REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	Technical Report	-

4. TITLE AND SUBTITLE
Mixture of Gaussian Models for Classification and Hypothesis Testingunder Differential Privacy

5a. CONTRACT NUMBER	W911NF-12-1-0558
5b. GRANT NUMBER	
5c. PROGRAM ELEMENT NUMBER	611102
5d. PROJECT NUMBER	
5e. TASK NUMBER	
5f. WORK UNIT NUMBER	

6. AUTHORS
Bowei Xi, Murat Kantarcioglu , Xiaosu Tong , Ali Inan

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Texas at Dallas
800 West Campbell Road, AD15

Richardson, TX 75080 -3021

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58345-CS.9

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Gaussian mixture models are an important tool in Bayesian decision theory. In this study, we focus on building such models over statistical database protected under differential privacy. Our approach involves querying necessary statistics from a database, and using the noise added responses generated according to differential privacy in classification and hypothesis test. We first formally analyze the sensitivity of our query set. Since there are multiple methods to query a statistic, either directly or indirectly, we analyze the sensitivities for different querying methods. We discover that adding Laplace noises may become

15. SUBJECT TERMS
Differential Privacy, Statistical Database, Mixture Model, Classification, Hypothesis Testing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Murat Kantarcioglu
UU	UU	UU	UU		19b. TELEPHONE NUMBER 972-883-6616

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

Proposal/Report Number: 58345.9-CS

Report Title: Mixture of Gaussian Models for Classification and Hypothesis Testingunder Differential Privacy
Report Type: Technical Report

Mixture of Gaussian Models for Classification and Hypothesis Testing under Differential Privacy

Bowei Xi[1], **Murat Kantarcıoğlu**[2], **Xiaosu Tong**[1], **Ali Inan**[3]

[1]*Department of Statistics, Purdue University*

[2]*Department of Computer Science, University of Texas at Dallas*

[2]*Department of Computer Engineering, Isik University*

Abstract: Gaussian mixture models are an important tool in Bayesian decision theory. In this study, we focus on building such models over statistical database protected under differential privacy. Our approach involves querying necessary statistics from a database, and using the noise added responses generated according to differential privacy in classification and hypothesis test. We first formally analyze the sensitivity of our query set. Since there are multiple methods to query a statistic, either directly or indirectly, we analyze the sensitivities for different querying methods. We discover that adding Laplace noises may become problematic. For example variance-covariance matrix after noise addition is no longer positive definite. We propose a heuristic algorithm to repair the noise added variance-covariance matrix. We then examine the Bayes error under differential privacy through experiments with both simulated data and real life data, and demonstrate under which condition the impact of the added noises can be reduced. We compute the type I and type II errors under differential privacy for one sample z test, one sample t test, and two sample t test with equal variances, and show when a hypothesis test becomes unreliable under differential privacy mechanism.

Keywords: Differential Privacy, Statistical Database, Mixture Model, Classification, Hypothesis Testing

1. INTRODUCTION

Mixture models are widely used, theoretically mature tools in statistical pattern recognition and pattern classification [2, 7]. The basic assumption behind mixture models is that the data are obtained by sampling a population consisting of several distinct sub-populations with their own distributions. Gaussian mixture models refer to the case where each model follows multivariate Gaussian distribution.

Mixture models are suitable for both unsupervised learning (e.g., clustering using the Expectation Maximization algorithm) and supervised learning (e.g., classification using the Bayes decision rule). In this study, we assume that the records of an input data set belongs to different categories and focus on classification and hypothesis testing tasks. We investigate the problem of building Gaussian mixture models in a privacy-preserving environment and try to establish theoretical and experimental results with differential privacy as the privacy protection mechanism.

Building Gaussian mixture models over a specific data set requires obtaining the mean vector and the covariance matrix for each class/category. This is often a straightforward task. However, when the data set in question contains sensitive information, special care has to be taken. Consider the following motivating scenario. A medical researcher believes that a certain disease (e.g., diabetes mellitus) can be diagnosed based on a series of attributes (e.g., blood pressure, weight, height, blood sugar, etc.) that is assumed to follow multivariate Gaussian distribution and is recorded for every patient admitted to a hospital. The researcher would like to build a Gaussian mixture model and empirically test this belief using the resulting classifier. Yet, the hospital database contains highly sensitive information (e.g., disease history of the patient) and should prevent direct access to the data, even for research purposes.

Instead of granting direct access, the data users (i.e., the researcher in our example) are provided with a sanitized view of the database containing private information[1]. Various alterna-

*Correspondence to: Bowei Xi (xbw@purdue.edu)

[1]Unless the data are distributed across multiple parties, methods based on

tive privacy protection mechanisms have been suggested for producing a sanitized view. Among the first were anonymization methods such as k-anonymity [14], ℓ-diversity [12], and t-closeness [11]. Anonymization methods try to break the association between data records and individuals by grouping together similar records. Once the groups are formed, through generalization, suppression or partitioning [15] a sanitized version of the data set is released to the data users. Most definitions of anonymity (e.g., k-anonymity, ℓ-diversity, etc.) differ in the way the groups are formed.

Anonymization methods protect privacy against adversaries with certain background information. Dwork proves in [3] that every privacy protection mechanism is vulnerable to some kind of background knowledge and "bad disclosures" might occur regardless of participation into the attacked database. Therefore, Dwork suggests that instead of tailoring privacy definitions against different types of background knowledge, one should minimize the risk of disclosure that arises from participation into a database. This notion is captured by the *differential privacy* protection mechanism [3]. Differential privacy restricts the access to a statistical interface, where users can only issue aggregate statistical queries to the database and the responses are perturbed with random noises. The magnitude of the noise depends on the privacy parameter (e.g., ε in ε-differential privacy) and sensitivity of the set of queries. Sensitivity is a function of the query set and not the database. As shown in [16], computing the sensitivity is NP-hard.

In this article, we develop a privacy preserving method for the mixture of Gaussian models. This is achieved by modeling the underlying database as a statistical database protected with differential privacy against disclosures, querying necessary statistics from the database, and either building a classifer or performing a hypothesis test with the noisy responses. The classification or test results based on the noisy responses are certainly less accurate than the exact results. Nonetheless they provide preliminary information about whether or not a task could be performed with reasonable results. If the results based on the noisy responses are promising, the users can then proceed to improve the accuracy of the results.

Main contributions of this article are as follows:

1. Sensitivity of statistical queries are formally analyzed. More accurate or exact bounds for sensitivity are established.

2. We propose a heuristic algorithm to repair the noise added variance-covariance matrix, which is no longer positive definite and cannot be directly used in building a Bayesian classifier.

3. For the univariate Gaussian case, we establish theoretical bounds on the Bayes error under differential privacy based on the Bhattacharyya bound [2].

4. We examine the Bayes error for the multivariate Gaussian case through experiments, using both simulated data and real-world data. The experiments demonstrate when the impact of the added noises can be reduced.

5. We provide theoretical results on the type I and type II errors under differential privacy for hypothesis tests. We also show when a hypothesis test becomes unreliable under differential privacy mechanism.

The rest of the paper is organized as follows. We formally define the problem in Section 1.1 and provide a brief overview of differential privacy as a protection mechanism in Section 1.2. Related work in the area is discussed in Section 2. In Section 3, we calculate the sensitivity of various query sets that retrieve necessary statistics from the database. Since the exact value of sensitivity depends on the number of records, our calculation is in terms of the database size. In Section 4 we provide an algorithm to repair the noise added variance-covariance matrix, and study the Bayes error through extensive experiments using both simulated and real life data. Section 5 provides theoretical results for hypothesis tests under differential privacy. Section 6 concludes our discussion and presents future directions of research. Then, in Appendix A, we establish theoretical bounds on the Bayes error under differential privacy as the privacy protection mechanism.

1.1. Problem Definition

Let $D = \{A_1, \cdots, A_d\}$ be a d-dimensional database such that the domain $Dom(A_i)$ of each attribute A_i, $i = 1, ..., d$, is continuous and bounded. For the analysis of sensitivity in Section 3, we assume that each domain is normalized to the range $[0,1]$ to simplify the expression of sensitivity. Assume the database D is comprised of n records. Without loss of generality, we assume that D is represented as a relation. Then the value of attribute A_i of record x_k, $k = 1, ..., n$, is denoted by $x_k[A_i]$.

We are interested in building mixture of Gaussian models over databases D that fit the above description. When privacy is not a concern, this is a straightforward task. Without delving into too much details of Gaussian mixture models, let us restrict the discussion to the following: one only needs to compute the expected values of each attribute A_i and the variance-covariance matrix Σ:

$$\Sigma_{ij} = cov(A_i, A_j) = E[(A_i - \mu_i)(A_j - \mu_j)],$$

where $\mu_i = E(A_i)$. More details follow in Section 4.

In our definition of the problem, we consider a database D that contains privacy-sensitive information that is protected through differential privacy. This provides us with a statistical database interface. The interface answers aggregate queries

Secure Multi-party Computation (SMC) do not apply here.

only (e.g. count, sum etc.) and to each response adds random noise [3, 5]. In what follows, we briefly review differential privacy and analyze the sensitivities of certain queries.

1.2. Differential Privacy

Given a set of queries $Q = \{Q_1, ..., Q_q\}$, differential privacy adds Laplace noise with λ magnitude to the true response. Magnitude λ is determined by two parameters: privacy parameter ε and query set sensitivity $S(Q)$. Here, ε is assumed to be set by the data curator (i.e. the party that holds the database D). Sensitivity $S(Q)$, on the other hand, is a function of the query set Q.

Sensitivity of a query set is defined over all possible pairs of databases that differ in only one record, referred to as sibling databases.

$$S(Q) = \max_{\forall \text{ sibling databases } D_1, D_2} \sum_{i=1}^{q} |Q_i^{D_1} - Q_i^{D_2}| \quad (1)$$

That is, sensitivity of Q is the maximum difference in the total L_1 norm that a single record update can possibly cause in the query responses. Notice that the definition is independent of the original database D.

Once ε and $S(Q)$ are known, λ can be set such that $\lambda \geq S(Q)/\varepsilon$ to facilitate uninterrupted querying[2]. The rest is straightforward. In response to each query Q_i, the database first computes the result Q_i^D over all records in D and then adds Laplace noise to obtain the noisy response R_i^D:

$$R_i^D = Q_i^D + r, \quad (2)$$

where $r \sim \text{Laplace}(\lambda)$. Obviously, the key to designing accurate differential privacy mechanism is to minimize the sensitivity $S(Q)$. In our problem definition, the query set Q is already fixed. However, there are multiple methods to query a statistic. Therefore we examine the sensitivities for different query approaches separately.

2. RELATED WORK

Gaussian mixture models are classical models that are widely used in practice [2, 7]. Despite their popularity in practice, so far, privacy issues related to building mixture models have received little attention. Merugu et al. propose in [13] that instead of perturbing original data to protect privacy, in distributed settings, statistical information describing mixture models can be released. The basic idea is to generate data samples based on mixture models and run data mining tasks over the samples. However, as discussed by Kantarcioglu et al. in [9], releasing (non-perturbed) two-class mixture models might violate individual privacy. Our approach is motivated by the results of [9].

Privacy preserving data mining has been studied extensively in recent years. Initial works in the area consisted mostly of two approaches: 1) perturbation methods (e.g., random noise addition method by Agrawal et al.[1]); 2) anonymization methods (e.g., k-anonymity method proposed by Sweeney [14]) that yield a *sanitized* version of the original data set. However, successful attack strategies against proposed solutions in both directions necessitated new definitions of privacy and anonymity. For example, Kargupta et al. shows in [10] that the random noise added according to [1] could be problematic since "in many cases the original data can be accurately estimated from the perturbed data". Similarly, ℓ-diversity [12] presents an attack scenario against k-anonymity definition of [14] based on lack of diversity over sensitive attributes. Such vulnerabilities have lead to the definition of differential privacy [3]. Dwork proves in [3] that for every privacy definition, there exists some background knowledge that results in disclosure of sensitive information and therefore violation of individual privacy. Consequently, a new and much stronger privacy definition that minimizes the risk of disclosure irrespective of attendance to a database is proposed, namely, differential privacy.

Differential privacy [3] models the database as a statistical database that only responds to statistical queries and adds to the responses random noise, whose magnitude is proportional to the privacy parameter ε and the sensitivity of the query set. Here, sensitivity is a function of the query set and not the database in question.

