STAT656 Final Report

Yuki Ohnishi, Yi-ting Hung and Satoshi Ido

Abstract

Local differential privacy is a differential privacy paradigm in which individuals first apply a privacy mechanism to their data (often by adding noise) before transmitting the result to a curator. The noise for privacy results in additional bias and variance in their analyses. Thus it is of great importance for analysts to incorporate the privacy noise into valid inference. In this article, we develop a Bayesian nonparametric methodology along with a blocked Gibbs sampling algorithm, which can be applied to any of privacy mechanisms, and which performs especially well in terms of MSE for tight privacy budgets. We also present simulation studies to evaluate the performance of our proposed frequentist and Bayesian methodologies for various privacy budgets, resulting in useful suggestions for performing causal inference for privatized data.

1 Introduction

Causal inference is a fundamental consideration across a wide range of domains in science, technology, engineering, and medicine. Researchers study experimental or observational data to unveil the causal effects of treatment assignment in an unbiased manner with valid uncertainty quantification. A traditional gold standard for performing causal inference is the classical randomized experiment (Imbens and Rubin, 2015). In this type of experiment, a great deal of control and precautions can be taken so as to eliminate events that would introduce instabilities and biases in causal inferences.

On the other hand, differential privacy (DP), introduced by Dwork et al. (2006), is another growing domain in science and business, as privacy protection has become a core concern for many organizations in the modern data-rich world. DP is a mathematical framework that provides a probabilistic guarantee that protects private information about individuals when publishing statistics about a dataset. This probabilistic guarantee is often achieved by adding random noise to the data. One DP model is the *central* differential privacy model, in which the data curators have access to the sensitive data and apply a DP mechanism to the data to produce the published outputs. A weakness of this model is that users are required to trust the data curators with their sensitive data. Another DP model is *local* differential privacy (LDP). In this model, the users do not directly provide their data to the data curator; instead, users apply the DP mechanism to their data locally before sending it to the curator. LDP is a preferable model if the data curators are not trusted by users. The LDP model has been adopted by various tasks and organizations, e.g., Google (Erlingsson et al., 2014) and Apple (Apple, 2017), for more stringent privacy protection.

Drawing causal conclusions from privatized data can be challenging. While the added random noise helps in safeguarding individuals' privacy, it distorts the actual patterns in the data. This distortion can lead to biased conclusions even in randomized experiments. This issue becomes even more pronounced in the LDP method, where each data point is individually altered before it is compiled. Therefore, when trying to understand cause-and-effect relationships using this protected data, researchers must exercise extra caution to ensure their interpretations remain accurate and unbiased.

In this article, we propose statistically valid Bayesian causal inferential methodologies under the local privacy scenarios. Our main contributions are as follows:

- We develop a flexible and efficient Bayesian nonparametric methodology. Additionally, we introduce a novel data augmentation Gibbs sampler, tailored for locally privatized observations under the potential outcome framework. This methodology is general and can be applied to all scenarios considered in the frequentist analyses. Our simulation studies show that our Bayesian approach performs well in terms of MSE, especially for tight privacy budgets.

- We present simulation studies to evaluate the Bayesian methodologies at various privacy budgets, resulting in useful suggestions for performing causal inference for privatized data. We then apply our methodologies to the real-life data from the evaluation of a cash transfer program conducted in Colombia. We show our methodologies successfully recover the non-private estimates for moderate privacy budgets.

1.1 Related Work

While DP is a rapidly growing field, the literature on causal inference methodologies for differentially privatized data remains sparse. The following work uses LDP for its DP mechanism. Agarwal and Singh (2021) introduced an end-to-end procedure for covariates cleaning, estimation, and inference, offering covariates cleaning-adjusted confidence intervals under the local differential privacy mechanism.

Some researchers have developed causal inference methodologies under the central DP model. D'Orazio et al. (2015) introduced the construction of central differential privacy mechanisms for summary statistics in causal inference. They then presented new algorithms for releasing differentially private estimates of causal effects and the generation of differentially private covariance matrices from which any least squares regression may be estimated. Lee et al. (2019) proposed a privacy-preserving inverse propensity score estimator for estimating the average treatment effect (ATE). Komarova and Nekipelov (2020) studied the impact of differential privacy on the identification of statistical models and demonstrated identification of causal parameters failed in regression discontinuity design under the central differential privacy. Niu et al. (2022) introduced a general meta-algorithm for privately estimating conditional average treatment effects. Kusner et al. (2016) tackles causal inference using a framework called the additive noise model (ANM), a more restrictive causal model than the Rubin Causal Model.

In non-causal domains, Evans and King (2022) offered statistically valid linear regression estimates and descriptive statistics for locally private data that can be interpreted as ordinary analyses of non-confidential data but with appropriately larger standard errors. Schein et al. (2019) presented an MCMC algorithm that approximates the posterior distribution over the latent variables conditioned on data that has been locally privatized by the geometric mechanism. Ju et al. (2022) proposed a general privacy-aware data augmentation MCMC framework to perform Bayesian inference from privatized data.

