Bayesian Inference of Differentially Private Datasets in Linear Regression Models

by

Yangdi Jiang

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

The demand for large dataset and demand of privacy protection are in constantly conflicts as the balance between the two is hard to keep. Differential privacy is a mathematical rigor definition that provides the balance bewteen these two opposite sides. It's developed with the purpose of making privacy-preserving analysis/inference [9]. Ever since the introduction of differential privacy, the literatures around it have been flourishing. However, the methodologies of statistical analysis and inference given the differentially private dataset are not much studied. In this thesis, we will tackle this problem using Bayesian method from the perspective of measurement error problems. Our simulation study shows that it outperforms the existing method (SIMEX) when applying to a similar specification of problem. Additionally, we will investigate briefly the question whether it's beneficial to generate multiple DP datasets.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Bei Jiang, whose expertise was invaluable in formulating the research questions and methodology. Thank you for your patience and support in helping me complete this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The demand for large dataset has been growing ever since the explosive growth of computer power from the beginning of this century. Along with the demand for large data, the privacy issue has become a increasing concern. These two opposite demands drive the development of differential privacy. On the one hand, Too much privacy protection or rather when privacy protection is not applied efficiently, the data will stop being useful. On the other hand, too little privacy protection, private information of individual is in risk of being discovered. Differential privacy provides a mathematical rigor definition suited towards the privacy-preserving analysis [9].

Differentially private has traditionally been applied to aggregate/summary statistics, and this has many shortcomings. First of all, as the data curator, the need to provide differentially private data for queries submitted constantly until the privacy budget runs out is not ideal. Secondly, since the privacy budget is pre-specified and finite, once it runs out, the data curator can no longer answer additional queries. Lastly, as the analyst/statistician, aggregate/summary statistics are often not enough to work with. To overcome these issues, attentions have been turn into the direction of differentially private data synthesis. That is, to generate/synthesize a dataset that is differentially private.

Even through the literatures on differential privacy have been flourishing, the methodology of statistical analysis and inference on differentially private data has hardly been studied. To elaborate, as a analyst/statistician, suppose

you are given a differentially private aggregate/summary statistic. How would you go about analyzing this statistic? Should you take the naive approach, that is, treat the dataset as if no privacy protection is added? Or should you develop a approach that takes the privacy protection into consideration? In turns out, the naive approach performs poorly comparing to the approach that considers the privacy protection [3]. Furthermore, how should you analyze a differentially private dataset? In this thesis, we will develop Bayesian inference for differentially private dataset in linear regression.

This thesis is organized as follow,

1. In chapter 2, we will discuss all the related works includes differential privacy, differentially private dataset synthesis, existing methods of inference on differentially private data and Bayesian inference.

2. In chapter 3, we will introduce the main method of this thesis. Inspired by the Bayesian method used in measurement error problem, we apply the Bayesian inference on the setting of differentially private dataset. Forthermor, the method will be extended to multiple differentially private dataset setting.

3. In chapter 4, we will discuss the setup and the result for our simulation study to evaluate the performance of the Bayesian inference and to compare using single differentially private dataset with using multiple differentially private datasets.

4. In chapter 5, we will end with a brief conclusion and directions on the future works.

Chapter 2

Related works

In this chapter, we will discuss all the relevant works in details. In section 2.1, we will give a overview of differential privacy, which includes the definition, a hypothesis testing point of view to help understanding the definition, two important properties of differential privacy and two mechanisms to create differentially privacy randomized algorithm. In the next section, we will focus on differentially private data synthesis, which is different from the traditional release of differentially private aggregate statistics. Specifically, MODIPS and Laplace mechanism for dataset synthesis will be considered. Moving on to section 2.3, we will review the existing method for making inference on differentially private data. Lastly, we will end this chapter with an overview of Bayesian method and Markov Chain Monte Carlo simulation in section 2.4.

2.1 Differential privacy

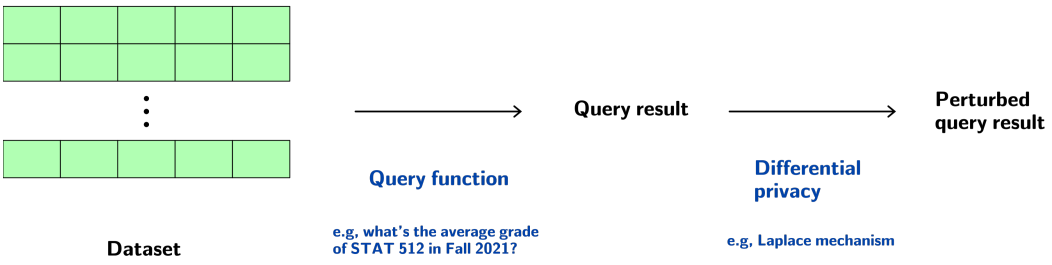

Figure 2.1: Illustration of differential privacy

Differential privacy is mainly concerned about a dataset and a specific

query. Suppose you are a dataset holder, and some outside party have some query regarding the dataset for one reason or another. As an example, a query could be "what's the average grade of the STAT 566 in Fall 2021?" or "What's the percentage of people that are full vaccinated returning to campus this Fall?". Simply providing the precise answer to these queries could pose privacy concerns as the Fundamental Law of Information Recovery states that overly accurate answers to too many questions will destroy privacy in a spectacular way [9]. Therefore, it's necessary to add some perturbation to the query result for privacy measure. However, how much perturbation is enough? Too little perturbation does not provide enough privacy protection. Too much perturbation would render the query result entirely useless. This is where differential privacy comes into play. It quantifies the privacy protection and provides guidance on how should we perturb the query result for given amount of privacy budget.

2.1.1 Why differential privacy?

[9] describes differential privacy as follow,

> "Differential privacy" describes a promise, made by a data holder, or curator, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available."

This quote highlights the main idea of the differential privacy, that is, it provides a balance between privacy and data utility. Its concept centers around the goal of making insightful inference regarding a population while learning nothing about any individual within that population. In other words, whether an individual (record) resides in the population (dataset) or not should not make a impact on the inference we make on the population (dataset). It is precisely this idea gives differential privacy one of the more appealing advantage over other privacy-preserving approach. That is, differential privacy prevent

linkage attack [9], which is referred to adversary using auxiliary information to identify an individual in a private dataset.

2.1.2 Definition of differential privacy

Before we introduce the definition of differential privacy, we need to go over some preliminary definitions.

Definition 1 (**Neighbouring datasets**)**.** Two datasets D and D' are said to be **neighboring datasets** iff they are differ only in one record/row.

Definition 2 (**Probability Simplex** [9])**.** Given a discrete set B, the probability simplex over B, denoted $\Delta(B)$ is defined to be:

$$\Delta(B) = \left\{ x \in \mathbb{R}^{|B|} : x_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{|B|} x_i = 1 \right\}$$

Definition 3 (**Randomized Algorithm** [9])**.** A randomized algorithm $\mathcal{M}$ with domain A and discrete range B is associated with a mapping $\mathcal{M} : A \to \Delta(B)$. On input $a \in A$, the algorithm $\mathcal{M}$ outputs $\mathcal{M}(a) = b$ with probability $(M(a))_b$ for each $b \in B$. The probability space is over the coin flips of the algorithm $\mathcal{M}$.

Now we are ready to introduce the formal definition for differential privacy.

Definition 4 (ϵ**-differential privacy [8]**)**.** A randomized algorithm $\mathcal{M}$ is ϵ**-differentially private** if for all neighborhood datasets D and D' and all possible result subset (event) Q, we have

$$\Pr(\mathcal{M}(D) \in Q) \leq e^{\epsilon} \cdot \Pr(\mathcal{M}(D') \in Q)$$

where ϵ is commonly referred as the privacy budget.

There exists a natural extension of ϵ-differential privacy, but we will mainly use ϵ-differential privacy throughout. For completeness, it's stated below.

