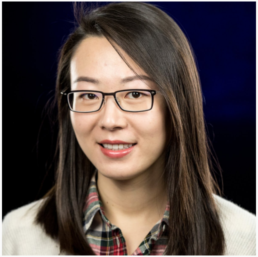
Data Augmentation MCMC for Bayesian Inference from Privatized Data

Jordan Awan

September 12, 2022

Department of Statistics, Purdue University

Collaborators



Nianqiao (Phyllis) Ju	**Ruobin (Robin) Gong**	**Vinayak Rao**
Purdue University	Rutgers University	Purdue University

Sept 28, 2019

CIVIC » Best Countries Best States Healthiest Communities Cities The Civic Report Photos News America 2020

HOME / CIVIC / U.S. NEWS®

Researchers Question Census Bureau's New Approach to Privacy

The U.S. Census Bureau is creating tighter privacy controls in response to new fears about prying by data snoopers.

By **Associated Press**, Wire Service Content Sept. 28, 2019, at 2:23 p.m.

To preserve privacy, the Bureau will use Differential Privacy by adding statistical "noise" to the 2020 data.

"But . . . social scientists, redistricting experts and others worry that it will make next year's census less accurate. They say the bureau's response is overkill."

DP Definition

What is Differential Privacy?

- We have a database $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$, where X_i is the private information of an individual
- We wish to output some (randomized) statistic T, a function of X
- We do not want the output to depend (much) on one individual
- Note that the mechanism M itself is not secret

Definition (Differential Privacy: DMNS06)
For $\epsilon > 0$, a mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ϵ-DP if for all

- databases $X, X' \in \mathcal{X}^n$ differing in one entry,
- all subsets $B \subset \mathcal{Y}$,

$$P(M(X) \in B) \leq e^{\epsilon} \cdot P(M(X') \in B)$$

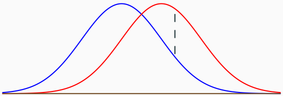
What is Differential Privacy?

- We have a database $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$, where X_i is the private information of an individual
- We wish to output some (randomized) statistic T, a function of X
- We do not want the output to depend (much) on one individual
- Note that the mechanism M itself is not secret

Definition (Differential Privacy: DMNS06)
For $\epsilon > 0$, a mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ϵ-DP if for all

- databases $X, X' \in \mathcal{X}^n$ differing in one entry,
- all subsets $B \subset \mathcal{Y}$,

$$P(M(X) \in B) \leq e^{\epsilon} \cdot P(M(X') \in B)$$



Two Statistical Frameworks

X is truth:

$$Z_{dp}|X \sim \eta(z_{dp} \mid X)$$

- Infer X based on Z_{dp}.
- η is the only source of randomness

X is a sample:

$$X_i \mid \theta \overset{\text{iid}}{\sim} f(x \mid \theta)$$
$$Z_{dp} \mid X \sim \eta(z_{dp} \mid X)$$

- Infer θ based on Z_{dp}.
- Randomness comes from η and f

Both are important frameworks, but require different techniques

The Marginal Likelihood

The likelihood function is a central concept in statistical inference, both for frequentist and Bayesian statistics.

The marginal likelihood is [WM10]

$$\mathcal{L}(\theta \mid Z_{dp}) = \int_{x \in \mathcal{X}^n} \eta(Z_{dp} \mid x) f^n(x \mid \theta)\, dx$$

The posterior distribution is

$$\pi(\theta \mid Z_{dp}) \propto \pi(\theta) \mathcal{L}(\theta \mid Z_{dp})$$

- If each individual contributes a d-dimensional vector, and there are n individuals, integrate over space of size nd

Prior Approaches

- Integrate exactly [AS18, AS20]
- Parametric bootstrap [FWS20]
- Asymptotic approximation [WKLK18]
- Variational approximation [KKS16]
- MCMC with latent sufficient statistics [BS18, BS19]

Solutions are typically either approximations,
or are only suitable for specific settings.

A General Gibbs Sampler

We propose a general Gibbs sampler, that targets the exact posterior distribution $\pi(\theta \mid Z_{dp}) \propto \pi(\theta)\mathcal{L}(\theta \mid Z_{dp})$.

- Only requires
  - the ability to sample from the model $f(x \mid \theta)$,
  - ability to evaluate $\eta(z_{dp} \mid x)$, and
  - any sampler for the non-private posterior distribution $\pi(\theta \mid X)$.
- User friendly – No tuning parameters
- Provably efficient – lower bound on acceptance probability
- Higher privacy means more efficiency
- Apply our sampler to privatized log-linear and linear regression models.

