Differential Privacy with Compression

Shuheng Zhou
Seminar für Statistik
ETH Zürich
Zürich, CH-8092, Switzerland
Email: zhou@stat.math.ethz.ch

Katrina Ligett
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: katrina@cs.cmu.edu

Larry Wasserman
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
Email: larry@stat.cmu.edu

***Abstract*—This work studies formal utility and privacy guarantees for a simple multiplicative database transformation, where the data are compressed by a random linear or affine transformation, reducing the number of data records substantially, while preserving the number of original input variables. We provide an analysis framework inspired by a recent concept known as *differential privacy*. Our goal is to show that, despite the general difficulty of achieving the differential privacy guarantee, it is possible to publish synthetic data that are useful for a number of common statistical learning applications. This includes high dimensional sparse regression [24], principal component analysis (PCA), and other statistical measures [16] based on the covariance of the initial data.**

I. Introduction

In statistical learning, privacy is increasingly a concern whenever large amounts of confidential data are manipulated within or published outside an organization. It is often important to allow researchers to analyze data *utility* without leaking information or compromising the *privacy* of individual records. In this work, we demonstrate that one can preserve utility for a variety of statistical applications while achieving a formal definition of privacy. The algorithm we study is a simple random projection by a matrix of independent Gaussian random variables that compresses the number of records in the database. Our goal is to preserve the privacy of every individual in the database, even if the number of records in the database is very large. In particular, we show how this randomized procedure can achieve a form of "differential privacy" [9, 8], while at the same time showing that the compressed data can be used for Principal Component Analysis (PCA) and other operations that rely on the accuracy of the empirical covariance matrix computed via the compressed data, compared to its population or the uncompressed correspondents. Toward this goal, we study "distributional privacy", which is more natural for many statistical inference tasks.

More specifically, the data are represented as an $n \times p$ matrix X. Each of the p columns is an attribute, and each of the n rows is the vector of attributes for an individual record. The data are compressed by a random linear transformation $X \mapsto \mathcal{X} \equiv \Phi X$, where Φ is a random $m \times n$ matrix with $m \ll n$. It is also natural to consider a random affine transformation $X \mapsto \mathcal{X} \equiv \Phi X + \Delta$, where Δ is a random $m \times p$ matrix, as considered in [24] for privacy analysis, the latter of which is beyond the scope of this paper and intended as future work. Such transformations have been called "matrix masking" in the privacy literature [7]. The entries of Φ are taken to be independent Gaussian random variables, but other distributions are possible. The resulting compressed data can then be made available for statistical analysis; that is, we think of $\mathcal{X}$ as "public," while Φ and Δ are private and only needed at the time of compression. However, even if Φ were revealed, recovering X from $\mathcal{X}$ requires solving a highly underdetermined linear system and comes with information theoretic privacy guarantees, as demonstrated in [24].

Informally, differential privacy [9, 8] limits the increase in the information that can be learned when any single entry is changed in the database. This limit implies [17] that allowing one's data to be included in the database is in some sense incentive-compatible. Differential privacy imposes a compelling and clear requirement, that when running a privacy-preserving algorithm on two neighboring databases that differ in only one entry, the probability of any possible outcome of the algorithm should be nearly (multiplicatively) equal. Many existing results in differential privacy use additive output perturbations by adding a small amount of random noise to the released information according to the sensitivity of the *query* function f on data X. In this work, we focus on a class $\mathcal{F}$ of Lipschitz functions that are bounded, up to a constant L, by the differences between two covariance matrices, (for example, for $\Sigma = \frac{X^T X}{n}$ and its compressed realization $\Sigma' = \frac{X^T \Phi^T \Phi X}{m}$ given Φ),

$$\mathcal{F}(L) = \left\{ f : \; |f(A) - f(D)| \leq L \, \|A - D\| \right\}, \qquad (1)$$

where A, D are positive definite matrices and $\|\cdot\|$ is understood to be any matrix norm (for example, PCA depends on $\|\Sigma - \Sigma'\|_F$). Hence we focus on releasing a multiplicative form of perturbation of the input data, such that for a particular type of functions as in (1), we achieve both utility and privacy. Due to the space limits, we only explore PCA in this paper.

