Principal Component Analysis in the Local Differential Privacy Model*

Di Wang†, Jinhui Xu

Department of Computer Science and Engineering
State University of New York at Buffalo, NY, USA.
{dwang45,jinhui}@buffalo.edu.

Abstract

In this paper, we study the Principal Component Analysis (PCA) problem under the (distributed) non-interactive local differential privacy model. For the low dimensional case, we show the optimal rate for the private minimax risk of the k-dimensional PCA using the squared subspace distance as the measurement. For the high dimensional row sparse case, we first give a lower bound on the private minimax risk. Then we provide an efficient algorithm to achieve a near optimal upper bound. Experiments on both synthetic and real world datasets confirm the theoretical guarantees of our algorithms.

1 Introduction

Principal Component Analysis (PCA) is a fundamental technique for dimension reduction in statistics, machine learning, and signal processing. As of today, it remains as one of the most commonly used tools in applications, especially in social sciences [Costello and Osborne, 2005], financial econometrics [Aït-Sahalia and Xiu, 2017], medicine [Barber et al., 1975], and genomics [Lu and Xu, 2013].

With the rapid development of information technologies, big data now ubiquitously exist in our daily life, which need to be analyzed (or learned) statistically by methods like regression and PCA. However, due to the presence of sensitive data (especially those in social science, biomedicine and genomics) and their distributed nature, such data are extremely difficult to aggregate and learn from. Consider a case where health records are scattered across multiple hospitals (or even countries), it is challenging to process the whole dataset in a central server due to privacy and ownership concerns. A better solution is to use some differentially private mechanisms to conduct the aggregation and learning tasks. Differential Privacy (DP) (Dwork et al., 2006b) is a commonly-accepted criterion that provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

Currently, there are mainly two user models available for differential privacy: the central model and the local one. In the central model, data are managed by a trusted central entity which is responsible for collecting them and for deciding which differentially private data analysis to perform and to release. A classical application of this model is the one of collecting census data. In the local model instead, each individual manages his/her proper data and discloses them to a server through some differentially private mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. A classical example of this model is the one aiming at collecting statistics from user devices like in the case of Google’s Chrome browser [Erlingsson et al., 2014], and Apple’s iOS-10 [Near, 2018].

In the local model, two basic types of protocols are often used: interactive and non-interactive. [Smith et al., 2017] have recently investigated the power of non-interactive differentially private protocols. This type of protocols is more natural for the classical use cases of the local model: both projects from Google and Apple use the non-interactive model. Moreover, implementing efficient interactive protocols in such applications is more difficult due to the latency of the network. Despite being used in industry, the local model has been much less studied than the central one. Part of the reason for this is that there are intrinsic limitations in what one can do in the local model. As a consequence, many basic questions, that are well studied in the central model, have not been completely understood in the local model, yet.

In this paper, we study PCA under the non-interactive local differential privacy model and aim to answer the following main question.

What are the limitations and the (near) optimal algorithms of PCA under the non-interactive local differential privacy model?

We summarize our main contributions as follows:

1. We first study the $k$-subspace PCA problem in the low dimensional setting and show that the minimax risk (measured by the squared subspace distance) under $\epsilon$ non-interactive local differential privacy (LDP) is lower bounded by $\Omega\left(\frac{1}{(k-1)^2} + \frac{1}{p} + \frac{1}{n}\right)$, where $p$ is the dimensionality of the data and $n$ is the number of data records. Moreover, we prove that the term $\Omega\left(\frac{pk}{n^2}\right)$ is optimal.

2. An undesirable issue of the above result is that the er-

---

*The research of this work was supported in part by NSF through grant CCF-1716400.
†Contact Author
ror bound could be too large in high dimensions (i.e., \( p \gg n \)). In such scenarios, a natural approach is to impose some additional structural constraints on the leading eigenvectors. A commonly used constraint is to assume that the leading eigenvectors are row sparse, which is referred to as sparse PCA in the literature and has been studied intensively in recent years [Vu et al., 2013b; Cai et al., 2013; Vu et al., 2013a]. Thus, for the high dimensional case, we consider the sparse PCA under the non-interactive local model and show that the private minimax risk (measured by the squared subspace distance) is lower bounded by \( \Omega\left(\frac{\lambda_1^2}{(\lambda_k-\lambda_{k+1})^2} \cdot \frac{1}{n^2} \cdot \log \frac{1}{\delta^2}\right) \), where \( \lambda_1, \lambda_k \) and \( \lambda_{k+1} \) are eigenvalues and \( s \) is the sparsity parameter of the eigenvectors. We also give an algorithm to achieve a near optimal upper bound of \( O\left(\frac{\lambda_1^2}{(\lambda_k-\lambda_{k+1})^2} \cdot \frac{1}{n^2} \cdot \log \frac{1}{\delta^2}\right) \).

