# Bayesian Inference Via Partitioning Under Differential Privacy

by

Gilad Amitai

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome Reiter, Supervisor

_____
Colin Rundel

_____
Galen Reeves

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2018

# Abstract

## Bayesian Inference Via Partitioning Under Differential Privacy

by

Gilad Amitai

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome Reiter, Supervisor

_____
Colin Rundel

_____
Galen Reeves

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2018

# Abstract

In this thesis, I develop differentially private methods to report posterior probabilities and posterior quantiles of linear regression coefficients. I accomplish this by randomly partitioning the data, taking an intermediate outcome of the data within each partition, aggregating the intermediate outcomes so that they approximate the statistic of interest, and adding Laplace noise to ensure differential privacy. I find the posterior probability by assuming that the variance of the posterior distribution given data from one partition is proportional to the variance of the posterior distribution given the full dataset. The mean posterior probability of the data within each partition is found as the intermediate outcome. The posterior probabilities given the data within one partition are averaged and the variance is rescaled so that the averaged probability approximates the posterior probability given the full dataset. Added noise ensures that the released quantity satisfies differential privacy. I find the posterior quantile by fitting the Bayesian model on the data within each partition where the likelihood has been inflated to rescale the posterior variance so that it approximates the posterior variance given the full dataset. The posterior quantile of the data within each partition is found as an intermediate outcome, and averaged to approximate the posterior quantile given the whole dataset. I add noise to ensure the released quantity satisfies differential privacy. Simulations show that both the partitioning methods and the noise mechanism can return accurate estimates of the statistics they are perturbing.

This research was supported by NSF Grants SES-1131897 and ACI-1443014.

# Contents

# List of Figures

# 1
# Introduction

Data stewards seek to provide researchers access to their data for analysis, but may be legally or ethically required to ensure the confidentiality of individuals providing data. Research has shown that removing directly identifying information alone does not guarantee privacy. For example, researchers were able to reveal the identity of anonymized medical records sold by Washington state by linking the sold data to newspaper stories containing the word "hospitalized" [Sweeney 2013]. Systems where the analyst has no access to individual-level data and queries a server for aggregate information can still put the identity of individuals at risk. It has been shown that given genotype data from an individual, it is possible to identify that individual's inclusion in aggregate genotype frequencies [Homer et al. 2008].

To address this, researchers have developed methods to release outputs that adhere to differential privacy [Dwork & Roth 2014], a formal definition of privacy that ensures the output of an algorithm will not reveal the inclusion of a particular record. This is achieved by releasing output with added noise that is large enough to obscure any particular record.

Differentially private methods have been adapted for use with frequentist statisti-

cal regression [Chen, et al 2016, Barrientos, et. al 2017], but have not been adapted yet to Bayesian regression. In the Bayesian framework, an analyst would define a likelihood and prior to obtain a posterior distribution of a regression coefficient. Typical summary statistics of a posterior distribution are the cumulative probability and quantile. To release differentially private versions of these statistics, we randomly split the dataset into exclusive partitions, take an intermediate outcome of the data within each partition, aggregate the intermediate outcomes so that they approximate the statistic of interest, and add Laplace noise to ensure differential privacy.

The remainder of this thesis is organized as follows. Section 2 will introduce differential privacy. Section 3 will introduce two methods to release differentially private posterior probabilities and show their effectiveness using simulations. The first method assumes the posterior variance of the regression coefficient given data from one partition is proportional to the variance of the posterior distribution given the full dataset. The posterior probabilities from each partition are averaged, and the $Z$-score of this value is scaled up so that it approximates the posterior probability with variance given the full dataset. A second method uses the sum of the $-2\log$'s of the posterior probabilities from each partition in a manner similar to Fisher's method to approximate the posterior probability given the full dataset. Section 4 will introduce a method to release differentially private posterior quantiles where the Bayesian model is fit within each partition with a likelihood that has been inflated so that the posterior variance within each partition matches the posterior variance using the full dataset. The posterior quantiles from each partition are averaged to estimate the posterior quantile found using the full dataset. Section 5 will conclude the thesis.

# 2

# Review of Differential Privacy

Differential privacy is a formal definition of privacy that ensures the output of an algorithm will not reveal the inclusion of any particular record. The principles of differential privacy are described in [Dwork & Roth 2014]. We review these here.

Let $\mathcal{D}$ be a confidential dataset. We define two databases $\mathcal{D}$ and $\mathcal{D}^*$ to be neighboring if the two differ in only one row. A randomized algorithm $\mathcal{A}$ satisfies differential privacy if for any pair of neighboring datasets $\mathcal{D}$ and $\mathcal{D}^*$ and any output $\mathcal{S} \in \text{range}(\mathcal{A})$, it is true that $\Pr\left(\mathcal{A}(\mathcal{D}) = \mathcal{S}\right) \leqslant \exp(\epsilon)\Pr\left(\mathcal{A}(\mathcal{D}^*) = \mathcal{S}\right)$. The value of $\epsilon$, called the privacy budget, controls the level of privacy offered by the algorithm $\mathcal{A}$, with lower values of $\epsilon$ corresponding to more privacy.

Differentially private algorithms satisfy the following composition properties. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be differentially private algorithms with privacy budgets $\epsilon_1$ and $\epsilon_2$. The property of sequential composition states that for any database $\mathcal{D}$, releasing the outputs of $\mathcal{A}_1(\mathcal{D})$ and $\mathcal{A}_2(\mathcal{D})$ ensures differential privacy with privacy budget $(\epsilon_1+\epsilon_2)$. The property of parallel composition states that for disjoint databases $\mathcal{D}_1$ and $\mathcal{D}_2$, releasing the outputs of both $\mathcal{A}_1(\mathcal{D}_1)$ and $\mathcal{A}_2(\mathcal{D}_2)$ satisfies differential privacy with budget $\max\{\epsilon_1, \epsilon_2\}$. The property of post-processing states that for any deterministic

function $j$, releasing $j(\mathcal{A}_1(\mathcal{D}))$ for any confidential database $\mathcal{D}$ still ensures differential privacy with budget $\epsilon_1$.

