
ADJUSTING QUERIES TO STATISTICAL PROCEDURES UNDER DIFFERENTIAL PRIVACY

Tomer Shoham

Department of Computer Science and Federmann Center for the Study of Rationality
The Hebrew University of Jerusalem
Tomer.Shoham@mail.huji.ac.il

Yosef Rinott

Department of Statistics and Federmann Center for the Study of Rationality
The Hebrew University of Jerusalem
Yosef.Rinott@mail.huji.ac.il

ABSTRACT

We consider a dataset S held by an agency, and a vector query of interest, $f(S) \in \mathbb{R}^k$, to be posed by an analyst, which contains the information required for certain planned statistical inference. The agency releases the requested vector query with noise that guarantees a given level of Differential Privacy – $\text{DP}(\varepsilon, \delta)$ – using the well-known Gaussian mechanism. The analyst can choose to pose the vector query $f(S)$ or to adjust it by a suitable transformation that can make the agency’s response more informative. For any given level of privacy $\text{DP}(\varepsilon, \delta)$ decided by the agency, we study natural situations where the analyst can achieve better statistical inference by adjusting the query with a suitable simple explicit transformation.

Keywords Gaussian Mechanism · Normal Datasets · Random Differential Privacy · Confidence Region

1 Introduction

Throughout the paper we consider a dataset or sample S given as an $n \times d$ matrix, where each row pertains to an individual in the sample, and d variables are measured for each of the n participant in the sample. The sample S is held by some agency and an analyst is interested in a vector function $f(S) = (f_1(S), \dots, f_k(S)) \in \mathbb{R}^k$ of the data, to be called a query. Thus, a query consists of k functions of the data to be posed to the agency by the analyst. We consider throughout the case $k > 1$. We assume that for privacy considerations, the agency releases the response to the query $f(S)$ with noise, using a standard Gaussian mechanism that adds independent $N(0, \sigma^2)$ noise to each coordinate of $f(S)$. The distribution of the added noise is always assumed to be known to the analyst, a standard assumption in the differential privacy literature. Two samples S and S' are said to be neighbors, denoted by $S \sim S'$, if they differ by a single individual, i.e., a single row. See, e.g., [7] for all needed details on Differential Privacy (henceforth DP). When we consider S and S' together we always assume that they are neighbors.

More generally, consider a noise mechanisms applied to S via a query $h(S) \in \mathbb{R}^k$ of the form $\mathcal{M}_h(S) = h(S) + U \in \mathbb{R}^k$, where U is a random vector. A mechanism \mathcal{M}_h is said to be $\text{DP}(\varepsilon, \delta)$ if for all (measurable) sets E we have

$$P(\mathcal{M}_h(S) \in E) \leq e^\varepsilon P(\mathcal{M}_h(S') \in E) + \delta \quad (1)$$

for all $S \sim S' \in \mathfrak{D}$, where the probability refers to the randomness of U , and \mathfrak{D} is the universe of potential datasets. For example, if S is a sample of a given size n from some population, then \mathfrak{D} is the universe of all samples that could

have been drawn and considered for dissemination. The standard definition of DP takes \mathfrak{D} to be a product C^n where C consists of all possible rows. Our results hold for any given $\varepsilon > 0$ and $\delta \in (0, 1)$, which we fix for the rest of this paper.

Our goal is to describe some simple natural examples where posing a linear transformation of the query $f(S)$, getting the agency's response via a mechanism that guarantees DP, and inverting the response to obtain the required information on $f(S)$ yields better inference on $f(S)$ in the case where S is a given fixed dataset, and on the model that generates $f(S)$, when S is a random sample.

Some related work: The principle of modifying queries for better results is not new. The Matrix Mechanism (MM) is put forward in a line of work that started with [15]. Further literature includes [8, 16, 17, 19] and numerous references therein. For given queries, MM linearly modifies the original data by applying a matrix that depends on the queries to be answered. The modified data is released with noise, and the answer to the original queries is computed. The above literature studies algorithms for finding optimal modifying matrices that minimizes the distance between the original queries and the mechanism's output relative to different utility metrics.

The difference between the above papers and ours is fourfold. First, we provide simple explicit transformations for the situations we consider rather than a numerical optimization algorithm; second, we consider continuous data while MM is directed mostly toward frequency tables and counting queries. Applying MM to continuous data leads to large and sparse tables of counts (contingency tables) and high complexity of the algorithm; third, our transformations are aimed toward specific statistical goals, rather than standard norms (metrics); and fourth, we consider also random datasets where inference is on the data-generating process.

For certain specific utility metrics and queries, optimal noise mechanisms have been found; see, e.g., [11, 12], but in general many researchers consider simple mechanisms with a well-known distribution (e.g., addition of Laplace or Gaussian iid noise) without considering optimality. In [20] a noise mechanism is proposed where the variance of the Laplace noise is random. Given a query and its sensitivity, an algorithm for optimal choice of the distribution of the Laplace variance from a certain class of distributions is provided. Optimality is with respect to the expected distance (metric) between the original query and the output of the mechanism.

When a dataset is randomly generated by some assumed distribution, it is well known that the analyst has to adjust the statistical procedure to the distribution of the observed data, taking the distribution of the added noise into account; see, e.g., [25, 22, 26, 4, 10] and references therein. Most of these results are asymptotic.

2 Fixed (non-random) datasets

Consider a dataset S held by an agency and an analyst who poses a query $f(S)$ in terms of measurement units of his choosing. For example, the components of $f(S)$ could be average age, average years of schooling, and median income in the sample S . The observed response is given with noise through a privacy mechanism applied by the data-holding agency. The analyst's goals are to construct a confidence region for $f(S)$ and to test simple hypotheses about it. For any given level ε, δ of DP, we show that instead of posing the query $f(S)$, the analyst can obtain a smaller confidence region for $f(S)$ by computing it from a query of the form $f_\xi(S) = \text{Diag}(\xi)^{1/2} f(S)$ for a suitable $\xi \in \mathbb{R}_{\geq 0}^k$ (a vector having nonnegative coordinates), where $\text{Diag}(\xi)$ is a diagonal matrix whose diagonal elements form the vector ξ . For the goal of testing hypotheses, it turns out that a different choice of ξ maximizes the power of the standard likelihood-ratio test. Thus, the analyst can achieve better inference by adjusting his queries to the planned statistical procedure.

Consider a row (x_1, \dots, x_d) in the sample S . For simplicity we assume that $x_i \in C_i$ for $i = 1, \dots, d$ for suitable sets C_i . In this case each row is in the Cartesian product $C := C_1 \times \dots \times C_d$ and we set $\mathfrak{D} := C^n$. We can also write $\mathfrak{D} := C_1 \times \dots \times C_d$, where C_i is the set of n -vectors whose coordinates are all in C_i . We assume that the agency releases data under $\text{DP}(\varepsilon, \delta)$ relative to this universe \mathfrak{D} , which is known to both the agency and the analyst.

In Section 2 we assume that the components f_i of the vector query $f = (f_1, \dots, f_k)$ are functions of disjoint sets of columns of S . This assumption is not needed in Section 3. The quantity

$$\Delta(f) := \max_{S \sim S' \in \mathfrak{D}} \|f(S) - f(S')\|,$$

where $\|\cdot\|$ denotes the L_2 norm, is known as the sensitivity of f ; higher sensitivity requires more noise for DP. Under simple assumptions on the functions f_i such as monotonicity, the agency can readily compute $\Delta(f)$, as well as

$$(\tilde{S}, \tilde{S}') := \arg\max_{S \sim S' \in \mathfrak{D}} \|f(S) - f(S')\|; \quad (2)$$

see Lemma 2.2, where it is shown that the maximization can be done separately for each coordinate of f . In general, the maximum in (2) is not unique, in which case $\arg \max$ is a set of pairs.