Definition 5 ((ϵ, δ)**-differential privacy**)**.** A randomized algorithm $\mathcal{M}$ is (ϵ, δ)**-differentially private** if for all neighborhood datasets D and D' and all possible result subset (event) Q, we have

$$\Pr(\mathcal{M}(D) \in Q) \leq e^{\epsilon} \cdot \Pr(\mathcal{M}(D') \in Q) + \delta$$

2.1.3 Hypothesis testing perspective

Let S and S' be two neighboring datasets. Denote the outcome of the random mechanism by Y, that is, the perturbed query result. Once Y is observed, we are dealing with the following hypothesis testing problem,

$$\boxed{\begin{array}{r} H_0 : \text{the underlying dataset is } D \\ H_\alpha : \text{the underlying dataset is } D' \end{array}}$$

Or equivalently, let P denotes the distribution of $\mathcal{M}(D)$ and Q denotes the distribution of $\mathcal{M}(D')$

$$\boxed{\begin{array}{l} H_0 : \text{the underlying distribution of } Y \text{ is } P \\ H_\alpha : \text{the underlying distribution of } Y \text{ is } Q \end{array}}$$

This is a simply (simply in the sense that both null hypothesis and alternative hypothesis consist of singleton) hypothesis testing problem. Due to Neyman-Pearson lemma, we know the most powerful test for this class of problem is the likelihood ratio test. That is, reject H_0 if $y \in C$, where

$$C = \left\{ y : \frac{\mathcal{L}(P; y)}{\mathcal{L}(Q; y)} \leq c \right\}$$

For a likelihood ratio test of size α, we have

$$\begin{aligned} & \mathbb{P}_P(Y \in C) = \alpha \\ \iff & \Pr(\mathcal{M}(D) \in C) = \alpha \end{aligned}$$

It follows that if $\mathcal{M}$ is (ϵ, δ)-DP, then we have

$$\begin{aligned} & \Pr(\mathcal{M}(D') \in C) \leq e^\epsilon \Pr(\mathcal{M}(D) \in C) + \delta \\ \iff & \mathbb{P}_Q(y \in C) \leq e^\epsilon \mathbb{P}_P(y \in C) + \delta \\ \iff & \mathbb{P}_Q(y \in C) \leq e^\epsilon \alpha + \delta \end{aligned}$$

That is, the power of the likelihood ratio test of size α is bounded above by $e^\epsilon \alpha + \delta$. Since the likelihood ratio test is the most powerful hypothesis test, all hypothesis tests of size α are bounded above by $e^\epsilon \alpha + \delta$. It follows that the smaller the ϵ and δ, the smaller the power upper bound will be. In other words, **the smaller the ϵ and δ, the harder to distinguish two neighbouring datasets given the perturbed query result** y.

2.1.4 Properties of differential privacy

A very important property of differential privacy is the post-processing property. Informally, this property says that any result obtained from the output of a ϵ-DP randomized algorithm, without the additional information from the original dataset, will be ϵ-DP as well.

Theorem 1 (**Post-processing property** [9])**.** Let $\mathcal{M}$ be a randomized algorithm that is (ε, δ)-differentially private. Let f be an arbitrary randomized mapping whose domain is a superset of the range of the randomized algorithm $\mathcal{M}$. Then $f \circ \mathcal{M}$ is (ε, δ) differentially private randomized algorithm.

Another important of differential privacy is the composition property. That is, the composition of two (ϵ, δ)-differentially private mechanisms is $(2\epsilon, 2\delta)$-differentially private. Formally, it's stated as below,

Theorem 2 (**Composition theorem [14]**)**.** Let $\{M_j\}$ be a sequence random mechanism that satisfies ϵ_j-DP, then the sequence $M_j(D)$ corporate on the same dataset D is, as a whole, a random mechanism that satisfies $\sum_j \epsilon_j$-DP.

This important property will be used many times throughout this report, especially in section 2.2.

2.1.5 Mechanisms of differential privacy

Laplace Mechanism

One of the most commonly used differential privacy method is the Laplace mechanism. It fulfills ϵ-DP by adding a certain size, depending on the size of ϵ, of Laplace noise to the query result.

Definition 6 (**Sensitivity**)**.** Let $f : \mathcal{D} \to \mathbb{R}^k$ be a query function. The sensitivity w.r.t f is defined as,

$$\Delta_f = \max_{D,D'} \|f(D) - f(D')\|_1$$

where D and D' are neighboring datasets.

Definition 7 (**Laplace Mechanism**)**.** A randomized algorithm M satisfies the ϵ-DP w.r.t the query function f if $M^{(i)}(D) \sim \text{Lap}(f^{(i)}(D), \frac{\Delta_f}{\epsilon})$ for $i = 1, 2, \ldots, k$, where $M^{(i)}$ and $f^{(i)}$ denote the i-th component of M and f respectively.

In other words, the Laplace mechanism perturb the query result with a Laplace noise of size $\frac{\Delta_f}{\epsilon}$.

Example 1 (**Counting query** [9])**.** Let consider the simply query, "how many records/rows in a dataset satisfies some specific property?". This type of query is referred as a counting query, and it's simplistic as its sensitivity will always be 1 as a single row difference will result in at most 1 count difference. It follows that to apply the Laplace mechanism is as simply as adding a Laplace noise with the scale parameter $\frac{1}{\epsilon}$.

Since the sensitivity Δ_f determines how much perturbation is required for a fixed amount of privacy budget ϵ, the implicit assumption that the sensitivity Δ_f is finite is made for Laplace mechanism to work appropriately. Is this assumption realistic? How should we deal with the potentially infinite/unbounded sensitivity?

Dealing with unbounded sensitivity

We will start by addressing the first question, is the assumption realistic? This is discussed in [12] and [4], and I will reiterate the point here briefly. The argument is that for most of the variables, their values are naturally bounded. For example, the human height, income and etc. This explanation seems fairly reasonable, however, it's not the case in simulation study as the frequent usage of gaussian distribution and other distributions with unbounded support.

This leads to the second question, how should we deal with unbounded variable? This is discussed in [2] and [16]. In our discussion, we will follow [2] and truncate the support of unbounded variables to a reasonable interval. For example, we would truncated a standard normal random variable to the interval $(-1.96, 1.96)$. It follows that the sensitivity is calculated using the interval.

Exponential mechanism

One shortcoming of the Laplace mechanism, aside from the potentially infinite sensitivity discussed above, is that it only applies to numerical data. Differ from the Laplace mechanism, the exponential mechanism does not add noise directly to the query result, thus it applies to all types of data.

Definition 8 (**Exponential Mechanism** [13])**.** In the Exponential mechanism, a utility function u assigns a score to each possible output $\mathbf{s}^*$ and releases $\mathbf{s}^*$ with probability

$$\frac{\exp\left(u\left(\mathbf{s}^* \mid D\right)\frac{\epsilon}{2\Delta_u}\right)}{\int \exp\left(u\left(\mathbf{s}^* \mid D\right)\frac{\epsilon}{2\Delta_u}\right) d\mathbf{s}^*}$$

to ensure ϵ-DP, where

$$\Delta_u = \max_{D,D'} \left|u\left(\mathbf{s}^* \mid D\right) - u\left(\mathbf{s}^* \mid D'\right)\right|$$

is the sensitivity of the score function u. Note, if $\mathbf{s}^*$ is discrete, the integral is replaced with summation.

However, we will only works with numerical data in this report. Therefore, no further discussion will be made regarding the exponential mechanism.

2.2 Differentially private dataset synthesis (DIPS)

In this section, we will consider synthesize a differentially private dataset, which is referred as DIPS (differentially private data synthesis) in [4], instead of generating differentially private aggregate statistics to answer queries submitted to data curator. As explained in [4], DIPS addressed one shortcoming of releasing aggregate/summary statistics, that is, since the privacy budget is often pre-specified, the data curator cannot answer any further query once the privacy budget is exhausted by answering a number of queries. DIPS bypasses this issue by release the dataset directly to perform statistical analysis/inference. Due to the post-pressing of differential privacy, any future query can be obtained using the synthesized dataset.

We will focus on two DIPS algorithms, binary datasets synthesis mechanism proposed in [1] and the model-based differentially private data synthesis (MODIPS) algorithm developed in [12].

