Journal of Machine Learning Research 14 (2013) 2905-2943 Submitted 6/12; Revised 5/13; Published 9/13

A Near-Optimal Algorithm for Differentially-Private Principal Components*

Kamalika Chaudhuri KCHAUDHURI@UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive MC 0404
La Jolla, CA, 92093-0404, USA

Anand D. Sarwate ASARWATE@ALUM.MIT.EDU
Toyota Technological Institute at Chicago
6045 S. Kenwood Ave
Chicago, IL 60637, USA

Kaushik Sinha KAUSHIK.SINHA@WICHITA.EDU
Department of Electrical Engineering and Computer Science
Wichita State University
1845 Fairmount (Campus Box 83)
Wichita, KS 67260-0083, USA

Editor: Gabor Lugosi

Abstract

The principal components analysis (PCA) algorithm is a standard tool for identifying good low-dimensional approximations to high-dimensional data. Many 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 show that the sample complexity of the proposed method differs from the existing procedure in the scaling with the data dimension, and that our method is nearly optimal in terms of this scaling. We furthermore illustrate our results, showing that on real data there is a large performance gap between the existing method and our method.

Keywords: differential privacy, principal components analysis, dimension reduction

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

*. A preliminary version of this work appeared at the Neural Information Processing Systems conference (Chaudhuri et al., 2012). This full version contains more experimental details, full proofs, and additional discussion.

use the transformed data for learning tasks such as classification or regression, and greatly reduce the space required to 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 second moment matrix A 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: examples include user or customer behaviors, demographic surveys, and test results from experimental subjects or patients. These data sets 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 ε_p 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 due to Dwork et al. (2006b) computes the desired algorithm (in our case, 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° 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 SUb Linear Queries (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 noise level required by SULQ may severely impact the quality of approximation, making it impractical for data sets of moderate size.

The goal of this paper is to characterize the problem of differentially private PCA. We assume that the algorithm is given n data points and a target dimension k and must produce a k-dimensional subspace that approximates that produced by the standard PCA problem. We propose a new algorithm, PPCA, which is an instance of the exponential mechanism of McSherry and Talwar (2007). Unlike SULQ, PPCA explicitly takes into account the quality of approximation—it outputs a k-dimensional subspace which is biased towards subspaces close to the output of PCA. In our case, the method corresponds to sampling from the matrix Bingham distribution. We implement PPCA using a Markov Chain Monte Carlo (MCMC) procedure due to Hoff (2009); simulations show that the subspace produced by PPCA captures more of the variance of A than SULQ. When the MCMC procedure converges, the algorithm provides differential privacy.

In order to understand the performance gap, we prove sample complexity bounds for the case of $k = 1$ 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 $\Omega(d^{3/2}\sqrt{\log d})$ for SULQ and as $O(d)$ for PPCA. Furthermore, we show that any differentially private algorithm requires $\Omega(d)$ samples. Therefore PPCA is nearly optimal in terms of sample complexity as a function of data dimension. These theoretical results suggest that our experiments demonstrate the limit of how well ε_p-differentially private algorithms can perform, and our experiments show that this gap should persist for general k. The result seems pessimistic for many applications, because

the sample complexity depends on the extrinsic dimension d rather than the intrinsic dimension k. However, we believe this is a consequence of the fact that we make minimal assumptions on the data; our results imply that, absent additional limitations on the data set, the sample complexity differentially private PCA must grow linearly with the data dimension.

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; investigating the impact of approximate sampling is an avenue for future work. A second set of issues is theoretical—while the privacy guarantees of PPCA hold for all k, our theoretical analysis of sample complexity applies only to $k = 1$ in which the distance and angles between vectors are related. An interesting direction is to develop theoretical bounds for general k; challenges here are providing the right notion of approximation of PCA, and extending the theory using packings of Grassmann or Stiefel manifolds. Finally, in this work we assume k is given to the algorithm, but in many applications k is chosen after looking at the data. Under differential privacy, the selection of k itself must be done in a differentially private manner.

1.1 Related Work

Differential privacy was first proposed by Dwork et al. (2006b). There has been an extensive literature following this work in the computer science theory, machine learning, and databases communities. A survey of some of the theoretical work can be found in the survey by Dwork and Smith (2009). 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 alternative definitions of privacy. In particular, so-called syntactic definitions of privacy (Sweeney, 2002; Machanavajjhala et al., 2006; Li et al., 2010) may be susceptible to attacks based on side-information about individuals in the database.

There are several general approaches to constructing differentially private approximations to some desired algorithm or computation. Input perturbation (Blum et al., 2005) adds noise to the data prior to performing the desired computation, whereas output perturbation (Dwork et al., 2006b) adds noise to the output of the desired computation. The exponential mechanism (McSherry and Talwar, 2007) can be used to perform differentially private selection based on a score function that measures the quality of different outputs. Objective perturbation (Chaudhuri et al., 2011) adds noise to the objective function for algorithms which are convex optimizations. These approaches and related ideas such as Nissim et al. (2007) and Dwork and Lei (2009) have been used to approximate a variety of statistical, machine learning, and data mining tasks under differential privacy (Barak et al., 2007; Wasserman and Zhou, 2010; Smith, 2011; McSherry and Mironov, 2009; Williams and McSherry, 2010; Chaudhuri et al., 2011; Rubinstein et al., 2012; Nissim et al., 2007; Blum et al., 2008; McSherry and Talwar, 2007; Friedman and Schuster, 2010; Hardt and Roth, 2012).

This paper deals with the problem of differentially private approximations to PCA. Prior to our work, the only proposed method for PCA was the Sub-Linear Queries (SULQ) method of Blum et al. (2005). This approach adds noise to the second moment matrix of the data before calculating the singular value decomposition. By contrast, our algorithm, PPCA, uses the exponential mechanism (McSherry and Talwar, 2007) to choose a k-dimensional subspace biased toward those which capture more of "energy" of the matrix. Subsequent to our work, Kapralov and Talwar (2013)

have proposed a dynamic programming algorithm for differentially private low rank matrix approximation which involves sampling from a distribution induced by the exponential mechanism. The running time of their algorithm is $O(d^6)$, where d is the data dimension, and it is unclear how this may affect its implementation. Hardt and Roth (Hardt and Roth, 2012, 2013) have studied low-rank matrix approximation under additional incoherence assumptions on the data. In particular, Hardt and Roth (2012) consider the problem of differentially-private low-rank matrix reconstruction for 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 $\|A-B\|_{\mathrm{F}}$ 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. Hardt and Roth (2013) show a method for guaranteeing (ε,δ)-differential privacy under an *entry-wise* neighborhood condition using the power method for calculating singular values. They, like Kapralov and Talwar (2013), also prove bounds under spectral norm perturbations, and their algorithm achieves the same error rates but with running time that is nearly linear in the number of non-zeros in the data.

In addition to these works, other researchers have examined the interplay between projections and differential privacy. Zhou et al. (2009) analyze a differentially private data release scheme where a random linear transformation is applied to data to preserve differential privacy, and then measures how much this transformation affects the utility of a PCA of the data. One example of a random linear transformation is random projection, popularized by the Johnson-Lindenstrauss (JL) transform. Blocki et al. (2012) show that the JL transform of the data preserves differential privacy provided the minimum singular value of the data matrix is large. Kenthapadi et al. (2013) study the problem of estimating the distance between data points with differential privacy using a random projection of the data points.

There has been significant work on other notions of privacy based on manipulating entries within the database (Sweeney, 2002; Machanavajjhala et al., 2006; Li et al., 2010), for example by reducing the resolution of certain features to create ambiguities. For more details on these and other alternative notions of privacy see Fung et al. (2010) for a survey with more references. An alternative line of privacy-preserving data-mining work (Zhan and Matwin, 2007) is in the Secure Multiparty Computation setting; one work (Han et al., 2009) studies privacy-preserving singular value decomposition in this model. 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.

2. Preliminaries

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 $A = \frac{1}{n} XX^T$ denote the $d \times d$ second moment matrix of the data. The matrix A is positive semidefinite, and has Frobenius norm $\|A\|_F$ 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 $\|A-\hat{A}\|_{\mathrm{F}}$, where k is much lower than the data dimension d. The Schmidt approximation theorem (Stewart, 1993) shows that the minimizer is given by the singular value decomposition, also known as the PCA algorithm in some areas of computer science.