The agency plans to release a response to the query $f(S)$ via a standard Gaussian mechanism; that is, the response is given by

$$\mathcal{M}(S) = f(S) + U \text{ where } U \sim N(0, \sigma^2 I)$$

and I is the $k \times k$ identity matrix. The variance σ^2 is the minimal variance such that the mechanism satisfies $\text{DP}(\varepsilon, \delta)$ for given ε, δ ; it can be determined by Lemma 2.3 below, which appears in [2]. This variance depends on $\Delta(f)$; however, here f is fixed and hence suppressed.

Consider a family of queries adjusted by $\text{Diag}(\xi)$:

$$f_\xi(S) := \text{Diag}(\xi)^{1/2} f(S) = (\xi_1^{1/2} f_1(S), \dots, \xi_k^{1/2} f_k(S)).$$

In particular, for the vector ξ whose components are all equal to one we have $f_{\xi=1} = f$. Given a query from this family, the agency returns a perturbed response using a Gaussian mechanism \mathcal{M}_ξ by adding to f_ξ a Gaussian vector $U \in \mathbb{R}^k$ where $U \sim N(0, \sigma^2 I)$, that is,

$$\mathcal{M}_\xi(S) = f_\xi(S) + U.$$

It is easy to see directly or from Lemma 2.3 that we can fix σ^2 and guarantee a given level of $\text{DP}(\varepsilon, \delta)$ by choosing ξ appropriately. Hence fixing σ^2 does not result in loss of generality. This is explained immediately following Theorem 2.1.

2.1 Confidence regions

The following discussion concerns the choice of $\xi \in \mathbb{R}_{>0}^k$ such that the standard confidence region CR_ξ^t for $\mu^* := f(S)$ given in formula (3) below, which is based on the observed $\mathcal{M}_\xi(S)$, has the smallest volume. It is easy to see that allowing the variance of U in \mathcal{M}_ξ to depend on ξ does not lead to smaller volumes.

The idea is simple: intuitively it appears efficient to add more noise to the more variable components of $f(S)$ rather than “waste noise” on components with low variability. Note that “more variable” depends on both the population being measured and the chosen units of measurement. Instead of asking the agency to adjust the noise to different components, we adjust the query, and thus the agency can use a standard Gaussian mechanism. This intuition, as the whole paper, is clearly relevant only for $k > 1$.

The analyst observes $\mathcal{M}_\xi(S) = f_\xi(S) + U$, where

$$\mathcal{M}_\xi(S) = (\text{Diag}(\xi)^{1/2} f(S) + U) \sim N(\text{Diag}(\xi)^{1/2} f(S), \sigma^2 I).$$

Thus,

$$\text{Diag}(\xi)^{-1/2} \mathcal{M}_\xi(S) = (f(S) + \text{Diag}(\xi)^{-1/2} U) \sim N(\mu^*, \text{Diag}(\xi)^{-1} \sigma^2),$$

where $\mu^* := f(S)$.

The standard confidence region for μ_x based on $X \sim N(\mu_x, \Sigma)$ is $\{\mu : (X - \mu)^T \Sigma^{-1} (X - \mu) \leq t\}$; see, e.g., [1], p. 79. Thus, the confidence region for $\mu^* = f(S)$ based on $\text{Diag}(\xi)^{-1/2} \mathcal{M}_\xi(S)$ becomes

$$CR_\xi^t = \{\mu \in \mathbb{R}^k : (\text{Diag}(\xi)^{-1/2} \mathcal{M}_\xi(S) - \mu)^T (\text{Diag}(\xi) \sigma^{-2}) (\text{Diag}(\xi)^{-1/2} \mathcal{M}_\xi(S) - \mu) \leq t\}. \quad (3)$$

For any $\xi \in \mathbb{R}_{>0}^k$ and any $\mu^* \in \mathbb{R}^k$, the coverage probability $P(\mu^* \in CR_\xi^t) = P(Y \leq t)$ where $Y \sim \chi_k^2$ (the chi-square distribution with k degrees of freedom), and thus all the regions CR_ξ^t have the same confidence (coverage) level. We denote the volume by $\text{Vol}(CR_\xi^t)$. For a discussion of the volume as a measure of utility of confidence regions see, e.g., [9]. We now need the notation

$$\psi := f(\tilde{S}) - f(\tilde{S}') = (f_1(\tilde{S}) - f_1(\tilde{S}'), \dots, f_k(\tilde{S}) - f_k(\tilde{S}')),$$

where (\tilde{S}, \tilde{S}') is any pair in the set defined in (2), and we assume that $\psi_i^2 > 0$ for all i .

Theorem 2.1.

(1) For any fixed t , all confidence regions CR_ξ^t defined in Equation (3) have the same confidence level; that is, the probability $P(\mu^* \in CR_\xi^t)$ depends only on t (and not on ξ).

(2) Set

$$\Lambda(\xi) = \frac{\sqrt{\psi^T \text{Diag}(\xi) \psi}}{\sigma}.$$

If for two vectors ξ_a and ξ_b the mechanisms \mathcal{M}_{ξ_a} and \mathcal{M}_{ξ_b} have the same level of DP (that is, the same ϵ and δ) then $\Lambda(\xi_a) = \Lambda(\xi_b)$.

(3) The choice $\xi = \xi^* := c(1/\psi_1^2, \dots, 1/\psi_k^2)$ with $c = \|\psi\|^2/k$ minimizes $\text{Vol}(CR_\xi^t)$ for any $t > 0$ over all vectors $\xi \in \mathbb{R}_{>0}^k$ and associated mechanisms \mathcal{M}_ξ having the same DP level. In particular,

$$\text{Vol}(CR_{\xi^*}^t) \leq \text{Vol}(CR_{\xi=1}^t),$$

with strict inequality when $\max_i(\psi_i) \neq \min_i(\psi_i)$. The right-hand side of the inequality pertains to the query f .

Fix σ^2 to be the smallest variance such that the Gaussian mechanism $\mathcal{M}_{\xi=1}(S)$ guarantees $DP(\epsilon, \delta)$ for the query f . Part (2) of Theorem 2.1 or of Theorem 2.4 below shows that for any \mathcal{M}_ξ to have the same DP level as $\mathcal{M}_{\xi=1}$ we must have $\frac{\sqrt{\psi^T \text{Diag}(\xi) \psi}}{\sigma} = \frac{\sqrt{\psi^T \psi}}{\sigma}$.

For the proof we need two lemmas that are given first.

Lemma 2.2. For any $\xi \in \mathbb{R}_{>0}^k$.

$$\Delta(f_\xi) \equiv \max_{S \sim S' \in \mathcal{D}} \|f_\xi(S) - f_\xi(S')\| = \|f_\xi(\tilde{S}) - f_\xi(\tilde{S}')\|,$$

where the pair (\tilde{S}, \tilde{S}') is defined in Equation (2).

The proof is given in the Appendix.

For an agency willing to release the query f , releasing f_ξ under the mechanism \mathcal{M}_ξ with the same DP level does not add any complications. The agency needs to compute the sensitivity defined by $\Delta(f_\xi) \equiv \max_{S \sim S' \in \mathcal{D}} \|f_\xi(S) - f_\xi(S')\|$. By Lemma 2.2, this amounts to computing $\|f_\xi(\tilde{S}) - f_\xi(\tilde{S}')\|$ using the pair (\tilde{S}, \tilde{S}') from Equation (2), which is needed to compute the sensitivity of f . In particular, $\Delta(f_\xi) = \sqrt{\psi^T \text{Diag}(\xi) \psi}$, and we shall see below that the quantity $\Delta(f_\xi)/\sigma$ is determined by the DP level. Given a $DP(\epsilon, \delta)$ level, the agency guarantees it by choosing σ using Lemma 2.3 below.

