Exact Statistical Inference for Differentially Private Data

Ruobin Gong

Rutgers University

U.S. Census Bureau

Jan 6, 2020

Differential privacy should be – and can be – modeled

- Statistical disclosure limitation mechanisms compliant with DP guarantee privacy with **provability** and **transparency**.
- Transparency enables **accurate statistical modeling** of the DP mechanism. This is the best way to ensure correctness in the resulting inference, when a (calculated) loss of statistical efficiency is present in the data.

Differential privacy: preliminaries

Definition (Dwork & Smith, 2009)

A random function $\boldsymbol{S} : \mathcal{X} \to \mathbb{R}^p$ is (ϵ, δ)-differentially private if for all neighboring datasets $\{(\boldsymbol{x}, \boldsymbol{x}') : d(\boldsymbol{x}, \boldsymbol{x}') = 1\}$ and all $A \in \mathscr{B}(\mathbb{R}^p)$,

$$Pr\left(\boldsymbol{S}\left(\boldsymbol{x}'\right) \in A\right) \leq e^{\epsilon} Pr\left(\boldsymbol{S}\left(\boldsymbol{x}\right) \in A\right) + \delta.$$

$\boldsymbol{S}$ is called ϵ-differentially private if it is $(\epsilon, 0)$-differentially private. ϵ and δ are called the *privacy loss budget*.

DP mechanism: output perturbation

For a dataset $\boldsymbol{x} \in \mathcal{X}$ and a deterministic function $\boldsymbol{s} : \mathcal{X} \to \mathbb{R}^p$, the random function $\boldsymbol{S}$ is a **perturbation mechanism** based on $\boldsymbol{s}$, if

$$\boldsymbol{S}(\boldsymbol{x}) \mid \boldsymbol{s}(\boldsymbol{x}) \sim \eta_{\text{dp}}(\cdot \mid \boldsymbol{s}(\boldsymbol{x})),$$

where η_{dp} is known and $\mathbb{E}(\boldsymbol{S}(\boldsymbol{x}) \mid \boldsymbol{s}(\boldsymbol{x})) = \boldsymbol{s}(\boldsymbol{x})$.

DP mechanism: output perturbation

For a dataset $\boldsymbol{x} \in \mathcal{X}$ and a deterministic function $\boldsymbol{s} : \mathcal{X} \to \mathbb{R}^p$, the random function $\boldsymbol{S}$ is a **perturbation mechanism** based on $\boldsymbol{s}$, if

$$\boldsymbol{S}(\boldsymbol{x}) \mid \boldsymbol{s}(\boldsymbol{x}) \sim \eta_{\mathrm{dp}}(\cdot \mid \boldsymbol{s}(\boldsymbol{x})),$$

where η_{dp} is known and $\mathbb{E}(\boldsymbol{S}(\boldsymbol{x}) \mid \boldsymbol{s}(\boldsymbol{x})) = \boldsymbol{s}(\boldsymbol{x})$. As a special case, $\boldsymbol{S}$ is said to be an **additive perturbation mechanism** if

$$\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{s}(\boldsymbol{x}) + h\boldsymbol{u}.$$

- $\boldsymbol{u}$: noise component with kernel density η and $\mathbb{E}(\boldsymbol{u}) = \mathbf{0}$, e.g. (multi-dimensional) Laplace, Normal, t, etc;
- $h = h(\epsilon, \delta, \boldsymbol{s}) > 0$: bandwidth parameter chosen as a function of the privacy loss budget (ϵ, δ) and summary function $\boldsymbol{s}(\cdot)$.

Private perturbation mechanisms: examples

$$\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{s}(\boldsymbol{x}) + h\boldsymbol{u}$$

1. ϵ-DP Laplace mechanism (Dwork et al., 2006):
   - $\boldsymbol{u} \sim \text{Lap}_p(1)$, a standard p-product Laplace,
   - $h = \epsilon^{-1} GS(\boldsymbol{s})$, where $GS(\boldsymbol{s})$ is the *global sensitivity* of $\boldsymbol{s}$.
2. (ϵ, δ)-DP Laplace mechanism (Nissim et al., 2007):
   - $\boldsymbol{u} \sim \text{Lap}_p(1)$,
   - $h = \epsilon^{-1} SS_\xi(t, \boldsymbol{x})$, where $SS_\xi(\boldsymbol{s}, \boldsymbol{x})$ is the ξ*-smooth sensitivity* of $\boldsymbol{s}$ at $\boldsymbol{x}$; $\xi = \epsilon \left\{4\left(d + \log\left(2/\delta\right)\right)\right\}^{-1} > 0$
3. (ϵ, δ)-DP Gaussian mechanism (Nissim et al., 2007):
   - $\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{I}_p)$,
   - $h = \epsilon^{-1} 5\sqrt{2\log(2/\delta)} SS_\xi(t, \boldsymbol{x})$, $\xi = \epsilon \left\{4\left(d + \log\left(2/\delta\right)\right)\right\}^{-1}$.