Definition 1 *Suppose A is a positive semidefinite matrix whose first k eigenvalues are distinct. Let the eigenvalues of A be $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A) \geq 0$ and let Λ 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. The top-k PCA subspace of A is the matrix

$$V_k(A) = [v_1 \; v_2 \; \cdots \; v_k], \tag{2}$$

where v_i is the i-th column of V in (1)*. The k-th eigengap is $\Delta_k = \lambda_k - \lambda_{k+1}$.*

Given the top-k subspace and the eigenvalue matrix Λ, we can form an approximation $A^{(k)} = V_k(A)\Lambda_k V_k(A)^T$ to A, where Λ_k contains the k largest eigenvalues in Λ. 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, and $\Delta = \Delta_1$ is the eigengap. For a $d \times k$ matrix $\hat{V}$ with orthonormal columns, the quality of $\hat{V}$ in approximating $V_k(A)$ can be measured by

$$q_{\mathrm{F}}(\hat{V}) = \operatorname{tr}\left(\hat{V}^T A \hat{V}\right). \tag{3}$$

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 on the utility of our PCA approximation apply to the special case $k = 1$. We prove results about the inner product between the output vector $\hat{v}_1$ and the true top eigenvector v_1:

$$q_{\mathrm{A}}(\hat{v}_1) = |\langle \hat{v}_1, v_1 \rangle|. \tag{4}$$

The utility in (4) is related to (3). 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 2 *A randomized algorithm $\mathcal{A}(\cdot)$ is an (ρ, η)-close approximation to the top eigenvector if for all data sets $\mathcal{D}$ of n points we have*

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

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

We study approximations to $\mathcal{A}$ to PCA 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 $\mathcal{A}$ applied to a data set $\mathcal{D}$.

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

$$\sup_{\mathcal{S}} \sup_{\mathcal{D},\mathcal{D}'} \frac{\mu(\mathcal{S} \mid \mathcal{B} = \mathcal{D})}{\mu(\mathcal{S} \mid \mathcal{B} = \mathcal{D}')} \leq e^{\varepsilon_p},$$

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 4 *An algorithm $\mathcal{A}(\mathcal{B})$ taking values in a set $\mathcal{T}$ provides (ε_p, δ)-differential privacy if*

$$\mathbb{P}(\mathcal{A}(\mathcal{D}) \in \mathcal{S}) \leq e^{\varepsilon_p} \mathbb{P}\left(\mathcal{A}(\mathcal{D}') \in \mathcal{S}\right) + \delta,$$

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

Here ε_p and δ are privacy parameters, where low ε_p and δ ensure more privacy (Dwork et al., 2006b; Wasserman and Zhou, 2010; Dwork et al., 2006a). The second privacy guarantee is weaker; the parameter δ bounds the probability of failure, and is typically chosen to be quite small. In our experiments we chose small but constant δ—Ganta et al. (2008) suggest $\delta < \frac{1}{n^2}$ is more appropriate.

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 ε_p and ρ, how large must the number of individuals n in the data set be such that the algorithm is both ε_p-differentially private and a (ρ, η)-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 the privacy ε_p and utility ρ scale with n and the tradeoff between them for fixed n. We show that the sample complexity depends on the eigengap Δ.

3. Algorithms and Results

In this section we describe differentially private techniques for approximating (2). The first is a modified version of the Sub-Linear Queries (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 are differentially private approximations to the top-k subspace: SULQ guarantees (ε_p, δ)-differential privacy and PPCA guarantees ε_p-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. More specifically, SULQ recommends adding a matrix N of i.i.d. Gaussian noise of variance $\frac{8d^2 \log^2(d/\delta)}{n^2 \varepsilon_p^2}$ and applies the PCA algorithm to $A + N$. This guarantees a weaker privacy definition known as (ε_p, δ)-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, as written, is not a good candidate for approximating PCA.

It is easy to modify SULQ to produce a an eigenvector with real entries that guarantees (ε_p, δ) differential privacy. In Algorithm 1, instead of adding an asymmetric Gaussian matrix, we add 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 β^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 in (5) of Algorithm 1 is given in Theorem 5. An alternative is to add Laplace noise of an appropriate variance to each entry—this would guarantee ε_p-differential privacy.

Algorithm 1: Algorithm MOD-SULQ (input pertubation)

inputs: $d \times n$ data matrix X, privacy parameter ε_p, parameter δ

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

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

2 Set

$$\beta = \frac{d+1}{n\varepsilon_p}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{n\sqrt{\varepsilon_p}}. \tag{5}$$

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

3 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 (Chikuse, 2003), which we denote by $\mathsf{BMF}_k(B)$. The algorithm and its privacy properties apply to general $k < d$ but our theoretical results on the utility focus on the special case $k = 1$. The matrix Bingham distribution takes values on the set of all k-dimensional subspaces of $\mathbb{R}^d$ and has a density equal to

$$f(V) = \frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}d, B\right)} \exp(\mathrm{tr}(V^T BV)), \tag{6}$$

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).

By combining results on the exponential mechanism 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 may be expensive 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

Algorithm 2: Algorithm PPCA (exponential mechanism)

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

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

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

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

(6). 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 sensitivity method 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 ε_p-differentially private algorithm. These results show that the PPCA is nearly optimal in terms of the scaling of the sample complexity with respect to the data dimension d, privacy parameter ε_p, and eigengap Δ. We further show that MOD-SULQ requires more samples as a function of d, despite having a slightly weaker privacy guarantee. Proofs are presented in Sections 4 and 5.

Even though both algorithms can output the top-k PCA subspace for general $k \leq d$, we prove results for the case $k = 1$. Finding the scaling behavior of the sample complexity with k is an interesting open problem that we leave for future work; challenges here are finding the right notion of approximation of the PCA, and extending the theory using packings of Grassman or Stiefel manifolds.

Theorem 5 *For the* β *in* (5) *Algorithm MOD-SULQ is* (ε_p, δ) *differentially private.*

Theorem 6 *Algorithm PPCA is* ε_p*-differentially private.*

The fact that these two algorithms are differentially private follows from some simple calculations. 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 grows linearly with the dimension d, inversely with ε_p, and inversely with the correlation gap $(1-\rho)$ and eigenvalue gap Δ. These sample complexity results hold for $k=1$.

Theorem 7 (Sample complexity of PPCA) *If*

$$n > \frac{d}{\varepsilon_p \Delta(1-\rho)}\left(4\frac{\log(1/\eta)}{d} + 2\log\frac{8\lambda_1}{(1-\rho^2)\Delta}\right),$$

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

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

That is, PPCA is a (ρ,η)-close approximation to PCA.

Our second result shows a lower bound on the number of samples required by *any* ε_p-differentially-private algorithm to guarantee a certain level of accuracy for a large class of data sets, and uses proof techniques in Chaudhuri and Hsu (2011, 2012).

Theorem 8 (Sample complexity lower bound) *Fix $d \geq 3$, ε_p, $\Delta \leq \frac{1}{2}$ and let $1-\phi = \exp\left(-2\cdot\frac{\ln 8+\ln(1+\exp(d))}{d-2}\right)$. For any $\rho \geq 1-\frac{1-\phi}{16}$, no ε_p-differentially private algorithm $\mathcal{A}$ can approximate PCA with expected utility greater than ρ on all databases with n points in dimension d having eigenvalue gap Δ, where*

$$n < \frac{d}{\varepsilon_p \Delta}\max\left\{1, \sqrt{\frac{1-\phi}{80(1-\rho)}}\right\}.$$

Theorem 7 shows that if n scales like $\frac{d}{\varepsilon_p\Delta(1-\rho)}\log\frac{1}{1-\rho^2}$ then PPCA produces an approximation $\hat{v}_1$ that has correlation ρ with v_1, whereas Theorem 8 shows that n must scale like $\frac{d}{\varepsilon_p\Delta\sqrt{(1-\rho)}}$ for any ε_p-differentially private algorithm. In terms of scaling with d, ε_p and Δ, the upper and lower bounds match, and they also match up to square-root 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 9 (Sample complexity lower bound for MOD-SULQ) *There are constants c and c' such that if*

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

then there is a data set of size n in dimension d such that the top PCA direction v and the output $\hat{v}$ of MOD-SULQ satisfy $\mathbb{E}[|\langle \hat{v}_1, v_1\rangle|] \leq \rho$.

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.