2.2.1 Binary data synthesis

As mentioned before, we will consider a binary dataset, $x_{1:n} = (x_1, \ldots, x_n)$ with $x_i \in \{0, 1\}$ for $i = 1, \ldots, n$. As usual, a binomial distribution is assumed for the data, and thus it's sufficient to represent the data $x_{1:n}$ by its sufficient statistic $x = \sum_i^n x_i$. The synthesis algorithm is to sample,

$$\tilde{p} \sim \text{Beta}(\alpha + x, \alpha + n - x)$$
$$\tilde{x} \sim \text{Binomial}(\tilde{n}, \tilde{p})$$

where the parameters α must satisfy $\alpha \geq \frac{\tilde{n}}{\exp(\epsilon)-1}$ to fulfill ϵ-DP. $\tilde{x}$ is the synthetic dataset that will be released. Note $\tilde{n}$ can be different from the original sample size n if we want to keep the sample size private as well.

Using the composition theorem, we can also extend this algorithm to generate multiple synthetic datasets,

$$\tilde{p}_m \sim \text{Beta}(\alpha + x, \alpha + n - x)$$
$$\tilde{x}_m \sim \text{Binomial}(\tilde{n}, \tilde{p}_m)$$

for $m = 1, \ldots, M$, where $\alpha \geq \frac{\tilde{n}}{\exp(\epsilon/M)-1}$

[6] provided the following insight regarding this synthesis mechanism,

> We can interpret this synthetic data generation process as generating from a perturbed posterior predictive distribution.

This interpretation connects nicely to the next synthesis algorithm, MODIPS, which is also based on sampling from a perturbed posterior predictive distribution.

2.2.2 MODIPS

Although the binary data synthesis mechanism works fine, the limitation to binary data is far too restrictive. For applying to a more general data, we

will discuss another model-based synthesis method, MODIPS. Using the post-processing property of differential privacy, MODIPS [12] synthesize differentially private dataset using posterior sampling. Similar to [2] and [3], differential privacy is applied to the sufficient statistics s. Sufficient statistics is important here since

$$p(\theta \mid x_{1:n}) = p(\theta \mid s)$$

That is, the posterior distribution given the full data $x_{1:n}$ is the same as the posterior distribution given the sufficient statistics s. Denote the perturbed (differentially private) sufficient statistics as y, then it follows that $p(\theta \mid y)$ is also differentially private by the post-processing property. Similarly, if we sample $\tilde{\theta}$ from $p(\theta \mid y)$, it will also be differentially private. Lastly, if we sample a synthetic dataset $\tilde{x}_{1:n}$ from $f(x \mid \tilde{\theta})$, then the dataset will be differentially private. Refer to figure 2.2 for the illustration of this process.

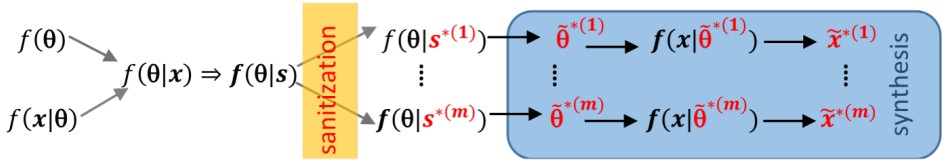

Figure 2.2: Illustration of MODIPS. Image from [12]

2.2.3 Laplace mechanism for dataset synthesis

MODIPS at its core is a model-based synthesis, and thus it requires a likelihood model. In practice, the specification of a likelihood model introduces additional uncertainty, which might not be necessary sometimes. In this section, we will introduce a more simplistic and model-free approach. That is, we will simply extend the Laplace mechanism to dataset generation.

Definition 9 (**Sensitivity for dataset**)**.** The sensitivity w.r.t dataset generation is defined as,

$$\Delta_f = \max_{D,D'} \max_i \|d_i - d_i'\|_1$$

where D and D' are neighboring datasets and d_i & d_i' denote their i-th rows.

Algorithm 1: MODIPS algorithm

Input:

- m: number of released data sets
- ϵ: overall privacy budget
- s: sufficient statistics in the Bayesian model assumed on original data $x_{1:n}$

For $k = 1, 2, \ldots, m$,

1. sanitize s via a differentially private mechanism with privacy budget ϵ/m to generate $s^{(k)*}$
2. draw $\theta^{(k)*}$ from the sanitized posterior distribution $f\left(\theta \mid s^{(k)*}\right)$
3. draw $\tilde{x}_{1:n}^{(k)*}$ from $f\left(x_{1:n} \mid \theta^{(k)*}\right)$.

Output: datasets: $\tilde{x}_{1:n}^{(1)*}, \ldots, \tilde{x}_{1:n}^{(m)*}$

Definition 10 (**Laplace mechanism for dataset**)**.** A randomized algorithm M satisfies the ϵ-DP w.r.t dataset generation f if $M^{(i,j)}(D) \sim \text{Lap}(d_{i,j}, \frac{\Delta_f}{\epsilon})$, where $M^{(i,j)}$ denote the (i,j)-th component of M and $d_{i,j}$ denotes the (i,j)-th element of D.

2.3 Inference on differentially private data

Over the last decade, there are many literatures regarding differential privacy. However, literatures on how to analyze and make inference on differentially private data are few and far bewteen. In this section, we will focus on 4 such papers. Two of these, [2] and [3], are about inferencing on differentially private sufficient statistics, and the other two, [6] and [7], are about analysis on differentially private datasets. Note, Bayesian methods are used in many parts of this section. For more details on Bayesian methods, refer to section 2.4.

2.3.1 Inference on DP sufficient statistics

Both [2] and [3] discusses DP Bayesian inference within the setting of exponential family as the sufficient statistics are easy to compute for exponential family. Since Bayesian inference is utilized, we are interested in obtain the posterior distribution.

Denote the data as $x_{1:n}$, then the density of the data is of the form,

$$p(x_{1:n} \mid \eta) = h(x_{1:n}) \exp\left(\eta^\top t(x_{1:n} - nA(\eta))\right)$$

where η is the natural parameter, and $t(x_{1:n})$ is the sufficient statistics, denote it as $s = t(x_{1:n})$. To ensure privacy protection, Laplace mechanism is applied to the sufficient statistics s, denote it by y, we have

$$y \sim \text{Lap}(s, \frac{\Delta_s}{\epsilon})$$

Denote the parameter of interested as θ, and it follows that we are interested in the posterior distribution $p(\theta \mid y)$, which can be written as,

$$\begin{aligned} p(\theta \mid y) &\propto p(\theta, y) \\ &= \int p(\theta, s, y) ds \\ &= \int p(\theta) p(s \mid \theta) p(y \mid s) \end{aligned}$$

Gibbs sampler is used to sample from this distribution, and as mention in the paper, the main difficulties centered around $p(s \mid \theta)$ and the full conditional distribution for s, $p(s \mid \theta, y)$.

In general, there no exact form of $p(s \mid \theta)$. To bypass this issue, [2] uses central limit theorem and the nice properties of exponential families to derive a normal approximation for $p(s \mid \theta) \approx \mathcal{N}(\mu, \Sigma)$ where

$$\mu = \mathbb{E}[t(x)] = \frac{\partial}{\partial \eta^\top} A(\eta), \quad \Sigma = \text{Var}[t(x)] = \frac{\partial^2}{\partial \eta \partial \eta^\top} A(\eta)$$

Gibbs sampler is based on sampling from the full conditional distributions, and thus the conditional distribution $p(s \mid \theta, y)$ is required. To obtain $p(s \mid \theta, y)$, the simply trick that Laplace distribution can be written as a scale mixture of normal distributions is used (For more details, refer to A.3).

