Gaussian Mixture Models for Classification and Hypothesis Tests Under Differential Privacy

Xiaosu Tong, Bowei Xi, Murat Kantarcioglu, Ali Inan

Gaussian Mixture Models for Classification and Hypothesis Tests under Differential Privacy

Xiaosu Tong[1], Bowei Xi[2⋆⋆], Murat Kantarcioglu[3], and Ali Inan[4]

[1] Amazon, Seattle WA
xiaosutong@gmail.com

[2] Department of Statistics, Purdue University,
xbw@purdue.edu

[3] Department of Computer Science, University of Texas at Dallas,
muratk@utdallas.edu

[4] Department of Computer Engineering,
Adana Science and Technology University, Adana, Turkey
ainan@adanabtu.edu.tr

Abstract. Many statistical models are constructed using very basic statistics: mean vectors, variances, and covariances. Gaussian mixture models are such models. When a data set contains sensitive information and cannot be directly released to users, such models can be easily constructed based on noise added query responses. The models nonetheless provide preliminary results to users. Although the queried basic statistics meet the differential privacy guarantee, the complex models constructed using these statistics may not meet the differential privacy guarantee. However it is up to the users to decide how to query a database and how to further utilize the queried results. In this article, our goal is to understand the impact of differential privacy mechanism on Gaussian mixture models. Our approach involves querying basic statistics from a database under differential privacy protection, and using the noise added responses to build classifier and perform hypothesis tests. We discover that adding Laplace noises may have a non-negligible effect on model outputs. 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 classification error using the noise added responses, through experiments with both simulated data and real life data, and demonstrate under which conditions the impact of the added noises can be reduced. We compute the exact 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. We then show under which condition a hypothesis test returns reliable result given differentially private means, variances and covariances.

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

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

1 Introduction

Building a model over a data set is often a straightforward task. However, when the data set contains sensitive information, special care has to be taken. Instead of having direct access to data, the users are provided with a sanitized view of the database containing private information, either through perturbed individual records or perturbed query responses.

From users' perspective, knowing the responses to their queries are perturbed, users may not want to directly query the output of a complex model. Many statistical models are constructed using very basic statistics. Knowing the values of means, variances and covariances, or equivalently the sums, the sums of squares and the sums of cross products, users can build least square regression models, conduct principal component analysis, construct hypothesis tests, and construct Bayesian classifiers under Gaussian mixture models, etc. Although the basic statistics (e.g., means, variances and covariances) satisfy differential privacy guarantee, the complex models constructed using these basic statistics may no longer meet the differential privacy guarantee.

We notice it is up to the users to decide how to query a database and how to further utilize the queried results. Building statistical models using the perturbed basic statistics provides quick initial estimates. If the results based on the perturbed query responses are promising, users can then proceed to improve the accuracy of the results.

In this article, our goal is to understand the impact of differential privacy mechanism for the mixture of Gaussian models. Gaussian mixture models refer to the case where each model follows multivariate Gaussian distribution. Hence users only need to obtain the mean vector and the variance-covariance matrix for each class. Out of all the statistical techniques that can be applied to Gaussian mixture models without further querying the database, we focus on building a classifier or performing a hypothesis test with the noisy responses. Through extensive experiments and theoretical discussions, we show when the classifiers and tests work reliably under privacy protection mechanism, in particular, differential privacy.

k-anonymity [24, 25, 29] and differential privacy [6] are two major privacy preserving models. Under *k*-anonymity model the perturbed individual records are released to the users, while under differential privacy model the perturbed query responses are released to the users. Recent work pointed out the two privacy preserving models are complimentary [4]. Main contributions of this article could be summarized as follows:

1. We provide theoretical results on the type I and type II errors under differential privacy for several hypothesis tests. We also show when a hypothesis test returns reliable result under differential privacy mechanism.
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. We examine the classification error for the multivariate Gaussian case through experiments. The experiments demonstrate when the impact of the added noise can be reduced.

The rest of the paper is organized as follows. Section 1.1 provides a brief overview of differential privacy mechanism. Related work is discussed in Section 2. Section 3

provides theoretical results for hypothesis tests under differential privacy. In Section 4 we provide an algorithm to repair the noise added variance-covariance matrix, and study the classification error through extensive experiments. Section 5 concludes our discussion.

1.1 Differential Privacy

Let $D = \{X_1, \cdots, X_d\}$ be a d-dimensional database with n observations, where the domain of each attribute X_i is continuous and bounded. We are interested in building Gaussian mixture models over database D. One only needs to compute the expected values of each attribute X_i and the variance-covariance matrix, $\Sigma_{ij} = cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$, where $\mu_i = E(X_i)$. More details follow in Section 4. Users obtain the values of μ_is and Σ_{ij}s by querying the database D. The query results are perturbed according to differential privacy mechanism. Next we briefly review differential privacy.

Given a set of queries $Q = \{Q_1, ..., Q_q\}$, Laplace mechanism for differential privacy adds Laplace noise with parameter λ to the actual value. λ is determined by privacy parameter ε and sensitivity $S(Q)$. Here, ε is a pre-determined parameter, selected by the database curator, while sensitivity $S(Q)$ is a function of the query Q. Hence differential privacy mechanism minimizes the risk of identifying individual records from a database containing sensitive information.

