International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 24(2) Special, December, 2024, pp 95-122
DOI: https://doi.org/10.3329/ijss.v24i20.78217

Multiple Imputation for Parametric Inference Under a Differentially Private Laplace Mechanism

Martin Klein[1*] and Bimal Sinha[2]

[1]Division of Biometrics VIII, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, USA
[2]Department of Mathematics and Statistics, University of Maryland, Baltimore County; and Center for Statistical Research and Methodology, U.S. Census Bureau, USA
Email: sinha@umbc.edu

[*]Correspondence should be addressed to Martin Klein
(Email: Martin.Klein@fda.hhs.gov)

[Received November 25, 2024; Accepted December 3, 2024]

Abstract

In this paper we consider the scenario where continuous microdata have been noise infused using a differentially private Laplace mechanism for the purpose of statistical disclosure control. We assume the original data are independent and identically distributed, having distribution within a parametric family of continuous distributions. We use a variant of the Laplace mechanism that allows the range of the original data to be unbounded by first truncating the original data and then adding appropriate Laplace random noise. We propose methodology to analyze the noise infused data using multiple imputation. This approach allows the data user to analyze the released data as if it were original, i.e., not noise infused, and then to obtain inference that accounts for the noise infusion mechanism using standard multiple imputation combining formulas. Methodology is presented for univariate data, and some simulation studies are presented to evaluate the performance of the proposed method. An extension of the proposed methodology to multivariate data is also presented.

Keywords: Differential Privacy, EM Algorithm, Multiple Imputation, Parametric Model, Statistical Disclosure Control.

AMS Classification: 62F99.

Disclaimer: This article reflects the views of the authors and should not be construed to represent the views and/or policies of neither the U.S. Food and Drug Administration or the U.S. Census Bureau.

1. Introduction

Research on privacy preserving statistical databases is essential for addressing the following situation. A data producer constructs a dataset $\boldsymbol{X}$ (for example, by conducting a survey) which contains useful information about the relevant population, and it is desirable for this information to be released. However, privacy and confidentiality concerns prevent the dataset $\boldsymbol{X}$ from being released. (In some settings a distinction is made between *privacy* and *confidentiality* [Nayak, Zhang, and Adeshiyan, 2015], but here we will consider these terms to be synonymous.) To resolve these competing objectives, a transformation (often randomized) is constructed from $\boldsymbol{X}$ to $\boldsymbol{Z}$, and the dataset $\boldsymbol{Z}$ is released, instead of $\boldsymbol{X}$. The goal is to construct this transformation such that by releasing $\boldsymbol{Z}$, the two objectives are satisfied: (1) individual's privacy is protected; and (2) the released dataset is useful for drawing inference on the relevant population. Statistical disclosure control methodology refers to the methodology used to make the transformation from $\boldsymbol{X}$ to $\boldsymbol{Z}$.

Differential privacy (Dwork et al., 2006, 2017) is a mathematical definition for quantifying the privacy protection provided by the transformation from $\boldsymbol{X}$ to $\boldsymbol{Z}$. The definition of differential privacy is designed to control the effect of any one individual's information on the released data. One may refer, for example, to Wasserman and Zhou (2010), Dwork and Roth (2014), Vadhan (2016), and Dwork et al. (2017) for discussion on the interpretation of differential privacy, as well as for some standard transformations (such as the Laplace mechanism and exponential mechanism) that satisfy differential privacy. Differential privacy possesses desirable properties such as closure under composition, closure under postprocessing, and group privacy (Dwork et al., 2017), and the approach has gained considerable popularity in recent years. Differential privacy has also been applied in practice, see for example, Machanavajjhala et al. (2008), Erlingsson, Pihur, and Korolova (2014), and Differential Privacy Team, Apple (2017).

In this paper we propose methodology that uses a differentially private noise infusion mechanism to protect the data, and multiple imputation to facilitate valid data analysis. An advantage of using multiple imputation is that data users can analyze the released, multiply imputed data as if it were original data, and then apply simple multiple imputation combination formulas to obtain valid inference that accounts for the extra variability due to the noise infusion. The methodology presented in this paper can be summarized as follows. One first applies random noise infusion to transform from $\boldsymbol{X}$ to $\boldsymbol{Z}$ in such a way that differential privacy is attained. Then one sets up a missing data problem, where the noise infused data $\boldsymbol{Z}$ are viewed as the observed data, and the original data $\boldsymbol{X}$ are viewed as the missing data. Based on the "observed" data $\boldsymbol{Z}$, the "missing" data $\boldsymbol{X}$ are multiply imputed to obtain $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, $m > 1$. Then the multiply imputed data $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$ are released. Because the transformation from $\boldsymbol{X}$ to $\boldsymbol{Z}$ is differentially private, it follows from the result on *closure under postprocessing* (Dwork et al. (2017), we also present a similar property in our setting in Result 1) that the overall transformation from $\boldsymbol{X}$ to $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$ is also differentially private under the parametric scenario that we will describe in Section 2.. For drawing inference on the underlying population, the data user

can then analyze each dataset $\boldsymbol{x}^{*(j)}$ as if it were the original data, and then apply multiple imputation combining formulas to obtain valid inference based on the entire released data $\{\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}\}$. Thus this approach enables the data user to obtain valid inference using standard methods and software, in conjunction with simple combination formulas.

In this paper we use Laplace random noise to make a transformation from $\boldsymbol{X}$ to $\boldsymbol{Z}$ that satisfies differential privacy, however, we employ a modification to the standard Laplace mechanism that incorporates truncation. The standard Laplace mechanism (Proposition 3.3 of Dwork et al., 2017) assumes that the L_1 Sensitivity (Definition 3.1 of Dwork et al., 2017) of the chosen query is finite. The query refers to the function of $\boldsymbol{X}$ that will be noise infused via Laplace additive noise. In this paper we assume the query is $\boldsymbol{X}$ itself, and we assume that the original data are independent and identically distributed (iid), having continuous distribution within a parametric family. Under many common parametric models the sample space (or support) is not bounded (e.g. normal, lognormal, exponential), and hence the L_1 sensitivity of $\boldsymbol{X}$ is not finite. We use a modified version of the Laplace mechanism that enables differential privacy to be attained even if the sample space of the original data is not bounded (Result 3) by first truncating the original data to a finite interval and then adding appropriate random noise. A similar variant of the Laplace mechanism that incorporates truncation is also discussed by Duchi et al. (2018).

Synthetic data methodology (Raghunathan, Reiter, and Rubin, 2003; Reiter, 2003; Reiter, 2005; Drechsler, 2011) is a well established form of statistical disclosure control methodology that also uses concepts of multiple imputation for missing data (Rubin, 1987) to enable valid inference to be drawn on the underlying population using the released data. The methodology proposed in this paper differs from established synthetic data methods, because the proposed method takes the explicit step of infusing noise using a differentially private mechanism before applying multiple imputation. The output of our methodology is a set of $m > 1$ multiply imputed datasets, which, to a user, would appear the same as the output of synthetic data methodology. As discussed by Rubin (1993), an advantage of synthetic data is that data users can analyze the released data using standard statistical procedures (in conjunction with multiple imputation combination formulas that are simple to apply). The proposed methodology shares this advantage, because the methods that the data user can apply for drawing inference are nearly the same (we propose to use the combination formulas of Rubin (1987) and Li, Raghunathan, and Rubin (1991) for missing data, instead of the formulas of Raghunathan, Reiter, and Rubin (2003), Reiter (2003), or Reiter (2005) for synthetic data, due to the nature of the missing data and imputation process) as those used for synthetic data. However, the extra step of infusing noise before applying multiple imputation allows the noise level to be controlled, and differential privacy to be attained at a desired privacy-loss budget. As with synthetic data, the most complicated part of the proposed methodology is carrying out the imputation, and this would be performed by the data producer, and not the data user. Furthermore, algorithms for generating these imputed values are developed in this paper.

In some settings differential privacy can be attained using synthetic data methodology via a specialized choice of the prior distribution in the Bayesian model specification; however,

in such settings it has been shown that the usual combination formulas may not be applicable (Charest, 2010). The methodology proposed in this paper uses ideas from our earlier work (Klein and Sinha, 2013) where we considered noise multiplication for privacy protection, and then applied multiple imputation for data analysis. There we suggested that the initial step of applying noise multiplication may be advantageous because it allowed explicit control over the level of noise infused into the released data. In this paper, we take direct advantage of this control over the level of noise, in fact, by infusing noise in such a way that differential privacy is satisfied, and then developing appropriate imputation procedures under this type of noise infusion.

The outline of the paper is as follows. In Section 2. we present the general setup, and some basic definitions and results concerning differential privacy that are used in the paper. Following Wasserman and Zhou (2010), we present these definitions and results using a statistical framework. In Section 3. we introduce the standard Laplace mechanism, and the modification to the Laplace mechanism using truncation, as applied in our setting. In Section 4. we derive methodology for applying multiple imputation to impute the original data, based on data that have been noise infused using the Laplace mechanism with truncation; and we review the combination formulas used for drawing inference. Specifically, in Section 4.1. we derive some basic distributional results, including a sampling algorithm that will be needed; in Section 4.2. we discuss how the distributional results can be applied for direct likelihood based analysis based on $\boldsymbol{Z}$; in Section 4.3. we apply the distributional results to develop procedures for carrying out multiple imputation; and in Section 4.4. we discuss how to use the multiply imputed data to draw inference. Section 5. presents some numerical results; specifically Section 5.1. presents an application of the proposed methodology when the original data are normally distributed; and Section 5.2. presents some simulation studies to assess the finite sample properties of the proposed methodology in this case. In Section 6. we extend the proposed methodology to multivariate data. Section 7. contains some concluding remarks. Proofs of the results and other technical details appear in the technical report Klein and Sinha (2019).

2. General Setup

Suppose the original data are $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\boldsymbol{\theta})$, where $f(x|\boldsymbol{\theta})$ is the probability density function (pdf) of a continuous distribution and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is an unknown parameter such that $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. Suppose each $X_i \in \mathcal{X}$ where $\mathcal{X} \subseteq \mathbb{R}$ denotes the support of the pdf $f(x|\boldsymbol{\theta})$ such that $f(x|\boldsymbol{\theta}) > 0$ if $x \in \mathcal{X}$, and $f(x|\boldsymbol{\theta}) = 0$ if $x \notin \mathcal{X}$. We assume that the original data $X_1, \ldots, X_n$ are sensitive and cannot be released. Instead, the original data are modified, using a randomized mechanism, to create a sanitized dataset, denoted by $Z_1, \ldots, Z_{\tilde{n}}$, that may then be released. Here $\tilde{n}$ and n can be unequal and each $Z_i \in \mathbb{R}$. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_{\tilde{n}})$. The randomized mechanism used to create the sanitized dataset induces a conditional distribution for $\boldsymbol{Z}$, given $\boldsymbol{X} = \boldsymbol{x}$.

Let $\mathcal{B}(\mathbb{R})$ be the class of Borel sets in $\mathbb{R}$, let $\mathcal{B}(\mathbb{R}^{\tilde{n}})$ be the class of Borel sets in $\mathbb{R}^{\tilde{n}}$, and let $\mathcal{X}^n = \mathcal{X} \times \cdots \times \mathcal{X} \subseteq \mathbb{R}^n$ be the n-fold Cartesian product. For two vectors

$\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ in $\mathcal{X}^n$, define $\delta(\boldsymbol{a}, \boldsymbol{b}) = |\{i : a_i \neq b_i\}|$, where for a set A having a finite number of elements, $|A|$ is the number of elements in A. We now state the definition of ϵ-differential privacy (Dwork et al., 2006, 2017), which is a condition imposed on the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$. The conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, is also referred to as the *data release mechanism* (Wasserman and Zhou, 2010). As discussed by Dwork et al. (2006, 2017), the definition of differential privacy requires the notion of neighboring datasets, where a dataset is an element of $\mathcal{X}^n$. For a given distance function, two datasets are called neighbors if and only if their distance equals 1. Following Wasserman and Zhou (2010), the definition stated below is based on the notion where $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}^n$ are *neighbors* if and only if $\delta(\boldsymbol{a}, \boldsymbol{b}) = 1$.