Algorithm 2: Gibbs sampler for noisy sufficient statistic

Initialize θ, s, σ^2

repeat

- $\theta \sim p\left(\theta; \lambda'\right)$ where $\lambda' =$ Conjugate-Update (λ, s, n).
- Calculate $\mu = \mathbb{E}[s]$ and $\Sigma = \mathrm{Var}[s]$
- $s \sim$ NormProduct $\left(n\mu, n\Sigma, y, \mathrm{diag}\left(\sigma^2\right)\right)$
- $1/\sigma_j^2 \sim$ InverseGaussian $\left(\frac{\epsilon}{\Delta_s |y - s|}, \frac{\epsilon^2}{\Delta_s^2}\right)$

Algorithm 3: NormProduct

input: $\mu_1, \Sigma_1, \mu_2, \Sigma_2$

$$\Sigma_3 = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$
$$\mu_3 = \Sigma_3(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$$

return: $\mathcal{N}(\mu_3, \Sigma_3)$

For completeness, the Gibbs sampler is listed below,

With all the details above, [2] develops a general algorithm that applies to any exponential family with bounded sufficient statistics. (Note that, a extension to unbounded sufficient statistics is also provided in [2]). To go a step further, applying the general framework in [2], [3] developed differentially private Bayesian linear regression. [3] demonstrated that the naive approach of ignoring the perturbation noise of the differential privacy mechanism is far inferior in realistic data settings comparing to the noise-aware methods that it developed as the noise-aware methods produce correct posteriors during a wide range of scenarios.

2.3.2 Analysis DP synthetic dataset

subsubsectionDP analysis on binary data

[6] is one of the first, if not the first, papers discussing how should one go about analyze differentially private dataset. It focused on binary data since the only algorithm (discussed in section 2.2.1) for generating differentially

private synthetic dataset at the time is for count data [1]. Given the synthetic dataset, [6] utilize the Bayesian method and make inference using the posterior distribution of the parameter of interest p.

As discussed in 2.2.1, the data generation model is,

$$\begin{aligned} x &\sim \text{Binomial}(n, p) \\ \tilde{p}_m &\sim \text{Beta}(\alpha + x, \alpha + n - x) \\ \tilde{x}_m &\sim \text{Binomial}(\tilde{n}, \tilde{p}_m) \end{aligned}$$

where x is the original dataset, which is not available to the analyst, and $\tilde{x}_1, \ldots, \tilde{x}_m$ are the synthetic differentially-private dataset, which is available to the analyst. Note that α and n are assumed known to the analyst.

To simplify the posterior computation, a conjugate prior is chosen, $p \sim \text{Beta}(\gamma_1, \gamma_2)$. There is no closed form for the posterior distribution of p, and thus we sample from the posterior distribution using MCMC. More specifically, to update x, a Metropolis-Hastings step can be used, and to update p and $\{tidlep_m\}_{m=1}^M$, the following simple Gibbs sampler can be used,

$$\begin{aligned} p \mid x, \tilde{p}, \{\tilde{x}_m\}_{m=1}^M &\sim \text{Beta}\left(\gamma_1 + x, \gamma_2 + n - x\right) \\ \tilde{p}_m \mid x, \tilde{x}_m, p &\sim \text{Beta}\left(\alpha_1 + \tilde{x}_m + x, \alpha_2 + \tilde{m} - \tilde{x}_m + n - x\right) \quad \text{for } m = 1, \cdots, M \end{aligned}$$

[6] demonstrated that by taking the data generation model/mechanism into consideration, it allows for accurate estimation for the parameter of interest p for moderately large privacy budget ϵ.

Simulation extrapolation (SIMEX)

Originally employed in measurement error problems, SIMEX is a method based on extrapolation as the name suggested. The idea is to introduce an additional parameter ζ. The naive estimator $\hat{\Theta}_{naive}$ is computed at $\zeta = 0$. By adding another perturbation noise to the predictor and computing the least square estimator, we obtain another estimator at $\zeta = 1$. Repeat the process, we obtain multiple estimators at a variety values of ζ. Given these estimators, a curved is fitted and by extrapolate the curve to $\zeta = -1$, we obtained the SIMEX estimator. For more details on SIMEX, refer to chapter 5 of [5].

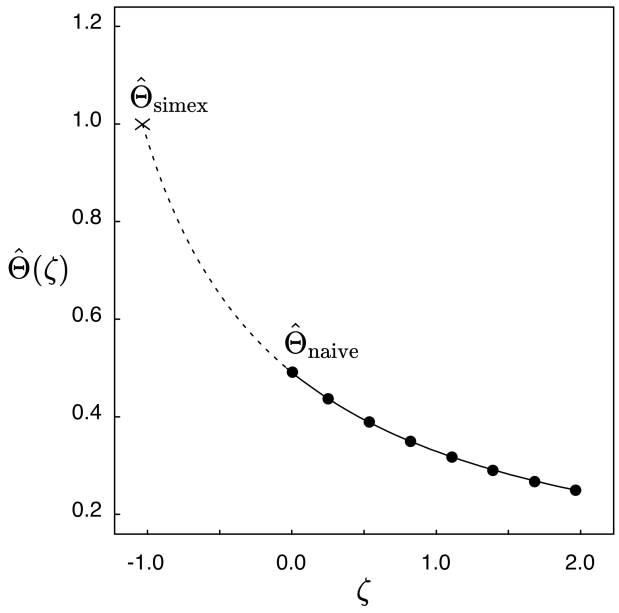

Figure 2.3: The SIMEX estimate is an extrapolation to $\zeta = -1$. The naive estimate occurs at $\zeta = 0$. Image from Page 99, Section 5.2 of [5].

As in the measurement error setting, usually only the predictor is perturbed and the perturbation noise is usually gaussian. [7] employees SIMEX method for DP inference by extending SIMEX to Laplace noise and the scenario where both predictor and response are perturbed.

Referring to the simulation result in [7], the performance of SIMEX in DP inference leaves a lot to be desired. It requires a large privacy budget ϵ, around 20, to obtain a estimator with reasonaly small bias. For this scale of privacy budget, it might no longer provide any meaning privacy protection. Additional, SIMEX as a method does not incorporate the model into estimation. Therefore, if the true data generation model is known, SIMEX will not utilize all the available information. This leads to the question, what measurement error method incorporate the model into consideration? The Bayesian approach is the obviously answer.

2.4 Bayesian inference

Bayesian approach differs from the frequentist approach as it involves the decision on prior specifications. In Bayesian perspective, the unknown parameters are treated as random variables. The prior knowledges about these parameters are embedded by selecting probability distributions for these parameter. These probability distributions are called prior distributions. Once prior distributions are specified, an experiment is conducted and data is collected. These components are called the likelihood. Using the prior distribution and the likelihood, we can obtain newly updated probability distributions on the unknown parameters through Bayes' rule. These updated distributions are called the posterior distributions.

2.4.1 Why Bayesian?

The advantage of the Bayesian approach lies in that it takes the prior knowledge about the unknown parameter into account and incorporate it into the prior distributions. However, it's at the same time the center of contravercy as it's commonly criticized due to the subjectivity of the prior distributions specification.

Another advantage of the Bayesian method, which highlights the distinct difference between frequentist and Bayesian, is the posterior distribution. Instead of point estimator or set/interval estimator often used in frequentist approach, we can obtain the posterior distribution, which is much more informative. However, at the same time, the posterior distribution is precisely where the difficulties with Bayesian method reside.

2.4.2 Difficulties with Bayesian method

Although the Bayesian approach has been around since the nineteenth century, it has only started gaining popularity in the recent decades. The reason of this surge of noterity centered around the computation of posterior distribution. A closed form for posterior distribution almost does not exists aside from a handful of simple model (conjugate models). As the complexity of model

increases, the computation of posterior distribution becomes exceedingly difficult. However, with a boom of the processing power since late 90s and advent of Markov chain Monte Carlo simulation, posterior distribution computation has become much more manageable.

2.4.3 Markov Chain Monte Carlo

Markov chain monte carlo (MCMC) is a class of methods for sampling from difficult/complex distribution. In the context of Bayesian method, the posterior distribution is often without closed form or it's difficult to obtain its closed form. In such scenarios, MCMC have proven to be quite useful.

The idea is to create a markov chain (X_t) such that its stationary distribution π is the desired distribution. By the converging theorem, the distribution of X_t is close to π for large enough t. For more details on MCMC, refer to section 3.1 of [11].

Gibbs Sampler

Gibbs sampler is one of the most commonly used method in MCMC. It utilize the conditional distributions, which should be easy to sample from, to sample from the full distribution, which is difficult to sample from. The following brief explanation is mainly from section 10.1 of [15].