Sensitivity is defined over sibling databases, which differ in only one observation.

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

That is, sensitivity of Q is the maximum difference in the L_1 norm of the query response caused by a single record update. Sensitivities for standard queries, such as sum, mean, variance-covariance are well established [8].

Once ε and $S(Q)$ are known, λ is set such that $\lambda \geq S(Q)/\varepsilon$. Then for each query Q, the database first computes the actual value Q^D in D, then adds Laplace noise to obtain the noisy response R^D, and return R^D to users: $R^D = Q^D + r$, where $r \sim \text{Laplace}(\lambda)$. There have been many work on sensitivity analysis. For querying mean, variance and covariance, we use the sensitivity results as in [31] in this article. Later in the experimental studies, the Laplace noises are added according to the results in [31]. Although there are other techniques to satisfy differential privacy (e.g., exponential mechanism [22]), for the three basic queries needed to build Gaussian mixture models, we leverage the Laplace mechanism discussed above.

2 Related Work

Gaussian mixture models are widely used in practice [5, 11]. Differential privacy mechanism [6] models the database as a statistical database. Random noises are added to the responses to user queries. The magnitude of random noise is proportional to the privacy parameter ε and the sensitivity of the query set. Different formulations of differential privacy have been proposed. One definition of sensitivity consider sibling

data sets that have the same size but differ in one record [6, 9]. Other studies have sibling data sets through insertion of a new record sets [8]. We follow the formulation in [6] in this article.

Classification under differential privacy has received some attention. In [10], Friedman et al. built a decision tree, a method of ID3 classification, through recursive queries retrieving the information gain. Jagannathan et al. [12] built multiple random decision trees using sum queries. [2] proposed perturbing the objective function before optimization for empirical risk minimization. The lower bounds of the noisy versions of convex optimization algorithms were studied. Privacy preserving optimization is an important component in some classifiers, such as regularized logistic regression and support vector machine (SVM). [16] extended the results in [2], and also proposed differentially private optimization algorithms for convex empirical risk minimization. [28] proposed a privacy preserving mechanism for SVM.

In [20] every component in a mixture population follows a Gaussian mixture distribution. A perturbation matrix was generated based on a gamma distribution. Gamma perturbations were included in the objective function as multipliers, and a classifier was learned through maximizing the perturbed objective function. On the other hand, we consider classifiers that can be constructed using very basic statistics, i.e., means, variances and covariances, and show how their performance is affected by the added noises. In this article, we present Bayes classifiers based on Gaussian mixture models by querying the mean vector and the variance-covariance matrix for each class.

[3] proposed an algorithm using a Markov Chain Monte Carlo procedure to produce principal components that satisfy differential privacy. It is a differentially private lower rank approximation to a semi-positive definite matrix. Typically the rank k is much smaller than the dimension d. [14] also proposed an algorithm to produce differentially private low rank approximation to a positive definite matrix. [21] focused on producing recommendations under differential privacy. In [21], the true ratings were perturbed. A variance-covariance matrix was computed using the perturbed ratings; noises were added to the resulting matrix; then a low rank approximation to the noise added matrix was computed. Compared to the existing work, we focus on the scenario where all the variables are used to learn a variance-covariance matrix and the subsequent classifier and dimension reduction is not needed.

[17] proposed the differentially private M-estimators, such as sample quantiles, maximum likelihood estimator, based on the perturbed histogram. Our work has a different focus. We examine the classifiers and hypothesis tests constructed using the differentially private sample means, variances and covariances. [30] derived rules for how to adjust sample sizes to achieve a pre-specified power for Pearson's χ^2 test of independence and the test of sample proportion. For the second test, when sample size is reasonably large, the sample proportion is approximately normally distributed. [30] developed sample size adjustment results based on the approximate normal distribution. Our work provides theoretical results to compute the exact type I and type II errors for one sample z test, one sample t test, and two sample t test. Both type I and type II errors are functions of ε and n. Hence with a known ε value users can obtain a minimum sample size required to achieve a pre-specified power while the exact type I error is controlled by a certain upper bound.

3 Hypothesis Tests 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 than classification. 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, users compute a test statistic using the noise added mean and variance, and make a decision. The exact type I and type II errors can be computed numerically for likely ε values, which provide a reliability check of the test for users. Here we show for what sample size the exact type I and type II errors are close to those without the added noises. 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. To construct a rejection region and compute the exact type I and type II errors merits more effort in this case. It is part of our future work.

3.1 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 $r \sim$ Laplace(λ). 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, GL$(\mu, \sigma^2, n, \lambda)$. It has the cumulative distribution function (CDF) $F_a(y|\mu)$ as follows.