We emphasize that although one could potentially release a version of the covariance matrix to preserve data privacy while performing PCA and functions as in (1), releasing the compressed data ΦX is more informative than releasing the perturbed covariance matrix (or other summaries) alone. For

example, Zhou et al. [24] demonstrated the utility of this random linear transformation by analyzing the asymptotic properties of a statistical estimator under random projection in the high dimensional setting for $n \ll p$. They showed that the relevant linear predictors can be learned from the compressed data almost as well as they could be from the original uncompressed data. Moreover, the actual predictions based on new examples are almost as accurate as they would be had the original data been made available. Finally, it is possible to release the compressed data plus some other features of the data to yield more information, although this is beyond the scope of the current paper. We note that in order to guarantee differential privacy, $p < n$ is required.

In the context of guarding privacy over a set of databases $\mathcal{S}_n = \{X_1, X_2, \ldots\}$, where $\Sigma_j = X_j^T X_j/n, \forall X_j$. we introduce an additional parameter in our privacy definition, $\Delta_{\max}(\mathcal{S}_n)$, which is an upper bound on pairwise distances between any two databases $X_1, X_2 \in \mathcal{S}_n$ (differing in any number of rows), according to a certain distance measure. In some sense, this parametrized approach of tuning the magnitude of the distance measure $\Delta_{\max}(\mathcal{S}_n)$ is the key idea we elaborate in Section III.

Toward these goals, we develop key ideas in Section IV, that include measure space truncation and renormalization for each measure $P_{\Sigma_j}, \forall j$ with Law $\mathcal{L}\left(\cdot|X_j\right) \sim N(0, \Sigma_j)$; these ideas are essential in order to guarantee differential privacy, which requires that even for very rare events, $\left|\ln P_{\Sigma_i}(\mathcal{E})/P_{\Sigma_j}(\mathcal{E})\right|$ remains small $\forall i, j$. We show that such rare events, when they happen not to be useful for the utilities that we explore, can be cut out entirely from the output space by simply discarding such outputs and regenerating a new $\mathcal{X}$. In this way, we provide a differential privacy guarantee by avoiding the comparisons made on these rare events. We conjecture that this is a common phenomenon rather than being specific to our analysis alone. In some sense, this observation is the inspiration for our *distributional* privacy definition: over a large number n of elements drawn from $\mathcal{D}$, the entire ocean of elements, the tail events are even more rare by the Law of Large Numbers, and hence we can safely truncate events whose measure $\mathbb{P}[\mathcal{E}]$ decreases as n increases.

Related work is summarized in Section I-A. Section II formalizes privacy definitions. Section III gives more detail of our probability model and summarizes our results on privacy and PCA (with proof in Section V). All proofs appear in the full version of the paper, available at http://arxiv.org/abs/0901.1365.

A. Related Work

Research on privacy in statistical data analysis has a long history, going back at least to [5]. We refer to [7] for discussion and further pointers into this literature; recent work includes [21]. Recent approaches to privacy include data swapping [14], k-anonymity [22], and cryptographic approaches (for instance, [19, 13]). Much of the work on data perturbation for privacy (for example, [12, 15, 23]) focuses on additive or multiplicative perturbation of individual records, which may not preserve similarities or other relationships within the database. Prior to [24], in [1], an information-theoretic quantification of privacy was proposed.

A body of recent work (for example, [6, 11, 3, 9, 8, 10, 18, 2, 17]) explores the tradeoffs between privacy and utility while developing the definitions and theory of *differential privacy*. The two main techniques used to achieve differential privacy to date have been additive perturbation of individual database queries by Laplace noise and the "exponential mechanism" [17]. In contrast, we provide a polynomial time non-interactive algorithm for guaranteeing differential privacy. Our goal is to show that, despite the general difficulty of achieving the differential privacy guarantee, it is possible to do so with an efficient algorithm for a specific class of functions.

The work of [16] and [24], like the work presented here, both consider low rank random linear transformations of the data X, and discuss privacy and utility. Liu et al. [16] argue heuristically that random projection should preserve utility for data mining procedures that exploit correlations or pairwise distances in the data. Their privacy analysis is restricted to observing that recovering X from ΦX requires solving an under-determined linear system. Zhou et al. [24] provide information-theoretic privacy guarantees, showing that the information rate $\frac{I(X;\mathcal{X})}{np} \to 0$ as $n \to \infty$. Their work casts privacy in terms of the rate of information communicated about X through $\mathcal{X}$, maximizing over all distributions on X. Hence their analysis provides privacy guarantees in an average sense, whereas in this work we prove differential privacy-style guarantees that aim to apply to every participant in the database semantically.