Suppose that for some $p > 1$, the random variable $\mathbf{X} \in \mathcal{X}$ can be written as $\mathbf{X} = (X_1, \ldots, X_p)$, where the X_i's are either uni- or multidimensional. Moreover, suppose that we can simulate from the corresponding univariate conditional densities $f_1, \ldots, f_p$, that is, we can simulate

$$(X_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p) \sim f_i\left(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p\right)$$

for $i = 1, 2, \ldots, p$. The densities $f_1, \ldots, f_p$ are called the full conditionals.

The associated Gibbs sampling algorithm (or Gibbs sampler) is given by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

Algorithm 4: Gibbs sampler

Given $\mathbf{x}^{(t)} = \left(x_1^{(t)}, \ldots, x_p^{(t)}\right)$, generate

$$\begin{aligned} X_1^{(t+1)} &\sim f_1\left(x_1 \mid x_2^{(t)}, \ldots, x_p^{(t)}\right) \\ X_2^{(t+1)} &\sim f_2\left(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}\right) \\ &\vdots \\ X_p^{(t+1)} &\sim f_p\left(x_p \mid x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)}\right) \end{aligned}$$

2.4.4 Bayesian method on measurement error problem

We will consider the setting of linear regression, where the predictor X and the response Y are both random and modeled as follow,

$$\begin{aligned} Y &= \beta_0 + X\beta_1 + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ X &\sim \mathcal{N}(\mu_x, \sigma_x^2) \end{aligned}$$

where $\beta_0, \beta_1, \mu_x, \sigma_x^2$ are the unknown parameters.

In most of the measurement error problems, only the predictor X is perturbed. The perturbed response scenario is not much considered with the reason being the perturbation noise is usually gaussian, which is usually the distribution for regression noise. Therefore, a perturbed response is equivalent to a unperturbed response with larger regression noise.

In order to simplify the posterior computation, the following priors are often chosen so that Gibbs sampler can be applied easily.

$$\begin{aligned} (\beta_0, \beta_1)^T &\sim \mathcal{N}(\mathbf{0}, \sigma_\beta \mathbf{I}) \\ \mu &\sim \mathcal{N}(0, \sigma_\mu^2) \\ \sigma_x^2 &\sim \mathrm{IG}(\eta_{1,x}, \eta_{2,x}) \\ \sigma_\varepsilon^2 &\sim \mathrm{IG}(\eta_{1,\varepsilon}, \eta_{2,\varepsilon}) \end{aligned}$$

where IG denotes the inverse gamma distribution.

Chapter 3

Bayesian inference on DP dataset

In this chapter, we will introduce the purposed Bayesian method of this report. First, the task DP inference is related to measurement error problem. Then, the Bayesian method for measurement error is extended to DP synthetic dataset. Lastly, we extend the method to multiple DP synthetic datasets.

3.1 Model setup

As mentioned in the introduction, we will be focusing on the setting of linear regression. More specifically, a random predictor, X, is used instead of a fixed predictor. The full model is stated as follow,

$$\begin{aligned} Y &= \beta_0 + X\beta + \varepsilon \\ X &\sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X) \\ \varepsilon &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \end{aligned}$$

Therefore, the dataset consists of both the predictor, X, and the response, Y, is considered. Denoted the perturbed predictors and perturbed response as $\tilde{X}$ and $\tilde{Y}$, we have,

$$\begin{aligned} \tilde{X} &= X + \varepsilon_x \\ \tilde{Y} &= Y + \varepsilon_y \end{aligned}$$

where in the case of Laplace mechanism,

$$\varepsilon_x, \varepsilon_y \sim \text{Lap}(0, \frac{\Delta_f}{\epsilon})$$

where Δ_f is the sensitivity of the dataset and ϵ is the privacy budget as usual. The main goal is to make inference on (β_0, β_1) given the DP dataset $(\tilde{X}, \tilde{Y})$.

3.2 DP Inference as a measurement error problem

Recall from section 2.2.3, to create a DP dataset is to perturb the dataset in a specific way, which is specified by DP and its privacy budget ϵ. That is, the task of making inference on a DP dataset is simply a subset of the measurement error problems. Although only the predictors are perturbed in most of the measurement problems, extending the method for measurement error problem to include a perturbed response takes little work as we will see later.

3.3 DP inference on linear regression

As explained in section 3.2, making inference on DP dataset can be viewed as a measurement error problem, and Bayesian method is commonly employed in measurement error problem as discussed in section 2.4.4. Therefore, it's natural to suspect that Bayesian method will be a good fit for making inference on DP dataset.

3.3.1 Choose the prior distributions

For the ease of simulation, only β_0 and β_1 are assumed unknown. Although it is not difficulty to extend to unknown μ_x, σ_x^2 and σ_ε, Δ_f would be also unknown as a result. It turns out that unknown Δ_f cause numerical overflow during the Gibbs sampler iterations. Therefore, we will only set the prior distribution for (β_0, β_1) as,

$$(\beta_0, \beta_1) \sim \mathcal{N}(0, \sigma_\beta^2)$$

Once again, the gaussian distribution is chosen so that Gibbs sampler can be applied.

3.3.2 Gibbs sampler

To obtain the Gibbs sampler, it mainly utilize the idea that Laplace distribution can be represented as a mixture of normal distribution. For more details on this, refer to section A.3.

Algorithm 5: Gibbs sampler for Bayesian DP inference

Given $\beta_0^{(t)}, \beta_1^{(t)}, \{x_i^{(t)}\}_{i=1}^n, \{y_i^{(t)}\}_{i=1}^n, \{u_i^{(t)}\}_{i=1}^n, \{v_i^{(t)}\}_{i=1}^n, \{\tilde{x}_i\}_{i=1}^n, \{\tilde{y}_i\}_{i=1}^n$, we sample the following,

1. $U_i^{(t+1)} \sim \text{InvGauss}(b_1/|\tilde{x}_i - x_i^{(t)}|, 1)$ for $i = 1, \ldots, n$.
2. $V_i^{(t+1)} \sim \text{InvGauss}(b_2/|\tilde{y}_i - y_i^{(t)}|, 1)$ for $i = 1, \ldots, n$.
3. $(X_i^{(t+1)}, Y_i^{(t+1)}) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, \ldots, n$, where (we will drop the subscript i and superscript (t) for notation convenient)

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \begin{bmatrix} A_2 \\ B_2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \frac{B_1}{A_1B_1 - C_1^2} & \frac{C_1}{A_1B_1 - C_1^2} \\ \frac{C_1}{A_1B_1 - C_1^2} & \frac{A_1}{A_1B_1 - C_1^2} \end{bmatrix}$$

with

$$A_1 = \frac{1}{\sigma_x^2} + \frac{\beta_1^2}{\sigma_\varepsilon^2} + \frac{U_i}{b_1^2}, \qquad A_2 = \frac{\mu_x}{\sigma_x^2} - \frac{\beta_0\beta_1}{\sigma_\varepsilon^2} + \frac{U_i^2\tilde{X}_i}{b_1^2}$$
$$B_1 = \frac{1}{\sigma_\varepsilon^2} + \frac{V_i^2}{b_2^2}, \qquad B_2 = \frac{\beta_0}{\sigma_\varepsilon^2} + \frac{V_i^2\tilde{Y}_i}{b_2^2}$$
$$C_1 = \frac{\beta_1}{\sigma_\varepsilon^2}$$

4. $(\beta_0^{(t+1)}, \beta_1^{(t+1)}) \sim \mathcal{N}\left((\frac{\sigma_\varepsilon^2}{\sigma_\beta^2}\mathbf{I} + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{y}, (\frac{1}{\sigma_\beta^2}\mathbf{I} + \frac{1}{\sigma_\varepsilon^2}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\right)$ where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & X_1^{(t+1)} \\ \vdots & \vdots \\ 1 & X_n^{(t+1)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} Y_1^{(t+1)} \\ \vdots \\ Y_n^{(t+1)} \end{bmatrix}$$

3.3.3 Inference on multiple DP dataset

Is multiple DP datasets more beneficial than a single DP dataset? That is, is creating multiple DP datasets with less privacy budget for each one more beneficial than creating only a single DP dataset, which will have more privacy budget for said dataset? We will answer this equation in the perspective of making inference.

