arXiv:1207.2812v1 [stat.ML] 12 Jul 2012

# Near-Optimal Algorithms for Differentially-Private Principal Components

Kamalika Chaudhuri* Anand D. Sarwate† Kaushik Sinha‡

July 13, 2012

**Abstract**

Principal components analysis (PCA) is a standard tool for identifying good low-dimensional approximations to data sets in high dimension. Many current data sets of interest contain private or sensitive information about individuals. Algorithms which operate on such data should be sensitive to the privacy risks in publishing their outputs. Differential privacy is a framework for developing tradeoffs between privacy and the utility of these outputs. In this paper we investigate the theory and empirical performance of differentially private approximations to PCA and propose a new method which explicitly optimizes the utility of the output. We demonstrate that on real data, there this a large performance gap between the existing methods and our method. We show that the sample complexity for the two procedures differs in the scaling with the data dimension, and that our method is nearly optimal in terms of this scaling.

## 1 Introduction

Dimensionality reduction is a fundamental tool for understanding complex data sets that arise in contemporary machine learning and data mining applications. Even though a single data point can be represented by hundreds or even thousands of features, the phenomena of interest are often intrinsically low-dimensional. By reducing the "extrinsic" dimension of the data to its "intrinsic" dimension, analysts can discover important structural relationships between features, more efficiently use the transformed data for learning tasks such as classification or regression, and greatly reduce the space required to effectively store the data. One of the oldest and most classical methods for dimensionality reduction is principal components analysis (PCA), which computes a low-rank approximation to the (positive semidefinite) second moment matrix of a set of points in $\mathbb{R}^d$. The rank $k$ of the approximation is chosen to be the intrinsic dimension of the data. We view this procedure as specifying a $k$-dimensional subspace of $\mathbb{R}^d$.

Much of today's machine-learning is performed on the vast amounts of personal information collected by private companies and government agencies about individuals, such as

*Department of Computer Science and Engineering, University of California, San Diego, kchaudhuri@ucsd.edu.

†Toyota Technological Institute at Chicago, asarwate@ttic.edu

‡Department of Computer Science and Engineering, University of California, San Diego, ksinha@cs.ucsd.edu

customers, users, and subjects. These datasets contain sensitive information about individuals and typically involve a large number of features. It is therefore important to design machine-learning algorithms which discover important structural relationships in the data while taking into account its sensitive nature. We study approximations to PCA which guarantee differential privacy, a cryptographically motivated definition of privacy [Dwork et al., 2006b] that has gained significant attention over the past few years in the machine-learning and data-mining communities [Machanavajjhala et al., 2008, McSherry and Mironov, 2009, McSherry, 2009, Friedman and Schuster, 2010, Mohammed et al., 2011]. Differential privacy measures privacy risk by a parameter $\alpha$ that bounds the log-likelihood ratio of output of a (private) algorithm under two databases differing in a single individual.

There are many general tools for providing differential privacy. The sensitivity method [Dwork et al., 2006b] computes the desired algorithm (PCA) on the data and then adds noise proportional to the maximum change than can be induced by changing a single point in the data set. The PCA algorithm is very sensitive in this sense because the top eigenvector can change by $90^\circ$ by changing one point in the data set. Relaxations such as smoothed sensitivity [Nissim et al., 2007] are difficult to compute in this setting as well. The SULQ method of Blum et al. [2005] adds noise to the second moment matrix and then runs PCA on the noisy matrix. As our experiments show, the amount of noise required is often quite severe and SULQ seems impractical for data sets of moderate size.

These general methods do not take into account the quality of approximation to the non-private PCA output. We address this by proposing a new method, PPCA, that is an instance of the exponential mechanism of McSherry and Talwar [2007]. This differentially private method outputs a $k$-dimensional subspace; the output is biased towards subspaces which are close to the output of PCA, and in our case corresponds to sampling from the matrix Bingham distribution. We implement this method using a Markov Chain Monte Carlo (MCMC) procedure due to Hoff [2009] and show that it achieves significantly better empirical performance.

In order to understand this gap in performance we prove sample complexity bounds for SULQ and PPCA, as well as a general lower bound on the sample complexity for any differentially private algorithm. We show that the sample complexity scales as $\Omega(d^{3/2}\sqrt{d})$ for SULQ and as $O(d)$ for PPCA. Furthermore, any differentially private algorithm requires $\Omega(d)$ samples, showing that PPCA is nearly optimal in terms of sample complexity. These theoretical results suggest that our experiments exhibit the limit of how well $\alpha$-differentially private algorithms can perform.

There are several interesting open questions suggested by this work. One set of issues is computational. Differentially privacy is a mathematical definition, but algorithms must be implemented using finite precision machines. Privacy and computation interact in many places, including pseudorandomness, numerical stability, optimization, and in the MCMC procedure we use to implement PPCA. A second set of issues is theoretical – our theoretical analysis for the sample complexity is for the special case of $k = 1$ in which the distance and angles between vectors are related. For general $k$ the measure of approximation may be application dependent; distances between subspaces and frames may be different and lead to different tradeoffs.

## Related work

There are many different models for privacy-preserving computation [Agrawal and Srikant, 2000, Evfimievski et al., 2003, Sweeney, 2002, Machanavajjhala et al., 2006, Li et al., 2010] – seeFung et al. [2010] for a survey with more references. Many of these models have been shown to be susceptible to composition attacks, which are attacks in which the adversary exploits prior knowledge to reidentify individuals [Ganta et al., 2008]. An alternative line of privacy-preserving data-mining work [Zhan and Matwin, 2007] is in the Secure Multiparty Computation setting due to Yao [1982]; one work studies privacy-preserving singular value decomposition in this model [Han et al., 2009]. Finally, dimension reduction through random projection has been considered as a technique for sanitizing data prior to publication [Liu et al., 2006]; our work differs from this line of work in that we offer differential privacy guarantees, and we only release the PCA subspace, not actual data.

Differential privacy was proposed by Dwork et al. [2006b], and has been used since in many works on privacy [Kasiviswanathan et al., 2008, Chaudhuri and Mishra, 2006, Barak et al., 2007, Machanavajjhala et al., 2008, McSherry and Mironov, 2009, Chaudhuri et al., 2011, Nissim et al., 2007, Blum et al., 2008, McSherry and Talwar, 2007, Friedman and Schuster, 2010, Hay et al., 2009, Williams and McSherry, 2010, Roy et al., 2010]. Differential privacy has been shown to have strong *semantic* guarantees [Dwork et al., 2006b, Kasiviswanathan and Smith, 2008] and is resistant to many attacks [Ganta et al., 2008] that succeed against some of the other aforementioned definitions of privacy. There are several standard approaches for designing differentially-private data-mining algorithms, including input perturbation [Blum et al., 2005], output perturbation [Dwork et al., 2006b], the exponential mechanism [McSherry and Talwar, 2007], and objective perturbation [Chaudhuri et al., 2011].

The SULQ method [Blum et al., 2005] gave a general differentially-private input perturbation algorithm and showed that it could be applied to PCA. Williams and McSherry [2010] described a method for inferring one-dimensional subspace from noisy measurements using differential privacy. Our work differs from Blum et al. [2005] these by tuning the mechanism to a utility function for PCA and from Williams and McSherry [2010] by not making modeling assumptions on the data. Independently of our work, Hardt and Roth [2012] consider the problem of differentially-private low-rank matrix reconstruction with a view towards applications to sparse matrices; provided certain coherence conditions hold, they provide an algorithm for constructing a rank $2k$ approximation $B$ to a matrix $A$ such that $\|A - B\|_F$ is $O(\|A - A_k\|)$ plus some additional terms which depend on $d$, $k$ and $n$; here $A_k$ is the best rank $k$ approximation to $A$. Because of their additional assumptions, their bounds are generally incomparable to ours; however, it can be shown that our bounds are superior for dense matrices $A$.

In many differentially-private algorithms that use input perturbation, the magnitude of the noise needed to guarantee differential privacy often grows linearly with the extrinsic dimension of the data. Examples include differentially-private means and covariances [Dwork et al., 2006b] and differentially-private PCA and $k$-means [Blum et al., 2005]. Another example is linear classification; Chaudhuri et al. [2011] present linear classification algorithms in which the magnitude of the added perturbation, as well as the sample requirement, grows linearly with the extrinsic dimension of the dataset. To the best of our knowledge, the only prior work which considered differentially-private PCA is SULQ [Blum et al., 2005]; SULQ adds noise to each entry of the empirical second moment matrix, and then outputs the top $k$ singular vectors of this perturbed matrix. Our experiments as well as theoretical results indicate that our solution has higher statistical efficiency, particularly when $k$ is small compared to the extrinsic data dimension. However, it has been recently shown by Hall et al. [2012] that $(\alpha, \delta)$-differential privacy can provide a significantly better dependence on the dimension than $\alpha$-differential privacy.

Lower bounds on the sample requirement of differentially private algorithms for statistical tasks is also an active area of research. Hardt and Talwar [2010] prove lower bounds on the amount of noise that needs to be added to a number of histogram count queries, and Chaudhuri and Hsu [2011] show lower bounds on the sample requirement of differentially-private classification as a function of the doubling dimension of the disagreement metric of the hypothesis class with respect to the data distribution. Our lower bounds on the sample requirement of any differentially-private algorithm extends ideas from these two papers.

## 2 Preliminaries

For an integer $d$, let $[d] = \{1, 2, \ldots, d\}$, let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, and let $I_d$ be the $d \times d$ identity matrix. The transpose of a matrix $A$ is denoted by $A^T$. If $A$ is positive definite we denote by $\lambda_i(A)$ the $i$-th largest eigenvalue of $A$. When $\lambda_i(A)$ has multiplicity one, we denote by $v_i(A)$ the corresponding eigenvector. We use $I_d$ to denote the $d \times d$ identity matrix. We will use different norms in this paper : $\|A\|_{\mathrm{F}} = \mathrm{tr}(AA^T)$ is the Frobenius norm of a matrix $A$ and $\|x\|_2$ is the Euclidean norm. We will write $\mathbf{KL}\left(f\|g\right) = \int f(x) \frac{f(x)}{g(x)} dx$ for the Kullback-Leibler divergence between two densities $f$ and $g$.

The data given to our algorithm is a set of $n$ vectors $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ where each $x_i$ corresponds to the private value of one individual, $x_i \in \mathbb{R}^d$, and $\|x_i\| \leq 1$ for all $i$. Let $X = [x_1, \ldots, x_n]$ be the matrix whose columns are the data vectors $\{x_i\}$. Let $A = \frac{1}{n} X X^T$ denote the $d \times d$ second moment matrix of the data. The matrix $A$ is positive semidefinite, and has Frobenius norm at most 1.

The problem of dimensionality reduction is to find a "good" low-rank approximation to $A$. A popular solution is to compute a rank-$k$ matrix $\hat{A}$ which minimizes the norm $\left\|A - \hat{A}\right\|_{\mathrm{F}}$, where $k$ is much lower than the data dimension $d$. The Schmidt approximation theorem [Stewart, 1993, Eckart and Young, 1936] shows that the minimizer is given by the singular value decomposition, also known as the PCA algorithm in some areas of computer science. Let the eigenvalues of $A$ be $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A) \geq 0$ and let $\Lambda$ be a diagonal matrix with $\Lambda_{ii} = \lambda_i(A)$. The matrix $A$ decomposes as

$$A = V \Lambda V^T, \tag{1}$$

where $V$ is an orthonormal matrix of eigenvectors.

**Definition 1.** *Suppose $A$ is a positive semidefinite matrix whose first $k$ eigenvalues are distinct. The top-$k$ subspace of $A$ is the matrix*

$$V_k(A) = \left[v_1 \ v_2 \ \cdots \ v_k\right]. \tag{2}$$

*where $\Lambda^{(k)}_{ii} = \lambda_i(A)$ for $i = 1, 2, \ldots, k$ and $\Lambda^{(k)}_{ii} = 0$ elsewhere, and $V$ is given by (1).*

Given the top-$k$ subspace and the eigenvalue matrix $\Lambda$, we can form an approximation to $A$:

$$A^{(k)} = [V_k(A)\ \mathbf{0}]\,\Lambda\,[V_k(A)\ \mathbf{0}]^T\,, \tag{3}$$

In the special case $k = 1$ we have $A^{(1)} = \lambda_1(A) v_1 v_1^T$, where $v_1$ is the eigenvector corresponding to $\lambda_1(A)$. We refer to $v_1$ as the *top eigenvector* of the data.