Various different formulations of differential privacy have been suggested. Initial definitions of sensitivity operate over sibling data sets that have the same size but differ in only record (i.e., one data set can be mapped to another by updating only one record) [3, 5]. Some later studies consider insertion of a new record when defining sibling data sets [4]. The distinction between the two approaches might appear minor. However, for most query sets, the prior definition asks for sensitivity computations twice that of the later. We follow [3] in our sensitivity computations.

Sensitivity calculations of many important functions are analyzed in [5], including some statistics used in this paper as well. However, the bounds achieved by [5] are admittedly crude. Dwork et al. calculate the sensitivity of querying the mean vector as $2\gamma/n$, where n is the number of records in the database and $\gamma = \max_x ||v(x)||_1$ (i.e., the maximum L_1 norm of any record). We establish the exact sensitivity on the same query, which equals to one half of the previously established bound: d/n, where d represents the dimensionality (i.e., the number of attributes)[3]. Similarly, [5] crudely calculates the

[2] If Q is not available ahead of the time and therefore $S(Q)$ cannot be computed, λ will be fixed heuristically. In such scenarios, the database must keep track of the sensitivity of the queries answered so far. If the pre-specified sensitivity threshold λ is exceeded, the database simply stops responding.

[3] We assume that all domains are normalized to the range [0,1], therefore having the value of γ to be fixed, $\gamma = d$. This is a trivial task if the domains

sensitivity of the variance-covariance matrix Σ. Here, we provide a complete, more formal analysis of the sensitivity of the query retrieving Σ, and establish much tighter bounds.

Privacy preserving classification with differential privacy as the underlying privacy protection mechanism has received little attention so far. In [6], Friedman et al. presented a method of ID3 classification that builds a decision tree through recursive queries retrieving the information gain across an attribute and the partitioning mechanism. A different solution to ID3 classification by Jagannathan et al. [8] builds multiple random decision trees using sum queries. In this study, we present a Bayes classifier based on Gaussian mixture models by querying the mean vector and the covariance matrix for each class category. To the best of our knowledge, we are the first to explore Bayes error for Gaussian mixture models in detail under differential privacy as the protection mechanism.

3. SENSITIVITY AS FUNCTIONS OF SAMPLE SIZE AND DIMENSIONALITY

Assume two sibling databases D_1 and D_2 have n records each, and they differ by one record. Next we establish the sensitivity of queries given sample size n and d attributes. [5] provided upper bounds for the sensitivity of querying mean and variance-covariance matrix. [5] defined $\gamma = max||x'||_1$. Since all the attributes are normalized to $[0,1]$, $\gamma = d$ in our setting. [5] showed that the sensitivity of directly querying the mean is smaller than or equal to $2d/n$, and the sensitivity of querying the variance-covariance matrix is smaller than or equal to $8d^2/n$. In this section we obtain the exact sensitivity of directly querying the mean, and indirectly through querying sum and sample size, or indirectly querying the median, which is the mean for symmetric distributions. We also obtain a much tighter upper bound for querying the variance-covariance matrix.

We notice there are multiple ways to query a statistic. For example, the value of sample mean can be obtained indirectly through the sample median for any symmetric distribution. The sample mean can also be obtained through the sum divided by the sample size. Users can attempt various methods to query a statistic and to reduce sensitivity. We discuss the different sensitivities associated with the different methods to query a statistic in this section. The following summarize the findings in this section:

1. The sensitivity of directly querying mean is d/n, which decreases with increasing sample size n.

2. The sensitivity of directly querying sum is d, not affected by the sample size n, so is the sensitivity of directly querying median.

3. Notice mean can be obtained indirectly through querying median for symmetric distributions, or through querying sum and sample size. These two indirect query methods for mean have sensitivity not affected by sample size.

4. Directly querying variance has sensitivity between $\frac{1}{n} - \frac{1}{n^2}$ and $\frac{3}{n} - \frac{3}{n^2}$, so does directly querying covariance. Directly querying variance-covariance matrix (upper triangle only) has sensitivity between $(\frac{1}{n} - \frac{1}{n^2})d(d+1)/2$ and $(\frac{3}{n} - \frac{3}{n^2})d(d+1)/2$.

3.1. Directly Querying Mean and Sum

We examine the sensitivity of directly querying the mean and the sum. These two statistics are closely related. One can be solved from another. Yet the sensitivity for querying these two statistics are quite different.

Theorem 3.1 *Assume we have two sibling databases and each has n records, i.e.* $|D_1| = |D_2| = n$, *where sample size* $n \geq 1$. *Let* $Q = \{Mean_1, ..., Mean_d\}$, *where* $d \geq 1$. *Hence*

$$S(Q) = d/n.$$

Proof: Let $Mean_i^{(n-1)}$ be the mean of A_i over the common $n-1$ records shared by D_1 and D_2. Let the unique record in D_1 be x_1 and the unique record in D_2 be x_2. Then the mean values of A_i in D_1 and D_2 are

$$Mean_i^{(n),1} = \frac{(n-1) \times Mean_i^{(n-1)} + x_1[A_i]}{n},$$

$$Mean_i^{(n),2} = \frac{(n-1) \times Mean_i^{(n-1)} + x_2[A_i]}{n}.$$

We have

$$|Mean_i^{(n),1} - Mean_i^{(n),2}| = \frac{|x_1[A_i] - x_2[A_i]|}{n}.$$

Then we have

$$\begin{aligned} & max_{\{D_1,D_2\}} \textstyle\sum_1^d |Mean_i^{(n),1} - Mean_i^{(n),2}| \\ = \quad & \left(max_{\{D_1,D_2\}} \textstyle\sum_{i=1}^d |x_1[A_i] - x_2[A_i]|\right)/n \\ = \quad & d/n \quad = \quad S(Q). \end{aligned}$$

When all the d attributes in the x_1 and x_2 differ by 1, we reach the maximum, which determines the sensitivity. ∎

Theorem 3.2 *Assume we have two sibling databases and each has n records, i.e.* $|D_1| = |D_2| = n$, *where sample size* $n \geq 1$. *Let* $Q = \{Sum_1, ..., Sum_d\}$, *where* $d \geq 1$. *Hence*

$$S(Q) = d.$$

are bounded, which has to be the case since differential privacy requires a bounded domain.

Proof: Let $Sum_i^{(n-1)}$ be the sum of attribute A_i over the common $n-1$ records shared by D_1 and D_2. Again let the unique record in D_1 be x_1 and the unique record in D_2 be x_2. Then the sum of A_i in D_1 and D_2 are

$$Sum_i^{(n),1} = Sum_i^{(n-1)} + x_1[A_i],$$

$$Sum_i^{(n),2} = Sum_i^{(n-1)} + x_2[A_i].$$

When all the d attributes in the x_1 and x_2 differ by 1, we have

$$\begin{aligned} & max_{\{D_1,D_2\}} \textstyle\sum_1^d |Sum_i^{(n),1} - Sum_i^{(n),2}| \\ = \quad & max_{\{D_1,D_2\}} \textstyle\sum_{i=1}^d |x_1[A_i] - x_2[A_i]| \\ = \quad & d \quad = \quad S(Q). \end{aligned}$$

∎

The two theorems do not rely on the distribution of A_i over the interval $[0,1]$. The sensitivity of $Q = \{Mean_1, ..., Mean_d\}$ improves linearly as the sample size n increases given a fixed d. It requires the sample size to be much larger than the dimensionality, $n >> d$, to have a small sensitivity. On the other hand increasing the sample size n will not improve the sensitivity of $Q = \{Sum_1, ..., Sum_d\}$, which is determined solely by dimensionality.

Since sensitivity is defined over all possible sibling databases with all possible sample sizes, the following corollary establishes the overall sensitivity of directly querying the mean.

Corollary 3.1 *Let $Q = \{Mean_1, ..., Mean_d\}$, where $d \geq 1$. $S(Q) = d$, for all possible pairs of sibling databases.*

Proof: Following Theorem 3.1, when we set n=1, we obtain the maximum change of L_1 norm over all possible sibling databases. The problem can be solved in a more straightforward fashion. Note $Mean_i$ has minimum value 0 and maximum value 1. Let D_1 and D_2 each contains 1 record. $x_1 = \vec{0}$ and $x_2 = \vec{1}$. Then D_1 has the minimum $Mean_i$ $\forall i = 1, ..., d$ and D_2 has the maximum $Mean_i$ $\forall i = 1, ..., d$. The maximum L_1 difference is $d = S(Q)$. ∎

3.2. Directly Querying Median

For Gaussian distribution, or in general any symmetric distribution, median equals to mean. However the sensitivity of directly querying the median is quite different than that of directly querying the mean. The sensitivity of directly querying the median of d attributes is a constant d, same as directly querying the sum, regardless of sample size n.

Theorem 3.3 *Let $Q = \{ Median_1, ..., Median_d \}$, such that $Median_i$ retrieves the median of attribute A_i. Hence the overall sensitivity for for all possible pairs of sibling databases is:*

$$S(Q) = d.$$

Proof: First consider one attribute A_i. Since attribute A_i is normalized to interval $[0,1]$, the minimum value of the median is 0 and the maximum is 1. Therefore, it is sufficient to show that there is a pair of sibling databases (D_1, D_2) such that the response to $Median_i$ shifts by 1.

Let database D_1 have $2m+1$ records, $m \geq 0$, where

$$x_j[A_i] = \begin{cases} 0, & if\ 1 \leq j \leq m+1 \\ 1, & otherwise. \end{cases}$$

Construct database D_2 by changing the value of $x_m[A_i]$ from 0 to 1. Notice the response to $Median_i$ over D_1 is 0, while it is 1 over D_2, which achieves the maximum L_1 difference.

For $Q = \{Median_1, ..., Median_d\}$, similarly we let D_1 have

$$x_j = \begin{cases} \vec{0}, & if\ 1 \leq j \leq m+1 \\ \vec{1}, & otherwise. \end{cases}$$

Construct database D_2 by changing the value of x_{m+1} from $\vec{0}$ to $\vec{1}$. Hence the responses to the query over D_1 and D_2 achieve the maximum difference in L_1 norm. We conclude $S(Q) = d$, $\forall n \geq 1$. ∎

3.3. Indirectly Querying Mean

There are multiple ways of estimating a statistic. For example, querying the median is equivalent to querying the mean for any symmetric distribution. Another choice is to issue two queries, one for sum and the other for sample size.

Theorem 3.4 *Assume we have two sibling databases and each has n records, i.e. $|D_1| = |D_2| = n$, where sample size $n \geq 1$. Let $Q = \{Sum_1, ..., Sum_d, SampleSize\}$, where $d \geq 1$. Hence $S(Q) = d$.*

Proof: The query for sample size has sensitivity 0, since both D_1 and D_2 have the same sample size. Then we only need to consider the sensitivity of Sum_i. Similar to the proof of Theorem 3.2, we obtain $S(Q) = d$. ∎

3.4. Directly Querying Variance and Covariance

Next we examine the sensitivity of directly querying variance, covariance, and the whole variance-covariance matrix. We establish much tighter bounds for the sensitivity in this section.

Theorem 3.5 *Assume we have two sibling databases and each has n records, i.e. $|D_1| = |D_2| = n$, where sample size $n \geq 2$. Without loss of generality let $Q = \{Var_1\}$ for attribute A_1. Then*

$$\frac{1}{n} - \frac{1}{n^2} \leq S(Q) \leq \frac{3}{n} - \frac{3}{n^2}.$$

Proof: Assume x_3, ..., x_{n+1} are the $n-1$ common records shared by the two databases D_1 and D_2. Let x_1 be the unique

record in D_1 and x_2 be the unique record in D_2. Here we estimate the sample variance as the following:

$$Var_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i[A_1]-\bar{x}[A_1])^2 = \frac{\sum_{i=1}^{n}x_i^2[A_1]}{n} - \bar{x}^2[A_1].$$

Let Var_1^i be the sample variance of database D_i, $i = 1,2$. Then we have

$$\begin{aligned} & Var_1^1 - Var_1^2 \\ = & \left[\frac{\sum_{i=3}^{n+1}x_i^2[A_1]+x_1^2[A_1]}{n} - (\frac{\sum_{i=3}^{n+1}x_i[A_1]+x_1[A_1]}{n})^2\right] \\ - & \left[\frac{\sum_{i=3}^{n+1}x_i^2[A_1]+x_2^2[A_1]}{n} - (\frac{\sum_{i=3}^{n+1}x_i[A_1]+x_2[A_1]}{n})^2\right] \\ = & (x_1^2[A_1]-x_2^2[A_1])(\frac{1}{n}-\frac{1}{n^2}) \\ + & \frac{2(x_2[A_1]-x_1[A_1])(\sum_{i=3}^{n+1}x_i[A_1])}{n^2} \end{aligned}$$

When $x_i[A_1] = 0$, $i = 3,...,n+1$, $x_1[A_1] = 1$, and $x_2[A_1] = 0$, we have

$$Var_1^1 - Var_1^2 = \frac{1}{n} - \frac{1}{n^2}.$$

This is a lower bound for $S(Q)$.