2 Related Work

There is a vast number of papers studying PCA under differential privacy, starting from the SULQ framework [Blum et al., 2005], [Dwork et al., 2014; Chaudhuri et al., 2013; Jiang et al., 2016; Gonem and Gilad-Bachrach, 2018; Ge et al., 2018; Balcan et al., 2016]. We compare only those private PCA results in distributed settings.

For the low dimensional case, Balcan et al. [Balcan et al., 2016] studied the private PCA problem under the interactive local differential privacy model and introduced an approach based on the noisy power method. They showed an upper bound which is suitable for general settings, while ours is mainly for statistical settings. It is worth pointing out that the output in [Balcan et al., 2016] is only an \( \mathcal{O}(k) \)-dimensional subspace, instead of an exact \( k \)-dimensional subspace; thus their result is incomparable with ours. Moreover, we provide, in this paper, a lower bound on the \( \varepsilon \) non-interactive private minimax risk.

For the private high dimensional sparse PCA, the work most closely related to ours is the one by Ge et al. [Ge et al., 2018]. The authors in this paper proposed a noisy iterative hard thresholding power method, which is an interactive LDP algorithm and proved an upper bound of \( O\left(\frac{\lambda_1 \lambda_k}{(\lambda_k-\lambda_{k+1})^2} \cdot \frac{1}{(1-p^2)^{k+1}} \right) \) for their method, where \( \rho \) is a parameter related to \( \varepsilon \). Specifically, they showed that there exists some "Privacy Free Region". However, several things need to be pointed out. Firstly, our method is for general \( \varepsilon \in (0, 1] \) and non-interactive settings, while Ge et al. considered the interactive setting with more restricted \( \varepsilon \). Secondly, the assumptions in our paper are less strict than the ones in [Ge et al., 2018]. Finally, we provide a lower bound on the private minimax risk.

The optimal procedure in our paper is based on perturbing the covariance by Gaussian matrices, which has been studied in [Dwork et al., 2014]. However, there are some major differences; firstly, we show the optimality of our algorithm under the non-interactive local model using subspace distance as the measurement, while [Dwork et al., 2014] showed the optimality under the \((\varepsilon, \delta)\) central model using variance as the measurement. It is notable that in [Dwork et al., 2014] the authors also provided an upper bound on the subspace distance. However, the lower bound is still unknown. Secondly, while the optimal algorithm for the low dimensional case is quite similar, we extend it to the high dimensional case. The optimal procedure in the high dimensional sparse case is quite different from that in [Dwork et al., 2014]. Thirdly, in this paper, since we focus the statistical setting while [Dwork et al., 2014] considered the general setting, the upper bound results are incomparable.

3 Preliminaries

3.1 Classical Minimax Risk

Since all of our lower bounds are in the form of private minimax risk, we first introduce the classical statistical minimax risk before discussing the locally private version.

Let \( P \) be a class of distributions over a data universe \( \mathcal{X} \). For each distribution \( P \in \mathcal{P} \), there is a deterministic function \( \theta(P) \in \Theta \), where \( \Theta \) is the parameter space. Let \( \rho : \Theta \times \Theta :\mapsto \mathbb{R}_+ \) be a semi-metric function on the space \( \Theta \) and \( \Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+ \) be a non-decreasing function with \( \Phi(0) = 0 \) (in this paper, we assume that \( \rho \) is the subspace distance and \( \Phi(x) = x^2 \) unless specified otherwise).

We further assume that \( \{X_i\}_{i=1}^n \) are \( n \) i.i.d observations drawn according to some distribution \( P \in \mathcal{P} \), and \( \hat{\theta} : \mathcal{X}^n \mapsto \Theta \) is some estimator. Then, the minimax risk in metric \( \Phi \rho \) is defined by the following saddle point problem:

\[
\mathcal{M}_\rho(\theta(P), \Phi \rho) := \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} [\Phi(\rho(\hat{\theta}(X_1, \ldots, X_n), \theta(P))],
\]

where the supremum is taken over distributions \( P \in \mathcal{P} \) and the infimum over all estimators \( \hat{\theta} \).

3.2 Local Differential Privacy and Private Minimax Risk

Since we consider the non-interactive local model in this paper, we will follow the definitions in [Duchi et al., 2013].

