Bayesian Methods to Improve The Accuracy of Differentially Private Measurements of Constrained Parameters

Ryan Janicki[1], **Scott H. Holan**[2], Kyle M. Irimata[1],

James Livsey[1], and Andrew Raim[1]

U.S. Census Bureau[1]

Department of Statistics
University of Missouri
and
U.S. Census Bureau[2]

October 22, 2024

Acknowledgement

This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau. The results from this presentation can be found at `https://arxiv.org/abs/2406.04448`

Outline

Outline

Motivation: Statistical Post Processing (SPP) and S-DHC

- The privacy of the truncated enumerated counts are protected using a form of differential privacy.
- Applying disclosure avoidance methodology results in a ("noisy") privacy protected truncated count.
- Privacy protected truncated counts can be statistically post processed (SPP) – outputs of the SPP inherit the associated privacy guarantees.
- **Motivation:** SPP reduces implausible results, improves accuracy, and provides measures of disclosure avoidance-related uncertainty.

Motivation/Overview: The SPP Approach

- SPP model starts with the ("noisy") privacy protected measurements.
- SPP uses a Bayesian approach along with the noisy measurements and logical constraints (such as non-negativity and lower bounds) from the enumerated data.
- SPP then generates a new set of estimates.
- The new estimates are typically more precise (relative to the truncated, enumerated value) than the noisy measurements.
- SPP corrects some of the implausible data and generates a measure of accuracy known as a "credible interval" (CI).

Motivation: Similarity to Area-level SAE Models

Area-level models:

- Treat the direct estimate as the response variable
- Usually incorporate smoothing through the model

Similar to Bradley et al. (2015):

- **Data Model:**

$$\begin{aligned} Z_i &= Y_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, D_i) \end{aligned}$$

- **Process Model:**

$$\begin{aligned} Y_i &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{s}_i^T \boldsymbol{\eta} \\ \boldsymbol{\eta} &\sim N_r\left(0, \sigma^2 \boldsymbol{K}\right) \end{aligned}$$

- **Goal:** Prediction of Y_i

Outline

Differential Privacy

The following formal definition of differential privacy (DP) can be found in Dwork and Roth (2014).

Definition: A randomized algorithm $\mathcal{M}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all databases $\boldsymbol{X}, \boldsymbol{X}'$ such that $\|\boldsymbol{X} - \boldsymbol{X}'\| \leq 1$,

$$P\left(\mathcal{M}(\boldsymbol{X}) \in \mathcal{S}\right) \leq \exp\{\varepsilon\} P\left(\mathcal{M}(\boldsymbol{X}') \in \mathcal{S}\right) + \delta,$$

for $\varepsilon > 0$ and $\delta \geq 0$.

- The privacy budget is given by the parameters ε and δ and determines the amount of privacy guarantee.
- Small values of ε and δ provide greater privacy protection but less accuracy, while larger values of ε and δ result in more accuracy in exchange for weaker privacy protection.

Differential Privacy Continued

- Roughly speaking, if an algorithm provides privacy protections,then the outputs should be similar when applied to similar databases, so that any one individual record is not overly influential and can not easily be recovered.
- Let Y be a tabulation of $\boldsymbol{X}$ that a statistical agency would like to publish.
- For example, Y could represent the number of households by relationship for the population under 18 years in a county in the U.S.
- The tabulation Y cannot be released to the public without first applying DA techniques.
- A simple privacy protection algorithm which achieves DP is adding statistical noise to Y and releasing this noisy version of Y.

DP Continued

The noisy measurement is denoted by Z, and can be generated as

$$Z = Y + \varepsilon, \tag{1}$$

where ε is sampled from a noise-generating (probability) distribution.

- Importantly, (1) is analogous to the area-level SAE model.
- Two of the most used distributions for ε are the Gaussian distribution and the Laplace distribution.
- The Laplace mechanism applied in this way preserves $(\varepsilon, 0)$-DP while the Gaussian mechanism preserves (ε, δ)-DP.
- The addition of statistical noise from a Laplace or Gaussian distribution guarantees that DP will be satisfied.

DP Continued

- Adding statistical noise reduces the utility of the data
- Specifically, the noisy measurements, Z, are less precise than the tabulations Y.
- The noisy measurements may also violate certain constraints that the unperturbed tabulations are known to satisfy.
    1. If Y is a count tabulation, then Y must be nonnegative.
    2. If Y is the ratio of two count tabulations, there may be a relationship between the numerator and the denominator that must be taken into account.