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

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 exact type I and type II errors can be done in a simpler fashion. Here and for the rest of this section we show the exact 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 exact 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 + e^{\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) - e^{\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), \\
\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{\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).
\end{aligned}
$$

3.2 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

$$T^a = \frac{Z^a}{X^a}, \text{ where } Z^a = \frac{\bar{Y}^a - \mu}{\sigma/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \text{ and } X^a = \sqrt{(S^a)^2/\sigma^2}.$$

We obtain the distribution of Z^a by rescaling a Gaussian-Laplace mixture distribution. Similarly we obtain the distribution of 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.

$$
\begin{aligned}
F_Z(z|\mu) = \Phi(z-\delta) &+ \frac{1}{2}\exp\{\frac{\sigma(z-\delta)}{\lambda_1\sqrt{n}} + \frac{\sigma^2}{2n\lambda_1^2}\} \times \Phi\left(-(z-\delta) - \frac{\sigma}{\lambda_1\sqrt{n}}\right) \\
&- \frac{1}{2}\exp\{-\frac{\sigma(z-\delta)}{\lambda_1\sqrt{n}} + \frac{\sigma^2}{2n\lambda_1^2}\} \times \Phi\left((z-\delta) - \frac{\sigma}{\lambda_1\sqrt{n}}\right),
\end{aligned} \tag{3}
$$

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.

$$\begin{aligned} f_X(x) = [\, & 2x f_g(x^2|\frac{n-1}{2},\theta_0) + \frac{\sigma^2 x}{\lambda_2} e^{-\frac{\sigma^2 x^2}{\lambda_2}} (\frac{\theta_1}{\theta_0})^{\frac{n-1}{2}} F_g(x^2|\frac{n-1}{2},\theta_1) \\ & - x e^{-\frac{\sigma^2 x^2}{\lambda_2}} (\frac{\theta_1}{\theta_0})^{\frac{n-1}{2}} f_g(x^2|\frac{n-1}{2},\theta_1) + \frac{\sigma^2 x}{\lambda_2} e^{\frac{\sigma^2 x^2}{\lambda_2}} (\frac{\theta_2}{\theta_0})^{\frac{n-1}{2}} (1 - F_g(x^2|\frac{n-1}{2},\theta_2)) \\ & - x e^{\frac{\sigma^2 x^2}{\lambda_2}} (\frac{\theta_2}{\theta_0})^{\frac{n-1}{2}} f_g(x^2|\frac{n-1}{2},\theta_2)\,] \,/\, [1 - \frac{1}{2}(\frac{\theta_2}{\theta_0})^{\frac{n-1}{2}}] \end{aligned} \tag{4}$$

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

$$F_T(t|\mu) = \begin{cases} \int_0^\infty F_Z(tx|\mu) f_X(x)\, \mathrm{d}x & t \geq 0 \\ \int_0^\infty (1 - F_Z(tx|\mu))\, f_X(x)\, \mathrm{d}x & t < 0 \end{cases} \tag{5}$$

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 exact 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.

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

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

3.3 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 $r_i \sim$ Laplace(λ_i), $i = 1 \sim 4$. We re-write the test statistic as

$$T^a = \frac{Z^a}{X^a}, \text{ where } Z^a = \frac{\bar{Y}_1^a - \bar{Y}_2^a - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ and } X^a = \frac{S^a}{\sigma}.$$

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, $i = 1 \sim 4$. We obtain their distributions under two separate 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 the CDF

$$
\begin{aligned}
&F_Z(z|\mu_1-\mu_2) = \Phi(z-\delta) - \frac{\lambda_2^2}{2(\lambda_1^2-\lambda_2^2)} e^{\tau_2(z-\delta)+\frac{\tau_2^2}{2}} (1-\Phi(z-\delta+\tau_2)) \\
&+ \frac{\lambda_2^2}{2(\lambda_1^2-\lambda_2^2)} e^{\frac{\tau_2^2}{2}-\tau_2(z-\delta)} \Phi(z-\delta-\tau_2) + \frac{\lambda_1^2}{2(\lambda_1^2-\lambda_2^2)} e^{\frac{\tau_1^2}{2}+\tau_1(z-\delta)} (1-\Phi(z-\delta+\tau_1)) \\
&- \frac{\lambda_1^2}{2(\lambda_1^2-\lambda_2^2)} e^{\frac{\tau_1^2}{2}-\tau_1(z-\delta)} \Phi(z-\delta-\tau_1)
\end{aligned}
\tag{6}
$$

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 the CDF

$$
\begin{aligned}
&F_Z(z|\mu_1-\mu_2) = \Phi(z-\delta) - \left(\frac{1}{2} + \frac{\tau(z-\delta)}{4} - \frac{\tau^2}{4}\right) e^{\frac{\tau^2}{2}-\tau(z-\delta)} \Phi(z-\delta-\tau) \\
&- \frac{\tau}{4\sqrt{2\pi}} e^{\frac{\tau^2}{2}-\tau(z-\delta)-\frac{(z-\delta-\tau)^2}{2}} + \frac{\tau}{4\sqrt{2\pi}} e^{\frac{\tau^2}{2}+\tau(z-\delta)-\frac{(z-\delta+\tau)^2}{2}} \\
&+ (\frac{1}{2} - \frac{\tau(z-\delta)}{4} - \frac{\tau^2}{4}) e^{\frac{\tau^2}{2}+\tau(z-\delta)} (1-\Phi(z-\delta+\tau))
\end{aligned}
\tag{7}
$$

