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.
Parameterized Complexity of PCA

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

$$\begin{align*}
\text{minimize} & \quad \|M - L\|_F^2 \\
\text{subject to} & \quad \text{rank}(L) \leq r.
\end{align*}$$

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

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

where $\|S\|_0$ denotes the number of non-zero entries in matrix $S$ and $\lambda$ 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 $\text{rank}(L) \leq r$ and $\|S\|_0 \leq k$. 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 $L, S \in \mathbb{R}^{n \times d}$, $\text{rank}(L) \leq r$ and $\|S\|_0 \leq k$, 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 $2^{O(r \cdot k \cdot \log(r \cdot k))} \cdot (nd)^{O(1)}$.

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{align}
\text{minimize} & \quad \text{rank}(L) + \lambda \|S\|_{0,r} \\
\text{subject to} & \quad M = L + S.
\end{align}

Here $\|S\|_{0,r}$ denotes the number of non-zero rows in matrix $S$ and $\lambda$ 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.

\begin{tabular}{|l|}
\hline
\textbf{PCA with Outliers} \\
\hline
\textbf{Input:} & Data matrix $M \in \mathbb{R}^{n \times d}$, integer parameters $r$ and $k$. \\
\textbf{Task:} & minimize $\|M - L - S\|_F^2$ \\
& subject to $L, S \in \mathbb{R}^{n \times d}$, \\
& $\text{rank}(L) \leq r$, and \\
& $S$ has at most $k$ non-zero rows. \\
\hline
\end{tabular}

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 $n^{O(d^2)}$ instances of PCA.

We also discuss some lower bounds for PCA with Outliers.
1:4 Parameterized Complexity of PCA

References


F. V. Fomin, P. A. Golovach, and K. Simonov