II. Definitions and preliminaries

Let a database D contain a set of n records. We focus on a non-interactive database access mechanism A such that $A(D)$ induces a distribution over sanitized output databases $\mathcal{D}'$. We first recall the differential privacy definition from [8].

Definition II.1. (α-Differential Privacy) *[8] A randomized function A gives α-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq \text{Range}(A)$, $\mathbb{P}[A(D_1) \in S] \leq e^{\alpha}\mathbb{P}[A(D_2) \in S]$.*

We now formalize our notation.

Notation: Let $\mathcal{D}$ be a collection of all records (potentially coming from some underlying distribution) and $\sigma(\mathcal{D})$ represent the entire set of input databases with elements drawn from $\mathcal{D}$. Let $\mathcal{S}_n = \{X_1, X_2, \ldots\} \subset \sigma(\mathcal{D})$, where $X_i \in \sigma(\mathcal{D}), \forall i$, denote a set of databases, each with n elements drawn from $\mathcal{D}$. Although differential privacy is defined with respect to all $D, E \in \sigma(D)$, we constrain the definition of distributional privacy to the scope of $\mathcal{S}_n$, which becomes clear in Definition II.4. We let $\mathcal{D}'$ be the entire set of possible output databases.

Definition II.2. *A privacy algorithm A takes an input database $D \in \sigma(\mathcal{D})$ and outputs a probability measure P_D on $\mathcal{D}'$, where $\mathcal{D}'$ is allowed to be different from $\sigma(\mathcal{D})$. Let $\mathcal{P}$ denote all probability measures on $\mathcal{D}'$. Then a privacy algorithm is a map $A : \sigma(\mathcal{D}) \to \mathcal{P}$ where $A(D) = P_D, \forall D \in \sigma(\mathcal{D})$.*

We now define differential privacy for continuous output. We introduce an additional parameter δ which measures how different two databases are according to V below.

Definition II.3. *Let $V(D, E)$ be the distance between D and E according to a certain metric, which is related to the utility we aim to provide. Let $d(D, E)$ denote the number of rows in which D and E differ. δ-constrained α-Differential ((α, δ)-Differential Privacy) requires the following condition,*

$$\sup_{D,E:d(D,E)=1, V(D,E)\leq\delta} \Delta(P_D, P_E) \leq e^{\alpha}, \tag{2}$$

where $\Delta(P, Q) = \operatorname{ess\,sup}_{D\in\mathcal{D}'} \frac{dP}{dQ}(D)$ denotes the essential supremum over $\mathcal{D}'$ for the Radon-Nikodym derivative dP/dQ.

Let $\mathcal{S}_n = \{X_1, X_2, \ldots\}$ be a set of databases of n records. Let $\Delta_{\max}(\mathcal{S}_n)$ bound the pairwise distance between $X_i, X_j \in \mathcal{S}_n, \forall i, j$. We now introduce a notion of distributional privacy, that is similar in spirit to that in [4].

Definition II.4. (DISTRIBUTIONAL PRIVACY FOR CONTINUOUS OUTCOME) *An algorithm A satisfies* (α, δ)-distributional privacy *on $\mathcal{S}_n$, for which a global parameter $\Delta_{\max}(\mathcal{S}_n)$ is specified, if for any two databases $X_1, X_2 \in \mathcal{S}_n$ such that each consists of n elements drawn from $\mathcal{D}$, where $X_1 \cap X_2$ may not be empty, and for all sanitized outputs $\mathcal{X} \in \mathcal{D}'$,*

$$f_{X_1}(\mathcal{X}) \leq e^{\alpha} f_{X_2}(\mathcal{X}), \quad \forall X_1, X_2 \text{ s.t. } V(X_1, X_2) \leq \delta \tag{3}$$

where $f_{X_j}(\cdot)$ is the density function for the conditional distribution with law $\mathcal{L}\left(\cdot|X_j\right), \forall i$ given X_j.

Note that this composes nicely if one is considering databases that differ in multiple rows. In particular, randomness in X_j is not directly exploited in the definition as we treat elements in $X_j \in \sigma(\mathcal{D})$ as fixed data. One could assume that they come from an underlying distribution, e.g., a multivariate Gaussian $N(0, \Sigma^*)$, and infer the distance between Σ_i and its population correspondent Σ^*. We now show that distributional privacy is a stronger concept than differential privacy.