We study approximations to $\hat{A}$ that preserve the privacy of the underlying data. The notion of privacy that we use is differential privacy, which quantifies the privacy guaranteed by a randomized algorithm $\mathcal{P}$ applied to a data set $\mathcal{D}$.

**Definition 2.** *An algorithm $\mathcal{A}(\mathcal{B})$ taking values in a set $\mathcal{T}$ provides $\alpha$-differential privacy if*

$$\sup_{\mathcal{S}} \sup_{\mathcal{D},\mathcal{D}'} \frac{\mu\left(\mathcal{S} \mid \mathcal{B} = \mathcal{D}\right)}{\mu\left(\mathcal{S} \mid \mathcal{B} = \mathcal{D}'\right)} \le e^{\alpha}, \tag{4}$$

*where the first supremum is over all measurable $\mathcal{S} \subseteq \mathcal{T}$, the second is over all data sets $\mathcal{D}$ and $\mathcal{D}'$ differing in a single entry, and $\mu(\cdot|\mathcal{B})$ is the conditional distribution (measure) on $\mathcal{T}$ induced by the output $\mathcal{A}(\mathcal{B})$ given a data set $\mathcal{B}$. The ratio is interpreted to be 1 whenever the numerator and denominator are both 0.*

**Definition 3.** *An algorithm $\mathcal{A}(\mathcal{B})$ taking values in a set $\mathcal{T}$ provides $(\alpha, \delta)$-differential privacy if*

$$\mathbb{P}\left(\mathcal{A}(\mathcal{D}) \in \mathcal{S}\right) \le e^{\alpha} \mathbb{P}\left(\mathcal{A}(\mathcal{D}') \in \mathcal{S}\right) + \delta, \tag{5}$$

*for all all measurable $\mathcal{S} \subseteq \mathcal{T}$ and all data sets $\mathcal{D}$ and $\mathcal{D}'$ differing in a single entry.*

Here $\alpha$ and $\delta$ are privacy parameters, where low $\alpha$ and $\delta$ ensure more privacy. For more details about these definitions, see Dwork et al. [2006b], Wasserman and Zhou [2010], Dwork et al. [2006a].

The purpose of PCA is often to approximate the data by a low dimensional subspace that captures most of the "energy" of the data. For a $d \times k$ matrix $\hat{V}$ with orthonormal columns, the quality of $\hat{V}$ in approximating $A$ can be measured by

$$q_{\mathrm{F}}(\hat{V}) = \mathrm{tr}\left(\hat{V}^T A \hat{V}\right). \tag{6}$$

The $\hat{V}$ which maximizes $q(\hat{V})$ has columns equal to $\{v_i : i \in [k]\}$, corresponding to the top $k$ eigenvectors of $A$.

Our theoretical results apply to the special case $k = 1$, although they can be extended for general $k$. For these results, we measure the inner product between the output vector $\hat{v}_1$ and the true top eigenvector $v_1$:

$$q_{\mathrm{A}}(\hat{v}_1) = \left|\langle \hat{v}_1, v_1 \rangle\right|. \tag{7}$$

This is related to (6). If we write $\hat{v}_1$ in the basis spanned by $\{v_i\}$, then

$$q_{\mathrm{F}}(\hat{v}_1) = \lambda_1 q_{\mathrm{A}}(\hat{v}_1)^2 + \sum_{i=2}^{d} \lambda_i \langle \hat{v}_1, v_i \rangle^2$$

Our proof techniques use the geometric properties of $q_{\mathrm{A}}(\cdot)$.

**Definition 4.** *A randomized algorithm $\mathcal{A}(\cdot)$ is an $(\rho, \eta)$-close approximation to the top eigenvector if for all data sets $\mathcal{D}$ of $n$ points,*

$$\mathbb{P}\left(q_{\mathrm{A}}(\mathcal{A}(\mathcal{D})) \geq \rho\right) \geq 1 - \eta, \tag{8}$$

*where the probability is taken over $\mathcal{A}(\cdot)$.*

In this paper we are interested in proving results on the sample complexity of differentially private algorithms that approximate PCA. That is, for a given $\alpha$ and $\rho$, how large must the number of individuals $n$ in the data set be such that it is $\alpha$-differentially private and also a $(\rho, \eta)$-close approximation to PCA? It is well known that as the number of individuals $n$ grows, it is easier to guarantee the same level of privacy with relatively less noise or perturbation, and therefore the utility of the approximation also improves. Our results characterize how privacy and utility scale with $n$ and the tradeoff between them for fixed $n$.

# 3 Algorithms and main results

In this section we describe differentially private techniques for approximating (2). The first is a modified version of the SULQ method [Blum et al., 2005]. Our new algorithm for differentially-private PCA, PPCA, is an instantiation of the exponential mechanism due to McSherry and Talwar [2007]. Both procedures provide differentially private approximations to the top-$k$ subspace: SULQ provides $(\alpha, \delta)$-differential privacy and PPCA provides $\alpha$-differential privacy.

## 3.1 Input perturbation

The only differentially-private approximation to PCA prior to this work is the SULQ method [Blum et al., 2005]. The SULQ method perturbs each entry of the empirical second moment matrix $A$ to ensure differential privacy and releases the top $k$ eigenvectors of this perturbed matrix. In particular, SULQ recommends adding a matrix $N$ of i.i.d. Gaussian noise of variance $\frac{8d^2 \log^2(d/\delta)}{n^2\alpha^2}$ and applies the PCA algorithm to $A+N$. This guarantees slightly a weaker privacy definition known as $(\alpha, \delta)$-differential privacy. One problem with this approach is that with probability 1 the matrix $A+N$ is not symmetric, so the largest eigenvalue may not be real and the entries of the corresponding eigenvector may be complex. Thus the SULQ algorithm is not a good candidate for practical privacy-preserving dimensionality reduction.

However, a simple modification to the basic SULQ approach does guarantee $(\alpha, \delta)$ differential privacy. Instead of adding a asymmetric Gaussian matrix, the algorithm can add the a symmetric matrix with i.i.d. Gaussian entries $N$. That is, for $1 \leq i \leq j \leq d$, the variable $N_{ij}$ is an independent Gaussian random variable with variance $\beta^2$. Note that this matrix is symmetric but not necessarily positive semidefinite, so some eigenvalues may be negative but the eigenvectors are all real. A derivation for the noise variance is given in Theorem 1.

**Algorithm 1:** Algorithm MOD-SULQ (input pertubation)

**inputs**: $d \times n$ data matrix $X$, privacy parameter $\alpha$, parameter $\delta$

**outputs**: $d \times k$ matrix $\hat{V}_k = [\hat{v}_1 \ \hat{v}_2 \ \cdots \ \hat{v}_k]$ with orthonormal columns

Set $A = \frac{1}{n} X X^T$.

Set

$$\beta = \frac{d+1}{n\alpha} \sqrt{2 \log \left( \frac{d^2 + d}{\delta 2 \sqrt{2\pi}} \right)} + \frac{1}{\sqrt{\alpha} n}. \tag{9}$$

Generate a $d \times d$ symmetric random matrix $N$ whose entries are i.i.d. drawn from $\mathcal{N}\left(0, \beta^2\right)$.

Compute $\hat{V}_k = V_k(A + N)$ according to (2).

### 3.2 Exponential mechanism

Our new method, PPCA, randomly samples a $k$-dimensional subspace from a distribution that ensures differential privacy and is biased towards high utility. The distribution from which our released subspace is sampled is known in the statistics literature as the matrix Bingham distribution [Bingham, 1974, Chikuse, 2003], which we denote by $\mathsf{BMF}_k(B)$. The algorithm is in terms of general $k < d$ but our theoretical results focus on the special case $k = 1$ where we wish to release a one-dimensional approximation to the data covariance matrix. The matrix Bingham distribution takes values on the set of all $k$-dimensional subspaces of $\mathbb{R}^d$ and has a density equal to

$$f(V) = \frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}d, B\right)} \exp(\mathrm{tr}(V^T B V)), \tag{10}$$

where $V$ is a $d \times k$ matrix whose columns are orthonormal and ${}_1F_1\left(\frac{1}{2}k, \frac{1}{2}d, B\right)$ is a confluent hypergeometric function [Chikuse, 2003, p.33]. Numerical approximations of this constant can be calculated using a variety of techniques [Butler and Wood, 2002, Koev and Edelman, 2006].

**Algorithm 2:** Algorithm PPCA (exponential mechanism)

**inputs**: $d \times n$ data matrix $X$, privacy parameter $\alpha$, dimension $k$

**outputs**: $d \times k$ matrix $\hat{V}_k = [\hat{v}_1 \ \hat{v}_2 \ \cdots \ \hat{v}_k]$ with orthonormal columns

Set $A = \frac{1}{n} X X^T$

Sample $\hat{V}_k = \mathsf{BMF}\left(n \frac{\alpha}{2} A\right)$

By combining results on the exponential mechanism [McSherry and Talwar, 2007] along with properties of PCA algorithm, we can show that this procedure is differentially private. In many cases, sampling from the distribution specified by the exponential mechanism distribution may be difficult computationally, especially for continuous-valued outputs. We implement PPCA using a recently-proposed Gibbs sampler due to Hoff [2009]. Gibbs sampling is a popular Markov Chain Monte Carlo (MCMC) technique in which samples are generated according to a Markov chain whose stationary distribution is the density in (10). Assessing the "burn-in time" and other factors for this procedure is an interesting question in its own right; further details are in Section 6.2.

### 3.3 Other approaches

There are other general algorithmic strategies for guaranteeing differential privacy. The sensitivity method [Dwork et al., 2006b] adds noise proportional to the maximum change that can be induced by changing a single point in the data set. Consider a data set $\mathcal{D}$ with $m+1$ copies of a unit vector $u$ and $m$ copies of a unit vector $u'$ with $u \perp u'$ and let $\mathcal{D}'$ have $m$ copies of $u$ and $m+1$ copies of $u'$. Then $v_1(\mathcal{D}) = u$ but $v_1(\mathcal{D}') = u'$, so $\|v_1(\mathcal{D}) - v_1(\mathcal{D}')\| = \sqrt{2}$. Thus the global sensitivity does not scale with the number of data points, so as $n$ increases the variance of the noise required by the Laplace mechanism [Dwork et al., 2006b] will not decrease. An alternative to global sensitivity is smooth sensitivity [Nissim et al., 2007]; except for special cases, such as the sample median, smooth sensitivity is difficult to compute for general functions. A third method for computing private, approximate solutions to high-dimensional optimization problems is objective perturbation [Chaudhuri et al., 2011]; to apply this method, we require the optimization problems to have certain properties (namely, strong convexity and bounded norms of gradients), which do not apply to PCA.

### 3.4 Main results

Our theoretical results are sample complexity bounds for PPCA and MOD-SULQ as well as a general lower bound on the sample complexity for any $\alpha$-differentially private algorithm. These results show that the PPCA is nearly optimal in terms the scaling of the sample complexity with respect to the data dimension $d$, privacy parameter $\alpha$, and correlation $\rho$. We further show that MOD-SULQ requires more samples as a function of $d$, despite having a slightly weaker privacy guarantee.

Even though both algorithms can output the top-$k$ PCA subspace for general $k \leq d$, we prove results for the case $k = 1$. Proving analogous results for general $k$ would require significantly more analysis and calculations, whereas for $k = 1$ the scaling behavior is cleaner to elucidate. Finding the scaling behavior of the sample complexity with $k$ is an interesting open problem that we leave for future work.

The fact that these two algorithms are differentially private follows from some simple calculations.

**Theorem 1.** *For the $\beta$ in (9), the MOD-SULQ algorithm is $(\alpha, \delta)$ differentially private.*

**Theorem 2.** *Algorithm PPCA is $\alpha$-differentially private.*

Our first sample complexity result provides an upper bound on the number of samples required by PPCA to guarantee a certain level of privacy and accuracy. The sample complexity of PPCA $n$ grows linearly with the dimension $d$, inversely with $\alpha$, and inversely with the correlation gap $(1 - \rho)$ and eigenvalue gap $\lambda_1(A) - \lambda_2(A)$.