Modeling Setup

- Let $\boldsymbol{Y} \in \mathbb{R}^m$ be a vector of tabulations of the database $\boldsymbol{X}$, and let $\boldsymbol{Z}$ be a privacy-protected measurement of $\boldsymbol{Y}$ obtained by independently adding noise to each component of $\boldsymbol{Y}$.

- Let f denote the noise generating distribution. Then

$$\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta} \sim \prod_{i=1}^{m} f\left(Z_i; Y_i, \boldsymbol{\theta}\right) \tag{2}$$

  where $\boldsymbol{\theta}$ is a vector of parameters determined by the DA algorithm which will be fully known to the analyst.

- Let p denote the number of known inequality constraints that must be satisfied by the components of $\boldsymbol{Y}$.

- Note that these constraints are acting on $\boldsymbol{Y}$.

Modeling Setup Continued

- We can summarize this information in terms of a vector of lower bounds, $\boldsymbol{l} \in \mathbb{R}^p$, a vector of upper bounds, $\boldsymbol{u} \in \mathbb{R}^p$, and a constraint matrix $\boldsymbol{D} \in \mathbb{R}^{p \times m}$,

$$\boldsymbol{l} \leq \boldsymbol{D}\boldsymbol{Y} \leq \boldsymbol{u}, \tag{3}$$

  where the inequalities are to be interpreted componentwise.

- A straightforward way to incorporate the constraints is to use a prior distribution on $\boldsymbol{Y}$ with support implied by the inequalities in (3).

- For our work we used an improper distribution

$$\pi\left(\boldsymbol{Y}\right) \propto I\left(\boldsymbol{l} \leq \boldsymbol{D}\boldsymbol{Y} \leq \boldsymbol{u}\right), \tag{4}$$

  where $I(\cdot)$ is the indicator function.

Modeling Setup Continued

- Combining (2) and (4) results in a posterior distribution

$$\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta} \propto \prod_{i=1}^{m} f\left(Z_i; Y_i, \boldsymbol{\theta}\right) I\left(\boldsymbol{l} \leq \boldsymbol{D}\boldsymbol{Y} \leq \boldsymbol{u}\right). \tag{5}$$

- Since the prior is improper when either the upper or lower bound is infinite, it does need to be verified that the expression in (5) is integrable.

- Fortunately, in most practical applications, the noise will be additive so that f is location invariant and (5) will be proper.

- For the special case when f is Gaussian, the posterior distribution is a multivariate truncated Gaussian distribution.

Outline

Empirical Example: S–DHC

- The S–DHC tables contain information about characteristics of persons, households, and person-household joins (tables which combine person data and household data).
- There are 8 nation/state-level S-DHC tables published for 2020:
  1. Average Household Size by Age (PH1),
  2. Household Type for the Population in Households (PH2),
  3. Households by Relationship For the Population Under 18 years (PH3),
  4. Population in Families by Age (PH4),
  5. Average Family Size by Age (PH5),
  6. Family Type and Age For Own Children Under 18 years (PH6),
  7. Total Population in Occupied Housing Units by Tenure (PH7),
  8. Average Household Size of Occupied Housing Units by Tenure (PH8).

S–DHC Example Continued

- We give an example using the 2010 version of the PH5 table, Average Family Size by Age, Race and Ethnicity in states in the U.S.
- The race and ethnicity iterations are White alone; Black or African American alone; Asian alone; American Indian and Alaska Native alone; Native Hawaiian and Other Pacific Islander alone; Some Other Race alone; Two or More Races; Hispanic or Latino; White alone, not Hispanic or Latino; and unattributed.
- The estimates that are produced in this table are:
  1. the ratio of number of persons 18 and under in families to the number of family households,
  2. the number of persons over 18 in families to the number of family households,
  3. the total number of persons in families to the number of family households.

S–DHC Example Continued

Table 1 shows the published 2010 state-level ratios for total population for five states.

		Alabama	Alaska	Arizona	Arkansas	California
Total:		3.02	3.21	3.19	3.00	3.45
	Under 18 years:	0.87	1.07	1.01	0.90	1.05
	18 years and over:	2.14	2.14	2.18	2.10	2.40

Table: Average family size by age: 2010 published decennial census state-level total population tabulations. Available at `data.census.gov`.

