Refined Complexity of PCA with Outliers

Fedor Fomin [*1] **Petr Golovach** [*1] **Fahad Panolan** [*1] **Kirill Simonov** [*1]

Abstract

Principal component analysis (PCA) is one of the most fundamental procedures in exploratory data analysis and is the basic step in applications ranging from quantitative finance and bioinformatics to image analysis and neuroscience. However, it is well-documented that the applicability of PCA in many real scenarios could be constrained by an "immune deficiency" to outliers such as corrupted observations. We consider the following algorithmic question about the PCA with outliers. For a set of n points in $\mathbb{R}^d$, how to learn a subset of points, say 1% of the total number of points, such that the remaining part of the points is best fit into some unknown r-dimensional subspace? We provide a rigorous algorithmic analysis of the problem. We show that the problem is solvable in time $n^{\mathcal{O}(d^2)}$. In particular, for constant dimension the problem is solvable in polynomial time. We complement the algorithmic result by the lower bound, showing that unless Exponential Time Hypothesis fails, in time $f(d)n^{o(d)}$, for any function f of d, it is impossible not only to solve the problem exactly but even to approximate it within a constant factor.

1. Introduction

Problem statement and motivation. Classical *principal component analysis* (PCA) is one of the most popular and successful techniques used for dimension reduction in data analysis and machine learning (Pearson, 1901; Hotelling, 1933; Eckart & Young, 1936). In PCA one seeks the best low-rank approximation of data matrix M by solving

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

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 (Eckart & Young, 1936), 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.

In this paper we consider a variant of PCA with outliers, where we wish to recover a low-rank matrix from large but sparse errors. 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. However, it is very likely that PCA of M will not reveal any reasonable information about non-corrupted observations—well-documented drawback of PCA is its vulnerability to even very small number of outliers, an example is shown in Figure 1.

Matrix formulation suggests the following interpretation: we seek a low-rank matrix L that, with an exception in few rows, approximates M best.

[*]Equal contribution [1]Department of Informatics, University of Bergen, Norway. Correspondence to: Fedor Fomin <fomin@ii.uib.no>, Petr Golovach <pgo041@uib.no>, Fahad Panolan <Fahad.Panolan@uib.no>, Kirill Simonov <Kirill.Simonov@uib.no>.

Proceedings of the 36^th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

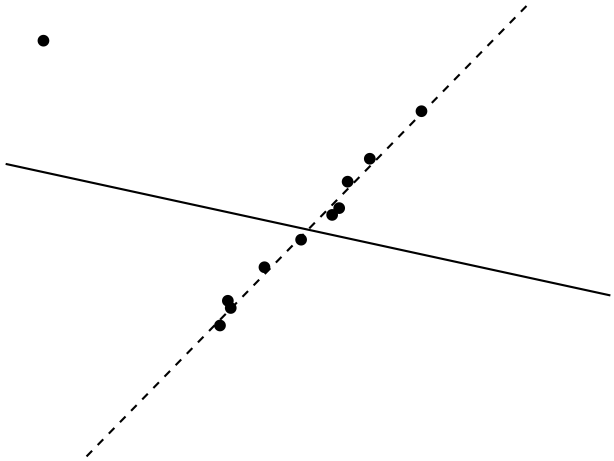

Figure 1. An illustration on how outliers impact PCA. The optimal approximation line (in dashed) of the given set of points without the evident outlier shows the linear structure of the dataset. However, when the outlier is present, the principal component (in solid) changes drastically.

PCA WITH OUTLIERS

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

Task:

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

The geometric interpretation of PCA WITH OUTLIERS is very natural: 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.

Related work. PCA WITH OUTLIERS belongs to the large class of extensively studied robust PCA problems, see e.g. (Vaswani & Narayanamurthy, 2018; Xu et al., 2010; Bouwmans et al., 2016). 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.

One of the approaches to robust PCA, which is relevant to our work, 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. (Candès et al., 2011), Wright et al. (Wright et al., 2009), and Chandrasekaran et al. (Chandrasekaran et al., 2011). 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

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

where $\|S\|_0$ denotes the number of non-zero columns in matrix S and λ is a regularizing parameter, recovers matrix L uniquely. While optimization problem (1) is NP-hard (Gillis & Vavasis, 2018), 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 (Grigoriev, 1976; 1980; Valiant, 1977). 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 $\operatorname{rank}(L) \le r$ and $\|S\|_0 \le k$. Fomin et al. (Fomin et al., 2018) gave an algorithm solving MATRIX RIGIITY in time $2^{\mathcal{O}(r \cdot k \cdot \log(r \cdot k))} \cdot (nd)^{\mathcal{O}(1)}$. On the other hand, they show that the problem is W[1]-hard parameterized by k. In particular, this implies that an algorithm of running time $f(k) \cdot (rnd)^{\mathcal{O}(1)}$ for this problem is highly unlikely for any function f of k only.

A natural extension of the robust PCA approach (1) is to consider the noisy version of robust PCA: Given $M = L + S + N$, where L, S, and N are unknown, but L is known to be low rank, S is known to have a few non-zero rows, and noise matrix N is of small Frobenius norm, recover L. Wright et al. (Wright et al., 2009) studied the following model of noisy robust PCA:

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(L) + \lambda \|S\|_0 \\ \text{subject to} \quad & \|M - L - S\|_F^2 \le \varepsilon. \end{aligned} \tag{2}$$

Thus (2) models the situations when we want to learn the principal components of n points in d-dimensional space under the assumption that a small number of coordinates is corrupted.