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 $\mu_1 - \mu_2$. Note $\upsilon = n_1 + n_2 - 2$. We have the PDF

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

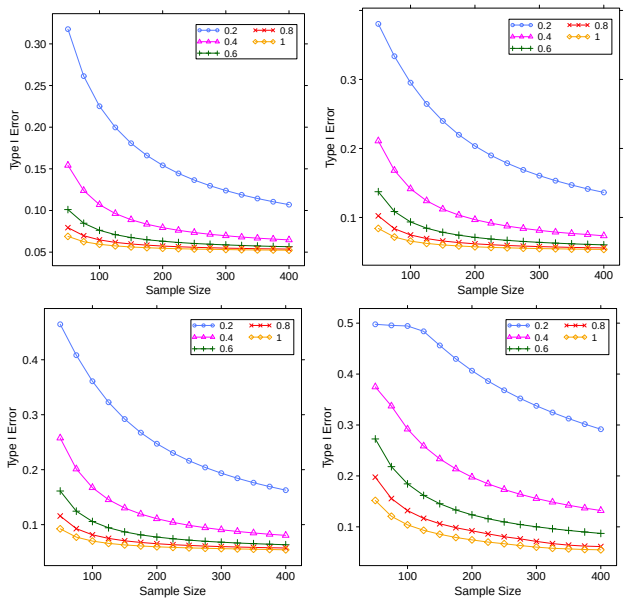

Fig. 1. Exact type I errors for increasing sample size n and five εs: 0.2, 0.4, 0.6, 0.8, and 1. Top left is one sample z test; top right is one sample t test; bottom left is two sample t test with equal sample size and equal variance; bottom right is two sample t test with unequal sample sizes and equal variance.

where $\tau_1 = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_1$, $\tau_2 = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}/\lambda_2$, $b_1 = \frac{(n_1+n_2-2)\sigma^2}{(n_1-1)\lambda_3}$, $b_2 = \frac{(n_1+n_2-2)\sigma^2}{(n_2-1)\lambda_4}$, $\theta_0 = \frac{2}{n_1+n_2-2}$, $\theta_1 = \frac{2}{n_1+n_2-2-2b_1}$, $\theta_2 = \frac{2}{n_1+n_2-2-2b_2}$, $\theta_3 = \frac{2}{n_1+n_2-2+2b_1}$, $\theta_4 = \frac{2}{n_1+n_2-2+2b_2}$.

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

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

where $b = 2\sigma^2/\lambda_3$, $\theta_0 = \frac{2}{n_1+n_2-2}$, $\theta_1 = \frac{2}{n_1+n_2-2-b}$, and $\theta_2 = \frac{2}{n_1+n_2-2+b}$.

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

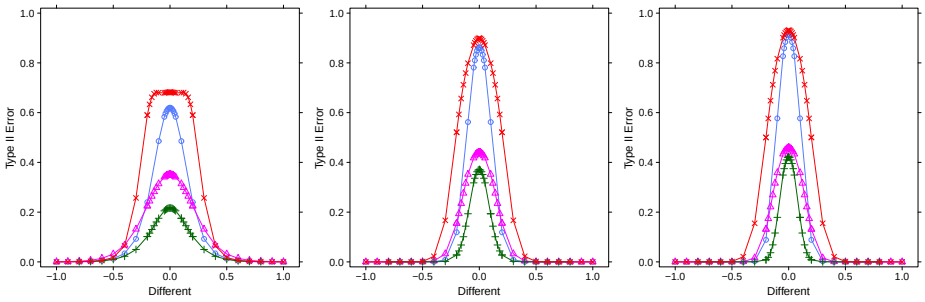

Fig. 2. Red line X is for one sample z test; blue line o is for one sample t test; pink line triangle is for two sample t test with equal sample size and equal variance; green line + is for two sample t test with unequal sample size and equal variance. $n = 50$. Left: $\varepsilon = 0.2$; Middle: $\varepsilon = 0.6$; Right:$\varepsilon = 1$.

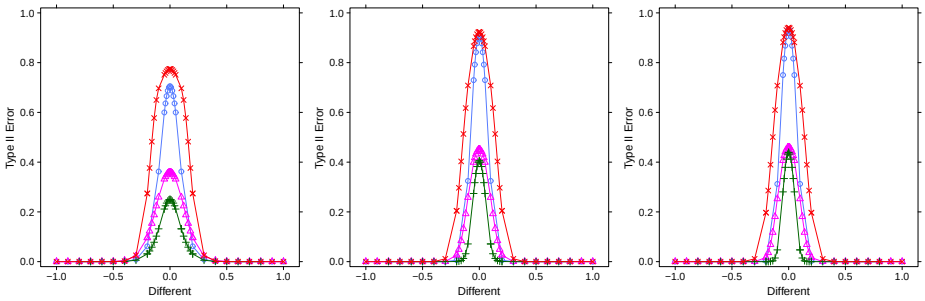

Fig. 3. Red line X is for one sample z test; blue line o is for one sample t test; pink line triangle is for two sample t test with equal sample size and equal variance; green line + is for two sample t test with unequal sample size and equal variance. $n = 100$. Left: $\varepsilon = 0.2$; Middle: $\varepsilon = 0.6$; Right:$\varepsilon = 1$.