The proof of Theorem 2.1 (and all other theorems) relies on the next lemma; it can be obtained readily from the results of [2], which hold for any query f .

Lemma 2.3. Let $\mathcal{M}(S) = f(S) + U$ be a Gaussian mechanism with $U \sim N(0, \sigma^2 I)$, and for given datasets S and S' set $D := D_{S, S'} = \|f(S) - f(S')\|$. (1) If

$$\Phi\left(\frac{D}{2\sigma} - \frac{\epsilon\sigma}{D}\right) - e^\epsilon \Phi\left(-\frac{D}{2\sigma} - \frac{\epsilon\sigma}{D}\right) \leq \delta, \quad (4)$$

then for all $E \subseteq \mathbb{R}^k$,

$$\mathbb{P}(\mathcal{M}(S) \in E) \leq e^\epsilon \mathbb{P}(\mathcal{M}(S') \in E) + \delta. \quad (5)$$

(2) Setting $\tilde{D} := \Delta(f) = \|f(\tilde{S}) - f(\tilde{S}')\|$, with (\tilde{S}, \tilde{S}') given in Equation (2), Equation (4) holds with D replaced by \tilde{D} if and only if the inequality (5) holds for all $S \sim S'$ and $E \subseteq \mathbb{R}^k$, that is, if and only if $DP(\epsilon, \delta)$ holds.

Part (2) of Lemma 2.3 coincides with Theorem 8 of [2], and the first part follows from their method of proof.

Proof of Theorem 2.1. Part (1) follows from the fact mentioned above that all these regions have confidence level $P(Y \leq t)$ where $Y \sim \mathcal{X}_k^2$. Part (2) is obtained by replacing f of Part (2) of Lemma 2.3 by f_ξ ; then (\tilde{D}/σ) becomes $\frac{\sqrt{\psi^T \text{Diag}(\xi) \psi}}{\sigma}$ and the result follows.

To prove Part (3), note that the confidence region for the adjusted query given in Equation (3) is an ellipsoid whose volume is given by:

$$\text{Vol}(CR_\xi^t) = V_k \cdot (\sigma^2 t)^{k/2} (\det[\text{Diag}(\xi)])^{-1/2}, \quad (6)$$

where V_k is the volume of the unit ball in k dimensions. By Part (2), we have to minimize the volume as a function of ξ subject to the constraint $\psi^T \text{Diag}(\xi) \psi = \psi^T \psi$, which we do by using Lagrange multipliers. See the Appendix for details. \square

Given a DP level, the volume is minimized by choosing ξ_i proportionally to $1/\psi_i^2$. Multiplying ξ by a suitable constant guarantees the desired DP level. It is easy to compute the ratio of the volumes of the optimal region and the one based on the original query f :

$$\frac{\text{Vol}(CR_{\xi^*}^t)}{\text{Vol}(CR_{\xi=1}^t)} = \left(\frac{(\prod_{i=1}^k \psi_i^2)^{1/k}}{\frac{1}{k} \sum_{i=1}^k \psi_i^2} \right)^{k/2}.$$

Clearly the ratio is bounded by one, which can be seen again by the arithmetic-geometric mean inequality. Also, if one of the coordinates ψ_i tends to zero, so does the ratio, implying the possibility of a substantial reduction in the volume obtained by using the optimally adjusted query f_{ξ^*} . We remark that the ratio is decreasing in the partial order of majorization applied to $(\psi_1^2, \dots, \psi_k^2)$; see [18].

2.2 Testing hypotheses: Likelihood-ratio test

As in Section 2.1, consider a query $f(S) \in \mathbb{R}^k$, which is observed with noise via a Gaussian privacy mechanism. Now the analyst's goal is to test the simple hypotheses $H_0 : f(S) = 0$, $H_1 : f(S) = \eta$. The null hypothesis is set at zero without loss of generality by a straightforward translation. For any $\xi \in \mathbb{R}_{\geq 0}$ (a vector with nonnegative components), let $f_{\xi}(S) = \text{Diag}(\xi)^{1/2} f(S)$ and let $\mathcal{M}_{\xi}(S) = f_{\xi}(S) + U$, where $U \sim N(0, \sigma^2 I)$ and σ^2 is the smallest variance such that the Gaussian mechanism $\mathcal{M}_{\xi=1}(S)$ guarantees $DP(\epsilon, \delta)$ for the query f .

Let $h_{\xi i}$ denote the density of $\mathcal{M}_{\xi}(S)$ under the hypothesis H_i , $i = 0, 1$. The log-likelihood ratio based on the observed $\mathcal{M}_{\xi}(S)$, $\log \left\{ \frac{h_{\xi 1}(\mathcal{M}_{\xi}(S))}{h_{\xi 0}(\mathcal{M}_{\xi}(S))} \right\}$, is proportional to $\frac{\mathcal{M}_{\xi}(S)^T \text{Diag}(\xi)^{1/2} \eta}{\sigma^2}$, which under H_0 has the $N(0, \frac{\eta^T \text{Diag}(\xi) \eta}{\sigma^2})$ distribution. The likelihood-ratio test (which by the Neyman–Pearson lemma has a well-known optimality property) rejects H_0 when the likelihood ratio is large. For a given significance level α , the rejection region has the form

$$R_{\xi} = \left\{ \mathcal{M}_{\xi}(S) : \frac{\mathcal{M}_{\xi}(S)^T \text{Diag}(\xi)^{1/2} \eta}{\sigma^2} > t \right\}, \text{ where } t = \Phi^{-1}(1 - \alpha) \frac{\sqrt{\eta^T \text{Diag}(\xi) \eta}}{\sigma}. \quad (7)$$

Let $\pi(R_{\xi}) := P_{H_1}(\mathcal{M}_{\xi}(S) \in R_{\xi})$ denote the power associated with the region R_{ξ} .

Theorem 2.4.

(1) For any fixed α and for all $\xi \in \mathbb{R}_{\geq 0}$, the rejection regions R_{ξ} defined in (7) have significance level α , that is, $P_{H_0}(R_{\xi}) = \alpha$.

(2) Assume that for two vectors ξ_a and ξ_b the mechanisms \mathcal{M}_{ξ_a} and \mathcal{M}_{ξ_b} have the same level of DP (same ϵ and δ); then $\Lambda(\xi_a) = \Lambda(\xi_b)$, where $\Lambda(\xi) = \frac{\sqrt{\psi^T \text{Diag}(\xi) \psi}}{\sigma}$.

(3) Let $j^* = \arg \max_i (\eta_i^2 / \psi_i^2)$, and define ξ^* by $\xi_{j^*}^* = \|\psi\|^2 / \psi_{j^*}^2$ and $\xi_i^* = 0 \ \forall i \neq j^*$; then the choice $\xi = \xi^*$ maximizes the power $\pi(R_{\xi})$ over all vectors $\xi \in \mathbb{R}_{\geq 0}^k$ and the associated mechanisms \mathcal{M}_{ξ} having the same DP level, and in particular $\pi(R_{\xi^*}) \geq \pi(R_{\xi=1})$, with strict inequality unless $\max_i (\eta_i^2 / \psi_i^2) = \min_i (\eta_i^2 / \psi_i^2)$.

The right-hand side of the latter inequality pertains to the original query f , and thus a query of just one coordinate of f , the one having the largest ratio of (loosely speaking) signal (η_i^2) to noise (ψ_i^2) maximizes the power of the test. Note the difference between the optimal query of Theorem 2.4 and that of Theorem 2.1, which uses all coordinates of f .