Theorem II.5. *Given $\mathcal{S}_n$, if A satisfies (α, δ)-distributional privacy as in Definition II.4 for all $X_j \in \mathcal{S}_n$, then A satisfies (α, δ)-Differential Privacy as in Definition II.3 for all $X_j \in \mathcal{S}$.*

Proof: For the same constraint parameter δ, if we guarantee that (3) is satisfied, for all $X_i, X_j \in \mathcal{S}_n$ that differ only in a single row such that $V(X_i, X_j) \leq \delta$, we have shown the α-differential privacy on $\mathcal{S}_n$; clearly, this type of guarantee is necessary in order to guarantee α-distributional privacy over all $X_i, X_j \in \mathcal{S}_n$ that satisfy the δ constraint. ∎

III. PROBABILITY MODEL AND SUMMARY OF RESULTS

Let (X_i) represent the matrix corresponding to $X_i \in \mathcal{S}_n$. By default, we use $(X_i)_j \in \mathbb{R}^p, \forall j = 1, \ldots, n$, and $(X_i^T)_j \in \mathbb{R}^n, \forall j = 1, \ldots p$ to denote row vectors and column vectors of matrix (X_i) respectively. Throughout this paper, we assume that given any $X_i \in \mathcal{S}_n$, columns are normalized,

$$\left\|(X_i^T)_j\right\|_2^2 = n, \forall j = 1, \ldots, p, \forall X_i \in \mathcal{S}_n \tag{4}$$

which can be taken as the first step of our sanitization scheme. Given X_j, $\Phi_{m\times n}$ induces a distribution over all $m \times p$ matrices in $\mathbb{R}^{m\times p}$ via $\mathcal{X} = \Phi X_j$, where $\Phi_{ij} \sim N(0, 1/n), \forall i, j$. Let $\mathcal{L}(\cdot|X_j)$ denote the conditional distribution given X_j and P_{Σ_j} denote its probability measure, where $\Sigma_j = X_j^T X_j/n, \forall X_j \in \mathcal{S}_n$. Hence $\mathcal{X} = (x_1, \ldots, x_m)^T$ is a Gaussian Ensemble composed of m i.i.d. random vectors with $\mathcal{L}\left(x_i|X_j\right) \sim N(0, \Sigma_j), \forall i = 1, \ldots, m$.

Given a set of databases $\mathcal{S}_n = \{X_1, X_2, \ldots\}$, we do assume there is a true parameter Σ^* such that $\Sigma_1, \Sigma_2, \ldots$, where $\Sigma_j = X_j^T X_j/n$, are just a sequence of empirical parameters computed from databases $X_1, X_2 \ldots \in \mathcal{S}_n$. Define

$$\Delta_{\max}(\mathcal{S}_n) := 2 \sup_{X_j\in\mathcal{S}_n} \max_{\ell,k} \left|\Sigma_j(\ell, k) - \Sigma^*(\ell, k)\right|. \tag{5}$$

Although we do not suppose we know Σ^*, we do compute $\Sigma_i, \forall i$. Thus $\Delta_{\max}(\mathcal{S}_n)$ provides an upper bound on the perturbations between any two databases $X_i, X_j \in \mathcal{S}_n$:

$$\max_{\ell,k} \left|\Sigma_i(\ell, k) - \Sigma_j(\ell, k)\right| \leq \Delta_{\max}(\mathcal{S}_n). \tag{6}$$

We now relate two other parameters that measure pairwise distances between elements in $\mathcal{S}_n$ to $\Delta_{\max}(\mathcal{S}_n)$. For a symmetric matrix M, $\lambda_{\min}(M)$, $\lambda_{\max}(M) = \|M\|_2$ are the smallest and largest eigenvalues respectively and the Frobenius norm is given by $\|M\|_F = \sqrt{\sum_i \sum_j M_{ij}^2}$.