Equation 5. Let $t_{\frac{\alpha}{2},\upsilon}$ be the $(1-\frac{\alpha}{2})$ quantile of a t distribution with υ degrees of freedom. The exact 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. Let $\delta = \mu_1 - \mu_2$.

$$\alpha^a = P\left(|T^a| > t_{\frac{\alpha}{2},\upsilon} \mid \delta = 0\right) = 1 - F_T(t_{\frac{\alpha}{2},\upsilon} \mid \delta = 0) + F_T(-t_{\frac{\alpha}{2},\upsilon} \mid \delta = 0),$$

$$\beta^a = P\left(|T^a| < t_{\frac{\alpha}{2},\upsilon} \mid \delta \neq 0\right) = F_T(t_{\frac{\alpha}{2},\upsilon} \mid \delta \neq 0) - F_T(-t_{\frac{\alpha}{2},\upsilon} \mid \delta \neq 0).$$

3.4 Experimental Evaluation

To examine when the exact type I and II errors are less reliable, we run a set of experiments and provide the results in the following tables and figures. For all the experiments we set $\alpha = 0.05$, increase sample size n from 50 to 400 by steps of 25, and examine five ε values, 0.2, 0.4, 0.6, 0.8 and 1. $\lambda = 1/(n_i\varepsilon)$.

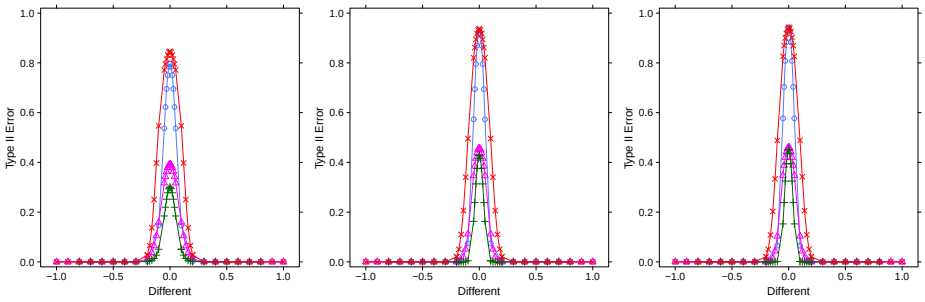

Fig. 4. Red line X is for one sample z test; blue line o is for one sample t test; pink line triangle is for two sample t test with equal sample size and equal variance; green line + is for two sample t test with unequal sample size and equal variance. $n = 200$. Left: $\varepsilon = 0.2$; Middle: $\varepsilon = 0.6$; Right:$\varepsilon = 1$.

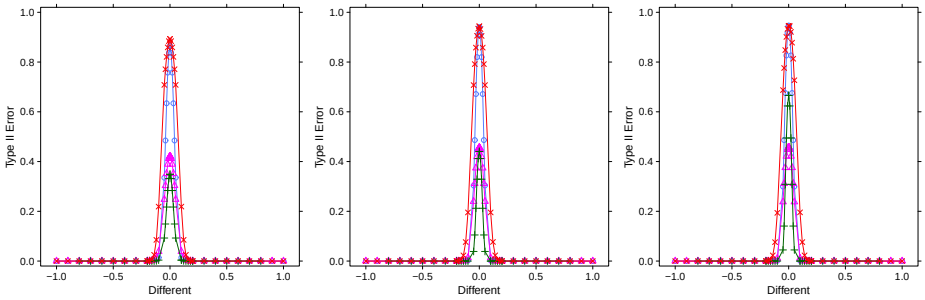

Fig. 5. Red line X is for one sample z test; blue line o is for one sample t test; pink line triangle is for two sample t test with equal sample size and equal variance; green line + is for two sample t test with unequal sample size and equal variance. $n = 400$. Left: $\varepsilon = 0.2$; Middle: $\varepsilon = 0.6$; Right:$\varepsilon = 1$.

In Table 1, we show the exact type I errors for selected sample size n: 50, 100, 200, 300, and 400. Figure 1 shows the exact type I errors for the tests with increasing sample size n. As sample size increases and ε becomes larger, the exact type I errors is approaching $\alpha = 0.05$. Considering the exact type I error only, when users construct a test statistic with noise added mean and variance, the sample size needs to 100 or larger to provide a reliable result for moderate to small noise. For large noise, i.e. $\varepsilon \leq 0.2$, the sample size needs to be 400 or larger for a reliable test.

Figures 2, 3, 4 and 5 show the type II errors with noise added query results for selected n: 50, 100, 200, 400 and ε: 0.2, 0.6, 1. Hypothesis tests often operate with far less samples than classification, since the test is always significant for large dataset. For the tests considered in this article, the type I errors based on noise added query results decrease sharply as sample size increases. Type II error depends on the difference between the true value and the hypothesized value. The type II error under differential privacy also improves significantly as sample size increases and ε becomes larger.