S–DHC Example Continued

- Three noisy measurements are generated for each geography and each race iteration for the PH5 table:
  1. the total population under 18 in families,
  2. the total population over 18 in families,
  3. and the number of family households.

- Y_{18-}, Y_{18+}, and Y_{FHH}, and the noisy measurements as Z_{18-}, Z_{18+}, and Z_{FHH}.

- The published values in PH5 are

$$\frac{Z_{18-}}{Z_{\text{FHH}}}, \quad \frac{Z_{18+}}{Z_{\text{FHH}}}, \quad \text{and} \quad \frac{Z_{18-} + Z_{18+}}{Z_{\text{FHH}}}, \tag{6}$$

  which are estimates of the ratios

$$\frac{Y_{18-}}{Y_{\text{FHH}}}, \quad \frac{Y_{18+}}{Y_{\text{FHH}}}, \quad \text{and} \quad \frac{Y_{18-} + Y_{18+}}{Y_{\text{FHH}}}. \tag{7}$$

S–DHC Example Continued

- The constraints that must be satisfied are
  1. $Y_{18-} \geq 0$,
  2. $Y_{18+} \geq 0$,
  3. $Y_{\text{FHH}} \geq 1$ (we only consider areas with at least one occupied household),
  4. $Y_{18+} + Y_{18-} \geq 2Y_{\text{FHH}}$ (the universe is family households).

- We have an additional constraint that is specific to our application that $Y_{18+} + Y_{18-} \leq \kappa Y_{\text{FHH}}$, where κ is a positive integer.

- This constraint is due to the privacy algorithm used by the U.S. Census Bureau for person-household join tables which truncates the family household universe to households with at most κ individuals.

- Let $\boldsymbol{Y}^{\intercal} = (Y_{18-}, Y_{18+}, Y_{\text{FHH}})$ and $\boldsymbol{Z}^{\intercal} = (Z_{18-}, Z_{18+}, Z_{\text{FHH}})$.

S–DHC Example Continued

- For this problem, the constraints in (3) are

$$\boldsymbol{l} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & -2 \\ -1 & -1 & \kappa \end{bmatrix}, \ \boldsymbol{u} = \begin{bmatrix} \infty \\ \infty \\ \infty \\ \infty \\ \infty \end{bmatrix}. \tag{8}$$

- The preliminary DA algorithm uses a privacy-loss budget which results in 90% margins of error of 200 and a truncation level of 10.

- This 90% margin of error is equivalent to a variance parameter of $\sigma^2 = 14{,}782$ when using a Gaussian noise distribution, or a scale parameter of $\lambda = 86.86$ when using a Laplace noise distribution.

S–DHC Example Continued

- We performed two experiments based on these parameter settings.
- We first generated a set of noisy measurements by adding independent Gaussian noise with variance $\sigma^2 = 14{,}782$ to the true 2010 census counts described above.
- We then drew 10,000 samples from the posterior distribution (5) using the correctly-specified Gaussian likelihood and constraints as in (8).
- We then repeated this experiment, but instead added independent Laplace noise with the scale parameter set to 86.86.

S–DHC Example Continued

Recall that the true value of each ratio must be between 0 and 10.

Mechanism	Estimate	MIN	MAX	BAD%	RMSE	COV	LEN
Gauss	NM	-5.2	17.2	1.4	0.7	87.5	1.2
	MB	0.5	6.2	0.0	0.2	89.3	0.3
Laplace	NM	-6.6	12.5	1.4	0.6	NA	NA
	MB	0.5	6.2	0.0	0.2	86.7	0.3

Table: The metrics shown are the maximum value (MAX), minimum value (MIN), the percent which are outside the constrained region (BAD%), root mean squared error (RMSE), coverage rate (COV) and interval length (LEN).

Outline

Summary

- We proposed a modeling approach for S–DHC (SPP).
- Demonstrated that SPP results in estimates which are more precise than the noisy measurements and belong to the constrained parameter space.
- The approach was illustrated with both Gaussian and Laplace noise.
- There are many opportunities to extend this approach for S–DHC and other data products:
  1. different geographies
  2. auxiliary information (covariates, admin. records, etc.)
  3. incorporating dependence
  4. machine learning methods
  5. "adaptive design"

Thank you!

holans@missouri.edu
scott.holan@census.gov

References

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Ann. Appl. Stat.*, 9(4):1761–1791.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.