Discussion of Differential Private Bayesian Linear Regression

Xiaozhu Zhang xiaozhu.zhang@duke.edu
Department of Statistical Science
Duke University
Durham, NC 27708, USA

Jerome P. Reiter jreiter@duke.edu
Department of Statistical Science
Duke University
Durham, NC 27708, USA

Editor:

Abstract

We implement a differentially private Bayesian linear regression using the Laplace mechanism, the Wishart mechanism, and the inverse Wishart mechanism respectively. However, after our simulation investigation, we believe that the Bayesian differential private framework offers acceptably accurate results only for large sample sizes, but the computation procedure for large sample sizes is unacceptably inefficient. This utility-computation tradeoff calls into question the differential private hierarchical model introduced by Bernstein and Sheldon (2019).

Keywords: High-Dimensional Linear Regression, Differential Privacy, Hierarchical Model

1. Introduction

Linear regression is one of the most widely used statistical methods. It is important to develop robust tools that can realize the benefits of regression analyses but maintain the privacy of individuals. Differential privacy (Dwork et al., 2006) is widely accepted formalism to provide algorithmic privacy guarantees: a differentially private algorithm randomizes its computation to provably limit the risk that its output discloses information about individuals.

Suppose that $X = (x_1, \cdots, x_n) \in \mathbb{R}^{n\times p}$ is a generic data set with n individuals and p features, and $X' = (x_{1:i-1}, x'_i, x_{i+1:n}) \in \mathbb{R}^{n\times p}$ for some i is the neighboring data set that differ by a single record from X.

Definition 1 (Differential Privacy, Dwork et al. (2006))**.** *A randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if for any input X, any X' and any subset of outputs $O \subset \text{Range}(\mathcal{A})$,*

$$\Pr[\mathcal{A}(X) \in O] \le \exp(\epsilon) \Pr[\mathcal{A}(X') \in O].$$

The above guarantee is ensured by randomizing $\mathcal{A}$. Note that differential privacy enjoys the post-processing property, indicating that any further processing on the output of a differentially private algorithm that does not access the original data retains the same privacy guarantees.

A key concept is the sensitivity of a function, which quantifies the impact an individual has on the output of the function.

Definition 2 (Sensitivity, Dwork et al. (2006)). *The sensitivity of a function, f is $\Delta_f = \sup_{X,X'} \|f(X) - f(X')\|_1$.*

The most popular mechanism that induces ϵ-differential privacy is the Laplace mechanism.

Definition 3 (Laplace Mechanism, Dwork et al. (2006)). *Given a function f that maps data sets to $\mathbb{R}^m$, the Laplace mechanism outputs the random variable $\mathcal{L}(X) \sim \mathsf{Lap}(f(X), \Delta_f/\epsilon)$ from the Laplace distribution, which has density $\mathsf{Lap}(z; u, b) = (2b)^{-m} \exp(-\|z - u\|_1/b)$. This corresponds to adding zero-mean independent noise $u_i \sim \mathsf{Lap}(0, \Delta_f/\epsilon)$ to each component of $f(X)$.*

Besides the Laplace mechanism, Sheffet (2019) introduces two more mechanisms inducing (ϵ, δ)-differential privacy, the Wishart mechanism and inverse Wishart mechanism, which preserve the positive semi-definite property specifically for covariance matrices. These two mechanisms can be applied to linear regression since the sufficient statistics of linear regression can be formulated as a covariance matrix.

Definition 4 (Wishart Mechanism, (Sheffet, 2019)). *Given a matrix $A \in \mathbb{R}^{n\times d}$ and a bound $B > 0$ on the ℓ_2-norm of any row in A, the Wishart mechanism outputs the positive semi-definite matrix $M = A^\top A + \sum_{i=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$ where $\boldsymbol{v}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_d, B^2 I_{d\times d})$ and $k = \lfloor d + 14/\epsilon^2 \cdot 2\log(4/\delta)\rfloor$.*

