Bayesian Model Uncertainty under Differential Privacy

Víctor Peña

Universitat Politècnica de Catalunya

https://arxiv.org/abs/2109.03949

SEIO 2023, Elche

Joint work with

Felipe Barrientos

Florida State University

Why should we care?

- Confidential data are increasingly common: electronic health records, activity in social media, wearable devices, etc.
- Legislation protecting data privacy is picking up: General Data Protection Regulation (GDPR) in the EU, increasing levels of state regulations in the US (e.g. CO, IA, IN, TN, MO, TX, DE, FL, OR have all passed legislation in 2023)

Why differential privacy?

- Clear mathematical definition
- Strong guarantees against identification and other attacks
- Most widely adopted notion of data privacy in the literature
- Adopted by the US Census Bureau and tech companies like Alphabet (Google, YouTube), Apple, Meta (Facebook, Whatsapp, Instagram), etc.

Some questions

For normal linear models...

- Can we find differentially private Bayes factors, posterior inclusion probabilities, etc. under differential privacy?
- Can we find methods that are easy to implement but still useful?
- Can we define methods that are also model-selection consistent? (i.e. asymptotically, can we retrieve the "true" model, if there is one?)

Differential Privacy

Differential Privacy (DP) in a Nutshell

- There are different flavors of differential privacy (DP) in the literature
- In practice, the implementation consists in perturbing statistics computed with the confidential data
- There is a tradeoff between statistical utility and privacy: if we want more privacy, we need to inject more noise, which reduces the statistical utility of the outputs

Introducing DP

- Two datasets D and $\tilde{D}$ are *neighbors* ($D \sim \tilde{D}$) if they only differ in one row
- Algorithms that ensure differential privacy are referred to as *mechanisms*. They are functions $\mathcal{M}$ that take data D as inputs and output a random $\mathcal{M}(D)$ that is "private"
- If our output is "private", we should have $\mathcal{M}(D) \approx \mathcal{M}(\tilde{D})$ for $D \sim \tilde{D}$. Otherwise, we could use that information to identify the difference between D and $\tilde{D}$

Definition of (ε, δ)-DP

Definition

Let $\varepsilon > 0$ and $0 \leq \delta \leq 1$. A mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy if, for all $D \sim \tilde{D}$ any $\mathcal{M}$-measurable set S, we have

$$P[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) P[\mathcal{M}(\tilde{D}) \in S] + \delta.$$

- Our privacy budget is (ε, δ): the higher the budget, the less privacy we need
- If $\mathcal{M}(D)$ is (ε, δ)-DP, so is $f(\mathcal{M}(D))$ for any function f that doesn't depend on D

Our methods

Linear models under Model Uncertainty

- We work with normal linear models under model uncertainty
- We have p variables and we don't which ones, if any, are active
- We express our model uncertainty through a binary vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)' \in \{0, 1\}^p$ such that $\gamma_j = 0$ if and only if $\beta_j = 0$.
- We use the notation $|\gamma| = \sum_{j=1}^{p} \gamma_j$ for the number of active coefficients in a model.

Linear model given γ

Given a model indexed by γ, we write the linear model as

$$Y = X_0\beta_0 + V_\gamma\beta_\gamma + \sigma W, \;\; W \sim N_n(0_n, I_n),$$

where $\beta_\gamma \in \mathbb{R}^{|\gamma|\times 1}$ is a vector including the β_j such that $\gamma_j = 1$ and $V_\gamma \in \mathbb{R}^{n\times|\gamma|}$ is a matrix with the active variables in γ.

We parameterize the model so that X_0 and V_γ are orthogonal.

Priors on parameters

Given γ, our prior specification is

$$\pi(\beta_0, \sigma^2) \propto 1/\sigma^2,$$

$$\pi(\beta_\gamma \mid \gamma, \sigma^2) = \int_0^\infty N_{|\gamma|}(\beta_\gamma \mid 0_p, g\sigma^2(V'V)^{-1})\, \pi(\mathrm{d}g).$$

The prior measure on g can depend on n and p.

The prior on γ can depend on p, but not on n or the design matrix.

Sufficient statistic

If we define the data matrix

$$D = [V\,;\,(I - P_{X_0})Y]$$

The Gram matrix

$$G = D'D = \begin{bmatrix} V'V & V'Y \\ Y'V & Y'(I - P_{X_0})Y \end{bmatrix}$$

is a sufficient statistic for the normal linear model.

We release a noisy version of G.

Bounding G

- Our methods assume that the data are bounded: all the entries in D lie within a bounded interval $[l, u]$
- We need the assumption because the error we add release a DP version of G depends on the range of the data
- If we know bounds, we can use them. If not, we can standardize the data.