4. Analysis of PPCA

In this section we provide theoretical guarantees on the performance of PPCA. The proof of Theorem 6 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 also prove a general lower bound on the sample complexity that holds for any ε_p-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 ε_p, and utility $q_{\mathrm{A}}(\cdot)$.

4.1 Privacy Guarantee

We first give a proof of Theorem 6.

Proof Let X be a data matrix whose i-th column is x_i and $A = \frac{1}{n}XX^T$. The PP-PCA algorithm is the exponential mechanism of McSherry and Talwar (2007) applied to the score function $n \cdot v^T A v$. Consider $X' = [x_1\ x_2\ \cdots\ x_{n-1}\ x_n']$ differing from X in a single column and let $A' = \frac{1}{n}X'X'^T$. We have

$$\begin{aligned}\max_{v\in\mathbb{S}^{d-1}} \left|n\cdot v^T A' v - n\cdot v^T A v\right| &\leq \left|v^T(x_n' x_n'^T - x_n x_n^T)v\right| \\ &\leq \left|\left\|v^T x_n'\right\|^2 - \left\|v^T x_n\right\|^2\right| \\ &\leq 1.\end{aligned}$$

The last step follows because $\|x_i\| \leq 1$ for all i. The result now follows immediately from McSherry and Talwar (2007, Theorem 6). ■

4.2 Upper Bound on Utility

The results on the exponential mechanism 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.

Lemma 10 (Lemmas 2.2 and 2.3 of Ball (1997)) *Let μ be the uniform measure on the unit sphere $\mathbb{S}^{d-1}$. For any $x \in \mathbb{S}^{d-1}$ and $0 \leq c < 1$ the following bounds hold:*

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

We are now ready to provide a proof of Theorem 7.

Proof Fix a privacy level ε_p, target correlation ρ, and probability η. Let X be the data matrix and $B = (\varepsilon_p/2)XX^T$ 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 $\langle u, v_1\rangle = \pm\rho$. On this boundary, 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

$$q_{\mathrm{F}}(u) \leq \frac{n\varepsilon_p}{2}(\rho^2\lambda_1 + (1-\rho^2)\lambda_2) \qquad u \in \overline{\mathcal{U}}_\rho, \tag{7}$$

$$q_{\mathrm{F}}(u) \geq \frac{n\varepsilon_p}{2}(\rho^2\lambda_1 + (1-\rho^2)\lambda_d) \qquad u \in \mathcal{U}_\rho. \tag{8}$$

Let $\mu(\cdot)$ denote the uniform measure on the unit sphere. Then fixing an $0 \leq b < 1$, using (7), (8), and the fact that $\lambda_d \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right) &\leq \frac{\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right)}{\mathbb{P}\left(\mathcal{U}_\sigma\right)} \\
&= \frac{\frac{1}{{}_1F_1\left(\frac{1}{2}k,\frac{1}{2}m,B\right)}\int_{\overline{\mathcal{U}}_\rho}\exp\left(u^TBu\right)d\mu}{\frac{1}{{}_1F_1\left(\frac{1}{2}k,\frac{1}{2}m,B\right)}\int_{\mathcal{U}_\sigma}\exp\left(u^TBu\right)d\mu} \\
&\leq \frac{\exp\left(n(\varepsilon_p/2)\left(\rho^2\lambda_1+(1-\rho^2)\lambda_2\right)\right)\cdot\mu\left(\overline{\mathcal{U}}_\rho\right)}{\exp\left(n(\varepsilon_p/2)\left(\sigma^2\lambda_1+(1-\sigma^2)\lambda_d\right)\right)\cdot\mu\left(\mathcal{U}_\sigma\right)} \\
&\leq \exp\left(-\frac{n\varepsilon_p}{2}\left(\sigma^2\lambda_1-(\rho^2\lambda_1+(1-\rho^2)\lambda_2)\right)\right)\cdot\frac{\mu\left(\overline{\mathcal{U}}_\rho\right)}{\mu\left(\mathcal{U}_\sigma\right)}.
\end{aligned} \tag{9}$$

Applying the lower bound from Lemma 10 to the denominator of (9) and the upper bound $\mu\left(\overline{\mathcal{U}}_\rho\right) \leq 1$ yields

$$\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right) \leq \exp\left(-\frac{n\varepsilon_p}{2}\left(\sigma^2\lambda_1-(\rho^2\lambda_1+(1-\rho^2)\lambda_2)\right)\right)\cdot\exp\left(\frac{d-1}{2}\log\frac{2}{1-\sigma}\right). \tag{10}$$

We must choose a $\sigma^2 > \rho^2$ to make the upper bound smaller than 1. 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

$$\log\frac{2}{1-\sigma} < \log\frac{4}{1-\sigma^2} = \log\frac{8\lambda_1}{(1-\rho^2)(\lambda_1-\lambda_2)}.$$

Setting the right hand side of (10) less than η yields

$$\frac{n\varepsilon_p}{4}(1-\rho^2)(\lambda_1-\lambda_2) > \log\frac{1}{\eta} + \frac{d-1}{2}\log\frac{8\lambda_1}{(1-\rho^2)(\lambda_1-\lambda_2)}.$$

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

$$n > \frac{d}{\varepsilon_p(1-\rho)(\lambda_1-\lambda_2)}\left(4\frac{\log(1/\eta)}{d} + 2\log\frac{8\lambda_1}{(1-\rho^2)(\lambda_1-\lambda_2)}\right),$$

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

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

■

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. For such a collection, Lemma 12 shows that for any differentially private mechanism, the average correlation over the collection cannot be too large. That is, any ε_p-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

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \tag{11}$$

has characteristic polynomial

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

and eigenvalues

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

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

We will also need the following Lemma, which is proved in the Appendix.

Lemma 11 (Simple packing set) *For $\phi \in [(2\pi d)^{-1/2}, 1)$, there exists a set of*

$$K = \frac{1}{8}\exp\left((d-1)\log\frac{1}{\sqrt{1-\phi^2}}\right) \tag{12}$$

vectors $\mathcal{C}$ in $\mathbb{S}^{d-1}$ such that for any pair $\mu, \nu \in \mathcal{C}$, the inner product between them is upper bounded by ϕ:

$$|\langle \mu, \nu \rangle| \leq \phi.$$

The following Lemma gives a lower bound on the expected utility averaged over a set of databases which differ in a "small" number of elements.

Lemma 12 *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ be K databases which differ in the value of at most $\frac{\ln(K-1)}{\varepsilon_p}$ 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 ε_p-differentially private algorithm, then,*

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

Proof Let

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

and $\mathcal{G}_i$ be the "double 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

$$\sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \geq \frac{1}{2}(K-1). \tag{13}$$

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,

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

which is a contradiction, so (13) 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

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

so

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

■

We can now prove Theorem 8.

Proof From Lemma 12, given a set of K databases differing in $\frac{\ln(K-1)}{\varepsilon_p}$ points with top eigenvectors $\{u_i : i = 1, 2, \ldots, K\}$, for at least one database i,

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

for any ε_p-differentially private algorithm. Setting the left side equal to some target ρ,

$$1 - \rho \geq \frac{1}{16}\left(1 - \max \left|\langle u_i, u_j \rangle\right|\right). \tag{14}$$

So our goal is construct these data bases such that the inner product between their eigenvectors is small.

Let $y = e_d$, the d-th coordinate vector, and let $\phi \in ((2\pi d)^{-1/2}, 1)$. Lemma 11 shows that there exists a packing $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ of the sphere $\mathbb{S}^{d-2}$ spanned by the first $d-1$ elementary vectors $\{e_1, e_2, \ldots, e_{d-1}\}$ such that $\max_{i \neq j} |\langle w_i, w_j \rangle| \leq \phi$, where

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

Choose ϕ such that $\ln(K-1) = d$. This means

$$1 - \phi = \exp\left(-2 \cdot \frac{\ln 8 + \ln(1 + \exp(d))}{d-2} \right).$$

The right side is minimized for $d = 3$ but this leads to a weak lower bound $1 - \phi > 3.5 \times 10^{-5}$. By contrast, for $d = 100$, the bound is $1 - \phi > 0.12$. In all cases, $1 - \phi$ is at least a constant value.