DP mechanism should not be ignored

Suppose a simple linear model between vector counts $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{e}.$$

Ordinary least squares produce consistent estimators

$$\hat{\beta}_0 \longrightarrow \beta_0, \qquad \hat{\beta}_1 \longrightarrow \beta_1.$$

DP mechanism should not be ignored

Suppose a simple linear model between vector counts $\boldsymbol{x}$ and $\boldsymbol{y}$:

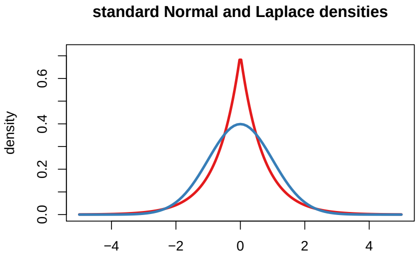
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{e}.$$

Ordinary least squares produce consistent estimators

$$\hat{\beta}_0 \longrightarrow \beta_0, \qquad \hat{\beta}_1 \longrightarrow \beta_1.$$

Treating $(\boldsymbol{x}, \boldsymbol{y})$ with ϵ-DP mechanism,

$$\boldsymbol{y}_{\text{dp}} = \boldsymbol{y} + \boldsymbol{w}, \qquad \boldsymbol{x}_{\text{dp}} = \boldsymbol{x} + \boldsymbol{z}, \quad \boldsymbol{w}, \boldsymbol{z} \sim Lap_n\left(\epsilon^{-1}\right)$$



DP mechanism should not be ignored

Naïvely fitting the original model to differentially privatized data

$$\boldsymbol{y}_{\mathrm{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\mathrm{dp}} + \boldsymbol{e},$$

the resulting least squares estimates will miss the mark:

$$\hat{\beta}_0^{\mathrm{dp}} \approx \beta_0 \underbrace{+\, \alpha_{x,z}\beta_1}, \qquad \hat{\beta}_1^{\mathrm{dp}} \approx \beta_1 \underbrace{-\, \gamma_{x,z}\beta_1},$$

where

$$\alpha_{x,z} = \gamma_{x,z}\bar{x} + (1 - \gamma_{x,z})\bar{z}, \qquad \gamma_{x,z} = \frac{\mathrm{SS}_{xz} + \mathrm{SS}_z}{\mathrm{SS}_{x+z}} \in (0, 1)\,.$$

DP mechanism should not be ignored

Naïvely fitting the original model to differentially privatized data

$$\boldsymbol{y}_{\text{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\text{dp}} + \boldsymbol{e},$$

the resulting least squares estimates will miss the mark:

$$\hat{\beta}_0^{\text{dp}} \approx \beta_0 \underbrace{+\, \alpha_{x,z}\beta_1}, \qquad \hat{\beta}_1^{\text{dp}} \approx \beta_1 \underbrace{-\, \gamma_{x,z}\beta_1},$$

where

$$\alpha_{x,z} = \gamma_{x,z}\bar{x} + (1 - \gamma_{x,z})\bar{z}, \qquad \gamma_{x,z} = \frac{\text{SS}_{xz} + \text{SS}_z}{\text{SS}_{x+z}} \in (0, 1)\,.$$

Ignoring the DP mechanism results in misguided inference:

- $\hat{\beta}_0^{\text{dp}}, \hat{\beta}_1^{\text{dp}}$ are systematically biased;
- Strength of association between $(\boldsymbol{x}, \boldsymbol{y})$ is underestimated (*attenuation* in the measurement error literature);
- Both estimates suffer inflated variance.

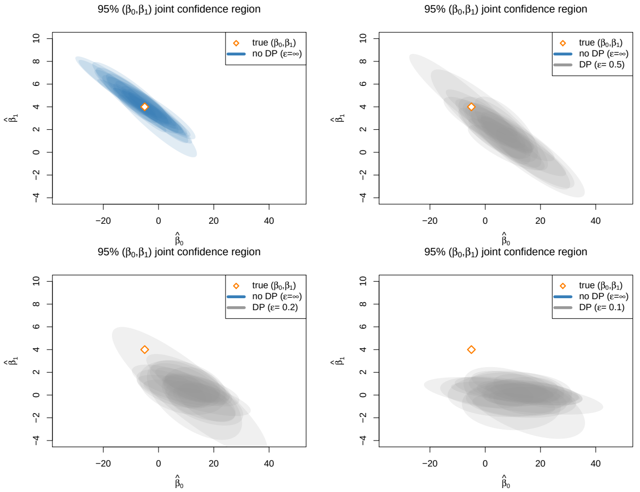


Figure: Naíve fitting with $\boldsymbol{x}_i \sim Pois\,(10)$, $\boldsymbol{y}_i = -5 + 4\boldsymbol{x}_i + \boldsymbol{e}_i$, $\boldsymbol{e}_i \sim N\left(0, 5^2\right)$, $n = 10$, at privacy budget levels $\epsilon = 0.5, 0.2, 0.1$, and ∞ (no privacy). Smaller ϵ induces more misguided confidence regions for (β_0, β_1). Each panel depicts 20 simulations.