2 Preliminaries

2.1 Rubin Causal Model

Causal inference is of fundamental importance across many scientific and engineering domains that require informed decision-making based on experiments. Throughout this manuscript, we adopt the Rubin Causal Model (RCM) as our causal paradigm. In the RCM it is critical to first carefully define the Science of a particular problem, i.e., to define the experimental units, covariates, treatments, and potential outcomes (Imbens and Rubin, 2015). We consider N experimental units, indexed by $i = 1, \dots, N$, that correspond to physical objects at a particular point in time. Each unit i has an observed outcome Y_i and treatment assignment W_i respectively. We consider a binary treatment $W_i \in \{0, 1\}$ with a fixed assignment probability, $p = P(W_i = 1)$, which is assumed to be known by the experimental design, and let $Y_i(w)$ denote a potential outcome for $w \in \{0, 1\}$. In this article, we consider the N units as a random sample from a large super-population, and we are interested in inferring the Population Average Treatment Effect (PATE): $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$. We invoke the common set of assumptions, which enable us to identify the PATE by the estimators derived in this manuscript (Imbens and Rubin, 2015).

Assumption 1. 1. *(Positivity) The probability of treatment assignment given the covariates is bounded away from zero and one:* $0 < P(W_i = 1) < 1$.

2. *(Random Assignment) The potential outcomes are independent of treatment assignment:* $\{Y_i(0), Y_i(1)\} \perp \perp W_i$.

3. *(Stable Unit Treatment Value Assumption [SUTVA]) There is neither interference nor hidden versions of treatment. The observed outcome is formally expressed as:* $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$.

Lemma 1. *The potential outcomes are conditionally independent of the privatized treatment assignments given the actual treatment assignment:*

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp \tilde{W}_i \mid W_i.$$

This result holds because the DP mechanism flips the given treatment independently. This result is subtle, but important because it plays a crucial role in the development of our Bayesian methodology.

2.2 Differential Privacy

In this article, we use the local differential privacy (LDP) model. Let $\mathcal{D}$ be the set of possible contributions from one individual in database D. In this paper, we only consider non-interactive local DP mechanisms. LDP is formally defined for any $\mathcal{D}$ as follows.

Definition 1 (Local Differential Privacy)**.** An algorithm $\mathcal{M}$ is said to be ϵ-locally differentially private (ϵ-LDP) if for any two data points $x, x' \in \mathcal{D}$, and any $S \subseteq \text{Range}(\mathcal{M})$,

$$P(\mathcal{M}(x) \in S) \leq \exp(\epsilon) P(\mathcal{M}(x') \in S).$$

Intuitively, if an individual were to change their value from x to x', the output distribution of M would be similar, making it difficult for an adversary to determine whether x or x' was the true value. The value ϵ is called the *privacy budget* and lower values indicate a stronger privacy guarantee. Two important properties of differential privacy are *composition* and *invariance to post-processing*. Composition allows one to derive the cumulative privacy cost when releasing the results of multiple privacy mechanisms: if $\mathcal{M}_1$ is ϵ_1-LDP and $\mathcal{M}_2$ is ϵ_2-DP, then the joint release $(\mathcal{M}_1(x), \mathcal{M}_2(x))$ satisfies $(\epsilon_1 + \epsilon_2)$-LDP. Invariance to post-processing ensures that applying a data-independent procedure to the output of a DP mechanism does not compromise the privacy guarantee: if $\mathcal{M}$ is ϵ-LDP with range $\mathcal{Y}$, and $f : \mathcal{Y} \to \mathcal{Z}$ is a (potentially randomized) function, then $f \circ \mathcal{M}$ is also ϵ-LDP. Invariance to post-processing is especially important in this paper, as all of our inference procedures can be expressed as post-processing of more basic DP quantities.

One of the most commonly used DP mechanisms is the Laplace mechanism, which adds noise to a function of interest. Importantly, the noise must be scaled proportionally to the *sensitivity* of the function, which measures the worst-case magnitude by which the function's value may change between two individuals. Formally, the ℓ_1-sensitivity of a function f: $\mathcal{D} \to \mathbb{R}^k$ is $\Delta_f = \sup_{x,y \in \mathcal{D}} ||f(x) - f(y)||_1$.

Proposition 1 (Laplace Mechanism)**.** *Let $f : \mathcal{D} \to \mathbb{R}^k$. The Laplace mechanism is defined as $M(x) = f(x) + (\nu_1, ..., \nu_k)^\top$, where the ν_i are independent Laplace random variables, $\nu_i \sim \text{Lap}(0, \Delta f/\epsilon)$, where the density of the Laplace distribution,* Lap(μ, b)*, is given by $f(\nu|\mu, b) = \frac{1}{2b}\exp(-\frac{|\nu-\mu|}{b})$. Then M satisfies ϵ-LDP.*

For a binary variable (e.g., treatment assignment), a common mechanism is the randomized response.

Proposition 2 (Randomized Response Mechanism)**.** *Let $Z_i \in \{0, 1\}$ be a binary variable. The randomized response mechanism is defined as*

$$M(Z_i) = \begin{cases} Z_i & w.p.\ \frac{\exp(\epsilon)}{1+\exp(\epsilon)} \\ 1 - Z_i & w.p.\ \frac{1}{1+\exp(\epsilon)}, \end{cases}$$

which satisfies ϵ-LDP.

Throughout this article, we consider a scenario where all variables are jointly and separately privatized. The observed outcomes are privatized by the Laplace mechanism. The privatized outcomes are $\tilde{Y}_i = Y_i + \nu_i^Y$, where $\nu_i^Y \sim \text{Lap}(1/\epsilon_y)$. The binary treatment variable W_i is privatized by the random response mechanism.

$$\tilde{W}_i = \begin{cases} W_i & \text{w.p. } q_{\epsilon_w} = \frac{\exp(\epsilon_w)}{1+\exp(\epsilon_w)} \\ 1 - W_i & \text{w.p. } 1 - q_{\epsilon_w} = \frac{1}{1+\exp(\epsilon_w)}. \end{cases}$$

By composition, the joint release of $(\tilde{Y}_i, \tilde{W}_i)_{i=1}^N$ satisfies $(\epsilon_y + \epsilon_w)$-LDP. $\tilde{Y}_i$ is observed after adding noise to Y_i, which is either $Y_i(0)$ or $Y_i(1)$, but we cannot identify which it is through the observed variables because W_i is also unobserved.

