near real time (need for dynamic PCA and robust PCA and subspace tracking); 4) data are often distributed and stored over multiple locations and one needs to perform PCA without communicating all the data to a central location (need for distributed PCA); and 5) many types of data are better represented as a tensor data set rather than a vector data set or matrix (need for tensor PCA).

PCA and its many extensions are used everywhere as summarized above. Moreover, PCA is also a key intermediate or initialization step in solving many convex and nonconvex optimization problems. All areas of electrical engineering and computer science (EECS) have benefited hugely from, and have contributed significantly to, solutions of PCA and extensions. Most people in EECS know certain aspects of PCA, not all, and this special issue helps bridge the gap in knowledge for EECS researchers from various backgrounds.

II. OVERVIEW OF PAPERS IN THIS ISSUE

A total of nine papers in this issue can be split into two broad categories. Each paper explains its specific problem setting and the applications where the problem occurs, the various solution approaches (algorithms), experimental comparisons (either from existing work or new ones), and theoretical guarantees (where they exist).

- PCA and sparse PCA.
  1) "PCA in high dimensions: An orientation," by Johnstone and Paul, provides a broad overview of the key phenomena associated with high-dimensional PCA. Its primary focus is on asymptotic results for the closeness of eigenvalues and eigenvectors of the sample covariance matrices to those of the population covariance matrix. However, it also reviews the more recent work on finite sample (nonasymptotic) guarantees for PCA, both under the spiked covariance model as well as under other more general data and noise models.
  2) "Streaming PCA and subspace tracking: The missing data case," by Balzano *et al.*, reviews both classical and recent algorithms, together with their performance guarantees, for solving the PCA problem in an online fashion under memory and computation constraints. This problem is typically referred to as "streaming PCA" or "subspace tracking." The former usually imposes strict memory constraints while the latter implies the need for online algorithms. This overview especially emphasizes algorithms that can handle missing data. As noted above, data with missing entries are ubiquitous in modern data science problems.
  3) "A selective overview of sparse principal component analysis," by Zou and Xue, provides a selective overview of methodological and theoretical developments of sparse PCA that produce principal components that are sparse, i.e., have only a few nonzero entries. Applications in scientific domains are also discussed.
  4) "A review of distributed algorithms for principal component analysis," by Wu *et al.*, discusses distributed PCA algorithms that are amenable when data are distributively acquired without communicating and accessing the entire data set locally. Distributed PCA algorithms effectively exploit local communications and network connectivity to overcome the conventional need of forming a sample covariance.
  5) "Extension of PCA to higher order data structures: An introduction to tensors, tensor decompositions, and tensor PCA," by Zare *et al.*, reviews the extension of PCA to tensors, which are multiway data that find important applications in many domains. This review paper presents tensor methods that aim to solve important challenges typically addressed by PCA such as dimensionality reduction, supervised learning, and robust low-rank tensor recovery.
- Robust PCA and subspace recovery: PCA with corrupted or missing data.
  1) "Static and dynamic robust PCA and matrix completion: A review," by Vaswani and Narayanamurthy, provides an exhaustive overview of the literature on robust PCA, with an emphasis on provably correct methods. Here "robust PCA" refers to PCA that is robust to "elementwise" outliers in the data vectors. It is thus equivalent to the problem of decomposing a given observed data matrix into the sum of a low-rank matrix (true data) and a sparse matrix (outliers). This article also reviews the more recent literature on dynamic/recursive/online robust PCA, often called "robust subspace tracking," and provides a quick overview of low-rank matrix completion or PCA with missing data (with a focus on recent approaches).
  2) "An overview of robust subspace recovery," by Lerman and Maunu, overviews the entire body of work on robust subspace recovery. Unlike "robust PCA" defined above, "robust subspace recovery" involves finding the low-dimensional subspace that best approximates a given data set in settings where the data set is corrupted by "columnwise" outlier vectors, i.e., an entire data vector is either an outlier or an inlier. While this problem is easy to state, it has been difficult to develop optimal algorithms for solving it due to its underlying nonconvexity.

This review emphasizes the advantages and disadvantages of the various proposed approaches on this topic and also discusses unsolved problems in the area.