**Theorem 3** (Sample complexity of PPCA)**.** *If*

$$n > \frac{d}{\alpha(1-\rho)(\lambda_1 - \lambda_2)} \left( \frac{\log(1/\eta)}{d} + \log \frac{4\lambda_1}{(1-\rho^2)(\lambda_1 - \lambda_2)} \right), \tag{11}$$

*then the top PCA direction $v_1$ and the output of PPCA $\hat{v}_1$ with privacy parameter $\alpha$ satisfy:*

$$\Pr(|\langle v_1, \hat{v}_1 \rangle| > \rho) \geq 1 - \eta.$$

*That is, PPCA is a $(\rho, \eta)$-close approximation to PCA.*

Our second result shows a lower bound on the number of samples required by *any* $\alpha$-differentially-private algorithm to guarantee a certain level of accuracy for a large class of datasets.

**Theorem 4** (Sample complexity lower bound)**.** *There are absolute constants $c$ and $c'$ such that if*

$$n < c \cdot \frac{d}{\alpha(1-\rho)} - c' \cdot \frac{1}{\alpha(1-\rho)}, \tag{12}$$

*then for any $\alpha$-differentially private algorithm $\mathcal{A}$ which computes an approximation $\hat{v}_1$ to the top PCA direction, there exists a data set for which $\lambda_1 - \lambda_2 \geq \frac{1}{2}$, and yet*

$$\mathbb{E}_{\mathcal{A}} \left[ |\langle v_1, \hat{v}_1 \rangle| \right] \leq \rho.$$

Theorem 3 shows that if $n$ scales like $\frac{d}{\alpha(1-\rho)}$ then PPCA produces an approximation $\hat{v}_1$ that has correlation $\rho$ with $v_1$, whereas Theorem 4 shows that $n$ must scale like $\frac{d}{\alpha(1-\rho)}$ for any $\alpha$-differentially private algorithm. In terms of scaling with $d$ and $\alpha$ the upper and lower bounds match, and they also match up to logarithmic factors with respect to the correlation. By contrast, the following lower bound on the number of samples required by MOD-SULQ to ensure a certain level of accuracy shows that MOD-SULQ has a less favorable scaling with dimension.

**Theorem 5** (Sample complexity lower bound for MOD-SULQ)**.** *There are constants $c$ and $c'$ such that if*

$$n < c \frac{d^{3/2} \sqrt{\log(d/\delta)}}{\alpha} (1 - c'(1-\rho)),$$

*then there exists a dataset of size $n$ in dimension $d$, such that the top PCA direction $v$ and the output $\hat{v}$ of MOD-SULQ satisfy*

$$\mathbf{E}[|\langle \hat{v}_1, v_1 \rangle|] \leq \rho.$$

Notice that the dependence on $n$ grows as $d^{3/2}$ in SULQ as opposed to $d$ in PPCA. Dimensionality reduction via PCA is often used in applications where the data points occupy a low dimensional space but are presented in high dimensions. These bounds suggest that PPCA is better suited to such applications than MOD-SULQ. We next turn to validating this intuition on real data. In the next two section we provide proofs of these results.

# 4 Analysis of PPCA

In this section we provide theoretical guarantees on the performance of PPCA. The proof of Theorem 2 follows from the results on the exponential mechanism [McSherry and Talwar, 2007]. To find the sample complexity of PPCA we bound the density of the Bingham distribution, leading to a sample complexity for $k = 1$ that depends on the gap $\lambda_1 - \lambda_2$ between the top two eigenvalues. We close with a general lower bound on the sample complexity that holds for any $\alpha$-differentially private algorithm. The lower bound matches our upper bound up to log factors, showing that PPCA is nearly optimal in terms of the scaling with dimension, privacy $\alpha$, and utility $q_{\mathrm{A}}(\cdot)$.

## 4.1 Privacy guarantee

*Proof of Theorem 2.* Let $X$ be a data matrix whose $i$-th column is $x_i$ and $A = \frac{1}{n} X X^T$. The PP-PCA algorithm is the exponential mechanism of McSherry and Talwar [2007] applied to the score function

$$q_{\mathrm{F}}(X, v) = n \cdot v^T A v$$

Consider $X' = [x_1 \; x_2 \; \cdots \; x_{n-1} \; x'_n]$ differ from $X$ in a single column and let $A' = \frac{1}{n} X' X'^T$. We have

$$\begin{aligned} \max_{v \in \mathbb{S}^{d-1}} \left| q_{\mathrm{F}}(X', v) - q_{\mathrm{F}}(X, v) \right| &\leq \left| v^T (x'_n x'^T_n - x_n x_n^T) v \right| \\ &\leq \left| \left\| v^T x'_n \right\|^2 - \left\| v^T x_n \right\|^2 \right| \\ &\leq 1. \end{aligned}$$

The result then follows immediately from the results of McSherry and Talwar [2007, Theorem 6]. □

## 4.2 Upper bound on utility

The results of McSherry and Talwar [2007] bound the gap between the value of the function $q_{\mathrm{F}}(\hat{v}_1) = n \cdot \hat{v}_1^T A \hat{v}_1$ evaluated at the output $\hat{v}_1$ of the mechanism and the optimal value $q(v_1) = n \cdot \lambda_1$. We derive a bound on the correlation $q_{\mathrm{A}}(\hat{v}_1) = |\langle \hat{v}_1, v_1 \rangle|$ via geometric arguments. To do this we need the a lower bound on the area of a circular cap.

**Lemma 6** (Lemmas 2.2 and 2.3 of [Ball, 1997])**.** *Let $\mu$ be the uniform measure on the unit sphere $\mathbb{S}^{d-1}$. For any $x \in \mathbb{S}^{d-1}$ and $0 \leq c < 1$*

$$\frac{1}{2} \exp\left( -\frac{d-1}{2} \log \frac{2}{1-c} \right) \leq \mu\left( \left\{ v \in \mathbb{S}^{d-1} : \langle v, x \rangle \geq c \right\} \right) \leq \exp\left( -dc^2/2 \right). \qquad (13)$$

*Proof of Theorem 3.* Fix a privacy level $\alpha$, target correlation $\rho$, and probability $\eta$. Let $X$ be the data matrix and $B = (\alpha/2) X X^T$ and

$$\mathcal{U}_\rho = \{ u : |\langle u, v_1 \rangle| \geq \rho \}.$$

be the union of the two spherical caps centered at $\pm v_1$. Let $\overline{\mathcal{U}}_\rho$ denote the complement of $\mathcal{U}_\rho$ in $\mathbb{S}^{d-1}$.

An output vector $\hat{v}_1$ is "good" if it is in $\mathcal{U}_\rho$. We first give some bounds on the score function $q_{\mathrm{F}}(u)$ on the boundary between $\mathcal{U}_\rho$ and $\overline{\mathcal{U}}_\rho$, where $\langle u, v_1 \rangle = \pm\rho$. The function $q_{\mathrm{F}}(u)$ is maximized when $u$ is a linear combination of $v_1$ and $v_2$, the top two eigenvectors of $A$. It minimized when $u$ is a linear combination of $v_1$ and $v_d$. Therefore

$$q_{\mathrm{F}}(u) \le \frac{n\alpha}{2}(\rho^2\lambda_1 + (1-\rho^2)\lambda_2) \qquad u \in \overline{\mathcal{U}}_\rho \tag{14}$$

$$q_{\mathrm{F}}(u) \ge \frac{n\alpha}{2}(\rho^2\lambda_1 + (1-\rho^2)\lambda_d) \qquad u \in \mathcal{U}_\rho. \tag{15}$$

Let $\mu(\cdot)$ denote the uniform measure on the unit sphere. Then fixing an $0 \le b < 1$, using (14), (15), and the fact that $\lambda_d \ge 0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right) &\le \frac{\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right)}{\mathbb{P}\left(\mathcal{U}_\sigma\right)} \\
&= \frac{\frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}m, B\right)} \int_{\overline{\mathcal{U}}_\rho} \exp\left(u^T B u\right) d\mu}{\frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}m, B\right)} \int_{\mathcal{U}_\sigma} \exp\left(u^T B u\right) d\mu} \\
&\le \frac{\exp\left(n(\alpha/2)\left(\rho^2\lambda_1 + (1-\rho^2)\lambda_2\right)\right) \cdot \mu\left(\overline{\mathcal{U}}_\rho\right)}{\exp\left(n(\alpha/2)\left(\sigma^2\lambda_1 + (1-\sigma^2)\lambda_d\right)\right) \cdot \mu\left(\mathcal{U}_\sigma\right)} \\
&\le \exp\left(-\frac{n\alpha}{2}\left(\sigma^2\lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2)\right)\right) \cdot \frac{\mu\left(\overline{\mathcal{U}}_\rho\right)}{\mu\left(\mathcal{U}_\sigma\right)}.
\end{aligned} \tag{16}$$

Applying the lower bound from Lemma 6 to the denominator of (16) and the upper bound $\mu\left(\overline{\mathcal{U}}_\rho\right) \le 1$ yields

$$\log \frac{1}{\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right)} \le \frac{n\alpha}{2}\left(\sigma^2\lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2)\right) - \frac{d-1}{2}\log\frac{2}{1-\sigma}. \tag{17}$$

We must choose a $\sigma^2 > \rho^2$ to make the exponent positive, but more precisely,

$$\begin{aligned}
\sigma^2 &> \rho^2 + (1-\rho^2)\frac{\lambda_2}{\lambda_1} \\
1 - \sigma^2 &< (1-\rho^2)\left(1 - \frac{\lambda_2}{\lambda_1}\right)
\end{aligned}$$

For simplicity, choose

$$1 - \sigma^2 = \frac{1}{2}(1-\rho^2)\left(1 - \frac{\lambda_2}{\lambda_1}\right).$$

So that

$$\begin{aligned}
\sigma^2\lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2) &= (1-\rho^2)\lambda_1 - (1-\sigma^2)\lambda_1 - (1-\rho^2)\lambda_2 \\
&= (1-\rho^2)\left(\lambda_1 - \frac{1}{2}(\lambda_1 - \lambda_2) - \lambda_2\right) \\
&= \frac{1}{2}(1-\rho^2)(\lambda_1 - \lambda_2),
\end{aligned}$$

and

$$\log \frac{2}{1-\sigma} < \log \frac{2}{1-\sigma^2}$$
$$= \log \frac{4\lambda_1}{(1-\rho^2)(\lambda_1 - \lambda_2)}.$$

Setting (17) greater than $\log(1/\eta)$ yields

$$\frac{n\alpha}{4}(1-\rho^2)(\lambda_1 - \lambda_2) > \log \frac{1}{\eta} + \frac{d-1}{2} \log \frac{4\lambda_1}{(1-\rho^2)(\lambda_1 - \lambda_2)}.$$

Since $1 - \rho < 1 - \rho^2$, if we choose

$$n > \frac{d}{\alpha(1-\rho)(\lambda_1 - \lambda_2)} \left( \frac{\log(1/\eta)}{d} + \log \frac{4\lambda_1}{(1-\rho^2)(\lambda_1 - \lambda_2)} \right),$$

then the output of $\mathsf{PPCA}$ will produce a $\hat{v}_1$ such that

$$\mathbb{P}\left(|\langle \hat{v}_1, v_1 \rangle| < \rho\right) \leq \eta.$$

□

The bound proved in the preceding theorem may be a little loose. It is difficult to measure the area on the unit sphere of the set $\{x : x^T A x \geq 1 - \gamma\}$. In the case where $A = I$ this is just a spherical cap, but for general $A$ it can have a more irregular shape. This is an artifact of the proof technique. A direct bound seems difficult because explicit bounds on the confluent hypergeometric function with matrix argument are not readily applicable; numerical approximation is still an open research question [Butler and Wood, 2002, Koev and Edelman, 2006].

### 4.3 Lower bound on utility

We now turn to a general lower bound on the sample complexity for any differentially private approximation to PCA. We construct $K$ databases which differ in a small number of points whose top eigenvectors are not too far from each other; yet the gap $\lambda_1 - \lambda_2$ between the top two eigenvalues is at least $\frac{1}{2}$. For such a collection, Lemma 7 shows that for any differentially private mechanism, the average correlation over the collection cannot be too large. That is, even if the eigengap between the first two eigenvalues is large, any $\alpha$-differentially private mechanism cannot have high utility on all $K$ data sets. The remainder of the argument is to construct these $K$ data sets.