3 Bayesian Approach

3.1 Overview of the Bayesian Methodology

Following the Bayesian paradigm of Rubin (1978), we consider deriving the posterior distributions of the causal estimands (Forastiere et al., 2016; Ohnishi and Sabbaghi, 2022a). The key idea is the data augmentation (Tanner and Wong, 1987) to obtain the posterior distribution of the causal estimands by imputing in turn the missing variables. The idea for estimating causal effects in the Bayesian paradigm is outlined in Rubin (1978); Imbens and Rubin (2015), but our unique challenges lie in the fact that neither treatment variable W nor either potential outcome $Y(0), Y(1)$ is observed.

To show how Bayesian inference proceeds in our framework, consider the following joint distribution of all observed variables $\tilde{\mathbf{O}}$ and missing variables $\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}$: $P(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}, \tilde{\mathbf{O}})$, where $\tilde{\mathbf{O}} = (\tilde{\mathbf{Y}}, \tilde{\mathbf{W}})$ for our scenario. Since causal effects are identifiable under randomization without covariate adjustment and incorporating covariates requires additional privacy costs for their release, we do not necessarily include covariates in our Bayesian methodologies, but the extension should be straightforward (e.g., Maceachern (1999)).

Under the super-population perspective, the observed and missing variables are considered as a joint draw from the population distribution. Bayesian inference considers the observed values of these quantities to be realizations of random variables and the missing values to be unobserved random variables. We also assume these quantities are unit exchangeable, then de Finetti's theorem implies that there exists a vector of parameters, $\boldsymbol{\theta}$, with the prior distribution $P(\boldsymbol{\theta})$ such that

$$\begin{aligned} P(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}, \tilde{\mathbf{Y}}, \tilde{\mathbf{W}}) &= \int P(\boldsymbol{\theta}) \prod_i P(Y_i(0), Y_i(1), W_i, \tilde{Y}_i, \tilde{W}_i \mid \boldsymbol{\theta}) d\boldsymbol{\theta} \\ = \int P(\boldsymbol{\theta}) \prod_i P(W_i) P(\tilde{W}_i \mid W_i) & P(Y_i(0), Y_i(1) \mid \boldsymbol{\theta}) P(\tilde{Y}_i \mid Y_i(0), Y_i(1), W_i) d\boldsymbol{\theta}, \end{aligned} \tag{1}$$

which follows from the conditional independence of potential outcomes and $\tilde{W}_i$ given W_i (Lemma 1) and the random assignment assumption. The distribution of $\tilde{Y}_i$ depends not only on $Y_i(0)$ and $Y_i(1)$ but also on W_i because the DP mechanism is applied to the observed outcome $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. Note that we know the DP mechanisms for W and Y, that is, $P(\tilde{Y}_i \mid Y_i(0), Y_i(1), W_i)$ and $P(\tilde{W}_i \mid W_i)$ have a known functional form. Therefore, the modeling effort is only required for $P(Y_i(0), Y_i(1) \mid \boldsymbol{\theta})$. Under this modeling strategy, our Bayesian approach is a valid inference for PATE. Note that PATE is a function of the parameters $\boldsymbol{\theta}$, which governs the potential outcomes. Thus, it suffices to obtain the posterior draws of the posterior of the $\boldsymbol{\theta}$ for the posterior draws of PATE.

A significant insight from (1) is that the treatment assignment mechanism is *not* ignorable. In conventional non-private settings, the treatment assignment model, represented as $P(W_i)$, is ignorable and falls out of the likelihood in Bayesian causal inference under randomization or unconfoundedness assumptions (Li et al., 2023). Yet, in a DP context, these treatment assignment variables are not directly observed. This necessitates the integration of both the treatment assignment models and their respective privacy mechanisms into our inferences. Additionally, a nuanced but crucial point is the necessity to model both $Y_i(0)$ and $Y_i(1)$. Typically, Bayesian causal inference for PATE estimation is performed via observable data (e.g., Zigler (2016); Stephens et al. (2023)). This is because the missing potential outcome eventually gets marginalized out from (1) under the assumption of prior parameter independence and unconfounded assignment, thus it does not influence parameter inference. In our scenario, however, it is uncertain whether $Y_i(0)$ or $Y_i(1)$ has been privatized to yield $\tilde{Y}_i$. This uncertainty calls for a data augmentation strategy for both potential outcomes.

We adopt the Dirichlet Process Mixture (DPM) to model $P(Y_i(0), Y_i(1) \mid W_i, \boldsymbol{\theta})$ for its flexibility. The DPM is a natural Bayesian choice for density estimation problems, which fits our needs that require

$P(Y_i(0), Y_i(1) \mid W_i, \boldsymbol{\theta})$ to be estimated without assuming strong parametric forms.

3.2 Algorithm Outlines

Equation (1) motivates the Gibbs sampling procedures to obtain the draws from the posterior distribution of $\boldsymbol{\theta}$. This section describes the key steps of the Gibbs sampler. Each step is derived from the corresponding components of (1). For inference of DPM parameters, denoted by $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{u})$, we adopt an approximated blocked Gibbs sampler based on the truncation of the stick-breaking representation (Ishwaran and Zarepour, 2000), due to its simplicity. In this algorithm, we set a conservatively large upper bound, $K \leq \infty$, on the number of components that units potentially belong to. Let $C_i \in \{1, ..., K\}$ denote the latent class indicators with a multinomial distribution, $C_i \sim \text{Multinomial}(\mathbf{u})$ where $\mathbf{u} = (u_1, ..., u_K)$ denote the weights of all components of the DPM. The algorithm proceeds as follows.

