Parameterized Complexity of PCA

Fedor V. Fomin
Department of Informatics, University of Bergen, Norway
Fedor.Fomin@uib.no

Petr A. Golovach
Department of Informatics, University of Bergen, Norway
Petr.Golovach@uib.no

Kirill Simonov
Department of Informatics, University of Bergen, Norway
Kirill.Simonov@uib.no

Abstract

We discuss some recent progress in the study of Principal Component Analysis (PCA) from the perspective of Parameterized Complexity.

2012 ACM Subject Classification Theory of computation → Parameterized complexity and exact algorithms

Keywords and phrases parameterized complexity, Robust PCA, outlier detection

Digital Object Identifier 10.4230/LIPIcs.SWAT.2020.1

Category Invited Talk

Funding This work is supported by the Research Council of Norway via the project "MULTIVAL".

1 Introduction

Worst-case running time analysis has been at the center of nearly all developments in theoretical computer science since the inception of the field. Nevertheless, the worst-case approach to measure algorithm efficiency has a serious drawback: For many fundamental problems it does not provide a reasonable explanation why in real life situations these problems are efficiently solvable. The dramatic gap between theory and practice calls for a more nuanced approach, beyond the worst-case case algorithmic analysis. The forthcoming book edited by Tim Roughgarden [23] provides a comprehensive introduction to this emerging area of algorithms.

A particularly successful attempt of building a mathematical model improving over worst-case analysis for *NP-hard* problems is the field of *parameterized complexity.* Originating in the late 80s from the foundational work of Downey and Fellows [7], parameterized complexity has experienced tremendous growth, and is now considered to be one of the central subfields of theoretical computer science, with several textbooks [8, 9, 11, 21], including the most recent book on parameterized algorithms [5] and kernelization [13].

However, so far the mainstream of parameterized complexity was devoted to the study of with NP-hard optimization problems, mostly on graphs and networks. In this talk we want to discuss the applicability of parameterized complexity to the problems involving data point, vectors and matrices.

17th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2020).
Editor: Susanne Albers; Article No. 1; pp. 1:1–1:5
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Robust PCA

Classical *principal component analysis* (PCA) is one of the most popular and successful techniques used for dimension reduction in data analysis and machine learning [22, 19, 10]. In PCA one seeks the best low-rank approximation of data matrix M by solving

$$\begin{aligned} \text{minimize } & \|M - L\|_F^2 \\ \text{subject to } & \text{rank}(L) \le r. \end{aligned}$$

Here $||A||_F^2 = \sum_{i,j} a_{ij}^2$ is the square of the Frobenius norm of matrix A. By the Eckart-Young theorem [10], PCA is efficiently solvable via Singular Value Decomposition (SVD). PCA is used as a preprocessing step in a great variety of modern applications including face recognition, data classification, and analysis of social networks. A well-documented drawback of PCA is its vulnerability to noise. Even when a small number of observations is corrupted, like a few elements or columns of matrix M are changed, PCA of M may not reveal any reasonable information about non-corrupted observations.

There is a large class of extensively studied various robust PCA problems, see e.g. [26, 28, 2]. In the robust PCA setting we observe a noisy version M of data matrix L whose principal components we have to discover. In the case when M is a "slightly" disturbed version of L, PCA performed on M provides a reasonable approximation for L. However, when M is very "noisy" version of L, like being corrupted by a few outliers, even one corrupted outlier can arbitrarily alter the quality of the approximation. Unfortunately, almost every natural mathematical model of robust PCA leads to an NP-hard computational problem, and hence computationally intractable from the perspective of the classical worst-case analysis.