The proof uses some simple eigenvalue and eigenvector computations. A matrix of positive entries

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \tag{18}$$

has characteristic polynomial

$$\det(A - \lambda I) = \lambda^2 - (a+c)\lambda + (ac - b^2)$$

and eigenvalues

$$\lambda = \frac{1}{2}(a+c) \pm \frac{1}{2}\sqrt{(a+c)^2 - 4(ac-b^2)}$$
$$= \frac{1}{2}(a+c) \pm \frac{1}{2}\sqrt{(a-c)^2 + 4b^2}. \tag{19}$$

The eigenvectors are in the directions $(b, -(a-\lambda))^T$.

**Lemma 7.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ be $K$ databases which differ in the value of at most $\frac{\ln(K-1)}{\alpha}$ points, and let $u_1, \ldots, u_K$ be the top eigenvectors of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. If $\mathcal{A}$ is any $\alpha$-differentially private algorithm, then,*

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}}\left[|\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle|\right] \leq K\left(1 - \frac{1}{16}(1 - \max|\langle u_i, u_j \rangle|)\right).$$

*Proof.* Let

$$t = \min_{i \neq j}(\|u_i - u_j\|, \|u_i + u_j\|),$$

and $\mathcal{G}_i$ be the cap around $\pm u_i$ of radius $t/2$:

$$\mathcal{G}_i = \{u : \|u - u_i\| < t/2\} \cup \{u : \|u + u_i\| < t/2\}.$$

We claim that

$$\sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \geq \frac{1}{2}(K-1). \tag{20}$$

The proof is by contradiction. Suppose the claim is false. Because all of the caps $\mathcal{G}_i$ are disjoint, and applying the definition of differential privacy,

$$\begin{aligned}
\frac{1}{2}(K-1) &> \sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \\
&\geq \sum_{i=1}^{K} \sum_{i' \neq i} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \in G_{i'}) \\
&\geq \sum_{i=1}^{K} \sum_{i' \neq i} e^{-\alpha \cdot \ln(K-1)/\alpha} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_{i'}) \in G_{i'}) \\
&\geq (K-1) \cdot \frac{1}{K-1} \cdot \sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \in G_i) \\
&\geq K - \frac{1}{2}(K-1),
\end{aligned}$$

which is a contradiction, so (20) holds. Therefore by the Markov inequality

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[ \min(\|\mathcal{A}(\mathcal{D}_i) - u_i\|^2, \|\mathcal{A}(\mathcal{D}_i) + u_i\|^2) \right] \geq \sum_{i=1}^{K} \mathbb{P}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \cdot \frac{t^2}{4}$$
$$\geq \frac{1}{8}(K-1)t^2.$$

Rewriting the norms in terms of inner products shows

$$2K - 2\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[ |\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \geq \frac{1}{8}(K-1)\left(2 - 2\max|\langle u_i, u_j \rangle|\right),$$

so

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[ |\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \leq K \left( 1 - \frac{1}{8} \frac{K-1}{K} (1 - \max|\langle u_i, u_j \rangle|) \right)$$
$$\leq K \left( 1 - \frac{1}{16} (1 - \max|\langle u_i, u_j \rangle|) \right).$$

□

*Proof of Theorem 4.* Let $y = e_d$, the $d$-th coordinate vector, and let $\phi \in ((2\pi d)^{-1/2}, 1)$. Lemma 12 shows that there exists a packing $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ of the sphere $\mathbb{S}^{d-2}$ spanned by $\{e_1, e_2, \ldots, e_{d-1}\}$ such that $\max_{i \neq j} |\langle w_i, w_j \rangle| \leq \phi$, where

$$K = \frac{1}{8}(1-\phi)^{-(d-2)/2}.$$

Let $\beta = \frac{\ln(K-1)}{n\alpha}$. For $i = 1, 2, \ldots, K$ let the data set $\mathcal{D}_i$ contain

- $n(1-\beta)$ copies of $y$
- $n\beta$ copies of $z_i = \sqrt{\beta} y + \sqrt{1-\beta} w_i$.

The second moment matrix $A_i$ of $\mathcal{D}_i$ is

$$A_i = (1 - \beta + \beta^2) y y^T + \beta\sqrt{\beta - \beta^2}(w_i^T y + y w_i^T) + \beta(1-\beta) w_i w_i^T.$$

By choosing an basis containing $y$ and $w_i$, we can write this as

$$A_i = \begin{bmatrix} 1 - \beta + \beta^2 & \beta\sqrt{\beta - \beta^2} & \mathbf{0} \\ \beta\sqrt{\beta - \beta^2} & \beta(1-\beta) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

This is in the form (18), with $a = 1 - \beta + \beta^2$, $b = \beta\sqrt{\beta - \beta^2}$, and $c = \beta(1-\beta)$. The top eigenvector is therefore

$$u_i = \frac{b}{\sqrt{b^2 + (a-\lambda)^2}} y + \frac{(a-\lambda)}{\sqrt{b^2 + (a-\lambda)^2}} w_i.$$

where $\lambda$ is given by (19). We observe that the gap between the top two eigenvalues is:

$$\sqrt{(a-c)^2+4b^2} = \sqrt{(1-2\beta(1-\beta))^2+4\beta^2(\beta-\beta^2)} \geq \frac{1}{2}$$

Therefore

$$\begin{aligned}\max_{i\neq j}|\langle u_i,u_j\rangle| &\leq \frac{b^2}{b^2+(a-\lambda)^2}+\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}\max_{i\neq j}|\langle w_i,w_j\rangle| \\ &\leq 1-\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}(1-\phi). \end{aligned} \tag{21}$$

To continue the upper bound we compute $\lambda$:

$$\begin{aligned}\lambda &= \frac{1}{2}(a+c)+\frac{1}{2}\sqrt{(a-c)^2+4b^2} \\ &= \frac{1}{2}\left(1+\sqrt{(1-2\beta+2\beta^2)^2+4(\beta^3-\beta^4)}\right) \\ &= \frac{1}{2}\left(1+\sqrt{1+4\beta^2+4\beta^4-4\beta+4\beta^2-8\beta^3+4\beta^3-4\beta^4}\right) \\ &= \frac{1}{2}\left(1+\sqrt{1-4\beta+8\beta^2-4\beta^3}\right) \\ &= \frac{1}{2}\left(1+\sqrt{(1-4\beta(1-\beta)^2}\right).\end{aligned}$$

A Taylor series expansion of $\sqrt{1-x}$ implies that for $x<1$,

$$1-x \leq \sqrt{1-x} \leq 1-\frac{x}{2}.$$

Applying this equation on $\lambda$, we get:

$$1-2\beta(1-\beta)^2 \leq \lambda \leq 1-\beta(1-\beta)^2. \tag{22}$$

Note that

$$\lambda > \frac{1}{2}(a+c)+\frac{1}{2}\sqrt{(a-c)^2} = a,$$

so $(a-\lambda)^2$ is maximized when $\lambda$ is as large as possible. Therefore

$$\begin{aligned}(a-\lambda)^2 &\leq (1-\beta+\beta^2-1+\beta(1-\beta)^2)^2 \\ &= (-\beta(1-\beta)+\beta(1-\beta)^2)^2 \\ &= \beta^4(1-\beta)^2.\end{aligned}$$

Moreover,

$$b^2 = \beta^2(\beta-\beta^2) = \beta^3(1-\beta).$$

Since for $\nu>0$, $x/(\nu+x)$ is monotonically increasing in $x$, we have

$$\begin{aligned}\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2} &\leq \frac{\beta^4(1-\beta)^2}{\beta^4(1-\beta)^2+\beta^3(1-\beta)} \\ &\leq \frac{\beta(1-\beta)}{1+\beta(1-\beta)} \\ &\leq \beta.\end{aligned}$$

Combining this with Equation 21, we get:

$$\max_{i \neq j} |\langle u_i, u_j \rangle| \leq 1 - \beta(1 - \phi).$$

Applying Lemma 7 shows that for at least one database $i$

$$\begin{aligned} \mathbb{E}_{\mathcal{A}} \left[ |\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] &\leq 1 - \frac{1}{16}(1 - (1 - \beta(1 - \phi))) \\ &= 1 - \frac{\beta}{16}(1 - \phi). \end{aligned}$$

Setting the left side equal to some target expected correlation $\rho$,

$$\begin{aligned} (1 - \rho) &\geq \frac{\beta}{16}(1 - \phi) \\ &= \frac{\ln(K-1)}{16 n \alpha}(1 - \phi), \end{aligned}$$

Thus

$$n > \frac{1}{\alpha(1-\rho)} \left( \frac{1-\phi}{16} \left( d \log \frac{1}{\sqrt{1-\phi}} - \log \frac{1}{1-\phi} - \log 16 \right) \right).$$

Setting $\phi$ as some constant close to 1 shows that there exist constants $c'$ and $c''$ such that

$$n > c' \frac{d}{\alpha(1-\rho)} - c'' \frac{1}{\alpha(1-\rho)}.$$

□

# 5 Analysis of MOD-SULQ

In this section we provide theoretical guarantees on the performance of the MOD-SULQ algorithm. Theorem 1 shows that MOD-SULQ is $(\alpha, \delta)$-differentially private. Theorem 14 provides a lower bound on the distance between the vector released by MOD-SULQ and the true top eigenvector in terms of the privacy parameters $\alpha$ and $\delta$ and the number of points $n$ in the data set. This implicitly gives a lower bound on the sample complexity of MOD-SULQ. We provide some graphical illustration of this tradeoff.

The following upper bound will be useful for future calculations : for two unit vectors $x$ and $y$,

$$\sum_{1 \leq i \leq j \leq d} (x_i x_j - y_i y_j)^2 \leq 2. \tag{23}$$

Note that this upper bound is achievable by setting $x$ and $y$ to be orthogonal elementary vectors.

### 5.1 Privacy guarantee

We first justify the choice of $\beta^2$ in the MOD-SULQ algorithm.

*Proof of Theorem 1.* Let $B$ and $\hat{B}$ be two independent symmetric random matrices where $\{B_{ij} : 1 \le i \le j \le d\}$ and $\{\hat{B}_{ij} : 1 \le i \le j \le d\}$ are each sets of i.i.d. Gaussian random variables with mean 0 and variance $\beta^2$. Consider two data sets $\mathcal{D} = \{x_i : i \in [n]\}$ and $\hat{\mathcal{D}} = \mathcal{D}_1 \cup \{\hat{x}_n\} \setminus \{x_n\}$ and let $A$ and $\hat{A}$ denote their second moment matrices. Let $G = A + B$ and $\hat{G} = \hat{A} + \hat{B}$. We first calculate the log ratio of the densities of $G$ and $\hat{G}$ at a point $H$:

$$\begin{aligned}\log \frac{f_G(H)}{f_{\hat{G}}(H)} &= \sum_{1 \le i \le j \le d} \left( -\frac{1}{2\beta^2}(H_{ij} - A_{ij})^2 + \frac{1}{2\beta^2}(H_{ij} - \hat{A}_{ij})^2 \right) \\ &= \frac{1}{2\beta^2} \sum_{1 \le i \le j \le d} \left( \frac{2}{n}(H_{ij} - A_{ij})(x_{n,i}x_{n,j} - \hat{x}_{n,i}\hat{x}_{n,j}) + \frac{1}{n^2}(\hat{x}_{n,i}\hat{x}_{n,j} - x_{n,i}x_{n,j})^2 \right).\end{aligned}$$

From (23) the last term is upper bounded by $2/n^2$. To upper bound the first term,

$$\begin{aligned}\sum_{1 \le i \le j \le d} |\hat{x}_{n,i}\hat{x}_{n,j} - x_{n,i}x_{n,j}| &\le 2 \max_{a:\|a\| \le 1} \sum_{1 \le i \le j \le d} a_i a_j \\ &\le 2 \cdot \frac{1}{2}(d^2 + d) \cdot \frac{1}{d} \\ &= d + 1.\end{aligned}$$

Note that this bound is not too loose – by taking $\hat{x} = d^{-1/2}\mathbf{1}$ and $x = (1, 0, \ldots, 0)^T$, this term is still linear in $d$.