Proof of Theorem 2.4. Part (1) follows from (7) and the discussion preceding it with standard calculations; Part (2) is similar to that of Theorem 2.1. The proof of Part (3) is given in the Appendix.

3 Random, normally distributed data

So far the dataset S was considered fixed, that is, nonrandom. Statisticians often view the data as random, model the data-generating process, and study the model's parameters. Accordingly, we now assume that the dataset, denoted by T , is randomly generated as follows: the rows of T , T_1, \dots, T_n are iid, where each row $T_{\ell} \in \mathbb{R}^d$ represents d measurements of an individual in the random sample T . We also assume that f is a linear query, that is,

$$f(T) = (f_1(T), \dots, f_k(T)) = \left(\frac{1}{n} \sum_{\ell=1}^n q_1(T_{\ell}), \dots, \frac{1}{n} \sum_{\ell=1}^n q_k(T_{\ell}) \right)$$

for some functions q_1, \dots, q_k . In addition, we assume that $q(T_{\ell}) := (q_1(T_{\ell}), \dots, q_k(T_{\ell})) \sim N(\mu^*, \Sigma)$ for some unknown μ^* and a known covariance matrix Σ . The normality assumption holds when the entries of T are themselves

normal, and q_i are linear functions. Assuming normality, possibly after transformation of the data, and iid observations is quite common in statistical analysis. It follows that $f(T) \sim N(\mu^*, \Sigma_n)$, where $\Sigma_n = \Sigma/n$. This may hold approximately by the central limit theorem even if normality of the dataset is not assumed. Here we assume that Σ is known. The case where it is obtained via a privatized query is beyond the scope of this paper. Assuming that Σ is known is sometimes natural. For example, when we test hypotheses on means of subpopulations, we sometimes use the covariance matrix estimated from the general population. In annual economic surveys, for example, the focus is on change between consecutive years in, say, average income or unemployment rate; however, one can assume that the past years' covariance matrix is roughly unchanged. In Section 4 we give an example of blood-test data, where we use a correlations matrix estimated from the general population.

Since the observed data will depend only on $q(T_\ell)$, we now redefine the dataset to be S , consisting of the n iid rows $S_\ell := q(T_\ell)$, $\ell = 1, \dots, n$. The assumption $q(T_\ell) \sim N(\mu^*, \Sigma)$ implies that these rows can take any value in $C := \mathbb{R}^k$. The universe of all such matrices S is $\mathfrak{D} := C^n = \mathbb{R}^{n \times k}$.

Our goal is to construct a confidence region for the model parameter μ^* and test hypotheses about it. This can be done via the query $f(S) = \frac{1}{n} \sum_{\ell=1}^n S_\ell$ having the distribution $N(\mu^*, \Sigma_n)$; however, we show that posing the query $g(S) := \Sigma_n^{-1/2} f(S)$ under the same Random Differential Privacy parameters (RDP, to be defined below) results in smaller confidence regions. We also compare the powers of certain tests of hypotheses.

We say that a query f is invariant if $f(S)$ is invariant under permutations of the rows of S . This happens trivially when f is a linear query as defined above. If f is invariant then the distribution of the output of any mechanism that operates on f is obviously unchanged by permutations of rows. In this case it suffices to consider neighbors $S \sim S'$ of the form $S = (S_1, \dots, S_{n-1}, S_n)$, $S' = (S_1, \dots, S_{n-1}, S_{n+1})$. We assume that S_1, \dots, S_{n+1} are iid rows having some distribution Q .

Definition 3.1. $(\varepsilon, \delta, \gamma)$ -Random Differential Privacy ([13]). A random perturbation mechanism \mathcal{M} whose distribution is invariant under permutations of rows is said to be $(\varepsilon, \delta, \gamma)$ -Randomly Differentially Private, denoted by $\text{RDP}(\varepsilon, \delta, \gamma)$, if

$$P_{S_1, \dots, S_{n+1}} \left(\forall E \subseteq \mathbb{R}^k, P(\mathcal{M}(S) \in E | S) \leq e^\varepsilon P(\mathcal{M}(S') \in E | S') + \delta \right) \geq 1 - \gamma,$$

where S and S' are neighbors as above, the probability $P_{S_1, \dots, S_{n+1}}$ is with respect to $S_1, \dots, S_{n+1} \stackrel{iid}{\sim} Q$ and the probability $P(\mathcal{M}(S) \in E | S)$ refers to the noise U after conditioning on S .

In words, instead of requiring the condition of differential privacy to hold for all $S \sim S' \in \mathfrak{D}$, we require that there be a "privacy set" in which any two random neighboring datasets satisfy the DP condition, and its probability is bounded below by $1 - \gamma$. An objection to this notion may arise from the fact that under RDP "extreme" participants, who are indeed rare, are not protected, even though they may be the ones who need privacy protection the most. Since RDP is not in the worst-case analysis spirit of DP, we remark that DP can be obtained if, instead of ignoring worst cases having small probability as in RDP, the agency trims them by either removing them from the dataset or by projecting them to a given ball (that is independent of the dataset) which determines the sensitivity. Such trimming, if its probability is indeed small, corresponding to a small γ , will not overly harm the data analysis.

To define a mechanism $\mathcal{M}_h(S) = h(S) + U$ (see (1)) that satisfies $\text{RDP}(\varepsilon, \delta, \gamma)$, we need to define a "privacy set" H , which is a subset of $\mathfrak{D} \times \mathfrak{D}$ consisting of neighboring pairs (S, S') , that satisfies two conditions. **(A)**: $P((S, S') \in H) = 1 - \gamma$, where the probability P is $P_{S_1, \dots, S_{n+1}}$ of Definition 3.1, and **(B)**: Equation (1) holds for all E , and any pair of neighboring datasets $(S, S') \in H$. We then say that $\mathcal{M}_h^H(S)$ is $\text{RDP}(\varepsilon, \delta, \gamma)$ with respect to the privacy set H and the query h .

To construct a suitable H satisfying condition **(A)** note that

$$f(S) - f(S') = \frac{1}{n} [q(S_n) - q(S_{n+1})] \sim N(0, 2\Sigma/n^2) \quad (8)$$

and

$$\|g(S) - g(S')\|^2 = \|\Sigma_n^{-1/2} [f(S) - f(S')]\|^2 = \|\Sigma^{-1/2} [q(S_n) - q(S_{n+1})]\|^2/n \sim 2\mathcal{X}_k^2/n.$$

Thus, if $Y \sim \mathcal{X}_k^2$ satisfies $P(Y \leq r^2) = 1 - \gamma$ then $P(\|g(S) - g(S')\|^2 \leq 2r^2/n) = 1 - \gamma$, and we can choose the set H to be

$$H_g := \{(S, S') \in \mathfrak{D} \times \mathfrak{D} : \|g(S) - g(S')\|^2 \leq 2r^2/n\}.$$

Also, by (8), we have that $\|f(S) - f(S')\|^2$ is distributed as $(2/n)Z^T \Sigma_n Z$ with $Z \sim N(0, I)$, and it is well known (by diagonalizing Σ_n) that the latter expression has the $(2/n) \sum_{i=1}^k \lambda_i X_i^2$ distribution, where λ_i denote the eigenvalues of Σ_n and X_i are iid $\mathcal{N}(0, 1)$.

Another privacy set we consider is given by

$$H_f := \{(S, S') \in \mathfrak{D} \times \mathfrak{D} : \|f(S) - f(S')\|^2 \leq C^2\},$$

where C is such that $P((S, S') \in H_f) = 1 - \gamma$, and by (8) the constant C depends on Σ and n .