Using the composition theorem from section 2.1.4, we can easily create multiple DP datasets that as a whole satisfies any given amount of privacy budget ϵ. Continuing with the linear regression from previous chapter, we can create multiple DP datasets as follow,

$$\tilde{Y}_{i,j} = Y_i + \varepsilon^x_{i,j}$$
$$\tilde{X}_{i,j} = X_i + \varepsilon^y_{i,j}$$

where $(\tilde{X}_{i,j}, \tilde{Y}_{i,j})$ indicates the j-th DP synthetic copy for the i-th record in the original dataset. In the case of Laplace mechanism,

$$\varepsilon^x_{i,j}, \varepsilon^y_{i,j} \sim \text{Lap}(0, m\frac{\Delta_f}{\epsilon})$$

where m is the number DP synthetic dataset.

Gibbs sampler

Here we state the Gibbs sampler extended to accommodate multiple DP synthetic dataset.

Algorithm 6: Gibbs sampler for Bayesian inference on multiple DP datasets

Given $\beta_0^{(t)}, \beta_1^{(t)}, \{x_i^{(t)}\}_{i=1}^n, \{y_i^{(t)}\}_{i=1}^n, \{u_i^{(t)}\}_{i=1}^n, \{v_i^{(t)}\}_{i=1}^n$, we sample the following

1. $U_{i,j}^{(t+1)} \sim \text{InvGauss}(b_1/|\tilde{x}_{i,j} - x_i^{(t)}|, 1)$ for $i = 1, \ldots, n$.

2. $V_{i,j}^{(t+1)} \sim \text{InvGauss}(b_2/|\tilde{y}_{i,j} - y_i^{(t)}|, 1)$ for $i = 1, \ldots, n$.

3. $(X_i^{(t+1)}, Y_i^{(t+1)}) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, \ldots, n$, where (we will drop the subscript i and superscript (t) for notation convenient)

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \begin{bmatrix} A_2 \\ B_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \beta_0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \frac{B_1}{A_1 B_1 - C_1^2} & \frac{C_1}{A_1 B_1 - C_1^2} \\ \frac{C_1}{A_1 B_1 - C_1^2} & \frac{A_1}{A_1 B_1 - C_1^2} \end{bmatrix}$$

with

$$A_1 = \frac{1}{\sigma_x^2} + \frac{\beta_1^2}{\sigma_\varepsilon^2} + \frac{\sum_{j=1}^J U_{i,j}}{b_1^2}, \qquad A_2 = \frac{\mu_x}{\sigma_x^2} + \frac{\sum_{j=1}^J U_{i,j} \tilde{X}_{i,j}}{b_1^2}$$
$$B_1 = \frac{1}{\sigma_\varepsilon^2} + \frac{\sum_{j=1}^J V_{i,j}}{b_2^2}, \qquad B_2 = \frac{\sum_{j=1}^J V_{i,j}(\tilde{Y}_{i,j} - \beta_0)}{b_2^2}$$
$$C_1 = \frac{\beta_1}{\sigma_\varepsilon^2}$$

4. $(\beta_0^{(t+1)}, \beta_1^{(t+1)}) \sim \mathcal{N}\left((\frac{\sigma_\varepsilon^2}{\sigma_\beta^2}\mathbf{I} + \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y}, (\frac{1}{\sigma_\beta^2}\mathbf{I} + \frac{1}{\sigma_\varepsilon^2}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \right)$ where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & X_1^{(t+1)} \\ \vdots & \vdots \\ 1 & X_n^{(t+1)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} Y_1^{(t+1)} \\ \vdots \\ Y_n^{(t+1)} \end{bmatrix}$$

Chapter 4

Simulation and results

In this chapter, we will discuss the simulation setup to study the effectiveness of the Bayesian inference on differentially private dataset in the linear regression setting. Next, the simulation results are summarized, and the comparison plots between inference on single differentially private dataset and inference on multiple (2) differentially private datasets are drawn.

4.1 Simulation setup

Using the linear regression models stated on chapter 3 and chapter 4,

$$\begin{aligned} X_i &\overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2) \\ Y_i \mid X_i &\overset{idd}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma_\varepsilon^2) \\ \tilde{X}_{i,j} \mid X_i &\overset{iid}{\sim} \text{Lap}(X_i, \Delta_f/\epsilon) \\ \tilde{Y}_{i,j} \mid Y_i &\overset{iid}{\sim} \text{Lap}(Y_i, \Delta_f/\epsilon) \end{aligned}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$ where $m = 1$ if only a single DP dataset is generated. Since Y and X both have unbounded support, we have to deal with infinite sensitivity Δ_f. Recall from section 2.1.5, using the truncated interval $(-1.96, 1.96)$ for a standard normal random variable, the dataset sensitivity Δ_f is computed as

$$\Delta_f = (1.96 - (-1.96)) * (\sigma_x + \sigma_\varepsilon)$$

With the model stated above, we set

$$(\beta_0, \beta_1) = (2, 2)$$
$$(\mu_x, \sigma_x^2) = (0, 1)$$
$$\sigma_\varepsilon = 0.5$$

with $n = 500$ and $m = 1, 2$ for single DP datasets and double DP datasets.

For this simulation, it's assumed that β_0 and β_1 are the only two unknown parameters. We set the prior distribution as,

$$(\beta_0, \beta_1) \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I})$$

with $\sigma_\beta = 10$.

4.2 Results

We generate the DP dataset(s) $(\tilde{X}_{i,j}, \tilde{Y}_{i,j})_{i,j}$ with $n = 500$, $m = 1$ for single DP dataset and $m = 2$ for double DP datasets. Next, we using the gibbs sampler described in section 3.3.2 & 3.3.3 to obtain posterior means for β_0 and β_1. Repeating this process 50 times, we summarize the result by computing the mean square error,

$$\frac{1}{50} \sum_{l=1}^{50} (\hat{\beta_0}^{(l)} - \beta_0)^2 + (\hat{\beta_1}^{(l)} - \beta_1)^2$$

and mean absolute relative error for β_1,

$$\frac{1}{50} \sum_{l=1}^{50} \frac{|\hat{\beta_1}^{(l)} - \beta_1|}{\beta_1}$$

where $\hat{\beta_0}^{(l)}$ and $\hat{\beta_1}^{(l)}$ denote the posterior means for β_0 and β_1 in l-th iteration, respectively.

The results are listed in the table 4.1 and table 4.2. Additionally, to compare the results between single DP synthetic dataset with multiple (2) DP synthetic datasets, 2 comparison plots are drawn.

It seems evident from the 4.1 and 4.2, using single DP synthetic dataset results in less bias in posterior mean estimator. The advantage of single DP

Privacy budget ϵ	Mean square error	Mean relative bias for slope
5	0.0280	0.0495
2.5	0.0403	0.0565
2	0.0745	0.0796
1.5	0.1379	0.1072
1	0.1338	0.0959
0.5	0.8162	0.2609
0.25	3.1700	0.4698

Table 4.1: Result for using only a single DP dataset

Privacy budget ϵ	Mean square error	Mean relative bias for slope
5	0.0366	0.0566
2.5	0.0498	0.0660
2	0.1041	0.0958
1.5	0.1416	0.1122
1	0.5135	0.2125
0.5	1.0955	0.2846
0.25	6.1115	0.7127

Table 4.2: Result for using two DP datasets

synthetic is much more pronounced when the privacy budget ϵ is small. However, in both case, a privacy budget of 2.5 seems to be enough to provide a reasonably accurate estimates.