Definition 1. For a given $\epsilon \geq 0$, the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, is said to satisfy ϵ-*differential privacy* if

$$P(\boldsymbol{Z} \in A \,|\, \boldsymbol{X} = \boldsymbol{a}) \leq e^{\epsilon} P(\boldsymbol{Z} \in A \,|\, \boldsymbol{X} = \boldsymbol{b}) \tag{1}$$

for all $A \in \mathcal{B}(\mathbb{R}^{\tilde{n}})$ and all $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}^n$ such that $\delta(\boldsymbol{a}, \boldsymbol{b}) = 1$.

Remark 1. Observe that in ϵ-differential privacy, small values of ϵ provide more privacy, while large values provide less privacy. The quantity ϵ is referred to as the *privacy-loss budget*.

Remark 2. It is said that the randomized transformation (or randomized mechanism) used to transform from $\boldsymbol{X}$ to $\boldsymbol{Z}$ satisfies ϵ-differential privacy, if under this transformation, the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy.

The following lemma presents an equivalent characterization of ϵ-differential privacy based on expectation.

Lemma 1. Let $\epsilon \geq 0$. The following statements are equivalent.

(a) The conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy.

(b) $E[h(\boldsymbol{Z}) \,|\, \boldsymbol{X} = \boldsymbol{a}] \leq e^{\epsilon} E[h(\boldsymbol{Z}) \,|\, \boldsymbol{X} = \boldsymbol{b}]$ for all nonnegative functions h that are measurable from $(\mathbb{R}^{\tilde{n}}, \mathcal{B}(\mathbb{R}^{\tilde{n}}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and all $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}^n$ such that $\delta(\boldsymbol{a}, \boldsymbol{b}) = 1$.

In the sequel, we will make use of the following result; we note that the result on *closure under postprocessing* (Proposition 2.4) of Dwork et al. (2017), and Part 2 of Lemma 2.6 of Wasserman and Zhou (2010) are both similar to this result

Result 1. Suppose that the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy, and let $\boldsymbol{Y}$ be a random vector in $\mathbb{R}^s$ such that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are conditionally independent, given $\boldsymbol{Z}$. Then the conditional distribution of $\boldsymbol{Y}$, given $\boldsymbol{X}$, also satisfies ϵ-differential privacy.

3. Laplace Mechanism, Standard Version and with Truncation

From now on $\tilde{n}$, the sample size of the sanitized dataset $\boldsymbol{Z}$, will be taken as equal to n, the sample size of the original dataset $\boldsymbol{X}$. Let $\text{Lap}(\mu, \sigma)$ denote the Laplace distribution having pdf $h_{\text{Lap}}(w) = \frac{1}{2\sigma} \exp\{-|w - \mu|/\sigma\}$, $-\infty < w < \infty$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. Recall that if $W \sim \text{Lap}(\mu, \sigma)$, then $E(W) = \mu$ and $\text{Var}(W) = 2\sigma^2$. Also recall that if $r_1, r_2 \in \mathbb{R}$, then

$$\big||r_1| - |r_2|\big| \leq |r_1 - r_2|, \tag{2}$$

which is a consequence of the triangle inequality: $|r_1 + r_2| \leq |r_1| + |r_2|$ (Bartle and Sherbert, 2000, page 31). The following is a standard data release mechanism that satisfies ϵ-differential privacy, referred to as the Laplace mechanism. While this result is well known (see, for example, Proposition 3.3 of Dwork et al., 2017, or Theorem 3.6 of Dwork and Roth, 2014), below we state the Laplace mechanism as it applies to our scenario.

Result 2. (standard Laplace mechanism) Let $\epsilon > 0$ and let $\Delta = \sup\{|\alpha - \beta| : \alpha, \beta \in \mathcal{X}\}$. If $\Delta \in (0, \infty)$, and

$$Z_i = X_i + R_i,\ i = 1, \ldots, n,\ \text{where } R_1, \ldots, R_n \overset{\text{iid}}{\sim} \text{Lap}\,(0, \Delta/\epsilon)\,, \tag{3}$$

then conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy.

For many choices of the original data distribution $f(x|\boldsymbol{\theta})$, the sample space $\mathcal{X}$ is not a bounded set and hence $\sup\{|\alpha - \beta| : \alpha, \beta \in \mathcal{X}\} = \infty$. For example if $f(x|\boldsymbol{\theta})$ is the normal pdf, then $\mathcal{X} = \mathbb{R}$ and $\sup\{|\alpha - \beta| : \alpha, \beta \in \mathbb{R}\} = \infty$; if $f(x|\boldsymbol{\theta})$ is the lognormal pdf, then $\mathcal{X} = (0, \infty)$ and $\sup\{|\alpha - \beta| : \alpha, \beta \in (0, \infty)\} = \infty$. If $\sup\{|\alpha - \beta| : \alpha, \beta \in \mathcal{X}\} = \infty$, then the Laplace mechanism as stated in Result 2 cannot be applied. To obtain ϵ-differential privacy using Laplace additive noise without requiring that $\sup\{|\alpha - \beta| : \alpha, \beta \in \mathcal{X}\}$ is finite, we consider the following modified version of the Laplace mechanism that incorporates truncation of the original data. A similar variant of the Laplace mechanism is also discussed by Duchi et al. (2018).

Result 3. (Laplace mechanism with truncation) Let $\epsilon > 0$, and let $L, U \in \mathbb{R}$ be such that $L < U$. If

$$Z_i = \left\{ \begin{array}{ll} L + R_i, & \text{if } X_i < L, \\ X_i + R_i, & \text{if } L \leq X_i \leq U, \\ U + R_i, & \text{if } X_i > U, \end{array} \right\},\ i = 1, \ldots, n,\ \text{where } R_1, \ldots, R_n \overset{\text{iid}}{\sim} \text{Lap}\left(0, \frac{U - L}{\epsilon}\right), \tag{4}$$

then the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy.

4. Methodology for Multiple Imputation Based Inference

Throughout this section we work under the notation of Section 1 with $\tilde{n} = n$, and we let $Z_1, \ldots, Z_n$ be defined as in Result 3.

4.1. Joint, Marginal, and Conditional Distributions

The conditional pdf of Z_i, given $X_i = x_i$, is

$$g_{Z|X}(z_i|x_i) = \begin{cases} \frac{1}{2c}e^{-|z_i-L|/c}, & \text{if } x_i < L \\ \frac{1}{2c}e^{-|z_i-x_i|/c}, & \text{if } L \le x_i \le U \\ \frac{1}{2c}e^{-|z_i-U|/c}, & \text{if } x_i > U \end{cases}$$
$$= I_{(-\infty,L)}(x_i)\frac{e^{-|z_i-L|/c}}{2c} + I_{[L,U]}(x_i)\frac{e^{-|z_i-x_i|/c}}{2c} + I_{(U,\infty)}(x_i)\frac{e^{-|z_i-U|/c}}{2c},$$

where $c = (U-L)/\epsilon$, and I_A is the indicator function for the set A. The joint pdf of (X_i, Z_i) is

$$g_{X,Z}(x_i, z_i|\boldsymbol{\theta}) = g_{Z|X}(z_i|x_i)f(x_i|\boldsymbol{\theta})$$
$$= I_{(-\infty,L)}(x_i)\frac{e^{-|z_i-L|/c}}{2c}f(x_i|\boldsymbol{\theta}) + I_{[L,U]}(x_i)\frac{e^{-|z_i-x_i|/c}}{2c}f(x_i|\boldsymbol{\theta}) + I_{(U,\infty)}(x_i)\frac{e^{-|z_i-U|/c}}{2c}f(x_i|\boldsymbol{\theta}).$$

The marginal pdf of Z_i is

$$g_Z(z_i|\boldsymbol{\theta}) = \int_{\mathbb{R}} g_{X,Z}(w, z_i|\boldsymbol{\theta})dw$$
$$= \int_{\mathbb{R}} \left\{ I_{(-\infty,L)}(w)\frac{e^{-\frac{|z_i-L|}{c}}}{2c}f(w|\boldsymbol{\theta}) + I_{[L,U]}(w)\frac{e^{-\frac{|z_i-w|}{c}}}{2c}f(w|\boldsymbol{\theta}) + I_{(U,\infty)}(w)\frac{e^{-\frac{|z_i-U|}{c}}}{2c}f(w|\boldsymbol{\theta}) \right\} dw,$$

and therefore,

$$g_Z(z_i|\boldsymbol{\theta}) = \frac{e^{-|z_i-L|/c}}{2c}\int_{-\infty}^{L} f(w|\boldsymbol{\theta})dw + \frac{1}{2c}\int_L^U e^{-|z_i-w|/c}f(w|\boldsymbol{\theta})dw + \frac{e^{-|z_i-U|/c}}{2c}\int_U^{\infty} f(w|\boldsymbol{\theta})dw$$
$$= \frac{e^{-|z_i-L|/c}}{2c}F(L|\boldsymbol{\theta}) + \frac{1}{2c}\int_L^U e^{-|z_i-w|/c}f(w|\boldsymbol{\theta})dw + \frac{e^{-|z_i-U|/c}}{2c}[1-F(U|\boldsymbol{\theta})],$$

where $F(w|\boldsymbol{\theta}) = \int_{-\infty}^{w} f(t|\boldsymbol{\theta})dt$ is the cumulative distribution function (cdf) of a random variable whose pdf is $f(x|\boldsymbol{\theta})$. Hence the marginal pdf of Z_i is

$$g_Z(z_i|\boldsymbol{\theta}) = \frac{e^{-|z_i-L|/c}}{2c}F(L|\boldsymbol{\theta}) + \frac{1}{2c}\int_L^U e^{-|z_i-w|/c}f(w|\boldsymbol{\theta})dw + \frac{e^{-|z_i-U|/c}}{2c}[1-F(U|\boldsymbol{\theta})]. \quad (5)$$

The conditional pdf of X_i, given $Z_i = z_i$, is

$$g_{X|Z}(x_i|z_i, \boldsymbol{\theta}) = \frac{g_{X,Z}(x_i, z_i|\boldsymbol{\theta})}{g_Z(z_i|\boldsymbol{\theta})}$$
$$= \frac{I_{(-\infty,L)}(x_i)\frac{e^{-|z_i-L|/c}}{2c}f(x_i|\boldsymbol{\theta}) + I_{[L,U]}(x_i)\frac{e^{-|z_i-x_i|/c}}{2c}f(x_i|\boldsymbol{\theta}) + I_{(U,\infty)}(x_i)\frac{e^{-|z_i-U|/c}}{2c}f(x_i|\boldsymbol{\theta})}{\frac{e^{-|z_i-L|/c}}{2c}F(L|\boldsymbol{\theta}) + \frac{1}{2c}\int_L^U e^{-|z_i-w|/c}f(w|\boldsymbol{\theta})dw + \frac{e^{-|z_i-U|/c}}{2c}[1-F(U|\boldsymbol{\theta})]},$$

and hence the conditional pdf of X_i, given $Z_i = z_i$, can be expressed as

$$g_{X|Z}(x_i|z_i,\boldsymbol{\theta}) = \frac{f(x_i|\boldsymbol{\theta})\left\{I_{(-\infty,L)}(x_i)e^{-|z_i-L|/c} + I_{[L,U]}(x_i)e^{-|z_i-x_i|/c} + I_{(U,\infty)}(x_i)e^{-|z_i-U|/c}\right\}}{e^{-|z_i-L|/c}F(L|\boldsymbol{\theta}) + \int_L^U e^{-|z_i-w|/c}f(w|\boldsymbol{\theta})dw + e^{-|z_i-U|/c}[1-F(U|\boldsymbol{\theta})]}. \tag{6}$$

Result 4. For given values of $z \in \mathbb{R}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the following algorithm produces a random variable X^* having the density $g_{X|Z}(x^*|z,\boldsymbol{\theta})$, where $g_{X|Z}$ is the pdf defined in Equation (6). Define

$$K(x,z) = \frac{I_{(-\infty,L)}(x)e^{-|z-L|/c} + I_{[L,U]}(x)e^{-|z-x|/c} + I_{(U,\infty)}(x)e^{-|z-U|/c}}{I_{(-\infty,L)}(z)e^{-|z-L|/c} + I_{[L,U]}(z) + I_{(U,\infty)}(z)e^{-|z-U|/c}}$$

and proceed as follows.