Definition 5 (Inverse Wishart Mechanism, Sheffet (2019)). *Given a matrix $A \in \mathbb{R}^{n\times d}$ and a bound $B > 0$ on the ℓ_2-norm of any row in A, the inverse Wishart mechanism outputs the positive semi-definite matrix $M \sim \mathcal{IW}_d((A^\top A + \psi \cdot I_{d\times d}, n + d)$, where $\psi = 2B^2/\epsilon \cdot \Big(2\sqrt{2(n + d)\log(4/\delta)} + 2\log(4/\delta)\Big)$.*

Although most existing work on differentially private linear regression focuses on frequentist approaches, Bernstein and Sheldon (2019) proposes a Bayesian hierarchical framework for differentially private Bayesian linear regression, where they choose the Laplace mechanism and add noise to sufficient statistics. Through the approximation by CLT and Gibbs sampling procedure, good coverage rates are claimed to be obtained even for small $\epsilon = 0.1$ and small moderate size $n = 1000$. However, our simulation results (described in section 3), which are obtained under the same experimental design in Bernstein and Sheldon (2019), show that the Bayesian framework is not efficient as claimed to be.

In this report, we will implement the Laplace mechanism, the Wishart mechanism, as well as the inverse Wishart mechanism to induce (ϵ, δ)-differential privacy for linear regression. Section 2 introduces the private Bayesian linear regression model. Section 3 describes the simulation design and shows the results. Section 4 concludes this report and makes further discussions.

2. Private Bayesian Linear Regression

Given a response vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times p}$, we assume that

$$y = X\theta^* + w, \quad w \sim \mathcal{N}(0, \sigma^2)$$

where $\theta^* \in \mathbb{R}^p$ is the true parameter and $w \in \mathbb{R}^n$ is the noise. The goal is to perform Bayesian linear regression in an (ϵ, δ)-differentially private manner. We ensure privacy by employing sufficient statistic perturbation (Foulds et al., 2016), in which the Laplace mechanism, the Wishart mechanism, and the inverse Wishart mechanism are used to inject noise into the sufficient statistics $\mathbf{s} = [X^\top X, X^\top y, y^\top y]$, making them fit for public release. The question is then how to compute the posterior over the model parameters θ^* given the noisy sufficient statistics.

A standard assumption in literature (Sheffet, 2017) is to assume X and y have known a priori lower and upper bounds, (X_l, X_u) and (y_l, y_u), with with b to be an uniform upper bound of the range of each column of X, and B to be an uniform upper bound of the ℓ_2-norm of each row of X.

We consider the hierarchical model represented as the DAG in Figure 1.

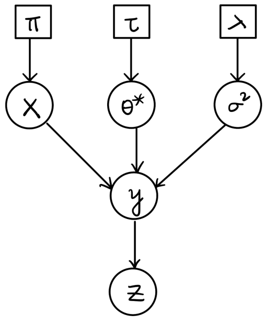


Figure 1: The DAG of differentially private linear regression.

Specifically,

$$X \sim \pi$$
$$\theta^* \sim \mathcal{N}(0, \tau^2)$$
$$\sigma^2 \sim \mathsf{HalfCauchy}(0, \lambda)$$
$$y | X, \theta^*, \sigma^2 \sim \mathcal{N}(X\theta^*, \sigma^2)$$

where π is the default non-informative prior in RStan. The observed values $\mathbf{z}$ is generated by

- Laplace mechanism: Note that here we keep the upper triangular of $X^\top X$ and reformulate $\mathbf{s} = [\text{vec}(\text{utri}(X^\top X)), X^\top y, y^\top y]$. Let $a = y_u - y_l$ and $\Delta_\mathbf{s} = b^2 p(p+1)/2 + abp + a^2$. Then we have $\mathbf{z}_i \sim \mathsf{Lap}(\mathbf{s}_i, \Delta_\mathbf{s}/\epsilon)$.
- Wishart mechanism: Let $A = [X; y]^\top [X; y]$, and then $Z = A + \sum_{i=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$ where $\boldsymbol{v}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_p, B^2 I_{p \times p})$ and $k = \lfloor p + 14/\epsilon^2 \cdot 2\log(4/\delta) \rfloor$.

- Inverse Wishart mechanism: Let $A = [X; y]^\top [X; y]$, and then $Z \sim \mathcal{IW}_p((A^\top A + \psi \cdot I_{p\times p}, n+p)$, where $\psi = 2B^2/\epsilon \cdot \left(2\sqrt{2(n+p)\log(4/\delta)} + 2\log(4/\delta)\right)$.