A method to ensure differential privacy is the Laplace Mechanism [Dwork et al. 2006]. For any function $f : \mathcal{D} \to \mathbb{R}^d$, let the global sensitivity be $\Delta(f) = \max_{\mathcal{D},\mathcal{D}^*} ||f(\mathcal{D}) - f(\mathcal{D}^*)||$, where $\mathcal{D}$ and $\mathcal{D}^*$ are neighboring datasets. This value measures the maximum $L_1$ distance of the outputs of the function $f$ for two neighboring datasets. The Laplace mechanism is given by $LM = f(\mathcal{D}) + \eta$ where $\eta$ is a $\mathbb{R}^d$ vector of independent draws from a Laplace distribution with density $p(x|\lambda) = \frac{1}{2\lambda} \exp\{-|x|/\lambda\}$ where $\lambda = \Delta(f)/\epsilon$.

As an example, let a confidential dataset be made up of $n$ elements $y_i \in \{0, 1\}$. An analyst wanting to know the proportion of ones would find $f(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} y_i$. Let $\mathcal{D} = \{1, \mathcal{D}_{-y_1}\}$ and $\mathcal{D}^* = \{0, \mathcal{D}_{-y_1}\}$ where $\mathcal{D}_{-y_1}$ is the same in both datasets so that they only differ in $y_1$. The two datasets are neighboring and the global sensitivity for the function $f$ is

$$\Delta(f) = \max_{\mathcal{D},\mathcal{D}^*} \left|\left| \frac{\sum \mathcal{D}^* + 1}{n} - \frac{\sum \mathcal{D}^* + 0}{n} \right|\right| = \frac{1}{n}. \tag{2.1}$$

Using the Laplace Mechanism, the analyst would receive $LM = f(\mathcal{D}) + \eta$ where $\eta$ is a draw from a Laplace $\left(\frac{1}{n\epsilon}\right)$.

It is useful to control the global sensitivity of $f$ by using the sample and aggregate technique [Nissim, et al. 2007]. We partition the dataset $\mathcal{D}$ into $M$ exclusive subsets $\mathcal{D}_1, \ldots, \mathcal{D}_M$ and compute some intermediate outcome of the data within each partition $h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)$. We find a related aggregation function $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$ of the intermediate outcomes where the global sensitivity of $g$ will be less than the global sensitivity of $f$ since a single observation can only affect one of the partitioned datasets. Noise is added to the aggregation function to preserve differential privacy. The aggregation function is chosen to reduce the global sensitivity and degree of noise

added. The output of the aggregation function can go through a post-processing step to best approximate $f(\mathcal{D})$.

We now go through an example of the sample and aggregate technique. We partition $\mathcal{D}$ into $M$ exclusive subsets $\mathcal{D}_1, \ldots, \mathcal{D}_M$ each of size $n^*$ where we want to release the output of a function $f$ with global sensitivity one. We apply the function $f$ to each partition of the data as an intermediate outcome, and take the mean of the intermediate outcomes as the aggregation function so that

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = \frac{1}{M} \sum_{j=1}^{M} f(\mathcal{D}_j). \tag{2.2}$$

For two neighboring datasets $\mathcal{D}$ and $\mathcal{D}^*$, only one of the partitions can be affected by the differing data point, let's say $\mathcal{D}_1$. Knowing this, we can find the global sensitivity of the aggregation function

$$\Delta(g) = \max \left|\left| \frac{f(\mathcal{D}_1) + \sum_{j \neq 1} f(\mathcal{D}_j)}{M} - \frac{f(\mathcal{D}_1^*) + \sum_{j \neq 1} f(\mathcal{D}_j)}{M} \right|\right| = \frac{1}{M} \max ||f(\mathcal{D}_1) - f(\mathcal{D}_1^*)|| = \frac{1}{M} \cdot 1. \tag{2.3}$$

The analyst would receive $LM = g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) + \eta$ where $\eta$ is a draw from a Laplace $\left(\frac{1}{M} \cdot \frac{1}{\epsilon}\right)$. If we had applied the function $f$ directly to the dataset $\mathcal{D}$, the scale parameter of our Laplace noise distribution would have been 1. Thus, we achieve less variance in the noise distribution if we use the sample and aggregate technique.

# 3

# Reporting Differentially Private Posterior Probabilities

If the analyst had access to the entire confidential dataset $\mathcal{D}$, he or she would be interested in the posterior probability of a regression coefficient $\beta$ such that $p(b) = \Pr(\beta \leqslant b|\mathcal{D})$. I examine two functions of the posterior probabilities from partitions of the data that approximate $p(b)$: one based on assumptions of the posterior variances, and the other based on the sum of the $-2\log$'s of the posterior probabilities from each partition of the data.