We assume that \( \{Z_i\}_{i=1}^n \) are the private observations transformed from \( \{X_i\}_{i=1}^n \) through some privacy mechanisms. When \( Z_i \) depends only on \( X_i \), the mechanism is called non-interactive and in this case we have a simpler form for the conditional distributions \( Q_i(Z_i \mid X_i = x_i) \). We now define local differential privacy by restricting the conditional distribution.

**Definition 1** ([Duchi et al., 2013]). For a given privacy parameter \( \varepsilon, \delta \geq 0 \), we say that the random variable \( Z_i \) is an \((\varepsilon, \delta)\) non-interactively locally differentially private (LDP) view of \( X_i \) if for any \( x_i, x'_i \in \mathcal{X} \):

\[
Q_i(Z_i \in S \mid X_i = x_i) \leq e^\varepsilon Q_i(Z_i \in S \mid X_i = x'_i) + \delta.
\]

When \( \delta = 0 \), we call it \( \varepsilon \) non-interactive LDP view. We say that the privacy mechanism \( Q = \{Q_i\}_{i=1}^n \) is an \((\varepsilon, \delta)\) \( (\varepsilon) \) non-interactively locally differentially private (LDP) if each \( Z_i \) is an \((\varepsilon, \delta)\) \((\varepsilon)\) non-interactively LDP view.

For a given privacy parameter \( \varepsilon > 0 \), let \( Q_\varepsilon \) be the set of conditional distributions that have the \( \varepsilon \)-LDP property. For a
given set of samples \( \{X_i\}_{i=1}^n \), let \( \{Z_i\}_{i=1}^n \) be the set of observations produced by any distribution \( Q \in Q_n \). Then, our estimator will be based on \( \{Z_i\}_{i=1}^n \), that is, \( \tilde{\theta}(Z_1, \ldots, Z_n) \). This yields a modified version of the minimax risk:
\[
\mathcal{M}_n(\theta(P), \Phi, Q) = \inf_{\tilde{\theta}} \sup_{P \in P} \mathbb{E}_P[\Phi(\tilde{\theta}(Z_1, \ldots, Z_n), \theta(P))].
\]

From the above definition, it is natural to seek the mechanism \( Q \in Q_n \) that has the smallest value for the minimax risk. This allows us to define functions that characterize the optimal rate of estimation in terms of privacy parameter \( \epsilon \).

**Definition 2.** Given a family of distributions \( \theta(P) \) and a privacy parameter \( \epsilon > 0 \), the \( \epsilon \)-non-interactive private minimax risk in the metric \( \Phi \) is:
\[
\mathcal{M}_n^{\text{Nati}}(\theta(P), \Phi, \epsilon) := \inf_{Q \in Q_n} \mathcal{M}_n(\theta(P), \Phi, Q),
\]
where \( Q_n \) is the set of all \( \epsilon \) non-interactive locally differentially private mechanisms.

### 3.3 Locally Private \( k \)-dimensional PCA

Let \( X \in \mathbb{R}^p \) a random vector with mean \( 0 \) and covariance matrix \( \Sigma \). \( k \)-dimensional PCA is to find a \( k \) dimensional subspace that optimizes the following problem:
\[
\min ||(I_p - \Pi_S)X||_2^2, \text{ s.t. } \Pi \in \mathcal{G}_{p,k},
\]
where \( \mathcal{G}_{p,k} \) is the Grassmann manifold of \( k \)-dimensional subspaces of \( \mathbb{R}^p \), and \( \Pi \) is the projection of \( X \). There always exists at least one solution; consider \( \Sigma = \sum_{j=1}^p \lambda_j v_j v_j^T \), where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \) are the eigenvalues of \( \Sigma \) and \( v_1, v_2, \ldots, v_p \in \mathbb{R}^p \) are the corresponding eigenvectors. If \( \lambda_k \geq \lambda_{k+1} \), then the \( k \)-dimensional principal subspace of \( \Sigma \), \( \text{i.e.} \) the subspace \( S \) spanned by \( v_1, \ldots, v_k \) solves the above optimization problem, where the orthogonal projector of \( S \) is given by \( \Pi_S = V_k V_k^T \), where \( V_k = [v_1, \ldots, v_k] \in \mathbb{V}_{p,k} \), \( \mathbb{V}_{p,k} \) is the set of all \( p \times k \) orthogonal matrices. For simplicity, we denote \( S = \text{col}(V_k) \), where \( \text{col}(M) \) denotes the subspace spanned by the columns vectors of \( M \).