We construct a database with n points for each w_i. Let $\beta = \frac{d}{n\varepsilon_p}$. For now, we assume that $\beta \leq \Delta \leq \frac{1}{2}$. The other case, when $\beta \geq \Delta$ will be considered later. Because $\beta \leq \Delta$, we have

$$n > \frac{d}{\varepsilon_p \Delta}.$$

The construction uses a parameter $0 \leq m \leq 1$ that will be set as a function of the eigenvalue gap Δ. We will derive conditions on n based on the requirements on d, ε_p, ρ, and Δ. For $i = 1, 2, \ldots, K$ let the data set $\mathcal{D}_i$ contain

- $n(1-\beta)$ copies of $\sqrt{m}y$
- $n\beta$ copies of $z_i = \frac{1}{\sqrt{2}}y + \frac{1}{\sqrt{2}}w_i$.

Thus data sets $\mathcal{D}_i$ and $\mathcal{D}_j$ differ in the values of $n\beta = \frac{\ln(K-1)}{n\varepsilon_p}$ individuals. The second moment matrix A_i of $\mathcal{D}_i$ is

$$A_i = ((1-\beta)m + \frac{1}{2}\beta)yy^T + \frac{1}{2}\beta(w_i^T y + yw_i^T) + \frac{1}{2}\beta w_i w_i^T.$$

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

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

This is in the form (11), with $a = (1-\beta)m + \frac{1}{2}\beta$, $b = \frac{1}{2}\beta$, and $c = \frac{1}{2}\beta$.

The matrix A_i has two nonzero eigenvalues given by

$$\begin{aligned} \lambda &= \frac{1}{2}(a+c) + \frac{1}{2}\sqrt{(a-c)^2 + 4b^2}, \\ \lambda' &= \frac{1}{2}(a+c) - \frac{1}{2}\sqrt{(a-c)^2 + 4b^2}, \end{aligned} \tag{15}$$

The gap Δ between the top two eigenvalues is:

$$\Delta = \sqrt{(a-c)^2 + 4b^2} = \sqrt{m^2(1-\beta)^2 + \beta^2}.$$

We can thus set m in the construction to ensure an eigengap of Δ:

$$m = \frac{\sqrt{(\Delta^2 - \beta^2)}}{1-\beta}. \tag{16}$$

The top eigenvector of A_i is given by

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

where λ is given by (15). Therefore

$$\begin{aligned}\max_{i\neq j}\left|\langle u_i,u_j\rangle\right| &\leq \frac{b^2}{b^2+(a-\lambda)^2}+\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}\max_{i\neq j}\left|\langle w_i,w_j\rangle\right| \\ &\leq 1-\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}(1-\phi). \end{aligned} \tag{17}$$

To obtain an upper bound on $\max_{i\neq j}\left|\langle u_i,u_j\rangle\right|$ we must lower bound $\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}$.

Since $x/(\nu+x)$ is monotonically increasing in x when $\nu>0$, we will find a lower bound on $(a-\lambda)$. Observe that from (15),

$$\lambda-a=\frac{b^2}{\lambda-c}.$$

So to lower bound $\lambda-a$ we need to upper bound $\lambda-c$. We have

$$\lambda-c=\frac{1}{2}(a-c)+\frac{1}{2}\Delta=\frac{1}{2}\left((1-\beta)m+\Delta\right).$$

Because $b=\beta/2$,

$$(\lambda-a)^2>\left(\frac{\beta^2}{2((1-\beta)m+\Delta)}\right)^2=\frac{\beta^4}{4((1-\beta)m+\Delta)^2}.$$

Now,

$$\begin{aligned}\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2} &> \frac{\beta^4}{\beta^2((1-\beta)m+\Delta)^2+\beta^4} \\ &= \frac{\beta^2}{\beta^2+((1-\beta)m+\Delta)^2} \\ &> \frac{\beta^2}{5\Delta^2}, \end{aligned} \tag{18}$$

where the last step follows by plugging in m from (16) and using the fact that $\beta\leq\Delta$. Putting it all together, we have from (14), (17), and (18), and using the fact that ϕ is such that $\ln(K-1)=d$ and $\beta=\frac{d}{n\varepsilon_p}$,

$$\begin{aligned}1-\rho &\geq \frac{1}{16}\cdot\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}(1-\phi) \\ &> \frac{1-\phi}{80}\frac{\beta^2}{\Delta^2} \\ &= \frac{1-\phi}{80}\cdot\frac{d^2}{n^2\varepsilon_p^2\Delta^2}, \end{aligned}$$

which implies

$$n>\frac{d}{\varepsilon_p\Delta}\sqrt{\frac{1-\phi}{80(1-\rho)}}.$$

Thus for $\beta \leq \Delta \leq 1/2$, any ε_p-differentially private algorithm needs $\Omega\left(\frac{d}{\varepsilon_p \Delta \sqrt{1-\rho}}\right)$ points to get expected inner product ρ on all data sets with eigengap Δ.

We now consider the case where $\beta > \Delta$. We choose a slightly different construction here. The i-th database now consists of $n(1-\beta)$ copies of the 0 vector, and $n\beta$ copies of $\frac{\Delta}{\beta} w_i$. Thus, every pair of databases differ in the values of $n\beta = \frac{\ln(K-1)}{\varepsilon_p}$ people, and the eigenvalue gap between the top two eigenvectors is $\beta \cdot \frac{\Delta}{\beta} = \Delta$.

As the top eigenvector of the i-th database is $u_i = w_i$,

$$\max_{i \neq j} |\langle u_i, u_j \rangle| = \max_{i \neq j} |\langle w_i, w_j \rangle| \leq \phi.$$

Combining this with (14), we obtain

$$1 - \rho \geq \frac{1}{16}(1-\phi),$$

which provides the additional condition in the Theorem. ■

5. Analysis of MOD-SULQ

In this section we provide theoretical guarantees on the performance of the MOD-SULQ algorithm. Theorem 5 shows that MOD-SULQ is (ε_p, δ)-differentially private. Theorem 15 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 ε_p and δ 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,

$$\sum_{1 \leq i \leq j \leq d} (x_i x_j - y_i y_j)^2 \leq 2. \tag{19}$$

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 β^2 in the MOD-SULQ algorithm by proving Theorem 5.

Proof Let B and $\hat{B}$ be two independent symmetric random matrices where $\{B_{ij} : 1 \leq i \leq j \leq d\}$ and $\{\hat{B}_{ij} : 1 \leq i \leq j \leq d\}$ are each sets of i.i.d. Gaussian random variables with mean 0 and variance β^2. Consider two data sets $\mathcal{D} = \{x_i : i = 1, 2, \ldots, n\}$ 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:

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

From (19) 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,

$$\mathbb{P}(G\in\mathcal{S}) \le \exp\left(\frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\gamma + \frac{3}{n^2}\right)\right)\mathbb{P}\left(\hat{G}\in\mathcal{S}\right) + \mathbb{P}(B_{ij} > \gamma \text{ for all } i,j). \tag{20}$$

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 γ 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 ε_p to make (20) imply (ε_p, δ) differential privacy:

$$\begin{aligned}\varepsilon_p &= \frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\gamma + \frac{2}{n^2}\right) \\ &= \frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\beta\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{2}{n^2}\right).\end{aligned}$$

Solving for β using the quadratic formula yields the particularly messy expression in (5):

$$\begin{aligned}\beta &= \frac{d+1}{2n\varepsilon_p}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{2n\varepsilon_p}\left(2(d+1)^2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right) + 4\varepsilon_p\right)^{1/2} \\ &\le \frac{d+1}{n\varepsilon_p}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{\sqrt{\varepsilon_p}n}.\end{aligned}$$

■

5.2 Proof of Theorem 9

In this section we provide theoretical guarantees on the performance of the MOD-SULQ algorithm. Theorem 5 shows that MOD-SULQ is (ε_p, δ)-differentially private. Theorem 15 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 ε_p and δ 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 main tool in our lower bound is a generalization by Yu (1997) of an information-theoretic inequality due to Fano.

Theorem 13 (Fano's inequality (Yu, 1997)) *Let $\mathcal{R}$ be a set and Θ 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 Θ. Let X have distribution $f \in \mathcal{F}$ with corresponding parameter θ and let $\hat{\theta}(X)$ be an estimate of θ. If, for all i and j*

$$d(\theta_i, \theta_j) \geq \tau$$

and

$$\mathbf{KL}(f_i \| f_j) \leq \gamma,$$

then

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

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.

Lemma 14 *Let Σ 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 Theorem 13 to derive a bound on the minimax error over the class.