## 3.1 Normality Method

Let $\mathcal{D}$ be a confidential dataset made up of $n$ observations $\{y_i, \boldsymbol{x}_i\}$ where $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\boldsymbol{x}_i$ being some fixed $p \times 1$ vector of fixed covariates. We fit a conjugate Bayesian model with priors

$$p(\sigma_\epsilon^2) \sim \text{inverse} - \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \quad \text{and} \quad p(\boldsymbol{\beta}|\sigma_\epsilon^2) \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0^{-1}\sigma_\epsilon^2\right). \tag{3.1}$$

Given the full dataset, the posterior distribution of the regression coefficients of this model is

$$p(\boldsymbol{\beta}|\sigma_\epsilon^2, \mathcal{D}) \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, (\Sigma_0^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma_\epsilon^2\right) \tag{3.2}$$

where $\hat{\boldsymbol{\beta}} = (\Sigma_0^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1}(\Sigma_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{y})$. Given only a partition of the data $\mathcal{D}_j = \{\boldsymbol{y}_j, \boldsymbol{X}_j\}$, the posterior distribution of the regression coefficient is

$$p(\boldsymbol{\beta}|\sigma_\epsilon^2, \mathcal{D}_j) \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_j, (\Sigma_0^{-1} + \boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\sigma_\epsilon^2\right). \tag{3.3}$$

where $\hat{\boldsymbol{\beta}}_p = (\Sigma_0^{-1} + \boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}(\Sigma_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}_j^T\boldsymbol{y}_j)$. I assume we are working with a large dataset so that the variance of the posterior distribution using the full dataset can be approximated with $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma_\epsilon^2$ and the variance for the posterior distribution of the data within a single partition can be approximated with $(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\sigma_\epsilon^2$. The marginal posterior distribution of regression coefficient $\beta_k$ given the full dataset is

$$p(\beta_k|\sigma_\epsilon^2, \mathcal{D}) \sim \mathcal{N}\left(\hat{\beta}_k, v_k^*\right) \tag{3.4}$$

where $\hat{\beta}_k$ is the $k$th entry of $\hat{\boldsymbol{\beta}}$ and $v_k^*$ is the $k$th diagonal entry of $\mathrm{Var}\left[\boldsymbol{\beta}|\sigma_\epsilon^2, \mathcal{D}\right]$. The marginal posterior distribution of regression coefficient $\beta_k$ given partition $j$ of the data is

$$p(\beta_k|\sigma_\epsilon^2, \mathcal{D}_j) \sim \mathcal{N}\left(\hat{\beta}_{j,k}, v_{j,k}\right) \tag{3.5}$$

where $\hat{\beta}_{j,k}$ is the $k$th entry of $\hat{\boldsymbol{\beta}}_j$ and $v_{j,k}$ is the $k$th diagonal entry of $\mathrm{Var}\left[\boldsymbol{\beta}|\sigma_\epsilon^2, \mathcal{D}_j\right]$.

After partitioning the data, I take the posterior probability within each partition, which given the approximate normality of the posterior distribution, is the value $\Phi\left(\frac{b-\hat{\beta}_{j,k}}{\sqrt{v_{j,k}}}\right)$. I take the mean of the posterior probabilities assuming that the differences between the posterior variances for each partition are small enough so that we can use a single variance $v_k$ to represent them. Thus, we assume

$$\bar{p} = \frac{1}{M}\sum_{j=1}^{M}\Phi\left(\frac{b-\hat{\beta}_{j,k}}{\sqrt{v_{j,k}}}\right) \approx \frac{1}{M}\sum_{j=1}^{M}\Phi\left(\frac{b-\hat{\beta}_{j,k}}{\sqrt{v_k}}\right). \tag{3.6}$$

I further assume that the mean of the posterior probabilities within each partition results in a posterior probability that is centered around the posterior mean using the full dataset

$$\bar{p} \approx \frac{1}{M} \sum_{j=1}^{M} \Phi\left(\frac{b - \hat{\beta}_{j,k}}{\sqrt{v_k}}\right) \approx \Phi\left(\frac{b - \hat{\beta}_k}{\sqrt{v_k}}\right). \tag{3.7}$$

I make the assumption that $Mv_k^* = v_k$ where $v_k^*$ is the posterior variance using the full dataset since $1/M$ of the data used to compute $v_k^*$ is used to compute $v_k$. I can rescale the variance of $\bar{p}$ to have variance $v_k^*$. I find the $Z$-score of the posterior probability and scale this to have the wider variance such that using Equation (3.7) we can say

$$\sqrt{M}\Phi^{-1}(\bar{p}) \approx \frac{b - \hat{\beta}}{\sqrt{v_k/M}} = \frac{b - \hat{\beta}}{\sqrt{v_k^*}}. \tag{3.8}$$

Our intermediate outcomes are $h(\mathcal{D}_j) = p_j(b) = \Pr(\beta \leqslant b|\mathcal{D}_j)$, the posterior probabilities from partition $j$, and our function to aggregate the data is

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = \frac{1}{M} \sum_{j=1}^{M} p_j(b). \tag{3.9}$$

We are working directly with probabilities $p_j(b)$ that have a global sensitivity of one because of their necessary bounds. Because we are averaging probabilities from partitions of the data, where each piece of data can manipulate only one partition, the global sensitivity of the aggregation function $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$ is $1/M$. We add Laplace noise to the aggregation function $g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M))$ to maintain differential privacy with scale parameter $\frac{1/M}{\epsilon}$, where $\epsilon$ is the entire privacy budget. The $Z$-score of this value is found, multiplied by a factor of $\sqrt{M}$, and then plugged into a normal CDF to find the best approximation of $p(b)$. The final version of the statistic

released to the analyst will be

$$LM_p = \Phi\left(\sqrt{M} \cdot \Phi^{-1}\left(\frac{1}{M}\sum_{j=1}^{M} p_j(b) + \eta\right)\right), \tag{3.10}$$

where $\eta$ is a draw from the Laplace distribution with mean parameter zero and scale parameter $\frac{1/M}{\epsilon}$ where $\epsilon$ is the entire privacy budget.