We consider three Gaussian mechanisms:

$$\begin{aligned}\mathcal{M}_g^{H_g}(S) &= g(S) + U, \text{ where } U \sim N(0, \sigma_g^2 I), \\ \mathcal{M}_f^{H_g}(S) &= f(S) + U \text{ with } U \sim N(0, \sigma_{fg}^2 I), \\ \mathcal{M}_f^{H_f}(S) &= f(S) + U \text{ with } U \sim N(0, \sigma_f^2 I),\end{aligned}$$

where the first two are with respect to the privacy set H_g , and the third is with respect to H_f . For each of the three, an appropriate noise variance σ_g^2 , σ_{fg}^2 , and σ_f^2 has to be computed, given the privacy set and the RDP parameters, so that condition **(B)** above holds. To determine the noise variance we have to compute the sensitivity of the query g on the set H_g and the sensitivity of f on both H_g and H_f .

Define the sensitivity of f and g on H_g , denoted by $D(fg)$ and $D(g)$, respectively, and the sensitivity of f on H_f , denoted by $D(f)$, as follows:

$$\begin{aligned}D(fg) &:= \max_{(S, S') \in H_g} \|f(S) - f(S')\|, \\ D(g) &:= \max_{(S, S') \in H_g} \|g(S) - g(S')\| = \sqrt{2}r/\sqrt{n}, \\ D(f) &:= \max_{(S, S') \in H_f} \|f(S) - f(S')\| = C.\end{aligned}\tag{9}$$

We compare the above three mechanisms with the same RDP level in terms of the volume of confidence regions and the power of tests of hypotheses for the model parameter μ^* , computed from data given by these mechanisms. We shall prove in Sections 3.1 and 3.2 that the mechanism $\mathcal{M}_g^{H_g}(S)$ is better than $\mathcal{M}_f^{H_g}(S)$ in terms of the volumes of confidence regions and the power of tests. It is easy to see that $D(f) \leq D(fg)$ and we show below that this implies that $\mathcal{M}_f^{H_f}(S)$ is better than $\mathcal{M}_f^{H_g}(S)$ both in terms of the volume of confidence regions, and the power of tests of simple hypotheses. We shall also show that for small γ we have $D(g) \leq D(f)$, and that $\mathcal{M}_g^{H_g}(S)$ is better than $\mathcal{M}_f^{H_f}(S)$ in terms of volume of confidence regions. The latter mechanism is discussed in Section 3.3.

By the definition of RDP, the mechanism $\mathcal{M}_f^{H_g}(S)$ satisfies $RDP(\varepsilon, \delta, \gamma)$ when (4) holds with $D = D(fg)$ and $\sigma = \sigma_{fg}$, as does the mechanisms $\mathcal{M}_g^{H_g}(S)$ with $D = D(g)$ and $\sigma = \sigma_g$, and likewise the mechanism $\mathcal{M}_f^{H_f}(S)$ with $D = D(f)$ and $\sigma = \sigma_f$.

Lemma 3.2. *If $\mathcal{M}_g^{H_g}$, $\mathcal{M}_f^{H_g}$, and $\mathcal{M}_f^{H_f}$ have the same RDP, then $D(g)/\sigma_g = D(fg)/\sigma_{fg} = D(f)/\sigma_f$. The first equality is equivalent to $\sigma_{fg}^2 = \lambda_{\max}(\Sigma_n)\sigma_g^2$, where $\lambda_{\max}(\Sigma_n)$ denotes the largest eigenvalue of Σ_n .*

Proof. The first part follows from Lemma 2.3 and the above discussion. For the second part it suffices to prove that $[D(fg)]^2 = \lambda_{\max}(\Sigma_n)[D(g)]^2$. To see this note that the maximization in (9) is equivalent to maximizing $(g(S) - g(S'))^T \Sigma (g(S) - g(S'))$ subject to $\|g(S) - g(S')\|^2 = [D(g)]^2$, and the result follows readily from Rayleigh's theorem; see, e.g., [14], Chapter 4. \square

3.1 Confidence regions

We have

$$\Sigma_n^{1/2} \mathcal{M}_g^{H_g}(S) \sim N(\mu^*, \Sigma_n(1 + \sigma_g^2)), \quad \mathcal{M}_f^{H_g}(S) \sim N(\mu^*, \Sigma_n + \sigma_{fg}^2 I), \quad \mathcal{M}_f^{H_f}(S) \sim N(\mu^*, \Sigma_n + \sigma_f^2 I).$$

The standard confidence regions for $\mu^* := E[f(S)]$ based on $\mathcal{M}_g^{H_g}(S)$, $\mathcal{M}_f^{H_g}(S)$, and $\mathcal{M}_f^{H_f}(S)$ are

$$\begin{aligned} CR_g^t &= \left\{ \mu \in \mathbb{R}^k : (\Sigma_n^{1/2} \mathcal{M}_g^{H_g}(S) - \mu)^T (\Sigma_n(1 + \sigma_g^2))^{-1} (\Sigma_n^{1/2} \mathcal{M}_g^{H_g}(S) - \mu) \leq t \right\}, \\ CR_{fg}^t &= \left\{ \mu \in \mathbb{R}^k : (\mathcal{M}_f^{H_g}(S) - \mu)^T (\Sigma_n + \sigma_{fg}^2 I)^{-1} (\mathcal{M}_f^{H_g}(S) - \mu) \leq t \right\}, \\ CR_f^t &= \left\{ \mu \in \mathbb{R}^k : (\mathcal{M}_f^{H_f}(S) - \mu)^T (\Sigma_n + \sigma_f^2 I)^{-1} (\mathcal{M}_f^{H_f}(S) - \mu) \leq t \right\}. \end{aligned}$$

The next theorem shows that confidence regions based on $\mathcal{M}_g^{H_g}$ have a smaller volume than those based on $\mathcal{M}_f^{H_g}$, and, for γ sufficiently small, also than those based on $\mathcal{M}_f^{H_f}$. Thus, of the three natural candidates we consider, $\mathcal{M}_g^{H_g}$ is the best mechanism for small γ .

Theorem 3.3.

- (1) For any fixed t , the confidence regions CR_g^t , CR_{fg}^t , and CR_f^t have the same confidence level; that is, for any μ^* we have $P(\mu^* \in CR_g^t) = P(\mu^* \in CR_{fg}^t) = P(\mu^* \in CR_f^t)$.
- (2) If the mechanisms $\mathcal{M}_g^{H_g}$, $\mathcal{M}_f^{H_g}$, and $\mathcal{M}_f^{H_f}$ have the same level of $RDP(\varepsilon, \delta, \gamma)$ then $D(g)/\sigma_g = D(fg)/\sigma_{fg} = D(f)/\sigma_f$.
- (3) $\text{Vol}(CR_g^t) \leq \text{Vol}(CR_{fg}^t)$, with strict inequality unless all the eigenvalues of Σ_n are equal.
- (4) For sufficiently small γ , $\text{Vol}(CR_g^t) \leq \text{Vol}(CR_f^t)$, with strict inequality, unless all the eigenvalues of Σ_n are equal.

Proof. Part (1) holds as in Theorem 2.1, and Part (2) holds by Lemma 3.2. The proof of Part (3), given in the Appendix, uses the relation $\sigma_{fg}^2 = \lambda_{\max}(\Sigma_n) \sigma_g^2$ of Lemma 3.2, and a straightforward eigenvalue comparison. The proof of Part (4) is somewhat more involved. It uses a comparison of distribution functions of weighted sums of independent gamma random variables and a majorization argument. Details and references are given in the Appendix.