In practice, \( \Sigma \) is unknown, and the only thing that we have is the set of observation data records \( \{X_1, \ldots, X_n\} \), which are i.i.d sampled from \( X \). Thus, the problem of (non-interactively) locally differentially private PCA is to find a \( k \)-dimensional subspace \( S^{\text{priv}} \) which is close to \( S \), where the algorithm that outputs \( S^{\text{priv}} \) must be \( \epsilon \) (non-interactively) locally differentially private. After obtaining a private estimator \( S^{\text{priv}} \), there are multiple ways to measure the success, such as variance guarantee [Dwork et al., 2014], low rank approximation error [Kapralov and Talwar, 2013], etc. In this paper, we will use the subspace distance as the measurement [Dwork et al., 2014; Ge et al., 2018].

Let \( S \) and \( S' \) be two \( k \)-dimensional subspaces in \( \mathbb{R}^p \). Also denote by \( E \) and \( F \), respectively, the orthogonal matrix corresponds to \( S \) and \( S' \). That is, \( E = VV^T \) and \( F = WW^T \) for some orthogonal matrices \( V \in \mathbb{V}_{p,k} \) and \( W \in \mathbb{V}_{p,k} \). Then, the squared subspace distance between \( S \) and \( S' \) is defined by the following [Stewart, 1990]:
\[
||\sin \Theta(S, S')||_F^2 = ||E - F||_F^2 = \frac{1}{2}||VV^T - WW^T||_F^2.
\]

where || · ||_F is the Frobenious norm. For simplicity, we will overload notation and write \( \Theta(S, S') = \sin \Theta(V, W) \).

### 4 Low Dimensional Case

In this section, we focus on the general case and always assume \( n \geq p \). We first derive a lower bound of the \( \epsilon \) non-interactive private minimax risk using the squared subspace distance as the measurement. By the definition of the \( \epsilon \)-private minimax risk, it is important to select an appropriate class of distributions.

#### 4.1 Class of Distributions

We assume that the observations \( \{X_i\}_{i=1}^n \) satisfies: \( X = \Sigma^\top Z \), where \( Z \in \mathbb{R}^p \) is a random vector satisfying equations \( E Z = 0 \) and \( \text{Var}(Z) = I_p \). Also, \( Z \) is sub-Gaussian and \( ||Z||_{\mathbb{F}_2} \leq 1 \) where
\[
||Z||_{\mathbb{F}_2} := \sup_{v : ||v||_2 \leq 1} \mathbb{E}[Z,v]^2 \leq 2,
\]
which means that all the one-dimensional marginals of \( X \) sub-Gaussian tails. We need to note that this assumption on \( X \) is commonly used in many papers on PCA in statistical settings, such as [Vu et al., 2013b; Ge et al., 2018].

In the study of private PCA, it is always assumed that the \( \mathcal{C}_\epsilon \) norm of each \( X_i \) is bounded by 1, as in [Dwork et al., 2014; Ge et al., 2018]. For convenience, we relax this assumption in the following way; for the random vector \( X \in \mathbb{R}^p \), we assume that \( ||X||_2 \leq 1 \) with a probability at least \( 1 - e^{-\Omega(p)} \).

Next, we give assumptions on the population covariance matrix \( \Sigma \). Firstly, we assume that for the target \( k \)-dimensional subspace, \( \lambda_k - \lambda_{k+1} > 0 \) so that the principal subspace is well defined. Next, we define the effective noise variance, which is proposed in [Vu et al., 2013b] and [Cai et al., 2013]:
\[
\sigma_k^2(\lambda_1, \lambda_2, \ldots, \lambda_p) := \frac{\lambda_1 \lambda_k \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}.
\]

For a given constant \( \sigma^2 > 0 \), we assume that \( \sigma_k^2 \leq \sigma^2 \). Also, we denote the collection of distributions by \( P(k, \sigma^2) \).

#### 4.2 Main Results

The next theorem shows a lower bound of \( \epsilon \) non-interactive private minimax risk under squared subspace distance.