1. Generate V and X independently such that $V \sim \text{Uniform}(0,1)$ and $X \sim f(x|\boldsymbol{\theta})$.
2. If $V \le K(X,z)$ then accept X and deliver $X^* = X$. Otherwise reject X and return to Step 1.

The expected number of iterations of Steps 1 and 2 required to obtain X^* is

$$M(z,\boldsymbol{\theta}) = C(z,\boldsymbol{\theta})\left\{I_{(-\infty,L)}(z)e^{-|z-L|/c} + I_{[L,U]}(z) + I_{(U,\infty)}(z)e^{-|z-U|/c}\right\},$$

where $C(z,\boldsymbol{\theta}) = \left\{e^{-|z-L|/c}F(L|\boldsymbol{\theta}) + \int_L^U e^{-|z-w|/c}f(w|\boldsymbol{\theta})dw + e^{-|z-U|/c}[1-F(U|\boldsymbol{\theta})]\right\}^{-1}$.

Remark 3. The pdf of the conditional distribution of X_i, given $Z_i = z_i$, can be expressed in the form of a mixture distribution; details appear in the technical report Klein and Sinha (2019).

4.2. A Look At Likelihood Based Data Analysis

By Equation (5) it follows that the likelihood function for $\boldsymbol{\theta}$ based the observed sanitized data $\boldsymbol{Z} = \boldsymbol{z}$ is

$$\begin{aligned}\mathcal{L}_{\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{z}) &= \prod_{i=1}^n g_Z(z_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^n \left\{\frac{e^{-|z_i-L|/c}}{2c}F(L|\boldsymbol{\theta}) + \frac{1}{2c}\int_L^U e^{-|z_i-w|/c}f(w|\boldsymbol{\theta})dw + \frac{e^{-|z_i-U|/c}}{2c}[1-F(U|\boldsymbol{\theta})]\right\}.\end{aligned} \tag{7}$$

The likelihood function for $\boldsymbol{\theta}$ based on the observed original data $\boldsymbol{X} = \boldsymbol{x}$ is

$$\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta}). \tag{8}$$

If the original data $\boldsymbol{X}$ were observed, then analysis for $\boldsymbol{\theta}$ can be conducted under the likelihood function (8). Generally, if $f(x|\boldsymbol{\theta})$ is a standard parametric model, then likelihood based analysis under (8) can be conducted in a straightforward manner using readily available software. However, since the sanitized data $\boldsymbol{Z}$ are observed instead of $\boldsymbol{X}$, the likelihood function is (7), which obviously differs from the original data likelihood function (8). One way to compute the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ under the likelihood function (7), is by means of the EM algorithm (Dempster, Laird, and Rubin, 1977; Little and Rubin, 2002). To apply the EM algorithm in this scenario, we setup a missing data problem with $\boldsymbol{X}$ as the missing data, $\boldsymbol{Z}$ as the observed data, and hence $(\boldsymbol{X}, \boldsymbol{Z})$ as the complete data. Then the complete data likelihood function is

$$\mathcal{L}_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z}) = \prod_{i=1}^{n} g_{X,Z}(x_i, z_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} g_{Z|X}(z_i|x_i) f(x_i|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}).$$

Therefore, with $\boldsymbol{\theta}^{(t)}$ denoting an estimate of $\boldsymbol{\theta}$ at iteration t, the E step of an EM algorithm for computing the MLE of $\boldsymbol{\theta}$ under (7), is to compute

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int_{\mathcal{X}^n} \ell_X(\boldsymbol{\theta}|\boldsymbol{x}) \prod_{i=1}^{n} g_{X|Z}(x_i|z_i, \boldsymbol{\theta}^{(t)}) d\boldsymbol{x},$$

where $\ell_X(\boldsymbol{\theta}|\boldsymbol{x}) = \log \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$. The M step of the EM algorithm is to compute $\boldsymbol{\theta}^{(t+1)}$ such that $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Depending on the choice of $f(x|\boldsymbol{\theta})$, the E and/or M steps may have a closed form.

4.3. Using the Sanitized Data to Multiply Impute the Original Data

Instead of a direct likelihood based approach, we now consider an alternative approach based on multiple imputation (Rubin, 1987). Under the multiple imputation based approach, the observed $\boldsymbol{Z} = \boldsymbol{z}$ will be used to create multiply imputed versions of the original data $\boldsymbol{X}$, that we denote by $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, where $m > 1$. An advantage of this approach is that each imputed data set $\boldsymbol{x}^{*(j)}$ can be analyzed as if it were the original data (analysis can be based on the simple likelihood function (8), instead of the more complicated likelihood (7)), and the overall inference is then obtained using standard multiple imputation combination formulas.

We setup a missing data problem (Little and Rubin, 2002) with $\boldsymbol{X}$ as the missing data, $\boldsymbol{Z}$ as the observed data, and hence $(\boldsymbol{X}, \boldsymbol{Z})$ as the complete data (as we did in the preceding section when discussing the EM algorithm). In order to apply the multiple imputation methodology of Rubin (1987), we will impute the missing data $\boldsymbol{X}$ under a Bayesian model, and to do so, we place a prior density $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$. The missing data $\boldsymbol{X}$ are multiply imputed by taking $m > 1$ random draws from the posterior predictive distribution of $\boldsymbol{X}$, given $\boldsymbol{Z} = \boldsymbol{z}$. A general procedure for drawing $(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$ from the joint posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$, is given in Algorithm 1.

Algorithm 1. Draw $(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$ from the joint posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$.

1. Draw $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_p^*)$ from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{Z} = \boldsymbol{z}$.
2. Draw $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$ from the conditional distribution of $\boldsymbol{X}$, given $\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{\theta} = \boldsymbol{\theta}^*$.

Return $(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$ as a draw from the posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$.

Observe that in Step 1 of Algorithm 1, we sample from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{Z} = \boldsymbol{z}$, and the pdf of this distribution is

$$\begin{aligned} &p_{\boldsymbol{\theta}|\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{z}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \mathcal{L}_{\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{z}) \\ &= p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \frac{e^{-|z_i - L|/c}}{2c} F(L|\boldsymbol{\theta}) + \frac{1}{2c} \int_L^U e^{-|z_i - w|/c} f(w|\boldsymbol{\theta}) dw + \frac{e^{-|z_i - U|/c}}{2c} [1 - F(U|\boldsymbol{\theta})] \right\}. \end{aligned} \tag{9}$$

In some cases, depending on the choice of the original data distribution $f(x|\boldsymbol{\theta})$, and prior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, a computationally efficient method for taking a random draw from the posterior density of $\boldsymbol{\theta}$, given $\boldsymbol{Z} = \boldsymbol{z}$, may not be readily available because the pdf in Equation (9) may have a complicated form (involving integrals that may not have closed form expressions). We can bypass direct sampling from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{Z} = \boldsymbol{z}$, by using the data augmentation algorithm (Tanner and Wong, 1987; Little and Rubin, 2002). The data augmentation algorithm is equivalent to a Gibbs Sampler, and it is an approach that uses Markov chain Monte Carlo (MCMC) to approximately sample from the posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$. A data augmentation algorithm for generating a Markov chain whose stationary distribution is the posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$, is presented in Algorithm 2.

Step 1 of Algorithm 2 requires sampling from the conditional distribution of $\boldsymbol{X}$, given $\boldsymbol{Z} = \boldsymbol{z}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. The conditional pdf of $\boldsymbol{X}$, given $\boldsymbol{Z} = \boldsymbol{z}$ and $\boldsymbol{\theta}$ is

$$g_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} g_{X|Z}(x_i|z_i, \boldsymbol{\theta}),$$

where $g_{X|Z}(x_i|z_i, \boldsymbol{\theta})$ is the pdf defined in Equation (6). Recall that Result 4 provides a general algorithm for drawing a random variable having the pdf (6). Hence Step 1 of Algorithm 2 can be implemented as follows.

1. Use Result 4 to draw $x_i^{(t+1)} \sim g_{X|Z}(x|z_i, \boldsymbol{\theta}^{(t)})$ for each $i = 1, \ldots, n$. Hence obtain $\boldsymbol{x}^{(t+1)} = (x_1^{(t+1)}, \ldots, x_n^{(t+1)})$.

Step 2 of Algorithm 2 requires sampling from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{X} = \boldsymbol{x}^{(t+1)}$ and $\boldsymbol{Z} = \boldsymbol{z}$. The posterior pdf of $\boldsymbol{\theta}$, given $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{Z} = \boldsymbol{z}$, is

$$p_{\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} g_{X,Z}(x_i, z_i|\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} \{g_{Z|X}(z_i|x_i) f(x_i|\boldsymbol{\theta})\} \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \tag{10}$$

Algorithm 2. General form of a data augmentation/Gibbs sampling algorithm for generating a Markov chain whose stationary distribution is the posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$. Algorithm 3 (below) provides specific details on how this procedure can be implemented in practice.

Set an initial value $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$.
Set a large natural number T.
Set $t = 0$.
While $t < T$, do the following $\{$

1. Draw $\boldsymbol{x}^{(t+1)} = (x_1^{(t+1)}, \ldots, x_n^{(t+1)})$ from the conditional distribution of $\boldsymbol{X}$, given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}, \boldsymbol{Z} = \boldsymbol{z}$.
2. Draw $\boldsymbol{\theta}^{(t+1)} = (\theta_1^{(t+1)}, \ldots, \theta_p^{(t+1)})$ from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{X} = \boldsymbol{x}^{(t+1)}, \boldsymbol{Z} = \boldsymbol{z}$.
3. Update $t = t + 1$.

$\}$
Return $(\boldsymbol{x}^{(1)}, \boldsymbol{\theta}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{\theta}^{(2)}), \ldots, (\boldsymbol{x}^{(T)}, \boldsymbol{\theta}^{(T)})$.

and hence, letting $p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$ denote the posterior density of $\boldsymbol{\theta}$, given $\boldsymbol{X} = \boldsymbol{x}$, we have

$$p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = p_{\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}).$$

That is, the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{X}$ and $\boldsymbol{Z}$, is the same as the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{X}$. Because $p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$ is the usual posterior density of $\boldsymbol{\theta}$, given the original data $\boldsymbol{X} = \boldsymbol{x}$, it will usually be known how to take a random draw from this distribution. Hence Step 2 of Algorithm 2 can be implemented as follows.

2. Draw $\boldsymbol{\theta}^{(t+1)} = (\theta_1^{(t+1)}, \ldots, \theta_p^{(t+1)}) \sim p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}^{(t+1)})$ where $p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}^{(t+1)})$ is the posterior density of $\boldsymbol{\theta}$, given $\boldsymbol{X} = \boldsymbol{x}^{(t+1)}$.

Thus the general procedure of Algorithm 2 can be implemented using the specific steps shown in Algorithm 3.

There are two standard ways in which Algorithm 3 can be applied to obtain m approximate draws from the posterior distribution of $(\boldsymbol{X}, \boldsymbol{\theta})$, given $\boldsymbol{Z} = \boldsymbol{z}$ (see, for example, Hu, Mitra, and Reiter (2013) for a general discussion on the use of MCMC for multiple imputation). The two ways of using Algorithm 3 are as follows.

(a) Run Algorithm 3 a total of m times, independently, to generate m independent Markov chains that have converged, and take the final value from each chain.