The hierarchical model above shows the full generation process of released data sets. From the users' prospective, only the contaminated covariance matrix Z, the assumed bounds b and B, the selected differential privacy mechanism, and the budget values (ϵ, δ) are regarded as known. In order to recover θ^* back and estimate the uncertainty with the given information described above, we have to use valid MCMC algorithms to construct Markov chains that converge to the posterior distribution $p(\theta^*|Z, X, y, \sigma^2)$. Although a Gibbs sampler can be obtained for the Laplace mechanism as shown by Bernstein and Sheldon (2019), it would much more difficult for the Wishart and inverse Wishart mechanisms. Consequently, we will choose the generic HMC algorithm and implement the model estimation in RStan directly.

3. Simulation

In this part, we will implement the hierarchical model that is introduced by Bernstein and Sheldon (2019) to perform differentially private linear regression. Apart from the Laplace mechanism, we will also try the Wishart mechanism and the inverse Wishart mechanism.

3.1 Simulation Design

We consider the following simulation design:

- Set the privacy budget to be $\epsilon = 0.1$ and $\delta = 0.1$, where δ only applies to the Wishart and inverse Wishart mechanisms.
- For the design matrix, we consider two cases: (1) Low-dimensional: $n \in \{10, 100, 1000\}$ and $p = 5$; (2) High-dimensional: $n = 30$ and $p \in \{30, 50, 100\}$. For each case, we generate the features to be independent and force each entry to reside in the range $[-1, 1]$. One of the possible strategies is to generate each entry iid from a truncated Gaussian distribution $\mathcal{N}(0, 1) \cdot \mathbf{1}_{[-1,1]}$.
- The true coefficients θ^* are from $\{0, 1\}^p$. For the low-dimensional case, we set all the coefficients to be 1. For the high-dimensional case, we set the first 20% coefficients to be 1 whereas others to be 0.
- The noise vector is given by $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.01)$ for all i.

We refer the low- and high-dimensional scenarios as

Case 1: Low-dimensional: $n = 10, p = 5$;

Case 2: Low-dimensional: $n = 100, p = 5$;

Case 3: Low-dimensional: $n = 1000, p = 5$;

Case 4: High-dimensional: $n = 30, p = 30$;

Case 5: High-dimensional: $n = 30, p = 50$;

Case 6: High-dimensional: $n = 30, p = 100$.

With the data sets obtained from the data generation process described above, we may begin to specify the prior placed for the DAG in Figure 1. In order to allow for a fully exploration of the parameter space for the Markov chain, we would like to choose rather large variances, say $\tau = 10$ and $\lambda = 10$. In addition, we will run two chains of MCMC each with 5000 iterations and 1000 burn-ins.

3.2 Simulation Results

It turns out that the results from the Bayesian hierarchical framework seldom provides efficient inference for the true coefficients θ^*. Specifically, for low-dimensional cases, none of the mechanisms manage to include the true coefficients $\theta_j^* = 1$, $j = 1, \ldots, 5$ within their 95% credible intervals, and most of the intervals cover the point 0. For high-dimensional cases, most of the credible intervals for true coefficients $\theta_j^* = 1$, $j \in \mathcal{S}$ cover both the point 0 and the true coefficients $\theta_j^* = 1$ for all j in the support. In addition, in the high-dimensional sub-figures of Figure 4 and Figure 5, we see that the intervals for true coefficients and null coefficients do not differentiate with each other too much. In summary, these credible intervals that are almost symmetric around 0 are not able to achieve neither feature selection nor proper estimation.

3.3 A Frequentist Implementation

One opinion about the Bayesian framework is that its results should not be too different from those in the Frequentist framework particularly when we are using non-informative priors, except that the Bayesian mechanism automatically provides a posterior distribution that is convenient for inference. In this case, the differential private linear regression should not provide efficient inference using the Frequentist framework either, which however is not consistent with the results in literature (Sheffet, 2019). Then the question becomes why and when the Laplace mechanism, the Wishart mechanism, and inverse Wishart mechanism lead to efficient inference.