3.2 Testing hypotheses: Likelihood-ratio test

With $E[f(S)] = \mu^*$ we consider the hypotheses $H_0 : \mu^* = 0$ and $H_1 : \mu^* = \eta$ and the mechanisms $\mathcal{M}_f^{H_g}(S)$ and $\mathcal{M}_g^{H_g}(S)$ defined above. If $\mathcal{M}_f^{H_g}(S)$ is observed then the rejection region R_{fg} of the likelihood-ratio test with significance level α has the form

$$R_{fg} = \{ \mathcal{M}_f^{H_g}(S) : \mathcal{M}_f^{H_g}(S)^T (\Sigma_n + \sigma_{fg}^2 I)^{-1} \eta > t \}, \text{ where } t = \Phi^{-1}(1 - \alpha) \sqrt{\eta^T (\Sigma_n + \sigma_{fg}^2 I)^{-1} \eta}.$$

If $\mathcal{M}_g^{H_g}(S)$ is observed then the testing problem becomes $H_0 : \mu^* = 0$ vs. $H_1 : \mu^* = \Sigma_n^{-1/2} \eta$, and the rejection region R_g of the likelihood-ratio test with significance level α has the form

$$R_g = \{ \mathcal{M}_g^{H_g}(S) : \mathcal{M}_g^{H_g}(S)^T [(1 + \sigma_g^2) I]^{-1} \Sigma_n^{-1/2} \eta > t \}, \text{ where } t = \Phi^{-1}(1 - \alpha) \sqrt{\frac{\eta^T \Sigma_n^{-1} \eta}{\sigma_g^2 + 1}}.$$

Theorem 3.4.

- (1) The rejection regions R_{fg} and R_g have the same significance level α .
- (2) If both mechanisms $\mathcal{M}_g^{H_g}$ and $\mathcal{M}_f^{H_g}$ have the same level of $RDP(\varepsilon, \delta, \gamma)$ then $D(g)/\sigma_g = D(fg)/\sigma_{fg}$.
- (3) Let $\pi(R_g)$ and $\pi(R_{fg})$ denote the power associated with the rejection regions R_g and R_{fg} , respectively; then $\pi(R_g) \geq \pi(R_{fg})$ with strict inequality, unless all the eigenvalues of Σ_n are equal.

Proof. Part (1) is similar to Part (1) of Theorem 2.4. Part (2) is already given in Theorem 3.3. The proof of Part (3), given in the Appendix, involves a simultaneous diagonalization argument and a comparison of eigenvalues using $\sigma_{fg}^2 = \lambda_{\max}(\Sigma_n) \sigma_g^2$. \square

3.3 The mechanism $\mathcal{M}_f^{H_f}(S)$

It is easy to see that the sensitivity $D(f)$ (on H_f) satisfies $D(f) \leq D(fg)$. An equal RDP level implies by Lemma 3.2 that $D(fg)/\sigma_{fg} = D(f)/\sigma_f$ and therefore $\sigma_{fg} \geq \sigma_f$. The power of the likelihood-ratio test based on $\mathcal{M}_f^{H_f}(S)$ is the same as in (11), with σ_{fg} replaced now with σ_f . This power is easily seen to be decreasing in σ and therefore

the likelihood-ratio test based on $\mathcal{M}_f^{H_f}(S)$ has a higher power than the test based on $\mathcal{M}_f^{H_g}(S)$. On the other hand, an extensive computational study shows that the power of the test based on $\mathcal{M}_g^{H_g}(S)$ may be higher or lower than that based on $\mathcal{M}_f^{H_f}(S)$ depending on the parameters involved. In the case of Figure 1 we see that the test based on $\mathcal{M}_g^{H_g}(S)$ has a higher power than that based on $\mathcal{M}_f^{H_f}(S)$.

3.4 Four degrees of naiveté

We consider the random query $f(S) \sim N(\mu^*, \Sigma_n)$. The discussion below applies in principle to other distributions. We discuss four ways of testing hypotheses on μ^* in the presence of perturbation noise having a known distribution. The discussion pertains either to a $\text{DP}(\varepsilon, \delta)$ mechanism \mathcal{M}_f , see (1), or to an $\text{RDP}(\varepsilon, \delta, \gamma)$ mechanism \mathcal{M}_f^H with some privacy set H , which together with the RDP parameters determines the variance of the added noise.

1. The most naive approach to analyzing perturbed data is to ignore the added noise altogether and determine rejection regions and significance levels as if $f(S)$ is observed without noise. In this case the test may not be optimal, and the significance level will be wrong. We call this approach *super-naive analysis*.
2. A less naive approach is to choose a test based as above on the wrong assumption that $f(S)$ is observed without noise, but to set its critical value t , which determines the significance level, according to the correct distribution of the observed data, taking the noise into account. In Figure 1 we depict the power curve (denoted by [a] in the example below) of this approach where RDP is with respect to the privacy set H_f . We call this approach *naive analysis*.
3. An even better approach is to choose the test optimally by computing the likelihood-ratio test and the significance level using the correct distribution of the observed response, taking the Gaussian noise into account. For simple queries and Gaussian noise this approach is feasible analytically. In Figure 1 we show two curves of the power under this approach, associated with \mathcal{M}_f^H when RDP is with respect to the privacy set H_g (denoted by [b]) and H_f (denoted by [c]), respectively. We call this approach *optimal analysis*.
4. In this paper we propose adjusting the query to the statistical goals of the analyst and then using the optimal rejection or confidence regions based on the observed response to the adjusted query, and on its correct distribution, taking the adjustment and the noise into account. This is accomplished by the mechanism $\mathcal{M}_g^{H_g}(S)$ discussed above. Its power curve is denoted by [d] in Figure 1. We call this approach *adjusted optimal analysis*.

The first two approaches are sometimes used by practitioners; they may be acceptable for very large samples. Their properties have been studied asymptotically.

4 A numerical example

We provide a simple data example. The privacy of medical data is of utmost importance. Consider a dataset consisting of blood-test results. A standard blood test contains 30-40 variables measured in different units, with ranges and variances that are very different. Some of these variables are highly correlated. Of the many blood-test measurements, we chose for the sake of our examples to focus on six variables: Cholesterol, High Density Lipoprotein, Apo Protein A-1, Low Density Lipoprotein, Total Lipid, and Glucose (all having the same units, MG/DL).

In this order, their covariance matrix Σ is given below. It is based on data from [21, 5, 3], and other Internet medical sources. Clearly, our goal is to provide a simple example to make our point, and not as a study of how to protect blood-test data.

$$\Sigma = \begin{pmatrix} 1600 & -160 & -400 & 840 & 800 & -40 \\ * & 400 & 160 & -175 & -200 & 0 \\ * & * & 1600 & 280 & 600 & 0 \\ * & * & * & 1225 & 700 & -35 \\ * & * & * & * & 2500 & -50 \\ * & * & * & * & * & 100 \end{pmatrix}$$

We consider the release of averages of the above six variables over a sample of size n , which will vary in our examples. It is quite standard for statisticians to assume (with justification by the central limit theorem) that such

vectors of averages are multivariate normal. We assume the agency releases data under $RDP(\varepsilon, \delta, \gamma)$. If instead of RDP we use DP and trimming as described in Section 3, we obtain essentially the same results, with $DP(4\varepsilon, \delta)$.

In the examples below we consider various parameters and alternatives.