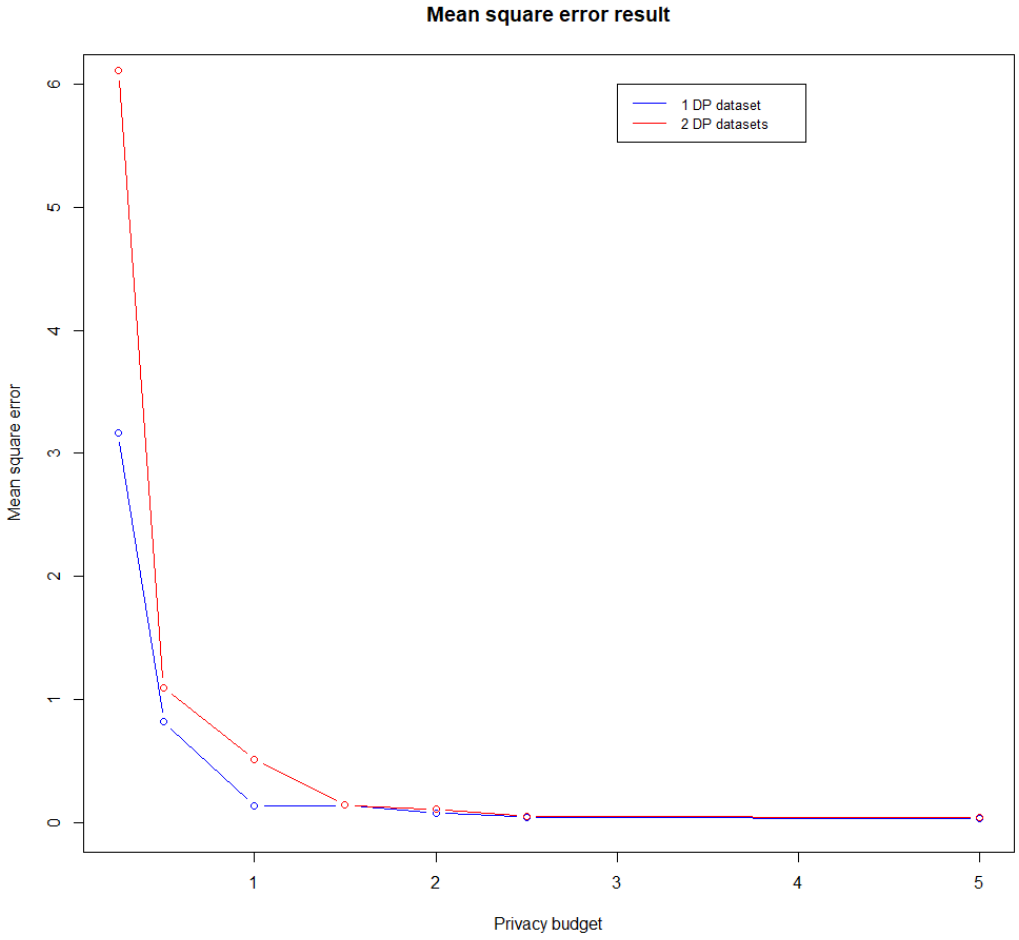

Figure 4.1: Mean square error for single DP dataset and double DP dataset

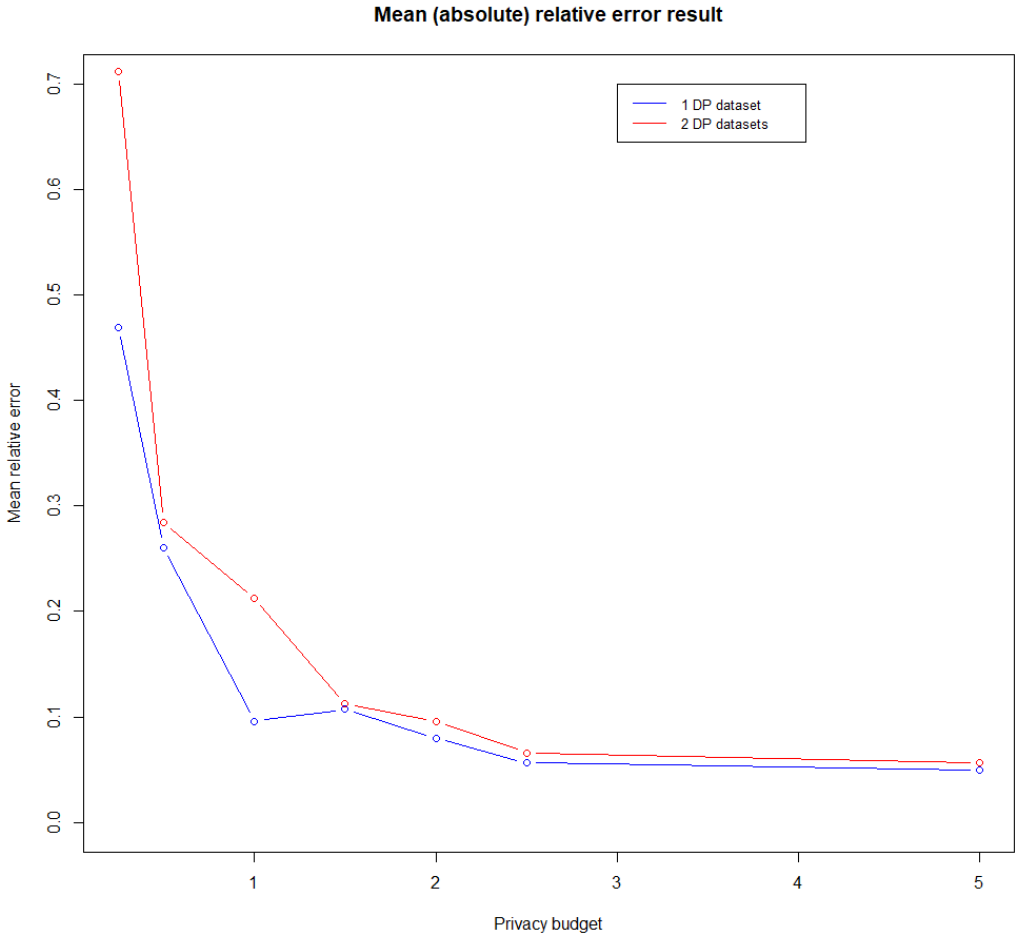

Figure 4.2: Mean relative error for single DP dataset and double DP dataset

Chapter 5

Conclusion and furture directions

Referring back to the simulation result (chapter 4), we were able to obtain point estimator with small bias at moderate large ϵ (around 2.5). Comparing to the result in [7], we see a great improvement as they obtain reasonable point estimator with much larger ϵ, around 20. Furthermore, generating two DP datasets resulted in slightly worser (much worser when privacy budget ϵ is small) result than generating a single DP dataset. One reason behind this could be the composition theorem does not provide a sharp privacy bound [10]. That is, when applying the composition theorem, there is a loss of the privacy budget.

However, the Bayesian method discussed in this work is far more complete. First, we assumed μ_x, σ_x^2 and σ_ε are known during the simulation. Although it's not much work to extend the model to this unknown parameters, the key issues are that Δ_f would be unknown as well as the result. It turns out that extending the model to unknown Δ_f is fairly problematic as it causes computational problem in Gibbs sampler. Next, we assumed that there is only one predictor, which is extremely restrictive. With multiple predictors, it will also introduce the scenario of correlated predictors. Last but not least, the gaussian distribution assumption on the predictors and the linear regression setting are extremely unrealistic in application.

References

[1] J. M. Abowd and L. Vilhuber, "How protective are synthetic data?" In *Privacy in Statistical Databases*, Springer Berlin Heidelberg, 2008, pp. 239–246.

[2] G. Bernstein and D. Sheldon, *Differentially private bayesian inference for exponential families*, 2018. arXiv: 1809.02188 [cs.LG].

[3] ——, *Differentially private bayesian linear regression*, 2019. arXiv: 1910.13153 [cs.LG].

[4] C. M. Bowen and F. Liu, "Comparative study of differentially private data synthesis methods," *Statistical Science*, vol. 35, no. 2, May 2020.

[5] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[6] A.-S. Charest, "How can we analyze differentially-private synthetic datasets?" *Journal of Privacy and Confidentiality*, vol. 2, pp. 21–33, Jan. 2010.

[7] A.-S. Charest and L. Nombo, "Analysis of differentially-private microdata using simex," in Sep. 2020, pp. 109–120.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, Springer Berlin Heidelberg, 2006, pp. 265–284.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, Aug. 2014.

[10] P. Kairouz, S. Oh, and P. Viswanath, *The composition theorem for differential privacy*, 2015. arXiv: 1311.0776 [cs.DS].