1. Given $Y_i(0), Y_i(1)$, draw each W_i from $P(W_i = 1|-) = \frac{r_1}{r_0+r_1}$, where $r_w = P(\tilde{Y}_i \mid Y_i(w))P(\tilde{W}_i \mid W_i = w)P(W_i = w)$ for $w = 0, 1$.

2. Given $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{u}$, C_i and W_i, draw each $Y_i(0)$ and $Y_i(1)$ according to:

$$\begin{aligned} P(Y_i(W_i)|-) &\propto P(Y_i(W_i) \mid \mu_{W_i}^{C_i}, \Sigma_{W_i}^{C_i})P(\tilde{Y}_i \mid Y_i(W_i)) \\ P(Y_i(1 - W_i)|-) &\propto P(Y_i(1 - W_i) \mid \mu_{1-W_i}^{C_i}, \Sigma_{1-W_i}^{C_i}). \end{aligned}$$

3. Given $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{u}$, $Y_i(0)$ and $Y_i(1)$, draw each C_i from

$$P(C_i = k|-) \propto u_k P(Y_i(0) \mid \mu_0^k, \Sigma_0^k)P(Y_i(1) \mid \mu_1^k, \Sigma_1^k).$$

4. Let $u'_K = 1$. Given α, $\mathbf{C}$, draw u'_k for $k \in \{1, ..., K - 1\}$ from

$$P(u'_k|-) \propto \text{Beta}\left(1 + \sum_{i:C_i=k} 1, \alpha + \sum_{i:C_i>k} 1\right).$$

   Then, update $u_k = u'_k \prod_{j<k}(1 - u'_j)$.

5. Given $\mathbf{C}$ and $\mathbf{u}'$, draw α from

$$P(\alpha|-) \propto P(\alpha) \prod_{k=1}^{K} f\left(u'_k \,\middle|\, 1 + \sum_{i:C_i=k} 1, \alpha + \sum_{i:C_i>k} 1\right),$$

   where f is the pdf of u'_k, the beta distribution. The standard Metropolis-Hastings algorithm is used for this step.

6. Given $\mathbf{Y}(0)$, $\mathbf{Y}(1)$ and $\mathbf{C}$, draw $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from

$$P(\mu_0^k, \Sigma_0^k|-) \propto H(\mu_0^k, \mu_1^k, \Sigma_0^k, \Sigma_1^k) \prod_{i:C_i=k} P(Y_i(0), Y_i(1) \mid \mu_0^k, \mu_1^k, \Sigma_0^k, \Sigma_1^k).$$

Remark. The key steps of this algorithm are 1 and 2, which correspond to the data augmentation steps, imputing the latent variables $Y_i(0), Y_i(1)$ and W_i. In Step 1, the probability $P(\tilde{Y}_i \mid Y_i(w))$ for $w = 0, 1$ indicates that $\tilde{Y}_i$ is observed via privatizing the potential outcome $Y_i(w)$, which would have been observed if we observed $W_i = w$. In step 2, given W_i, the corresponding potential outcome $Y_i(W_i)$ is considered to be privatized, but the other missing potential outcome $Y_i(1-W_i)$ should not be associated with the observed $\tilde{Y}_i$ within the iteration. Therefore, the posterior distribution of $Y_i(W_i)$ cannot be obtained in a closed form as it is weighted by the privacy mechanism $P(\tilde{Y}_i \mid Y_i(W_i))$, whereas the missing potential outcomes $Y_i(1-W_i)$ are just generated from the outcome model $P(Y_i(1 - W_i) \mid \boldsymbol{\theta})$. We adopt the privacy-aware Metropolis-within-Gibbs algorithm proposed in Ju et al. (2022) for the posterior draws of $Y_i(W_i)$. They proposed a generic data augmentation approach of updating confidential data that exploits the privacy guarantee of the mechanism to ensure efficiency. Their algorithm has guarantees on mixing performance, indicating that the acceptance probability is lower bounded by $\exp(-\epsilon_y)$. Another advantage of their approach is that we may utilize the outcome model to sample a proposal value from $P(Y_i(W_i)|\theta)$ at the current value of θ, rather than specifying a custom proposal distribution and step size for the Metropolis-Hastings step. Finally, Steps 3–6 updates all the parameters of the DPM that govern the potential outcomes, using standard techniques; see the following sections for details of the DPM, full details of the algorithm.