Then for any measurable set $\mathcal{S}$ of matrices,

$$\mathbb{P}(G \in \mathcal{S}) \le \exp\left( \frac{1}{2\beta^2} \left( \frac{2}{n}(d+1)\gamma + \frac{3}{n^2} \right) \right) \mathbb{P}\left( \hat{G} \in \mathcal{S} \right) + \mathbb{P}(B_{ij} > \gamma \text{ for all } i, j). \quad (24)$$

To handle the last term, use a union bound over the $(d^2 + d)/2$ variables $\{B_{ij}\}$ together with the tail bound, which holds for $\gamma > \beta$:

$$\mathbb{P}(B_{ij} > \gamma) \le \frac{1}{\sqrt{2\pi}} e^{-\gamma^2/2\beta^2}.$$

Thus setting $\mathbb{P}(B_{ij} > \gamma \text{ for some } i, j) = \delta$ yields the condition

$$\delta = \frac{d^2 + d}{2\sqrt{2\pi}} e^{-\gamma^2/2\beta^2}.$$

Rearranging to solve for $\gamma$ gives

$$\gamma = \max\left( \beta, \beta \sqrt{2 \log\left( \frac{d^2 + d}{\delta 2\sqrt{2\pi}} \right)} \right) = \beta \sqrt{2 \log\left( \frac{d^2 + d}{\delta 2\sqrt{2\pi}} \right)}$$

for $d > 1$ and $\delta < 3/\sqrt{2\pi e}$. This then gives an expression for $\alpha$ to make (24) imply $(\alpha, \delta)$ differential privacy:

$$\alpha = \frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\gamma + \frac{2}{n^2}\right)$$
$$= \frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\beta\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{2}{n^2}\right).$$

Solving for $\beta$ using the quadratic formula yields a particularly messy expression:

$$\beta = \frac{d+1}{2n\alpha}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{2n\alpha}\left(2(d+1)^2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right) + 4\alpha\right)^{1/2}$$
$$\leq \frac{d+1}{n\alpha}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{\sqrt{\alpha}n}. \tag{25}$$

□

## 5.2 Lower bound on utility

The main tool in our lower bound is a generalization by Yu [1997] of an information-theoretic inequality due to Fano.

**Theorem 8** (Fano's inequality [Yu, 1997]). *Let $\mathcal{R}$ be a set and $\Theta$ be a parameter space with a pseudo-metric $d(\cdot)$. Let $\mathcal{F}$ be a set of $r$ densities $\{f_1, \ldots, f_r\}$ on $\mathcal{R}$ corresponding to parameter values $\{\theta_1, \ldots, \theta_r\}$ in $\Theta$. Let $X$ have distribution $f \in \mathcal{F}$ with corresponding parameter $\theta$ and let $\hat{\theta}(X)$ be an estimate of $\theta$. If, for all $i$ and $j$*

$$d(\theta_i, \theta_j) \geq \tau \tag{26}$$

*and*

$$\mathbf{KL}\left(f_i \| f_j\right) \leq \gamma, \tag{27}$$

*then*

$$\max_j \mathbb{E}_j[d(\hat{\theta}, \theta_j)] \geq \frac{\tau}{2}\left(1 - \frac{\gamma + \log 2}{\log r}\right), \tag{28}$$

*where $\mathbb{E}_j[\cdot]$ denotes the expectation with respect to distribution $f_j$.*

To use this inequality, we will construct a set of densities on the set of covariance matrices corresponding distribution of the random matrix in the MOD-SULQ algorithm under different inputs. These inputs will be chosen using a set of unit vectors which are a packing on the surface of the unit sphere.

### 5.3 A packing lemma

The proof of this lemma is relatively straightforward. The following is a slight refinement of a lemma due to Csiszár and Narayan [1988] (See also Csiszár and Narayan [1991]).

**Lemma 9.** *Let* $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ *be arbitrary random variables and let* $f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)$ *be arbitrary with* $0 \le f_i \le 1$*,* $i = 1, 2, \ldots, N$*. Then the condition*

$$\mathbb{E}\left[f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)|\mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}\right] \le a_i \text{ a.s.}, \qquad i = 1, 2, \ldots, N \tag{29}$$

*implies that*

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \le \exp\left(-Nt(\log 2) + \sum_{i=1}^{N} a_i\right). \tag{30}$$

*Proof.* First apply Markov's inequality:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \\
&\qquad = \mathbb{P}\left(2^{\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)} > 2^{Nt}\right) \\
&\qquad \le 2^{-Nt}\mathbb{E}\left[2^{\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)}\right] \\
&\qquad \le 2^{-Nt}\mathbb{E}\left[2^{\sum_{i=1}^{N-1} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)}\mathbb{E}\left[2^{f_N(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)}|\mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\right]\right].
\end{aligned}$$

Now note that for $b \in [0, 1]$ we have $2^b \le 1 + b$, so

$$\begin{aligned}
\mathbb{E}\left[2^{f_N(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)}|\mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\right] &\le \mathbb{E}\left[1 + f_N(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)|\mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\right] \\
&\le (1 + a_N) \\
&\le \exp(a_N).
\end{aligned}$$

Therefore

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \le \exp(-Nt(\log 2) + a_N)\mathbb{E}\left[2^{\sum_{i=1}^{N-1} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)}\right].$$

Continuing in the same way yields

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \le \exp\left(-Nt(\log 2) + \sum_{i=1}^{N} a_i\right).$$

□

The second technical lemma [Csiszár and Narayan, 1991, Lemma 2] is a basic result about the distribution of inner product between a randomly chosen unit vector and any other fixed vector. It is a consequence of a result of Shannon [1959] on the distribution of the angle between a uniformly distributed unit vector and a fixed unit vector.

**Lemma 10** (Lemma 2 of [Csiszár and Narayan, 1991])**.** *Let* $\mathbf{U}$ *be uniformly distributed on the unit sphere* $\mathbb{S}^{d-1}$ *in* $\mathbb{R}^d$*. Then for every unit vector* $\mathbf{u}$ *on this sphere and any* $\phi \in [(2\pi d)^{-1/2}, 1)$*, the following inequality holds:*

$$\mathbb{P}\left(\langle \mathbf{U}, \mathbf{u} \rangle \geq \phi\right) \leq (1-\phi^2)^{(d-1)/2}. \tag{31}$$

**Lemma 11** (Packing set on the unit sphere)**.** *Let a dimension* $d$ *and* $\phi \in [(2\pi d)^{-1/2}, 1)$ *be given. For* $N$ *and* $t$ *satisfying*

$$-Nt(\log 2) + N(N-1)(1-\phi^2)^{(d-1)/2} < 1, \tag{32}$$

*there exists a set of* $K = \lfloor (1-t)N \rfloor$ *unit vectors* $\mathcal{C}$ *such that for all distinct pairs* $\mu, \nu \in \mathcal{C}$*,*

$$|\langle \mu, \nu \rangle| < \phi. \tag{33}$$

*Proof.* The goal is to generate a set of $N$ unit vectors on the surface of the sphere $\mathbb{S}^{d-1}$ such that they have large pairwise distances, or correspondingly small pairwise inner products. To that end, define $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ i.i.d. uniformly distributed on $\mathbb{S}^{d-1}$ and

$$f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) = \mathbf{1}\left(|\langle \mathbf{Z}_i, \mathbf{Z}_j \rangle| > \phi,\ j < i\right). \tag{34}$$

That is, $f_i = 1$ if $\mathbf{Z}_i$ has large inner product with any $\mathbf{Z}_j$ for $j < i$. The conditional expectation, by a union bound and Lemma 10, is

$$\mathbb{E}\left[f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) | \mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}\right] \leq 2(i-1)(1-\phi^2)^{(d-1)/2}. \tag{35}$$

Let $a_i = (i-1)(1-\phi^2)^{(d-1)/2}$. Then

$$\sum_{i=1}^{N} a_i = N(N-1)(1-\phi^2)^{(d-1)/2}. \tag{36}$$

Then Lemma 9 shows

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \leq \exp\left(-Nt(\log 2) + N(N-1)(1-\alpha^2)^{(d-1)/2}\right). \tag{37}$$

This inequality implies that as long as

$$-Nt(\log 2) + N(N-1)(1-\phi^2)^{(d-1)/2} < 1, \tag{38}$$

then there is a finite probability that $\{\mathbf{Z}_i\}$ contains a subset $\{\mathbf{Z}'_i\}$ of size $\lfloor (1-t)N \rfloor$ such that $\left|\langle \mathbf{Z}'_i, \mathbf{Z}'_j \rangle\right| < \phi$ for all $(i, j)$. Therefore such a set exists. □

A simple setting of the parameters gives the following packing.

**Lemma 12** (Simple packing set)**.** *For* $\phi \in [(2\pi d)^{-1/2}, 1)$*, there exists a set of*

$$K = \frac{1}{8}\exp\left((d-1)\log\frac{1}{\sqrt{1-\phi^2}}\right) \tag{39}$$

*vectors* $\mathcal{C}$ *in* $\mathbb{S}^{d-1}$ *such that for any pair* $\mu, \nu \in \mathcal{C}$*, the inner product between them satisfies*

$$|\langle \mu, \nu \rangle| \leq \phi. \tag{40}$$

*Proof.* Applying Lemma 11 yields a set of $K$ vectors $\mathcal{C}$ satisfying (32) and (33). To get a simple bound that's easy to work with, we can set

$$-Nt(\log 2) + N(N-1)(1-\phi^2)^{(d-1)/2} - 1 = 0, \tag{41}$$

and find an $N$ close to this. Setting $\psi = (1-\phi^2)^{(d-1)/2}$, the quadratic formula solving for $N$ yields

$$\begin{aligned} N &= \frac{1}{\psi}\left(t\log 2 + \psi + \left((t\log 2+\psi)^2 + 4\psi\right)^{1/2}\right) \\ &> \frac{t}{2\psi}. \end{aligned}$$

Now setting $K = \frac{t(1-t)}{2\psi}$ and $t = 1/2$ gives (39). So there exists a set of $K$ vectors on $\mathbb{S}^{d-1}$ whose pairwise inner products are smaller than $\phi$. □

The maximum set of points that can be selected on a sphere of dimension $d$ such that their pairwise inner products are bounded by $\phi$ is an open question. These sets are sometimes referred to as spherical codes [Conway and Sloane, 1998] because they correspond to a set of signaling points of dimension $d$ that can be perfectly decoded over a channel with bounded noise. The bounds here are from a probabilistic construction and can be tightened for smaller $d$. However, in terms of scaling with $d$ this construction is essentially optimal [Shannon, 1959].

### 5.4 Lower bounds for input perturbation

**Lemma 13.** *Let $\Sigma$ be a positive definite matrix and let $f$ denote the density $\mathcal{N}(a, \Sigma)$ and $g$ denote the density $\mathcal{N}(b, \Sigma)$. Then $\mathbf{KL}(f\|g) = \frac{1}{2}(a-b)^T\Sigma^{-1}(a-b)$.*

*Proof.* This is a simple calculation:

$$\begin{aligned} \mathbf{KL}(f\|g) &= \mathbb{E}_{x\sim f}\left[-\frac{1}{2}(x-a)^T\Sigma^{-1}(x-a) + \frac{1}{2}(x-b)\Sigma^{-1}(x-b)\right] \\ &= \frac{1}{2}\left(a^T\Sigma^{-1}a - a^T\Sigma^{-1}b - b^T\Sigma^{-1}a + b^T\Sigma^{-1}b\right) \\ &= \frac{1}{2}(a-b)^T\Sigma^{-1}(a-b). \end{aligned}$$

□

The next theorem is a lower bound on the expected distance between the vector output by MOD-SULQ and the true top eigenvector. In order to get this lower bound, we construct a class of data sets and use Fano's inequality to derive a bound on the minimax error over the class.