Proposition III.1. *Subject to normalization as in* (4)*, w.l.o.g., for any two databases X_1, X_j, let $\Delta = \Sigma_1 - \Sigma_j$ and $\Gamma = \Sigma_j^{-1} - \Sigma_1^{-1} = \Sigma_j^{-1}(\Sigma_1 - \Sigma_j)\Sigma_1^{-1} = \Sigma_j^{-1}\Delta\Sigma_1^{-1}$. Suppose $\max_{\ell,k}\left|(\Sigma_1 - \Sigma_j)_{\ell k}\right| \leq \Delta_{\max}(\mathcal{S}_n), \forall j$ then*

$$\|\Delta\|_F \leq p\Delta_{\max}(\mathcal{S}_n) \quad \text{and} \tag{7}$$

$$\|\Gamma\|_F \leq \frac{\|\Delta\|_F}{\lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_j)}. \tag{8}$$

Suppose we choose a reference point Σ_1 which can be thought of as an approximation to the true value Σ^*.

Assumption 1**:** Let $\lambda_{\min}(\Sigma_1^{-1}) = \frac{1}{\lambda_{\max}(\Sigma_1)} \geq C_{\min}$ for some constant $C_{\min} > 0$. Suppose $\|\Gamma\|_2 = o(1)$ and $\|\Delta\|_2 = o(1)$.

Assumption 1 is crucial in the sense that it guarantees that all matrices in $\mathcal{S}_n$ stay away from being singular (see Lemma III.3). We are now ready to state the first main result.

Theorem III.2. *Suppose Assumption* 1 *holds. Assuming that $\|\Sigma_1\|_2, \lambda_{\min}(\Sigma_1)$ and $\lambda_{\min}(\Sigma_i), \forall X_i \in \mathcal{S}_n$ are all in the same order, and $m \geq \Omega(\ln 2np)$. Consider the worst case realization when $\|\Delta\|_F = \Theta(p\Delta_{\max}(\mathcal{S}_n))$, where $\Delta_{\max} < 1$.*

In order to guard (distributional) privacy for all $X_i \in \mathcal{S}_n$ in the sense of Definition II.4, it is sufficient if

$$\Delta_{\max}(\mathcal{S}_n) = o\left(1/(p^2\sqrt{m \ln 2np})\right). \tag{9}$$

The following lemma is a standard result on existence conditions for Σ_j^{-1} given Σ_1^{-1}. It also shows that all eigenvalue conditions in Theorem III.2 indeed hold given Assumption 1.

Lemma III.3. *Let $\lambda_{\min}(\Sigma_1) > 0$. Let $\Delta = \Sigma_1 - \Sigma_j$ and $\|\Delta\|_2 < \lambda_{\min}(\Sigma_1)$. Then $\lambda_{\min}(\Sigma_j) \geq \lambda_{\min}(\Sigma_1) - \|\Delta\|_2$.*

Next we use the result by Zwald and Blanchard for PCA as an instance from (1) to illustrate the tradeoff between parameters. Proof of Theorem III.5 appears in Section V.

Proposition III.4. ([25]) *Let A be a symmetric positive Hilbert-Schmidt operator of Hilbert space $\mathcal{H}$ with simple nonzero eigenvalues $\lambda_1 > \lambda_2 > \ldots$. Let $D > 0$ be an integer such that $\lambda_D > 0$ and $\delta_D = \frac{1}{2}(\lambda_D - \lambda_{D+1})$. Let $B \in HS(\mathcal{H})$ be another symmetric operator such that $\|B\|_F \leq \delta_D/2$ and $A+B$ is still a positive operator. Let $P^D(A)$ (resp. $P^D(A+B)$) denote the orthogonal projector onto the subspace spanned by the first D eigenvectors A (resp. $(A+B)$). Then these satisfy*

$$\|P^D(A) - P^D(A+B)\|_F \leq \|B\|_F/\delta_D. \tag{10}$$

Subject to measure truncation of at most $1/n^2$ in each $P_{\Sigma_j}, \forall X \in \mathcal{S}_n$, as we show in Section IV, we have,

Theorem III.5. *Suppose Assumption 1 holds. If we allow $\Delta_{\max}(\mathcal{S}_n) = O(\sqrt{\log p/n})$, then we essentially perform PCA on the compressed sample covariance matrix $\mathcal{X}^T\mathcal{X}/m$ effectively in the sense of Proposition III.4: that is, in the form of (10) with $A = \frac{X^TX}{n}$ and $B = \frac{\mathcal{X}^T\mathcal{X}}{m} - A$, where $\|B\|_F = o(1)$ for $m = \Omega(p^2 \ln 2np)$. On the other hand, the databases in $\mathcal{S}_n$ are private in the sense of Definition II.4, so long as $p^2 = O\left(\sqrt{n/m}/\log n\right)$. Hence in the worst case, we require*

$$p = o\left(n^{1/6}/\sqrt{\ln 2np}\right).$$

As a special case, we look at the following example.