The study of the natural, and seemingly more difficult extension of (1) to the PCA with outliers, was initiated by Xu et al. (Xu et al., 2010), who introduced the following idealization of the problem.

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(L) + \lambda \|S\|_{0,r} \\ \text{subject to} \quad & M = L + S. \end{aligned} \tag{3}$$

Here $\|S\|_{0,r}$ denotes the number of non-zero columns in matrix S and λ is a regularizing parameter. Xu et al. (Xu et al., 2010) approached this problem by building its convex surrogate and applying efficient convex optimization-based algorithm for the surrogate. Chen et al. (Chen et al., 2011) studied the variant of the problem with the partially observed data. Similar as (2) is the noisy version of the robust PCA model (1), the PCA WITH OUTLIERS problem studied in our work can be seen as a noisy version of (3).

Our results. Even though PCA WITH OUTLIERS was assumed to be NP-hard, to the best of our knowledge, this has never been studied formally. While NP-hardness is a serious strike against the tractability of the problem, on the other hand, it only says that in the worst case the problem is not tractable. But since the complexity of the problem could be governed by several parameters like the rank r of L, the number of outliers k or dimension d of M, it is natural to ask how these parameters influence the complexity of the problem. For example, when k is a small constant, we can guess which points are outliers and run PCA for the remaining points. This will bring us to n^k calls of PCA which is polynomial for constant k and is exponential when k is a fraction of n.

In this paper we give an algorithm solving PCA WITH OUTLIERS roughly in time $|M|^{\mathcal{O}(d^2)}$, where $|M|$ is the size of the input matrix M. Thus for fixed dimension d, the problem is solvable in polynomial time. The algorithms works in polynomial time for any number of outliers k and the rank r of the recovered matrix L. Our algorithm strongly relies on the tools developed in computational algebraic geometry, in particular, for handling arrangements of algebraic surfaces in $\mathbb{R}^d$ defined by polynomials of bounded degree.

We complement our algorithmic result by a complexity lower bound. Our lower bound not only implies that the problem is NP-hard when dimension d is part of the input, it also rules out a possibility of certain type of algorithms for PCA WITH OUTLIERS. More precisely, assuming the Exponential Time Hypothesis (ETH),[1] we show that for any constant $\omega \geq 1$, PCA WITH OUTLIERS cannot be ω-approximated in time $f(d)|M|^{o(d)}$, for any function f of d only.

Our algorithm is, foremost, of theoretical interest, especially in the presense of the nearly-matching lower bound showing that doing something essentially better is next to impossible. In practice, PCA is often applied to reduce high-dimensional datasets, and for this task the running time exponential in d is not practical. However, there are cases where such an algorithm could be useful. One example could be the visualization of low-dimensional data, where the number of dimensions, even if it is small already, needs to be lowered down to two to actually draw the dataset. Another example could be when we suspect a small subset of features to be highly correlated, and we want to reduce them to one dimension in order to get rid of the redundancy in data. This potential application is well illustrated by the popular PCA tutorial (Shlens, 2014), where essentially one-dimensional movement of a spring-mass is captured by three cameras, resulting in 6 features.

[1]ETH of Impagliazzo, Paturi, and Zane (Impagliazzo et al., 2001) is that 3-SAT with n-variables is not solvable in time $2^{o(n)}$.

2. Polynomial algorithm for bounded dimension

2.1. Preliminaries

As a subroutine in our algorihm, we use a standard result about sampling points from cells of an arrangement of algebraic surfaces, so first we state some definitions and an algorithm from (Basu et al., 2006).

We denote the ring of polynomials in variables $X_1, \ldots, X_d$ with coefficients in $\mathbb{R}$ by $\mathbb{R}[X_1, \ldots, X_d]$. By saying that an algebraic set V in $\mathbb{R}^d$ is defined by $Q \in \mathbb{R}[X_1, \ldots, X_d]$, we mean that $V = \{x \in \mathbb{R}^d | Q(x_1, \ldots, x_d) = 0\}$. For a set of s polynomials $\mathcal{P} = \{P_1, \ldots, P_s\} \subset \mathbb{R}[X_1, \ldots, X_d]$, a sign condition is specified by a sign vector $\sigma \in \{-1, 0, +1\}^s$, and the sign condition is non-empty over V with respect to $\mathcal{P}$ if there is a point $x \in V$ such that

$$\sigma = (\mathrm{sign}(P_1(x)), \ldots, \mathrm{sign}(P_s(x))),$$

where $\mathrm{sign}(x)$ is the sign function on real numbers defined as

$$\mathrm{sign}(x) = \begin{cases} 1, \text{ if } x > 0, \\ 0, \text{ if } x = 0, \\ -1, \text{ if } x < 0 \end{cases}$$

for $x \in \mathbb{R}$.

The *realization space* of $\sigma \in \{-1, 0, +1\}^s$ over V is the set

$$R(\sigma) = \{x | x \in V, \sigma = (\mathrm{sign}(P_1(x)), \ldots, \mathrm{sign}(P_s(x)))\}.$$

If $R(\sigma)$ is not empty then each of its non-empty semi-algebrically connected (which is equivalent to just connected on semi-algebraic sets as proven in (Basu et al., 2006), Theorem 5.22) components is a *cell* of $\mathcal{P}$ over V.

For an algebraic set W its real dimension is the maximal integer d' such that there is a homeomorphism of $[0, 1]^{d'}$ in W. Naturally, if $W \subset \mathbb{R}^d$, then $d' \leq d$.

The following theorem from (Basu et al., 2006) gives an algorithm to compute a point in each cell of $\mathcal{P}$ over V.