Notice users cannot know how much noises are added to the query results. Small noises can cause major distortion to the test results. 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.

n	$\varepsilon = 0.2$	$\varepsilon = 0.4$	$\varepsilon = 0.6$	$\varepsilon = 0.8$	$\varepsilon = 1$	n	$\varepsilon = 0.2$	$\varepsilon = 0.4$	$\varepsilon = 0.6$	$\varepsilon = 0.8$	$\varepsilon = 1$
a						b					
50	0.3177	0.1542	0.1011	0.0794	0.0689	50	0.3805	0.2109	0.1373	0.1023	0.0841
100	0.2251	0.1070	0.0762	0.0647	0.0594	100	0.2953	0.1415	0.0935	0.0746	0.0657
200	0.1542	0.0794	0.0631	0.0573	0.0546	200	0.2035	0.0968	0.0711	0.0618	0.0575
300	0.1239	0.0697	0.0587	0.0548	0.0531	300	0.1606	0.0813	0.0639	0.0578	0.0549
400	0.1070	0.0647	0.0565	0.0536	0.0523	400	0.1363	0.0734	0.0603	0.0559	0.0537
c						d					
50	0.4645	0.2576	0.1612	0.1157	0.0924	50	0.4977	0.3748	0.2726	0.1975	0.1518
100	0.3609	0.1673	0.1057	0.0815	0.0701	100	0.4944	0.2920	0.1844	0.1319	0.1039
200	0.2472	0.1108	0.0774	0.0653	0.0597	200	0.4066	0.1976	0.1236	0.0922	0.0744
300	0.1936	0.0907	0.0681	0.0601	0.0564	300	0.3376	0.1557	0.1001	0.0714	0.0602
400	0.1627	0.0805	0.0634	0.0575	0.0542	400	0.2917	0.1320	0.0871	0.0613	0.0549

Table 1. (a) Z test type I error with added noises. $\sigma = 0.5$. (b) One sample t test type I error with added noises. $\sigma = 0.4$. (c) Two sample t test with equal sample size type I error with added noises. $\sigma_1 = \sigma_2 = 0.35$. $n_1 = n_2 = n$. (d) Two sample t test with unequal sample size type I error with added noises. $\sigma_1 = \sigma_2 = 0.2$. $n_1 = n$ and $n_2 = 1.1n$.

4 Differentially Private Bayesian Classifier for Gaussian Mixture Models

Let database $D = \{X_1, \ldots, X_d, W\}$, where W is a binary class label, $Dom(W) = \{w_1, w_2\}$, and each X_i, $1 \leq i \leq d$ is a continuous attribute. A Bayesian classifier has the following decision rule:

$$\text{Assign a record } \mathbf{x} \text{ to } w_1 \text{ if } P(w_1|\mathbf{x}) > P(w_2|\mathbf{x}) \text{ ; otherwise assign it to } w_2.$$

The probabilities $P(w_i|\mathbf{x})$ can be calculated as: $P(w_i|\mathbf{x}) = p(\mathbf{x}|w_i)P(w_i)/p(\mathbf{x})$. If $p(\mathbf{x}|w_i)$ follows multivariate Gaussian distribution, it is known as the Gaussian mixture model [5]. For each class w_i, its mean μ_i and the variance-covariance matrix Σ_i of $p(\mathbf{x}|w_i) \sim N(\mu_i, \Sigma_i)$ are estimated from the data set D. For binary case, the Bayes error (i.e., the classification error) is calculated as [5]:

$$\text{Bayes Error } = \int_{\mathscr{R}_1} p(\mathbf{x}|w_2)P(w_2)d\mathbf{x} + \int_{\mathscr{R}_2} p(\mathbf{x}|w_1)P(w_1)d\mathbf{x}.$$

$\mathscr{R}_1$ is the region where records are labeled as class 1, and $\mathscr{R}_2$ is the region where records are labeled as class 2.

In this article we examine Bayes error for Gaussian mixture models under differential privacy protection. The database D only needs to return the following for users to build a Bayesian classifier:

- The sample size in D, which has sensitivity 0,
- The proportions of the two classes, i.e., $P(w_1)$ and $P(w_2)$,
- For each category, mean μ_i and variance-covariance Σ_i of the multivariate Gaussian distribution for $p(\mathbf{x}|w_i)$.

Bounded variables fit well into differential privacy mechanism. With unbounded variables one very large or small record can cause a significant increase the sensitivity. Notice Gaussian distribution is unbounded. Hence we work with truncated Gaussian distribution over interval $[\mu-6\sigma, \mu+6\sigma]$, a probability range of 0.999999998. Truncated Gaussian has density $I_{\{\mu-6\sigma\le x\le\mu+6\sigma\}}(x)\frac{f(x)}{\Phi(6)-\Phi(-6)}$.