Example III.6. *Let $X_1 = \{\vec{x}_1, \ldots, \vec{x}_n\}^T$ be a matrix of $\{-1, 1\}^{n\times p}$. A neighboring matrix X_2 is any matrix obtained via changing the signs on τp bits, where $0 \leq \tau < 1$, on any $\vec{x}_i$.*

Corollary III.7. *For the Example III.6, it suffices if $p = o(n/\log n)^{1/4}$, in order to conduct PCA on compressed data, (subject to measure truncation of at most $1/n^2$ in each $P_{\Sigma_j}, \forall X \in \mathcal{S}_n$,) effectively in the sense of Proposition III.4, while preserve the α-differential privacy for $\alpha = o(1)$.*

IV. Distributional privacy with bounded $\Delta_{\max}(\mathcal{S}_n)$

In this section, we show how we can *modify* the output events $\mathcal{X}$ to effectively hide some large-tail events. We make it clear how these tail events are connected to a particular type of utility. Given X_i, let $\mathcal{X} = \Phi X_i = (x_1, \ldots, x_m)^T$. Let $f_{\Sigma_i}(x_j) = \exp\left\{-\frac{1}{2}x_j^T\Sigma_i^{-1}x_j\right\}/|\Sigma_i|^{1/2}(2\pi)^{p/2}$ be the density for Gaussian distribution $N(0, \Sigma_i)$. Before modification, the density function $f_{\Sigma_i}(\mathcal{X})$ is

$$f_{\Sigma_i}(\mathcal{X}) = \prod_{j=1}^{m} f_{\Sigma_i}(x_j). \tag{11}$$

We focus on defining two procedures that lead to both distributional and differential types of privacy. Indeed, the proof of Theorem IV.6 applies to both, as the distance metric $V(X_1, X_2)$ does not specify how many rows X_1 and X_2 differ in. We use $\Delta_{\max}$ as a shorthand for $\Delta_{\max}(\mathcal{S}_n)$ when it is clear.

Procedure IV.1. (Truncation of the Tail for Random Vectors in $\mathbb{R}^p$) We require Φ to be an independent random draw each time we generate a $\mathcal{X}$ for compression (or when we apply it to the same dataset for handling a truncation event). W.l.o.g, we choose Σ_1 to be a reference point. Now we only examine output databases $\mathcal{X} \in \mathbb{R}^{m\times p}$ such that for $C = \sqrt{2(C_1 + C_2)}$, where $C_1 \approx 2.5$ and $C_2 \approx 7.7$,

$$\max_{j,k}\left|(\mathcal{X}^T\mathcal{X}/m)_{jk} - \Sigma_1(j,k)\right| \leq C\sqrt{\ln 2np/m} + \Delta_{\max}, \tag{12}$$

where $\Delta_{\max}(\mathcal{S}_n) = O\left(\sqrt{\log n/n}\right)$. Algorithmically, one can imagine that for an input X, each time we see an output $\mathcal{X} = \Phi X$ that does not satisfy our need in the sense of (12), we throw the output database $\mathcal{X}$ away, and generate a new random draw Φ' to calculate $\Phi' X$ and repeat until $\Phi' X$ indeed satisfies (12). We also note that the adversary neither sees the databases we throw away nor finds out that we did so.

Given $X_i \in \mathcal{S}_n$, let $\mathbb{P}_{\Sigma_i}$ be the probability measure over random outcomes of ΦX_i. Upon truncation,

Procedure IV.2. (Renormalization) We set $f'_{\Sigma_i}(\mathcal{X}) = 0$ for all $\mathcal{X} \in \mathbb{R}^{m\times p}$ belonging to set E, where $E =$

$$\left\{\mathcal{X} : \max_{j,k}\left|\left(\frac{\mathcal{X}^T\mathcal{X}}{m}\right)_{jk} - \Sigma_1(j,k)\right| > C\sqrt{\frac{\ln 2np}{m}} + \Delta_{\max}\right\}, \tag{13}$$

corresponds to the bad events that we truncate from the outcome in Procedure IV.1; We then renormalize the density as in (11) on the remaining $\mathcal{X}$ that satisfies (12) to obtain:

$$f'_{\Sigma_i}(\mathcal{X}) = \frac{f_{\Sigma_i}(\mathcal{X})}{1 - \mathbb{P}_{\Sigma_i}[E]}. \tag{14}$$