Proposition 1 ((Basu et al., 2006), Theorem 13.22). *Let V be an algebraic set in $\mathbb{R}^d$ of real dimension d' defined by $Q(X_1, \ldots, X_d) = 0$, where Q is a polynomial in $\mathbb{R}[X_1, \ldots, X_d]$ of degree at most b, and let $\mathcal{P} \subset$*

$\mathbb{R}[X_1, \ldots, X_d]$ be a finite set of s polynomials with each $P \in \mathcal{P}$ also of degree at most b. Let D be a ring generated by the coefficients of Q and the polynomials in $\mathcal{P}$. There is an algorithm which takes as input Q, d' and $\mathcal{P}$ and computes a set of points meeting every non-empty cell of V over $\mathcal{P}$. The algorithm uses at most $s^{d'} b^{O(d)}$ arithmetic operations in D.

On the practical side, we note that a number of routines from (Basu et al., 2006) is implemented in the SARAG library (Caruso, 2006).

2.2. Algorithm

First, we emphasize on a folklore observation that geometrically the low-rank approximation matrix L is defined as orthogonal projection of rows of M on some r-dimensional subspace of $\mathbb{R}^d$. For the proof see e.g. (Blum et al., 2017).

Proposition 2. *Given a matrix $M \in \mathbb{R}^{n\times d}$ with rows m_1, $\ldots$, m_n, the task of finding a matrix L of rank at most r which minimizes $||M - L||_F^2$ is equivalent to finding an r-dimensional subspace of $\mathbb{R}^d$ which minimizes the total squared distance from rows of M treated as points in $\mathbb{R}^d$:*

$$\min_{\substack{L \in \mathbb{R}^{n\times d} \\ \operatorname{rank} L \le r}} ||M - L||_F^2 = \min_{\substack{U \subset \mathbb{R}^d \\ U \text{ is a linear subspace of } \dim r}} \sum_{i=1}^{n} ||m_i - \operatorname{proj}_U m_i||_F^2,$$

where $\operatorname{proj}_U x$ is the orthogonal projection of x on U for $x \in \mathbb{R}^d$.

By Proposition 2, if we fix an r-dimensional subspace U containing the span of rows of L, then the outliers are automatically defined as k farthest points from U among $\{m_i\}_{i=1}^n$. In the next proposition, we give a precise statement of this.

Proposition 3. *The optimization objective of* PCA WITH OUTLIERS *for a given matrix $M \in \mathbb{R}^{n\times d}$ with rows m_1, $\ldots$, m_n can be equivalently redefined as follows.*

$$\min_{\substack{L, S \in \mathbb{R}^{n\times d} \\ \operatorname{rank} L \le r \\ S \text{ has at most } k \text{ non-zero rows}}} ||M - L - S||_F^2 = \min_{\substack{U \subset \mathbb{R}^d \\ U \text{ is a linear subspace of } \dim r}} ||M - L_U - S_U||_F^2,$$

where S_U has k non-zero rows which are k rows of M with the largest value of $||m_i - \operatorname{proj}_U m_i||_F^2$, and L_U is the orthogonal projection of the rows of $(M - S_U)$ on U.

$$S_U = \begin{pmatrix} m_1 \\ \vdots \\ m_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad L_U = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \operatorname{proj}_U m_{k+1} \\ \vdots \\ \operatorname{proj}_U m_n \end{pmatrix},$$

assuming that rows of M are ordered by descending $||m_i - \operatorname{proj}_U m_i||_F^2$.

So for a fixed U we may determine S_U easily and then solve the classical PCA for the matrix $(M - S_U)$. The intuition behind our algorithm is that the set of k farthest points is the same for many subspaces, and solving PCA for $(M - S_U)$ treats all these subspaces. The crucial point is to bound the number of different matrices S_U we have to consider. There is of course a trivial bound of n^k since S_U is always obtained by choosing k rows of M. But the number of different S_U is also geometrically limited, and exploiting this we are able to obtain another bound of $n^{O(d^2)}$, resulting in the following theorem.

Theorem 1. *Solving* PCA WITH OUTLIERS *is reducible to solving*

$$\binom{n}{2}^{\min(rd,(d-r)d)} 2^{\mathcal{O}(d)} = n^{\mathcal{O}(d^2)}$$

instances of PCA. This reduction can be computed in the number of operations over $\mathbb{R}$ bounded by the expression above.

First, a note about the statement of Theorem 1. Our algorithm relies on solving the classical PCA, and since only iterative algorithms for PCA and SVD exist, we could not claim that our algorithm solves PCA WITH OUTLIERS in some fixed number of operations. However, if we are only interested in solving the problem up to some constant precision, for example machine epsilon, then PCA is solvable in polynomial number of operations and so by Theorem 1, PCA WITH OUTLIERS is solvable in $n^{\mathcal{O}(d^2)}$ operations.

Proof of Theorem 1. We start with associating r-dimensional subspaces of $\mathbb{R}^d$ with points of a certain algebraic set. Consider the matrix space $\mathbb{R}^{(d-r)\times d}$, and for an element $V \in \mathbb{R}^{(d-r)\times d}$, $V = \{v_{ij}\}_{i,j}$, the following polynomial conditions:

$$Q_{i,j}^O(V) := \sum_{l=1}^{d} v_{il} v_{jl} = 0, \text{ for } 1 \le i < j \le (d-r),$$

$$Q_i^N(V) := \left(\sum_{l=1}^{d} v_{il}^2\right) - 1 = 0, \text{ for } 1 \le i \le (d-r),$$

where condition $Q^O_{i,j}(V) = 0$ requires rows i, j of V to be pairwise orthogonal and condition $Q^N_j(V) = 0$ requires row j of V to have length 1. We may write all these conditions as a single polynomial condition $Q(V) = 0$ by taking the sum of squares:

$$Q(V) = \sum_{1\le i<j\le(d-r)} (Q^O_{i,j}(V))^2 + \sum_{i=1}^{d-r}(Q^N_i(V))^2.$$

Thus $Q(V) = 0$ if and only if each of $Q^O_{i,j}(V)$ and each of $Q^N_i(V)$ is 0.