One of the popular approaches to robust PCA, is to model outliers as additive sparse matrix. Thus we have a data $d \times n$ matrix M, which is the superposition of a low-rank component L and a sparse component S. That is, $M = L + S$. This approach became popular after the works of Candès et al. [3], Wright et al. [27], and Chandrasekaran et al. [4]. A significant body of work on the robust PCA problem has been centered around proving that, under some feasibility assumptions on M, L, and S, a solution to

$$\begin{aligned} \text{minimize} \quad & \text{rank}(L) + \lambda\|S\|_0 \\ \text{subject to} \quad & M = L + S, \end{aligned} \tag{1}$$

where $\|S\|_0$ denotes the number of non-zero entries in matrix S and λ is a regularizing parameter, recovers matrix L uniquely. While optimization problem (1) is NP-hard [15], it is possible to show that under certain assumptions on L and S, its convex relaxation can recover these matrices efficiently.

The problem strongly related to (1) was studied in computational complexity under the name Matrix Rigidity [16, 17, 25]. Here, for a given matrix M, and integers r and k, the task is to decide whether at most k entries of M can be changes so that the rank of the resulting matrix is at most r. Equivalently, this is the problem to decide whether a given matrix $M = L + S$, where $\text{rank}(L) \le r$ and $\|S\|_0 \le k$. Thus we define the following problem.

Robust PCA

Input:	Data matrix $M \in \mathbb{R}^{n \times d}$, integer parameters r and k.
Task:	Decide whether there are $L, S \in \mathbb{R}^{n \times d}$, $\text{rank}(L) \le r$ and $\|S\|_0 \le k$, such that $M = L + S$.

We first look at Robust PCA from the perspective of parameterized complexity and discuss when the problem is tractable and when it is not.

▶ **Theorem 1** ([12]). *Robust PCA is solvable in time* $2^{\mathcal{O}(r\cdot k\cdot \log(r\cdot k))}\cdot (nd)^{\mathcal{O}(1)}$.

The proof of the theorem requires ideas from kernelization, linear algebra and algebraic geometry. Thus Robust PCA is fixed-parameter tractable when parameterized by $k+d$. It is also worth to note that the theorem is tight in the following sense: The problem is NP-hard for every $r \geq 1$ [14, 6] and is W[1]-hard parameterized by k [12].

3 PCA with Outliers

Another popular variant of robust PCA is PCA with outliers. Suppose that we have n points (observations) in d-dimensional space. We know that a part of the points are arbitrarily located (say, produced by corrupted observations) while the remaining points are close to an r-dimensional true subspace. We do not have any information about the true subspace and about the corrupted observations. Our task is to learn the true subspace and to identify the outliers. As a common practice, we collect the points into $n \times d$ matrix M, thus each of the rows of M is a point and the columns of M are the coordinates.

Xu et al. [28] introduced the following idealization of this problem.

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(L) + \lambda \|S\|_{0,r} \\ \text{subject to} \quad & M = L + S. \end{aligned} \tag{2}$$

Here $\|S\|_{0,r}$ denotes the number of non-zero raws in matrix S and λ is a regularizing parameter. Xu et al. [28] approached this problem by building its convex surrogate and applying efficient convex optimization-based algorithm for the surrogate. A huge body of work exists on a variant of this problem, called Robust Subspace Recovery, see e.g. [20] for a survey. In this problem for the set of given n points in r-dimensional space, the task is to find an r-dimensional subspace containing the maximum number of points. Hardt and Moitra [18] prove non-approximability of the optimization version of Robust Subspace Recovery under Small Set Expansion conjecture.

An approximation variant of (2) and of Robust Subspace Recovery is the following problem. Given n points in $\mathbb{R}^d$, we seek for a set of k points whose removal leaves the remaining $n-k$ points as close as possible to some r-dimensional subspace. Here is the reformulation of the problem in terms of matrices.

PCA with Outliers

Input: Data matrix $M \in \mathbb{R}^{n\times d}$, integer parameters r and k.

Task:

$$\begin{aligned} \text{minimize} \quad & \|M - L - S\|_F^2 \\ \text{subject to} \quad & L, S \in \mathbb{R}^{n\times d}, \\ & \operatorname{rank}(L) \leq r, \text{ and} \\ & S \text{ has at most } k \text{ non-zero rows.} \end{aligned}$$

We will see how the tools from Real Algebraic Geometry [1] can be used to prove the following theorem.