We believe that the sample size matters. Figure 2 shows the bootstrap results for higher dimensional data (case 4, varying sample size) with the metric $\min\{\frac{\|\hat{\theta}-\theta^*\|_2}{\|\theta^*\|_2}, 1\}$. In other words, we did not run the MCMC for the hierarchical model we built in section 2; instead, we use the differentially private but contaminated point estimate $(\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top \tilde{y}$ directly, where $(\tilde{X}^\top \tilde{X}, \tilde{X}^\top \tilde{y})$ is obtained from the matrix Z. Based on the results, we see that the performance of inference for θ^* using the differentially private estimator improves with increasing sample size. In addition, compared with the Wishart and the inverse Wishart mechanisms, the Laplace mechanism performs much better particularly when the sample size is larger than 1,000,000. In summary, the efficiency of the differentially private estimator is acceptable only when the sample size is large.

3.4 The utility-computation trade-off

In order to improve the utility of the Bayesian framework for differentially private linear regression, we have to increase the sample size so that the estimator may be more efficient

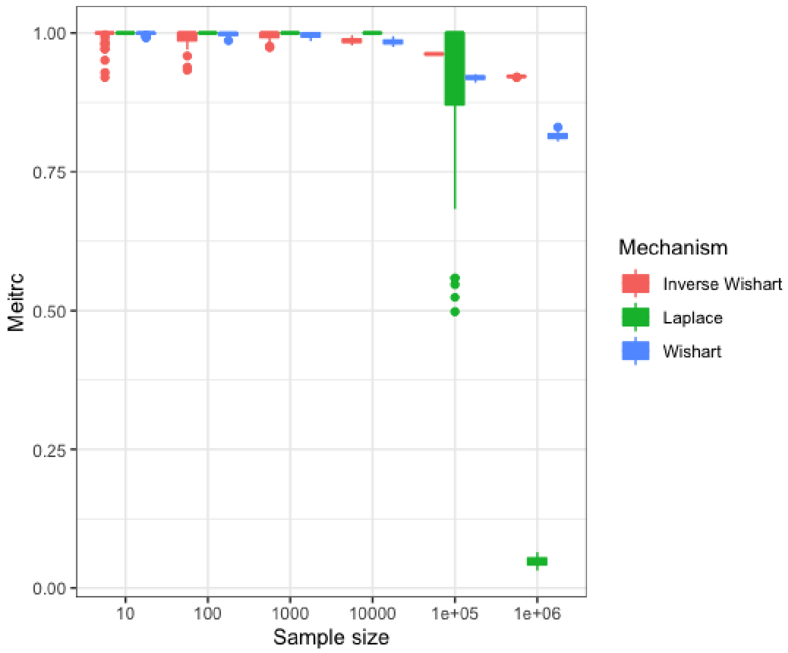


Figure 2: The bootstrap results for high dimensional data with the metric $\min\{\frac{\|\hat{\theta}-\theta^*\|_2}{\|\theta^*\|_2}, 1\}$.

in inference. However, the running time might be unacceptably longer, and the mixing might be unacceptably worse with increasing sample size.

Table 1 summaries the computation results for case 4 (varying sample size) using the Bayesian Laplace mechanism, where we run a MCMC with 2 chains and each chain has 5000 iterations with 1000 burn-ins. Here we use a computer with an Apple M1 chip, 16GB memory, and 4 cores. Based on the Table 1, we see that the mixing becomes unacceptably terrible when the sample size is $n = 1000$. Specifically, the $\hat{R}$ for the log-posterior is 1.36 and the effective sample size is only 2. Figure 3 shows the trace-plot of the log-posterior when $n = 1000$.

Table 1: The computational summary of MCMC procedure.

n	Time (sec)	$\max \hat{R}$	$\min \hat{R}$	$\hat{R}$ of lp	ESS of lp
10	107.415	1.01	1.00	1.00	882
100	530.719	1.03	1.00	1.03	149
1000	6137.39	1.44	1.00	1.36	2

In addition, when $n = 1000$, the sampling procedure took more than 1.5 hours, and the gradient evaluation for each chain of the HMC took 0.001025 seconds and 0.001166 seconds respectively. This duration might be acceptable, however, when $n = 10000$ the gradient evaluation for each chain would take 0.011773 seconds and 0.015447 seconds. In this case, the estimated time might be 10 times of 1.5 hours.