Theorem 15 (Utility bound for MOD-SULQ) *Let d, n, and $\varepsilon_p > 0$ be given and let β be given by Algorithm 1 so that the output of MOD-SULQ is (ε_p, δ)-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:*

$$\mathbb{E}\left[|\langle \hat{v}_1, v_1 \rangle|\right] \leq \min_{\phi \in \Phi} \left(1 - \frac{(1-\phi)}{4} \left(1 - \frac{1/\beta^2 + \log 2}{(d-1)\log \frac{1}{\sqrt{1-\phi^2}} - \log(8)} \right)^2 \right), \tag{21}$$

where

$$\Phi \in \left[\max \left\{ \frac{1}{\sqrt{2\pi d}}, \sqrt{1 - \exp\left(-\frac{2\log(8d)}{d-1} \right)}, \sqrt{1 - \exp\left(-\frac{2/\beta^2 + \log(256)}{d-1} \right)} \right\}, 1 \right). \tag{22}$$

Proof For $\phi \in [(2\pi d)^{-1/2}, 1)$, Lemma 11 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 (12). Note that for small ϕ this setting of K is loose, but any orthonormal basis provides d unit vectors which are orthogonal, setting $K = d$ and solving for ϕ yields

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

Setting the lower bound on ϕ to the maximum of these two yields the set of ϕ and K which we will consider in (22).

For any unit vector μ, let

$$A(\mu) = \mu\mu^T + N,$$

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 β^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

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

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

For $\mu, \nu \in \mathcal{C}$, the divergence between f_μ and f_ν can be calculated using Lemma 14:

$$\begin{aligned} \mathbf{KL}\left(f_\mu \| f_\nu\right) &= \frac{1}{2}(\bar{\mu} - \bar{\nu})^T \Sigma^{-1} (\bar{\mu} - \bar{\nu}) \\ &= \frac{1}{2\beta^2} \|\bar{\mu} - \bar{\nu}\|^2 \\ &\leq \frac{1}{\beta^2}. \end{aligned} \tag{23}$$

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:

$$\|\mu - \nu\| \geq \sqrt{2(1-\phi)}. \tag{24}$$

Let $\Theta = \mathbb{S}^{d-1}$ with the Euclidean norm and $\mathcal{R}$ be the set of distributions $\{A(\mu) : \mu \in \Theta\}$. From (24) and (23), the set $\mathcal{C}$ satisfies the conditions of Theorem 13 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,

$$\max_{\mu \in \mathcal{C}} \mathbb{E}_{f_\mu}[\|\hat{v} - \mu\|] \geq \frac{\sqrt{2(1-\phi)}}{2}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right). \tag{25}$$

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

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

This yields the third term in (22). Note that for larger n this term will dominate the others.

Using Jensen's inequality on the the left side of (25):

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

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

$$\mathbb{E}_{f_\mu}[|\langle \hat{v}, \mu \rangle|] \leq 1 - \frac{(1-\phi)}{4}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right)^2. \tag{26}$$

Consider the data set consisting of n copies of μ. 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 μ and hence satisfies (26). Minimizing over ϕ gives the desired bound. ■

The minimization over ϕ in (21) does not lead to analytically pretty results, so we plotted the results in Figure 1 in order to get a sense of the bounds. Figure 1 shows the lower bound on the expected correlation $\mathbb{E}[|\langle \hat{v}_1, v_1 \rangle|]$ as a function of the number of data points (given on a logarithmic scale). Each panel shows a different dimension, from $d = 50$ to $d = 1000$, and plots are given for different values of ε_p ranging from 0.01 to 2. In all experiments we set $\delta = 0.01$. In high dimension, the lower bound 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 in Theorem 9

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

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

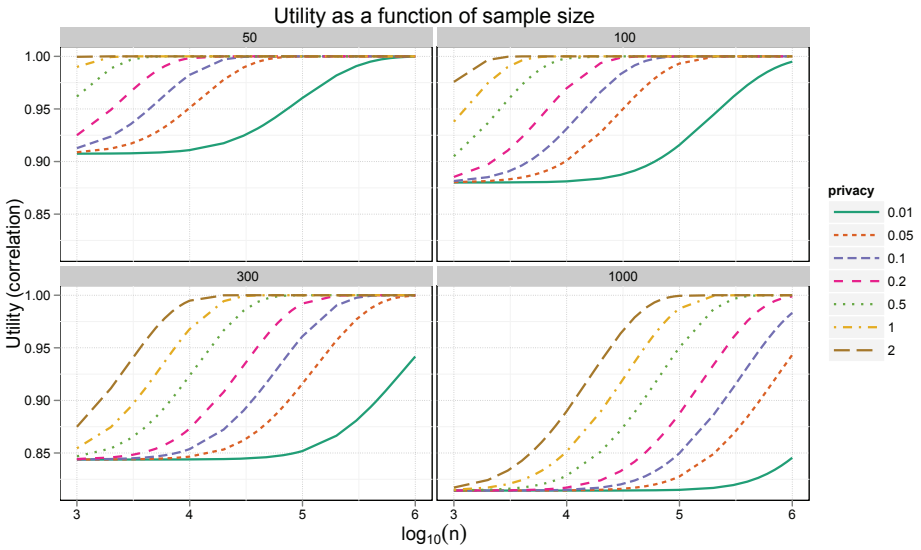


Figure 1: Upper bound from Theorem 15 on the expected correlation between the true top eigenvector and the $\hat{v}_1$ produced by MOD-SULQ. The horizontal axis is $\log_{10}(n)$ and the vertical axis shows the lower bound in (21). The four panels correspond to different values of the dimension d, from 50 to 1000. Each panel contains plots of the bound for different values of ε_p.

Setting ϕ such that $(d-1)\log\frac{1}{\sqrt{1-\phi^2}} - \log(8) = 2(1/\beta^2 + \log 2)$ we have

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

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

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

so

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

From Algorithm 1, to get for some constant c_1, we have the following lower bound on β:

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

Substituting, we get for some constants c_2 and c_3 that

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

Now solving for n shows

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

■

6. Experiments

We next turn to validating our theoretical results on real data. We implemented MOD-SULQ and PPCA in order to test our theoretical bounds. 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 data sets. We chose four data sets from four different domains—`kddcup99` (Bache and Lichman, 2013), which includes features of 494,021 network connections, `census` (Bache and Lichman, 2013), a demographic data set on 199,523 individuals, `localization` (Kaluža et al., 2010), a medical data set with 164,860 instances of sensor readings on individuals engaged in different activities, and `insurance` (van der Putten and van Someren, 2000), a data set on product usage and demographics of 9,822 individuals.

These data sets contain a mix of continuous and categorical features. We preprocessed each data set by converting a feature with q discrete values to a vector in $\{0,1\}^q$; after preprocessing, the data sets `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 $q_{\mathrm{F}}(V_k)$ of the top-k PCA subspace of the data matrix accounts for at least 80% of $\|A\|_{\mathrm{F}}$. Thus, all four data sets, although fairly high dimensional, have good low-dimensional representations. The properties of each data set are summarized in Table 1.

6.2 Implementation of Gibbs Sampling

The theoretical analysis of PPCA uses properties of the Bingham distribution $\mathsf{BMF}_k(\cdot)$ given in (6). 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 (6).

Data Set	#instances	#dimensions	k	$q_{\mathrm{F}}(V_k)$	$q_{\mathrm{F}}(V_k)/\|A\|_{\mathrm{F}}$
kddcup	494,021	116	4	0.6587	0.96
census	199,523	513	8	0.7321	0.81
localization	164,860	44	10	0.5672	0.81
insurance	9,822	150	11	0.5118	0.81

Table 1: Parameters of each data set. The second column is the number of dimensions after preprocessing. k is the dimensionality of the PCA, the third column contains $q_{\mathrm{F}}(V_k)$, where V_k is the top-k PCA subspace, and the fifth column is the normalized utility $q_{\mathrm{F}}(V_k)/\|A\|_{\mathrm{F}}$.

Gibbs sampling and other MCMC procedures are widely used in statistics, scientific modeling, and machine learning to estimate properties of complex distributions (Brooks, 1998).