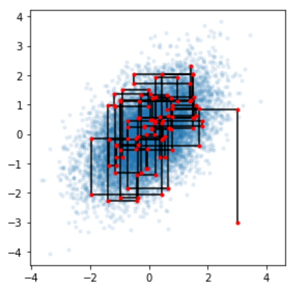
A Traditional Gibbs Sampler

- Our observed data is Z_{dp}
- Our parameters are $\omega = (\theta, X_1, \ldots, X_n)$.

Iterate the following steps:

1: Sample $\theta \mid (X, Z_{dp}) \stackrel{d}{=} \theta \mid X$
2: **for** $i = 1, \ldots, n$ **do**
3: Sample $X_i \mid (X_{-i}, \theta, Z_{dp})$
4: **end for**

- $\theta \mid X$ is the non-private posterior. We assume we can sample
- The challenge is sampling $X_i \mid (X_{-i}, \theta, Z_{dp})$.



https://jessicastringham.net/2018/05/09/gibbs-sampling/

Tailoring the Sampler for DP

Idea: $X_i \mid (X_{-i}, \theta, Z_{dp})$ is mechanism specific, but $X_i \mid \theta$ is mechanism independent.

However, since it is not the correct distribution, accept the proposed sample with probability

$$\min\left\{\frac{\eta(Z_{dp} \mid X_1, \ldots, X_i', \ldots, X_n)}{\eta(Z_{dp} \mid X_1, \ldots, X_i, \ldots, X_n)}, 1\right\}$$

If η satisfies ϵ-DP, acceptance probability is greater than $\exp(-\epsilon)$.

For $\epsilon = 1$, this ensures $\approx 36.7\%$ acceptance rate!

https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/applyingdifferentialprivacyprotect

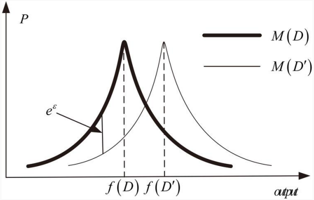
Tailoring the Sampler for DP

Idea: $X_i \mid (X_{-i}, \theta, Z_{dp})$ is mechanism specific, but $X_i \mid \theta$ is mechanism independent.

However, since it is not the correct distribution, accept the proposed sample with probability

$$\min\left\{\frac{\eta(Z_{dp} \mid X_1, \ldots, X_i', \ldots, X_n)}{\eta(Z_{dp} \mid X_1, \ldots, X_i, \ldots, X_n)}, 1\right\}$$

If η satisfies ϵ-DP, acceptance probability is greater than $\exp(-\epsilon)$.



For $\epsilon = 1$, this ensures $\approx 36.7\%$ acceptance rate!

https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/applyingdifferentialprivacyprotect

Evaluating the Acceptance Threshold

- One may be worried that computing $\eta(Z_{dp} \mid X_1, \ldots, X_i', \ldots, X_n)$ would take $O(n)$ time.
- However, many mechanisms only depend on an empirical quantity of the form $T(Z_{dp}, X) = \sum_{i=1}^{n} t(Z_{dp}, X_i)$

  (e.g., empirical risk minimization, sufficient statistics of exponential families)
- Updating $T(Z_{dp}, X)$ to $T(Z_{dp}, X')$ takes $O(1)$ time if X and X' are adjacent.
- One round of our sampler takes $O(n)$ time, same as a non-private sampler

Mixing

- A lower bound $\exp(-\epsilon)$ on the acceptance rate ensures that the chain mixes well.
- As ϵ decreases, acceptance rate improves (most important case!)
- The average acceptance rate is observed to be significantly better
- Under mild assumptions, we show that our sampler is *ergodic*

Application: Naive Bayes

- $X = (X_1, \dots, X_K)$ are *features*, each taking values in $\{1, \dots, J_K\}$
- $Y \in \{1, \dots, I\}$ is the *class*
- The non-private data consists of n i.i.d. copies of (X, Y).
- We are interested in estimating $P(Y \mid X)$.
- The *Naive Bayes Classifier* assumes $P(X \mid Y) = \prod_{k=1}^{K} P(X_k \mid Y)$
- Release $Z_{dp} = \{n_{ijk} + \text{Laplace}(2K/\epsilon)\}_{ijk}$, the noisy counts

		X_1		X_2			X_K	
		1	2	1	2		1	2
Y	1	n^1_{11}	n^1_{12}	n^2_{11}	n^2_{12}	$\dots$	n^K_{11}	n^K_{12}
	2	n^1_{21}	n^1_{22}	n^2_{21}	n^2_{22}		n^K_{21}	n^K_{22}

TABLE 1
Sufficient statistics of the Naive Bayes model.