The table 2 gives estimated running time for each sample size. We see that when $n = 10,000$, the MCMC would take less than a day; when $n = 100,000$, the MCMC would

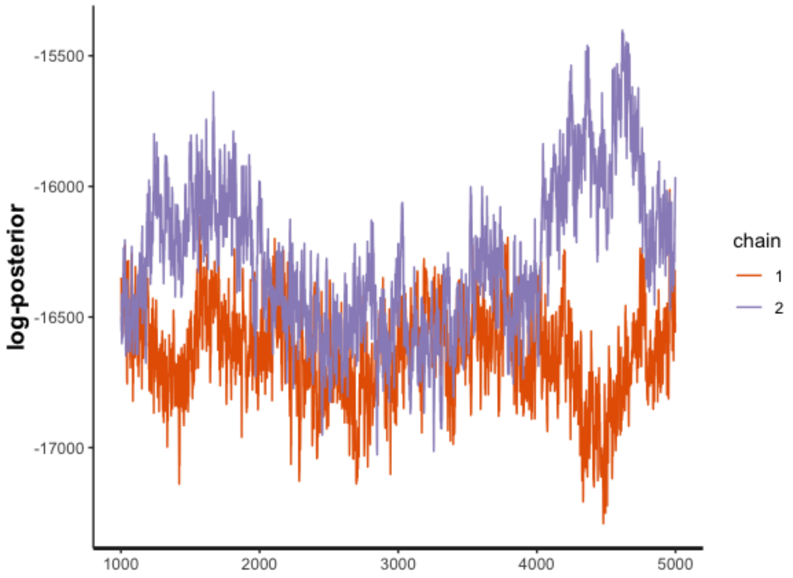


Figure 3: The trace-plot of log posterior for case 4 when $n = 1000$.

Table 2: The estimated running time for the HMC gradient of each chain, and the estimated total running time for each sample size.

n	Gradient of chain 1	Gradient of chain 2	Estimated Total (hours)
10,000	0.011773	0.015447	15
100,000	0.216508	0.228947	300
1,000,000	3.65513	3.67765	4800

take around 12 days; when $n = 1,000,000$, the MCMC would take around half a year! Even worse, as the sample size increasing, 5000 iterations might be far from enough if we need good mixing, implying the running time have to be even longer. You may wonder if it is proper if we choose a smaller sample size, however, according to the Figure 2 the sample size must be larger than 1,000,000 so that the inference can be efficient.

In summary, we see a trade-off between data utility and computation. When the sample size is not larger enough, the differentially private inference for θ^* is far from useful, however, for a sample size large enough the running time of the Bayesian framework is unacceptably long. This results in a death sentence of the differentially private linear regression introduced in section 2, and we consequently doubt the methodology and results described in Bernstein and Sheldon (2019).

4. Discussion

In this report we implemented the differentially private Bayesian linear regression using the Laplace mechanism, the Wishart mechanism, and the inverse Wishart mechanism respectively. However, after our simulation investigation, we believe that the Bayesian differential private framework offers acceptably accurate results only for large sample sizes. This creates computational challenges and makes the Bayesian framework introduced by Bernstein and Sheldon (2019) an impractical attempt.

One caveat is that in Bernstein and Sheldon (2019), the prior knowledge of the predictors' second and fourth crossed population moments is obtained by privately querying the corresponding sampling moments. This practice may somewhat reduce the running time, however, as pointed by Barrientos et al. (2021), "Privately querying the sample moments will require an exponentially increasing portion of the total privacy budget, because the number of required moments grows exponentially with the number of predictors."

In conclusion, significant improvement is needed for this differentially private Bayesian framework. Possible future directions may include underlying parameter expansion and data augmentation to speed up the computation process and improve mixing.

References

Andrés F Barrientos, Aaron R Williams, Joshua Snoke, and Claire McKay Bowen. A feasibility study of differentially private summary statistics and regression analyses for administrative tax data. *arXiv preprint arXiv:2110.12055*, 2021.