3) "Efficient optimization algorithms for robust principal component analysis and its variants," by Ma and Aybat, reviews specialized efficient optimization algorithms that have been developed to solve convex relaxations of various optimization programs that can be defined to solve robust PCA and related problems. The pros and cons of the various methods as well as their convergence properties are also discussed.
4) "On the applications of robust PCA in image and video processing," by Bouwmans *et al.* surveys the applications of RPCA in computer vision and biomedical imaging by reviewing representative image processing applications (low-level imaging, biomedical imaging, 3-D computer vision), and video processing applications such as background/foreground separation. Both solutions for standard robust PCA as well as those that use extra image and video processing problem-specific constraints are described along with a discussion of their applicability to various imaging problems.

Acknowledgments

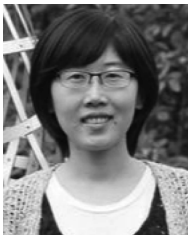
The guest editors would like to thank all the contributors for their hard work, exciting articles, and prompt responses to revision requests. They would also like to thank Prof. Joel H. Trussell and the entire Proceedings of the IEEE Editorial Board for their helpful feedback and suggestions on the contents of this special issue. Finally, special thanks are due to Vaishali Damle, Managing Editor, Proceedings of the IEEE and IEEE Press and Jo Sun, Senior Publications Editor, Proceedings of the IEEE, for their guidance, support,and immediate responses to all requests.

About the Authors

Namrata Vaswani received the B.Tech. degree from Indian Institute of Technology (IIT-Delhi), Delhi, India, in 1999 and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2004.

Currently, she is a Professor of Electrical and Computer Engineering, and (by courtesy) of Mathematics, at Iowa State University, Ames, IA, USA. Her research interests lie at the intersection of statistical machine learning and data science, computer vision, and signal processing. Her recent research has been on provable online algorithms for various dynamic structured high-dimensional (big) data recovery problems such as dynamic compressive sensing (CS), dynamic robust principal component analysis (RPCA), and more recently, phase retrieval.

Prof. Vaswani is an Area Editor for IEEE Signal Processing Magazine, has served twice as an Associate Editor for the IEEE Transactions on Signal Processing, and is the Lead Guest Editor for the Proceedings of the IEEE Special Issue on Rethinking PCA for Modern Data Sets. She is also the Chair of the Women in Signal Processing (WiSP) Committee and a steering committee member of SPS's Data Science Initiative. In 2014, she received the IEEE Signal Processing Society (SPS) Best Paper Award for her Modified-CS work that was coauthored with her graduate student Lu in the IEEE Transactions on Signal Processing in 2010.



Yuejie Chi (Senior Member, IEEE) received the B.E. (honors) degree in electrical engineering from Tsinghua University, Beijing, China, in 2007 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2012.

Since January 2018, she has been an Associate Professor in the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, where she holds the Robert E. Doherty Early Career Professorship. Her research interests include signal processing, machine learning, large-scale optimization, and their applications in data science, inverse problems, imaging, and sensing systems.

Prof. Chi is the recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2013 and the Best Paper Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2012. She received the National Science Foundation (NSF) CAREER Award in 2017, Young Investigator Program Awards from the Air Force Office of Scientific Research (AFOSR) and the U.S. Office of Naval Research (ONR) in 2015, Ralph E. Powe Junior Faculty Enhancement Award from Oak Ridge Associated Universities in 2014, Google Faculty Research Award in 2013, and Roberto Padovani scholarship from Qualcomm Inc. in 2010. She has been an Elected Member of the MLSP and SPTM Technical Committees of the IEEE Signal Processing Society since January 2016.

Thierry Bouwmans received the Diploma of HDR for full professor position in 2014.

He has been an Associate Professor at the University of La Rochelle, La Rochelle, France, since 2000. His research interests consist mainly in the detection of moving objects in challenging environments. He has coauthored two books: *Background Modeling and Foreground Detection for Video Surveillance* (Boca Raton, FL, USA: CRC Press, 2014) and *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image Video Processing* (Boca Raton, FL, USA: CRC Press, 2016). His research investigated particularly the use of robust PCA in video surveillance.

Prof. Bouwmans is the main organizer of the Workshop on Robust Subspace Learning and Computer Vision (RSL-CV) hosted at the 2015 and 2017 International Conference on Computer Vision (ICCV). He is a reviewer for international journals including the IEEE Transactions on Image Processing, the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, the *International Journal of Computer Vision*, *Machine Vision and Applications*, *Computer Vision and Image Understanding*, *Pattern Recognition*, *Pattern Recognition Letters*, and top-level conferences such as the Conference on Computer Vision and Pattern Recognition (CVPR), the International Conference on Pattern Recognition (ICPR), the International Conference on Image Processing (ICIP), the IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS), and more.