3.3 Details of the DPM

We say the probability measure H is generated from a Dirichlet Process, $\mathrm{DP}(\alpha, H_0)$, with a concentration parameter $\alpha > 0$ and a base probability measure H_0 over a measurable space $(\Theta, \mathcal{S})$ (Ferguson, 1974) if, for any finite partition $(S_1, ..., S_k)$ of $\mathcal{S}$, we have

$$\big(H(S_1), ..., H(S_k)\big) \sim \mathrm{Dir}\big(\alpha H_0(S_1), ..., \alpha H_0(S_k)\big),$$

where $\mathrm{Dir}(\alpha_1, ..., \alpha_k)$ denotes the Dirichlet distribution with positive parameters $\alpha_1, ..., \alpha_k$. The DPM is specified as

$$\begin{aligned} \{Y_1(0), Y_1(1)\}, ..., \{Y_N(0), Y_N(1)\} \mid \Phi_1, ..., \Phi_N &\overset{ind}{\sim} p(Y_i(0), Y_i(1)|\Phi_i), \\ \Phi_1, ..., \Phi_N|H &\overset{ind}{\sim} H, \\ H &\overset{ind}{\sim} DP(\alpha, H_0). \end{aligned}$$

We write $\overset{ind}{\sim}$ to say *independently distributed.* This model has unit-level parameters Φ_i for $i = 1, ..., N$, but the discreteness of the Dirichlet process (DP) distributed prior implies that the vector $\mathbf{\Phi} = (\Phi_1, ..., \Phi_N)$ can be rewritten in terms of its unique values $\mathbf{\Phi}^* = (\Phi_1^*, ..., \Phi_K^*)$. In particular, this can be represented in the following stick-breaking process.

$$H = \sum_{k=1}^{\infty} u_k \delta_{\Phi_k}, \;\; u_k = v_k \prod_{l<k} [1 - v_l], \;\; v_l \overset{ind}{\sim} \mathrm{Beta}(1, \alpha).$$

More specifically, the outcome model is specified by the following model.

$$P(Y_i(w)|\boldsymbol{\mu}, \mathbf{\Sigma}) \propto \sum_{k=1}^{\infty} u_k \mathrm{TN}(\mu_w^k, \Sigma_w^k, 0, 1), \tag{2}$$

where $\mathrm{TN}(\mu, \sigma^2, u, l)$ denotes the truncated normal distribution with the mean, variance, upper bound and lower bound parameters. The atoms $\Phi_k = (\mu_0^k, \mu_1^k, \Sigma_0^k, \Sigma_1^k)$ and the weight parameters u_k are nonparametrically specified via $\mathrm{DP}(\alpha, H_0)$. This can be regarded as the infinite mixture of normal distributions, where μ_w^k and Σ_w^k is the location parameter and variance parameter of each component respectively.

For inference, we adopt an approximated blocked Gibbs sampler based on a truncation of the stick-breaking representation of the DP proposed by Ishwaran and Zarepour (2000), due to its simplicity. In this algorithm, we first set a conservatively large upper bound, $K \leq \infty$, on the number of components that units potentially belong to. Let $C_i \in \{1, ..., K\}$ denote the latent class indicators with a multinomial distribution, $C_i \sim MN(\mathbf{w})$ where $\mathbf{u} = (u_1, ..., u_K)$ denote the weights of all components of the DPM. Conditional on $C_i = k$, (2) is greatly simplified to

$$P(Y_i(w)|\boldsymbol{\mu}, \mathbf{\Sigma}) \propto \mathrm{TN}(\mu_w^k, \Sigma_w^k, 0, 1).$$

Ishwaran and James (2001) showed that an accurate approximation to the exact DP is obtained as long as K is chosen sufficiently large. The DPM provides an automatic selection mechanism for the number of active components $K^* < K$. To ensure that K is sufficiently large, we run several MCMC iterations with different values of K. If the current iteration occupies all components, then K is not large enough, so we increase K for the next iteration. We conduct this iterative process until the number of the occupied components is below K.

3.4 Detailed Steps of Gibbs Sampler

In this section we present the detailed steps of the Gibbs sampler that is described in Section 3.2. The algorithm is inspired by Schwartz et al. (2011) and Ohnishi and Sabbaghi (2022b).

1. Given $Y_i(0), Y_i(1)$, draw each W_i from

$$P(W_i = 1|-) = \frac{r_1}{r_0 + r_1},$$

   where, for unit i with $\tilde{W}_i = 0$,

$$r_0 = \mathrm{Lap}(\tilde{\mathrm{Y}}_\mathrm{i} \mid \mathrm{Y_i}(0), 1/\epsilon_\mathrm{y}) q_{\epsilon_w}(1-p) \text{ and } r_1 = \mathrm{Lap}(\tilde{\mathrm{Y}}_\mathrm{i} \mid \mathrm{Y_i}(1), 1/\epsilon_\mathrm{y})(1 - q_{\epsilon_w})p,$$

   and for unit i with $\tilde{W}_i = 1$,

$$r_0 = \mathrm{Lap}(\tilde{\mathrm{Y}}_\mathrm{i} \mid \mathrm{Y_i}(0), 1/\epsilon_\mathrm{y})(1 - q_{\epsilon_w})(1-p) \text{ and } r_1 = \mathrm{Lap}(\tilde{\mathrm{Y}}_\mathrm{i} \mid \mathrm{Y_i}(1), 1/\epsilon_\mathrm{y}) q_{\epsilon_w} p.$$

   where $\mathrm{Lap}(\mathrm{y} \mid \mu, \sigma)$ is the pdf of the laplace distribution evaluated at y with the location parameter μ and scale parameter σ.

2. Given $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{u}$, C_i and $W_i = w$, draw $Y_i(1-w)$ according to:

$$Y_i(1-w) \sim \mathrm{TN}(\mu_{1-w}^{C_i}, \Sigma_{1-w}^{C_i}, 0, 1),$$

   where $\mathrm{TN}(\mu, \sigma^2, u, l)$ denotes the truncated normal distribution with the mean, variance, upper bound and lower bound parameters.

   Then, draw $Y_i(w)$ using the following Privacy-Aware Metropolis-within-Gibbs sampler Ju et al. (2022):

   (a) Draw a proposal: $y* \sim \mathrm{TN}(\mu_w^{C_i}, \Sigma_w^{C_i}, 0, 1)$.