		X_1		X_2			X_K	
		1	2	1	2		1	2
Y	1	p^1_{11}	p^1_{12}	p^2_{11}	p^2_{12}	$\dots$	p^K_{11}	p^K_{12}
	2	p^1_{21}	p^1_{22}	p^2_{21}	p^2_{22}		p^K_{21}	p^K_{22}

TABLE 2
An example of the parameters of the Naive Bayes model for a $2 \times 2 \times K$ table.

Figure from [KKS16]

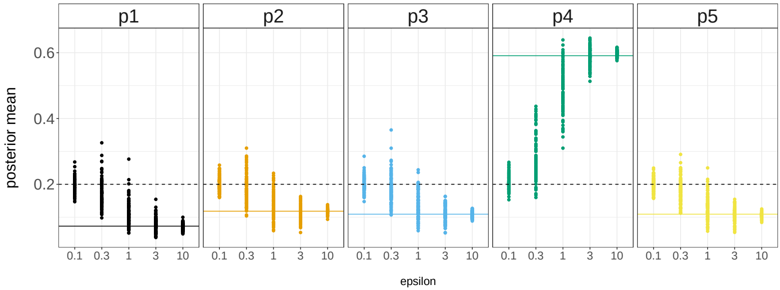
Simulation Setup

For the simulation, set

- $N = 100$ (number of samples)
- $I = 5$ (number of classes)
- $K = 5$ (number of questions)
- $J_K = 3$ (possible answers)
- $\epsilon \in \{.1, .3, 1, 3, 10\}$
- Prior for all parameters $\text{Dirichlet}(2, \ldots, 2)$.

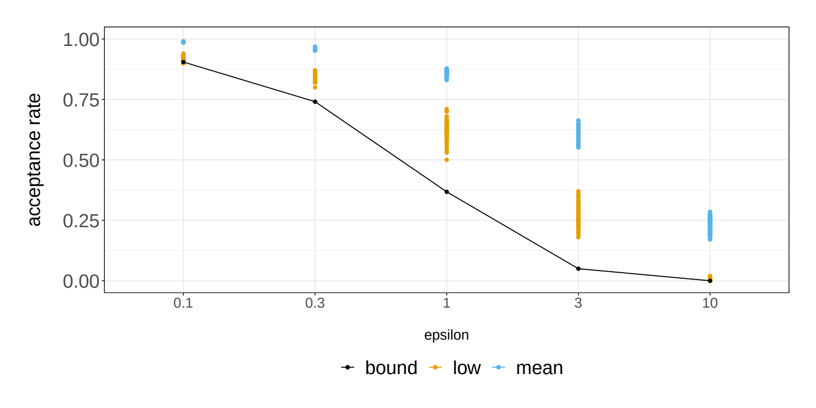
Posterior Mean

- Fix a non-private dataset
- Run 100 chains at each ϵ value
- each chain ran for 10,000 iterations.
- For each chain, calculate posterior mean



Acceptance Rate

- 100 chains at each ϵ value
- each chain ran for 10,000 iterations.
- For each chain, calculate minimum and mean acceptance rate



Coverage: Naive Bayes

We consider the frequentist coverage of a 90% credible interval for the probabilities $P(Y = i)$ for $i = 1, \ldots, 5$. 100 replicates per ϵ value.

ϵ	$p_1 = .097$	$p_2 = .148$	$p_3 = .145$	$p_4 = .446$	$p_5 = .163$
.1	1	1	1	**.36**	1
.3	.97	1	1	**.59**	1
1	.94	.99	.97	**.83**	.98
3	.95	.91	.97	.89	.93
10	.92	.88	.94	.92	.9

Table 1: Coverage of $p_i = P(Y = i)$ for different ϵ. Average $= .914$. For $\epsilon = .1$, average $= .872$.

Application: Linear Regression

- Observe n i.i.d. copies of (x_0^i, y^i)
- Write $x = (\underline{1}, x_0)$ for the design matrix, where x_0 are the
- Model the response as $y|x \sim N(x\beta, \sigma^2 I_n)$
- Model the predictors as $x_0^i \sim N(\mu, \Sigma^2)$
- Before adding noise for privacy, we first clamp the predictors and response, and then normalize them to take values in $[-1, 1]$:

$$\widetilde{x}_0^i := (b_i - a_i)^{-1} 2([x_0^i]_{a_i}^{b_i} - a_i) - 1$$

$$\widetilde{y} := (b_y - a_y)^{-1} 2([y]_{a_y}^{b_y} - a_y) - 1.$$

Call

$$\widetilde{x} := [\underline{1}, \widetilde{x}_0^1, \widetilde{x}_0^2, \ldots, \widetilde{x}_0^p]$$

$$Z_{dp} := (\widetilde{x}^\top \widetilde{y}, \widetilde{y}^\top \widetilde{y}, \widetilde{x}^\top \widetilde{x}) + \text{Laplace}.$$