4.1 Repair Noise Added Variance-Covariance Matrix

Let $\hat{\Sigma}=(\hat{\sigma}_{ij})_{d\times d}$ be the sample variance-covariance 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 variance-covariance matrix, which is the results that users can easily obtain to test their model. Σ_Q is still symmetric but seize to be positive definite. In order to have a valid variance-covariance matrix, we repair the noise added variance-covariance 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, $l_1\ge l_2\ge ...\ge l_d$. 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 to find a Σ_+ that minimizes the determinant 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 a fine grid search over wide intervals to obtain positive eigenvalues that

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

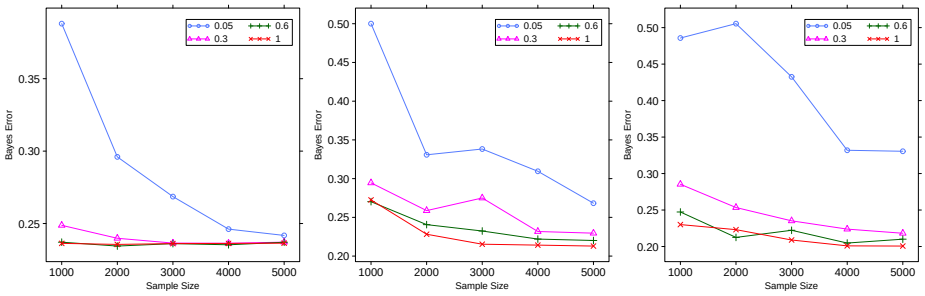

Fig. 6. Small training sample LDA Bayes error. Left: 2 dimension; Middle: 5 dimension; Right: 10 dimension.

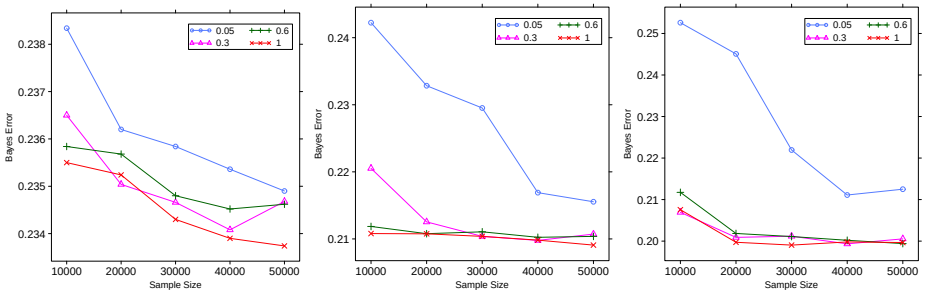

Fig. 7. Large training sample LDA Bayes error. Left: 2 dimension; Middle: 5 dimension; Right: 10 dimension.

4.2 Experimental Evaluation

We have conducted extensive experiments in this section. We consider binary classification scenario. To understand how differential privacy affects the Bayes error, we do not want to introduce any other errors. Note Gaussian distribution may not represent the underlying data accurately. To avoid 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 two experiments, and synthetic in the rest.

In Equation 4, if the two Gaussian distributions have the same variance-covariance matrix, we perform a linear discriminant analysis (LDA). If the two Gaussian distributions have different variance-covariance 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 to 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 according to [31], sample means and variance-covariance matrices. Then we add independent Laplace noises to each Gaussian component.

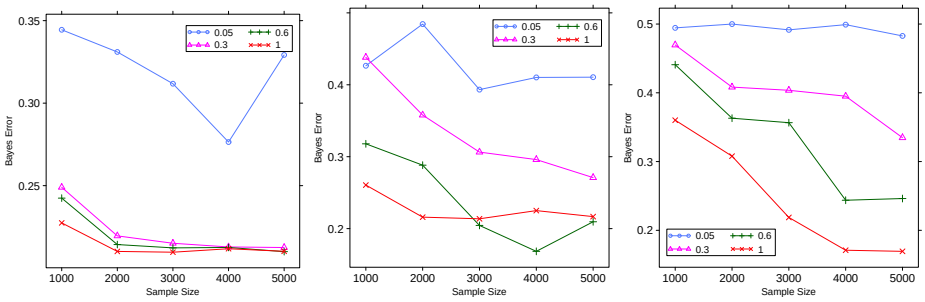

Fig. 8. Small training sample QDA Bayes error. Left: 2 dimension; Middle: 5 dimension; Right: 10 dimension.

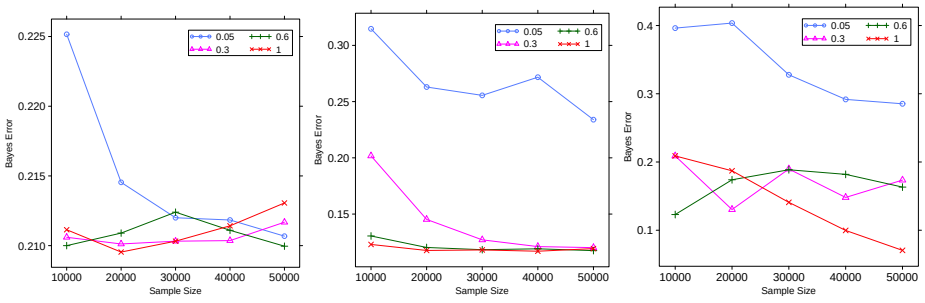

Fig. 9. Large training sample QDA Bayes error. Left: 2 dimension; Middle: 5 dimension; Right: 10 dimension.