On the other hand we have

$$\begin{aligned} |Var_1^1 - Var_1^2| \leq & \left|(x_1^2[A_1]-x_2^2[A_1])(\frac{1}{n}-\frac{1}{n^2})\right| \\ + & \left|\frac{2(x_2[A_1]-x_1[A_1])(\sum_{i=3}^{n+1}x_i[A_1])}{n^2}\right| \end{aligned}$$

We obtain an upper bound by letting every component on the right hand side of the above inequality reach their maximum individually.

$$\begin{aligned} max|Var_1^1 - Var_1^2| \leq & 1\times(\frac{1}{n}-\frac{1}{n^2}) + \frac{2\times 1\times(n-1)}{n^2} \\ = & \frac{3}{n} - \frac{3}{n^2} \end{aligned}$$

Therefore we have

$$\frac{1}{n} - \frac{1}{n^2} \leq S(Q) \leq \frac{3}{n} - \frac{3}{n^2}.$$

∎

Theorem 3.6 *Assume we have two sibling databases and each has n records, i.e. $|D_1| = |D_2| = n$, where sample size $n \geq 2$. Without loss of generality let $Q = \{Cov_{1,2}\}$ for attributes A_1 and A_2. Then*

$$\frac{1}{n} - \frac{1}{n^2} \leq S(Q) \leq \frac{3}{n} - \frac{3}{n^2}.$$

Proof: Again assume x_3, ..., x_{n+1} are the $n-1$ common records shared by the two databases D_1 and D_2. Let x_1 be the unique record in D_1 and x_2 be the unique record in D_2. The sample covariance is the following:

$$\begin{aligned} Cov_{1,2} = & \frac{1}{n}\sum_{i=1}^{n}(x_i[A_1]-\bar{x}[A_1])(x_i[A_2]-\bar{x}[A_2]) \\ = & \frac{\sum_{i=1}^{n}x_i[A_1]x_i[A_2]}{n} - \bar{x}[A_1]\bar{x}[A_2]. \end{aligned}$$

We have the difference as

$$\begin{aligned} & Cov_{1,2}^1 - Cov_{1,2}^2 \\ = & \frac{\sum_{i=3}^{n+1}x_i[A_1]x_i[A_2]+x_1[A_1]x_1[A_2]}{n} \\ - & (\frac{\sum_{i=3}^{n+1}x_i[A_1]+x_1[A_1]}{n})\times(\frac{\sum_{i=3}^{n+1}x_i[A_2]+x_1[A_2]}{n}) \\ - & \frac{\sum_{i=3}^{n+1}x_i[A_1]x_i[A_2]+x_2[A_1]x_2[A_2]}{n} \\ + & (\frac{\sum_{i=3}^{n+1}x_i[A_1]+x_2[A_1]}{n})\times(\frac{\sum_{i=3}^{n+1}x_i[A_2]+x_2[A_2]}{n}) \end{aligned}$$

Cleaning up the above expression we have

$$\begin{aligned} & Cov_{1,2}^1 - Cov_{1,2}^2 \\ = & (x_1[A_1]x_1[A_2]-x_2[A_1]x_2[A_2])(\frac{1}{n}-\frac{1}{n^2}) \\ - & (x_1[A_1]-x_2[A_1])\left(\frac{\sum_{i=3}^{n+1}x_i[A_2]}{n^2}\right) \\ - & (x_1[A_2]-x_2[A_2])\left(\frac{\sum_{i=3}^{n+1}x_i[A_1]}{n^2}\right) \end{aligned}$$

Let $x_i[A_1] = x_i[A_2] = 0$ for $i = 3,...,n+1$, $x_1[A_1] = x_1[A_2] = 1$, and $x_2[A_1] = x_2[A_2] = 0$. We have $Cov_{1,2}^1 - Cov_{1,2}^2 = 1/n - 1/n^2$. Hence this is a lower bound of $S(Q)$.

We also have

$$\begin{aligned} & |Cov_{1,2}^1 - Cov_{1,2}^2| \\ \leq & |x_1[A_1]x_1[A_2]-x_2[A_1]x_2[A_2]|(\frac{1}{n}-\frac{1}{n^2}) \\ + & |x_1[A_1]-x_2[A_1]|\left|\frac{\sum_{i=3}^{n+1}x_i[A_2]}{n^2}\right| \\ + & |x_1[A_2]-x_2[A_2]|\left|\frac{\sum_{i=3}^{n+1}x_i[A_1]}{n^2}\right| \end{aligned}$$

Let every component reach their maximum values, we have

$$\begin{aligned} max|Cov_{1,2}^1 - Cov_{1,2}^2| \leq & (\frac{1}{n}-\frac{1}{n^2}) + \frac{n-1}{n^2} + \frac{n-1}{n^2} \\ = & \frac{3}{n} - \frac{3}{n^2}. \end{aligned}$$

Therefore we have

$$\frac{1}{n}-\frac{1}{n^2} \leq S(Q) \leq \frac{3}{n}-\frac{3}{n^2}.$$

■

For large sample size n, the above result shows the sensitivity of a single variance or a single covariance decreases as $O(1/n)$. Next we consider querying the whole variance-covariance matrix.

Theorem 3.7 *Assume we have two sibling databases and each has n records, i.e.* $|D_1| = |D_2| = n$, *where sample size* $n \geq 2$. *Without loss of generality let* $Q = \{\Sigma\}$ *for d attributes. We consider only the upper triangle. Then*

$$(\frac{1}{n}-\frac{1}{n^2})\frac{d(d+1)}{2} \leq S(Q) \leq (\frac{3}{n}-\frac{3}{n^2})\frac{d(d+1)}{2}.$$

Proof: Again assume x_3, ..., x_{n+1} are the $n-1$ common records shared by the two databases D_1 and D_2. Let x_1 be the unique record in D_1 and x_2 be the unique record in D_2. We follow the thread in the above two theorems. Then we have

$$\begin{aligned}
& |Q^1 - Q^2| \\
= & \sum_{k=1}^{d-1}\sum_{l=k+1}^{d} |(x_1[A_k]x_1[A_l] - x_2[A_k]x_2[A_l])(\frac{1}{n}-\frac{1}{n^2}) \\
- & (x_1[A_k]-x_2[A_k])\left(\frac{\sum_{i=3}^{n+1} x_i[A_l]}{n^2}\right) \\
- & (x_1[A_l]-x_2[A_l])\left(\frac{\sum_{i=3}^{n+1} x_i[A_k]}{n^2}\right)| \\
+ & \sum_{k=1}^{d} |(x_1^2[A_k]-x_2^2[A_k])(\frac{1}{n}-\frac{1}{n^2}) \\
- & 2(x_1[A_k]-x_2[A_k])\left(\frac{\sum_{i=3}^{n+1} x_i[A_k]}{n^2}\right)|
\end{aligned}$$

When $x_3 = ... = x_{n+1} = \vec{0}$, $x_2 = \vec{0}$, and $x_1 = \vec{1}$, we have the above sum equal to $(\frac{1}{n}-\frac{1}{n^2})\frac{d(d+1)}{2}$. This forms a lower bound of $S(Q)$. We also have

$$\begin{aligned}
& |Q^1 - Q^2| \\
\leq & \sum_{k=1}^{d-1}\sum_{l=k+1}^{d} \{\ |x_1[A_k]x_1[A_l] - x_2[A_k]x_2[A_l]| \times (\frac{1}{n}-\frac{1}{n^2}) \\
+ & |x_1[A_k]-x_2[A_k]| \times \left|\frac{\sum_{i=3}^{n+1} x_i[A_l]}{n^2}\right| \\
+ & |x_1[A_l]-x_2[A_l]| \times \left|\frac{\sum_{i=3}^{n+1} x_i[A_k]}{n^2}\right| \ \} \\
+ & \sum_{k=1}^{d} \{\ |x_1^2[A_k]-x_2^2[A_k]| \times (\frac{1}{n}-\frac{1}{n^2}) \\
+ & 2|x_1[A_k]-x_2[A_k]| \times |\frac{\sum_{i=3}^{n+1} x_i[A_k]}{n^2}| \ \}
\end{aligned}$$

Let each component reach their maximum values (i.e. $x_3 = ... = x_{n+1} = \vec{1}$), we have

$$max|Q^1 - Q^2| \leq (\frac{3}{n}-\frac{3}{n^2})\frac{d(d+1)}{2}.$$

Hence we establish an upper bound for $S(Q)$ too. Combining the lower and upper bounds we have:

$$(\frac{1}{n}-\frac{1}{n^2})\frac{d(d+1)}{2} \leq S(Q) \leq (\frac{3}{n}-\frac{3}{n^2})\frac{d(d+1)}{2}.$$

■

We obtain a much tighter bound for querying the variance-covariance matrix. The above result indicates that in order to reduce sensitivity for querying the whole variance-covariance matrix, we need the sample size to be much larger than d^2, $n >> d^2$. Next as what we do for directly querying the mean, we can obtain an upper bound for the maximum change in L_1 norm for querying the variance-covariance matrix for all possible sibling databases with all possible sample sizes. The following establishes an upper bound for the overall sensitivity of directly querying the variance-covariance matrix.

Corollary 3.2 *Let* $Q = \{\Sigma\}$, *where* Σ *retrieves the variance-covariance matrix.* $S(Q) \leq 3d(d+1)/8$.

Proof: We let $n = 2$ in the upper bound specified by Theorem 3.7. We then obtain the overall upper bound for all possible sample size: $S(Q) \leq 3d(d+1)/8$. ■

The primary reason behind high overall sensitivity in Corollaries 3.1 and 3.2 calculations is the small sample size of the databases. Even though any databases that will be used to build Gaussian mixture models would contain thousands if not millions of records, by definition sensitivity is calculated over all possible sibling databases.

3.5. Multiple Querying Methods for A Statistic and The Effect on Sensitivity

Different methods to issue the queries for the same statistic are associated with very different sensitivity values. To obtain the sample mean, we can query the median instead if the attribute is from a symmetric distribution, or we can query the sum and the sample size. Based on the above theorems, we discover that querying the median or the sum together with sample size has sensitivity d, which is not affected by sample size n. Directly querying the mean has sensitivity d/n, fast approaching 0 as sample size increases. Some indirect queries can result in high sensitivity.

There are also alternative methods to issues a set of queries to construct variance, covariance, and a variance-covariance matrix, instead of directly querying the statistics. For example, for attribute A_1, we can query the sums and the sample size, i.e. $\sum_{i=1}^{n} x_i[A_1]$, $\sum_{i=1}^{n} x_i^2[A_1]$, and n. Another method is to query the means, i.e. $(\sum_{i=1}^{n} x_i[A_1])/n$ and $(\sum_{i=1}^{n} x_i^2[A_1])/n$.

We then construct the variance from the sums or the means. However querying the sums and querying the means have very different sensitivity values.

While working with differential privacy, we usually try to come up with query methods that will perturb the results as little as possible. However, most accurate results need not be computed with query sets of smaller sensitivities. Comparing the direct query for mean in Corollary 3.1 and the indirect query in Theorem 3.4, we observe the indirect query is more resilient to noise. Any positive or negative noise with magnitude larger than 1 completely disguise the mean value retrieved by direct querying (as in Corollary 3.1). Yet Laplace distribution has support over $(-\infty,\infty)$. The conclusion we would like to draw is that, directly querying a statistic may not always be the best idea, especially for databases with small sample size.

Later we examine databases with various sample sizes and we apply sensitivity values of directly querying the mean and variance-covariance matrix, after adjusting for the range.