(b) Run Algorithm 3 a single time to generate a single Markov chain, and after the chain has converged, take m draws from the converged part of the chain. The m draws should be far enough apart from each other within the chain so that the draws are approximately independent.

Algorithm 3. Specific details on how the general data augmentation method of Algorithm 2 can be implemented in practice.

Set an initial value $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$.
Set a large natural number T.
Set $t = 0$.
While $t < T$, do the following $\{$
1. Use Result 4 to draw $x_i^{(t+1)} \sim g_{X|Z}(x|z_i, \boldsymbol{\theta}^{(t)})$ for each $i = 1, \ldots, n$. Hence obtain $\boldsymbol{x}^{(t+1)} = (x_1^{(t+1)}, \ldots, x_n^{(t+1)})$.
2. Draw $\boldsymbol{\theta}^{(t+1)} = (\theta_1^{(t+1)}, \ldots, \theta_p^{(t+1)}) \sim p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}^{(t+1)})$ where $p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}^{(t+1)})$ is the posterior density of $\boldsymbol{\theta}$, given $\boldsymbol{X} = \boldsymbol{x}^{(t+1)}$; that is, $p_{\boldsymbol{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}^{(t+1)}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{i=1}^{n} f(x_i^{(t+1)}|\boldsymbol{\theta})$.
3. Update $t = t + 1$.

$\}$
Return $(\boldsymbol{x}^{(1)}, \boldsymbol{\theta}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{\theta}^{(2)}), \ldots, (\boldsymbol{x}^{(T)}, \boldsymbol{\theta}^{(T)})$.

Remark 4. The multiply imputed values of $\boldsymbol{X}$, namely $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, can be generated using either Algorithm 1, 2, or 3. Each of these algorithms uses the observed value $\boldsymbol{Z} = \boldsymbol{z}$, but not the observed value $\boldsymbol{X} = \boldsymbol{x}$. Therefore, we may conclude that $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ is conditionally independent of $\boldsymbol{X}$, given $\boldsymbol{Z}$. Because the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ satisfies ϵ-differential privacy, Result 1 implies that the conditional distribution of $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$, given $\boldsymbol{X}$, also satisfies ϵ-differential privacy. Regarding the conditional independence of $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ and $\boldsymbol{X}$, given $\boldsymbol{Z}$, we have two additional comments.

(a) Algorithms 2 and 3 require an initial value $\boldsymbol{\theta}^{(0)}$. The initial value $\boldsymbol{\theta}^{(0)}$ can be a function of $\boldsymbol{Z}$, but it should not be a function of $\boldsymbol{X}$, so that the required conditional independence between $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ and $\boldsymbol{X}$, given $\boldsymbol{Z}$, holds.

(b) We are working under the paradigm where the original data are known to be iid from a distribution within the parametric family $\mathcal{F} = \{f(x|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, and it is only the parameter $\boldsymbol{\theta}$ that is unknown and must be estimated from the observed data. The Algorithms 1, 2, or 3 do not use $\boldsymbol{X}$, but instead use $\boldsymbol{Z}$ along with the knowledge that the original data are iid from a distribution within the parametric family $\mathcal{F}$, to generate multiply imputed values of $\boldsymbol{X}$.

4.4. Data Analysis Using Multiple Imputation Combination Formulas

Suppose that a data analyst observes the multiply imputed data $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, and wants to draw inference on the unknown parameter $\boldsymbol{\theta}$, or on a function of $\boldsymbol{\theta}$. Below we review the inferential procedures of Rubin (1987) for a scalar-valued estimand, and the procedures of Rubin (1987) and Li, Raghunathan, and Rubin (1991) for a vector-valued estimand. In addition to Rubin (1987) and Li, Raghunathan, and Rubin (1991), one may also refer to

Rubin (1996), Schafer (1997), Little and Rubin (2002) and Reiter and Raghunathan (2007) for more information on these procedures, and for additional procedures.

Scalar-Valued Estimand. Suppose that $Q = Q(\boldsymbol{\theta})$ is the scalar-valued parameter of interest. Let $\hat{Q}(\boldsymbol{X})$ be an estimator of $Q(\boldsymbol{\theta})$ based on the original data $\boldsymbol{X}$, and let $\hat{V}(\boldsymbol{X})$ be an estimator of the variance of $\hat{Q}(\boldsymbol{X})$, also based on the original data $\boldsymbol{X}$. Let $Q_j^* = \hat{Q}(\boldsymbol{x}^{*(j)})$ and $V_j^* = \hat{V}(\boldsymbol{x}^{*(j)})$ be the values of $\hat{Q}$ and $\hat{V}$ when computed on the jth imputed dataset $\boldsymbol{x}^{*(j)}$ for $j = 1, \ldots, m$. Define the following

$$\bar{Q}_m = \frac{1}{m}\sum_{j=1}^{m} Q_j^*, \quad \bar{V}_m = \frac{1}{m}\sum_{j=1}^{m} V_j^*, \quad B_m = \frac{1}{m-1}\sum_{j=1}^{m}(Q_j^* - \bar{Q}_m)^2.$$

Then $\bar{Q}_m$ is an estimator of Q, and the variance of $\bar{Q}_m$ is estimated by $T_m = \bar{V}_m + \left(1 + \frac{1}{m}\right) B_m$. The distribution of $(\bar{Q}_m - Q)/\sqrt{T_m}$ is approximated by a t distribution with $w = (m-1)(1 + r_m^{-1})^2$ degrees of freedom where $r_m = (1 + m^{-1})B_m \bar{V}_m^{-1}$. Hence, to obtain a test of significance for Q, or a confidence interval, one can use $(\bar{Q}_m - Q)/\sqrt{T_m}$ along with its approximate t distribution.

Vector-Valued Estimand. Suppose that $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$ is the $k \times 1$ dimensional vector-valued parameter of interest. Let $\hat{\boldsymbol{Q}}(\boldsymbol{X})$ be a $k \times 1$ dimensional estimator of $\boldsymbol{Q}(\boldsymbol{\theta})$ based on the original data $\boldsymbol{X}$, and let $\hat{\boldsymbol{V}}(\boldsymbol{X})$ be a $k \times k$ dimensional estimator of the covariance matrix of $\hat{\boldsymbol{Q}}(\boldsymbol{X})$, also based on the original data $\boldsymbol{X}$. Let $\boldsymbol{Q}_j^* = \hat{\boldsymbol{Q}}(\boldsymbol{x}^{*(j)})$ and $\boldsymbol{V}_j^* = \hat{\boldsymbol{V}}(\boldsymbol{x}^{*(j)})$ be the values of $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{V}}$ when computed on the jth imputed dataset $\boldsymbol{x}^{*(j)}$ for $j = 1, \ldots, m$. Define the following

$$\bar{\boldsymbol{Q}}_m = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{Q}_j^*, \quad \bar{\boldsymbol{V}}_m = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{V}_j^*, \quad \boldsymbol{B}_m = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{Q}_j^* - \bar{\boldsymbol{Q}}_m)(\boldsymbol{Q}_j^* - \bar{\boldsymbol{Q}}_m)'.$$

Then $\bar{\boldsymbol{Q}}_m$ is an estimator of $\boldsymbol{Q}$, and the covariance matrix of $\bar{\boldsymbol{Q}}_m$ is estimated by $\boldsymbol{T}_m = \bar{\boldsymbol{V}}_m + \left(1 + \frac{1}{m}\right) \boldsymbol{B}_m$. Define the quantity

$$S_m = \frac{(\bar{\boldsymbol{Q}}_m - \boldsymbol{Q})' \bar{\boldsymbol{V}}_m^{-1} (\bar{\boldsymbol{Q}}_m - \boldsymbol{Q})}{k(1 + r_m)}$$

where $r_m = (1 + m^{-1})\mathrm{tr}(\boldsymbol{B}_m \bar{\boldsymbol{V}}_m^{-1})/k$. The distribution of S_m is approximated by an $F_{k,w}$ distribution with

$$w = \begin{cases} 4 + [k(m-1) - 4]\left[1 + r_m^{-1}\left(1 - \frac{2}{k(m-1)}\right)\right]^2, & \text{if } k(m-1) > 4, \\ (m-1)\left(\frac{k+1}{2}\right)(1 + r_m^{-1})^2, & \text{if } k(m-1) \leq 4. \end{cases}$$

Hence, to obtain a test of significance for $\boldsymbol{Q}$, or a confidence ellipsoid, one can use S_m along with its approximate F distribution.

5. Simulation Studies Under the Normal Model

In this section we present simulation results to evaluate finite sample properties of the proposed method in the case that $f(x|\boldsymbol{\theta})$ is the normal model. In Section 5.1. we present details to illustrate how the proposed methodology can be applied under the normal model, and in Section 5.2. we summarize some results of the simulation study.

5.1. Application to the Normal Model

Suppose $f(x|\boldsymbol{\theta})$ is the normal pdf with mean μ and variance σ^2, that is,

$$f(x|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma^2 \in (0,\infty),$$

and hence $\boldsymbol{\theta} = (\mu, \sigma^2)$, $\mathcal{X} = \mathbb{R}$, $\boldsymbol{\Theta} = \mathbb{R} \times (0,\infty)$. We specify the following conjugate prior distribution for (μ, σ^2) (see, for example, Gelman et al., 2014),

$$\mu \,|\, \sigma^2 \sim N\left(\lambda_0, \frac{\sigma^2}{\kappa_0}\right), \quad \sigma^2 \sim \frac{\tau_0}{\chi^2_{\nu_0}}, \tag{11}$$

where $\lambda_0 \in \mathbb{R}$, $\kappa_0 \in (0,\infty)$, $\nu_0 \in (0,\infty)$, $\tau_0 \in (0,\infty)$ are parameters of the prior distribution. Under this specification, the prior density on (μ, σ^2) is

$$p_{\boldsymbol{\theta}}(\mu,\sigma^2) = \left\{\left(\frac{2\pi\sigma^2}{\kappa_0}\right)^{-1/2} \exp\left[-\frac{\kappa_0}{2\sigma^2}(\mu-\lambda_0)^2\right]\right\} \left\{\frac{\left(\frac{\tau_0}{2}\right)^{\nu_0/2}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma^2)^{-(\nu_0/2)-1} \exp\left[-\frac{\tau_0}{2\sigma^2}\right]\right\},$$

for $\mu \in \mathbb{R}$ and $\sigma^2 \in (0,\infty)$. Then the posterior distribution of (μ, σ^2), given $\boldsymbol{X} = \boldsymbol{x}$, can be represented as follows:

$$\mu \,|\, \sigma^2, \boldsymbol{X} = \boldsymbol{x} \sim N\left(\lambda_n, \frac{\sigma^2}{\kappa_n}\right), \quad \sigma^2 \,|\, \boldsymbol{X} = \boldsymbol{x} \sim \frac{\tau_n}{\chi^2_{\nu_n}}, \tag{12}$$

where $\lambda_n = \frac{\kappa_0\lambda_0 + n\bar{x}}{\kappa_0+n}$, $\kappa_n = \kappa_0 + n$, $\tau_n = \tau_0 + (n-1)s_x^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{x}-\lambda_0)^2$, $\nu_n = \nu_0 + n$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$. Hence the posterior density of (μ, σ^2), given $\boldsymbol{X} = \boldsymbol{x}$, is

$$p_{\boldsymbol{\theta}|\boldsymbol{X}}(\mu,\sigma^2|\boldsymbol{x}) = \left\{\left(\frac{2\pi\sigma^2}{\kappa_n}\right)^{-1/2} \exp\left[-\frac{\kappa_n}{2\sigma^2}(\mu-\lambda_n)^2\right]\right\} \left\{\frac{\left(\frac{\tau_n}{2}\right)^{\nu_n/2}}{\Gamma\left(\frac{\nu_n}{2}\right)} (\sigma^2)^{-(\nu_n/2)-1} \exp\left[-\frac{\tau_n}{2\sigma^2}\right]\right\},$$