DP mechanism should be modeled

A model adequate for the confidential data $\boldsymbol{s} = (\boldsymbol{x}, \boldsymbol{y})$, if naïvely fitted to the privatized data $\boldsymbol{s}_{\text{dp}} = (\boldsymbol{x}_{\text{dp}}, \boldsymbol{y}_{\text{dp}})$, will almost certainly be inadequate:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{e} \qquad \not\Longrightarrow \qquad \boldsymbol{y}_{\text{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\text{dp}} + \boldsymbol{e}.$$

Instead, **augment** the original model with the DP mechanism:

$$\Longrightarrow \qquad \left(\boldsymbol{y}_{\text{dp}} - \boldsymbol{w}\right) = \beta_0 + \beta_1 \left(\boldsymbol{x}_{\text{dp}} - \boldsymbol{z}\right) + \boldsymbol{e}, \quad \boldsymbol{w}, \boldsymbol{z} \sim \mathit{Lap}\left(\epsilon^{-1}\right)$$

DP mechanism should be modeled

A model adequate for the confidential data $\boldsymbol{s} = (\boldsymbol{x}, \boldsymbol{y})$, if naïvely fitted to the privatized data $\boldsymbol{s}_{\text{dp}} = (\boldsymbol{x}_{\text{dp}}, \boldsymbol{y}_{\text{dp}})$, will almost certainly be inadequate:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{e} \qquad \not\Longrightarrow \qquad \boldsymbol{y}_{\text{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\text{dp}} + \boldsymbol{e}.$$

Instead, **augment** the original model with the DP mechanism:

$$\Longrightarrow \qquad \left(\boldsymbol{y}_{\text{dp}} - \boldsymbol{w}\right) = \beta_0 + \beta_1 \left(\boldsymbol{x}_{\text{dp}} - \boldsymbol{z}\right) + \boldsymbol{e}, \quad \boldsymbol{w}, \boldsymbol{z} \sim Lap\left(\epsilon^{-1}\right)$$

A general construction

Likelihood for $\boldsymbol{\beta}$ based on privatized data $\boldsymbol{s}_{\text{dp}}$ (observed) is integrated over the confidential data $\boldsymbol{s}$ (missing), with respect to the DP mechanism:

$$L\left(\boldsymbol{\beta}; \boldsymbol{s}_{\text{dp}}\right) = \int \underbrace{\eta_{\text{dp}}\left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}\right)}_{\text{DP mechanism}} \underbrace{\pi\left(\boldsymbol{s} \mid \boldsymbol{\beta}\right)}_{\text{original model}} \partial \boldsymbol{s}$$

Transparency of the DP mechanism enables accurate modeling.

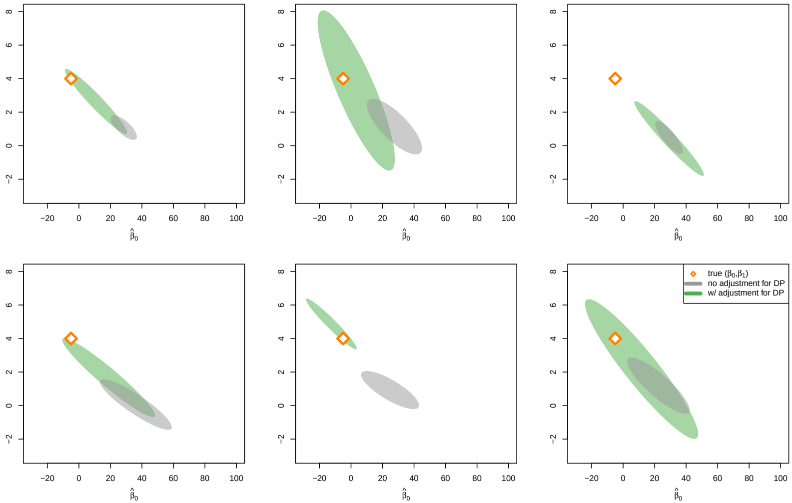


Figure: Correct model (green) fitted via Monte Carlo EM (G. 2019) vs. naïve model (gray) on six instances of DP protected datasets ($\epsilon = 0.2$). Displayed 95% confidence ellipses are based on normal approximations at the MLE.