▶ **Theorem 2** ([24]). *Solving PCA with Outliers is reducible to solving* $n^{\mathcal{O}(d^2)}$ *instances of PCA.*

We also discuss some lower bounds for PCA with Outliers.

References

1 Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry (Algorithms and Computation in Mathematics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

2 Thierry Bouwmans, Necdet Serhat Aybat, and El-hadi Zahzah. *Handbook of robust low-rank and sparse matrix decomposition: Applications in image and video processing*. Chapman and Hall/CRC, 2016.

3 Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011. doi:10.1145/1970392.1970395.

4 Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. doi:10.1137/090761793.

5 Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. doi:10.1007/978-3-319-21275-3.

6 Chen Dan, Kristoffer Arnsfelt Hansen, He Jiang, Liwei Wang, and Yuchen Zhou. On low rank approximation of binary matrices. *CoRR*, abs/1511.01699, 2015. arXiv:1511.01699.

7 Rodney G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness. *Proceedings of the 21st Manitoba Conference on Numerical Mathematics and Computing. Congr. Numer.*, 87:161–178, 1992.

8 Rodney G. Downey and Michael R. Fellows. *Parameterized complexity*. Springer-Verlag, New York, 1999.

9 Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer, 2013.

10 Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

11 Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer-Verlag, Berlin, 2006.

12 Fedor V. Fomin, Daniel Lokshtanov, Syed Mohammad Meesum, Saket Saurabh, and Meirav Zehavi. Matrix rigidity from the viewpoint of parameterized complexity. *SIAM J. Discrete Math.*, 32(2):966–985, 2018. doi:10.1137/17M112258X.

13 Fedor V. Fomin, Daniel Lokshtanov, Saket Saurabh, and Meirav Zehavi. *Kernelization. Theory of Parameterized Preprocessing*. Cambridge University Press, 2019.

14 Nicolas Gillis and Stephen A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *CoRR*, abs/1509.09236, 2015. URL: http://arxiv.org/abs/1509.09236, arXiv:1509.09236.

15 Nicolas Gillis and Stephen A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *Math. Oper. Res.*, 43(4):1072–1084, 2018. doi:10.1287/moor.2017.0895.

16 Dmitry Grigoriev. Using the notions of separability and independence for proving the lower bounds on the circuit complexity (in russian). *Notes of the Leningrad branch of the Steklov Mathematical Institute, Nauka*, 1976.

17 Dmitry Grigoriev. Using the notions of separability and independence for proving the lower bounds on the circuit complexity. *Journal of Soviet Math.*, 14(5):1450–1456, 1980.

18 Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, volume 30 of *JMLR Proceedings*, pages 354–375. JMLR.org, 2013.

19 Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

20 Gilad Lerman and Tyler Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018.

21 Rolf Niedermeier. *Invitation to fixed-parameter algorithms*, volume 31 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, 2006.

22 Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

23 Tim Roughgarden, editor. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, 2020.

24 Kirill Simonov, Fedor V. Fomin, Petr A. Golovach, and Fahad Panolan. Refined complexity of PCA with outliers. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, volume 97, pages 5818–5826. PMLR, 2019. URL: `http://proceedings.mlr.press/v97/simonov19a.html`.

25 L G Valiant. Graph-theoretic arguments in low-level complexity. In *MFCS*, pages 162–176, 1977.

26 Namrata Vaswani and Praneeth Narayanamurthy. Static and dynamic robust PCA and matrix completion: A review. *Proceedings of the IEEE*, 106(8):1359–1379, 2018. `doi:10.1109/JPROC.2018.2844126`.

27 John Wright, Arvind Ganesh, Shankar R. Rao, YiGang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Proceedings of 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2080–2088. Curran Associates, Inc., 2009. URL: `http://papers.nips.cc/paper/3704-robust-principal-component-analysis-exact-recovery-of-corrupted-low-rank-matrices-via-convex-optimization`.

28 Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2496–2504. Curran Associates, Inc., 2010. URL: `http://papers.nips.cc/paper/4005-robust-pca-via-outlier-pursuit`.