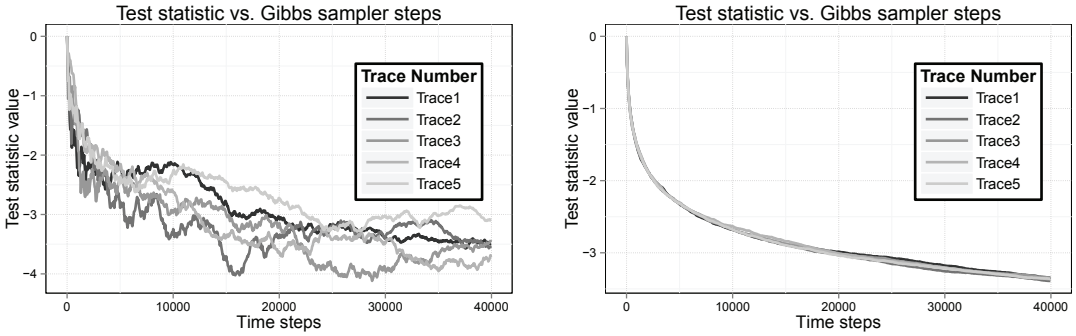
Finding the speed of convergence of MCMC methods is still an open area of research. There has been much theoretical work on estimating convegence times (Jones and Hobart, 2004; Douc et al., 2004; Jones and Hobart, 2001; Roberts, 1999; Roberts and Sahu, 2001; Roberts, 1999; Roberts and Sahu, 2001; Rosenthal, 1995; Kolassa, 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 an important question for future work.

Because the MCMC procedure of Hoff (2009) does not come with convergence-time guarantees, for our experiments we had to choose an appropriate burn-in time. The "ideal" execution of PPCA provides ε_p-differential privacy, but because our implementation only approximates sampling from the Bingham distribution, we cannot guarantee that this implementation provides the privacy guarantee. As noted by Mironov (2012), even current implementations of floating-point arithmetic may suffer from privacy problems, so there is still significant work to do between theory and implementation. For this paper we tried to find a burn-in time that was sufficiently long so that we could be confident that the empirical performance of PPCA was not affected by the initial conditions of the sampler.

In order to choose an appropriate burn-in time, we examined the *time series trace* of the Markov Chain. We ran l copies, or traces, 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. We used 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.



(a) `kddcup` ($k = 4$) (b) `insurance` ($k = 11$)

Figure 2: Plots of $\log F_k^i(T)$ for five different traces (values of i) on two different data sets. Figure 2(a) shows $\log F_k^i(T)$ for for $k = 4$ as a function of iteration T for $40,000$ steps of the Gibbs sampler on the `kddcup` data set. Figure 2(b) shows the same for the `insurance` data set.

Figure 2 illustrates the behavior of the Gibbs sampler. The plots show the value of $\log F_k^i(T)$ as a function of the Markov chain iteration for 5 different restarts of the MCMC procedure for two data sets, `kddcup` and `insurance`. The initial starting points were chosen 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 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` data set has a much larger number of samples, which makes its stationary distribution farther from the initial distribution of the sampler. Based on these and other 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 data sets.

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.

To investigate how well these different algorithms performed on real data, for each data set we subsampled data sets of different sizes n uniformly and ran the algorithms on the subsets. We chose $\varepsilon_p = 0.1$ for this experiment, and for MOD-SULQ we used $\delta = 0.01$. We averaged over 5 such subsets and over several instances of the randomized algorithms (10 restarts for PPCA and 100 for MOD-SULQ and random projections). For each subset and instance we calculated the resulting utility $q_{\mathrm{F}}(\cdot)$ of the output subspace.

Figures 3(a), 3(b), 4(a), and 4(b) show $q_{\mathrm{F}}(U)$ as a function of the subsampled data set sizes. The bars indicate the standard deviation over the restarts (from subsampling the data and random sampling for privacy). The non-private algorithm achieved $q_{\mathrm{F}}(V_k)$ for nearly all subset sizes (see Table 1 for the values). These plots illustrate how additional data can improve the utility of the output for a fixed privacy level ε_p. As n increases, the dashed blue line indicating the utility of PPCA begins to approach $q_{\mathrm{F}}(V_k)$, the utility of the optimal subspace.

These experiments also show that the performance of PPCA is significantly better than that of MOD-SULQ, and MOD-SULQ produces subspaces whose utility is on par with randomly choosing a subspace. The only exception to this latter point is `localization`, We believe this is because d is much lower for this data set ($d = 44$), which shows that for low dimension and large n, MOD-SULQ may produce subspaces with reasonable utility. Furthermore, MOD-SULQ is simpler and hence runs faster than PPCA, which requires running the Gibbs sampler past the burn-in time. Our theoretical results suggest that the performance of differentially private PCA cannot be significantly improved over the performance of PPCA but since those results hold for $k = 1$ they do not immediately apply here.

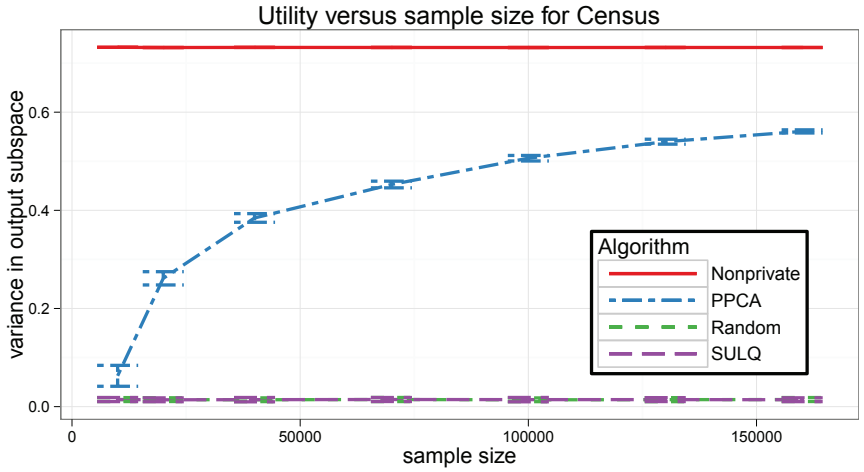
6.4 Effect of Privacy on Classification

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 2. 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).

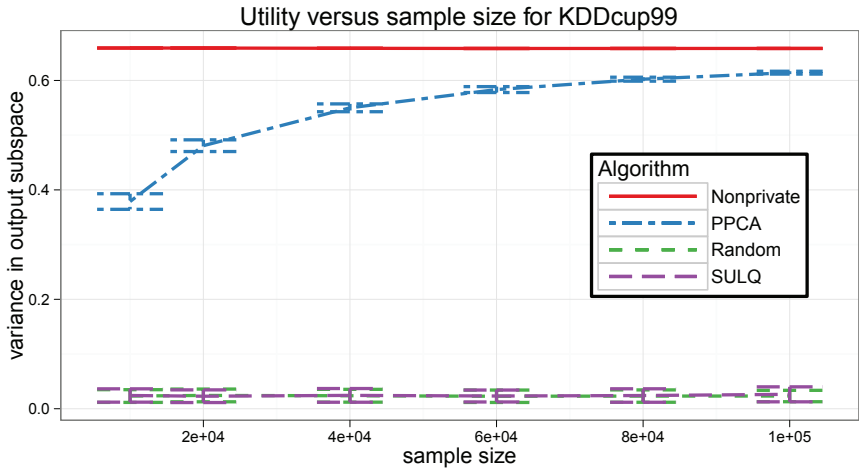
For the classification experiments, we used half of the data as a holdout set for computing a projection subspace. We projected the classification data onto the subspace computed based on the holdout set; 10% of this data was used for training and parameter-tuning, and the rest for testing. We repeated 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

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



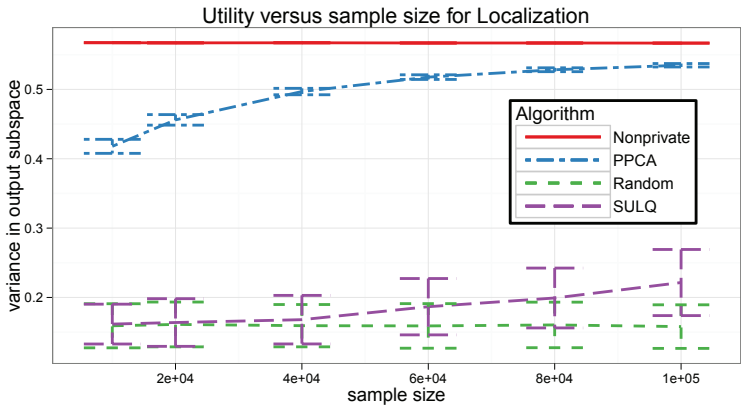
(a) `census` ($k = 8$)