Approximate computation in Bayesian inference

A Bayesian model is posited:

- prior: $\theta \sim \pi_0(\theta)$
- likelihood: $\boldsymbol{x} \mid \theta \sim \pi(\boldsymbol{x} \mid \theta)$
- posterior:

$$\pi\left(\theta \mid \boldsymbol{x}\right) \propto \pi_0\left(\theta\right) \pi\left(\boldsymbol{x} \mid \theta\right)$$

Approximate computation in Bayesian inference

A Bayesian model is posited:

- prior: $\theta \sim \pi_0(\theta)$
- likelihood: $\boldsymbol{x} \mid \theta \sim \pi(\boldsymbol{x} \mid \theta)$
- posterior:

$$\pi\left(\theta \mid \boldsymbol{x}\right) \propto \pi_0\left(\theta\right) \pi\left(\boldsymbol{x} \mid \theta\right)$$

Sampling from the posterior via Monte Carlo requires that it at least can be evaluated. This is not the case for complex models.

- Case in point: intractable or implicit likelihood $\pi\left(\boldsymbol{x} \mid \theta\right)$ (e.g. the Lokta-Volterra/predator-prey model)

A "likelihood-free" method

ALGORITHM 1

Input: observed data $\boldsymbol{x}_0$, integer $N > 0$;

Iterate: for $i = 1, \ldots, N$:

  step 1, simulate $\theta_i \sim \pi_0(\theta)$;

  step 2, simulate $\boldsymbol{x}_i \sim \pi(\boldsymbol{x} \mid \theta_i)$;

  step 3, accept θ_i if $\boldsymbol{x}_i = \boldsymbol{x}_0$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^N$.

A "likelihood-free" method

Algorithm 1

Input: observed data $\boldsymbol{x}_0$, integer $N > 0$;

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\theta_i \sim \pi_0(\theta)$;

step 2, simulate $\boldsymbol{x}_i \sim \pi(\boldsymbol{x} \mid \theta_i)$;

step 3, accept θ_i if $\boldsymbol{x}_i = \boldsymbol{x}_0$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^N$.

Algorithm 1 draws $\theta_i \sim \pi(\theta \mid \boldsymbol{x}_0)$, i.i.d.

A “likelihood-free” method

ALGORITHM 1

Input: observed data $\boldsymbol{x}_0$, integer $N > 0$;

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\theta_i \sim \pi_0(\theta)$;

step 2, simulate $\boldsymbol{x}_i \sim \pi(\boldsymbol{x} \mid \theta_i)$;

step 3, accept θ_i if $\boldsymbol{x}_i = \boldsymbol{x}_0$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^N$.

Algorithm 1 draws $\theta_i \sim \pi(\theta \mid \boldsymbol{x}_0)$, i.i.d.

However, exact matching $\boldsymbol{x}_i = \boldsymbol{x}_0$ may not be practical.

- $\boldsymbol{x}_0$ may not be discrete;
- $\boldsymbol{x}_0$ may be high dimensional.

Approximate Bayesian Computation (ABC)

ALGORITHM 2

Input: observed summary data $\boldsymbol{s}_0 = \boldsymbol{s}(\boldsymbol{x}_0)$, integer $N > 0$,

a kernel density η with bandwidth $h > 0$;

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\theta_i \sim \pi_0(\theta)$;

step 2, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s}(\boldsymbol{x}) \mid \theta_i)$;

step 3, accept θ_i with probability $c\eta\left(\left(\boldsymbol{s}_i - \boldsymbol{s}_0\right)/h\right)$

where $c^{-1} = \max\{\eta(\cdot)\}$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^{N}$.

Approximate Bayesian Computation (ABC)

ALGORITHM 2

Input: observed summary data $\boldsymbol{s}_0 = \boldsymbol{s}(\boldsymbol{x}_0)$, integer $N > 0$,
a kernel density η with bandwidth $h > 0$;

Iterate: for $i = 1, \dots, N$:

step 1, simulate $\theta_i \sim \pi_0(\theta)$;

step 2, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s}(\boldsymbol{x}) \mid \theta_i)$;

step 3, accept θ_i with probability $c\eta\left((\boldsymbol{s}_i - \boldsymbol{s}_0)/h\right)$
where $c^{-1} = \max\{\eta(\cdot)\}$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^N$.

$\theta_i \sim \pi_{ABC}(\theta \mid \boldsymbol{s}_0)$: two layers of approximation

1. From $\pi(\theta \mid \boldsymbol{x}_0)$ to $\pi(\theta \mid \boldsymbol{s}_0)$: choice of $\boldsymbol{s}(\cdot)$;
2. From $\pi(\theta \mid \boldsymbol{s}_0)$ to $\pi_{ABC}(\theta \mid \boldsymbol{s}_0)$: choice of $\eta(\cdot)$ and h

Modeling differentially private queries

The Bayesian model is modified to:

- prior: $\theta \sim \pi_0(\theta)$
- confidential query likelihood: $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$
- privacy mechanism: $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}, \cancel{\theta} \sim \eta_{\text{dp}} \left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}\right)$ $\leftarrow$ ignorability
- observed/private posterior:

$$\pi\left(\theta \mid \boldsymbol{s}_{\text{dp}}\right) = \frac{\pi_0\left(\theta\right) \int \eta_{\text{dp}}(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s}}{\int \pi_0\left(\theta\right) \int \eta_{\text{dp}}(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s} d\theta}.$$