Consider an algebraic set $W \subset \mathbb{R}^{(d-r)\times d}$ defined as the zero set of $Q(V)$. For any $V \in W$ with rows $v_1, \ldots, v_{d-r}$, consider the r-dimensional subspace $\mathrm{comp}(V) := \mathrm{span}(\{v_1, \cdots, v_{d-r}\})^\perp \subset \mathbb{R}^d$ which is the orthogonal complement of the span of the rows of V. Since $Q(V) = 0$, the rows of V are pairwise orthogonal and are of length 1. Then, the dimension of $\mathrm{comp}(V)$ is r and for any point $x \in \mathbb{R}^d$ the squared distance from x to $\mathrm{comp}(V)$ is equal to

$$\sum_{i=1}^{d-r}(v_i\cdot x)^2 = ||Vx^T||_F^2,$$

assuming that x is a row vector.

Each $V \in W$ defines an r-dimensional subspace $\mathrm{comp}(V) \subset \mathbb{R}^d$ and each r-dimensional subspace $U \subset \mathbb{R}^d$ is of this form for some $V \in W$ since there exists an orthonormal basis of the orthogonal complement of U. Then we can reformulate Proposition 3 in terms of elements of W as follows.

$$\begin{aligned}&\min_{\substack{L,S\in\mathbb{R}^{n\times d}\\ \mathrm{rank}\, L\le r\\ S \text{ has at most } k \text{ non-zero rows}}} ||M-L-S||_F^2\\ &\quad= \min_{V\in W} ||M - S_{\mathrm{comp}(V)} - L_{\mathrm{comp}(V)}||_F^2, \quad (4)\end{aligned}$$

where $S_{\mathrm{comp}(V)}$ and $L_{\mathrm{comp}(V)}$ are defined in accordance with notation in Proposition 3. Let $m_1, \ldots, m_n$ be the rows of the input matrix M; $S_{\mathrm{comp}(V)}$ has k non-zero rows which are k rows of M with the largest value of $||m_i - \mathrm{proj}_{\mathrm{comp}(V)}\, m_i||_F^2 = ||Vm_i^T||_F^2$, and $L_{\mathrm{comp}(V)}$ is the orthogonal projection of the rows of $(M - S_{\mathrm{comp}(V)})$ on $\mathrm{comp}(V)$. Denote $S_{\mathrm{comp}(V)}$ by S_V and $L_{\mathrm{comp}(V)}$ by L_V.

Now, consider the set of polynomials $\mathcal{P} = \{P_{i,j}\}_{1\le i<j\le n}$ defined on W, where

$$P_{i,j}(V) = ||Vm_i^T||_F^2 - ||Vm_j^T||_F^2.$$

Consider the partition $\mathcal{C}$ of W on cells over $\mathcal{P}$. For each cell C, the sign condition with respect to $\mathcal{P}$ is constant over C, meaning that for every pair $1 \le i < j \le n$, the sign of

$$||Vm_i^T||_F^2 - ||Vm_j^T||_F^2$$

is the same for all $V \in C$. So the relative order on $\{||Vm_i^T||_F^2\}_{i=1}^n$ is also the same for all $V \in C$. Since $||Vm_i^T||_F^2$ is exactly the squared distance from m_i to V, k rows of M which are the farthest are also the same for all $V \in C$. Then S_V is constant over $V \in C$, denote this common value as S_C. We can rewrite (4) as

$$\begin{aligned}&\min_{V\in W} ||M - S_V - L_V||_F^2\\ &\quad= \min_{C\in\mathcal{C}}\min_{V\in C} ||M - S_V - L_V||_F^2\\ &\quad= \min_{C\in\mathcal{C}}\min_{V\in C} ||(M - S_C) - L_V||_F^2.\end{aligned}$$

Note that

$$\begin{aligned}&\min_{C\in\mathcal{C}}\min_{V\in C} ||(M - S_C) - L_V||_F^2\\ &\quad= \min_{C\in\mathcal{C}}\min_{V\in W} ||(M - S_C) - L_V||_F^2, \quad (5)\end{aligned}$$

as for any $C \in \mathcal{C}$, $\min_{V\in C} ||(M - S_C) - L_V||_F^2 \ge \min_{V\in W} ||(M - S_C) - L_V||_F^2$ since $C \subset W$. Also, any (S_C, L_V) in the right-hand side of (5) is still a valid choice of (S, L) for the original problem, and the optimum of the original problem is equal to the left-hand side of (5).

For a fixed $C \in \mathcal{C}$ computing right-hand side of (5) is equivalent to solving an instance $(M - S_C, r)$ of the classical PCA by Proposition 2:

$$\begin{aligned}&\min_{V\in W} ||(M - S_C) - L_V||_F^2\\ &\quad= \min_{L\in\mathbb{R}^{n\times d},\, \mathrm{rank}(L)\le r} ||(M - S_C) - L||_F^2.\end{aligned}$$

By the reasoning above, the optimum of the original instance of PCA WITH OUTLIERS is reached on one of the constructed instances $\{(M - S_C, r)\}_{C\in\mathcal{C}}$ of PCA. A toy example of an algebraic set W and its partitioning is shown in Figure 2.