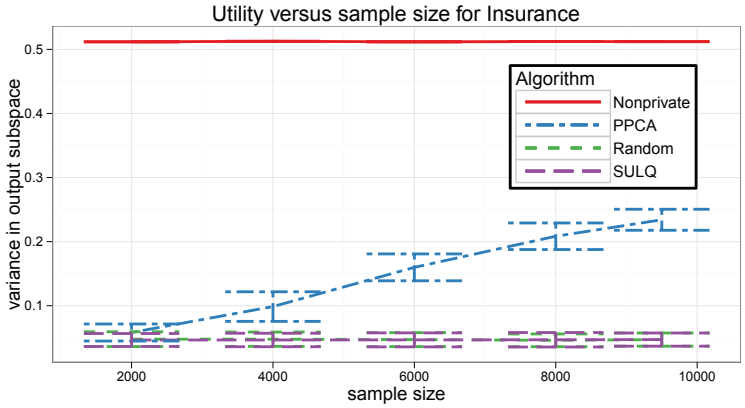
(b) `kddcup` ($k = 4$)

Figure 3: Plot of the unnormalized utility $q_{\mathrm{F}}(U)$ versus the sample size n, averaged over random subsets of the data and randomness in the algorithms. The bars are at one standard deviation about the mean. The top red line is the PCA algorithm without privacy constraints. The dashed line in blue is the utility for PPCA. The green and purple dashed lines are nearly indistinguishable and represent the utility from random projections and MOD-SULQ, respectively. In these plots $\varepsilon_p = 0.1$ and $\delta = 0.01$.

is because the projections tend to be farther from the top-k PCA subspace, making the classification error more variable.



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



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

Figure 4: Plot of the unnormalized utility $q_{\mathrm{F}}(U)$ versus the sample size n, averaged over random subsets of the data and randomness in the algorithms. The bars are at one standard deviation about the mean. The top red line is the PCA algorithm without privacy constraints. The dashed line in blue is the utility for PPCA. The green and purple dashed lines are nearly indistinguishable for `insurance` but diverge for `localization`—they represent the utility from random projections and MOD-SULQ, respectively. In these plots $\varepsilon_p = 0.1$ and $\delta = 0.01$.

6.5 Effect of the Privacy Requirement

How to choose ε_p is important open question for many applications. We wanted to understand the impact of varying ε_p on the utility of the subspace. We did this via a synthetic data set—we generated $n = 5{,}000$ points drawn from a Gaussian distribution in $d = 10$ with mean $\mathbf{0}$ and covariance matrix with eigenvalues

$$\{0.5, 0.30, 0.04, 0.03, 0.02, 0.01, 0.004, 0.003, 0.001, 0.001\}. \tag{27}$$

	kddcup99	**localization**
Non-private PCA	98.97 ± 0.05	100 ± 0
PPCA	98.95 ± 0.05	100 ± 0
MOD-SULQ	98.18 ± 0.65	97.06 ± 2.17
Random Projections	98.23 ± 0.49	96.28 ± 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.

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 ε_p is shown in Figure 5. The privacy requirement relaxes as ε_p increases, and both MOD-SULQ and PPCA approach the utility of PCA without privacy constraints. However, for moderate ε_p PPCA still captures most of the utility, whereas the gap between MOD-SULQ and PPCA becomes quite large.

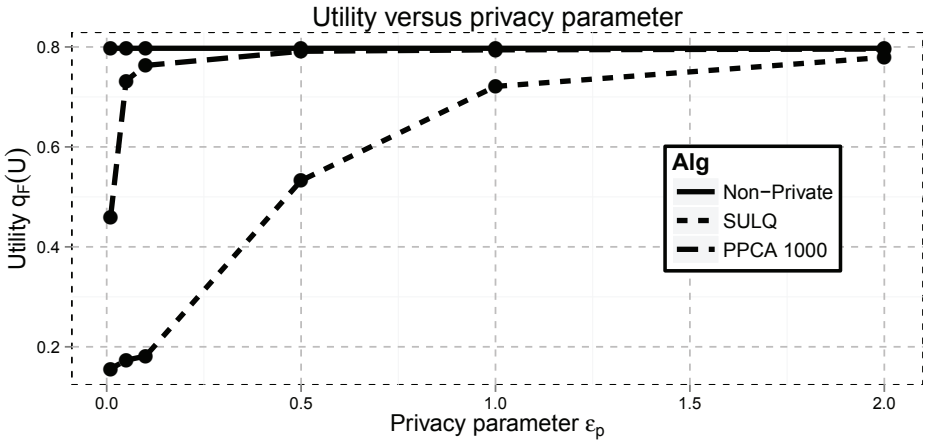


Figure 5: Plot of $q_{\mathrm{F}}(U)$ versus ε_p for a synthetic data set with $n = 5{,}000$, $d = 10$, and $k = 2$. The data has a Gaussian distribution whose covariance matrix has eigenvalues given by (27).

7. Conclusion

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 scales as $d^{3/2}\sqrt{\log d}$ whereas PPCA scales as 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 data sets had too few points 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.

Our investigation brought up many interesting issues to consider for future work. The description of differentially private algorithms assume an ideal model of computation : real systems require additional security assumptions that have to be verified. The difference between truly random noise and pseudorandomness and the effects of finite precision can lead to a gap between the theoretical ideal and practice. Numerical optimization methods used in some privacy methods (Chaudhuri et al., 2011) can only produce approximate solutions; they may also have complex termination conditions unaccounted for in the theoretical analysis. MCMC sampling is similar : if we can guarantee that the sampler's distribution has total variation distance δ from the Bingham distribution, then sampler can guarantee (ε_p, δ) differential privacy. However, we do not yet have such analytical bounds on the convergence rate; we must determine the Gibbs sampler's convergence empirically. Accounting for these effects is an interesting avenue for future work that can bring theory and practice together.

For PCA more specifically, it would be interesting to extend the sample complexity results to general $k > 1$. For $k = 1$ the utility functions $q_{\mathrm{F}}(\cdot)$ and $q_{\mathrm{A}}(\cdot)$ are related, but for larger k it is not immediately clear what metric best captures the idea of "approximating" the top-k PCA subspace. For minimax lower bounds, it may be possible to construct a packing with respect to a general utility metric. For example, Kapralov and Talwar (2013) use properties of packings on the Grassmann manifold. Upper bounds on the sample complexity for PPCA may be possible by performing a more careful analysis of the Bingham distribution or by finding better approximations for its normalizing constant. Developing a framework for analyzing general approximations to PCA may be of interest more broadly in machine learning.

Acknowledgments

The authors would like to thank the reviewers for their detailed comments, which greatly improved the quality and readability of the manuscript, and the action editor, Gabor Lugosi, for his patience during the revision process. 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.

Appendix A. 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, 1991).

Lemma 16 *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 $0 \leq f_i \leq 1$, $i = 1, 2, \ldots, N$. Then the condition*

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

implies that

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

Proof First apply Markov's inequality:

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

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

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

Therefore

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

Continuing in the same way yields

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

■

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 (Shannon, 1959) on the distribution of the angle between a uniformly distributed unit vector and a fixed unit vector.

Lemma 17 (Lemma 2 of Csiszár and Narayan (1991)) *Let* $\mathbf{U}$ *be a random vector distributed uniformly 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:*

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

Lemma 18 (Packing set on the unit sphere) *Let the dimension* d *and parameter* $\phi \in [(2\pi d)^{-1/2}, 1)$ *be given. For* N *and* t *satisfying*

$$-Nt(\log 2) + N(N-1)(1-\phi^2)^{(d-1)/2} < 0 \tag{28}$$

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

$$|\langle \mu, \nu \rangle| < \phi. \tag{29}$$

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, equivalently, small pairwise inner products. To that end, define i.i.d. random variables $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ uniformly distributed on $\mathbb{S}^{d-1}$ and functions

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

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 17, is

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

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

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

Then Lemma 16 shows

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

This inequality implies that as long as

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

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 packing in Lemma 11.