Table 1: Examples 1–4

Example	η	n	δ	γ
(1)	(10, 5, 10, 8.75, 12.5, 2.5)	50	0.0200	10^{-4}
(2)	(10, 5, 10, 8.75, 12.5, 2.5)	50	0.0004	10^{-6}
(3)	(0, 0, 20, 0, 25, 5)	50	0.0004	10^{-6}
(4)	(0, 0, 20, 0, 25, 5)	100	0.0001	10^{-6}

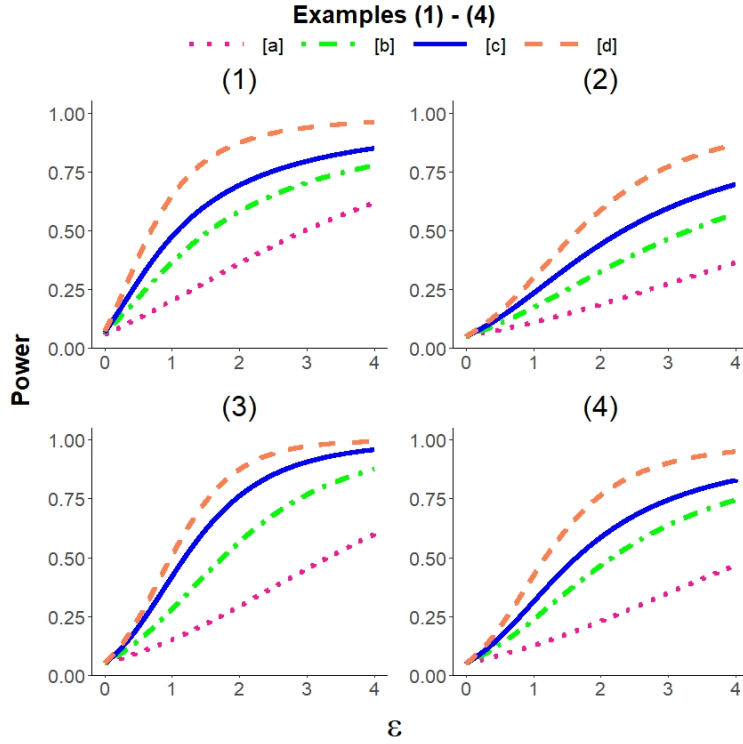


Figure 1: Comparison of power of the likelihood-ratio test with significance level $\alpha = 0.05$ as a function of ε for naive, optimal, and adjusted optimal analyses.

Acknowledgments

This work was supported in part by a gift to the McCourt School of Public Policy and Georgetown University, Simons Foundation Collaboration 733792, and Israel Science Foundation (ISF) grants 1044/16 and 2861/20.

We are grateful to Katrina Ligett and Moshe Shenfeld for very useful discussions and suggestions.

5 Appendix

Proof of Theorem 2.1 Part (3). The confidence region for the adjusted query given in (3) is an ellipsoid whose volume is given by

$$\text{Vol}(CR_\xi^t) = V_k \cdot (\sigma^2 t)^{k/2} (\det[\text{Diag}(\xi)])^{-1/2}, \quad (10)$$

where V_k is the volume of the unit ball in k dimensions. In view of Part (2) we minimize the log of the volume as a function of ξ subject to the constraint $\psi^T \text{Diag}(\xi) \psi = \psi^T \psi$. We consider the Lagrangian

$$\mathcal{L}(\xi_1, \dots, \xi_k, \lambda) = - \sum_{i=1}^k \log(\xi_i) - \lambda \left[\sum_{i=1}^k \psi_i^2 \xi_i - \sum_{i=1}^k \psi_i^2 \right].$$

We are minimizing a strictly convex function subject to a linear constraint. Differentiating and setting the Lagrangian to zero we readily obtain the unique minimum when ξ_i is proportional to $1/\psi_i^2$. The constraint $\psi^T \text{Diag}(\xi) \psi = \psi^T \psi$, which by Part (2) guarantees the same DP level, implies that $\xi^* = c(1/\psi_1^2, \dots, 1/\psi_k^2)$ with $c = \|\psi\|^2/k$. \square

Proof of Lemma 2.2. It is easy to see that $\Delta^2(f_\xi) = \max_{S \sim S' \in \mathfrak{D}} \sum_{i=1}^k \xi_i (f_i(S) - f_i(S'))^2$. We assumed that the components of the vector query $f = (f_1, \dots, f_k)$ are functions of disjoint sets of columns of S . Since $\mathfrak{D} = \mathcal{C}_1 \times \dots \times \mathcal{C}_d$, the sum is maximized by maximizing each summand individually. Multiplying each summand by a positive constant does not change the point where the maximum is achieved. \square

Proof of Theorem 2.4 Part (3). Note first that for \mathcal{M}_ξ and $\mathcal{M}_{\xi=1}$ to have the same $\text{DP}(\varepsilon, \delta)$ level we must have

$$\Lambda(\xi) = \frac{\sqrt{\psi^T \text{Diag}(\xi) \psi}}{\sigma} = \Lambda(1) = \frac{\sqrt{\psi^T \psi}}{\sigma}.$$

The power of the rejection region R_ξ is

$$\pi(R_\xi) = P_{H_1} \left(\frac{\mathcal{M}_\xi(S)^T \xi^{1/2} \eta}{\sigma^2} > \Phi^{-1}(1 - \alpha) \frac{\sqrt{\eta^T \text{Diag}(\xi) \eta}}{\sigma} \right) = 1 - \Phi \left(\Phi^{-1}(1 - \alpha) - \frac{\sqrt{\eta^T \text{Diag}(\xi) \eta}}{\sigma} \right),$$

which is increasing in $\eta^T \text{Diag}(\xi) \eta$. Thus in order to maximize the power we have to maximize $\eta^T \text{Diag}(\xi) \eta$ over ξ subject to $\psi^T \text{Diag}(\xi) \psi = \psi^T \psi$. Defining $v_i = \xi_i \psi_i^2$ the problem now is to maximize $\sum_i v_i \frac{\eta_i^2}{\psi_i^2}$ over $v_i \geq 0$, subject to $\sum_i v_i = \|\psi\|^2$. Clearly the maximum is attained when $v_{j^*} = \|\psi\|^2$, where $j^* = \arg \max_i (\eta_i^2 / \psi_i^2)$, and $v_i = 0$ for $i \neq j^*$, completing the proof. \square

Proof of Theorem 3.3. The proof of Part (3) uses the fact that

$$\text{Vol}(CR_{fg}^t) = b_k \sqrt{\det[\Sigma_n + \sigma_{fg}^2 I]} \quad \text{and} \quad \text{Vol}(CR_g^t) = b_k \sqrt{\det[\Sigma_n(1 + \sigma_g^2)]},$$

where $b_k = t^{k/2} V_k$ and V_k is the volume of the k -dimensional unit ball, and the relation $\sigma_{fg}^2 = \lambda_{\max}(\Sigma_n) \sigma_g^2$. The required inequality follows from the relations

$$\det(\Sigma_n(\sigma_g^2 + 1)) = \prod_{i=1}^k \left[\frac{\lambda_i}{\lambda_{\max}(\Sigma_n)} \sigma_{fg}^2 + \lambda_i \right] \leq \prod_{i=1}^k [\sigma_{fg}^2 + \lambda_i] = \det(\Sigma_n + \sigma_{fg}^2 I),$$

where λ_i denote the eigenvalues of Σ_n .

To prove Part (4), we need the following fact, which is a special case of a result stated in [6], Proposition 2.7 and Equation (10). The last part is given by [23], p. 999, and [24], Theorem 2.2.

Fact. Let $X_i \sim \mathcal{X}_1^2$ be iid. Without loss of generality assume that $\lambda_1, \dots, \lambda_k$ with $\lambda_i > 0$ satisfy $\bar{\lambda} := \sum_{i=1}^k \lambda_i / k = 1$. Define $F_\lambda(x) = P(\sum_{i=1}^k \lambda_i X_i \leq x)$ and let $F(x)$ denote the distribution function of \mathcal{X}_k^2 . Then for sufficiently large x we have $F_\lambda(x) \leq F(x)$.

More specifically, the latter inequality holds for $x > 2k$. The latter lower bound, given by [23, 24], is far from being tight, as suggested by numerical computations.