4. BAYES ERROR OF GAUSSIAN MIXTURE MODELS UNDER DIFFERENTIAL PRIVACY

Let $D = \{A_1,\ldots,A_d,W\}$ be a database of n records, where W represents a binary class attribute with the domain $Dom(W) = \{w_1,w_2\}$, and each attribute A_i, $1 \le i \le d$ represents a continuous attribute with the domain $Dom(A_i) = \mathbf{R}$.

Our purpose is to build a classifier using D that, given a non-classified record in terms of a d-dimensional feature vector $\mathbf{x} \in \mathbf{R}^d$, assigns a class value to $\mathbf{x}$ such that the probability of mis-classification

$$P(error|\mathbf{x}) = \begin{cases} P(w_1|\mathbf{x}) & \text{if } \mathbf{x} \in w_2 \\ P(w_2|\mathbf{x}) & \text{if } \mathbf{x} \in w_1 \end{cases}$$

is minimized. The following Bayes' decision rule describes one such classifier:

$$\text{Assign } w_1 \text{ if } P(w_1|\mathbf{x}) > P(w_2|\mathbf{x}) \text{ ; otherwise assign } w_2. \quad (3)$$

Here, the probabilities $P(w_i|\mathbf{x})$ can easily be calculated based on Bayes' theorem:

$$P(w_i|\mathbf{x}) = \frac{p(\mathbf{x}|w_i)P(w_i)}{p(\mathbf{x})}.$$

The specific case where $p(\mathbf{x}|w_i)$ follows multivariate normal (Gaussian) density is known as the "mixture of Gaussian models" problem and it has been studied extensively due to its tractability [2]. For each class value w_i, the mean μ_i and the covariance matrix Σ_i of the distribution of $p(\mathbf{x}|w_i) \sim N(\mu_i,\Sigma_i)$ are estimated from the data set D. Based on the parameters of these distributions, the feature space $\mathbf{R}^d$ can be partitioned into possibly disconnected decision regions $\mathscr{R}_i$ such that $\mathbf{x} \in \mathscr{R}_i$ implies $\mathbf{x}$ will be classified as w_i.

The Bayes error is calculated by integrating the probability of incorrect decision(s) over decision regions. For binary classification, this implies [2]:

$$\begin{aligned} &\text{Bayes Error } = P(\mathbf{x} \in R_1, w_2) + P(\mathbf{x} \in R_2, w_1) \\ &= P(\mathbf{x} \in R_1|w_2)P(w_2) + P(\mathbf{x} \in R_2|w_1)P(w_1) \\ &= \int_{\mathscr{R}_1} p(\mathbf{x}|w_2)P(w_2)d\mathbf{x} + \int_{\mathscr{R}_2} p(\mathbf{x}|w_1)P(w_1)d\mathbf{x} \end{aligned}$$

In mixture of Gaussian models, such error can be bounded from above using the *Chernoff* bound or the *Bhattacharyya* bound as explained in [2]. Among these two approaches, the Chernoff bound is never looser than the Bhattacharyya but computationally more complex.

Our purpose is to calculate similar error bounds for privacy preserving Gaussian mixture models. Specifically, data set D acts as a statistical database that only responds to aggregate queries about the records. Using differential privacy as the underlying privacy protection mechanism, all responses to the queries will be perturbed with independent Laplace noise Laplace(λ), where $\lambda \ge S(Q)/\varepsilon$ is the magnitude of the added noise, $S(Q)$ is the sensitivity of the query set issued to the database (as defined in [3]) and ε is the privacy parameter.

In order to build a Gaussian mixture model, the query set Q including the following statistical information has to be issued to the database D:

- The number of records in D, which has sensitivity 0,
- The distribution of the classes, i.e., $P(w_1)$ and $P(w_2)$,
- For each category, parameters of the multivariate Gaussian distribution, i.e., $p(\mathbf{x}|w_i)$, in terms of μ_i and Σ_i.

4.1. Truncated Gaussian Distribution

Differential privacy works well for bounded variables. For unbounded variables one extremely large or small record has the ability to cause an extremely large change in any statistic queried and inflate the sensitivity. However Gaussian distribution has support over the entire real line. Assume we truncate a Gaussian variable to interval $[\mu - k\sigma, \mu + k\sigma]$ and the original Gaussian variable $X \sim N(\mu,\sigma^2)$ has density $f(x)$. The truncated Gaussian variable has density:

$$I_{\{\mu-k\sigma \le x \le \mu+k\sigma\}}(x)\frac{f(x)}{Z(k) - Z(-k)},$$

where $Z(\cdot)$ is the cumulative distribution function of the standard normal variable, and $I_{\mu-k\sigma\le x\le\mu+k\sigma}(x)$ is an indicator function. If we choose sufficiently large k, $Z(k) - Z(-k)$ is almost 1, and the truncated Gaussian variable and the genuine Gaussian variable have almost identical properties, such as density, mean, variance etc. We notice a Gaussian variable has probability 0.999999998 to fall into the bounded interval

$[\mu-6\sigma,\mu+6\sigma]$. Therefore in the simulation study we choose $k=6$.

Appendix A shows a theoretical upper bound for the one dimensional Bayes error with Gaussian mixture models under differential privacy for binary classes. We are not able to develop theoretical results for multivariate Gaussian distribution. We obtain information for high dimensional Bayes error through experiments.

4.2. Repair Noise Added Variance-Covariance Matrix

Let $\hat{\Sigma}=(\hat{\sigma}_{ij})_{d\times d}$ be the sample var-cov matrix. When users query variances and covariances separately, independent Laplace noises are added to every element of $\hat{\Sigma}$. Let $A=(r_{ij})_{d\times d}$ be the matrix of independent Laplace noises, where $r_{ij}=r_{ji}$. The returned query result is

$$\Sigma_Q=\hat{\Sigma}+A.$$

Σ_Q is the noise added var-cov matrix, which is the results from data that users can obtain easily to test their initial model. Σ_Q is still symmetric but seize to be positive definite. In order to have a valid var-cov matrix, we repair the noise added var-cox matrix, and obtain a positive definite matrix Σ_+ close to Σ_Q, since $\hat{\Sigma}$ is not disclosed to users under differential privacy.

Let (l_j,e_j), $j=1,...,d$ be the eigenvalue and eigenvector pairs of Σ_Q, where the eigenvalues follow the decreasing order, $l_1>l_2>...>l_d$. The last several eigenvalues of Σ_Q are negative. Let $l_k,...,l_d$ be the negative eigenvalues. The positive definite matrix Σ_+ has eigenvalue and eigenvector pairs as the following: (l_1,e_1), ..., (l_{k-1},e_{k-1}), (l_k^+,e_k), ..., (l_d^+,e_d). We keep the eigenvectors, and use an optimization algorithm to search over positive eigenvalues for a Σ_+ that minimizes the determinent of $\Sigma_+-\Sigma_Q$.

$$(l_k^+,...,l_d^+)=\text{argmin}\,|\Sigma_+-\Sigma_Q|.$$

Let $E_j=e_je_j'$, $j=1,...,d$. We have

$$\Sigma_+-\Sigma_Q=\sum_{j=k}^{d}(l_j^+-l_j)E_j.$$

Therefore we perform an exhaustive search over wide intervals to obtain positive eigenvalues that

$$(l_k^+,...,l_d^+)=\text{argmin}_{\{w_k>0,...,w_d>0\}}|\sum_{j=k}^{d}(w_j-l_j)E_j|.$$

4.3. Experimental Evaluation

In order to evaluate the performance of Gaussian mixture models learned from data under differential privacy, we have conducted extensive experiments in this section. We consider binary classification scenario in this section. Since our goal is to understand how differential privacy affects the Bayes error of Gaussian mixture models, we try to avoid introducing any other type of errors. Clearly, one of the issues with using Gaussian mixture models in practice is that Gaussian distribution may not represent the underlying data accurately. To sidestep this issue, and to make sure that we do not have additional errors due to modeling real data distribution inaccurately, we generate data sets from known Gaussian mixture parameters. The parameters are estimated from real life data in one experiment, and synthetic in the rest. By using such generated data sets, we ensure that we do not introduce errors due to wrong distribution model selection.

In Equation 3, if the two Gaussian distributions have the same var-cov matrix, we perform a linear discriminant analysis (LDA). If the two Gaussian distributions have different var-cov matrices, we perform a quadratic discriminant analysis (QDA). Every experimental run has the following steps.

1. Given the parameters of the Gaussian mixture models, we generate a training set of n samples. We truncate the training samples using the $\mu\pm 6\sigma$ interval, throwing away samples that fall out of the interval.

2. Using the truncated training set which has less than n samples, given a pre-specified ε, we compute the sensitivity values, sample means and var-cov matrices. Then we add Laplace noise to each Gaussian component.

3. We repair the noise added var-cov matrices, and obtain positive definite matrices.

4. We generate a separate test data set of size 50,000 using the original parameters without the noises, and report the effectiveness of the Gaussian mixture models using the noise added sample means and the positive definite matrices from the previous step. Test data set of size 50,000 is chosen to make sure that the estimated Bayes errors are accurate.

	2-D	5-D	10-D
Bayes error	0.2351	0.2100	0.1996

Table 1: LDA Bayes error

Experiment 1. We set $\mu_1=0.75\times 1_d$ and $\mu_2=0.25\times 1_d$, where 1_d is a d-dimensional vector with elements all equal to 1. The two d-dimensional Gaussian distributions have the same var-cov matrix Σ, where $\sigma_{ii}=0.8^2$ and $\sigma_{ij}=0.5\times 0.8^2$. The prior is $p_1=p_2=0.5$. We pool the two classes to estimate the sample var-cov matrix. We compute the sensitivity for variances and covariances using the range of the pooled data. The sample means and the sensitivity values for sample means are computed separately. We run the experiments

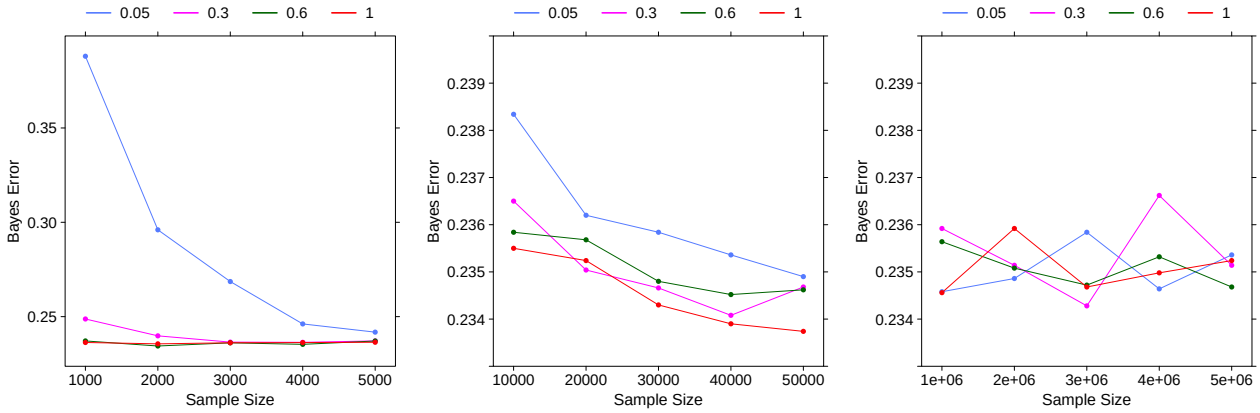

Figure 1: 2-dimensional LDA bayes error. Left: small sample size; Middle: median sample size; Right: large sample size.