**Theorem 1.** Let \( \{X_i\}_{i=1}^n \) be samples from \( P \in P(k, \sigma^2) \). If \( \frac{n}{k} \leq k \leq \frac{2n}{k^2} \), \( \epsilon \in (0, \frac{1}{2}) \) and \( n \geq \Omega\left(\frac{1}{\epsilon^2} \left(\frac{\lambda_1 \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}\right) \min(k, p - k)\right) \), then the \( \epsilon \)-non-interactive private minimax risk in the metric of squared subspace distance satisfies:
\[
\mathcal{M}_n^{\text{Nati}}(P(k, \sigma^2); \Phi, \epsilon) \geq \Omega\left(\frac{\lambda_1 \lambda_{k+1}}{\lambda_k - \lambda_{k+1}} \frac{kp}{n^2} \right).
\]

**Remark 1.** We note that for the non-private case, the minimax risk is lower bounded by \( \Omega\left(\frac{\lambda_1 \lambda_{k+1}}{\lambda_k - \lambda_{k+1}} \frac{kp}{n^2} \right) \) [Cai et al., 2013]. Thus, in this case, the impact of the local differential privacy is to change the number of efficient sample from \( n \) to \( n^2 \). However, the collection of the considered distributions needs another assumption which says that \( ||X||_2 \) is bounded by 1 with high probability. This is not necessary in the non-private case [Cai et al., 2013], but needed in ours for showing the upper bound.
We also note that in the central differential privacy model, [Dwork et al., 2014] showed that the lower bound of the $k$-dimensional PCA is $\Omega\left(\frac{kp\log(1/\delta)}{n^{2}}\right)$ for $(\epsilon, \delta)$-differential privacy. However, this lower bound is measured by the variance of $X = (X_{1}^{T}, \ldots, X_{n}^{T})^{T} \in \mathbb{R}^{np}$, not the squared subspace distance used in this paper. Although [Dwork et al., 2014] gave an upper bound of $O\left(\frac{kp\log(1/\delta)}{(\lambda_{k}-\lambda_{k+1})^{2}n^{2}}\right)$ in the general setting using the squared subspace distance as measurement, it is still unknown whether the bound is optimal. Also, their lower bound omits the parameters related to the eigenvalues.

For the $\epsilon$ differential privacy in the central model, [Chaudhuri et al., 2013] showed that the lower bound is $\Omega\left(\frac{\epsilon^{2}}{n^{2}(\lambda_{1}-\lambda_{2})^{2}}\right)$ in the special case of $k = 1$. However, it is still unknown for the general case of $k$. Thus, from the above discussion, we can see that the lower bound of $\epsilon$ non-interactively locally differentially private PCA is similar to the $(\epsilon, \delta)$ differentially private PCA in the central model.

One of the main questions is whether the lower bound in Theorem 1 is tight. In the following, we show that the term $\Omega\left(\frac{kp\log(1/\delta)}{n^{2}}\right)$ is tight. By our definition of the parameter space, we know that for any $X \sim P \in P(\sigma^{2}, k)$, $\|X\|_{2} \leq 1$ with high probability. Thus, we always assume that the event of each $\|X_{i}\|_{2} \leq 1$ holds. Note that this assumption also appears in [Ge et al., 2018; Dwork et al., 2014; Balcan et al., 2016]. The idea is the same as in [Dwork et al., 2014], where each $X_{i}$ perturbs its covariance and aggregates the noisy version of covariance, see Algorithm 1 for details.

**Theorem 2.** For any $\epsilon, \delta > 0$, Algorithm 1 is $(\epsilon, \delta)$ (non-interactively) locally differentially private. Furthermore, with probability at least $1 - e^{-C_{1}} \frac{1}{\epsilon^{2}}$, the output satisfies:

$$
\|\sin(\Theta(\tilde{V}_{k}, V_{k}))\|_{F}^{2} \leq O\left(\frac{\lambda_{k}^{2}kp\log(1/\delta)}{(\lambda_{k} - \lambda_{k+1})^{2}n^{2}}\right),
$$

(2)

where $C_{1}, C_{2}$ are some universal constants.

**Algorithm 1** Local Gaussian Mechanism

**Input:** data records $\{X_{i}\}_{i=1}^{n} \sim P^{n}$ for $P \in P(\sigma^{2}, k)$, and for $i \in [n], \|X_{i}\|_{2} \leq 1$. $\epsilon, \delta$ are the privacy parameters.

1. for Each $i \in [n]$ do
2. Denote $\tilde{X}_{i} = X_{i} + Z_{i}$, where $Z_{i} \in \mathbb{R}^{p \times p}$ is a symmetric matrix where the upper triangle (including the diagonal) is i.i.d samples from $\mathcal{N}(0, \sigma_{1}^{2})$; here $\sigma_{1}^{2} = \frac{2\ln(5/\delta + 1)}{\epsilon^{2}}$, and each lower triangle entry is copied from its upper triangle counterpart.
3. end for
4. Compute $\tilde{S} = \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_{i}\tilde{X}_{i}^{T}$.
5. Output col($\tilde{V}_{k}$) where $\tilde{V}_{k} \in \mathbb{R}^{p \times k}$ is the principal rank $k$ subspace of $\tilde{S}$.