Putting all together, our algorithm proceeds as follows.

1. Using the algorithm from Proposition 1, obtain a point V_C from each cell C of W over $\mathcal{P}$.
2. For each V_C, compute the optimal S_{V_C}: select the k rows of M with the largest value of $||V_C m_i^T||_F^2$. Construct the instance $(M - S_{V_C}, r)$ of PCA.
3. The solution to the original instance of PCA WITH OUTLIERS is the best solution among the solutions of all the constructed PCA instances.

Since degrees of Q and polynomials from $\mathcal{P}$ are at most 4, $|\mathcal{P}| = \binom{n}{2}$, and the real dimension of W is at most $(d-r)d$, which is the dimension of $\mathbb{R}^{(d-r)\times d} \supset W$, the algorithm from Proposition 1 does at most

$$t = \binom{n}{2}^{(d-r)d} 2^{\mathcal{O}(d)}$$

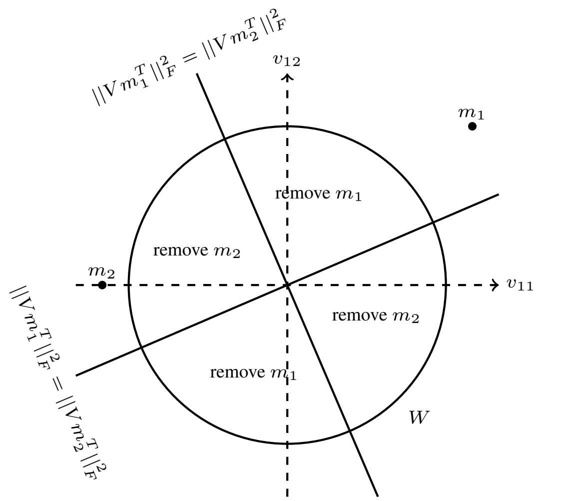

Figure 2. An example with $d = 2$, $n = 2$, $r = 1$, $k = 1$. Two rows of the input matrix M are represented as two points m_1 and m_2 on the plane. The same plane represents the choice of 1-dimensional approximation subspace through the selection of a vector V orthogonal to it. The algebraic set W is the unit circle since length of V must be 1. The diagonal lines mark the values of V for which m_1 and m_2 are equidistant. They split W into four one-dimensional and four zero-dimensional cells. For each of the one-dimensional cells it is shown in the corresponding sector which of the points is the outlier and hence is removed.

operations and produces at most t points V_C, and our algorithm produces one instance of PCA for each computed point.[2]

We are also able to obtain a reduction to

$$\binom{n}{2}^{rd} 2^{\mathcal{O}(d)}$$

instances of PCA by proceeding in the same manner for slightly different characterization of r-dimensional subspaces. Intuitively, now points on the algebraic set define the orthonormal basis of the subspace itself, and not of its orthogonal complement as in the previous part.

Now the matrix space is $\mathbb{R}^{r\times d}$, the conditions that an element $V \in \mathbb{R}^{r\times d}$ defines an orthonormal basis of size r are analogous:

$$\bar{Q}^O_{i,j}(V) := \sum_{l=1}^{d} v_{il}v_{jl} = 0, \text{ for } 1 \le i < j \le r,$$

$$\bar{Q}^N_i(V) := \left(\sum_{l=1}^{d} v_{il}^2\right) - 1 = 0, \text{ for } 1 \le i \le r.$$

Again, we may write them as a single polynomial condition $\bar{Q}(V) = 0$ where

$$\bar{Q}(V) = \sum_{1\le i<j\le r} (\bar{Q}^O_{i,j}(V))^2 + \sum_{i=1}^{r} (\bar{Q}^N_i(V))^2,$$

Consider an algebraic set $\bar{W} \subset \mathbb{R}^{r\times d}$ defined as the set of zeroes of $\bar{Q}(V)$. Similarly, any $V \in \bar{W}$ defines an r-dimensional subspace $U \in \mathbb{R}^d$ which is the span of the rows of V. Since the rows of V form an orthonormal basis of U, for any point $x \in \mathbb{R}^d$ the squared distance from x to U is equal to

$$||x||_F^2 - \sum_{i=1}^{r} (v_i \cdot x)^2 = ||x||_F^2 - ||Vx^T||_F^2.$$

The new distance formula leads to a slightly different set of polynomials $\bar{\mathcal{P}} = \{\bar{P_{i,j}}\}_{1\le i<j\le n}$ on W, comparing the distance from m_i and from m_j,

$$\bar{P_{i,j}}(V) = (||m_i||_F^2 - ||Vm_i^T||_F^2) - (||m_j||_F^2 - ||Vm_j^T||_F^2).$$

Again, the k farthest points and the matrix S_V are the same over any cell in the partition of $\bar{W}$ over $\bar{\mathcal{P}}$. So by the same reasoning as in the first part, it suffices to take a point V from each cell, compute the outlier matrix S_V and solve PCA for $(M - S_V, r)$.