Perturbing the sufficient statistic

- We release $G^* = G + E$, where E is a zero-mean perturbation that ensures DP.
- To ensure ε-DP, the entries of E are drawn from a Laplace distribution (Laplace mechanism)
- For (ε, δ)-DP with $0 < \varepsilon < 1$ and $0 < \delta \leq 1$, E is Wishart-distributed (Sheffet, 2019).
- The scale of the perturbation increases in p and the range of the data

Transforming the perturbed statistic

Two comments about G^*

1. G^* has noise added, so we have to take it into account somehow
2. If we use the Laplace mechanism, G^* may not be positive-definite

For these reasons, we regularize G^*

Regularized G^*

We regularize G^* in two ways:

1. **Hard-thresholding** (off-diagonal):

$$G^{**}_{ij} = G^*_{ij} \mathbb{1}(i = j \text{ or } |G^*_{ij}| \geq e_\lambda),$$

   where e_λ is a large quantile of e_{ij}

2. **Ridge** (diagonal):

$$G^*_r = G^* + rI \text{ or } G^{**}_r = G^{**}_r + rI.$$

   r is big enough to ensure positive-definiteness

Applying the methods

If we didn't have privacy constraints,

$$B_{\gamma 0} = \int_0^\infty \frac{(g+1)^{(n-p-p_0)/2}}{[1+g(1-R^2_\gamma)^{(n-p_0)/2}]}\,\pi(\mathrm{d}g).$$

We plug-in G^*_r or G^{**}_r as our estimates of G:

$$B_{\gamma 0,\mathrm{DP}} = \int_0^\infty \frac{(g+1)^{(n-p-p_0)/2}}{[1+g(1-R^2_{\gamma,\mathrm{DP}})^{(n-p_0)/2}]}\,\pi(\mathrm{d}g),$$

where $R^2_{\gamma,\mathrm{DP}}$ comes from G^*_r or G^{**}_r

How does this work in practice?

1. Users obtain G^*, which is DP
2. They compute $\mathcal{G}$, which is G_r^* or G_r^{**} (they could compute both)
3. With $\mathcal{G}$, they can find $B_{\gamma 0,\mathrm{DP}}$ and other quantities of interest

Easy implementation!

- When X_0 is an intercept, we can compute a synthetic dataset $\mathcal{D}$ whose Gram matrix is $\mathcal{G}$ with the formula

$$\mathcal{D} = \mathcal{M}(\mathcal{M}'\mathcal{M})^{-1/2}\mathcal{G}^{1/2}$$
$$\mathcal{M} = (I_n - 1_n 1_n'/n)\mathcal{U},$$

  where $\mathcal{U} \in \mathbb{R}^{n\times(p+1)}$ is an arbitrary full-rank matrix.

- Given $\mathcal{D}$, we can run the usual `R` packages to obtain $B_{\gamma 0,\mathrm{DP}}$ and any other statistic of interest

Model-selection consistency

The Bayes factors are misspecified:

1. We assume the data are bounded, and the Bayes factors are derived assuming normality
2. The Gram matrix G has been replaced by an estimate $\mathcal{G}$

Despite this fact, they are model-selection consistent under common assumptions!

Empirical results

- In our article, we did simulation studies and an application
- Hard-thresholding is especially useful if the truth is sparse
- We recommend comparing results with G_r^{**} (hard-thresholding) and G_r^*

Conclusions

Conclusions

- We propose releasing a perturbed Gram matrix that satisfies DP. Then, we transform it and plug-in the result into standard machinery
- The methods are consistent and work fairly well in practice
- Interesting extensions:
  - GLMs, survival models, etc. Work with approximate sufficient statistics instead.
  - Build a Bayesian model for G

Thanks!

Assumptions for consistency (1 out of 2)

- **Boundedness:** The data D are within the interval $[l, u]$ for finite l and u.
- **Regression mean and variance:** $\mathbb{E}(Y \mid X_0, V) = X_0\psi + V_T\beta_T$ and $\mathrm{Var}(Y \mid X_0, V) = \sigma_T^2 I_n$, where $V_T \in \mathbb{R}^{n \times p_T}$ is a matrix that contains the truly active predictors.
- **Privacy parameters:** The privacy parameters ε and δ are such that the matrix perturbation term $E/n \to_P 0$.

Assumptions for consistency (2 out of 2)

- **Regularization parameters:** e_λ is fixed and r is so that $\lim_{n\to\infty} r/n = 0$.
- **Design matrices:** $\lim_{n\to\infty} V'V/n = S_1$, where S_1 is symmetric and positive-definite.
- **Priors on g:** the prior $\pi(g)$ satisfies

$$\lim_{n\to\infty} \int_0^\infty n^{p/2} \, (g+1)^{-|\gamma|/2} \pi(g) \, \nu(\mathrm{d}g) < \infty$$

$$\lim_{n\to\infty} \int_n^\infty n^{p/2} (g+1)^{-|\gamma|/2} \pi(g) \, \nu(\mathrm{d}g) > 0.$$