Recall that $C = D(f)$ satisfies $P((2/n) \sum_{i=1}^k \lambda_i X_i \leq C^2) = 1 - \gamma$. For the rest of the proof set $D = D(g)$; then D is defined by $P((2/n)Y \leq D^2) = 1 - \gamma$, where $Y \sim \mathcal{X}_k^2$; see Equation (9). For $\bar{\lambda} = 1$, which can be assumed without loss of generality, the above Fact immediately implies that $C^2 \geq D^2$ for sufficiently small γ . By the last part of the above fact, for $k = 6, 10, 20$, and 30 , sufficiently small means $\gamma \leq 1 - P(Y < 12) = 0.062$ and $\gamma \leq 0.03, 0.005$, and 0.001 , respectively.

Proof of Part (4). As in the proof of Part (3), by Lemma 3.2 and then for sufficiently small γ such that $C^2 \geq D^2 \bar{\lambda}$ (where $\bar{\lambda} = 1$), we have $\det(\Sigma_n(\sigma_g^2 + 1)) = \prod_{i=1}^k [\lambda_i + \frac{\lambda_i D^2}{C^2} \sigma_f^2] \leq \prod_{i=1}^k [\lambda_i + \lambda_i \sigma_f^2 / \bar{\lambda}]$ and now it remains to show that the latter product is bounded above by $\prod_{i=1}^k [\lambda_i + \sigma_f^2] = \det(\Sigma_n + \sigma_f^2 I)$. Dividing by $\prod_{i=1}^k \lambda_i$ and taking log, we see that the required bound is equivalent to $\sum_{i=1}^k \log[1 + \sigma_f^2 / \bar{\lambda}] \leq \sum_{i=1}^k \log[1 + \sigma_f^2 / \lambda_i]$. This follows from the fact that $\log[1 + \sigma_f^2 / \lambda]$ is convex in λ and therefore $\sum_{i=1}^k \log[1 + \sigma_f^2 / \lambda_i]$ is a Schur convex function; see [18]. \square

Proof of Theorem 3.4 Part (3). The power of the rejection region R_{fg} is given by

$$\begin{aligned} \pi(R_{fg}) &= P_{H_1} \left(\mathcal{M}_f^{H_g}(S)^T (\Sigma_n + I \sigma_{fg}^2)^{-1} \eta > \Phi^{-1}(1 - \alpha) \sqrt{\eta^T (\Sigma_n + I \sigma_{fg}^2)^{-1} \eta} \right) \\ &= 1 - \Phi \left(\Phi^{-1}(1 - \alpha) - \sqrt{\eta^T (\Sigma_n + I \sigma_{fg}^2)^{-1} \eta} \right) \\ &= 1 - \Phi \left(\Phi^{-1}(1 - \alpha) - \sqrt{\eta^T (\Sigma_n + I \lambda_{max}(\Sigma_n) \sigma_g^2)^{-1} \eta} \right). \end{aligned} \quad (11)$$

Likewise $\pi(R_g) = 1 - \Phi \left(\Phi^{-1}(1 - \alpha) - \sqrt{\frac{\eta^T \Sigma_n^{-1} \eta}{\sigma_g^2 + 1}} \right)$. Therefore, $\pi(R_g) \geq \pi(R_{fg})$ if and only if

$$\frac{\eta^T \Sigma_n^{-1} \eta}{\sigma_g^2 + 1} \geq \eta^T (\Sigma_n + I \lambda_{max}(\Sigma_n) \sigma_g^2)^{-1} \eta.$$

Diagonalizing the two matrices $(1 + \sigma_g^2) \Sigma_n$ and $\Sigma_n + I \lambda_{max}(\Sigma_n) \sigma_g^2$ by the common orthogonal matrix of their eigenvectors, we see that the diagonal terms, that is, the eigenvalues $(\sigma_g^2 + 1) \lambda_i$ of the first matrix are less than or equal to those of the second one, $\lambda_i + \lambda_{max} \sigma_g^2$. It follows that $\frac{\Sigma_n^{-1}}{\sigma_g^2 + 1} \succeq (\Sigma_n + I \lambda_{max}(\Sigma_n) \sigma_g^2)^{-1}$, where $A \succeq B$ means that $A - B$ is nonnegative definite; see [14], Chapter 4. The result follows. \square

References

- [1] Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, 2003.
- [2] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.
- [3] Conrad B Blum, RB Dell, RH Palmer, R Ramakrishnan, AH Seplowitz, and DS Goodman. Relationship of the parameters of body cholesterol metabolism with plasma levels of hdl cholesterol and the major hdl apoproteins. *Journal of lipid research*, 26(9):1079–1088, 1985.
- [4] Clément L Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 310–321, 2019.
- [5] William P Castelli, Gerald R Cooper, Joseph T Doyle, Mario Garcia-Palmieri, Tavia Gordon, Curtis Hames, Steven B Hulley, Abraham Kagan, Myron Kuchmak, Daniel McGee, et al. Distribution of triglyceride and total, ldl and hdl cholesterol in several populations: a cooperative lipoprotein phenotyping study. *Journal of chronic diseases*, 30(3):147–169, 1977.
- [6] Persi Diaconis and Michael D Perlman. Bounds for tail probabilities of weighted sums of independent gamma random variables. *Lecture Notes-Monograph Series*, pages 147–166, 1990.
- [7] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2014.
- [8] Alexander Edmonds, Aleksandar Nikolov, and Jonathan Ullman. The power of factorization mechanisms in local and central differential privacy. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 425–438, 2020.
- [9] Bradley Efron. Minimum volume confidence regions for a multivariate normal mean vector. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):655–670, 2006.
- [10] Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on machine learning*, pages 2111–2120. PMLR, 2016.
- [11] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.
- [12] Mangesh Gupte and Mukund Sundararajan. Universally optimal privacy mechanisms for minimax agents. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 135–146, 2010.
- [13] Robert Hall, Larry Wasserman, and Alessandro Rinaldo. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2), 2013.
- [14] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [15] Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. Optimizing histogram queries under differential privacy. *arXiv preprint arXiv:0912.4742*, 2009.
- [16] Chao Li and Gerome Miklau. An adaptive mechanism for accurate query answering under differential privacy. *Proceedings of the VLDB Endowment*, 5(6), 2012.
- [17] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB journal*, 24(6):757–781, 2015.
- [18] Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.
- [19] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *Proceedings of the VLDB Endowment*, 11(10), 2018.

- [20] Meisam Mohammady, Shangyu Xie, Yuan Hong, Mengyuan Zhang, Lingyu Wang, Makan Pourzandi, and Mourad Debbabi. R2dp: A universal and automated approach to optimizing the randomization mechanisms of differential privacy for utility metrics with no known optimal distributions. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 677–696, 2020.
- [21] Naeem Ahmed Qureshi, Velo Suthar, Habibullah Magsi, Muhammad Javed Sheikh, Mubeena Pathan, and Barkatullah Qureshi. Application of principal component analysis (pca) to medical data. *Indian Journal of Science and Technology*, 10(20):1–9, 2017.
- [22] Ryan Rogers and Daniel Kifer. A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics*, pages 991–1000. PMLR, 2017.
- [23] Farbod Roosta-Khorasani and Gábor J Székely. Schur properties of convolutions of gamma random variables. *Metrika*, 78(8):997–1014, 2015.
- [24] Farbod Roosta-Khorasani, Gábor J Székely, and Uri M Ascher. Assessing stochastic algorithms for large scale nonlinear least squares problems using extremal probabilities of linear combinations of gamma random variables. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):61–90, 2015.
- [25] Eftychia Solea. Differentially private hypothesis testing for normal random variables. 2014.
- [26] Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.