### *3.1.1 Evaluation*

To evaluate the utility and accuracy of the normality method, we want to ensure the differentially private posterior probability released to the user is approximately the posterior probability found using the full dataset. It is useful to also confirm that the posterior probability found using the normality method is approximately that found using the entire dataset. We want

$$p(b) \approx \Phi\left(\sqrt{M} \cdot \Phi^{-1}\left(\frac{1}{M}\sum_{j=1}^{M} p_j(b)\right)\right) \approx LM_p \tag{3.11}$$

for any confidential dataset used. I will show this by using calibration probabilities [Rubin 1984] that show, under any dataset, that the expected value of the posterior probability, the expected value of the posterior probability using the normality method, and the expected value of the differentially private posterior probability are all approximately the same,

$$\mathrm{E}\left[\int p(b) d\mathcal{D}\right] \approx \mathrm{E}\left[\int \Phi\left(\sqrt{M} \cdot \Phi^{-1}\left(\frac{1}{M}\sum_{j=1}^{M} p_j(b)\right)\right) d\mathcal{D}\right] \approx \mathrm{E}\left[\int LM_p d\mathcal{D}\right]. \tag{3.12}$$

To show this empirically, I simulate artificial datasets of size 100000 from $y_i = 1 + 2x_i + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$ and split the dataset into partitions of size 100. The posterior density of the regression coefficients using the full dataset and within each

partition is found using a Zellner's $g$-prior where $g$ is the length of the dataset so that the posterior density is

$$p(\boldsymbol{\beta}|\sigma_\epsilon^2, \mathcal{D}) \sim \mathcal{N}\left(\frac{100000}{100001}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \frac{100000}{100001}\sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

and

$$p(\boldsymbol{\beta}|\sigma_\epsilon^2, \mathcal{D}_j) \sim \mathcal{N}\left(\frac{100}{101}(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^T\boldsymbol{y}_j, \frac{100}{101}\sigma_\epsilon^2(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\right).$$

The distribution under random datasets of the posterior probability, of the posterior probability found using the normality method, and of the differentially private posterior probability where $\epsilon = 0.25$ is displayed in Figure 3.1. The mean of the distributions for different cutoff values $b$ are similar. We also examine that as the cutoff value of $b$ is increased, the means of the distributions increase in a similar pattern within each method. The means within each method follow a normal CDF curve as the cutoff value $b$ increases, which we expect since the posterior distribution of the regression coefficient is approximately normal.

To evaluate how well the posterior probability found using the normality method and the differentially private posterior probability approximate the posterior distribution using the full dataset, I take the differences between these values found in one dataset. In Figure 3.2, I plot

$$p(b) - \Phi\left(\sqrt{M} \cdot \Phi^{-1}\left(\frac{1}{M}\sum_{j=1}^{M} p_j(b)\right)\right) \quad \text{and} \quad p(b) - LM_p \tag{3.13}$$

for each of the datasets I simulated. We see that there are bigger variances in differences when the cutoff value is closer to the true value of the regression coefficient. The normality method overestimates for values smaller than the true value until the difference converges to zero when the value $b$ is sufficiently smaller than the true value. The normality method underestimates for values greater than the true value

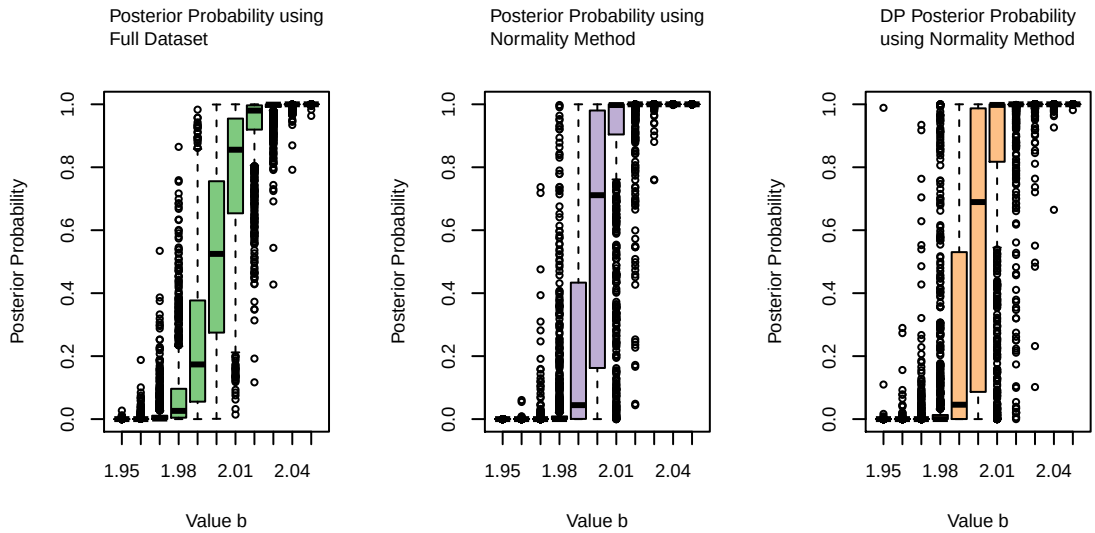


FIGURE 3.1: These graphs in order are: the distributions of the posterior probability found for various cutoff values $b$ using the full dataset, the distributions of the posterior probability found for various cutoff values $b$ using the normality method on partitioned data, and the distributions of the posterior probability found for various cutoff values $b$ on partitioned data using the normality method with noise added to ensure differential privacy where $\epsilon = 0.25$.

until the difference converges to zero when the value $b$ is sufficiently larger than the true value. Overall, the differences are small and confirm we are getting reasonably accurate estimates.