**Theorem 14** (Utility bound for MOD-SULQ)**.** *Let $d$, $n$, and $\alpha > 0$ be given and let $\beta$ be given by* (9) *so that the output of MOD-SULQ is $(\alpha, \delta)$-differentially private for all data sets in $\mathbb{R}^d$ with $n$ elements. Then there exists a data set with $n$ elements such that if $\hat{v}_1$ denotes* *the output of* MOD-SULQ *and* $v_1$ *is the top eigenvector of the empirical covariance matrix of the data set, the expected correlation* $\langle \hat{v}_1, v_1 \rangle$ *is upper bounded:*

$$\mathbb{E}\left[|\langle \hat{v}_1, v_1 \rangle|\right] \leq \min_{\phi \in \Phi} \left( 1 - \frac{(1-\phi)}{4} \left( 1 - \frac{1/\beta^2 + \log 2}{(d-1)\log \frac{1}{\sqrt{1-\phi^2}} - \log(8)} \right)^2 \right) \tag{42}$$

*where*

$$\Phi \in \left[ \max\left\{ \frac{1}{\sqrt{2\pi d}}, \sqrt{1 - \exp\left(-\frac{2\log(8d)}{d-1}\right)}, \sqrt{1 - \exp\left(-\frac{2/\beta^2 + \log(256)}{d-1}\right)} \right\}, 1 \right). \tag{43}$$

*Proof.* For $\phi \in [(2\pi d)^{-1/2}, 1)$, Lemma 12 shows there exists a set of $K$ unit vectors $\mathcal{C}$ such that for $\mu, \nu \in \mathcal{C}$, the inner product between them satisfies $|\langle \mu, \nu \rangle| < \phi$, where $K$ is given by (39). Note that for small $\phi$ this setting of $K$ is loose, but any orthonormal basis provides $d$ unit vectors which are orthogonal, setting $K = d$ and solving for $\phi$ yields

$$\left(1 - \exp\left(-\frac{2\log(8d)}{d-1}\right)\right)^{1/2}.$$

Setting the lower bound on $\phi$ to the maximum of these two yields the set of $\phi$ and $K$ which we will consider in (43).

For any unit vector $\mu$, let

$$A(\mu) = \mu\mu^T + N, \tag{44}$$

where $N$ is a $d \times d$ symmetric random matrix such that $\{N_{ij} : 1 \leq i \leq j \leq d\}$ are i.i.d. $\mathcal{N}(0, \beta^2)$, where $\beta^2$ is the noise variance used in the MOD-SULQ algorithm. Due to symmetry, the matrix $A(\mu)$ can be thought of as a jointly Gaussian random vector on the $d(d+1)/2$ variables $\{A_{ij}(\mu) : 1 \leq i \leq j \leq d\}$. The mean of this vector is

$$\bar{\mu} = \left(\mu_1^2, \mu_2^2, \ldots, \mu_d^2, \mu_1\mu_2, \mu_1\mu_3, \ldots, \mu_{d-1}\mu_d\right)^T, \tag{45}$$

and the covariance is $\beta^2 I_{d(d+1)/2}$. Let $f_\mu$ denote the density of this vector.

For $\mu, \nu \in \mathcal{C}$, the divergence between $f_\mu$ and $f_\nu$ can be calculated using Lemma 13:

$$\begin{aligned} \mathbf{KL}\left(f_\mu \| f_\nu\right) &= \frac{1}{2}(\bar{\mu} - \bar{\nu})^T \Sigma^{-1} (\bar{\mu} - \bar{\nu}) \\ &= \frac{1}{2\beta^2} \left\|\bar{\mu} - \bar{\nu}\right\|^2 \\ &\leq \frac{1}{\beta^2}. \end{aligned} \tag{46}$$

The last line follows from the fact that the vectors in $\mathcal{C}$ are unit norm.

For any two vectors $\mu, \nu \in \mathcal{C}$, lower bound the Euclidean distance between them using the upper bound on the inner product:

$$\|\mu - \nu\| \geq \sqrt{2(1-\phi)}. \tag{47}$$

Let $\Theta = \mathbb{S}^{d-1}$ with the Euclidean norm and $\mathcal{R}$ be the set of distributions $\{A(\mu) : \mu \in \Theta\}$. From (47) and (46), the set $\mathcal{C}$ satisfies the conditions of Theorem 8 with $\mathcal{F} = \{f_\mu : \mu \in \mathcal{C}\}$, $r = K$, $\tau = \sqrt{2(1-\phi)}$, and $\gamma = \frac{1}{\beta^2}$. The conclusion of the Theorem shows that for MOD-SULQ,

$$\max_{\mu \in \mathcal{C}} \mathbb{E}_{f_\mu}\left[\|\hat{v} - \mu\|\right] \geq \frac{\sqrt{2(1-\phi)}}{2}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right). \tag{48}$$

This lower bound is vacuous when the term inside the parenthesis is negative, which imposes further conditions on $\phi$. Setting $\log K = 1/\beta^2 + \log 2$, we can solve to find another lower bound on $\phi$:

$$\phi \geq \sqrt{1 - \exp\left(-\frac{2/\beta^2 + \log(256)}{d-1}\right)}. \tag{49}$$

This yields the third term in (43). Note that for larger $n$ this term will dominate the others.

Using Jensen's inequality on the the left side of (48):

$$\max_{\mu \in \mathcal{C}} \mathbb{E}_{f_\mu}\left[2(1 - |\langle \hat{v}, \mu\rangle|)\right] \geq \frac{(1-\phi)}{2}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right)^2.$$

So there exists a $\mu \in \mathcal{C}$ such that

$$\mathbb{E}_{f_\mu}\left[|\langle \hat{v}, \mu\rangle|\right] \leq 1 - \frac{(1-\phi)}{4}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right)^2. \tag{50}$$

Consider the data set consisting of $n$ copies of $\mu$. The corresponding covariance matrix is $\mu\mu^T$ with top eigenvector $v_1 = \mu$. The output of the algorithm MOD-SULQ applied to this data set is an estimator of $\mu$ and hence satisfies (50). Minimizing over $\phi$ gives the desired bound. □

The minimization over $\phi$ in (42) does not lead to analytically pretty results, but numerical optimization can give some insight into these bounds. In all experiments we set $\delta = 0.01$. Figure 1 shows the correlation as a function of $n$ for different dimensions and different values of $\alpha$. In high dimension, the lower bound is shows that the expected performance of MOD-SULQ is poor when there are a small number of data points. This limitation may be particularly acute when the data lies in a very low dimensional subspace but is presented in very high dimension. In such "sparse" settings, perturbing the input as in MOD-SULQ is not a good approach. However, in lower dimensions and data-rich regimes, the performance may be more favorable.

A little calculation yields the sample complexity bound.

*Proof of Theorem 5.* Suppose $\mathbb{E}\left[|\langle \hat{v}_1, v_1\rangle|\right] = \rho$. Then a little algebra shows

$$2\sqrt{1-\rho} \geq \min_{\phi \in \Phi} \sqrt{1-\phi}\left(1 - \frac{1/\beta^2 + \log 2}{(d-1)\log\frac{1}{\sqrt{1-\phi^2}} - \log(8)}\right).$$

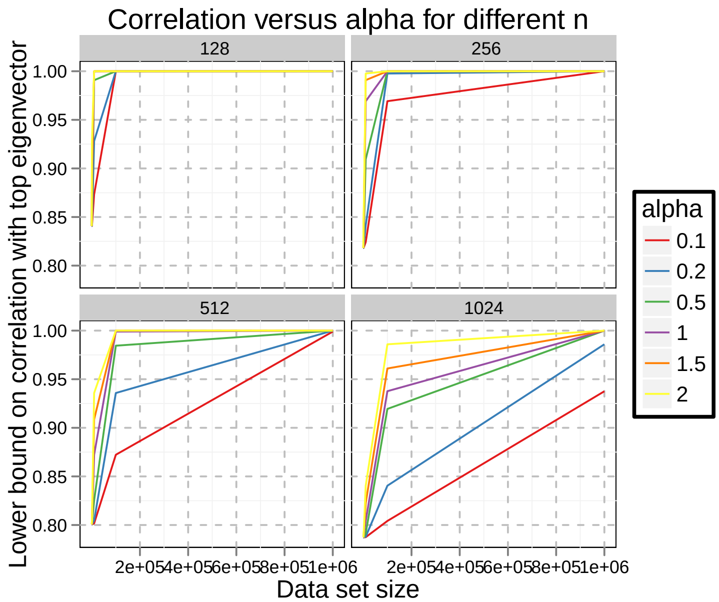



Figure 1: Upper bound on the correlation between $|\langle \hat{v}_1, v_1 \rangle|$ for MOD-SULQ. The horizontal axis is the size of the data set $n$, and the vertical axis is the correlation. The four panels correspond to values of $d = 64$, 128, 256, and 1024.

Set $\phi$ such that $(d-1)\log \frac{1}{\sqrt{1-\phi^2}} - \log(8) = 2(1/\beta^2 + \log 2)$ to get

$$4\sqrt{1-\rho} \geq \sqrt{1-\phi}.$$

Since we are concerned with the scaling behavior for large $d$ and $n$,

$$\log \frac{1}{\sqrt{1-\phi^2}} = \Theta\left(\frac{1}{\beta^2 d}\right),$$

so

$$\begin{aligned}\phi &= \sqrt{1 - \exp\left(-\Theta\left(\frac{1}{\beta^2 d}\right)\right)} \\ &= \Theta\left(\sqrt{\frac{1}{\beta^2 d}}\right).\end{aligned}$$

Lower bound $\beta$ in (9) to get for some constant $c_1$,

$$\beta^2 > c_1 \frac{d^2}{n^2\alpha^2} \log(d/\delta).$$

Substituting this we get for some constant $c_2$ that

$$(1 - c_2(1-\rho)) \leq c_3 \frac{n^2\alpha^2}{d^3 \log(d/\delta)}.$$

Now solving for $n$ shows

$$n \geq c\frac{d^{3/2}\sqrt{\log(d/\delta)}}{\alpha}\left(1 - c'(1-\rho)\right).$$

□

# 6 Implementation and Experiments

We implemented MOD-SULQ and PPCA in order to test our theoretical results. Implementing PPCA involved using a Gibbs sampling procedure [Hoff, 2009]. A crucial parameter in MCMC procedures is the burn-in time, which is how long the chain must be run for it to reach its stationary distribution. Theoretically, chains reach their stationary distribution only in the limit; however, in practice MCMC users must sample after some finite time. In order to use this procedure appropriately, we determined a burn-in time using our data sets. The interaction of MCMC procedures and differential privacy is a rich area for future research.

## 6.1 Data and preprocessing

We report on the performance of our algorithm on some real datasets. We chose four datasets from four different domains – `kddcup99` [Hettich and Bay, 1999], which includes features of 494,021 network connections, `census` [Asuncion and Newman, 2007], a demographic data set on 199,523 individuals, `localization` [Kaluža et al., 2010], a medical dataset with 164,860 instances of sensor readings on individuals engaged in different activities, and `insurance` [van der Putten and van Someren, 2000], a dataset on product usage and demographics of 9,822 individuals.

These datasets contain a mix of continuous and categorical features. We preprocessed each dataset by converting a feature with $q$ discrete values to a vector in $\{0,1\}^q$; after preprocessing, the datasets `kddcup99`, `census`, `localization` and `insurance` have dimensions 116, 513, 44 and 150 respectively. We also normalized each row so that each entry has maximum value 1, and normalize each column such that the maximum (Euclidean) column norm is 1. We choose $k = 4$ for `kddcup`, $k = 8$ for `census`, $k = 10$ for `localization` and $k = 11$ for `insurance`; in each case, the utility $q_{\mathrm{F}}(U_k)$ of the top-$k$ PCA subspace of the data matrix accounts for at least 80% of $\|A\|_{\mathrm{F}}$. Thus, all four datasets, although fairly high dimensional, have good low-dimensional representations. The properties of each dataset are summarized in Table 6.1.

## 6.2 Implementation of Gibbs sampling

The theoretical analysis of PPCA uses properties of the Bingham distribution $\mathsf{BMF}_k(\cdot)$ given in (10). To implement this algorithm for experiments we use a Gibbs sampler due to Hoff [2009]. The Gibbs sampling scheme induces a Markov Chain, the stationary distribution of which is the density in (10). Gibbs sampling and other MCMC procedures are widely used in statistics, scientific modeling, and machine learning to estimate properties of complex distributions Brooks [1998].

| Dataset | #instances | #dimensions | $k$ | $q_{\mathrm{F}}(U)/\left\|A\right\|_{\mathrm{F}}$ |
|---|---|---|---|---|
| `kddcup` | 494,021 | 116 | 4 | 0.96 |
| `census` | 199,523 | 513 | 8 | 0.81 |
| `localization` | 164,860 | 44 | 10 | 0.81 |
| `insurance` | 9,822 | 150 | 11 | 0.81 |

Table 1: Parameters of each dataset. The second column is the number of dimensions after preprocessing. $k$ is the dimensionality of the PCA, and the fourth column contains $q_{\mathrm{F}}(U)/\left\|A\right\|_{\mathrm{F}}$ where $U$ is the top $k$ PCA subspace.