for $\mu \in \mathbb{R}$ and $\sigma^2 \in (0,\infty)$. Drawing a random sample from the posterior distribution of (μ, σ^2), given $\boldsymbol{X} = \boldsymbol{x}$, is readily accomplished using the representation in (12), and hence Algorithm 3 is readily applied to obtain $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$ as discussed in Section 4.3.. For

the starting value $\boldsymbol{\theta}^{(0)}$ needed to run Algorithm 3, one use the MLE of $\boldsymbol{\theta}$ based on $\boldsymbol{Z}$. As discussed in Section 4.2., the MLE of $\boldsymbol{\theta}$ based on $\boldsymbol{Z}$ can be computed using the EM algorithm. In this case, the log-likelihood function for $\boldsymbol{\theta}$ based on $\boldsymbol{X} = \boldsymbol{x}$ is

$$\begin{aligned}\ell_{\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2}\end{aligned}$$

and the E step of the EM algorithm is to compute

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \int_{\mathbb{R}^n} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) \prod_{i=1}^{n} g_{X|Z}(x_i|z_i,\boldsymbol{\theta}^{(t)})d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \left\{ -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2} \right\} \prod_{i=1}^{n} g_{X|Z}(x_i|z_i,\boldsymbol{\theta}^{(t)})d\boldsymbol{x} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\psi_2(z_i,\boldsymbol{\theta}^{(t)}) + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}\psi_1(z_i,\boldsymbol{\theta}^{(t)}) - \frac{n\mu^2}{2\sigma^2},\end{aligned}$$

where

$$\begin{aligned}\psi_1(z_i,\boldsymbol{\theta}^{(t)}) &= \int_{-\infty}^{\infty} w g_{X|Z}(w|z_i,\boldsymbol{\theta}^{(t)})dw \\ &= \frac{e^{-|z_i-L|/c}\int_{-\infty}^{L} w f(w|\boldsymbol{\theta}^{(t)})dw + \int_{L}^{U} w f(w|\boldsymbol{\theta}^{(t)}) e^{-|z_i-w|/c}dw + e^{-|z_i-U|/c}\int_{U}^{\infty} w f(w|\boldsymbol{\theta}^{(t)})dw}{e^{-|z_i-L|/c}F(L|\boldsymbol{\theta}^{(t)}) + \int_{L}^{U} e^{-|z_i-w|/c} f(w|\boldsymbol{\theta}^{(t)})dw + e^{-|z_i-U|/c}[1-F(U|\boldsymbol{\theta}^{(t)})]},\end{aligned}$$

and

$$\begin{aligned}\psi_2(z_i,\boldsymbol{\theta}^{(t)}) &= \int_{-\infty}^{\infty} w^2 g_{X|Z}(w|z_i,\boldsymbol{\theta}^{(t)})dw \\ &= \frac{e^{-|z_i-L|/c}\int_{-\infty}^{L} w^2 f(w|\boldsymbol{\theta}^{(t)})dw + \int_{L}^{U} w^2 f(w|\boldsymbol{\theta}^{(t)}) e^{-|z_i-w|/c}dw + e^{-|z_i-U|/c}\int_{U}^{\infty} w^2 f(w|\boldsymbol{\theta}^{(t)})dw}{e^{-|z_i-L|/c}F(L|\boldsymbol{\theta}^{(t)}) + \int_{L}^{U} e^{-|z_i-w|/c} f(w|\boldsymbol{\theta}^{(t)})dw + e^{-|z_i-U|/c}[1-F(U|\boldsymbol{\theta}^{(t)})]}.\end{aligned}$$

By maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$, we obtain the following equations which define the sequence of EM iterations:

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\psi_1(z_i,\boldsymbol{\theta}^{(t)}), \qquad (\sigma^2)^{(t+1)} = \frac{1}{n}\left[\{\sum_{i=1}^{n}\psi_2(z_i,\boldsymbol{\theta}^{(t)})\} - n(\mu^{(t+1)})^2\right].$$

Here the integrals that appear in $\psi_1(z_i,\boldsymbol{\theta}^{(t)})$ and $\psi_2(z_i,\boldsymbol{\theta}^{(t)})$ can be expressed in closed form, up to the standard normal cdf; the relevant formulas appear in the technical report Klein and Sinha (2019).

Remark 5. Observe that in this case we are using a proper prior distribution for $\boldsymbol{\theta}$. If an improper prior distribution is used, then one must be careful, because even if the posterior distribution of $\boldsymbol{\theta}$, given the original data $\boldsymbol{X}$, is proper, the posterior distribution of $\boldsymbol{\theta}$, given the noise infused data $\boldsymbol{Z}$, may be improper. Consider the normal example with $n > 1$, and improper prior $p_{\boldsymbol{\theta}}(\mu, \sigma^2) \propto 1/\sigma^2$, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$. Then the posterior distribution of (μ, σ^2), given the original data $\boldsymbol{X}$, is proper (Klein and Sinha, 2013). However, the posterior distribution (μ, σ^2), given the noise infused data $\boldsymbol{Z}$, is not proper. Proof that the posterior distribution of (μ, σ^2), given $\boldsymbol{Z}$, is not proper is provided in the technical report Klein and Sinha (2019).

5.2. Simulation Results

We now present numerical evaluations to study the finite sample properties of the proposed methodology under the normal scenario of Section 5.1.. The simulation results are displayed in Tables 1, 2, and 3, and we set $\mu = 0$, $\sigma^2 = 1$ in all cases, and the parameters in the prior distribution are set to $\lambda_0 = 1$, $\kappa_0 = 0.1$, $\tau_0 = 10$, $\nu_0 = 5$. Each of these tables shows results when the parameter of interest is the mean μ, and also when the parameter of interest is the variance σ^2; and each table also shows results for each $\epsilon \in \{1, 2, 3, 4, 5\}$, and each $[L, U] \in \{[-3, 3], [-4, 4]\}$. Tables 1, 2, and 3 show results for (the number of imputations) $m = 10$, 30, and 50, respectively. For each setting the tables display Monte Carlo estimates (multiplied by 100), based on 2500 iterations, of root mean squared error of the estimator (RMSE), bias of the estimator (Bias), and standard deviation of the estimator (SD), as well as the average of the standard deviation estimator over the simulation runs ($\widehat{\text{SD}}$), empirical coverage probability of the nominal 0.95 confidence interval (Cvg), and average length of this confidence interval over the simulation runs (EL).

We use the methodology of Section 4.4. to obtain an estimator for the parameter of interest, along with an associated standard deviation estimator, and confidence interval. Let $\hat{\mu}_x = \bar{x}$ and $\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n} s_x^2$, i.e., $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ are the MLEs of μ and σ^2 based on the original data. In applying the methodology of Section 4.4., when the parameter of interest is $Q(\boldsymbol{\theta}) = \mu$, we take $\hat{Q}(\boldsymbol{x}) = \bar{x}$, $\hat{V}(\boldsymbol{x}) = \hat{\sigma}_x^2/n$; and when the parameter of interest is $Q(\boldsymbol{\theta}) = \sigma^2$, we take $\hat{Q}(\boldsymbol{x}) = \hat{\sigma}_x^2$, $\hat{V}(\boldsymbol{x}) = 2(\hat{\sigma}_x^2)^2/n$. Furthermore, we use Algorithm 3 to generate the multiply imputed data $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$. For the initial value $\boldsymbol{\theta}^{(0)}$ needed to run Algorithm 3, we use the MLE of $\boldsymbol{\theta}$ based on $\boldsymbol{Z}$, which we compute via the EM algorithm. We used the statistical computing software R (R Core Team, 2018) to obtain the simulation results. The EM algorithm and Algorithm 3 are the most computationally intensive parts of the simulation. Using Rcpp (Eddelbuettel and Francois, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2017) we implemented the EM algorithm, Algorithm 3, and the Laplace mechanism with truncation (4) in C++, and called the C++ implementations from R. Within each iteration of the simulation we ran Algorithm 3 a single time with a burn-in of 1000 iterations, and subsequently we sampled every 100th iteration until we obtained the desired number of m imputed values. For the EM algorithm we stopped when either a maximum number of 1000 iterations was reached,

or the convergence criterion $\{[\mu^{(t)} - \mu^{(t+1)}]^2 + [(\sigma^2)^{(t)} - (\sigma^2)^{(t+1)}]^2\}^{1/2} \leq 10^{-4}$ was satisfied. The following is a summary of the simulation results of Tables 1, 2, and 3.

1. *When the parameter of interest is* μ: We observe that Cvg tends to be close to the nominal value of 95%, Bias is close to 0, and $\widehat{\text{SD}}$ is close to SD. In this sense, we are able to obtain inference for μ that is approximately valid in all the scenarios considered, even those with smaller ϵ. Because $\widehat{\text{SD}}$ is close to SD, the results indicate that the multiple imputation standard deviation estimator $\sqrt{T_m}$ is an approximately unbiased estimator for the standard deviation of $\bar{Q}_m$. As expected, for given values of n, L, U, and m, we observe in the tables that RMSE, SD, $\widehat{\text{SD}}$, and EL tend to decrease as ϵ increases; and as ϵ gets larger the amount of decrease in these quantities reduces. We also observe that for given values of n, ϵ, L, and U, the quantities RMSE, SD, $\widehat{\text{SD}}$, and EL tend to decrease as m increases; though there are some exceptions, and the amount of decrease reduces as m increases.

2. *When the parameter of interest is* σ^2: We observe that for the smaller values of ϵ considered, Cvg tends to exceed the nominal value of 95%, $\widehat{\text{SD}}$ is greater than SD, and the Bias is also not approximately equal to 0. In Table 1 ($m = 10$), when $n = 5000$, $\epsilon = 1$ we also see that Cvg is 90.44% and 87.12% in the $[L, U] = [-3, 3]$ and $[L, U] = [-4, 4]$ cases, respectively; in these same scenarios when $m = 30$ or $m = 50$ the value of Cvg is more than 95%. In the Tables, we observe that Cvg tends to get closer to 95%, and Bias gets closer to 0, as ϵ increases. When Cvg exceeds 95%, we observe that $\widehat{\text{SD}}$ is greater than SD, indicating that the multiple imputation standard deviation estimator is positively biased for the standard deviation of $\bar{Q}_m$. We also observe that when the sample size increases from $n = 1000$ to $n = 5000$, Bias and Cvg appear to get closer to 0 and 95%, respectively, and also $\widehat{\text{SD}}$ is closer to SD.

3. *On the choice of* L *and* U: Looking at the Laplace mechanism with truncation in (4), recall that in addition to ϵ, this data release mechanism also has the tuning parameters L and U. Result 3 shows that the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, under this mechanism satisfies ϵ-differential privacy, irrespective of the choice of L and U. We observe that if $[L, U]$ is too narrow then most of the Z_i values will equal $L + R_i$ or $U + R_i$, which seems to cause a large amount of information loss in the data compared to a wider choice of $[L, U]$ which would give more Z_i values equal to $X_i + R_i$. However, the variance of the noise variable R_i is $2[(U - L)/\epsilon]^2$, and thus if the interval $[L, U]$ is too wide, then the variance of R_i will be large, which seems to cause a large amount of information loss in the data compared to a narrower choice of $[L, U]$ which would give a smaller variance for R_i. Thus it appears to be undesirable to choose $[L, U]$ to be too narrow, and it also appears undesirable to choose $[L, U]$ to be too wide. In the simulation studies we have considered the cases $[L, U] = [-3, 3]$ and $[L, U] = [-4, 4]$, and the original data are distributed as $N(0, 1)$ so that $P(X_i \in [-3, 3]) \approx 0.99730$ and $P(X_i \in [-4, 4]) \approx 0.99994$. Clearly under both of these choices, most of the original values are in the interval $[L, U]$. We observe in the Tables that inference tends to be less

accurate in the $[L, U] = [-4, 4]$ cases compared to the corresponding $[L, U] = [-3, 3]$. Thus in the simulation scenarios, the interval $[-3, 3]$ contains most of the original data, and while $[-4, 4]$ covers even more, the increase in variance of R_i seems to offset any benefit.