## 3.2 $-2\log p$ Method

Fisher's method [Fisher 1932] is relevant to our goal of combining separate probabilities. Fisher's method combines and tests probabilities in the form of the $p$-values from independent hypothesis tests. It states that after conducting $k$ independent hypothesis tests, if the null hypothesis is true and the $p$-values follow a uniform distribution on $(0, 1)$, then

$$-2\sum_{l=1}^{k}\log(\rho_l) \sim \chi^2_{2k}. \tag{3.14}$$

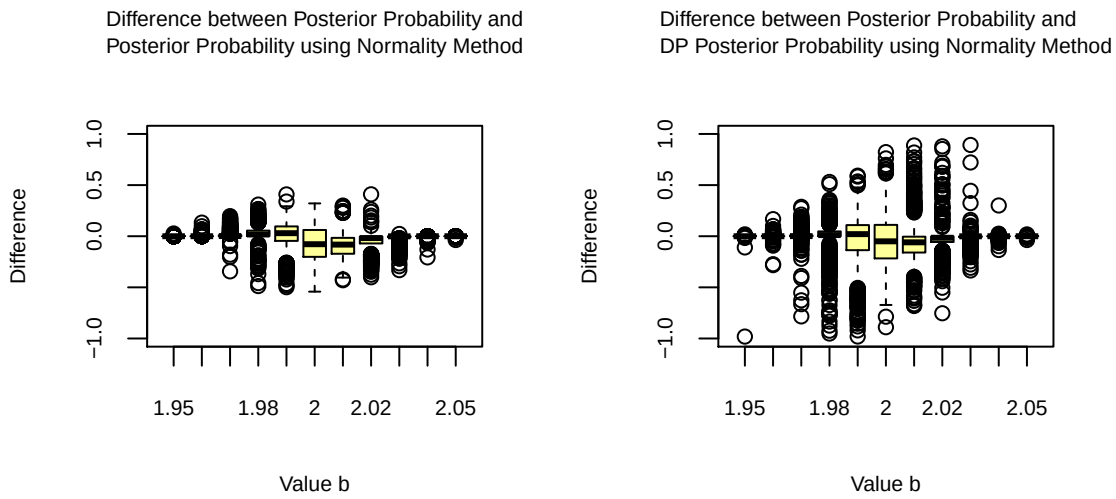


FIGURE 3.2: The graphs in order are: The distribution of the difference within one simulated dataset between the posterior probability given the full dataset and the posterior probability found using the normality method, and the distribution of the difference within one simulated dataset between the posterior probability given the full dataset and the posterior probability found using the normality method with noise added to ensure differential privacy where $\epsilon = 0.25$.

where $\rho_l$ is the $p$-value from hypothesis test $l$. While we have no guarantee that posterior probabilities given the data within a partition follow a uniform distribution, we have found through empirical tests that using the sum of the $-2\log$'s of the probabilities is effective in combining the posterior probabilities from within each partition. Let the intermediate outcome be $h(\mathcal{D}_j) = -2\log(p_j(b))$ where $p_j(b)$ is the posterior probability given the data within partition $j$. Let the aggregation function be

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = -2\sum_{j=1}^{M} \log(p_j(b)). \tag{3.15}$$

We are working with the $-2\log$ of probabilities, which are bounded between zero and infinity. To control the global sensitivity of the $-2\log$ of probabilities, we clip the $-2\log$ of probabilities at a reasonable value [Barrientos, et. al 2017]. The researcher

defines a probability threshold $p^*$ such that the analyst is comfortable with treating posterior probabilities smaller than $p^*$ as equivalent to $p^*$. The resulting $-2\log$ of posterior probabilities within each partition are clipped so that if $-2\log(p_j(b)) > -2\log(p^*)$, then $-2\log(p_j(b))$ is equal to $-2\log(p^*)$. The global sensitivity of the $-2\log$ of a posterior probability from within any partition is $-2\log(p^*)$, and the sensitivity of the aggregation function $-2\sum\log(p_j(b))$ is $\frac{-2\log(p^*)}{M}$.

The final version of the statistic release to the analyst is

$$LM_p = 1 - F_{\chi^2_{2m}}\left(-2\sum_{j=1}^{M}\log(p_j(b)) + \eta\right) \tag{3.16}$$

where $F_{\chi^2_k}(x)$ is the CDF of a $\chi^2_k$ distribution and $\eta$ is a draw from the Laplace distribution with mean parameter zero and scale parameter $\frac{-2\log(p^*)}{M\epsilon}$ where $\epsilon$ be the entire privacy budget.

### *3.2.1 Evaluation*

To confirm the utility and accuracy of the $-2\log p$ method, we want

$$p(b) \approx 1 - F_{\chi^2_{2m}}\left(-2\sum_{j=1}^{M}\log(p_j(b))\right) \approx LM_p \tag{3.17}$$

for any confidential dataset. I will show this by using calibration probabilities [Rubin 1984] that show, under any dataset, that the expected value of the posterior probability, the expected value of the posterior probability using the $-2\log p$ method, and the expected value of the differentially private posterior probability are all approximately the same,

$$\mathrm{E}\left[\int p(b)d\mathcal{D}\right] \approx \mathrm{E}\left[\int 1 - F_{\chi^2_{2m}}\left(-2\sum_{j=1}^{M}\log(p_j(b))\right)d\mathcal{D}\right] \approx \mathrm{E}\left[\int LM_p d\mathcal{D}\right]. \tag{3.18}$$

To show this empirically, I simulate artificial datasets of size 100000 from $y_i = 1 + 2x_i + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$ and split the dataset into partitions of size 100. I use a Zellner's $g$-prior, with $g$ being the length of the dataset, to find the posterior distributions. The distribution under random datasets of the posterior probability, of the posterior probability found using the $-2 \log p$ method where $p^* = 0.001$, and of the differentially private posterior probability where $\epsilon = 0.25$ and $p^* = 0.001$ is displayed in Figure 3.3. We see that the means of the distributions are approximately the same for different cutoff values $b$. The means within each distribution follow a normal CDF curve as the cutoff value $b$ increases, which we expect since the posterior distribution of the regression coefficient is approximately normal.