Proof Applying Lemma 18 yields a set of K vectors $\mathcal{C}$ satisfying (28) and (29). To get a simple bound that's easy to work with, we can set

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

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

$$N = 1 + \frac{t \log 2}{\psi} > \frac{t}{2\psi}.$$

Now setting $K = \frac{t(1-t)}{2\psi}$ and $t = 1/2$ gives (12). So there exists a set of K vectors on $\mathbb{S}^{d-1}$ whose pairwise inner products are smaller than ϕ. ■

The maximum set of points that can be selected on a sphere of dimension d such that their pairwise inner products are bounded by ϕ 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).

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. URL http://dx.doi.org/10.1145/335191.335438.

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

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 '07)*, pages 273–282, New York, NY, USA, 2007. ACM. doi: 10.1145/1265530.1265569. URL http://dx.doi.org/10.1145/1265530.1265569.

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson-Lindenstrauss Transform itself preserves differential privacy. In *IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 410–419, October 2012. doi: 10.1109/FOCS.2012.67. URL http://dx.doi.org/10.1109/FOCS.2012.67.

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 '05)*, pages 128–138, New York, NY, USA, 2005. ACM. doi: 10.1145/1065167.1065184. URL http://dx.doi.org/10.1145/1065167.1065184.

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 Annual ACM Symposium on Theory of Computing (STOC '08)*, pages 609–618, New York, NY, USA, 2008. ACM. doi: 10.1145/1374376.1374464. URL http://dx.doi.org/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. URL http://dx.doi.org/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. URL http://dx.doi.org/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. URL http://dx.doi.org/10.1023/A:1008820505350.

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 Sham Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory (COLT '11)*, volume 19 of *JMLR Workshop and Conference Proceedings*, pages 155–186, Budapest, Hungary, June 2011. URL http://www.jmlr.org/proceedings/papers/v19/chaudhuri11a/chaudhuri11a.pdf.

Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1327–1334, New York, NY, USA, July 2012. Omnipress. URL http://icml.cc/2012/papers/663.pdf.

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, Berlin, Heidelberg, August 2006. Springer-Verlag. doi: 10.1007/11818175_12. URL http://dx.doi.org/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. URL http://jmlr.csail.mit.edu/papers/v12/chaudhuri11a.html.

Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 998–1006, 2012. URL http://books.nips.cc/papers/files/nips25/NIPS2012_0482.pdf.

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. URL http://dx.doi.org/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. doi: 10.1109/18.2627. URL http://dx.doi.org/10.1109/18.2627.

Imre Csiszár and Prakash Narayan. Capacity of the Gaussian arbitrarily varying channel. *IEEE Transactions on Information Theory*, 37(1):18–26, 1991. doi: 10.1109/18.61125. URL http://dx.doi.org/10.1109/18.61125.

Randal Douc, Eric 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. URL http://dx.doi.org/10.1214/105051604000000620.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09)*, pages 371–380, New York, NY, USA, 2009. ACM. doi: 10.1145/1536414.1536466. URL http://dx.doi.org/10.1145/1536414.1536466.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2009. URL http://repository.cmu.edu/jpc/vol1/iss2/2.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503, Berlin, Heidelberg, 2006a. Springer-Verlag. doi: 10.1007/11761679_29. URL http://dx.doi.org/10.1007/11761679_29.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284, Berlin, Heidelberg, March 4–7 2006b. Springer. doi: 10.1007/11681878_14. URL http://dx.doi.org/10.1007/11681878_14.

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. URL http://dx.doi.org/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. URL http://dx.doi.org/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 Mining (KDD '10)*, pages 493–502, New York, NY, USA, 2010. ACM. doi: 10.1145/1835804.1835868. URL http://dx.doi.org/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 Computing Surveys*, 42(4):14:1–14:53, June 2010. doi: 10.1145/1749603.1749605. URL http://dx.doi.org/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 Mining (KDD '08)*, pages 265–273, New York, NY, USA, 2008. ACM. doi: 10.1145/1401890.1401926. URL http://dx.doi.org/10.1145/1401890.1401926.

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. URL http://dx.doi.org/10.1109/ICDE.2009.217.

Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC '12)*, pages 1255–1268, New York, NY, USA, 2012. ACM. doi: 10.1145/2213977.2214088. URL http://dx.doi.org/10.1145/2213977.2214088.

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC '13)*, pages 331–340, New York, NY, USA, June 2013. ACM. doi: 10.1145/2488608.2488650. URL http://dx.doi.org/10.1145/2488608.2488650.

Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining (ICDM '09)*, pages 169–178, 2009. doi: 10.1109/ICDM.2009.11. URL http://dx.doi.org/10.1109/ICDM.2009.11.

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. doi: 10.1198/jcgs.2009.07177. URL http://dx.doi.org/10.1198/jcgs.2009.07177.

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. URL http://dx.doi.org/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, April 2004. doi: 10.1214/009053604000000184. URL http://dx.doi.org/10.1214/009053604000000184.

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. URL http://dx.doi.org/10.1007/978-3-642-16917-5_18.

Mikhail Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '13)*, pages 1395–1414, New Orleans, LA, USA, January 2013.

Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. Technical Report arXiv:0803.3946v1 [cs.CR], ArXiV, March 2008. URL http://arxiv.org/abs/0803.3946.

Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the Johnson-Lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71, 2013. URL http://repository.cmu.edu/jpc/vol5/iss1/2.

John E. Kolassa. Convergence and accuracy of Gibbs sampling for conditional distributions in generalized linear models. *The Annals of Statistics*, 27(1):129–142, 1999. doi: 10.1214/aos/1018031104. URL http://dx.doi.org/10.1214/aos/1018031104.

John E. Kolassa. Explicit bounds for geometric covergence of Markov chains. *Journal of Applied Probability*, 37(3):642–651, 2000. doi: 10.1239/jap/1014842825. URL http://dx.doi.org/10.1239/jap/1014842825.

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. doi: 10.1109/TKDE.2009.139. URL http://dx.doi.org/10.1109/TKDE.2009.139.

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. doi: 10.1109/TKDE.2006.14. URL http://dx.doi.org/10.1109/TKDE.2006.14.

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. doi: 10.1109/ICDE.2006.1. URL http://dx.doi.org/10.1109/ICDE.2006.1.

Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering (ICDE)*, pages 277–286, April 2008. doi: 10.1109/ICDE.2008.4497436. URL http://dx.doi.org/10.1109/ICDE.2008.4497436.

Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD Conference*, pages 19–30, 2009. doi: 10.1145/1559845.1559850. URL http://dx.doi.org/10.1145/1559845.1559850.

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. URL http://dx.doi.org/10.1145/1557019.1557090.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*, pages 94–103, October 2007. doi: 10.1109/FOCS.2007.41. URL http://dx.doi.org/10.1109/FOCS.2007.41.

Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS '12)*, pages 650–661, New York, NY, USA, 2012. ACM. doi: 10.1145/2382196.2382264. URL http://dx.doi.org/10.1145/2382196.2382264.

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 Mining (KDD '11)*, pages 493–501, New York, NY, USA,

2011. ACM. doi: 10.1145/2020408.2020487. URL http://dx.doi.org/10.1145/2020408.2020487.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing (STOC '07)*, pages 75–84, New York, NY, USA, 2007. ACM. doi: 10.1145/1250790.1250803. URL http://dx.doi.org/10.1145/1250790.1250803.

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. URL http://dx.doi.org/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. URL http://dx.doi.org/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. URL http://dx.doi.org/10.2307/2291067.

Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012. URL http://repository.cmu.edu/jpc/vol4/iss1/4/.

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

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC '11)*, pages 813–822, New York, NY, USA, 2011. ACM. doi: 10.1145/1993636.1993743. URL http://dx.doi.org/10.1145/1993636.1993743.

Gilbert 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*, 10(5):557–570, October 2002. doi: 10.1142/S0218488502001648. URL http://dx.doi.org/10.1142/S0218488502001648.

Peter van der Putten and Maarten 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. doi: 10.1198/jasa.2009.tm08651. URL http://dx.doi.org/10.1198/jasa.2009.tm08651.

Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2451–245, 2010. URL `http://books.nips.cc/papers/files/nips23/NIPS2010_1276.pdf`.

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. doi: 10.1504/IJIIDS.2007.016686.

Shuheng Zhou, Katrina Ligett, and Larry Wasserman. Differential privacy with compression. In *Proceedings of the 2009 International Symposium on Information Theory (ISIT)*, pages 2718–2722, Seoul, South Korea, June–July 2009. doi: 10.1109/ISIT.2009.5205863.