Modeling differentially private queries

The Bayesian model is modified to:

- prior: $\theta \sim \pi_0(\theta)$
- confidential query likelihood: $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$
- privacy mechanism: $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}, \cancel{\theta} \sim \eta\left(\left(\boldsymbol{s}_{\text{dp}} - \boldsymbol{s}\right)/h\right)$, if additive
- observed/private posterior:

$$\pi\left(\theta \mid \boldsymbol{s}_{\text{dp}}\right) = \frac{\pi_0\left(\theta\right) \int \eta\left(\left(\boldsymbol{s}_{\text{dp}} - \boldsymbol{s}\right)/h\right) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s}}{\int \pi_0\left(\theta\right) \int \eta\left(\left(\boldsymbol{s}_{\text{dp}} - \boldsymbol{s}\right)/h\right) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s} d\theta}.$$

ABC produces exact posterior draws for DP data

Algorithm 3

Input: private query $\boldsymbol{s}_{\text{dp}} = \boldsymbol{S}(\boldsymbol{x}_0)$, integer $N > 0$, perturbation
mechanism w/ density η and bandwidth $h(\epsilon, \delta, \boldsymbol{s}) > 0$;

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\theta_i \sim \pi_0(\theta)$;

step 2, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s} \mid \theta_i)$;

step 3, accept θ_i with probability $c\eta\left(\left(\boldsymbol{s}_{\text{dp}} - \boldsymbol{s}_i\right)/h\right)$
where $c^{-1} = \max\{\eta(\cdot)\}$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^N$.

ABC produces exact posterior draws for DP data

ALGORITHM 3

Input: private query $\boldsymbol{s}_{\text{dp}} = \boldsymbol{S}(\boldsymbol{x}_0)$, integer $N > 0$, perturbation mechanism w/ density η and bandwidth $h(\epsilon, \delta, \boldsymbol{s}) > 0$;

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\theta_i \sim \pi_0(\theta)$;

step 2, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s} \mid \theta_i)$;

step 3, accept θ_i with probability $c\eta\left(\left(\boldsymbol{s}_{\text{dp}} - \boldsymbol{s}_i\right)/h\right)$ where $c^{-1} = \max\{\eta(\cdot)\}$, otherwise go to step 1;

Output: a set of parameter values $\{\theta_i\}_{i=1}^N$.

Theorem (G. 2019)

Algorithm 3 draws $\theta_i \sim \pi(\theta \mid \boldsymbol{s}_{\text{dp}})$, i.i.d.

* Noisy ABC (Fearnhead & Prangle, 2012);
* ABC under the assumption of model error (Wilkinson, 2013).

Numerical example: privatized count data

- prior: $\theta \sim \mathit{Gamma}(\alpha, \beta)$
- noiseless query likelihood: $\boldsymbol{s} \mid \theta \sim \mathit{Pois}(\theta)$
- ϵ-Laplace privacy mechanism: $\boldsymbol{s}_{\mathrm{dp}} \mid \boldsymbol{s} \sim \epsilon^{-1}\mathit{Lap}(1)$
- Private posterior:

$$\pi\left(\theta \mid \boldsymbol{s}_{\mathrm{dp}}\right) \propto \theta^{\alpha-1} e^{-(\beta+1)\theta} \left[\frac{\Gamma\left(\left\lceil \boldsymbol{s}_{\mathrm{dp}} \right\rceil, \theta_{\epsilon}^{+}\right)}{\Gamma\left(\left\lceil \boldsymbol{s}_{\mathrm{dp}} \right\rceil\right)} e^{\theta_{\epsilon}^{+} - \epsilon \boldsymbol{s}_{\mathrm{dp}}} + \frac{\gamma\left(\left\lceil \boldsymbol{s}_{\mathrm{dp}} \right\rceil, \theta_{\epsilon}^{-}\right)}{\Gamma\left(\left\lceil \boldsymbol{s}_{\mathrm{dp}} \right\rceil\right)} e^{\theta_{\epsilon}^{-} + \epsilon \boldsymbol{s}_{\mathrm{dp}}} \right]$$

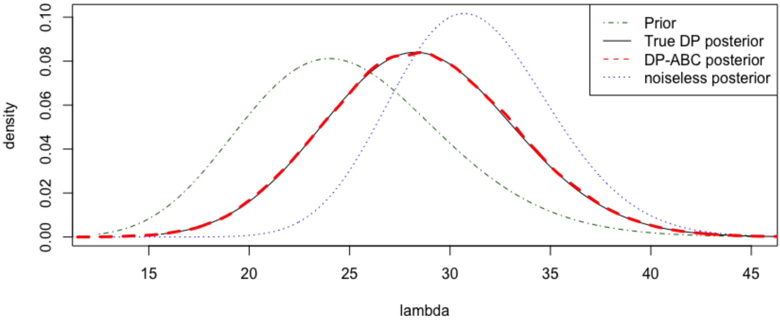


Figure: Algorithm 3 produces draws (red dashed density, estimated w/ $N = 10^6$) exactly from the true posterior (black solid density), and is different from the incorrect posterior (blue dotted density) which treats $\boldsymbol{s}_{\text{dp}} = 37.4$ as if confidential. Green dot-dash density is the prior. $\alpha = 25, \beta = 1, \epsilon = 0.2$.