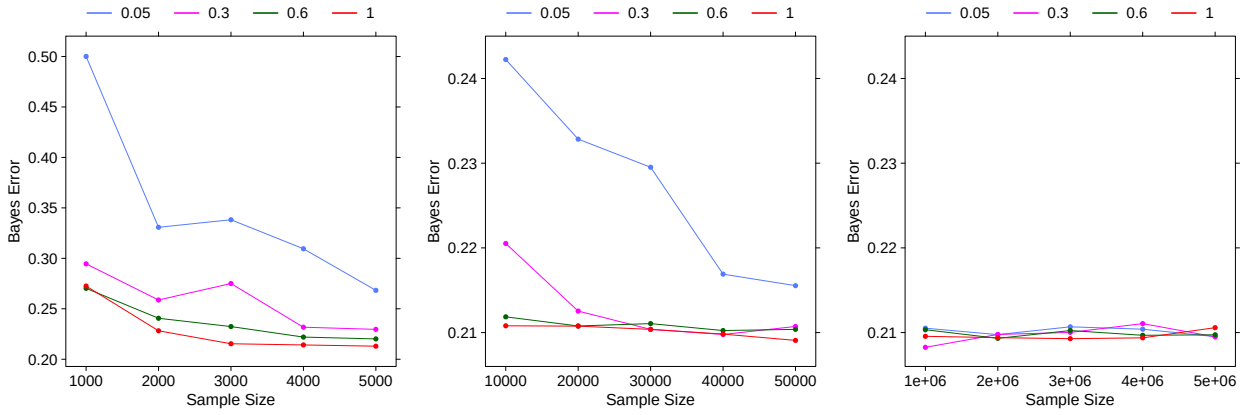

Figure 2: 5-dimensional LDA bayes error. Left: small sample size; Middle: median sample size; Right: large sample size.

in 2-dimension, 5-dimension, and 10-dimention, $d = 2, 5, 10$. We have four ε values, $\varepsilon = 0.05, 0.3, 0.6, 1$. Meanwhile we gradually increase the training set size.

Using the prespecified parameter values, we have the true LDA classification rule, following Equation 3. We generate 5 million samples using the prespecified parameter values without truncation, using the true LDA classification rules to estimate Bayes error. We take the average Bayes error of four such runs as the actual LDA Bayes error, shown in Table 1.

Figures 1, 2 and 3 show the Bayes error rate for LDA experiment in increasing dimensions. For each combination (ε, n, d), we perform five runs. The average Bayes error of five runs is shown on the Figures.

When two classes have the same var-cov matrix, the LDA Bayes error in general is not significantly affected by the noise added query results used in the classifier. For ε from 0.3 to 1, several thousand training samples are sufficient to return a preliminary Bayes error estimate which is very close to the actual LDA Bayes error. For this special case, we can obtain a fairly accurate idea about how well the LDA classifier performs using the noise added query results.

Experiment 2. We set $\mu_1 = 0.75 \times 1_d$ and $\mu_2 = 0.25 \times 1_d$. We set $\Sigma_1 = I_d$, where I_d is a d-dimensional identity matrix, and set Σ_2 as the one in Experiment 1. The prior is $p_1 = p_2 = 0.5$. The sample means, variances, covariances, and the sensitivity values are computed separately. Again, we run the experiments in 2-dimension, 5-dimension, and 10-dimention, $d = 2, 5, 10$. We have four ε values, $\varepsilon = 0.05, 0.3, 0.6, 1$. Meanwhile we gradually increase the training set size.

Using the prespecified parameter values, we have the true QDA classification rule, following Equation 3. We generate 5 million samples using the prespecified parameter values without truncation, using the true QDA classification rules to estimate Bayes error. We take the average Bayes error of four such runs as the actual QDA Bayes error, shown in Table 2.

Figures 4, 5 and 6 show the Bayes error rate for QDA experiment in increasing dimensions. For each combination (ε, n, d), we perform five runs. The average Bayes error of five runs is shown on the Figures.

When two classes have different var-cov matrices, dimensionality has a large impact on the Bayes error obtained under differential privacy. For ε from 0.3 to 1, 2 dimensional experiment shows that three thousand training samples is sufficient

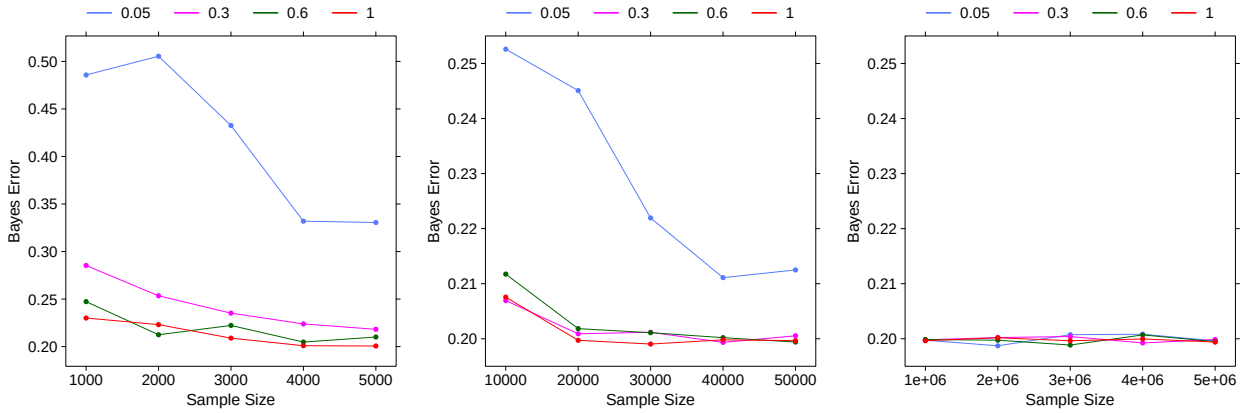

Figure 3: 10-dimensional LDA bayes error. Left: small sample size; Middle: median sample size; Right: large sample size.

to return a reasonable estimate of the actual Bayes error. 5 dimensional experiment needs 40,000 training samples to eliminate the impact of the added noises. 10 dimensional experiment needs even more training samples to return a reasonable estimate of the actual Bayes error under differential privacy.

	2-D	5-D	10-D
Bayes error	0.2105	0.1170	0.0589

Table 2: QDA Bayes error

Experiment 3. Finally, we used the Parkinson data set from the UCI Machine learning repository (https://archive.ics.uci.edu/ml/datasets/Parkinsons). We computed the mean and var-cov matrix of each class in the Parkinson data set and used these parameters in our Gaussian mixture models. In all of the experiments, we set $\varepsilon = 0.6$. For the Parkinson data set, a classifier that put all the records into the majority class has Bayes error 0.2462. Without differential privacy mechanism, directly using the sample estimates, the Bayes error is less than 0.01. On the other hand, the Gaussian mixture models with increasing sample sizes under differential privacy have Bayes error decreasing from 0.246 to 0.198. The Bayes error 0.198 is obtained from 50,000 training samples. The above results confirm that direct noise addition to Gaussian mixture parameters could cause significant distortion in higher dimensional space. As dimensionality increases, we need more training samples to reduce the impact of the added noises.

5. HYPOTHESIS TESTING UNDER DIFFERENTIAL PRIVACY

Differential privacy mechanism has a big impact on hypothesis tests because the test statistic is now created using the noise added query results, and hypothesis tests often apply to data with smaller sample size. Next we provide the distributions for the noise added test statistic under the null value and an alternative value.

Only when we know the true λs for the Laplace noises, we can numerically compute the exact p-value given a noise added test statistic. The true λs are unknown to the users querying a database. Hence in this section we examine a more realistic scenario: A rejection region is constructed using the critical values from a Gaussian distribution or a T distribution as usual, and the revised type I and type II errors can be computed numerically. In this section we consider the most commonly used hypothesis tests: the one sample z test, the one sample t test, the two sample t test with equal variance.

For the two sample t test with unequal variances, the degrees of freedom for the standard test is also affected by the added Laplace noises. Therefore we cannot simply use the critical values from a t distribution with the noise added degrees of freedom. How to contruct a proper rejection region merits more effort in this case. It is part of our future work.

5.1. One sample z test

Assume n samples $Y_1, Y_2, ..., Y_n$ i.i.d $\sim N(\mu, \sigma^2)$, where σ^2 is known. The null hypothesis is $H_0 : \mu = \mu_0$. We consider the common two-sided alternative hypothesis $H_a : \mu \neq \mu_0$ or the one-sided $H_a : \mu > \mu_0$ and $H_a : \mu < \mu_0$.

The test statistic is based on the noise added sample mean. $\bar{Y}^a = \bar{Y} + r$, where $r \sim \text{Laplace}(\lambda)$. The test statistic under differential privacy is

$$Z = \frac{\bar{Y}^a - \mu_0}{\sigma/\sqrt{n}}.$$

$\bar{Y}^a$ follows a Gaussian-Laplace mixture distribution.

$$\bar{Y}^a \sim \text{GL}(\mu, \sigma^2, n, \lambda),$$

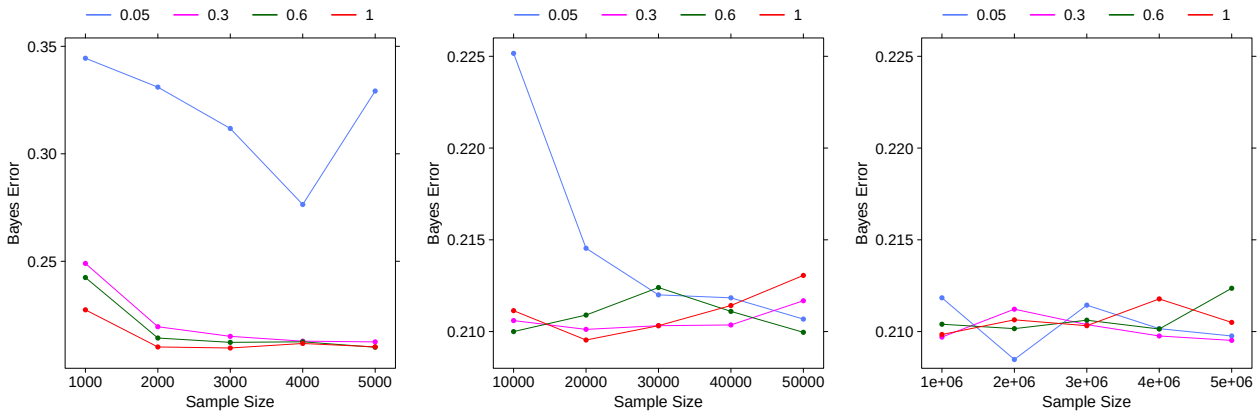

Figure 4: 2-dimensional QDA Bayes error. Left: small sample size; Middle: median sample size; Right: large sample size.

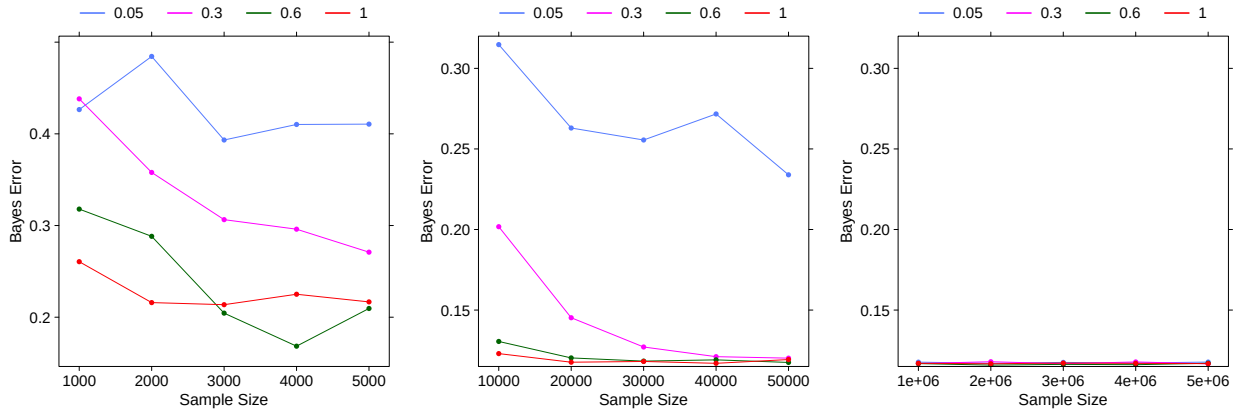

Figure 5: 5-dimensional QDA Bayes error. Left: small sample size; Middle: median sample size; Right: large sample size.

with the cumulative distribution function (CDF) as follows.

$$
\begin{aligned}
F_a(y|\mu) = \Phi\left(\frac{y-\mu}{\sigma/\sqrt{n}}\right) \\
+\frac{1}{2}\exp\{\frac{y-\mu}{\lambda}+\frac{\sigma^2}{2n\lambda^2}\}\Phi\left(-\frac{y-\mu}{\sigma/\sqrt{n}}-\frac{\sigma}{\lambda\sqrt{n}}\right) \\
-\frac{1}{2}\exp\{\frac{-y+\mu}{\lambda}+\frac{\sigma^2}{2n\lambda^2}\}\Phi\left(\frac{y-\mu}{\sigma/\sqrt{n}}-\frac{\sigma}{\lambda\sqrt{n}}\right),
\end{aligned}
$$

where $\Phi(\cdot)$ is the CDF of the unit Gaussian distribution.