In Theorem 7 of [Dwork et al., 2014], the authors provided a similar upper bound for the $(\epsilon, \delta)$-differential privacy in the central model. However, they need to assume that the eigenvalues satisfy the condition $\lambda_{k}^{2} - \lambda_{k+1}^{2} = o(\sqrt{p})$, which is not needed in our Theorem 2.

From Theorems 1 and 2, we can see that there is still a gap of $O(\frac{1}{\epsilon^{2}\delta^{2}k})$ between the lower and upper bounds. We leave it as an open problem to determine whether these bounds are tight or not.

## 5 High Dimensional Sparse Case

From Theorem 1, we can see that for the high dimensional case, i.e. $p \gg n$, the bound in (2) becomes trivial. Thus, to avoid this issue, we need some additional assumption on the parameter space. One of the commonly used assumption is sparsity. There are many definitions of sparsity on PCA and we use the row sparsity in this paper, which has also been studied in [Vu et al., 2013b; Cai et al., 2013; Ge et al., 2018].

We first define the $(2,0)$-norm of a $p \times k$ matrix $A$ as the usual $\ell_{2}$ norm of the vector of row-wise $\ell_{0}$ norms of $A$:

$$
\|A\|_{2,0} := \|\|a_{1,1}\|_{2}, \|a_{2,1}\|_{2}, \ldots, \|a_{p,1}\|_{2}\|_{0},
$$

(3)

where $a_{j, l}$ denotes the $j$-th row of $A$. Note that $\|\cdot\|_{2,0}$ is coordinate independent, i.e. $\|A\|_{2,0} = \|A\|_{2,0}$ for any orthogonal matrix $O \in \mathbb{R}^{k \times k}$. We define the row sparse principal subspace as follows.

**Definition 3.** Let $s$ be the sparsity level parameter satisfying the condition of $k \leq s \leq p$. The $s$- (row) sparse subspace is defined as follows

$$
\mathcal{M}_{s}(\Sigma) = \{Col(U), U \in \mathbb{R}^{p \times k} \text{ and orthogonal }, \|U\|_{2,0} \leq s\}.
$$

We define our parameter space, $P(s, k, \sigma^{2})$, to be the same as in the previous section with an additional condition that $S \in \mathcal{M}_{s}(\Sigma)$, where $S$ is the $k$-dimensional principal subspace of covariance matrix $\Sigma$.

Below, we will first derive a lower bound of the non-interactive locally differentially private PCA in the high dimensional sparse case.

**Theorem 3.** Let $\{X_{i}\}_{i=1}^{n}$ be the observations sampled from a distribution $P \in P(s, k, \sigma^{2})$. If the privacy parameter $\epsilon \in (0, \frac{1}{\sqrt{n}})$, $n \geq C(s-k)\sigma^{2}(k+\log p)$ for a universal constant $C > 0$, and $2k \leq s - k \leq p - k$, then for all $k \in [p]$ satisfying the condition of $k \leq s - 4$, the $\epsilon$ non-interactive private minimax risk in the metric of squared subspace distance satisfies the following

$$
\mathcal{M}_{n}^{\text{inf}}(S(\mathcal{P}(s, k, \sigma^{2})), e) \geq \Omega\left(\frac{\lambda_{1}\lambda_{k+1}}{(\lambda_{k} - k_{k+1})^{2}n^{2}}\right).
$$

Note that in the non-private case, the optimal minimax risk is $O\left(\frac{\lambda_{1}\lambda_{k+1}}{(\lambda_{k} - k_{k+1})^{2}n}\right)$. Thus, same as in the low dimensional case, the impact of the privacy constraint is to change the efficient samples from $n$ to $ne^{2}$.

Next, we consider the upper bound. In the non-private case, the optimal procedure is to solve the following NP-hard optimization problem [Vu et al., 2013b]:

$$
\max <S, UU^{T}> \quad \text{subject to } U^{T}U = I_{k}, U \in \mathbb{R}^{p \times k} \text{ and } \|U\|_{2,0} \leq s,
$$

(4)
where $S$ is the empirical covariance matrix. Our upper bound is based on (4). However, instead of solving (4) on the perturbed version of the empirical covariance matrix, we perturb the covariance matrix and solve the following optimization problem on the convex hull of the constraints in (4), that is:

$$\hat{X} = \arg \max < S, X > - \lambda \| X \|_1$$

subject to $X \in P^k := \{ X : 0 \leq X \leq I$ and $\text{Tr}(X) = k \}$,

where $< S, X > = \text{Tr}(S X^T)$. Note that the constraints in (5), which is called Fantope [Bhatia, 2013] [Vu et al., 2013a], is the convex hull of the constrains in (4). Also, since the constraints in (5) only guarantees that the rank of the output is $\geq k$, the output $\hat{X}$ needs not be a matrix with exact rank of $k$. Thus, in order to obtain a proper $k$-dimensional subspace, we just output the $k$-PCA of $\hat{X}$.