[11] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, 2nd ed. American Mathematical Society, 2017.

[12] F. Liu, *Model-based differentially private data synthesis and statistical inference in multiply synthetic differentially private data*, 2016. arXiv: 1606.08052 [stat.ME].

[13] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 2007, pp. 94–103. DOI: 10.1109/FOCS.2007.66.

[14] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, 2009, pp. 19–30.

[15] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Chapman and Hall/CRC, 2004.

[16] A. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, 2011, pp. 813–822.

Appendix A

Background Material

A.1 Laplace distribution

Definition 11 (**Laplace distribution**). A random variable X is said to be a Laplace random variable with location parameter m and scale parameter b, denoted as Lap(m, b), if it has the following probability density function,

$$f(x \mid m, b) = \frac{1}{2b} \exp\left(-\frac{|x - m|}{b}\right)$$

A.2 Derivation for Laplace mechanism

A.2.1 Laplace mechanism for regular query

Proof. Let $f : \mathcal{D} \to \mathbb{R}^k$ be the query function with $f = (f_1, f_2, \ldots, f_k)$, and fix a dataset $D \in \mathbb{R}^{n \times p}$ with its neigboring dataset by D'. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_k) \in \mathbb{R}^k$, we have

$$\begin{aligned}
&\log\left(\frac{\Pr(M(D) = \boldsymbol{y})}{\Pr(M(D') = \boldsymbol{y})}\right) \\
=&\log\left(\frac{\prod_{i=1}^{k} \operatorname{Lap}(y_i \mid f_i(D), \Delta_f/\epsilon)}{\prod_{i=1}^{k} \operatorname{Lap}(y_i \mid f_i(D'), \Delta_f/\epsilon)}\right) \\
=&\sum_{i=1}^{k} \log\left(\frac{\operatorname{Lap}(y_i \mid f_i(D), \Delta_f/\epsilon)}{\operatorname{Lap}(y_i \mid f_i(D'), \Delta_f/\epsilon)}\right) \\
=&\sum_{i=1}^{k} \log\left(\exp\left(\frac{-|y_i - f_i(D)| + |y_i - f_i(D')|}{\Delta_f/\epsilon}\right)\right) \\
=&\frac{1}{\Delta_f/\epsilon} \sum_{i=1}^{k} -|y_i - f_i(D)| + |y_i - f_i(D')|
\end{aligned}$$

Then it follows that

$$\begin{aligned}
&\left|\log\left(\frac{\Pr(M(D)=y)}{\Pr(M(D')=y)}\right)\right| \\
\leq &\frac{1}{\Delta_f/\epsilon}\sum_{i=1}^{k}|f_i(D)-f_i(D')| \\
\leq &\epsilon
\end{aligned}$$

□

A.2.2 Laplace mechanism for DP dataset

Proof. Fix a dataset $D \in \mathbb{R}^{n\times k}$, where denotes its row by $\boldsymbol{d}_i$. Denotes its neigboring dataset by D' and its row by $\boldsymbol{d}'_i$. Let j be the index of the row that D and D' differ.

$$\begin{aligned}
&\log\left(\frac{\Pr(M(D)=D^*)}{\Pr(M(D')=D^*)}\right) \\
=&\log\left(\frac{\prod_{i=1}^{n}\Pr(\boldsymbol{m}_i=\boldsymbol{d}_i^*)}{\prod_{i=1}^{n}\Pr(\boldsymbol{m}'_i=\boldsymbol{d}_i^*)}\right) \\
=&\log\left(\frac{\Pr(\boldsymbol{m}_j=\boldsymbol{d}_j^*)}{\Pr(\boldsymbol{m}'_j=\boldsymbol{d}_j^*)}\right), \text{ since } D \text{ and } D' \text{ only differ in 1 row} \\
=&\log\left(\frac{\prod_{i=1}^{k}\Pr(m_{j,i}=d_{j,i}^*)}{\prod_{i=1}^{k}\Pr(m'_{j,i}=d_{j,i}^*)}\right) \\
=&\sum_{i=1}^{K}\log\left(\frac{\Pr(m_{j,i}=d_{j,i}^*)}{\Pr(m'_{j,i}=d_{j,i}^*)}\right) \\
=&\sum_{i=1}^{K}\log\left(\frac{\mathrm{Lap}(d_{j,i}^* \mid d_{j,i},\Delta_f/\epsilon)}{\mathrm{Lap}(d_{j,i}^* \mid d'_{j,i},\Delta_f/\epsilon)}\right) \\
=&\sum_{i=1}^{K}\log\left(\exp\left(\frac{-|d_{j,i}^*-d_{j,i}|+|d_{j,i}^*-d'_{j,i}|}{\Delta_f/\epsilon}\right)\right) \\
=&\frac{1}{\Delta_f/\epsilon}\sum_{i=1}^{K}-|d_{j,i}^*-d_{j,i}|+|d_{j,i}^*-d'_{j,i}|
\end{aligned}$$

Then it follows that,

$$\begin{aligned}
&\log\left(\frac{\Pr(M(D)=D^*)}{\Pr(M(D')=D^*)}\right)\\
\leq&\frac{1}{\Delta_f/\epsilon}\sum_{i=1}^{K}|d_{j,i}-d'_{j,i}|\\
\leq&\epsilon
\end{aligned}$$

□

A.3 Laplace as a mixture of gaussian

Theorem 3. Fix a $b>0$, we have

$$\left.\begin{aligned}X\mid W&\sim\mathcal{N}(0,W)\\ W&\sim\mathrm{Exp}\left(\frac{1}{2b^2}\right)\end{aligned}\right\}\implies X\sim\mathrm{Lap}(0,b)$$

Note $f_W(w)=\frac{1}{2b^2}\exp\left(-\frac{w}{2b^2}\right)$ for $w\geq 0$.

Proof.

$$\begin{aligned}
&\int_0^\infty f_{X|W=w}(x)f_W(w)dw\\
=&\int_0^\infty\frac{1}{\sqrt{2\pi w}}\exp\left(-\frac{x^2}{2w}\right)\frac{1}{2b^2}\exp\left(-\frac{w}{2b^2}\right)dw\\
=&\frac{1}{2b^2}\int_0^\infty\frac{1}{\sqrt{2\pi w}}\exp\left(-\frac{w^2+|x|^2b^2}{2b^2w}\right)dw\\
=&\frac{1}{2b^2}\int_0^\infty\frac{1}{\sqrt{2\pi w}}\exp\left(-\frac{(w-|x|b)^2}{2b^2w}-\frac{2w|x|b}{2b^2w}\right)dw\\
=&\frac{1}{2b^2}e^{-|x|/b}\int_0^\infty\frac{1}{\sqrt{2\pi w}}\exp\left(-\frac{|x|^2(w-|x|b)^2}{2|x|^2b^2w}\right)dw\\
=&\frac{1}{2b^2}e^{-|x|/b}\frac{1}{\sqrt{\lambda}}\int_0^\infty w\underbrace{\frac{\sqrt{\lambda}}{\sqrt{2\pi w^3}}\exp\left(-\frac{\lambda(w-\mu)^2}{2\mu^2w}\right)}_{\mathrm{InvGauss}(\mu,\lambda)}dw,\ \text{where }\lambda=|x|^2,\ \mu=|x|b\\
=&\frac{1}{2b^2}e^{-|x|/b}\frac{1}{|x|}\mathbb{E}[Z],\ \text{where }Z\sim\mathrm{InvGauss}(\mu,\lambda)\\
=&\frac{1}{2|x|b^2}e^{-|x|/b}\mu\\
=&\frac{1}{2b}e^{-|x|/b}
\end{aligned}$$

□

Corollary 1. Fix a $b > 0$, we have

$$\left.\begin{array}{l} X \mid W \sim \mathcal{N}(0, \frac{b^2}{W}) \\ W \sim \mathrm{IG}\left(1, \frac{1}{2}\right) \end{array}\right\} \implies X \sim \mathrm{Lap}(0, b)$$

where IG denotes the inverse gamma distribution. In addition, we have

$$W \mid X \sim \mathrm{InvGauss}(\frac{b}{X}, 1)$$