We can easily derive the distribution of the test statistic under the null value and an alternative value by re-scaling $\bar{Y}^a$. However for the one sample z test the computation of the revised type I and type II errors can be done in a simpler fashion. Here and for the rest of this section we show the revised type I and type II errors for the two-sided alternative $H_a : \mu \neq \mu_0$. The results for the one-sided alternatives can be derived similarly.

Let α be the significance level of the test. Let $z_{\frac{\alpha}{2}}$ be the $(1-\frac{\alpha}{2})$ quantile of the unit Gaussian distribution (i.e., the upper quantile). α and β are the type I and type II errors for the standard test, without the added Laplace noise. For the test under differential privacy, we have the revised type I error, α^a, and type II error, β^a, as follows.

$$
\begin{aligned}
\alpha^a =& P\left(\left|\frac{\bar{Y}^a-\mu_0}{\sigma/\sqrt{n}}\right| > z_{\frac{\alpha}{2}} | H_0\right) \\
=& 1 - F_a\left(\mu_0 + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \mid \mu_0\right) + F_a\left(\mu_0 - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \mid \mu_0\right) \\
=& \alpha + \exp\{\frac{-z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}}+\frac{\sigma^2}{2n\lambda^2}\}\Phi\left(z_{\frac{\alpha}{2}}-\frac{\sigma}{\lambda\sqrt{n}}\right) \\
& - \exp\{\frac{z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}}+\frac{\sigma^2}{2n\lambda^2}\}\Phi\left(-z_{\frac{\alpha}{2}}-\frac{\sigma}{\lambda\sqrt{n}}\right)
\end{aligned}
$$

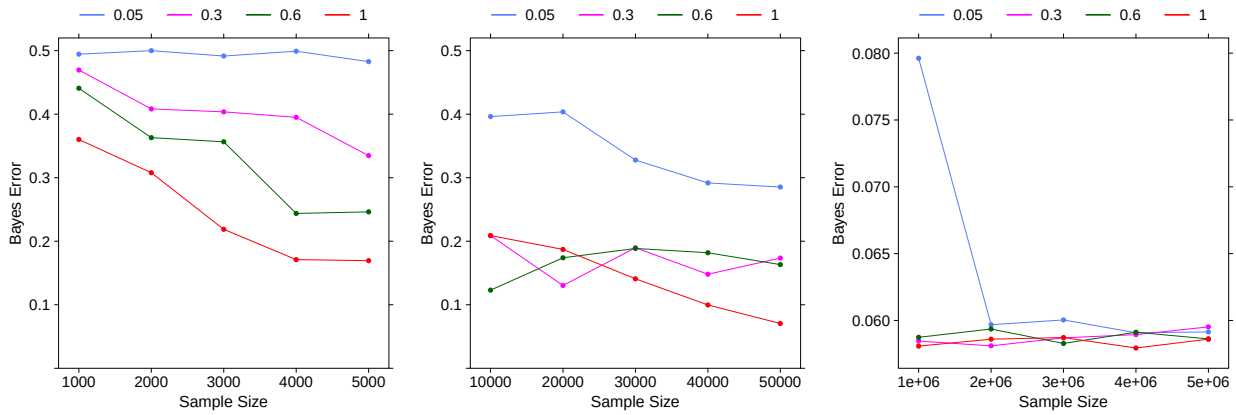

Figure 6: 10-dimensional QDA Bayes error. Left: small sample size; Middle: median sample size; Right: large sample size.

$$
\begin{aligned}
\beta^a &= P\left(\left|\frac{\bar{Y}^a-\mu_0}{\sigma/\sqrt{n}}\right| < z_{\frac{\alpha}{2}} | H_a\right)\\
&= F_a\left(\mu_0 + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \mid \mu_a\right) - F_a\left(\mu_0 - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \mid \mu_a\right)\\
&= \beta + \frac{1}{2}\exp\{\frac{-z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}} + \frac{\mu_0-\mu_a}{\lambda} + \frac{\sigma^2}{2n\lambda^2}\}\\
&\times \Phi\left(-z_{\frac{\alpha}{2}} + \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} + \frac{\sigma}{\lambda\sqrt{n}}\right)\\
&+ \frac{1}{2}\exp\{\frac{z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}} - \frac{\mu_0-\mu_a}{\lambda} + \frac{\sigma^2}{2n\lambda^2}\}\\
&\times \Phi\left(-z_{\frac{\alpha}{2}} + \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} - \frac{\sigma}{\lambda\sqrt{n}}\right)\\
&- \frac{1}{2}\exp\{\frac{\sigma^2}{2n\lambda^2} + \frac{\mu_0-\mu_a}{\lambda} - \frac{z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}}\}\\
&- \frac{1}{2}\exp\{\frac{z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}} + \frac{\mu_0-\mu_a}{\lambda} + \frac{\sigma^2}{2n\lambda^2}\}\\
&\times \Phi\left(z_{\frac{\alpha}{2}} + \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} + \frac{\sigma}{\lambda\sqrt{n}}\right)\\
&- \frac{1}{2}\exp\{\frac{-z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}} - \frac{\mu_0-\mu_a}{\lambda} + \frac{\sigma^2}{2n\lambda^2}\}\\
&\times \Phi\left(z_{\frac{\alpha}{2}} + \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} - \frac{\sigma}{\lambda\sqrt{n}}\right)\\
&+ \frac{1}{2}\exp\{\frac{\sigma^2}{2n\lambda^2} + \frac{\mu_0-\mu_a}{\lambda} + \frac{z_{\frac{\alpha}{2}}\sigma}{\lambda\sqrt{n}}\}
\end{aligned}
$$

5.2. One sample t test

Assume n samples $Y_1, Y_2, ..., Y_n$ i.i.d $\sim N(\mu, \sigma^2)$, where σ^2 is unknown. The null hypothesis is $H_0 : \mu = \mu_0$. The common alternative hypotheses are $H_a : \mu \neq \mu_0$, $H_a : \mu > \mu_0$, or $H_a : \mu < \mu_0$.

Suppose users query the sample mean and the sample variance. Then the test statistic involves two noise added sample statistics,

$$T^a = \frac{\bar{Y}^a - \mu_0}{S^a/\sqrt{n}},$$

where $Y^a = \bar{Y} + r_1$ with $r_1 \sim \text{Laplace}(\lambda_1)$, and $S^a = \sqrt{S^2 + r_2}$ with $r_2 \sim \text{Laplace}(\lambda_2)$.

To obtain the distribution of the test statistic under either the null value or an alternative value, we re-write the test statistic as

$$T^a = \frac{Z^a}{X^a},$$

where $Z^a = \frac{\bar{Y}^a-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$ and $X^a = \sqrt{(S^a)^2/\sigma^2}$. We obtain the distribution of the numerator Z^a by rescaling a Gaussian-Laplace mixture distribution. Similarly we obtain the distribution of the denominator X^a based on a Chi-Square-Laplace mixture distribution. Let $F_Z(z)$ be the CDF of Z^a and $f_X(x)$ be the PDF of X^a.

$$
\begin{aligned}
F_Z(z|\mu) &= \Phi(z-\delta)\\
&+ \frac{1}{2}\exp\{\frac{\sigma(z-\delta)}{\lambda_1\sqrt{n}} + \frac{\sigma^2}{2n\lambda_1^2}\}\Phi\left(-(z-\delta) - \frac{\sigma}{\lambda_1\sqrt{n}}\right)\\
&- \frac{1}{2}\exp\{-\frac{\sigma(z-\delta)}{\lambda_1\sqrt{n}} + \frac{\sigma^2}{2n\lambda_1^2}\}\Phi\left((z-\delta) - \frac{\sigma}{\lambda_1\sqrt{n}}\right),
\end{aligned}
$$

where $\delta = \frac{\mu_0-\mu}{\sigma/\sqrt{n}}$. δ equals to 0 under the null and does not equal to 0 under the alternative.

The distribution of X^a does not depend on the mean.

$$
\begin{aligned}
f_X(x) = [\ & 2x f_g(x^2|\frac{n-1}{2},\theta_0) \\
& + \frac{\sigma^2 x}{\lambda_2}\exp\{-\frac{\sigma^2x^2}{\lambda_2}\}\left(\frac{\theta_1}{\theta_0}\right)^{\frac{n-1}{2}} F_g(x^2|\frac{n-1}{2},\theta_1) \\
& - x\exp\{-\frac{\sigma^2x^2}{\lambda_2}\}\left(\frac{\theta_1}{\theta_0}\right)^{\frac{n-1}{2}} f_g(x^2|\frac{n-1}{2},\theta_1) \\
& + \frac{\sigma^2 x}{\lambda_2}\exp\{\frac{\sigma^2x^2}{\lambda_2}\}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{n-1}{2}} \left(1-F_g(x^2|\frac{n-1}{2},\theta_2)\right) \\
& - x\exp\{\frac{\sigma^2x^2}{\lambda_2}\}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{n-1}{2}} f_g(x^2|\frac{n-1}{2},\theta_2)\ ]/ \\
& [1-\frac{1}{2}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{n-1}{2}}]
\end{aligned}
$$

where $\theta_0 = \frac{2}{n-1}$, $\theta_1 = \frac{2}{n-1-2\sigma^2/\lambda_2}$, $\theta_2 = \frac{2}{n-1+2\sigma^2/\lambda_2}$, and F_g and f_g are the CDF and PDF of a gamma distribution respectively.

The distribution of the test statistic T^a given mean μ is

$$
F_T(t|\mu) = \begin{cases} \int_0^\infty F_Z(tx|\mu) f_X(x)\,\mathrm{d}x & t \geq 0 \\ \int_0^\infty (1-F_Z(tx|\mu))\, f_X(x)\,\mathrm{d}x & t < 0 \end{cases} \tag{4}
$$

Let $t_{\frac{\alpha}{2},n-1}$ be the $(1-\frac{\alpha}{2})$ quantile of a T distribution with $n-1$ degrees of freedom. The revised type I and type II errors can be computed numerically. Again we just show α^a and β^a under the two sided alternative. Similarly we can obtain the revised errors for the one sided alternatives.

$$
\begin{aligned}
\alpha^a =& P\left(|T^a| > t_{\frac{\alpha}{2},n-1}\middle|\mu=\mu_0\right) \\
=& 1 - F_T(t_{\frac{\alpha}{2},n-1}|\mu_0) + F_T(-t_{\frac{\alpha}{2},n-1}|\mu_0)
\end{aligned}
$$

$$
\begin{aligned}
\beta^a =& P\left(|T^a| < t_{\frac{\alpha}{2},n-1}\middle|\mu=\mu_a\right) \\
=& F_T\left(t_{\frac{\alpha}{2},n-1}|\mu_a\right) - F_T\left(-t_{\frac{\alpha}{2},n-1}|\mu_a\right)
\end{aligned}
$$

5.3. Two sample t test with equal variance

Assume n_1 samples $Y_1^1, Y_2^1, ..., Y_{n_1}^1$ i.i.d $\sim N(\mu_1,\sigma^2)$, n_2 samples $Y_1^2, Y_2^2, ..., Y_{n_2}^2$ i.i.d $\sim N(\mu_2,\sigma^2)$, where σ^2 is unknown. The null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$. The common alternative hypotheses are $H_a : \mu_1 - \mu_2 \neq 0$, $H_a : \mu_1 - \mu_2 > 0$, or $H_a : \mu_1 - \mu_2 < 0$.

Suppose users query the sample means and the sample variances. Then the test statistic involves multiple noise added sample statistics.

$$
T^a = \frac{\bar{Y}_1^a - \bar{Y}_2^a}{S^a\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},
$$

where $\bar{Y}_1^a = \bar{Y}_1 + r_1$, $\bar{Y}_2^a = \bar{Y}_2 + r_2$, and

$$
S^a = \sqrt{\frac{(n_1-1)(S_1^2+r_3)+(n_2-1)(S_2^2+r_4)}{n_1+n_2-2}},
$$

with $r_i \sim \text{Laplace}(\lambda_i)$, $i = 1 \sim 4$. We re-write the test statistic as

$$
T^a = \frac{Z^a}{X^a},
$$

where

$$
Z^a = \frac{\bar{Y}_1^a - \bar{Y}_2^a - (\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} + \frac{(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},
$$

and $X^a = S^a/\sigma$. Since the Laplace noises are added independently, we can then obtain the distribution of the numerator by convoluting Gaussian and Laplace distributions. The distribution of X^a is based on convolution of chi-square and Laplace distributions. The distributions of Z^a and X^a depend on the Laplace noise parameters λ_i, $i = 1 \sim 4$. Whether the sample sizes are equal or not no longer matters. We obtain their distributions under two seperate cases. Let $\upsilon = n_1 + n_2 - 2$. Let

$$
\delta = \frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.
$$

δ equals to 0 under H_0 and is non-zero under H_a.