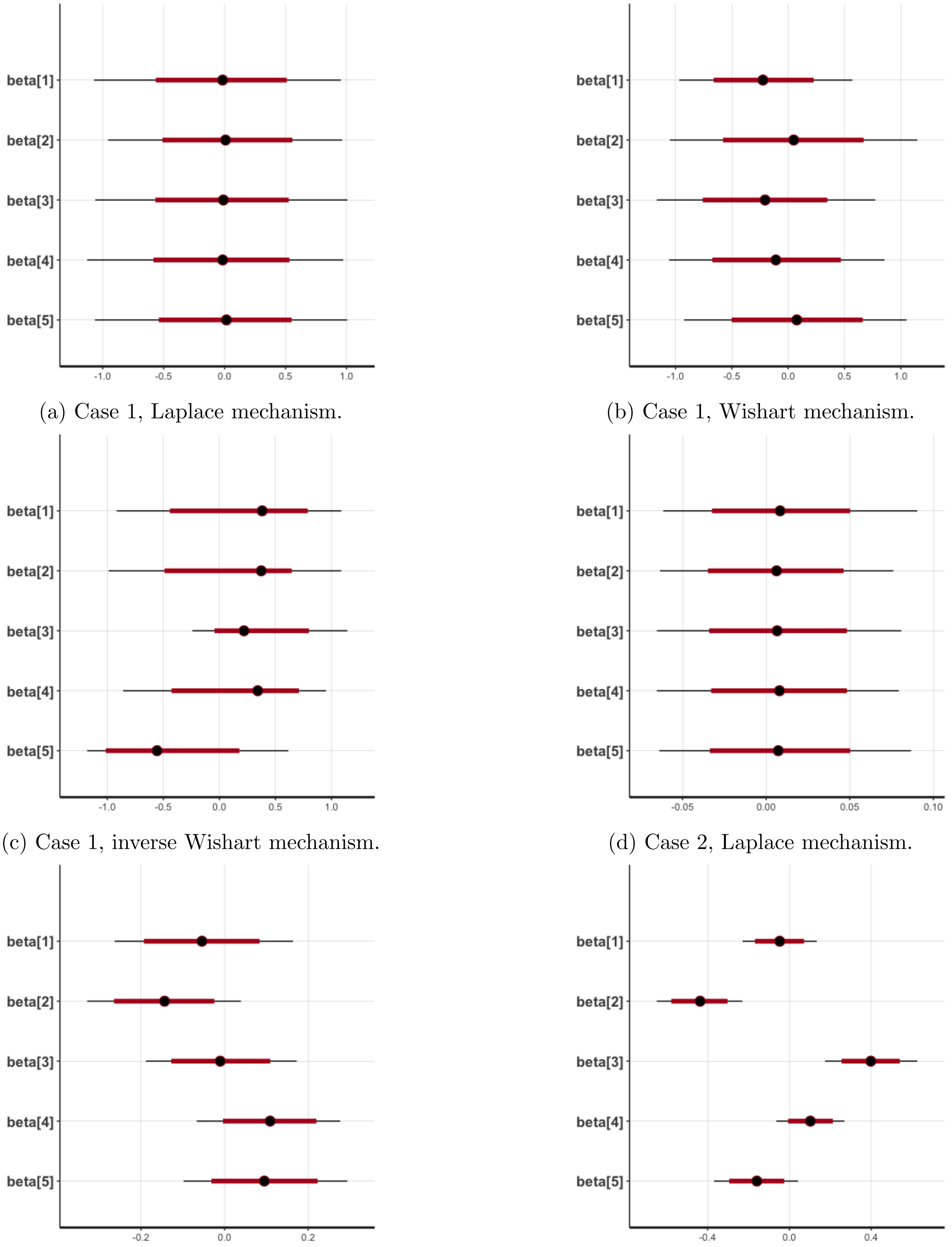
Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.

Or Sheffet. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pages 789–827. PMLR, 2019.



(a) Case 1, Laplace mechanism.

(b) Case 1, Wishart mechanism.