Finding the speed of convergence of MCMC methods is still an open area of research. There has been much theoretical work on estimating convegence times [Jones and Hobart, 2004, Douc et al., 2004, Jones and Hobart, 2001, Roberts, 1999, Roberts and Sahu, 2001, Roberts, 1999, Roberts and Sahu, 2001, Rosenthal, 1995, Kolasa, 1999, 2000], but unfortunately, most theoretical guarantees are available only in special cases and are often too weak for practical use. In lieu of theoretical guarantees, users of MCMC methods empirically estimate the *burn-in time*, or the number of iterations after which the chain is sufficiently close to its stationary distribution. Statisticians employ a range of diagnostic methods and statistical tests to empirically determine if the Markov chain is close to stationarity [Cowles and Carlin, 1996, Brooks and Roberts, 1998, Brooks and Gelman, 1998, El Adlouni et al., 2006]. These tests do not provide a sufficient guarantee of stationarity, and there is no "best test" to use. In practice, the convergence of derived statistics is used to estimate an appropriate the burn-in time. In the case of the Bingham distribution, Hoff [2009] performs qualitative measures of convergence. Developing a better characterization of the convergence of this Gibbs sampler is also an important question for future work.

To choose an appropriate burn-in time, we examined the *time series trace* of the Markov Chain. We ran $l$ copies of the chain, starting from $l$ different initial locations drawn uniformly from the set of all $d \times k$ matrices with orthonormal columns. Let $X^i(t)$ be the output of the $i$-th copy at iteration $t$, and let $U$ be the top $k$ PCA subspace of the data. For each $i$, we plot the magnitude of the projection of $X^i(t)$ onto $U$. After a number of iterations, the projections should converge to the same value.

For each copy, we also plot the following statistic as a function of iteration $T$:

$$F_k^i(T) = \frac{1}{\sqrt{k}} \left\| \frac{1}{T} \sum_{t=1}^{T} X^i(t) \right\|_F ,$$

where $|| \cdot ||_F$ is the Frobenius norm. The matrix Bingham distribution has mean 0, and hence with increasing $T$, the statistic $F_k^i(T)$ should converge to 0.

In our simulations, we observed that the Gibbs sampler converges to $F_k(t) < 0.01$ at $t = 20{,}000$ when run on data with a few hundred dimensions and with $k$ between 5 and 10; we thus chose to run the Gibbs sampler for $T = 20{,}000$ timesteps for all the datasets. We show $\log F_k^i(T)$ as a function of iteration $T$ for datasets `insurance` and `kddcup` in Figures 3 and 2 respectively; the plots are over 5 trajectories of the Markov Chain, which are initialized at 5 locations drawn uniformly from the set of all $d \times k$ matrices with orthonormal columns. The plots show that $F_k^i(T)$ decreases rapidly after a few thousand iterations, and is less

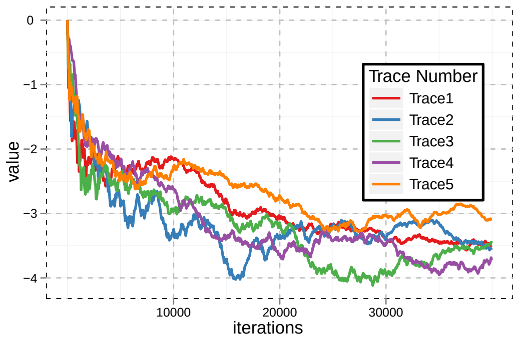



Figure 2: Plot of $\log F_k(T)$ for $k = 4$ as a function of iteration $T$ for $40,000$ iterations of the Gibbs sampler for the `kddcup` dataset.

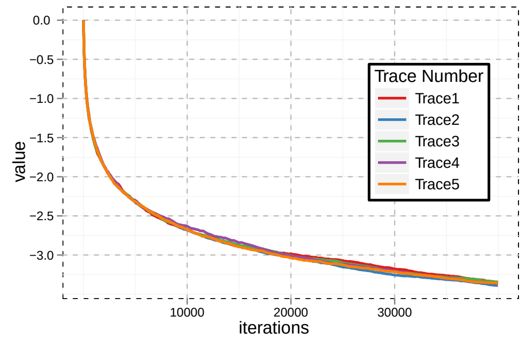



Figure 3: Plot of $\log F_k(T)$ for $k = 11$ as a function of iteration $T$ for $40,000$ iterations of the Gibbs sampler for the `insurance` dataset.

than 0.01 after $T = 20{,}000$ in both cases. $\log F_k^i(T)$ also appears to have a larger variance for `kddcup` than for `insurance`; this is explained by the fact that the `kddcup` dataset has a much larger number of samples, which makes its stationary distribution farther from the initial distribution of the sampler.

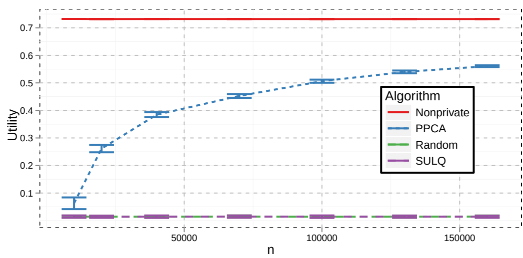
Our simulations indicate that the chains converge fairly rapidly, particularly when $\|A - A_k\|_{\mathrm{F}}$ is small so that $A_k$ is a good approximation to $A$. Convergence is slower for larger $n$ when the initial state is chosen from the uniform distribution over all $k \times d$ matrices with orthonormal columns; this is explained by the fact that for larger $n$, the stationary distribution is farther in variation distance from the starting distribution, which results in a longer convergence time.

### 6.3 Scaling with data set size

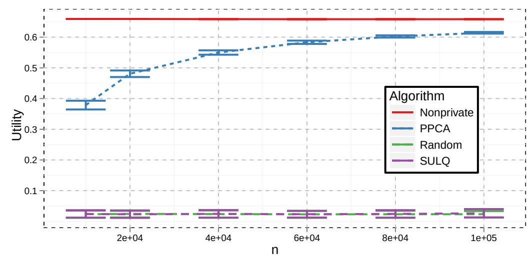
We ran three algorithms on these data sets : standard (non-private) PCA, MOD-SULQ, and PPCA. As a sanity check, we also tried a uniformly generated random projection – since this projection is data-independent we would expect it to have low utility. We measured the utility $q_{\mathrm{F}}(U)$, where $U$ is the $k$-dimensional subspace output by the algorithm; $q_{\mathrm{F}}U$ is maximized when $U$ is the top-$k$ PCA subspace, and thus this reflects how close the output subspace is to the true PCA subspace in terms of representing the data. Although our theoretical results hold for $q_{\mathrm{A}}(\cdot)$, the "energy" $q_{\mathrm{F}}(\cdot)$ is more relevant in practice for larger $k$.

The results for utility $q_{\mathrm{F}}(U)$ are shown in Figures 4(a), 4(b), 4(c), and 4(d); each plot shows $q_{\mathrm{F}}(U)$ as a function of sample size for the $k$-dimensional subspace output by PPCA, MOD-SULQ, non-private PCA, and random projections. PPCA was run with privacy parameter $\alpha = 0.1$; MOD-SULQ with $\alpha = 0.1$ and $\delta = 0.01$. Each value in the figure is an average over 5 random permutations of the data, as well as 10 random starting points of the Gibbs sampler per permutation (for PPCA), and 100 random runs per permutation (for MOD-SULQ and random projections).

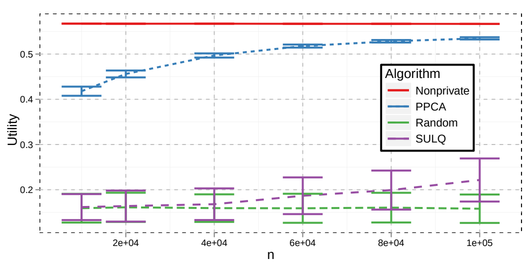
The plots show that PPCA always outperforms MOD-SULQ, and approaches the performance of non-private PCA with increasing sample size. By contrast, for most of the problems and sample sizes considered by our experiments, MOD-SULQ does not perform much better than random projections. The only exception is `localization`, which has much lower dimension (44). This confirms that MOD-SULQ does not scale very well with the data dimension $d$. The performance of both MOD-SULQ and PPCA improve as the sample size increases; the improvement is faster for PPCA than for MOD-SULQ. However, to be fair, MOD-SULQ is simpler and hence runs faster than PPCA. At the sample sizes in our experiments, the performance of non-private PCA does not improve much with a further increase in samples. Our theoretical results suggest that the performance of differentially private PCA cannot be significantly improved over these experiments.



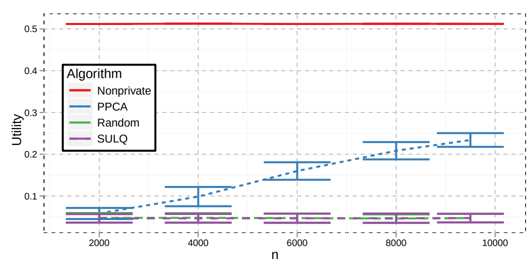

(a) `census` ($k = 8$)




(b) `kddcup` ($k = 4$)




(c) `localization` ($k = 10$)




(d) `insurance` ($k = 11$)

Figure 4: Utility $q_{\mathrm{F}}(U)$ for different data sets.

## 6.4 Effect of privacy on classification

A common use of a dimension reduction algorithm is as a precursor to classification or clustering; to evaluate the effectiveness of the different algorithms, we projected the data onto the subspace output by the algorithms, and measured the classification accuracy using the projected data. The classification results are summarized in Table 6.4. We chose the *normal* vs. all classification task in `kddcup99`, and the *falling* vs. all classification task in `localization`. [1] We used a linear SVM for all classification tasks, which is implemented by libSVM [Chang and Lin, 2011].

For the classification experiments, we used half of the data as a holdout set for computing a projection subspace. We projected the classification data onto the subspace computed

[1]For the other two datasets, `census` and `insurance`, the classification accuracy of linear SVM after (non-private) PCAs is as low as always predicting the majority label.

| | KDDCUP | LOCALIZATION |
|---|---|---|
| Non-private PCA | $98.97 \pm 0.05$ | $100 \pm 0$ |
| PPCA | $98.95 \pm 0.05$ | $100 \pm 0$ |
| SULQ | $98.18 \pm 0.65$ | $97.06 \pm 2.17$ |
| Random Projections | $98.23 \pm 0.49$ | $96.28 \pm 2.34$ |

Table 2: Classification accuracy in the $k$-dimensional subspaces for `kddcup99`($k = 4$), and `localization`($k = 10$) in the $k$-dimensional subspaces reported by the different algorithms.

based on the holdout set; 10% of this data was used for training and parameter-tuning, and the rest for testing. We repeated the classification process 5 times for 5 different (random) projections for each algorithm, and then ran the entire procedure over 5 random permutations of the data. Each value in the figure is thus an average over $5 \times 5 = 25$ rounds of classification.

The classification results show that our algorithm performs almost as well as non-private PCA for classification in the top $k$ PCA subspace, while the performance of MOD-SULQ and random projections are a little worse. The classification accuracy while using MOD-SULQ and random projections also appears to have higher variance compared to our algorithm and non-private PCA; this can be explained by the fact that these projections tend to be farther from the PCA subspace, in which the data has higher classification accuracy.

### 6.5 Effect of the privacy requirement

To check the effect of the privacy requirement, we generated a synthetic data set of $n = 5{,}000$ points drawn from a Gaussian distribution in $d = 10$ with mean $\mathbf{0}$ and whose covariance matrix had eigenvalues

$$\{0.5, 0.30, 0.04, 0.03, 0.02, 0.01, 0.004, 0.003, 0.001, 0.001\}.$$

In this case the space spanned by the top two eigenvectors has most of the energy, so we chose $k = 2$ and plotted the utility $q_{\mathrm{F}}(\cdot)$ for non-private PCA, MOD-SULQ with $\delta = 0.05$, and PPCA with a burn-in time of $T = 1000$. We drew 100 samples from each privacy-preserving algorithm and the plot of the average utility versus $\alpha$ is shown in Figure 5. The privacy requirement relaxes as $\alpha$ increases, and both MOD-SULQ and PPCA approach the utility of PCA without privacy constraints. However, for moderate $\alpha$ PPCA still captures most of the utility, whereas the gap between MOD-SULQ and PPCA becomes quite large.