As we can choose the most effecient of the two subspace representations, we can reduce PCA WITH OUTLIERS to

$$\binom{n}{2}^{\min(rd,(d-r)d)} 2^{\mathcal{O}(d)}$$

instances of PCA.[3] □

3. ETH lower bound

In this section we show that we cannot avoid the dependence on d in the exponent of the running time of a constant factor approximation algorithm for PCA WITH OUTLIERS unless *Exponential Time Hypothesis* (ETH) is false. Recall that ETH is the conjecture stated by Impagliazzo, Paturi and Zane (Impagliazzo et al., 2001) in 2001 that for every integer $k \ge 3$, there is a positive constant δ_k such that the k-SATISFIABILITY problem with n variables and m clauses cannot be solved in time $\mathcal{O}(2^{\delta_k n} \cdot (n+m)^{\mathcal{O}(1)})$. This means that k-SATISFIABILITY cannot be solved in subexponential in n time.

[2] As W is restricted by $Q(V) = 0$, its dimension is actually smaller. It could be bounded more precisely as $(d-r)(d+r-1)/2$, but we omit the calculation in order not to unnecessarily complicate the text.

[3] As with W, the dimension of $\bar{W}$ could be bounded more precisely as $r(2d - r - 1)/2$, and with these dimension bounds PCA WITH OUTLIERS reduces to $\binom{n}{2}^{\min(r(2d-r-1)/2,(d-r)(d+r-1)/2)} 2^{\mathcal{O}(d)}$ instances of PCA.

Let (M, r, k) be an instance of PCA WITH OUTLIERS where M is an $n \times d$ matrix. We say that a pair of $n \times d$ matrices (L, S) is a *feasible solution* for (M, r, k) if $\operatorname{rank}(L) \leq r$ and S has at most k non-zero rows. A feasible solution (L^*, S^*) is *optimal* if $\|M - L^* - S^*\|_F^2$ is minimum, and we denote $\mathsf{Opt}(M, r, k) = \|M - L^* - S^*\|_F^2$.

Theorem 2. *For any $\omega \geq 1$, there is no ω-approximation algorithm for* PCA WITH OUTLIERS *with running time $f(d) \cdot N^{o(d)}$ for any computable function f unless ETH fails, where N is the bitsize of the input matrix M.*

Sketch of the proof. Due to space constraints, we only sketch the proof; the details can be found in the full version of the paper (Fomin et al., 2019). Let $\omega \geq 1$. We reduce from the MULTICOLORED CLIQUE problem:

MULTICOLORED CLIQUE parameterized by r

Input: A graph G with a partition $V_1, \ldots, V_r$ of the vertex set.

Task: Decide whether there is a clique X in G with $|X \cap V_i| = 1$.

This problem cannot be solved in time $f(r) \cdot |V(G)|^{o(r)}$ for any computable function f unless ETH fails (Cygan et al., 2015; Lokshtanov et al., 2011)).

Let $(G, V_1, \ldots, V_r)$ be an instance of MULTICOLORED CLIQUE. We assume without loss of generality that for $i \in \{1, \ldots, r\}$, $|V_i| = n$ (otherwise, we can add dummy isolated vertices to insure the property; clearly, the claim that the problem cannot be solved in time $f(r) \cdot |V(G)|^{o(r)}$ up to ETH remains correct) and each V_i is an independent set. We denote $v_1^i, \ldots, v_n^i$ the vertices of V_i for $i \in \{1, \ldots, r\}$ and let $m = |E(G)|$. We also assume that $r \geq 4$.

We set $a = 4(r+1)^2 mn^2 \omega$ and $c = 9a$. We define $m \times r$ matrices $P = (p_{ij})$ and $Q = (q_{ij})$ whose rows are indexed by the edges of G as follows: for every $e = v_i^s v_j^t \in E(G)$,

- set $p_{es} = c(a - i)$, $p_{et} = c(a - j)$,
- set $q_{es} = q_{et} = ca$,
- for $h \in \{1, \ldots, r\}$ such that $h \neq i, j$, set $p_{ei} = q_{ej} = 0$.

For $e \in E(G)$, p_e and q_e denotes the e-th row of P and Q, respectively. We define the $r \times r$ matrix $A = aI_r$. Let $a_1, \ldots, a_r$ be the rows of A. We construct $2m$ copies of A denoted by A_i, A'_i for $i \in \{1, \ldots, m\}$. We construct an $(r+1)m \times 2r$ matrix M using P, Q and A_i, A'_i for $i \in \{1, \ldots, m\}$ as blocks:

$$M = \left(\begin{array}{c|c} A_1 & A'_1 \\ \hline \vdots & \vdots \\ \hline A_m & A'_m \\ \hline P & Q \end{array}\right). \tag{6}$$

To simplify notation, we index the rows of M corresponding to $(P \mid Q)$ by the edges of G and use integers for other indices. We follow the same convention of the matrices of similar structure that are considered further. We define $k = m - \binom{r}{2}$ and set $D = rmn^2$ and $D' = D\omega$. Note that the dimension $d = 2r$ Observe also that a and c are chosen in such a way that $c >> a >> D'$ and this is crucial for our reduction. The main property of the constructed instance of PCA WITH OUTLIERS is stated in the following claim.

Claim 1. *If $(G, V_1, \ldots, V_r)$ is a yes-instance of* MULTICOLORED CLIQUE*, then $\mathsf{Opt}(M, r, k) \leq D$, and if there is a feasible solution (L, S) for (M, r, k) with $\|M - L - S\|_F^2 \leq \omega D$, then $(G, V_1, \ldots, V_r)$ is a yes-instance of* MULTICOLORED CLIQUE.