3. We repair the noise added variance-covariance 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.

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 variance-covariance 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 variance-covariance matrix. We compute the sensitivity for variances and covariances adjusted to the range of the pooled data. The sample means and the sensitivity values for sample means are computed. We run the experiments in 2-dimension, 5-dimension, and 10-dimension, $d = 2, 5, 10$. 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 4. We generate 5 million samples using the prespecified parameter values without truncation, using the true LDA classification rules to estimate Bayes

	2-D	5-D	10-D		2-D	5-D	10-D
LDA Bayes error	0.2351	0.2100	0.1996	QDA Bayes error	0.2105	0.1170	0.0589

Table 2. True LDA and QDA Bayes errors

error. We take the average Bayes error of four such runs as the actual LDA Bayes error, shown in Table 2.

Figures 6 and 7 show the Bayes error under differential privacy 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 variance-covariance 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 used in Experiment 1. We set the prior as $p_1 = p_2 = 0.5$. The sample means, variances, covariances, and the sensitivity values are computed. Again, we run the experiments in 2-dimension, 5-dimension, and 10-dimension, $d = 2, 5, 10$. 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 4. 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 1.

Figures 8 and 9 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 variance-covariance matrices, dimensionality has a large impact on the Bayes error estimates obtained under differential privacy. For ε from 0.3 to 1, 2 dimensional experiment shows that three thousand training samples is sufficient 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 Bayes error under differential privacy.

Experiment 3. 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 variance-covariance matrix of each class in the Parkinson data and used these parameters in our Gaussian mixture models. In all of the experiments, we set $\varepsilon = 0.6$. For the Parkinson data, the majority class equals to 75.38% of the total. There are 197 observations and 21 numerical variables besides the class label. 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 when two classes have different variance-covariance matrices. As dimensionality increases, we need a very large number of training samples to reduce the impact of the added noises.

Experiment 4. We also used the Adult data set from the UCI Machine learning repository (https://archive.ics.uci.edu/ml/datasets/Adult). The Adult data is much larger than the Parkinson data, with 32,561 observations. We used all the numerical variables in this experiment, i.e., 6 variables. We computed the mean and variance-covariance matrix of each class in the Adult data and used these parameters in our Gaussian mixture models. Again we set $\varepsilon = 0.6$. For the Adult data, the majority class equals to 75.92% of the total, similar to the Parkinson data. Without differential privacy mechanism, directly using the sample estimates, the Bayes error is 0.0309. With 50,000 training samples, the Gaussian mixture model under differential privacy has the Bayes error equal to 0.0747. The impact of the added noises is less severe for this lower dimensional data. Training sample size around 50,000 provides a reasonable result.

5 Summary

In this article we examine the performance of Bayesian classifier using the noise added mean and variance-covariance matrices. We also study the exact 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.

This is a general issue for random noise addition. Users may simply assemble basic query results without directly querying a complex statistic. Then adding noises causes the assembled result to no longer satisfy certain constraints. The query results need to be further modified in order to be used in subsequent studies.

Bibliography

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

[2] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research* 12: 1069-1109, 2011

[3] K. Chaudhuri, A. D. Sarwate, and K. Sinha. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Res*, 14:2905-2943, 2013.

[4] C. Clifton, and T. Tassa. On syntactic anonymity and differential privacy. *Transactions on Data Privacy*, 6(2), 161-183, 2013.

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

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

[7] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. *51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 51-60, 2010.

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

[9] 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.

[10] 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.

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

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

[13] 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.

[14] M. Kapralov and K. Talwar. On differentially private low rank approximation. In *SODA '13: Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1395–1414, New Orleans, Louisiana, 2013. SIAM.

[15] 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.

[16] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research* 23:1-41, 2012

[17] J. Lei. Differentially private M-estimators. In *Advances in Neural Information Processing Systems*, pages 361–369, 2011.

[18] 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.

[19] 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.

[20] M. A. Pathak and B. Raj. Large margin Gaussian mixture models with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 9(4):463–469, 2012.

[21] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the Netflix prize contenders. In *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 627–636, Paris, France, 2009. ACM.

[22] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS '07: 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94-103, Providence, Rhode Island, 2007. IEEE.

[23] 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.

[24] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. *Proc. of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, Seattle, WA, USA, June 1-3, 1998

[25] P. Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6), 1010-1027, 2001.

[26] J. Soria-Comas and Josep Domingo-Ferrer. Connecting privacy models: synergies between k-anonymity, t-closeness and differential privacy. *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2013.

[27] J. Domingo-Ferrer. On the connection between t-closeness and differential privacy for data releases. *IEEE International Conference on Security and Cryptography (SECRYPT)*, 2013.

[28] B. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning. *Journal of Privacy and Confidentiality*, 4(1): 65-100, 2012.

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

[30] D. Vu and A. Slavkovic. Differential privacy for clinical trial data: preliminary evaluations. In *IEEE 13th International Conference on Data Mining Workshops*, pages 138-143, Los Alamitos, CA, USA, 2009. IEEE.

[31] B. Xi, M. Kantarcioglu, and A. Inan. Mixture of Gaussian models and Bayes error under differential privacy. In *Proceedings of the first ACM conference on Data and application security and privacy*, pages 179-190. ACM, 2011.

[32] 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.

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