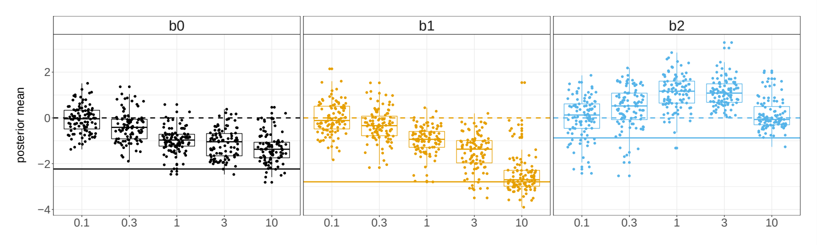
Simulation Setup

For the simulation, set

- $N = 100$ (number of samples)
- $p = 2$ (number of predictors)
- Fixed $\Sigma = I$, $\sigma^2 = 2$
- Sampled $m_j \overset{\text{iid}}{\sim} N(0, 1)$, and then fixed at $(.9, -1.17)$
- $\epsilon \in \{.1, .3, 1, 3, 10\}$
- Prior $\beta_i \overset{\text{iid}}{\sim} N(0, \tau^2 = 2)$

Posterior Mean

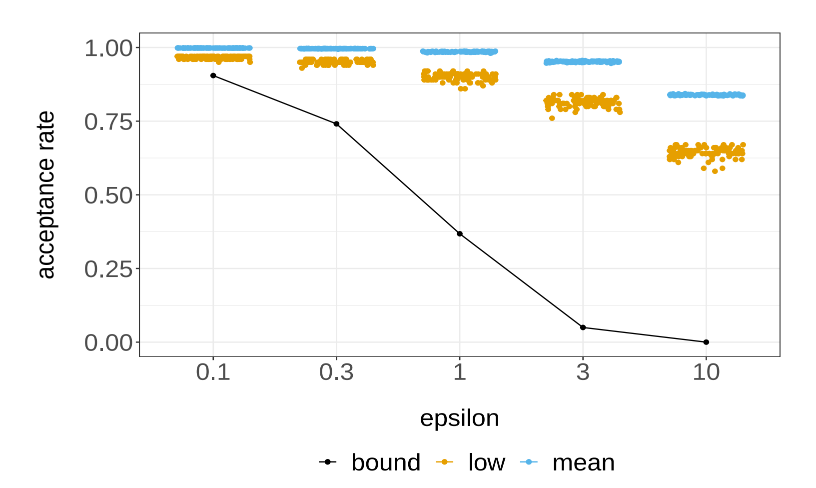
- Fix a non-private dataset
- Run 100 chains at each ϵ value
- each chain ran for 10,000 iterations.
- For each chain, calculate posterior mean



- Loss of information due to clamping remains

Acceptance Rate

- 100 chains at each ϵ value
- each chain ran for 10,000 iterations.
- For each chain, calculate minimum and mean acceptance rate



Conclusions

- A Gibbs sampler with the correct target distribution
- User-friendly implementation: mechanism independent
- Privacy implies efficiency: smaller ϵ gives higher acceptance rate
- Application to a log-linear model
- Application to a linear regression model

Thank You!

Ju, Nianqiao, Jordan Awan, Ruobin Gong, and Vinayak Rao. "Data Augmentation MCMC for Bayesian Inference from Privatized Data." arXiv preprint arXiv:2206.00710 (2022).

jawan@purdue.edu

References

[AS18] Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4208–4218. Curran Associates, Inc., 2018.

[AS20] Jordan Alexander Awan and Aleksandra Slavkovic. Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, 10(1), 2020.

[BS18] Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian inference for exponential families. *Advances in Neural Information Processing Systems*, 31, 2018.

[BS19] Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. *Calibrating Noise to Sensitivity in Private Data Analysis*, pages 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[FWS20] Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. General-purpose differentially-private confidence intervals. *arXiv preprint arXiv:2006.07749*, 2020.

[KKS16] Vishesh Karwa, Dan Kifer, and Aleksandra Slavković. Private posterior distributions from variational approximations. *NIPS 2015 Workshop on Learning and Privacy with Incomplete Data and Weak Supervision*, 2016.

[WKLK18] Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.

[WM10] Oliver Williams and Frank Mcsherry. Probabilistic inference and differential privacy. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2451–2459. Curran Associates, Inc., 2010.