Exact likelihood inference with Monte Carlo EM

Expectation-Maximization (Dempster et al., 1977) in the context of differential privacy:

- complete data is $(\boldsymbol{s}, \boldsymbol{s}_{\text{dp}})$;
- missing data is $\boldsymbol{s}$, where $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$; ← data analyst
- observed data is $\boldsymbol{s}_{\text{dp}}$, where $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s} \sim \eta_{\text{dp}}(\cdot \mid \boldsymbol{s})$. ← data curator

Exact likelihood inference with Monte Carlo EM

Expectation-Maximization (Dempster et al., 1977) in the context of differential privacy:

- complete data is $(\boldsymbol{s}, \boldsymbol{s}_{\text{dp}})$;
- missing data is $\boldsymbol{s}$, where $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$; $\leftarrow$ data analyst
- observed data is $\boldsymbol{s}_{\text{dp}}$, where $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s} \sim \eta_{\text{dp}}(\cdot \mid \boldsymbol{s})$. $\leftarrow$ data curator

Iterate till convergence:

- E-step:

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= \mathbb{E}\left(\log L(\theta; \boldsymbol{s}, \boldsymbol{s}_{\text{dp}}) \mid \boldsymbol{s}_{\text{dp}}, \theta^{(t)}\right) \\ &= \mathbb{E}\left(\log \pi(\boldsymbol{s} \mid \theta) \mid \boldsymbol{s}_{\text{dp}}, \theta^{(t)}\right) + \text{const.} \end{aligned}$$

- M-step:

$$\theta^{(t+1)} = \text{argmax}_{\theta} Q(\theta; \theta^{(t)}).$$

Exact likelihood inference with Monte Carlo EM

E-Step via Importance Sampling

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s} \mid \theta^{(t)})$; ← data analyst

step 2, assign weight $\omega_i = \eta_{\text{dp}}\left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}_i\right)$; ← data curator

Output: a set of weighted samples $\{\boldsymbol{s}_i, \omega_i\}_{i=1}^{N}$.

Exact likelihood inference with Monte Carlo EM

E-Step via Importance Sampling

Iterate: for $i = 1, \ldots, N$:

step 1, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s} \mid \theta^{(t)})$; ← data analyst

step 2, assign weight $\omega_i = \eta_{\text{dp}}\left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}_i\right)$; ← data curator

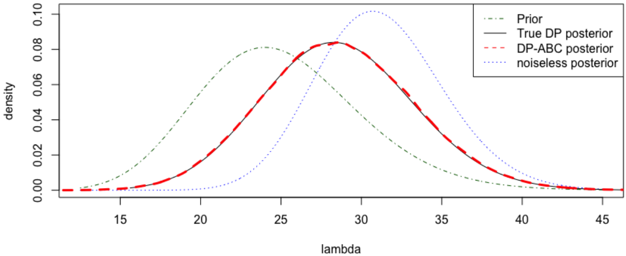
Output: a set of weighted samples $\{\boldsymbol{s}_i, \omega_i\}_{i=1}^N$.

$$\sum_{i=1}^{N} \omega_i b\left(\boldsymbol{s}_i\right) / \sum_{i=1}^{N} \omega_i \xrightarrow{p} \mathbb{E}\left(b(\boldsymbol{s}) \mid \boldsymbol{s}_{\text{dp}}, \theta^{(t)}\right), \qquad \text{as } N \to \infty.$$

Take $b(\boldsymbol{s})$ to be...

- sufficient statistic for θ, if $\pi(\boldsymbol{s} \mid \theta)$ is exponential family;
- $\log \pi(\boldsymbol{s} \mid \theta)$ in general;
- $\nabla_\theta \log \pi(\boldsymbol{s} \mid \theta)$ and $\nabla_\theta^2 \log \pi(\boldsymbol{s} \mid \theta)$, towards estimating observed score function and Fisher information.

Numerical example revisited: privatized count data



$$\boldsymbol{s} \mid \theta \sim Pois(\theta), \quad \boldsymbol{s}_{\mathrm{dp}} \mid \boldsymbol{s} \sim \epsilon^{-1} Lap(1), \quad \epsilon = 0.2, \quad \boldsymbol{s}_{\mathrm{dp}} = 37.4.$$

Monte Carlo EM gives

- $\hat{\theta}_{\mathrm{dp}} = 37.237$, $\hat{I}_{\mathrm{dp}} = 1.582 \times 10^{-2}$;
- Compared to incorrectly treating $\boldsymbol{s}_{\mathrm{dp}}$ as if confidential: $\hat{\theta} = 37.4$, $\hat{I} = 2.674 \times 10^{-2} \approx 169\% \times \hat{I}_{\mathrm{dp}}$.