Suppose that $(G, V_1, \ldots, V_r)$ is a yes-instance of MULTICOLORED CLIQUE, that is, there is a clique X of G such that $|X \cap V_i| = 1$ for $i \in \{1, \ldots, r\}$. Let $X = \{v_{i_1}^1, \ldots, v_{i_r}^r\}$ and $R = \{v_{i_s}^s v_{i_t}^t \mid 1 \leq s < t \leq r\}$, that is, $R = E(G[X])$. We show that $\mathsf{Opt}(M, r, k) \leq D$. For this, we construct a feasible solution (L, S) such that $\|M - L - S\|_F^2 \leq D$.

We define the $m \times r$ matrices P^* and Q^* by setting the elements of P and Q respectively that are in the rows for $e \in E(G) \setminus R$ to be zero and the other elements are the same as in P and Q, that is, for $e \in R$, the rows of P^* and Q^* are p_e and q_e respectively and other rows are zero-rows. We define an $r \times r$ matrix B as follows:

$$B = \begin{pmatrix} a - i_1 & 0 & \ldots & 0 \\ 0 & a - i_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a - i_r \end{pmatrix},$$

that is, the diagonal elements of B are $a - i_1, \ldots, a - i_r$ and the other elements are zeros. Denote by $b_1, \ldots, b_r$ the rows of B. We construct m copies $B_1, \ldots, B_m$ of B and the $(r+1)m \times 2k$ matrix L using P^*, Q^* and B_i, A'_i for $i \in \{1, \ldots, m\}$ as blocks:

$$L = \left(\begin{array}{c|c} B_1 & A'_1 \\ \hline \vdots & \vdots \\ \hline B_m & A'_m \\ \hline P^* & Q^* \end{array}\right). \tag{7}$$

It is straightforward to verify that $\operatorname{rank}(L) \leq r$. Indeed, the rank of the each submatrix $(B_i \mid A'_i)$ is r as only diagonal elements of B_i and A'_i are non-zero. Also for each $e = v^s_{i_s} v^t_{i_t} \in R$, we have that $p_e = c(b_{i_s} + b_{i_t})$ and $q_e = c(a_{i_s} + a_{i_t})$, i.e., each row of L indexed by $e \in R$ is a linear combination of two rows of $(B_1 \mid A'_1)$.

Let $\mathbb{0}$ be $r \times r$ zero matrix. We construct $2m$ copies $\mathbb{0}_1, \ldots, \mathbb{0}_m$ and $\mathbb{0}'_1, \ldots, \mathbb{0}'_m$ of $\mathbb{0}$ and define

$$S = \left(\begin{array}{c|c} \mathbb{0}_1 & \mathbb{0}'_1 \\ \hline \vdots & \vdots \\ \hline \mathbb{0}_m & \mathbb{0}'_m \\ \hline P - P^* & Q - Q^* \end{array}\right). \tag{8}$$

Clearly, this matrix has at most k non-zero rows that are indexed by $e \in R$.

Combining (6)–(8), we have that

$$\begin{aligned} M - S - L \\ &= \left(\begin{array}{c|c} A_1 & A'_1 \\ \hline \vdots & \vdots \\ \hline A_m & A'_m \\ \hline P & Q \end{array}\right) - \left(\begin{array}{c|c} B_1 & A'_1 \\ \hline \vdots & \vdots \\ \hline B_m & A'_m \\ \hline P^* & Q^* \end{array}\right) \\ &- \left(\begin{array}{c|c} \mathbb{0}_1 & \mathbb{0}'_1 \\ \hline \vdots & \vdots \\ \hline \mathbb{0}_m & \mathbb{0}'_m \\ \hline P - P^* & Q - Q^* \end{array}\right) \\ &= \left(\begin{array}{c|c} A_1 - B_1 & \mathbb{0} \\ \hline \vdots & \vdots \\ \hline A_m - B_m & \mathbb{0} \\ \hline \mathbb{0}' & \mathbb{0}' \end{array}\right), \end{aligned}$$

where $\mathbb{0}'$ is the $m \times r$ zero matrix, and

$$\begin{aligned} \|M - S - L\|_F^2 =& m\|A - B\|_F^2 = m(i_1^2 + \ldots + i_r^2) \\ \leq& rmn^2 = D. \end{aligned}$$

We conclude that (L, S) is a feasible solution for the considered instance (M, r, k) of PCA WITH OUTLIERS with $\|M - S - L\|_F^2 \leq D$. Therefore, $\mathsf{Opt}(M, r, k) \leq D$.

Suppose now that (L, S) is a feasible solution for (M, r, k) of PCA WITH OUTLIERS with $\|M - S - L\|_F^2 \leq \omega D = D'$. We prove that $(G, V_1, \ldots, V_r)$ is a yes-instance of MULTICOLORED CLIQUE.

Recall that S has at most $k = m - \binom{r}{2}$ non-zero rows, Hence, there is a set $R \subseteq E(G)$ with $|R| = m - k = \binom{r}{2}$ such that the rows of S indexed by $e \in R$ are zero-rows. We claim that the edges of R form the set of edges of a complete graph. More formally, we show the following.

Claim 2. *There are $i_1, \ldots, i_r \in \{1, \ldots, n\}$ such that $R = \{v^s_{i_s} v^t_{i_t} \mid 1 \leq s < t \leq r\}$.*

The proof of Claim 2 exploits the property that $c >> a >> D'$. It demands a lot of technicalities and, therefore is omitted.