**Algorithm 2** Local Gaussian Mechanism-High Dimension

**Input:** data records $\{X_i\}_{i=1}^n \sim P^n$ for $P \in P(s, \sigma^2, k)$, and for $i \in [n], \| X \|_2 \leq 1$. $\epsilon, \delta$ are privacy parameters. $\rho >$ is a constant.

1: for Each $i \in [n]$ do
2: Denote $\tilde{X}_i \tilde{X}_i^T = X_i X_i^T + Z_i$, where $Z_i \in \mathbb{R}^{p \times p}$ is a symmetric matrix where the upper triangle (including the diagonal) is i.i.d samples from $\mathcal{N}(0, \sigma_i^2)$; here $\sigma_i^2 = \frac{2 \ln(1, 25/\delta)}{\epsilon^2}$, and each lower triangle entry is copied from its upper triangle counterpart.
3: end for
4: Compute $\tilde{S} = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T$.
5: Get the optimal solution $\hat{X}$ in (5) or do as the followings
6: Setting $Y(0) = 0, U(0) = 0$
7: for $i = 1, 2, \ldots$ do
8: $X^{(i+1)} = P_{\rho, s}(Y^{(i)} - U^{(i)} + \tilde{S})$
9: $Y^{(i+1)} = S_{\lambda/\rho}(X^{(i+1)} + U^{(i)})$ where $S$ is the entrywise soft thresholding operator defined as $S_{\lambda/\rho}(x) = \text{sign}(x) \max(\text{max}(x) - \lambda/\rho, 0)$.
10: $U^{(i+1)} = Y^{(i)} - X^{(i+1)} - Y^{(i+1)}$
11: Return $Y^{(i)}$
12: end for
13: Let $k$-dimensional principal component of $\hat{X}$ or $Y^{(i)}$ be $\hat{V}_k$, output $\hat{S} = \text{col}(\hat{V}_k)$.

**Theorem 4.** For any given $0 < \epsilon, \delta < 1$, if $\{X_i\}_{i=1}^n \sim P^n$ for $P \in P(s, \sigma^2, k)$ and $\| X \|_2 \leq 1$ for all $i \in [n]$, then the solution to the optimization problem (5) is $(\epsilon, \delta)$ non-interactive locally differentially private. Moreover, if let $\hat{V}_k$ denote the $k$-dimensional principal component subspace of $\hat{X}$ and set $\lambda \leq O(\lambda^k \sqrt{\log p / nc^2})$, then with probability at least $1 - \frac{1}{\rho^2} - \frac{1}{p}$, the following holds

$$\| \sin(\Theta(\hat{V}_k, V_k)) \|_F^2 \leq O\left(\frac{\lambda^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s^2 \log p}{nc^2}\right)$$

where $c$ is a universal constant.

Since the optimization problem (5) is convex, we can follow the approach in [Vu et al., 2013a] to solve it by using ADMM method (see Algorithm 2 for the details).

Comparing with the lower bound of the private minimax risk in Theorem 3, we can see that the bound in Theorem 4 is roughly larger than the optimal rate by a factor of $O(\frac{\epsilon}{\lambda^{k+1}})$. This means that the upper bound is only near optimal [Vu et al., 2013a]. A remaining open problem is to determine whether it is possible to get a tighter upper bound that does not contain the term of $\frac{\epsilon}{k}$ in the gap.

### 6 Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>$s$</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer RNA-Seq</td>
<td>(801, 20531)</td>
<td>10</td>
<td>3.162</td>
</tr>
<tr>
<td>Leukemia</td>
<td>(72, 7128)</td>
<td>10</td>
<td>3.162</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>(60, 2000)</td>
<td>10</td>
<td>2.449</td>
</tr>
<tr>
<td>isolet5</td>
<td>(1559, 617)</td>
<td>20</td>
<td>2.023</td>
</tr>
<tr>
<td>lung</td>
<td>(203, 3312)</td>
<td>20</td>
<td>2.858</td>
</tr>
<tr>
<td>NIPS</td>
<td>(11463, 5811)</td>
<td>20</td>
<td>3.643</td>
</tr>
</tbody>
</table>

Table 1: Results with different sparsity $s$ for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions $k$ is set to be $k = 10$ and $\epsilon = 2$.