## 7 Discussion

In this paper we investigated the theoretical and empirical performance of differentially private approximations to PCA. Empirically, we showed that MOD-SULQ and PPCA differ markedly in how well they approximate the top-$k$ subspace of the data. The reason for this, theoretically, is that the sample complexity of MOD-SULQ is $\Omega(d^{3/2}\sqrt{\log d})$ whereas the sample complexity of PPCA is $O(d)$. Because PPCA uses the exponential mechanism with $q_{\mathrm{F}}(\cdot)$ as the utility function, it is not surprising that it performs well. However, MOD-SULQ often had a performance comparable to random projections, indicating that the real

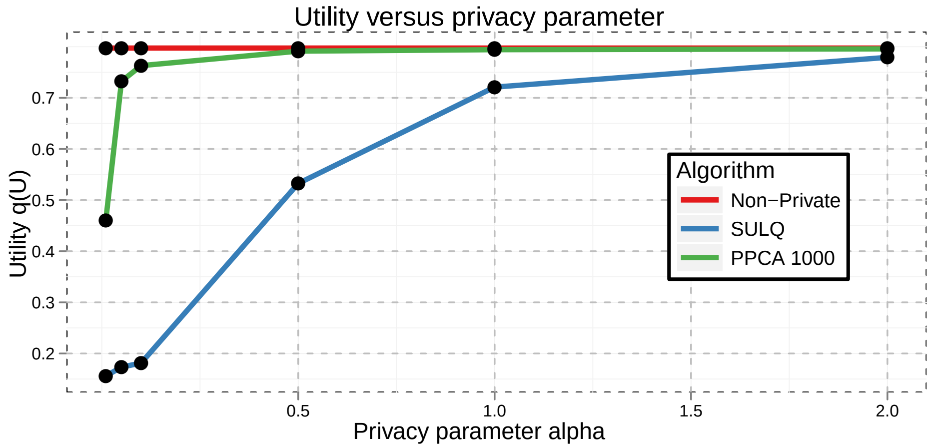



Figure 5: Plot of $q_{\mathrm{F}}(U)$ versus $\alpha$ for a synthetic data set with $n = 5{,}000$, $d = 10$, and $k = 2$.

data sets we used were too small for it to be effective. We furthermore showed that PPCA is nearly optimal, in that any differentially private approximation to PCA must use $\Omega(d)$ samples. An open question is to find algorithms that guarantee weaker $(\alpha, \delta)$-differential privacy in exchange for better dependence on the dimension $d$, analogous to Hall et al. [2012].

Our investigation brought up many interesting issues to consider for future work. Because differential privacy is a mathematical definition, the description of differentially private procedure makes a number of idealizations regarding computation. Some of these idealizations are related to running the algorithm in a real environment; the designers of differentially private systems such as Airavat Roy et al. [2010] require additional security assumptions that have to be verified. At a more basic level, the difference between truly random noise and pseudorandomness Mironov et al. [2009], McGregor et al. [2010] and the effects of finite precision can lead to a gap between the theoretical ideal and practice. Finally, implementation of private algorithms can lead to further gaps between theory and practice. For example, implementing objective perturbation Chaudhuri et al. [2011] uses numerical optimization tools for approximate solutions to convex optimization problems, which have complex termination conditions that are not part of the accompanying theoretical analysis. Our implementation of PPCA does not sample exactly from the Bingham distribution, and we leave a theoretical investigation of the impact of approximate sampling for future work.

Finally, more germane to the work on PCA here is to prove sample complexity results for general $k$ rather than the case $k = 1$ here. For $k = 1$ the utility functions $q_{\mathrm{F}}(\cdot)$ and $q_{\mathrm{A}}(\cdot)$ are related, but for general $k$ it is not immediately clear what metric best captures the idea of "approximating" PCA. In particular, there is a difference between producing an approximation to the top-$k$ PCA subspace (the problem considered here) and an approximation to the top $k$ eigenvectors. Developing a framework for such approximations is of interest more generally in machine learning.

### Acknowledgements

KC and KS would like to thank NIH for research support under U54-HL108460. The experimental results were made possible by support from the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. ADS was supported in part by the California Institute for Telecommunications and Information Technology (CALIT2) at UC San Diego.

## References

Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *SIGMOD Record*, 29(2):439–450, 2000. ISSN 0163-5808. doi: 10.1145/335191.335438.

Arthur Asuncion and David J. Newman. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2007. URL `http://www.ics.uci.edu/~mlearn/{MLR}epository.html`.

Keith M. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, volume 31 of *Mathematical Sciences Research Institute Publications*, pages 1–58. Cambridge University Press, 1997.

Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 273–282, 2007.

Charles Bingham. An antipodally symmetric distribution on the sphere. *Annals of Statistics*, 2(6):1201–1225, 1974.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 128–138, 2005.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In R. E. Ladner and C. Dwork, editors, *Proceedings of the 40th ACM Symposium on the Theory of Computation (STOC)*, pages 609–618, 2008. doi: 10.1145/1374376.1374464.

Stephen P. Brooks. Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):69–100, April 1998. ISSN 00390526. doi: 10.1111/1467-9884.00117.

Stephen P. Brooks and Andrew Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, December 1998. doi: 10.2307/1390675.

Stephen P. Brooks and Gareth O. Roberts. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8(4):319–335, December 1998. doi: 10.1023/A:1008820505350.

Roland W. Butler and Andrew T. A. Wood. Laplace approximations for hypergeometric functions with matrix argument. *Annals of Statistics*, 30(4):1155–1177, 2002.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2011.

Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In Cynthia Dwork, editor, *Advances in Cryptology - CRYPTO 2006*, volume 4117 of *Lecture Notes in Computer Science*, pages 198–213. Springer-Verlag, 2006. doi: 10.1007/11818175_12.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, March 2011.

Yasuko Chikuse. *Statistics on Special Manifolds.* Number 174 in Lecture Notes in Statistics. Springer, New York, 2003.

John H. Conway and Neil J. A. Sloane. *Sphere Packing, Lattices and Groups*. Springer-Verlag, New York, 1998.

Mary Kathryn Cowles and Bradley P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91 (434):883, June 1996. ISSN 01621459. doi: 10.2307/2291683.

Imre Csiszár and Prakash Narayan. The capacity of the arbitrarily varying channel revisited : Positivity, constraints. *IEEE Transactions on Information Theory*, 34(2):181–193, 1988.

Imre Csiszár and Prakash Narayan. Capacity of the Gaussian arbitrarily varying channel. *IEEE Transactions on Information Theory*, 37(1):18–26, 1991.

R. Douc, E. Moulines, and Jeffrey S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *The Annals of Applied Probability*, 14(4):1643–1665, November 2004. ISSN 1050-5164. doi: 10.1214/105051604000000620.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography (TCC)*, pages 265–284, 2006b. doi: 10.1007/11681878_14.

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

Salaheddine El Adlouni, Anne-Catherine Favre, and Bernard Bobée. Comparison of methodologies to assess the convergence of Markov chain Monte Carlo methods. *Computational Statistics & Data Analysis*, 50(10):2685–2701, June 2006. ISSN 01679473. doi: 10.1016/j.csda.2005.04.018.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 211–222, 2003. doi: 10.1145/773153.773174.

Arik Friedman and Assaf Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD)*, pages 493–502, 2010. doi: 10.1145/1835804.1835868.

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):53 pages, June 2010. doi: 10.1145/1749603.1749605.

Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD)*, pages 265–273, 2008.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. arXiv:1203.2570v1 [stat.ML], March 2012.

Shuguo Han, Wee Keong Ng, and P.S. Yu. Privacy-preserving singular value decomposition. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE)*, pages 1267 –1270, 2009. doi: 10.1109/ICDE.2009.217.

Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *Proceedings of the 44th ACM Symposium on the Theory of Computation (STOC)*, 2012.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on the Theory of Computation (STOC)*, pages 705–714, 2010. doi: 10.1145/1806689.1806786.

Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. In *Proceedings of the Ninth IEEE International Conference on Data Mining (ICDM '09)*, pages 169–178, 2009.

S. Hettich and S.D. Bay. The UCI KDD Archive. University of California, Irvine, Department of Information and Computer Science, 1999. URL `http://kdd.ics.uci.edu`.

Peter D. Hoff. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009. ISSN 1061-8600.

Galin L. Jones and James P. Hobart. Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo. *Statistical Science*, 16(4):312–334, 2001. doi: 10.1214/ss/1015346317.

Galin L. Jones and James P. Hobart. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32(2):784–817, 2004.

Boštjan Kaluža, Violeta Mirchevska, Erik Dovgan, Mitja Luštrek, and Matjaž Gams. An agent-based approach to care in independent living. In B. de Ruyter et al., editor, *International Joint Conference on Ambient Intelligence (AmI-10)*, volume 6439/2010 of *Lecture Notes in Computer Science*, pages 177–186. Springer-Verlag, Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-16917-5_18.

Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. ArXiV Preprint arXiv:0803.3946v1 [cs.CR], March 2008.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *Proceedings of the 49ths Annual IEEE Symposium on Foundations of Computer Science (FOCS '08)*, 2008.

Plamen Koev and Alan Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75(254):883–846, January 2006.

John E. Kolasa. Convergence and Accuracy of Gibbs Sampling for Condirtional Distributions in Generalized Linear Models. *The Annals of Statistics*, 27(1):129–142, 1999.

John E. Kolasa. Explicit Bounds for Geometric Covergence of Markov Chains. *Journal of Applied Probability*, 37(3):642–651, 2000.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):943–956, 2010.

Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):92–106, 2006.

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE)*, page 24, 2006.

Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE)*, pages 277–286, 2008.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '10)*, pages 23–26, October 2010.

Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD Conference*, pages 19–30, 2009.

Frank McSherry and Ilya Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD)*, pages 627–636, 2009. doi: 10.1145/1557019.1557090.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.

Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil Vadhan. Computational differential privacy. In Shai Halevi, editor, *Advances in Cryptology - CRYPTO 2009*, volume 5677 of *Lecture Notes in Computer Science*, pages 126–142. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-03355-1. doi: 10.1007/978-3-642-03356-8_8.

Noman Mohammed, Rui Chen, Benjamin C. M. Fung, and Philip S. Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD)*, pages 493–501, 2011. doi: 10.1145/2020408.2020487.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th ACM Symposium on the Theory of Computation (STOC)*, pages 75–84, 2007.

Gareth O. Roberts. Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and their Applications*, 80(2):211–229, April 1999. ISSN 03044149. doi: 10.1016/S0304-4149(98)00085-4.

Gareth O Roberts and Sujit K Sahu. Approximate Predetermined Convergence Properties of the Gibbs Sampler. *Journal of Computational and Graphical Statistics*, 10(2):216–229, June 2001. ISSN 1061-8600. doi: 10.1198/10618600152627915.

Jeffrey S. Rosenthal. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, June 1995. ISSN 01621459. doi: 10.2307/2291067.

Indrajit Roy, Srinath T. V. Setty, Ann Kilzer, Vitaly Shmatikov, and Emmett Witchel. Airavat: Security and privacy for mapreduce. In *Proceedings of the 7th Usenix Symposium on Networked Systems Design and Implementation (NSDI)*, 2010.

Claude. E. Shannon. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal*, 38:611–656, 1959.

G.W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35 (4):551–566, December 1993.

Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

P. van der Putten and M. van Someren. CoIL Challenge 2000: The Insurance Company Case, 2000. URL `http://www.liacs.nl/~putten/library/cc2000/`. Leiden Institute of Advanced Computer Science Technical Report 2000-09.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Andrew Chi-Chih Yao. Protocols for secure computations (extended abstract). In *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science (FOCS '82)*, pages 160–164, 1982.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang, editors, *Festschrift for Lucien Le Cam*, Research Papers in Probability and Statistics, chapter 29, pages 423–425. Springer-Verlag, 1997.

Justin Z. Zhan and Stan Matwin. Privacy-preserving support vector machine classification. *International Journal of Intelligent Information and Database Systems*, 1(3/4):356–385, 2007.