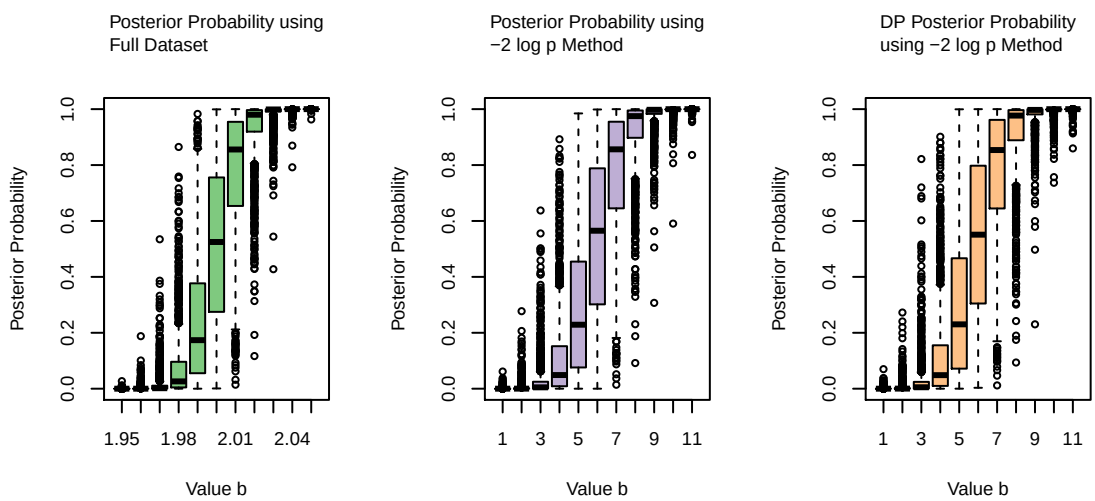


FIGURE 3.3: These graphs in order are: the distributions of the posterior probability found for various cutoff values $b$ using the full dataset, the distributions of the posterior probability found for various cutoff values $b$ using the $-2 \log p$ method on partitioned data where $p^* = 0.001$, and the distributions of the posterior probability found for various cutoff values $b$ on partitioned data using the $-2 \log p$ method with noise added to ensure differential privacy where $\epsilon = 0.25$ and $p^* = 0.001$.

To evaluate how well the posterior probability found using the $-2 \log p$ method and the differentially private posterior probability approximate the posterior distribution using the full dataset, I take the differences between these values found in

one dataset. In Figure 3.4, I plot

$$p(b) - 1 - F_{\chi^2_{2m}}\left(-2\sum_{j=1}^{M}\log(p_j(b))\right) \quad \text{and} \quad p(b) - LM_p. \tag{3.19}$$

We see that there are bigger variances in differences when the cutoff value is closer to the true value of the regression coefficient. There is no other clear pattern of underestimating or overestimating. Overall, the differences are small and confirm we are getting reasonably accurate estimates.

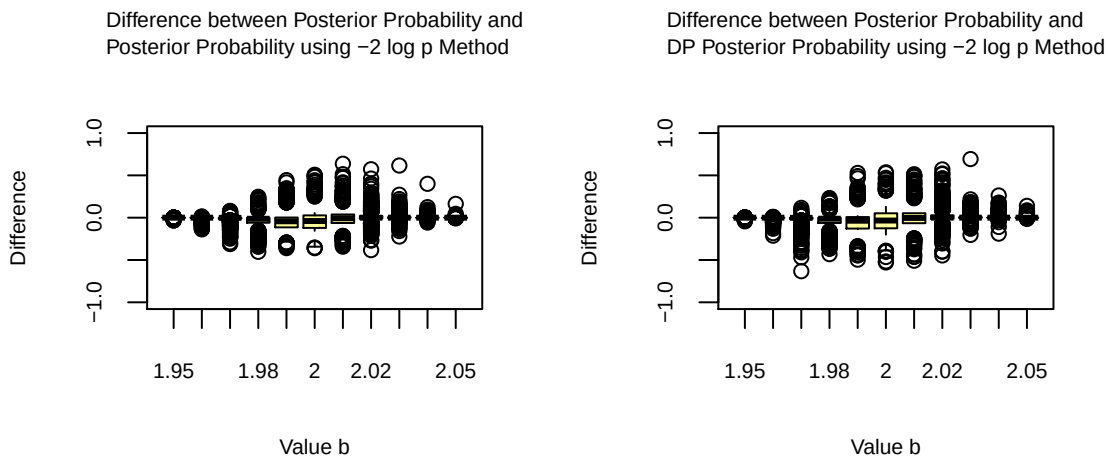


FIGURE 3.4: These graphs in order are: The distribution of the difference within one simulated dataset between the posterior probability given the full dataset and the posterior probability found using the $-2\log p$ method where $p^* = 0.001$, and the distribution of the difference within one simulated dataset between the posterior probability given the full dataset and the posterior probability found using the $-2\log p$ method with noise added to ensure differential privacy where $\epsilon = 0.25$ and $p^* = 0.001$.

# 4

# Reporting Differentially Private Posterior Quantiles

If the analyst had access to an entire confidential dataset $\mathcal{D}$, he or she would be interested in the posterior quantile of a regression coefficient that is $Q(p) = \{b : \Pr(\beta \leqslant b|\mathcal{D}) = p\}$.

To use the sample and aggregate technique, I adapt the method from [Cheng, et al. 2017] to report posterior quantiles on partitioned data. The posterior distribution within a partition is found with the likelihood function raised to the power of $M$, the number of partitions. This serves to rescale the variance of the posterior distribution within a partition to match that of the posterior distribution using the full dataset. A quantile is found on each posterior distribution within one partition, and then averaged to find an estimate of the posterior quantile found on the whole dataset.

Let $\mathcal{D}$ be a confidential dataset made up of $n$ observations $\{y_i, \boldsymbol{x}_i\}$ where $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\boldsymbol{x}_i$ being some fixed $p \times 1$ vector of fixed covariates.