4. *Comparison with original data inference*: In this setting, with $X_1, \ldots, X_n$ iid as $N(0, 1)$, $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, $\hat{\sigma}_X^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, we note that $100 \times [\mathrm{Var}(\bar{X})]^{1/2} = 100 \times 1/\sqrt{n} \approx 3.16$ if $n = 1000$, ≈ 1.41 if $n = 5000$; and $100 \times [\mathrm{Var}(\hat{\sigma}_X^2)]^{1/2} = 100 \times \sqrt{2(n-1)/n^2} \approx 4.47$ if $n = 1000$, ≈ 2.00 if $n = 5000$. One can compare these values with the values of SD given in the tables to assess the amount of increase in standard deviation due to the combination of the Laplace mechanism with truncation (4), and multiple imputation based methodology.

6. Extension to Multivariate Data

In this section we extend the proposed methodology to the case of multivariate data. Suppose the original data consist of q-dimensional random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} f(\boldsymbol{x}|\boldsymbol{\theta})$. As in Section 2., assume $f(\boldsymbol{x}|\boldsymbol{\theta})$ is the pdf of a continuous distribution and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is an unknown parameter. In this setting, each $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq}) \in \boldsymbol{\mathcal{X}}$, where $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^q$ is the support of $f(\boldsymbol{x}|\boldsymbol{\theta})$ such that $f(\boldsymbol{x}|\boldsymbol{\theta}) > 0$ if $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and $f(\boldsymbol{x}|\boldsymbol{\theta}) = 0$ if $\boldsymbol{x} \notin \boldsymbol{\mathcal{X}}$. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\tilde{n}}$ denote the sanitized dataset, where each $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iq}) \in \mathbb{R}^q$. Let $\mathcal{B}(\mathbb{R}^{q\tilde{n}})$ be the class of Borel sets in $\mathbb{R}^{q\tilde{n}}$, and $\boldsymbol{\mathcal{X}}^n = \boldsymbol{\mathcal{X}} \times \cdots \times \boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^{qn}$ be the n-fold Cartesian product. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{\tilde{n}})$. For two vectors $\boldsymbol{a} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) = ((a_{11}, \ldots, a_{1q}), \ldots, (a_{n1}, \ldots, a_{nq})) \in \boldsymbol{\mathcal{X}}^n$ and $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n) = ((b_{11}, \ldots, b_{1q}), \ldots, (b_{n1}, \ldots, b_{nq})) \in \boldsymbol{\mathcal{X}}^n$, define $\delta(\boldsymbol{a}, \boldsymbol{b}) = |\{i : \boldsymbol{a}_i \neq \boldsymbol{b}_i\}|$. Below we state the definition of ϵ-differential privacy (Dwork et al., 2006, 2017), as it applies in the present multivariate scenario. As before, we follow Wasserman and Zhou (2010), treating $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{\mathcal{X}}^n$ as *neighbors* if and only if $\delta(\boldsymbol{a}, \boldsymbol{b}) = 1$.

Definition 2. For a given $\epsilon \geq 0$, the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, is said to satisfy ϵ-*differential privacy* if

$$P(\boldsymbol{Z} \in A \,|\, \boldsymbol{X} = \boldsymbol{a}) \leq e^{\epsilon} P(\boldsymbol{Z} \in A \,|\, \boldsymbol{X} = \boldsymbol{b}) \tag{13}$$

for all $A \in \mathcal{B}(\mathbb{R}^{q\tilde{n}})$ and all $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{\mathcal{X}}^n$ such that $\delta(\boldsymbol{a}, \boldsymbol{b}) = 1$.

Lemma 1 and Result 1 continue to hold in the multivariate scenario, with the obvious adjustments. Below we state a version of the Laplace mechanism (Dwork and Roth, 2014; Dwork et al., 2017) for the present multivariate scenario. We state the Laplace mechanism below in such a way that the random noise variables are independent, but not identically distributed.

Result 5. (standard Laplace mechanism for multivariate data) For each $j = 1, \ldots, q$, let $\epsilon_j > 0$ and $\Delta_j = \sup\{|\alpha_j - \beta_j| : \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q) \in \boldsymbol{\mathcal{X}}, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_q) \in \boldsymbol{\mathcal{X}}\}$. If $\Delta_j \in (0, \infty)$ for all $j = 1, \ldots, q$, and

$$Z_{ij} = X_{ij} + R_{ij}, \tag{14}$$

where $R_{ij} \sim \text{Lap}\,(0, \Delta_j/\epsilon_j)$, independently, for $j = 1, \ldots, q$ and $i = 1, \ldots, n$, then conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy with $\epsilon = \sum_{j=1}^{q} \epsilon_j$.

Remark 6. In Result 5, the Laplace random variables R_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, q$ are independent, but not identically distributed; instead, the noise variable R_{ij} is scaled to X_{ij} according to $R_{ij} \sim \text{Lap}\,(0, \Delta_j/\epsilon_j)$, leading to $(\sum_{j=1}^{q} \epsilon_j)$-differential privacy. In Result 5 one can instead draw $R_{ij} \overset{\text{iid}}{\sim} \text{Lap}\left(0, \epsilon^{-1} \sum_{j=1}^{q} \Delta_j\right)$, $j = 1, \ldots, q$, $i = 1, \ldots, n$, in which case the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, will satisfy ϵ-differential privacy. In this version with $R_{ij} \overset{iid}{\sim} \text{Lap}\left(0, \epsilon^{-1} \sum_{j=1}^{q} \Delta_j\right)$, one would only specify a single value $\epsilon > 0$, instead of the q positive values $\epsilon_1, \ldots, \epsilon_q$.

As in the univariate scenario, for many choices of the original data distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$, the sample space $\boldsymbol{\mathcal{X}}$ is such that Δ_j is not finite for all $j = 1, \ldots, q$, and hence the Laplace Mechanism in Result 5 cannot be applied. For example if $f(\boldsymbol{x}|\boldsymbol{\theta})$ is the multivariate normal pdf, then $\boldsymbol{\mathcal{X}} = \mathbb{R}^q$, and hence $\Delta_j = \infty$ for $j = 1, \ldots, q$. We now consider a modification of Result 5, analogous to the Laplace mechanism with truncation of Result 3, enabling ϵ-differential privacy to be obtained in the multivariate scenario without requiring that Δ_j is finite for all $j = 1, \ldots, q$.

Result 6. (Laplace mechanism with truncation for multivariate data) For each $j = 1, \ldots, q$, let $\epsilon_j > 0$, and let $L_j, U_j \in \mathbb{R}$ be such that $L_j < U_j$. If

$$Z_{ij} = \left\{ \begin{array}{ll} L_j + R_{ij}, & \text{if } X_{ij} < L_j, \\ X_{ij} + R_{ij}, & \text{if } L_j \leq X_{ij} \leq U_j, \\ U_j + R_{ij}, & \text{if } X_{ij} > U_j, \end{array} \right\}, \; j = 1, \ldots, q, \; i = 1, \ldots, n, \tag{15}$$

where $R_{ij} \sim \text{Lap}\,(0, (U_j - L_j)/\epsilon_j)$ independently for $j = 1, \ldots, q$ and $i = 1, \ldots, n$, then the conditional distribution of $\boldsymbol{Z}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy with $\epsilon = \sum_{j=1}^{q} \epsilon_j$.

For the rest of this section, let $\tilde{n} = n$, and let $\boldsymbol{Z}$ be defined as in Result 6. The conditional pdf of Z_{ij}, given $X_{ij} = x_{ij}$, is

$$g_{j,Z|X}(z_{ij}|x_{ij}) = \begin{cases} \frac{1}{2c_j} e^{-|z_{ij} - L_j|/c_j}, & \text{if } x_{ij} < L_j \\ \frac{1}{2c_j} e^{-|z_{ij} - x_{ij}|/c_j}, & \text{if } L_j \leq x_{ij} \leq U_j \\ \frac{1}{2c_j} e^{-|z_{ij} - U_j|/c_j}, & \text{if } x_{ij} > U_j \end{cases}$$

$$= I_{(-\infty, L_j)}(x_{ij}) \frac{e^{-|z_{ij} - L_j|/c_j}}{2c_j} + I_{[L_j, U_j]}(x_{ij}) \frac{e^{-|z_{ij} - x_{ij}|/c_j}}{2c_j} + I_{(U_j, \infty)}(x_{ij}) \frac{e^{-|z_{ij} - U_j|/c_j}}{2c_j},$$

where $c_j = (U_j - L_j)/\epsilon_j$, and I_A is the indicator function for the set A. The conditional pdf of $\boldsymbol{Z}_i$, given $\boldsymbol{X}_i = \boldsymbol{x}_i$, is

$$g_{Z|X}(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(x_{ij}) \frac{e^{-|z_{ij}-L_j|/c_j}}{2c_j} + I_{[L_j,U_j]}(x_{ij}) \frac{e^{-|z_{ij}-x_{ij}|/c_j}}{2c_j} + I_{(U_j,\infty)}(x_{ij}) \frac{e^{-|z_{ij}-U_j|/c_j}}{2c_j} \right\}.$$

The joint pdf of $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ is

$$\begin{aligned} g_{X,Z}(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) &= f(\boldsymbol{x}_i|\boldsymbol{\theta}) g_{Z|X}(\boldsymbol{z}_i|\boldsymbol{x}_i) \\ &= f(\boldsymbol{x}_i|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(x_{ij}) \frac{e^{-\frac{|z_{ij}-L_j|}{c_j}}}{2c_j} + I_{[L_j,U_j]}(x_{ij}) \frac{e^{-\frac{|z_{ij}-x_{ij}|}{c_j}}}{2c_j} + I_{(U_j,\infty)}(x_{ij}) \frac{e^{-\frac{|z_{ij}-U_j|}{c_j}}}{2c_j} \right\}. \end{aligned}$$

The marginal pdf of $\boldsymbol{Z}_i$ is

$$\begin{aligned} g_Z(\boldsymbol{z}_i|\boldsymbol{\theta}) &= \int_{\mathbb{R}^q} g_{X,Z}(\boldsymbol{w}, \boldsymbol{z}_i|\boldsymbol{\theta}) d\boldsymbol{w} \\ &= \int_{\mathbb{R}^q} f(\boldsymbol{w}|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(w_j) \frac{e^{-\frac{|z_{ij}-L_j|}{c_j}}}{2c_j} + I_{[L_j,U_j]}(w_j) \frac{e^{-\frac{|z_{ij}-w_j|}{c_j}}}{2c_j} + I_{(U_j,\infty)}(w_j) \frac{e^{-\frac{|z_{ij}-U_j|}{c_j}}}{2c_j} \right\} d\boldsymbol{w}. \end{aligned}$$

The conditional pdf of $\boldsymbol{X}_i$, given $\boldsymbol{Z}_i = \boldsymbol{z}_i$, is

$$\begin{aligned} g_{X|Z}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) &= \frac{g_{X,Z}(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{g_Z(\boldsymbol{z}_i|\boldsymbol{\theta})} \qquad (16) \\ &= \frac{f(\boldsymbol{x}_i|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(x_{ij}) \frac{e^{-\frac{|z_{ij}-L_j|}{c_j}}}{2c_j} + I_{[L_j,U_j]}(x_{ij}) \frac{e^{-\frac{|z_{ij}-x_{ij}|}{c_j}}}{2c_j} + I_{(U_j,\infty)}(x_{ij}) \frac{e^{-\frac{|z_{ij}-U_j|}{c_j}}}{2c_j} \right\}}{\int_{\mathbb{R}^q} f(\boldsymbol{w}|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(w_j) \frac{e^{-\frac{|z_{ij}-L_j|}{c_j}}}{2c_j} + I_{[L_j,U_j]}(w_j) \frac{e^{-\frac{|z_{ij}-w_j|}{c_j}}}{2c_j} + I_{(U_j,\infty)}(w_j) \frac{e^{-\frac{|z_{ij}-U_j|}{c_j}}}{2c_j} \right\} d\boldsymbol{w}} \\ &= \frac{f(\boldsymbol{x}_i|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(x_{ij}) e^{-\frac{|z_{ij}-L_j|}{c_j}} + I_{[L_j,U_j]}(x_{ij}) e^{-\frac{|z_{ij}-x_{ij}|}{c_j}} + I_{(U_j,\infty)}(x_{ij}) e^{-\frac{|z_{ij}-U_j|}{c_j}} \right\}}{\int_{\mathbb{R}^q} f(\boldsymbol{w}|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(w_j) e^{-\frac{|z_{ij}-L_j|}{c_j}} + I_{[L_j,U_j]}(w_j) e^{-\frac{|z_{ij}-w_j|}{c_j}} + I_{(U_j,\infty)}(w_j) e^{-\frac{|z_{ij}-U_j|}{c_j}} \right\} d\boldsymbol{w}}. \end{aligned}$$