Distribution of Z^a, $\lambda_1 \neq \lambda_2$: We have

$$
\begin{aligned}
F_Z(z|\mu_1-\mu_2) =& \Phi(z-\delta) \\
& - \frac{\lambda_2^2}{2(\lambda_1^2-\lambda_2^2)}\exp\{\tau_2(z-\delta)+\frac{\tau_2^2}{2}\}(1-\Phi(z-\delta+\tau_2)) \\
& + \frac{\lambda_2^2}{2(\lambda_1^2-\lambda_2^2)}\exp\{\frac{\tau_2^2}{2}-\tau_2(z-\delta)\}\Phi(z-\delta-\tau_2) \\
& + \frac{\lambda_1^2}{2(\lambda_1^2-\lambda_2^2)}\exp\{\frac{\tau_1^2}{2}+\tau_1(z-\delta)\}(1-\Phi(z-\delta+\tau_1)) \\
& - \frac{\lambda_1^2}{2(\lambda_1^2-\lambda_2^2)}\exp\{\frac{\tau_1^2}{2}-\tau_1(z-\delta)\}\Phi(z-\delta-\tau_1)
\end{aligned}
$$

where

$$
\tau_1 = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_1,
$$

and

$$
\tau_2 = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_2.
$$

Distribution of Z^a, $\lambda_1 = \lambda_2$: We have

$$
\begin{aligned}
&F_Z(z|\mu_1-\mu_2) = \Phi(z-\delta) \\
&-\left(\frac{1}{2}+\frac{\tau(z-\delta)}{4}-\frac{\tau^2}{4}\right)\exp\{\frac{\tau^2}{2}-\tau(z-\delta)\}\Phi(z-\delta-\tau) \\
&-\frac{\tau}{4\sqrt{2\pi}}\exp\{\frac{\tau^2}{2}-\tau(z-\delta)-\frac{(z-\delta-\tau)^2}{2}\} \\
&+\frac{\tau}{4\sqrt{2\pi}}\exp\{\frac{\tau^2}{2}+\tau(z-\delta)-\frac{(z-\delta+\tau)^2}{2}\} \\
&+\left(\frac{1}{2}-\frac{\tau(z-\delta)}{4}-\frac{\tau^2}{4}\right)\exp\{\frac{\tau^2}{2}+\tau(z-\delta)\}\left(1-\Phi(z-\delta+\tau)\right)
\end{aligned}
$$

where

$$\tau = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_1.$$

Distribution of X^a, $\lambda_3 \neq \lambda_4$: It does not depend on $\mu_1-\mu_2$. $\upsilon = n_1+n_2-2$. We have

$$
\begin{aligned}
&f_X(x) = [2xf_G(x^2;\frac{\upsilon}{2},\theta_0) \\
&+\frac{b_2^2}{b_2^2-b_1^2}\exp\{-b_1x^2\}(b_1x)\left(\frac{\theta_1}{\theta_0}\right)^{\frac{\upsilon}{2}}F_G(x^2;\frac{\upsilon}{2},\theta_1) \\
&-\frac{b_2^2}{b_2^2-b_1^2}\exp\{-b_1x^2\}(x)\left(\frac{\theta_1}{\theta_0}\right)^{\frac{\upsilon}{2}}f_G(x^2;\frac{\upsilon}{2},\theta_1) \\
&-\frac{b_1^2}{b_2^2-b_1^2}\exp\{-b_2x^2\}(b_2x)\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon}{2}}F_G(x^2;\frac{\upsilon}{2},\theta_2) \\
&+\frac{b_1^2}{b_2^2-b_1^2}\exp\{-b_2x^2\}(x)\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon}{2}}f_G(x^2;\frac{\upsilon}{2},\theta_2) \\
&+\frac{b_2^2}{b_2^2-b_1^2}\exp\{b_1x^2\}(b_1x)\left(\frac{\theta_3}{\theta_0}\right)^{\frac{\upsilon}{2}}\left(1-F_G(x^2;\frac{\upsilon}{2},\theta_3)\right) \\
&-\frac{b_2^2}{b_2^2-b_1^2}\exp\{b_1x^2\}(x)\left(\frac{\theta_3}{\theta_0}\right)^{\frac{\upsilon}{2}}f_G(x^2;\frac{\upsilon}{2},\theta_3) \\
&-\frac{b_1^2}{b_2^2-b_1^2}\exp\{b_2x^2\}(b_2x)\left(\frac{\theta_4}{\theta_0}\right)^{\frac{\upsilon}{2}}\left(1-F_G(x^2;\frac{\upsilon}{2},\theta_4)\right) \\
&+\frac{b_1^2}{b_2^2-b_1^2}\exp\{b_2x^2\}(x)\left(\frac{\theta_4}{\theta_0}\right)^{\frac{\upsilon}{2}}f_G(x^2;\frac{\upsilon}{2},\theta_4)]/ \\
&[1-\frac{b_2^2}{2(b_2^2-b_1^2)}\left(\frac{\theta_3}{\theta_0}\right)^{\frac{\upsilon}{2}}+\frac{b_1^2}{2(b_2^2-b_1^2)}\left(\frac{\theta_4}{\theta_0}\right)^{\frac{\upsilon}{2}}]
\end{aligned}
$$

where

$$
\begin{aligned}
\tau_1 &= \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_1, \\
\tau_2 &= \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_2, \\
b_1 &= \frac{(n_1+n_2-2)\sigma^2}{(n_1-1)\lambda_3}, \\
b_2 &= \frac{(n_1+n_2-2)\sigma^2}{(n_2-1)\lambda_4}, \\
\theta_0 &= \frac{2}{n_1+n_2-2}, \\
\theta_1 &= \frac{2}{n_1+n_2-2-2b_1}, \\
\theta_2 &= \frac{2}{n_1+n_2-2-2b_2}, \\
\theta_3 &= \frac{2}{n_1+n_2-2+2b_1},
\end{aligned}
$$

and

$$\theta_4 = \frac{2}{n_1+n_2-2+2b_2}.$$

Distribution of X^a, $\lambda_3 = \lambda_4$: Again, it does not depend on $\mu_1-\mu_2$. We have

$$
\begin{aligned}
&f_X(x) = [2xf_G(x^2;\frac{\upsilon}{2},\theta_0) \\
&+\left(\frac{b^2x^3+bx}{2}\right)\exp\{-bx^2\}\left(\frac{\theta_1}{\theta_0}\right)^{\frac{\upsilon}{2}}F_G(x^2;\frac{\upsilon}{2},\theta_1) \\
&-\left(\frac{2x+bx^3}{2}\right)\exp\{-bx^2\}\left(\frac{\theta_1}{\theta_0}\right)^{\frac{\upsilon}{2}}f_G(x^2;\frac{\upsilon}{2},\theta_1) \\
&-\left(\frac{b^2x}{2}\right)\exp\{-bx^2\}\left(\frac{\theta_1}{\theta_0}\right)^{\frac{\upsilon+2}{2}}F_G(x^2;\frac{\upsilon+2}{2},\theta_1) \\
&+\left(\frac{bx}{2}\right)\exp\{-bx^2\}\left(\frac{\theta_1}{\theta_0}\right)^{\frac{\upsilon+2}{2}}f_G(x^2;\frac{\upsilon+2}{2},\theta_1) \\
&+\left(\frac{bx-b^2x^3}{2}\right)\exp\{bx^2\}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon}{2}}\left(1-F_G(x^2;\frac{\upsilon}{2},\theta_2)\right) \\
&-\left(\frac{2x-bx^3}{2}\right)\exp\{bx^2\}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon}{2}}f_G(x^2;\frac{\upsilon}{2},\theta_2) \\
&+\left(\frac{b^2x}{2}\right)\exp\{bx^2\}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon+2}{2}}\left(1-F_G(x^2;\frac{\upsilon+2}{2},\theta_2)\right) \\
&-\left(\frac{bx}{2}\right)\exp\{bx^2\}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon+2}{2}}f_G(x^2;\frac{\upsilon+2}{2},\theta_2)]/ \\
&\left[1-\frac{1}{2}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon}{2}}-\frac{b}{4}\left(\frac{\theta_2}{\theta_0}\right)^{\frac{\upsilon+2}{2}}\right]
\end{aligned}
$$

where

$$
\begin{aligned}
b &= 2\sigma^2/\lambda_3, \\
\theta_0 &= \frac{2}{n_1+n_2-2}, \\
\theta_1 &= \frac{2}{n_1+n_2-2-b},
\end{aligned}
$$

and

$$\theta_2 = \frac{2}{n_1+n_2-2+b}.$$

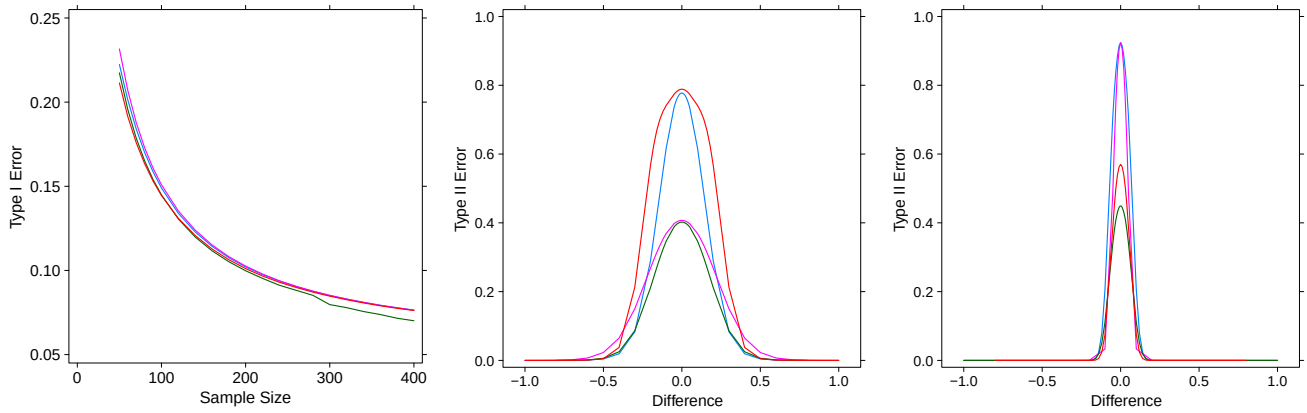

Figure 7: $\varepsilon = 0.3$. We set λ_i equal to $1/(n_i\varepsilon)$ and fix $\alpha = 0.05$. Red line is for one sample Z test; blue line is for one sample T test; pink line is for two sample T test with equal sample size and equal variance; green line is for two sample T test with unequal sample size and equal variance. Left: Type I error plotted against sample size; Middle: n=50, Type II error plotted against the difference between the true value and the hypothesized value; Right: n=400, Type II error plotted against the difference between the true value and the hypothesized value.