   (b) Accept the proposal with probability $\alpha = \min\left(1, \frac{\mathrm{Lap}(\mathrm{y}* | \tilde{\mathrm{Y}}_\mathrm{i}, 1/\epsilon_\mathrm{y})}{\mathrm{Lap}(\mathrm{y}^{\mathrm{prev}} | \tilde{\mathrm{Y}}_\mathrm{i}, 1/\epsilon_\mathrm{y})}\right)$,

   where y^{prev} is the value of $Y_i(w)$ in the previous step.

3. Given $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{u}$, $Y_i(0)$ and $Y_i(1)$, draw each C_i from

$$P(C_i = k|-) \propto u_k \mathrm{TN}(Y_i(0) \mid \mu_0^k, \Sigma_0^k, 0, 1) \mathrm{TN}(Y_i(1) \mid \mu_1^k, \Sigma_1^k, 0, 1).$$

   This is a multinomial distribution.

4. Let $u'_K = 1$. Given α, $\mathbf{C}$, draw u'_k for $k \in \{1, ..., K-1\}$ from

$$P(u'_k|-) \propto \mathrm{Beta}\left(1 + \sum_{i:C_i = k} 1, \alpha + \sum_{i:C_i > k} 1\right).$$

   Then, update $u_k = u'_k \prod_{j<k}(1 - u'_j)$.

5. Given $\mathbf{C}$ and $\mathbf{u}'$, draw α from

$$P(\alpha|-) \propto P(\alpha) \prod_{k=1}^{K} f\left(u'_k \middle| 1 + \sum_{i:C_i = k} 1, \alpha + \sum_{i:C_i > k} 1\right),$$

   where f is the pdf of u'_k, the beta distribution. The Metropolis-Hastings algorithm is used for this step with a proposal distribution $\mathrm{TN}(\alpha^{prev}, 1.0, 0, \infty)$. α^{prev} is the value of α in the previous step.

6. Given $\mathbf{Y}(0)$, $\mathbf{Y}(1)$ and $\mathbf{C}$, draw $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from

   (a) If $N_k = \sum_{i=1}^{N} \mathbb{1}(C_i = k) > 0$, draw Σ_w^k from $\mathrm{IG}(2+0.5N_k, 0.2^2+0.5s_w^k)$ where $s_w^k = \sum_{i:C_i=k}(Y_i(w) - \mu_w^k)^2$ for $w = 0, 1$. If $N_k = 0$,then draw Σ_w^k from the prior $\mathrm{IG}(2, 0.2^2)$.

   (b) If $N_k > 0$, draw μ_w^k from

   $$\mathrm{TN}\left(\frac{0.5 * \Sigma_w^k + 9.0 s_w}{\Sigma_w^k + 9.0N_k}, \frac{9.0\Sigma_w^k}{\Sigma_w^k + 9.0N_k}, 0, 1\right),$$

   where $s_w = \sum_{i=1}^{N} Y_i(w)$. If $N_k = 0$, draw μ_w^k from

   $$\mathrm{TN}(0.5, 9.0, 0, 1).$$

   We use a common choice of the base measure H_0: the Normal-Inverse-Gamma conjugate $\mathrm{N}(\mu_0, \sigma_0^2)\mathrm{N}(\mu_0, \sigma_0^2)\mathrm{IG}(a_0, b_0)\mathrm{IG}(a_0, b_0)$. The specific values of the hyperparameters in this step are: $\mu_0 = 0.5$, $\sigma_0 = 3.0$, $a_0 = 2.0$ and $b_0 = 0.2^2$ for both $w = 0, 1$.

4 Simulation Studies

We evaluate the frequentist properties of our methodologies for various privacy budgets. The evaluation metrics that we consider are bias and mean square error (MSE) in estimating a causal estimand, coverage of an interval estimator for a causal estimand, and the interval length. Bias, MSE and coverage are generally defined as $\sum_{m=1}^{M} (\tau - \hat{\tau}_m)/M$, $\sum_{m=1}^{M} (\tau - \hat{\tau}_m)^2/M$ and $\sum_{m=1}^{M} \mathbb{1}\left(\hat{\tau}_m^l \leq \tau \leq \hat{\tau}_m^u\right)/M$ respectively, where M denotes the number of simulated datasets, τ denotes the true causal estimand, $\hat{\tau}_m$, $\hat{\tau}_m^l$ and $\hat{\tau}_m^u$ denote the estimate of the causal estimand, 95% lower and upper end of the interval estimator of the causal estimand using dataset $m = 1, \ldots, M$. Our summary of the interval length is the mean of the lengths of the intervals computed from M simulated datasets. For our Bayesian method, the point estimator is the mean of the posterior distribution of a causal estimand, and the interval estimator is the 95% central credible interval. We ran the MCMC algorithm for $100,000$ iterations using a burn-in of $50,000$. The iteration numbers were chosen after experimentation to deliver stable results over multiple runs.