Result 7. For given values of $\boldsymbol{z} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the following algorithm produces a random vector $\boldsymbol{X}^*$ having the density $g_{X|Z}(\boldsymbol{x}^*|\boldsymbol{z}, \boldsymbol{\theta})$, where $g_{X|Z}$ is the pdf defined in Equation (16). Define

$$K(\boldsymbol{x}, \boldsymbol{z}) = \frac{\prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(x_j) e^{-\frac{|z_j-L_j|}{c_j}} + I_{[L_j,U_j]}(x_j) e^{-\frac{|z_j-x_j|}{c_j}} + I_{(U_j,\infty)}(x_j) e^{-\frac{|z_j-U_j|}{c_j}} \right\}}{\prod_{j=1}^{q} \left\{ I_{(-\infty,L_j)}(z_j) e^{-\frac{|z_j-L_j|}{c_j}} + I_{[L_j,U_j]}(z_j) + I_{(U_j,\infty)}(z_j) e^{-\frac{|z_j-U_j|}{c_j}} \right\}}$$

and proceed as follows.

1. Generate V and $\boldsymbol{X}$ independently such that $V \sim \text{Uniform}(0,1)$ and $\boldsymbol{X} \sim f(\boldsymbol{x}|\boldsymbol{\theta})$.
2. If $V \leq K(\boldsymbol{X}, \boldsymbol{z})$ then accept $\boldsymbol{X}$ and deliver $\boldsymbol{X}^* = \boldsymbol{X}$. Otherwise reject $\boldsymbol{X}$ and return to Step 1.

The expected number of iterations of Steps 1 and 2 required to obtain $\boldsymbol{X}^*$ is

$$M(\boldsymbol{z}, \boldsymbol{\theta}) = C(\boldsymbol{z}, \boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty, L_j)}(z_j) e^{-|z_j - L_j|/c_j} + I_{[L_j, U_j]}(z_j) + I_{(U_j, \infty)}(z_j) e^{-|z_j - U_j|/c_j} \right\},$$

where

$$\frac{1}{C(\boldsymbol{z}, \boldsymbol{\theta})} = \int_{\mathbb{R}^q} f(\boldsymbol{w}|\boldsymbol{\theta}) \prod_{j=1}^{q} \left\{ I_{(-\infty, L_j)}(w_j) e^{-\frac{|z_j - L_j|}{c_j}} + I_{[L_j, U_j]}(w_j) e^{-\frac{|z_j - w_j|}{c_j}} + I_{(U_j, \infty)}(w_j) e^{-\frac{|z_j - U_j|}{c_j}} \right\} d\boldsymbol{w}.$$

Using the distributional results above, the likelihood based data analysis and imputation methods can be derived in a completely analogous way as how these results were derived in Sections 4.2. and 4.3., respectively. For the multivariate version of Algorithm 3, one would use Result 7 in place of Result 4 in step 1; but otherwise, extending each of the Algorithms 1, 2, and 3 to the multivariate scenario requires only the obvious modifications to account for vector-valued $\boldsymbol{X}_i$, instead of real-valued X_i. The methods described in Section 4.4. can be applied for data analysis based on the multiply imputed data. The (multivariate version of) Result 1 can be used to conclude that the conditional distribution of the multiply imputed data $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy under the assumed parametric scenario.

7. Conclusion

In this paper we have proposed methodology to randomly transform an original dataset $\boldsymbol{X}$ to $\boldsymbol{Z}$ using the Laplace mechanism with truncation as stated in Result 3. The modified version of the Laplace mechanism using truncation enables ϵ-differential privacy to be obtained without requiring the sample space of the original data to be bounded. We proposed to setup a missing data problem with $\boldsymbol{X}$ as the missing data, and $\boldsymbol{Z}$ as the observed data, and hence to multiply impute $\boldsymbol{X}$, based on $\boldsymbol{Z}$, thus obtaining $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$. Because the imputation process (Algorithm 1, 2, or 3) only uses $\boldsymbol{Z}$, but not $\boldsymbol{X}$, it follows that $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ is conditionally independent of $\boldsymbol{X}$, given $\boldsymbol{Z}$. Therefore, by Result 1, we may conclude that the conditional distribution of $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy. By releasing the multiply imputed data $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, data users are able to obtain valid inference for the unknown parameter $\boldsymbol{Q}(\boldsymbol{\theta})$ (scalar or vector) using standard statistical methods in conjunction with easily applicable multiple imputation combining formulas as explained in Section 4.4.. Indeed, an advantage of this approach is the ability to attain ϵ-differential privacy, while enabling data users to obtain valid inference using easily applied multiple imputation combining formulas. The data user does

not need to apply complicated or specialized inferential methodology to account for the noise mechanism, which, as discussed by Rubin (1993), is an advantage of using multiple imputation. In fact, upon observing $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, the data user does not even require knowledge about the noise infusion mechanism in Result 3 to obtain valid inference using the multiple imputation combining formulas; for example, the data user does not require the values of ϵ, L, and U to analyze $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$. The most complicated part of the proposed methodology is generating the imputed values of $\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)}$, and generating these values would be the job of the data producer, and not the data user. Furthermore, algorithms for generating these imputed values are provided in Section 4.3..

Throughout we have worked under the assumption that the original data $\boldsymbol{X}$, prior to collection, are known to be independent and identically distributed (iid) from a distribution within the parametric family $\mathcal{F} = \{f(x|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, and it is only the parameter $\boldsymbol{\theta}$ that is unknown and must be estimated from the data. We mention two comments regarding this parametric modeling assumption.

1. As discussed in Remark 4, the imputation procedures (Algorithms 1, 2, and 3) make use of the knowledge that the original data are iid from a distribution within $\mathcal{F}$, but do not make explicit use of $\boldsymbol{X}$, to generate the imputed values. Because the parametric family $\mathcal{F}$ is known in advance of data collection, and these Algorithms make use of $\boldsymbol{Z}$, but not $\boldsymbol{X}$, we are able to conclude that $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ is conditionally independent of $\boldsymbol{X}$, given $\boldsymbol{Z}$. Therefore Result 1 implies that the conditional distribution of $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy. However, if the parametric family $\mathcal{F}$ was not known a priori, but instead the imputer used $\boldsymbol{X}$ to model the parametric family $\mathcal{F}$, and then applied the imputation procedures of Section 4.3. using $\mathcal{F}$; then such a use of $\boldsymbol{X}$ to inform about $\mathcal{F}$, could cause $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ and $\boldsymbol{X}$, to be conditionally dependent, given $\boldsymbol{Z}$. If $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$ and $\boldsymbol{X}$ are conditionally dependent, given $\boldsymbol{Z}$, then Result 1 no longer applies, and we cannot conclude that the conditional distribution of $(\boldsymbol{x}^{*(1)}, \ldots, \boldsymbol{x}^{*(m)})$, given $\boldsymbol{X}$, satisfies ϵ-differential privacy.

2. In practice there is, of course, the possibility that the parametric family $\mathcal{F}$ is misspecified by the imputer and/or the data analyst. Such misspecification can lead to invalid inference, and we refer to Meng (1994) and Robins and Wang (2000) for further discussion.

The two points above indicate the role of the parametric modeling assumption in this paper. As future research one could study ways of relaxing the parametric assumption, perhaps through semi-parametric or non-parametric modeling approaches. Under the proposed methodology, one could also study the effects of model selection on differential privacy, and the effects of model misspecification on inference.

We are thankful to an anonymous reviewer who pointed out that it would be of interest to evaluate the possibility of a loss of accuracy of inference resulting from the proposed multiple imputation-based data analysis, as compared with approaches that analyze the

sanitized data directly. As future work one could also take up this research problem. With respect to asymptotic comparisons, Wang and Robins (1998) provide asymptotic theory for parametric multiple imputation procedures, and Robins and Wang (2000) provide further asymptotic theory for imputation estimators.

Acknowledgments

We thank Tommy Wright for encouragement, and we thank the reviewer for helpful comments. The original version of this paper was written while Martin Klein was a Principal Researcher, Center for Statistical Research and Methodology, U.S. Census Bureau.

References

Bartle R.G., Sherbert D.R. (2000). Introduction to Real Analysis, Third Edition. John Wiley and Sons, Inc, New York.

Charest, A.-S. (2010). How Can We Analyze Differentially-Private Synthetic Datasets? Journal of Privacy and Confidentiality. 2(2): 21-33.

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. 39(1): 1-38.

Differential Privacy Team, Apple. (2017). Learning with Privacy at Scale. Apple Machine Learning Journal. 1(8).

Drechsler J. (2011). Synthetic Datasets for Statistical Disclosure Control. Springer, New York.

Duchi J.C., Jordan M.I., Wainwright M.J. (2018). Minimax Optimal Procedures for Locally Private Estimation. Journal of the American Statistical Association. 113(521): 185-215.

Dwork C., McSherry F., Nissim K., Smith A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In Third Theory of Cryptography Conference, TCC: 265-284.

Dwork C., McSherry F., Nissim K., Smith A. (2017). Calibrating Noise to Sensitivity in Private Data Analysis. Journal of Privacy and Confidentiality. 7(3): 17-51.

Dwork C., Roth A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

Eddelbuettel D., Francois R. (2011). Rcpp: Seamless R and C++ Integration. Journal of Statistical Software, 40(8): 1-18. URL `http://www.jstatsoft.org/v40/i08/`.

Eddelbuettel D. (2013). Seamless R and C++ Integration with Rcpp. Springer, New York.

Eddelbuettel D., Balamuta J.J. (2017). Extending R with C++: A Brief Introduction to Rcpp. PeerJ Preprints 5:e3188v1. URL https://doi.org/10.7287/peerj.preprints.3188v1.

Erlingsson Ú., Pihur, V., Korolova A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security: 1054-1067.

Gelman A., Carlin J.B., Stern H.S., Dunson D.B., Vehtari A., Rubin D.B. (2014). Bayesian Data Analysis. Third Edition. Chapman and Hall/CRC, New York.

Hu J., Mitra R, Reiter J. (2013). Are Independent Parameter Draws Necessary for Multiple Imputation? The American Statistician. 67(3), 143-149.

Klein M., Sinha B. (2013). Statistical Analysis of Noise-Multiplied Data Using Multiple Imputation. Journal of Official Statistics. 29(3): 425-465.

Klein M., Sinha B. (2019). Multiple Imputation for Parametric Inference Under a Differentially Private Laplace Mechanism. Research Report Series (Statistics #2019-05), Center for Statistical Research and Methodology, U.S. Census Bureau. URL https://www.census.gov/library/working-papers/2019/adrm/RRS2019-05.html.

Li K.H., Raghunathan T.E., Rubin D.B. (1991). Large-Sample Significance Levels From Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution. Journal of the American Statistical Association. 86(416): 1065-1073.

Little R.A., Rubin D.B. (2002). Statistical Analysis With Missing Data. Second Edition. John Wiley and Sons, Inc, New Jersey.