Contribution & takeaway

- Theoretically exact statistical inference for general likelihood and Bayesian models with DP data;
- Applicable to all proper Bayesian priors;
- Fully amenable to computing acceleration for specific applications.

The analogy at play here:

$$\text{approximate computation on exact data} \\ \Updownarrow \\ \text{exact computation on approximate data}$$

such that the statistical tradeoff (efficiency vs privacy) becomes aligned with the computational tradeoff (approximation vs exactness).

Caveats & further research

1. Framework is overly general
   - ▶ Computing acceleration is possible, but requires domain knowledge;
   - ? How to afford accessible and (approximately) correct analysis tools to many DP data users?
   - ! Bias correction for popular models and code implementation
2. How to account for invariants imposed on the DP mechanism

Bias correction: a quick remedy

Naïvely fitting the original model to privatized data

$$\boldsymbol{y}_{\text{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\text{dp}} + \boldsymbol{e}$$

results in biased least squares estimates

$$\hat{\beta}_0^{\text{dp}} \approx \beta_0 \underbrace{+\,\alpha_{x,z}\beta_1}, \qquad \hat{\beta}_1^{\text{dp}} \approx \beta_1 \underbrace{-\,\gamma_{x,z}\beta_1},$$

where

$$\alpha_{x,z} = \gamma_{x,z}\bar{x} + (1-\gamma_{x,z})\bar{z}, \qquad \gamma_{x,z} = \frac{\text{SS}_{xz} + \text{SS}_z}{\text{SS}_{x+z}} \in (0,1)\,.$$

- Both $\alpha_{x,z}$ and $\gamma_{x,z}$ can be estimated using the privatized data and knowledge of the DP mechanism.
- General bias correction strategies for measurement error models:
  - regression calibration
  - simulation extrapolation

Imposing invariants on DP mechanisms

Invariants are exact statistics computed from the confidential micro-data (Ashmead et al., 2019), with which the DP releases should be congruent.

Two ways to impose a set of invariants $\mathcal{C}$ onto a given DP mechanism S:

1. Co-processing:
$$S_{\mathcal{C}}(x) \overset{d}{=} S(x) \mid S(x) \in \mathcal{C}.$$
2. Post-processing:
$$\tilde{S}_{\mathcal{C}}(x) = \text{argmin}_{a \in \mathcal{C}} \Delta(S(x), a),$$
for Δ some discrepancy measure (L_2, L_1, etc.)

Co-processing guarantee: a result

Let S be an ϵ-DP mechanism based on the confidential query $s : \mathcal{X} \to \mathbb{R}^p$. $\mathcal{C} \in \mathscr{B}(\mathbb{R}^p)$ is a set of invariants, and $S_\mathcal{C}$ a modified privatization mechanism such that

$$S_\mathcal{C} \overset{d}{=} S \mid S \in \mathcal{C}.$$

Then for all k-neighboring and $\mathcal{C}$-conforming pairs of datasets $\{(x, x') : d(x, x') = k, s(x) \in \mathcal{C}, s(x') \in \mathcal{C}\}$, and all $A \in \mathscr{B}(\mathbb{R}^p)$,

$$P(S_\mathcal{C}(x) \in A) \leq \exp(2k\epsilon) P(S_\mathcal{C}(x') \in A).$$

Caution: Due to the constraint $\mathcal{C}$ imposes on $\mathcal{X}$, neighboring dataset pairs with $k = 1$ (original DP definition) may no longer be feasible.

Proof:

$$\begin{aligned}\frac{P(S_\mathcal{C}(x) \in A)}{P(S_\mathcal{C}(x') \in A)} &= \underbrace{\frac{P(S(x) \in A \cap \mathcal{C})}{P(S(x') \in A \cap \mathcal{C})}}_{\leq \exp(k\epsilon)} \cdot \underbrace{\frac{P(S(x') \in \mathcal{C})}{P(S(x) \in \mathcal{C})}}_{\leq \exp(k\epsilon)} \\ &\leq \exp(2k\epsilon).\end{aligned}$$

Co-processing guarantee: a result

Let S be an ϵ-DP mechanism based on the confidential query $s : \mathcal{X} \to \mathbb{R}^p$. $\mathcal{C} \in \mathscr{B}(\mathbb{R}^p)$ is a set of invariants, and $S_\mathcal{C}$ a modified privatization mechanism such that

$$S_\mathcal{C} \overset{d}{=} S \mid S \in \mathcal{C}.$$

Then for all k-neighboring and $\mathcal{C}$-conforming pairs of datasets $\{(x, x') : d(x, x') = k, s(x) \in \mathcal{C}, s(x') \in \mathcal{C}\}$, and all $A \in \mathscr{B}(\mathbb{R}^p)$,

$$P(S_\mathcal{C}(x) \in A) \leq \exp(2k\epsilon) P(S_\mathcal{C}(x') \in A).$$

Caution: Due to the constraint $\mathcal{C}$ imposes on $\mathcal{X}$, neighboring dataset pairs with $k = 1$ (original DP definition) may no longer be feasible.