Remark IV.3. Hence $\frac{f'_{\Sigma_1}(\mathcal{X})}{f'_{\Sigma_2}(\mathcal{X})} = \frac{f_{\Sigma_1}(\mathcal{X})(1-\mathbb{P}_{\Sigma_2}[E])}{f_{\Sigma_2}(\mathcal{X})(1-\mathbb{P}_{\Sigma_1}[E])}$, which changes $\alpha(m, \delta)$ that we bounded below based on original density prior to truncation of E by a constant in the order of $\ln(1+\epsilon) = O(\epsilon)$, where $\epsilon = O(1/n^2)$. Hence we safely ignore this normalization issue given it only changes $\alpha(m, \delta)$ by $O(1/n^2)$.

The following lemma bounds the probability on the events that we truncate in Procedure IV.1.

Lemma IV.4. *According to any individual probability measure $\mathbb{P}_{\Sigma_i}$ which corresponds to the sample space for outcomes of ΦX_i, suppose that the columns of (X_i) have been normalized to have $\left\|(X_i^T)_j\right\|_2^2 = n, \forall i, j = 1, \ldots, p$ and $m \geq 2(C_1 + C_2)\ln 2np$, then for E as defined in (13), $\mathbb{P}_{\Sigma_i}[E] \leq \frac{1}{n^2}$.*

As hinted after Definition II.4 regarding distributional privacy, we can think of the input data as coming from a distribution, such that $\Delta_{\max}(\mathcal{S}_n)$ in (5) can be derived with a typical large deviation bound between the sample and population covariances. For example, for multivariate Gaussian,

Lemma IV.5. ([20]) *Suppose $(X_i)_j \sim N(0, \Sigma^*), \forall j = 1, \ldots, n$ for all $X_i \in \mathcal{S}_n$, then $\Delta_{\max}(\mathcal{S}_n) = O_P\left(\sqrt{\log p/n}\right)$.*

We now state the main result of this section.

Theorem IV.6. *Under Assumption* 1*, let* m *and* $\left\|(X_i^T)_j\right\|_2, \forall i, j$ *satisfy conditions in Lemma IV.4. By truncating a subset of measure at most* $1/n^2$ *from each* $\mathbb{P}_{\Sigma_i}$*, in the sense of Procedure IV.1 and renormalizing the density functions according to Procedure IV.2, we have*

$$\alpha(m,\delta) \quad \leq \quad \frac{mp\|\Delta\|_F}{2\lambda_{\min}(\Sigma_i)\lambda_{\min}(\Sigma_1)} \cdot \Xi + o(1), \qquad (15)$$

$$\textit{where } \Xi = C\sqrt{\frac{\ln 2np}{m}} + \Delta_{\max} + \frac{2\|\Delta\|_F \|\Sigma_1\|_2^2}{p\lambda_{\min}(\Sigma_i)\lambda_{\min}(\Sigma_1)},$$

when we compare each $X_i \in \mathcal{S}_n$ *with* X_1*, where* $\|\Gamma\|_F$ *is bounded as as in* (7) *for* $i = 2$*.*

Remark IV.7. While the theorem only states results for comparing $\frac{f_{\Sigma_1}(\mathcal{X})}{f_{\Sigma_i}(\mathcal{X})}$, we note $\forall X_k, X_j \in \mathcal{S}_n$,

$$\left|\ln\frac{f_{\Sigma_k}(\cdot)}{f_{\Sigma_j}(\cdot)}\right| = \left|\ln\frac{f_{\Sigma_k}(\cdot)}{f_{\Sigma_1}(\cdot)}\cdot\frac{f_{\Sigma_1}(\cdot)}{f_{\Sigma_j}(\cdot)}\right| \leq \left|\ln\frac{f_{\Sigma_1}(\cdot)}{f_{\Sigma_k}(\cdot)}\right| + \left|\ln\frac{f_{\Sigma_1}(\cdot)}{f_{\Sigma_j}(\cdot)}\right|,$$

which is simply a sum of terms as bounded as in (15).