### 6.1 Low Dimensional Case

For synthetic datasets, we generate the data samples $\{X_i\}_{i=1}^n$ independently from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma = \frac{1}{5p(k+1)}V V^T + \frac{1}{5p(k+1)}I_p$ for $V \in \mathbb{R}^{p \times k}$. It can be shown that $\| X \|_2 \leq 1$ for all $i \in [n]$ with high probability (see supplemental material). We choose $n = 10^3$, $p = 40$, $k = \{5, 10, 15, 20\}$, $\epsilon = 0.5$, $\delta = 10^{-4}$, and $\lambda = 1$. For real world datasets, we run Algorithm 1 on Covertype and Buzz datasets [Dheeru and Karra Taniskidou, 2017] with normalized rows for each dataset. The error is measured by the subspace distance $\| \hat{V}_k V_k^T - V_k V_k^T \|_F$. For each experiment, we repeat 20 times and take the average as the final result.

Figure 1 is the result for the synthetic datasets. Figure indicates that 1) the error decreases as the sample size increases or $i$ increases (i.e., becomes less private); 2) the error increases as the dimensionality $p$ increases or the dimensionality $k$ of the target subspace increases.
Figure 1: LDP-PCA in high dimensional case on synthetic datasets. The left one is for different target dimensions $k$ over sample size $n$ with fixed $\epsilon = 2$ and $p = 400$. The middle one is for different dimensions with fixed $n = 2000$ and $\epsilon = 2$. The right one is for different level of privacy with fixed $n = 2000$ and $p = 400$.

Figure 2: LDP-PCA in high dimensional case on synthetic datasets. The left one is for different target dimensions $k$ over sample size $n$ with fixed $\epsilon = 2$ and $p = 400$. The middle one is for different dimensions with fixed $n = 2000$ and $\epsilon = 2$. The right one is for different level of privacy with fixed $n = 2000$ and $p = 400$.

6.2 High Dimensional Case

For the high dimensional case, we consider the same distributions as in the low dimensional case and generate the target subspace $V$ in the following way. For a given sparsity parameter $s$, we first generate a random orthogonal matrix $\tilde{V} \in \mathbb{R}^{s \times k}$, then pad it with rows of zeros, and finally randomly permute the matrix. We set $k = 10$, $n = 2000$, $p = 400$, $s = \{15, 20, 40, 80\}$ and $\epsilon = 1$.

Besides the synthetic datasets, we also test our algorithm on some real world datasets in [Dheeru and Karra Taniskidou, 2017] and [Li et al., 2017]. We can see that 1) as the term of $\frac{k}{n}$ increases ($n$ decreases), the error increases accordingly; 2) the error slightly increases when the dimensionality $p$ increases, which is due to the fact that the upper bound in Theorem 4 depends only logarithmically on $p$ (i.e., $\log p$); 3) the error decreases when $\epsilon$ increases. Table 1 and 2 show the results of the error with different sparsity and privacy, respectively.

References


| Dataset       | Size            | $\epsilon$ | Error  
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer RNA-Seq</td>
<td>(801, 20531)</td>
<td>1</td>
<td>3.559</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>3.710</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>3.967</td>
</tr>
<tr>
<td>Leukemia</td>
<td>(72, 7128)</td>
<td>1</td>
<td>4.375</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>4.403</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>4.518</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>(60, 2000)</td>
<td>1</td>
<td>3.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>4.237</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>4.310</td>
</tr>
<tr>
<td>isoelet5</td>
<td>(1559, 617)</td>
<td>1</td>
<td>2.884</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>3.405</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>3.896</td>
</tr>
<tr>
<td>lung</td>
<td>(203, 3312)</td>
<td>1</td>
<td>4.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>4.275</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>4.362</td>
</tr>
<tr>
<td>NIPS</td>
<td>(11463, 5811)</td>
<td>1</td>
<td>4.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>4.052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>4.472</td>
</tr>
</tbody>
</table>

Table 2: Results with different privacy levels $\epsilon$ for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions $k$ is set to be $k = 10$ and $s = 20$. 


[Costello and Osborne, 2005] Anna B Costello and Jason W Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical assessment, research & evaluation, 10(7):1–9, 2005.