Proof:

$$\frac{P(S_\mathcal{C}(x) \in A)}{P(S_\mathcal{C}(x') \in A)} = \underbrace{\frac{P(S(x) \in A \cap \mathcal{C})}{P(S(x') \in A \cap \mathcal{C})}}_{\leq \exp(k\epsilon)} \cdot \underbrace{\frac{P(S(x') \in \mathcal{C})}{P(S(x) \in \mathcal{C})}}_{= \exp(\alpha k \epsilon)}$$

$$\leq \exp\{(1+\alpha)k\epsilon\}, \quad \alpha \in [0, 1].$$

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

1. Co-processing:

$$S_{\mathcal{C}}(\mathbf{x}) \stackrel{d}{=} (x_1 + u_1, x_2 + u_2) \mid u_1 + u_2 = 0.$$

The density of $S_{\mathcal{C}}$ is

$$p(S_{\mathcal{C}}(\mathbf{x}) = (s, n - s)) = \epsilon \exp\{-2\epsilon |s - x_1|\}.$$

That is, simulate $s_1 \sim x_1 + Lap\left((2\epsilon)^{-1}\right)$ and set $s_2 = n - s_1$, or equivalently, simulate $s_2 \sim x_2 + Lap\left((2\epsilon)^{-1}\right)$ and set $s_1 = n - s_2$.

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

2. Post-processing (L_2):

$$\tilde{S}_{\mathcal{C}}^{L_2}(\mathbf{x}) = \text{argmin}_{a \in \mathcal{C}} \|S(x) - a\|_2 \overset{d}{=} (\bar{x} + \tilde{u}, \bar{x} - \tilde{u}),$$

where $\tilde{u}$ is a 50%-50% mixture of:

- a Laplace distribution with scale $(2\epsilon)^{-1}$, and
- a signed Gamma distribution (i.e. a regular Gamma distribution times a fair random sign) with shape $k = 2$ and scale $(2\epsilon)^{-1}$.

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

3. Post-processing (L_1):

$$\tilde{S}_{\mathcal{C}}^{L_1}(\mathbf{x}) = \text{argmin}_{a \in \mathcal{C}} \|S(x) - a\|_1 = (\tilde{s}, n - \tilde{s})$$

is not a unique mechanism, only having to satisfy

$$\tilde{s} \in [x_1 + \min(u_1, u_2), x_1 + \max(u_1, u_2)].$$

In particular, $\tilde{s} = x_1 + u_1$ is always a solution, i.e. simply add $Lap(\epsilon^{-1})$ noise to the first entry, and subtract the same amount from the second.

Co-processing vs. Post-processing

Suppose the confidential dataset has just two count entries: $\mathbf{x} = (x_1, x_2)$.
The DP mechanism $S(\mathbf{x}) = (s_1, s_2) = (x_1 + u_1, x_2 + u_2)$, $u_i \sim Lap(\epsilon^{-1})$.
The invariant information is the total count $x_1 + x_2 = n$.

3. Post-processing (L_1):

$$\tilde{S}_{\mathcal{C}}^{L_1}(\mathbf{x}) = \text{argmin}_{a \in \mathcal{C}} \|S(x) - a\|_1 = (\tilde{s}, n - \tilde{s})$$

is not a unique mechanism, only having to satisfy

$$\tilde{s} \in [x_1 + \min(u_1, u_2), x_1 + \max(u_1, u_2)].$$

In particular, $\tilde{s} = x_1 + u_1$ is always a solution, i.e. simply add $Lap(\epsilon^{-1})$ noise to the first entry, and subtract the same amount from the second.

Question: privacy guarantees for post-processing?

Bibliography

Gong, R. (2019). Exact Inference with Approximate Computation for Differentially Private Data via Perturbations. *arXiv:1909.12237*

Ashmead, R., Kifer, D., Leclerc, P., Machanavajjhala, A., & Sexton, W. (2019). Effective privacy after adjusting for invariants with applications to the 2020 census.

Avella-Medina, M. (2018). Privacy-preserving parametric inference: a case for robust statistics.

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(2), 235–269.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

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

Dwork, C., & Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, *1*(2), 135–154.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 226–233.

Nissim, K., Raskhodnikova, S., & Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual acm symposium on theory of computing* (pp. 75–84).

Wei, G. C., & Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, *85*(411), 699–704.

Wilkinson, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, *12*(2), 129–141.