4.1 Data-generating Mechanisms

For our simulations, we consider a Bernoulli randomized experiment with treatment assignment and covariates for unit i generated according to:

$$W_i \sim \mathrm{Bernoulli}(0.5), X_{i,1} \sim \mathrm{Uniform}(0,1), X_{i,2} \sim \mathrm{Beta}(2,5), X_{i,3} \sim \mathrm{Bernoulli}(0.7).$$

To generate potential outcomes, we adopt the Beta regression Ferrari and Cribari-Neto (2004): $Y_i(w) \sim \mathrm{Beta}(\mu_i(w)\phi, (1-\mu_i(w))\phi)$, where $\mu_i(w)$ and ϕ are a location parameter and scale parameter respectively with $\mu_i(w) = \mathrm{expit}(1.0 - 0.8X_1 + 0.5X_2 - 2.0X_3 + 0.5w)$ and $\phi = 50$. We consider $X_{i,d}$ to generate Y_i but do not release the privatized $\tilde{X}_{i,d}$. This model is beneficial for our simulations because the generated data automatically satisfy the following sensitivity: $\Delta_Y = 1$. Then, we obtain the private data $\tilde{Y}_i$, $\tilde{W}_i$ by applying the corresponding privacy mechanisms. The actual value of PATE can be obtained in a closed form, which is necessary to calculate bias, MSE, and coverage. Under the data-generating processes, the expectations of each potential outcome are expressed as:

$$\mathbb{E}[Y(0)] = \mathbb{E}_{X_1,X_2,X_3}[\mu(0)] = \mathbb{E}_{X_1,X_2,X_3}\left[\frac{\exp(1.0 - 0.8X_1 + 0.5X_2 - 2.0X_3)}{1 + \exp(1.0 - 0.8X_1 + 0.5X_2 - 2.0X_3)}\right] = 0.359613,$$

$$\mathbb{E}[Y(1)] = \mathbb{E}_{X_1,X_2,X_3}[\mu(1)] = \mathbb{E}_{X_1,X_2,X_3}\left[\frac{\exp(1.5 - 0.8X_1 + 0.5X_2 - 2.0X_3)}{1 + \exp(1.5 - 0.8X_1 + 0.5X_2 - 2.0X_3)}\right] = 0.457068.$$

We refer readers to Ferrari and Cribari-Neto (2004) for further details about the Beta regression.

Table 1: Evaluation metrics of Bayesian estimators for $N = 10000, N_{sim} = 1000$. N_{sim} denotes the number of simulations. ϵ_{tot} denotes the total privacy budget.

	Coverage	Bias	MSE	Interval Width
ϵ_{tot}				
0.1	96.4%	−0.0949	0.0099	0.34
0.3	96.9%	−0.0953	0.0099	0.342
1.0	93.4%	−0.0691	0.0077	0.32
3.0	93.2%	−0.0081	0.0006	0.093
10.0	95.0%	−0.0023	0.0	0.026

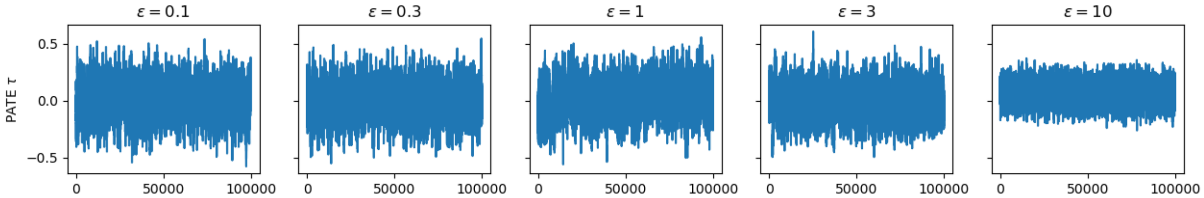

Figure 1: Traceplots for different privacy budgets.

4.2 Results

Table 1 presents the performance evaluation of our estimators for $N = 10000$ with various privacy budgets for ϵ_{tot}. We let $\epsilon_{tot} = \epsilon_y + \epsilon_w$, where $\epsilon_y = \epsilon_w$. All scenarios achieve about 95% coverage. We see that the Bayes estimator yields well-calibrated coverage probabilities and smaller MSE and bias for most cases. The differences in MSE between frequentist estimators and Bayesian estimators become negligible as ϵ_{tot} gets large ($\epsilon_{\text{tot}} = 3.0, 10.0$). When the privacy budget is tight, the Bayesian methodology outperforms the frequentist approach in all metrics. Specifically, the interval length of the Bayes estimator for $\epsilon_{\text{tot}} = 0.1$ is around 0.35 for all scenarios, which is informative enough about the estimands.

5 Real Data Analysis

We applied our methodology to a real-world causal inference task. We analyzed a randomized experiment that examined the impact of a cash transfer program on students' attendance rates (Barrera-Osorio et al., 2011). Conducted at San Cristobal in Colombia, the study recruited households with one to five school children, randomly assigning children to either participate in the cash transfer program or not with probability $p = 0.628$. The number of recruited students is $N = 5240$. With known treatment assignment, we assessed the treatment effect of the program on the attendance rate of the students, with eligible students receiving cash subsidies if they attended school at least 80% of the time in a given month. We utilized the privatization techniques as outlined in Section 2, setting ϵ_{tot} to values of 0.1, 0.3, 1.0, 3.0, and 10.0. Our methodologies were then benchmarked against non-private baseline methods, which offer target values for our private estimates. For the non-private frequentist baseline, we employed the standard IPW estimator.

5.1 Discussion

First, Figure 1 presents the traceplots of the MCMC chains across different privacy budgets. We observe that the MCMC chains converge for all privacy budgets.

Table 5.1 presents point mean estimators alongside the lower (2.5%) and upper (97.5%) bounds for interval estimators across each methodology. For the interval estimators, we used central confidence intervals for the frequentist approach and credible intervals for the Bayesian approach. Both frequentist and Bayesian non-private interval estimators highlighted a positive interval, indicating a significant effect. Given these results, our expectation for the private methodologies is, at best, to approximate the non-private values, since better inferences are unlikely with privatized data. Note that as the experimental data is fixed, the only randomness in this study is the privacy mechanisms. The point estimates for our Bayesian methodologies are similar to their non-private results when $\epsilon_{\text{tot}} \geq 3.0$. In particular, we observe

Table 2: Empirical analysis evaluating privatized cash transfer programs in Colombia. In the "Non-private" columns, "Freq" represents the standard DM estimator, while "Bayes" represents the standard Dirichlet process mixture models for non-private data.