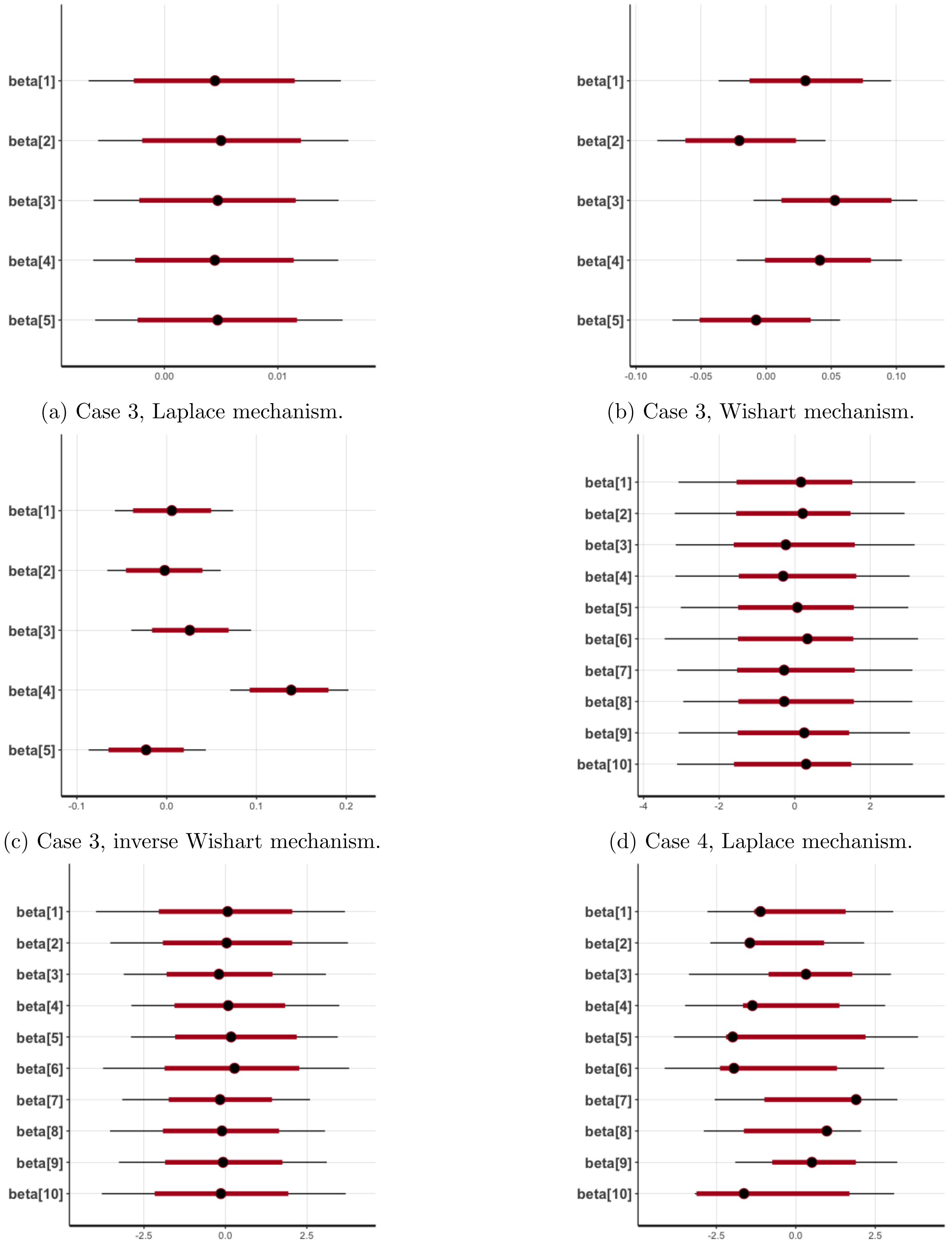
(c) Case 1, inverse Wishart mechanism.

(d) Case 2, Laplace mechanism.

(e) Case 2, Wishart mechanism.

(f) Case 2, inverse Wishart mechanism.

Figure 4: The 95% credible interval of posterior coefficients (part 1).



(a) Case 3, Laplace mechanism.

(b) Case 3, Wishart mechanism.