Given the Laplace noise parameters λ_i, we select the CDF and PDF of Z^a and X^a respectively. The distribution of the test statistic T^a given the value of $\mu_1 - \mu_2$ follows Equation 4. Let $t_{\frac{\alpha}{2},\upsilon}$ be the $(1-\frac{\alpha}{2})$ quantile of a t distribution with υ degrees of freedom. The revised type I and type II errors again can be computed numerically. We show α^a and β^a under the two sided alternative. Similarly we can obtain the revised errors for the one sided alternatives.

$$\begin{aligned}\alpha^a &= P\left(|T^a| > t_{\frac{\alpha}{2},\upsilon} \mid \mu_1 - \mu_2 = 0\right) \\ &= 1 - F_T(t_{\frac{\alpha}{2},\upsilon} \mid \mu_1 - \mu_2 = 0) + F_T(-t_{\frac{\alpha}{2},\upsilon} \mid \mu_1 - \mu_2 = 0)\end{aligned}$$

$$\begin{aligned}\beta^a =& P\left(|T^a| < t_{\frac{\alpha}{2},\upsilon} \mid \mu_1 - \mu_2 \neq 0\right) \\ &= F_T(t_{\frac{\alpha}{2},\upsilon} \mid \mu_1 - \mu_2 \neq 0) - F_T(-t_{\frac{\alpha}{2},\upsilon} \mid \mu_1 - \mu_2 \neq 0)\end{aligned}$$

Figure 7 show the type I and type II errors under differential privacy. Hypothesis tests often operate with far less samples than classification, since the test is always significant for large dataset. On the left panel of Figure 7, we plot the type I error under differential privacy against sample size, up to 400 samples. For the tests considered in this article, the type I errors under differential privacy decrease sharply as sample size increases. Type II error depends on the difference between the true value and the hypothesized value. The left panel of Figure 7 shows the type II error under differential privacy for sample size 50. The right panel of Figure 7 shows the type II error under differential privacy for sample size 400. Again the type II error under differential privacy improves significantly as sample size increases. We must apply differential privacy query results with caution in hypothesis tests. Often users have only a handful or a few dozen samples in a test, the direct noise addition makes the test result unreliable. With very small datasets, users need the clean query results or direct access to the raw data for a reliable output.

6. SUMMARY

In this article we calculate the sensitivities of various statistics queried from a statistical database. We examine the performance of Bayesian classifier using the noise added mean and variance-covariance matrix. We also study the type I and type II errors under differential privacy for various hypothesis tests. In the process we identify an interesting issue associated with random noise addition: The variance-covariance matrix without the added noise is positive definite. However simply adding noise can only return a symmetric matrix, which is no longer positive definite. Consequently the query result cannot be used to construct a classifier. We implement a heuristic algorithm to repair the noise added matrix to achieve positive definiteness in the experiments.

This is a general issue for random noise addition. Adding noise to a statistic which must satisfy certain constraint may return query results that no longer satisfy the constraint. The query results need to be further modified in order to be used in subsequent studies. An interesting question is how to provide query results that are helpful for subsequent studies while safely protecting database participant's privacy. Each constrained statistic may need an algorithm to achieve its original properties after noise addition.

APPENDIX A. ONE DIMENSIONAL BAYES ERROR BOUND

We can obtain an upper bound for the one dimensional Bayes error with Gaussian mixture models under differential privacy for binary classes. Assume class $\omega_1 \sim N(\mu_1, \sigma_1^2)$ and class $\omega_2 \sim N(\mu_2, \sigma_2^2)$. Further assume class ω_1 has n_1 records and class ω_2 has n_2 records. First note the *Bhattacharyya* bound [2] states that

$$\text{Bayes Error} \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-K}, \tag{5}$$

where

$$K = \frac{1}{4}\frac{\mu_1^2+\mu_2^2-2\mu_1\mu_2}{\sigma_1^2+\sigma_2^2} + \frac{\log(\sigma_1^2+\sigma_2^2)}{2} - \frac{\log(4\sigma_1^2\sigma_2^2)}{4}. \tag{6}$$

Considering the Laplace noises added to the queries of mean and variances in each class, we have the following theorem.

Theorem 6.1 *The Gaussian mixture models are as specified above. Assume under differential privacy the query responses are the sample means and the sample variances plus independent Laplace noises:*

$$\hat{\mu}_1 = \bar{x}_1 + r_1,\ \hat{\mu}_2 = \bar{x}_2 + r_2,\ \hat{\sigma}_1^2 = S_1^2 + r_3,\ \hat{\sigma}_2^2 = S_2^2 + r_4.$$

Since there are multiple ways to query a statistic, we simply assume the independent Laplace noises $r_i \sim L(0,\lambda_i)$ for a general result. We have for $0 < p < 1$,

$$P(K^L(p) < K < K^U(p)) = p^8,$$

and

$$Pr(\text{Bayes Error} < \sqrt{P(\omega_1)P(\omega_2)}e^{-K^L(p)}) \geq p^8,$$

where

$$\begin{aligned} K^U(p) \ =& \\ & \frac{\sum_{i=1}^2\{\mu_i + \sqrt{\frac{\sigma_i^2}{n_i}}Z(1-\frac{p}{2}) - \lambda_i\log(1-2|\frac{1-p}{2}|)\}^2}{4\{\sum_{i=1}^2\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(\frac{p}{2}) + \sum_{i=3}^4\lambda_i\log(1-2|\frac{1-p}{2}|)\}} \\ & - \frac{\prod_{i=1}^2\{\mu_i - \sqrt{\frac{\sigma_i^2}{n_i}}Z(1-\frac{p}{2}) + \lambda_i\log(1-2|\frac{1-p}{2}|)\}^2}{2\{\sum_{i=1}^2\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(1-\frac{p}{2}) - \sum_{i=3}^4\lambda_i\log(1-2|\frac{1-p}{2}|)\}} \\ & + \frac{\log\{\sum_{i=1}^2\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(1-\frac{p}{2}) - \sum_{i=3}^4\lambda_i\log(1-2|\frac{1-p}{2}|)\}}{2} \\ & - \frac{\log\{4\prod_{i=1}^2[\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(\frac{p}{2}) + \lambda_{i+2}\log(1-2|\frac{1-p}{2}|)]\}}{4}, \end{aligned}$$

and

$$\begin{aligned} K^L(p) \ =& \\ & \frac{\sum_{i=1}^2\{\mu_i - \sqrt{\frac{\sigma_i^2}{n_i}}Z(1-\frac{p}{2}) + \lambda_i\log(1-2|\frac{1-p}{2}|)\}^2}{4\{\sum_{i=1}^2\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(1-\frac{p}{2}) - \sum_{i=3}^4\lambda_i\log(1-2|\frac{1-p}{2}|)\}} \\ & - \frac{\prod_{i=1}^2\{\mu_i + \sqrt{\frac{\sigma_i^2}{n_i}}Z(1-\frac{p}{2}) - \lambda_i\log(1-2|\frac{1-p}{2}|)\}^2}{2\{\sum_{i=1}^2\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(\frac{p}{2}) + \sum_{i=3}^4\lambda_i\log(1-2|\frac{1-p}{2}|)\}} \\ & + \frac{\log\{\sum_{i=1}^2\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(\frac{p}{2}) + \sum_{i=3}^4\lambda_i\log(1-2|\frac{1-p}{2}|)\}}{2} \\ & - \frac{\log\{4\prod_{i=1}^2[\frac{\sigma_i^2}{n_i}\chi^2_{n_i-1}(1-\frac{p}{2}) - \lambda_{i+2}\log(1-2|\frac{1-p}{2}|)]\}}{4}. \end{aligned}$$

$Z(r)$ is the r quantile of the standard normal distribution. $\chi^2_{n-1}(r)$ is the r quantile of χ^2_{n-1}. $\lambda\log(1-2|\frac{1-p}{2}|)$ and $-\lambda\log(1-2|\frac{1-p}{2}|)$ are $p/2$ and $(1-p/2)$ quantile of Laplace distribution $L(0,\lambda)$.

Proof: Since both classes follow Gaussian distribution, we have the following distribution for the sample means and the sample variances:

$$\begin{aligned} \bar{x}_1 &\sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \\ \bar{x}_2 &\sim N(\mu_2, \frac{\sigma_2^2}{n_2}), \\ \frac{n_1S_1^2}{\sigma_1^2} = \frac{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2}{\sigma_1^2} &\sim \chi^2_{n_1-1}, \\ \frac{n_2S_2^2}{\sigma_2^2} = \frac{\sum_{i=1}^{n_2}(x_{2,i}-\bar{x}_2)^2}{\sigma_2^2} &\sim \chi^2_{n_2-1}. \end{aligned}$$

Note the sample means and the sample variances are independent. Also note we add independent Laplace noises $r_i \sim L(0,\lambda_i)$,

$$\hat{\mu}_1 = \bar{x}_1 + r_1,\ \hat{\mu}_2 = \bar{x}_2 + r_2,\ \hat{\sigma}_1^2 = S_1^2 + r_3,\ \hat{\sigma}_2^2 = S_2^2 + r_4.$$

With probability p (for example $p = 0.90, 0.95$, etc.), we have:

$$\mu_i - \sqrt{\frac{\sigma_i^2}{n_i}} \times Z(1-\frac{p}{2}) < \bar{x}_i < \mu_i + \sqrt{\frac{\sigma_i^2}{n_i}} \times Z(1-\frac{p}{2}),\ i = 1,2,$$

$$\frac{\sigma_i^2}{n_i} \times \chi^2_{n_i-1}(\frac{p}{2}) < S_i^2 < \frac{\sigma_i^2}{n_i} \times \chi^2_{n_i-1}(1-\frac{p}{2}),\ i = 1,2,$$

$$\lambda_i\log(1-2|\frac{1-p}{2}|) < r_i < -\lambda_i\log(1-2|\frac{1-p}{2}|),\ i = 1-4,$$

where $Z(1-p/2)$ is the $(1-p/2)$ quantile of the standard normal distribution, $\chi^2_{n_i-1}(r)$ is the r quantile of $\chi^2_{n_i-1}$

($r = p/2$ *or* $1-p/2$), and $\lambda_i \log(1-2|\frac{1-p}{2}|)$ and $-\lambda_i \log(1-2|\frac{1-p}{2}|)$ are $p/2$ and $(1-p/2)$ quantile of Laplace distribution $L(0,\lambda_i)$.

In Equation 6, plugging in the bounds of the sample means, the sample variances, and the Laplace noises, we have:

$$Pr(K^L(p) < K < K^U(p)) = p^8,$$

where $K^L(p)$ and $K^U(p)$ are specified in the main theorem. Because $Pr(K^L(p) < K) \geq p^8$, we have

$$Pr(\text{Bayes Error} < \sqrt{P(\omega_1)P(\omega_2)}e^{-K^L(p)}) \geq p^8.$$

∎

The proof is based on genuine Gaussian distributions. Bhattacharyya bound can be applied to truncated Gaussian distribution [2] and we can obtain useful information if k is sufficiently large.

APPENDIX B. NEW MATERIAL SINCE THE CONFERENCE PAPER

The 2011 conference paper has a crude algorithm to fix the noise added variance-covariance matrix in Section 4, with preliminary experiments. In this expanded journal article, we implement a new algorithm to fix the noise added variance-covariance matrix, which is much more effective. We then carry on extensive experiments using both simulated and real life data.

Section 5, hypothesis testing under differential privacy, is new material since the 2011 conference paper. We have revised abstract, introduction, related work, and summary accordingly. Only Section 3 is the same as in the 2011 conference paper.

Acknowledgments: This work was supported in part by National Science Foundation DMS-1228348, Air Force Office of Scientific Research FA9550-12-1-0082, National Institutes of Health Grants 1R0-1LM009989 and 1R01HG006844, National Science Foundation Grants Career-CNS-0845803, CNS-1111529, CNS-1228198 and Army Research Office Grant W911NF-12-1-0558.

References

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, 2000.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2. edition, 2001.

[3] C. Dwork. Differential privacy. In *ICALP (2)*, pages 1–12. Springer, 2006.

[4] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer Berlin / Heidelberg, 2008.

[5] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[6] A. Friedman and A. Schuster. Data mining with differential privacy. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502, New York, NY, USA, 2010. ACM.

[7] K. Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[8] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright. A practical differentially private random decision tree classifier. In *ICDM Workshops*, pages 114–121, 2009.

[9] M. Kantarcioğlu, J. Jin, and C. Clifton. When do data mining results violate privacy? In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 599–604, New York, NY, USA, 2004. ACM.

[10] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, page 99, Washington, DC, USA, 2003. IEEE Computer Society.

[11] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE '07*, pages 106–115, Istanbul, Turkey, 2007. IEEE.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE '06*, page 24, Atlanta, GA, USA, 2006. IEEE Computer Society.

[13] S. Merugu and J. Ghosh. Privacy-preserving distributed clustering using generative models. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, page 211, Washington, DC, USA, 2003. IEEE Computer Society.

[14] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[15] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In *VLDB '06: Proceedings of the*

32nd international conference on Very large data bases, pages 139–150, Seoul, Korea, 2006. VLDB Endowment.

[16] X. Xiao and Y. Tao. Output perturbation with query relaxation. *PVLDB*, 1(1):857–869, 2008.