Given only a partition of the data $\mathcal{D}_j = \{\boldsymbol{y}_j, \boldsymbol{X}_j\}$, let

$$p^*(\beta|\mathcal{D}_j) \propto \left[\prod_{i=1}^{n/M} p(Y_{j,i}|\beta)\right]^M p(\beta) \tag{4.1}$$

and let our intermediate outcomes be

$$h(\mathcal{D}_j) = Q_j^*(p) = \{b : \Pr\left(p^*(\beta|\mathcal{D}_j) \leqslant b\right) = p\}. \tag{4.2}$$

Our aggregation function is

$$g(h(\mathcal{D}_1), \ldots, h(\mathcal{D}_M)) = \sum_{j=1}^{M} Q_j^*(p). \tag{4.3}$$

The quantiles are unbounded and thus have indeterminate global sensitivity. To derive the bounds on quantiles found within one partition, I use an adaptation of the sparse vector technique [Dwork & Roth 2014] [Chen, et al 2016]. This will use use $\epsilon_1$ of the privacy budget. Given a list of the quantiles, the algorithm iterates through bounds $[-\mu \cdot 2^t, \mu \cdot 2^t]$ for $t = 0, 1, 2, \ldots$ until it finds $t$ where a noisy number of observations in the bounds $\tilde{q}_t = q_t + \eta_1$, where $\eta_1$ is randomly sampled from a Laplace distribution with scale parameter $4/\epsilon_1$, exceeds a noisy fraction $\theta$ of the data $\tilde{N} = \theta \cdot n + \eta_2$, where $\eta_2$ is randomly sampled from a Laplace distribution with scale parameter $2/\epsilon_1$. Quantiles found within a partition that are outside the found bounds are clipped to be within the bounds $(-b, b)$. The global sensitivity of the clipped partition quantiles are then $2b$, and using the sample and aggregate technique, the global sensitivity of the mean of the clipped quantiles found within a partition are $\frac{2b}{M}$. The remaining privacy budget $\epsilon_2 = \epsilon - \epsilon_1$ is in the Laplace mechanism.

The current algorithm can result in wide bounds for the quantiles found using data within a partition since the bounds are centered around zero, while the quantiles are not necessarily centered around zero. While these wide bounds can be used, they

potentially inject too much noise in the Laplace Mechanism, or require us to increase the number of partitions, the size of the dataset, or the privacy budget. Future work will derive a method to use part of the privacy budget to find the center of the quantiles in order to center the bounds correctly, which will minimize the global sensitivity and the Laplace noise added.

The differentially private statistic released to the analyst is

$$LM_q = \frac{1}{M} \sum_{j=1}^{M} Q_j^*(p) + \eta \tag{4.4}$$

where $\eta$ is a draw from the Laplace distribution with mean zero and scale parameter $\frac{2b}{M\epsilon_2}$.

## 4.1 Evaluation

To evaluate the utility and accuracy of our method, we want to ensure that the differentially private posterior quantile released to the user is approximately the posterior quantile found using the full dataset. It is useful to also confirm that the posterior quantile found using [Cheng, et al. 2017] is approximately that found using the entire dataset. We want

$$q(p) \approx \sum_{j=1}^{M} Q_j^*(p) \approx LM_q \tag{4.5}$$

for any confidential dataset used. I will show this by using calibration probabilities [Rubin 1984] that show, under any dataset, that the expected value of the posterior quantile, the expected value of the posterior quantile using [Cheng, et al. 2017], and the expected value of the differentially private posterior quantile are all approximately the same,

$$\mathrm{E}\left[\int q(p) d\mathcal{D}\right] \approx \mathrm{E}\left[\int \sum_{j=1}^{M} Q_j^*(p)\, d\mathcal{D}\right] \approx \mathrm{E}\left[\int LM_q d\mathcal{D}\right]. \tag{4.6}$$

To show this empirically, I simulate artificial datasets from $y_i = 1 + 2x_i + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 100)$ of size 100000 and split the dataset into partitions of size 1000. A larger dataset, larger partition size, and larger error variance were necessary due to the wide bounds found around the quantiles. I used a semi-conjugate prior to find the posterior distribution of the slope coefficient using the full dataset and within partitions of the dataset but with the likelihood raised to the power of $M$. I found the posterior quantile using the full dataset, using the partitioned data, and using the partitioned data where noise is added to preserve differential privacy. The privacy budget used to find the bounds was $\epsilon_1 = 0.45$ and the privacy budget to add noise was $\epsilon_2 = 0.45$. Both of these values were used because they yielded good results. The distribution of the posterior quantiles are shown in 4.1 for various probability values $p$. The means of the distributions are similar across different values of $p$, and the curve made when increasing $p$ is the inverse of a normal cumulative distribution function, which we expect since the regression coefficient is distributed approximately normally.

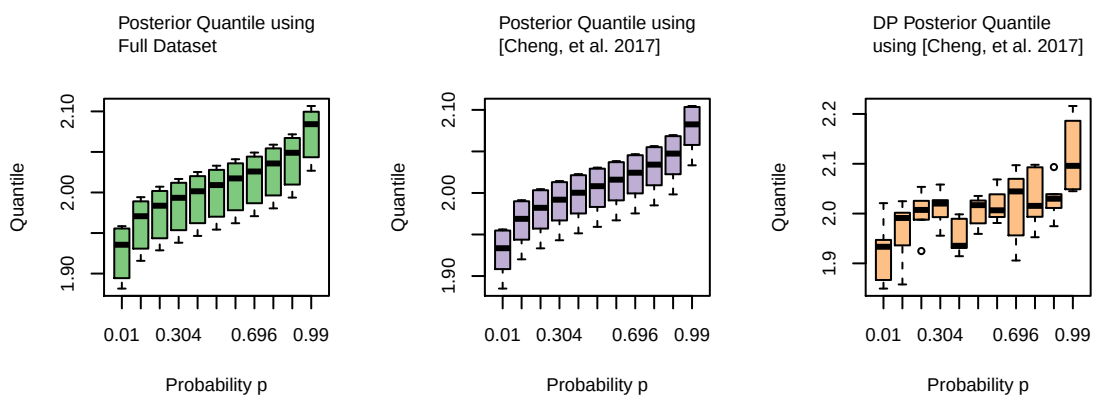


FIGURE 4.1: These figures in order are: the distributions of the posterior quantile found for various probability values $p$ using the full dataset, the distributions of the posterior quantile found for various probability values $p$ using [Cheng, et al. 2017] on partitioned data, and the distributions of the posterior quantile found for various probability values $p$ on partitioned data using [Cheng, et al. 2017] with noise added to ensure differential privacy where $\epsilon_1 = 0.45$ and $\epsilon_2 = 0.45$.