	Non-private						Private		
	Freq			Bayes			Bayes		
ϵ_{tot}	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
0.1	0.006	0.001	0.009	0.005	0.001	0.008	0.011	-0.178	0.145
0.3	0.006	0.001	0.009	0.005	0.001	0.008	0.049	-0.082	0.190
1.0	0.006	0.001	0.009	0.005	0.001	0.008	0.041	-0.022	0.111
3.0	0.006	0.001	0.009	0.005	0.001	0.008	0.018	-0.007	0.034
10.0	0.006	0.001	0.009	0.005	0.001	0.008	0.007	0.000	0.015

that the Bayesian methodology demonstrated strong performance across all scenarios. These observations align with our simulation studies, further validating the efficacy of our methodologies.

6 Concluding Remarks

In this article we proposed causal inferential methodologies to analyze differential private data under the Rubin Causal Model. We considered a distinct local privacy scenarios that have practical relevance, where the outcome variables are privatized by the Laplace mechanism and the treatment variables are privatized by the randomized response mechanism. We presented a Bayesian methodology and its sampling algorithm as an alternative to the frequentist methodologies. Additionally, our Bayesian algorithm works effectively across a broad spectrum of privacy mechanisms if the privacy mechanism has a known likelihood. Finally, we validated the performance of our estimators via simulation studies and empirical analyses using real-world data.

A direction for future research is to develop an analytical framework for unbounded variables. Our framework is restricted to bounded variables due to considerations of the sensitivity of DP mechanisms. Another direction of future work would be to develop methodologies for the PATE estimation in observational studies.

References

Agarwal, A. and R. Singh (2021). Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. arXiv preprint arXiv:2107.02780.

Apple, D. (2017). Learning with privacy at scale. Apple Machine Learning Journal 1(8).

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle (2011, April). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. American Economic Journal: Applied Economics 3(2), 167–195.

D'Orazio, V., J. Honaker, and G. King (2015, 01). Differential privacy for social science inference. SSRN Electronic Journal.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference, pp. 265–284. Springer.

Erlingsson, Ú., V. Pihur, and A. Korolova (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp. 1054–1067.

Evans, G. and G. King (2022, 2021). Statistically valid inferences from differentially private data releases, with application to the Facebook urls dataset. Political Analysis, 1–21.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. The Annals of Statistics 2(4), 615 – 629.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics 31(7), 799–815.

Forastiere, L., F. Mealli, and T. J. VanderWeele (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. Journal of the American Statistical Association 111, 510–525.

Imbens, G. W. and D. B. Rubin (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96(453), 161–173.

Ishwaran, H. and M. Zarepour (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. Biometrika 87(2), 371–390.

Ju, N., J. Awan, R. Gong, and V. Rao (2022). Data augmentation MCMC for Bayesian inference from privatized data. Advances in Neural Information Processing Systems 35, 12732–12743.

Komarova, T. and D. Nekipelov (2020). Identification and formal privacy guarantees. arXiv preprint arXiv:2006.14732.

Kusner, M. J., Y. Sun, K. Sridharan, and K. Q. Weinberger (2016). Private causal inference. International Conference on Artificial Intelligence and Statistics 51, 1308–1317.

Lee, S. K., L. Gresele, M. Park, and K. Muandet (2019). Privacy-preserving causal inference via inverse probability weighting. arXiv preprint arXiv:1905.12592.

Li, F., P. Ding, and F. Mealli (2023). Bayesian causal inference: A critical review. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 381(2247), 20220153.

Maceachern, S. (1999, 01). Dependent nonparametric processes. Proceedings of the Section on Bayesian Statistical Science, American Statistical Association, 50–55.

Niu, F., H. Nori, B. Quistorff, R. Caruana, D. Ngwe, and A. Kannan (2022). Differentially private estimation of heterogeneous causal effects. In First Conference on Causal Learning and Reasoning.

Ohnishi, Y. and A. Sabbaghi (2022a). A Bayesian analysis of two-stage randomized experiments in the presence of interference, treatment nonadherence, and missing outcomes. Bayesian Analysis, 1 – 30.

Ohnishi, Y. and A. Sabbaghi (2022b). Degree of interference: A general framework for causal inference under interference. arXiv preprint arXiv:2210.17516.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics 6(1), 34 – 58.

Schein, A., Z. S. Wu, A. Schofield, M. Zhou, and H. Wallach (2019, 09–15 Jun). Locally private Bayesian inference for count models. In K. Chaudhuri and R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Volume 97 of Proceedings of Machine Learning Research, pp. 5638–5648. PMLR.

Schwartz, S. L., F. Li, and F. Mealli (2011, 12). A Bayesian semiparametric approach to intermediate variables in causal inference. Journal of the American Statistical Association 106, 1331–1344.

Stephens, D. A., W. S. Nobre, E. E. M. Moodie, and A. M. Schmidt (2023). Causal Inference Under Mis-Specification: Adjustment Based on the Propensity Score (with Discussion). Bayesian Analysis 18(2), 639 – 694.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82(398), 528–540.

Zigler, C. M. (2016, March). The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. The American Statistician 70(1), 47–54.