To complete the proof of Theorem 2, recall that M is $(r+1)m \times 2r$ integer matrix and the absolute value of each element is at most $c = \mathcal{O}(r^2 mn^2)$. Therefore, the bitsize N of M is $\mathcal{O}(|V(G)|^4 \log |V(G)|)$. Observe that, given $(G, V_1, \ldots, V_r)$, M can be constructed in polynomial time. Assume that there is a ω-approximation algorithm $\mathcal{A}$ for PCA WITH OUTLIERS with running time $f(d) \cdot N^{o(d)}$ for a computable function f. If $(G, V_1, \ldots, V_r)$ is a yes-instance of MULTICOLORED CLIQUE, then $\mathsf{Opt}(G, r, k) \leq D$. Therefore, $\mathcal{A}$ applied to (M, r, k) reports that there is a feasible solution (L, S) with $\|M - L - S\|_F^2 \leq \omega \mathsf{Opt}(G, r, k) \leq \omega D$ by Claim 1. For the opposite direction, if $\mathcal{A}$ reports that there is a feasible solution (L, S) with $\|M - L - S\|_F^2 \leq \omega D$, then $(G, V_1, \ldots, V_r)$ is a yes-instance of MULTICOLORED CLIQUE by Claim 1. Hence, $\mathcal{A}$ reports the existence of a feasible solution (L, S) with $\|M - L - S\|_F^2 \leq \omega D$ if and only if $(G, V_1, \ldots, V_r)$ is a yes-instance of MULTICOLORED CLIQUE. Since $N = \mathcal{O}(|V(G)|^4 \log |V(G)|)$ and $2r$, we obtain that $\mathcal{A}$ solves MULTICOLORED CLIQUE in time $f(2k) \cdot |V(G)|^{o(r)}$ contradicting ETH. □

As MULTICOLORED CLIQUE is well-known to be W[1]-hard (see (Fellows et al., 2009; Cygan et al., 2015)), our reduction gives the following corollary based on the weaker conjecture that $\mathsf{FPT} \neq \mathsf{W}[1]$. We refer to the book (Cygan et al., 2015) for the formal definitions of the parameterized complexity classes FPT and W[1]. Note that ETH implies that $\mathsf{FPT} \neq \mathsf{W}[1]$ but not the other way around.

Corollary 1. *For any $\omega \geq 1$, there is no ω-approximation algorithm for* PCA WITH OUTLIERS *with running time $f(d) \cdot N^{\mathcal{O}(1)}$ for any computable function f unless* $\mathsf{FPT} = \mathsf{W}[1]$*, where N is the bitsize of the input matrix M.*

Acknowledgements

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

References

Basu, S., Pollack, R., and Roy, M.-F. *Algorithms in Real Algebraic Geometry (Algorithms and Computation in Mathematics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3540330984.

Blum, A., Hopcroft, J., and Kannan, R. *Foundations of Data Science*. June 2017. URL https://www.microsoft.com/en-us/research/publication/foundations-of-data-science-2/.

Bouwmans, T., Aybat, N. S., and Zahzah, E.-h. *Handbook of robust low-rank and sparse matrix decomposition: Applications in image and video processing*. Chapman and Hall/CRC, 2016.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011. doi: 10.1145/1970392.1970395. URL http://doi.acm.org/10.1145/1970392.1970395.

Caruso, F. The SARAG library: Some algorithms in real algebraic geometry. In Iglesias, A. and Takayama, N. (eds.), *Mathematical Software - ICMS 2006*, pp. 122–131, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-38086-3.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. doi: 10.1137/090761793. URL https://doi.org/10.1137/090761793.

Chen, Y., Xu, H., Caramanis, C., and Sanghavi, S. Robust matrix completion and corrupted columns. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 873–880, 2011.

Cygan, M., Fomin, F. V., Kowalik, L., Lokshtanov, D., Marx, D., Pilipczuk, M., Pilipczuk, M., and Saurabh, S. *Parameterized Algorithms*. Springer, 2015. ISBN 978-3-319-21274-6. doi: 10.1007/978-3-319-21275-3. URL https://doi.org/10.1007/978-3-319-21275-3.

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

Fellows, M. R., Hermelin, D., Rosamond, F. A., and Vialette, S. On the parameterized complexity of multiple-interval graph problems. *Theoretical Computer Science*, 410(1): 53–61, 2009.

Fomin, F. V., Lokshtanov, D., Meesum, S. M., Saurabh, S., and Zehavi, M. Matrix rigidity from the viewpoint of parameterized complexity. *SIAM J. Discrete Math.*, 32 (2):966–985, 2018. doi: 10.1137/17M112258X. URL https://doi.org/10.1137/17M112258X.

Fomin, F. V., Golovach, P. A., Panolan, F., and Simonov, K. Refined complexity of PCA with outliers. *CoRR*, abs/1905.04124, 2019. URL http://arxiv.org/abs/1905.04124.

Gillis, N. and Vavasis, S. A. 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. URL https://doi.org/10.1287/moor.2017.0895.

Grigoriev, D. 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.

Grigoriev, D. 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.

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

Impagliazzo, R., Paturi, R., and Zane, F. Which problems have strongly exponential complexity. *J. Computer and System Sciences*, 63(4):512–530, 2001.

Lokshtanov, D., Marx, D., and Saurabh, S. Lower bounds based on the exponential time hypothesis. *Bulletin of the EATCS*, 105:41–72, 2011. URL http://eatcs.org/beatcs/index.php/beatcs/article/view/92.

Pearson, K. 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.

Shlens, J. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014. URL http://arxiv.org/abs/1404.1100.

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

Vaswani, N. and Narayanamurthy, P. 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. URL https://doi.org/10.1109/JPROC.2018.2844126.

Wright, J., Ganesh, A., Rao, S. R., Peng, Y., and Ma, Y. 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)*, pp. 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.

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