Machanavajjhala A., Kifer D., Abowd J., Gehrke J., Vilhuber L. (2008). Privacy: Theory meets Practice on the Map. Proceedings of the 2008 IEEE 24th International Conference on Data Engineering: 277-286.

Meng X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science. 9(4): 538-573.

Nayak T.K., Zhang C., Adeshiyan S.A. (2015). Emerging Applications of Randomized Response Concepts and Some Related Issues. Model Assisted Statistics and Applications. 10(4); 335-344.

R Core Team. (2018). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raghunathan T.W., Reiter J.P., Rubin D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics. 19(1): 1-16.

Reiter J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology. 29(2): 181-188.

Reiter J.P. (2005). Significance Tests for Multi-Component Estimands from Multiply Imputed, Synthetic Microdata. Journal of Statistical Planning and Inference. 131(2): 365-377.

Reiter J.P., Raghunathan T.E. (2007). The Multiple Adaptations of Multiple Imputation. Journal of the American Statistical Association. 102(480): 1462-1471.

Robins J.M., Wang N. (2000). Inference for Imputation Estimators. Biometrika. 87(1): 113-124.

Rubin D.B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons, Inc, New Jersey.

Rubin D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics. 9(2): 461-468.

Rubin D.B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association. 91(434): 473-489.

Schafer J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman and Hall/CRC, New York.

Tanner M.A., Wong W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association. 82(398): 528-540.

Vadhan S. (2016) The Complexity of Differential Privacy. URL `https://privacytools.seas.harvard.edu/publications/complexity-differential-privacy`.

Wang N., Robins, J.M. (1998). Large-Sample Theory for Parametric Multiple Imputation Procedures. Biometrika. 85(4): 935-948.

Wasserman L., Zhou S. (2010). A Statistical Framework for Differential Privacy. Journal of the American Statistical Association. 105(489): 375-389.

Table 1: Simulation results under $N(\mu = 0, \sigma^2 = 1)$ with $m = 10$.

				Parameter of interest is μ						Parameter of interest is σ^2					
				RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg	EL	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg	EL
n	L	U	ϵ	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	%	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	%	$\times 10^2$
1000	-3	3	1	23.25	-0.13	23.25	22.51	94.04	100.62	139.02	98.82	97.78	98.81	98.64	445.62
			2	11.98	0.63	11.97	11.97	95.24	52.69	50.26	34.58	36.48	42.85	97.96	192.55
			3	8.60	0.15	8.60	8.62	94.60	37.35	28.28	14.66	24.19	26.80	97.24	119.81
			4	6.97	-0.13	6.97	6.94	95.28	29.62	20.21	7.24	18.87	19.70	96.08	87.49
			5	5.80	-0.01	5.80	5.95	95.64	25.02	15.61	4.25	15.02	15.59	96.44	68.67
			6	5.38	-0.01	5.38	5.33	94.48	22.14	12.96	2.56	12.70	12.81	95.20	55.83
1000	-4	4	1	28.80	0.51	28.80	28.31	94.92	127.04	149.38	110.59	100.43	108.01	98.60	487.20
			2	15.47	-0.33	15.47	15.10	94.76	66.99	66.72	49.78	44.42	54.47	98.36	245.11
			3	10.95	0.19	10.95	10.67	95.12	46.76	38.89	24.93	29.85	35.23	97.80	158.04
			4	8.71	0.02	8.71	8.54	94.48	37.00	27.90	15.03	23.50	26.40	96.80	117.98
			5	7.28	0.09	7.28	7.21	94.76	30.84	21.04	7.98	19.47	20.85	95.96	92.74
			6	6.41	0.08	6.41	6.36	95.04	26.93	17.02	5.05	16.25	17.23	96.72	76.21
5000	-3	3	1	9.92	0.21	9.92	9.69	94.96	43.44	64.58	42.11	48.97	38.74	90.44	174.94
			2	5.32	-0.03	5.32	5.28	95.80	23.35	22.10	9.05	20.17	20.19	95.24	91.00
			3	3.85	-0.07	3.85	3.80	94.64	16.52	13.22	3.38	12.78	12.81	95.40	57.49
			4	3.17	-0.09	3.17	3.08	94.36	13.15	9.35	1.31	9.26	9.12	95.48	40.63
			5	2.68	0.06	2.68	2.65	94.48	11.14	7.29	0.99	7.22	7.17	95.00	31.68
			6	2.43	-0.05	2.43	2.39	94.56	9.92	5.89	0.43	5.88	5.84	95.08	25.51
5000	-4	4	1	12.71	-0.08	12.71	12.21	94.00	54.87	85.91	58.43	62.98	46.32	87.12	209.21
			2	6.64	0.16	6.64	6.61	95.44	29.41	31.91	17.23	26.86	25.80	93.72	116.39
			3	4.77	0.10	4.77	4.72	95.00	20.78	18.20	6.13	17.13	17.53	95.84	78.91
			4	3.79	0.06	3.78	3.79	95.00	16.47	13.07	3.01	12.72	12.78	95.76	57.36
			5	3.20	-0.08	3.19	3.20	94.92	13.71	10.12	1.58	10.00	9.90	95.24	44.21
			6	2.86	-0.05	2.86	2.83	94.92	12.00	8.20	1.35	8.09	7.95	95.20	35.27

Table 2: Simulation results under $N(\mu = 0, \sigma^2 = 1)$ with $m = 30$.

				Parameter of interest is μ						Parameter of interest is σ^2					
				RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg	EL	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg	EL
n	L	U	ϵ	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	%	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	%	$\times 10^2$
1000	-3	3	1	22.34	0.39	22.34	22.33	95.32	91.05	131.42	101.49	83.49	110.65	99.92	452.27
			2	11.56	-0.39	11.55	11.73	95.48	47.60	47.90	33.71	34.03	42.55	99.28	173.73
			3	8.39	0.24	8.39	8.41	95.96	33.97	27.07	14.20	23.04	26.57	97.56	108.32
			4	6.65	-0.02	6.65	6.84	95.92	27.48	18.81	6.99	17.46	19.50	97.52	79.34
			5	5.96	-0.11	5.96	5.86	94.36	23.46	14.91	4.07	14.34	15.31	96.68	62.12
			6	5.26	-0.20	5.26	5.26	94.88	20.98	12.28	2.09	12.10	12.70	96.08	51.38
1000	-4	4	1	28.18	0.35	28.18	28.08	95.36	114.61	141.82	110.32	89.11	119.72	100.00	489.34
			2	14.54	0.31	14.53	14.92	96.04	60.70	63.71	49.81	39.73	55.25	99.32	225.68
			3	10.29	0.02	10.29	10.57	95.36	42.84	37.50	24.59	28.31	35.08	98.16	143.17
			4	8.37	0.24	8.37	8.41	94.68	33.95	26.60	14.15	22.52	25.94	97.40	105.75
			5	7.37	0.02	7.37	7.14	94.08	28.75	20.21	8.58	18.29	20.55	96.88	83.64
			6	6.27	0.05	6.27	6.30	95.60	25.26	16.99	5.37	16.12	16.87	95.56	68.55
5000	-3	3	1	9.67	-0.13	9.67	9.59	95.16	39.13	61.50	43.83	43.15	45.74	96.44	187.04
			2	5.20	0.06	5.20	5.18	95.28	21.05	21.66	9.39	19.52	20.46	95.72	83.60
			3	3.64	0.07	3.64	3.73	95.88	15.07	12.80	3.01	12.44	12.85	95.72	52.42
			4	3.00	-0.01	3.00	3.02	95.20	12.15	9.15	1.41	9.05	9.09	95.00	37.03
			5	2.66	-0.06	2.66	2.62	94.60	10.49	7.02	0.53	7.00	7.07	95.80	28.71
			6	2.36	0.09	2.36	2.35	95.16	9.38	5.75	0.49	5.73	5.78	95.08	23.40
5000	-4	4	1	12.48	0.18	12.48	12.32	94.52	50.30	80.76	61.08	52.83	56.23	95.44	229.91
			2	6.44	0.01	6.44	6.55	95.80	26.68	29.69	16.02	25.00	27.15	96.00	110.98
			3	4.72	0.08	4.72	4.70	94.60	19.06	18.15	6.47	16.96	17.61	95.56	71.94
			4	3.70	-0.11	3.70	3.73	95.44	15.08	12.47	3.21	12.05	12.63	95.64	51.55
			5	3.23	0.07	3.23	3.16	95.12	12.72	9.84	1.97	9.64	9.75	95.24	39.72
			6	2.81	-0.04	2.81	2.80	95.00	11.24	7.84	0.97	7.78	7.86	95.36	31.97

Table 3: Simulation results under $N(\mu = 0, \sigma^2 = 1)$ with $m = 50$.

				Parameter of interest is μ						Parameter of interest is σ^2					
				RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg	EL	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg	EL
n	L	U	ϵ	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	%	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	$\times 10^2$	%	$\times 10^2$
1000	-3	3	1	22.16	0.28	22.16	22.34	95.88	89.61	123.52	99.72	72.89	110.56	100.00	444.16
			2	11.71	0.14	11.70	11.74	94.80	46.98	47.70	34.59	32.84	43.27	99.40	173.71
			3	8.31	-0.06	8.31	8.39	95.32	33.46	26.50	14.16	22.40	26.52	98.20	106.40
			4	6.74	0.04	6.74	6.80	95.12	27.05	18.70	6.79	17.42	19.37	96.88	77.61
			5	5.96	-0.08	5.96	5.87	94.80	23.30	14.85	4.04	14.29	15.21	96.64	60.87
			6	5.24	-0.16	5.24	5.27	95.60	20.84	12.49	3.03	12.12	12.68	96.44	50.65
1000	-4	4	1	28.32	-0.40	28.32	28.03	94.72	112.50	144.26	114.81	87.36	127.46	100.00	512.07
			2	14.51	0.32	14.51	14.94	95.64	59.87	64.66	50.59	40.26	56.05	99.44	225.08
			3	10.48	0.35	10.47	10.52	94.88	42.07	37.13	24.70	27.73	34.95	98.44	140.28
			4	8.13	-0.15	8.13	8.40	95.40	33.51	25.25	13.53	21.32	25.83	97.92	103.61
			5	7.15	0.12	7.15	7.11	94.76	28.29	20.48	8.67	18.56	20.47	97.12	82.02
			6	6.29	-0.07	6.29	6.26	95.32	24.87	16.10	5.31	15.20	16.85	96.84	67.45
5000	-3	3	1	9.49	0.14	9.49	9.59	95.28	38.46	58.66	42.99	39.90	47.28	97.68	190.00
			2	5.29	-0.00	5.29	5.14	94.44	20.57	20.78	9.13	18.67	20.63	96.48	82.88
			3	3.72	-0.12	3.72	3.72	95.08	14.85	12.92	3.20	12.51	12.85	94.92	51.58
			4	3.00	-0.09	3.00	3.02	95.12	12.02	9.03	1.46	8.91	9.09	95.32	36.45
			5	2.56	0.04	2.56	2.61	95.44	10.34	7.00	0.79	6.95	7.03	95.40	28.15
			6	2.35	0.06	2.35	2.34	95.20	9.26	5.68	0.37	5.67	5.74	96.08	22.93
5000	-4	4	1	12.39	0.33	12.38	12.29	94.60	49.34	78.65	59.90	50.97	58.62	97.72	235.58
			2	6.45	0.02	6.45	6.57	95.92	26.33	29.88	17.47	24.23	27.66	97.08	111.11
			3	4.66	0.11	4.66	4.66	95.36	18.66	17.07	6.04	15.96	17.63	97.20	70.82
			4	3.68	0.12	3.68	3.71	95.16	14.81	12.45	3.12	12.06	12.64	96.04	50.72
			5	3.14	0.06	3.14	3.16	95.32	12.59	9.61	1.84	9.43	9.75	95.28	39.12
			6	2.81	-0.02	2.81	2.79	94.68	11.09	7.73	1.00	7.67	7.85	95.48	31.45