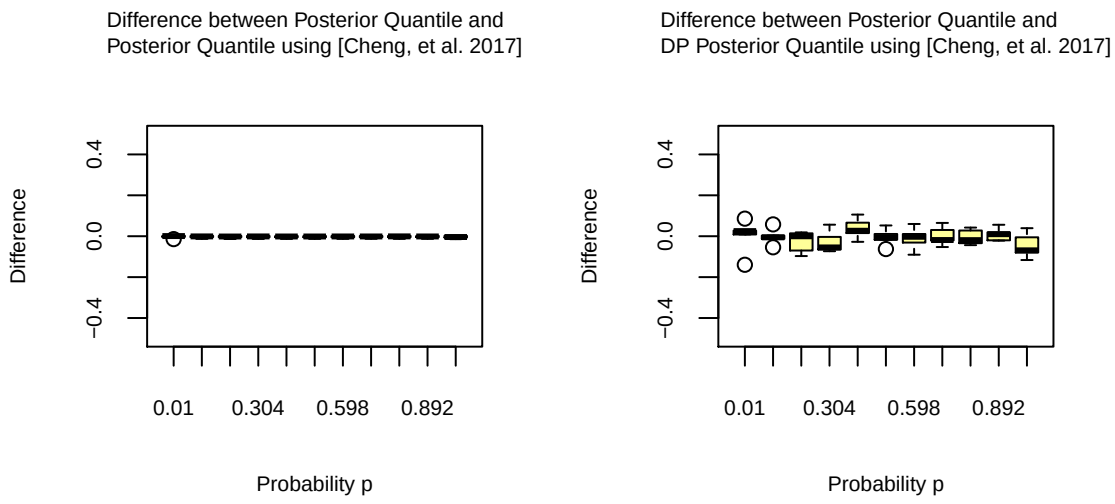


FIGURE 4.2: The graphs in order are: The distribution of the difference within one simulated dataset between the posterior quantile using unpartitioned data and the posterior quantile using partitioned data, and the distribution of the difference within one simulated dataset between the posterior quantile using unpartitioned data and the posterior quantile using partitioned data with Laplace noise added to preserve differential privacy where $\epsilon_1 = 0.45$ and $\epsilon_2 = 0.45$.

To evaluate how well the posterior quantile found using [Cheng, et al. 2017] and the differentially private posterior quantile approximate the posterior quantile using the full dataset, I take the difference between these values found in one dataset. In Figure 4.2, I plot

$$q(p) - \frac{1}{M}\sum_{j=1}^{M} Q_j^*(p) \quad \text{and} \quad q(p) - LM_q \tag{4.7}$$

for each of the datasets I simulated. We see between the posterior quantile using the full dataset and the posterior quantile using [Cheng, et al. 2017] that there are no clear pattern in the differences. Between the posterior quantile using the full dataset and the differentially private posterior quantile, there is more noise, which is expected because of the wide bounds on posterior quantiles. Overall, the differences are small and we are getting reasonably accurate differentially private quantiles.

# 5

# Conclusion

I have presented methods for releasing summary statistics on posterior distributions of regression coefficients while preserving differential privacy. To release differentially private posterior probabilities, the normality method works well, but was limited due to the assumptions made on the variances of the posterior distributions. The $-2\log p$ method has been shown empirically to work well, but lacks current theoretical justification. Future work will focus on easing the restrictions of the normality method and proving the effectiveness of the $-2\log p$ method. The differentially private method to report quantiles is effective and theoretically sound, but is limited due to the need to find bounds on the quantiles. Future work will use a fraction of the privacy budget to estimate the center of the quantiles and find tighter bounds.

# Bibliography


[Chen, et al 2016] Chen, Y., Barrientos, A. F., Machanavajjhala, A., & Reiter, J. P. (2016). Differentially Private Regression Diagnostics. *2016 IEEE 16th International Conference on Data Mining (ICDM).*

[Cheng, et al. 2017] Li, C., Srivastava, S., & Dunson, D. B. (2017). Simple, Scalable and Accurate Posterior Interval Estimation. *Biometrika*, 104(3), 665-680.

[Barrientos, et. al 2017] Barrientos, A. F., Reiter, J. P., Machanavajjhala, A., & Chen, Y. (2017). Differentially private significance tests for regression coefficients.

[Dwork et al. 2006] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference '06*, 265284.

[Dwork & Roth 2014] Dwork, C., and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy.* Now Publ.

[Fisher 1932] Fisher R. A. (1932).*Statistical Methods for Research Workers.* London: Oliver and Boyd.

[Homer et al. 2008] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genetics* 4(8): e1000167. https://doi.org/10.1371/journal.pgen.1000167

[Nissim, et al. 2007] Nissim, K., Raskhodnikova, S., & Smith, A. (2007). Smooth Sensitivity and Sampling in Private Data Analysis. *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing - STOC 07.*

[Rubin 1984] Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4), 1151-1172.

[Sweeney 2013] Sweeney, L. (2013). Matching known patients to health records in Washington state data. Tech. rep. Data Privacy Lab, Harvard University.