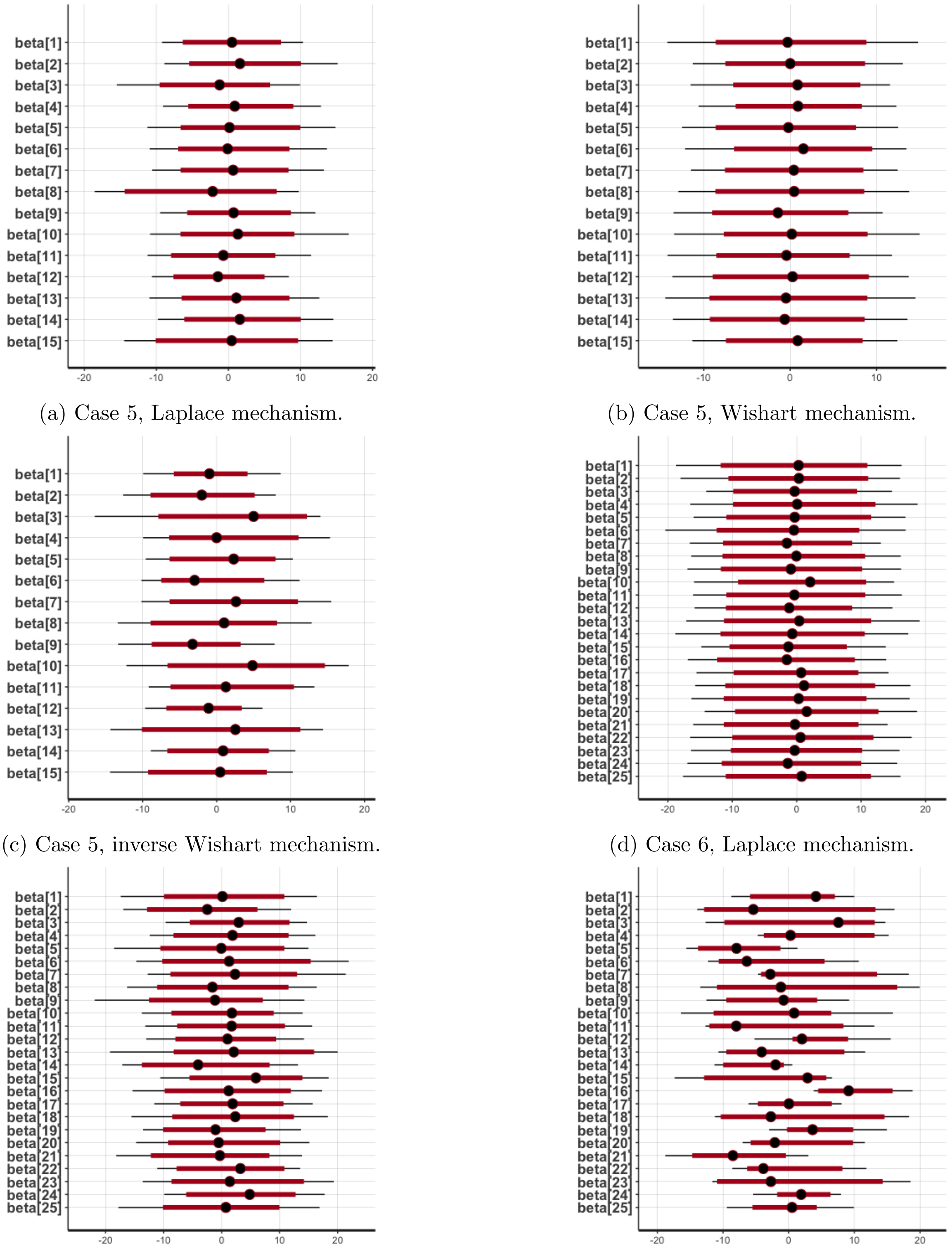
(c) Case 3, inverse Wishart mechanism.

(d) Case 4, Laplace mechanism.

(e) Case 4, Wishart mechanism.

(f) Case 4, inverse Wishart mechanism.

Figure 5: The 95% credible interval of posterior coefficients (part 2).



(a) Case 5, Laplace mechanism.

(b) Case 5, Wishart mechanism.

(c) Case 5, inverse Wishart mechanism.

(d) Case 6, Laplace mechanism.

(e) Case 6, Wishart mechanism.

(f) Case 6, inverse Wishart mechanism.

Figure 6: The 95% credible interval of posterior coefficients (part 3).