V. Proof of Theorem III.5

Combining the following theorem, which illustrates the tradeoff between the parameters n, p and m for PCA, with Theorem III.2, we obtain Theorem III.5.

Theorem V.1. *For a database* $X \in \mathcal{S}_n$*, let* A*,* $A+B$ *be the original and compressed sample covariance matrices respectively:* $A = \frac{X^TX}{n}$ *and* $B = \frac{\mathcal{X}^T\mathcal{X}}{m} - \frac{X^TX}{n}$*, where* $\mathcal{X}$ *is generated via Procedure IV.1. By requiring that* $m = \Omega(p^2 \ln 2np)$*, we can achieve meaningful bounds in the form of* (10)*.*

Proof: We know that A and $A+B$ are both positive definite, and B is symmetric. We first obtain a bound on $\|B\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p B_{ij}^2} \leq p\tau$, where

$$\begin{aligned}
\tau \quad &:= \quad \max_{jk} B_{jk} = \max_{jk}\left|(\mathcal{X}^T\mathcal{X}/m)_{jk} - A_{jk}\right| \\
&\leq \quad \max_{jk}\left|(\mathcal{X}^T\mathcal{X}/m)_{jk} - \Sigma_1(j,k)\right| + \left|\Sigma_1(j,k) - A_{jk}\right| \\
&\leq \quad C\sqrt{\ln 2np/m} + 2\Delta_{\max}(\mathcal{S}_n),
\end{aligned}$$

by (12), (6), and the triangle inequality, for $\mathcal{X} = \Phi X$. The theorem follows by Proposition III.4 given that $\|B\|_F = o(1)$ for $m = \Omega(p^2 \ln 2np)$. ∎

Acknowledgments. We thank Avrim Blum and John Lafferty for helpful discussions. KL is supported in part by an NSF Graduate Research Fellowship. LW and SZ's research was supported by NSF grant CCF-0625879, a Google research grant and a grant from Carnegie Mellon's Cylab.

References

[1] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *In Proceedings of the 20th PODS*, May 2001.

[2] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *Proceedings of the twenty-sixth ACM PODS*, 2007.

[3] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *In Proceedings of the 24th PODS*, 2005.

[4] A. Blum, K. Ligett, and A. Roth. A Learning theory approach to non-interactive database privacy. *Proceedings of the 40th STOC*, 2008.

[5] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

[6] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *In Proceedings of the 22nd PODS*, 2003.

[7] G. Duncan and R. Pearson. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6(3):219–232, August 1991.

[8] C. Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming–ICALP 2006*, pages 1–12, 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Proceedings of the 3rd Theory of Cryptography Conference*, 2006.

[10] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of LP decoding. *Proceedings of the 39th STOC*, 2007.

[11] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. *Proc. CRYPTO*, 2004.

[12] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. *Information Systems*, 29(4), 2004.

[13] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. J. Strauss, and R. N. Wright. Secure multiparty computation of approximations. *ACM Trans. Algorithms*, 2(3):435–472, 2006.

[14] S. Fienberg and J. McIntyre. Data swapping: variations on a theme by Dalenius and Reiss. *Privacy in Statistical Databases*, 3050, 2004.

[15] J. Kim and W. Winkler. Multiplicative noise for masking continuous data. *Statistics*, 2003.

[16] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. on Knowledge and Data Engineering*, 18(1), January 2006.

[17] F. McSherry and K. Talwar. Mechanism design via differential privacy. *Proceedings of the 48th FOCS*, 2007.

[18] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. *Proceedings of the 39th STOC*, 2007.

[19] B. Pinkas. Cryptographic techniques for privacy-preserving data mining. *ACM SIGKDD Explorations Newsletter*, 4(2), 2002.

[20] A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation, 2007. Technical report 467, Dept. of Statistics, Univ. of Michigan.

[21] A. P. Sanil, A. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modeling via distributed computation. In *In Proceedings of Tenth ACM SIGKDD* 2004.

[22] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 2002.

[23] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 1965.

[24] S. Zhou, J. Lafferty, and L. Wasserman. Compressed and Privacy Sensitive Sparse Regression. *IEEE Trans. Info. Theory*, 55(2), February 2009.

[25] L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. *In Advances in Neural Information Processing Systems (NIPS) 17*, 2005.