Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 425–465

Statistical Analysis of Noise-Multiplied Data Using Multiple Imputation

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

A statistical analysis of data that have been multiplied by randomly drawn noise variables in order to protect the confidentiality of individual values has recently drawn some attention. If the distribution generating the noise variables has low to moderate variance, then noise-multiplied data have been shown to yield accurate inferences in several typical parametric models under a formal likelihood-based analysis. However, the likelihood-based analysis is generally complicated due to the nonstandard and often complex nature of the distribution of the noise-perturbed sample even when the parent distribution is simple. This complexity places a burden on data users who must either develop the required statistical methods or implement the methods if already available or have access to specialized software perhaps yet to be developed. In this article we propose an alternate analysis of noise-multiplied data based on multiple imputation. Some advantages of this approach are that (1) the data user can analyze the released data as if it were never perturbed, and (2) the distribution of the noise variables does not need to be disclosed to the data user.

Key words: Combining rules; confidentiality; rejection sampling; statistical disclosure limitation; top coded data.

1. Introduction

When survey organizations and statistical agencies such as the U.S. Census Bureau release microdata to the public, a major concern is the control of disclosure risk, while ensuring fairly high quality and utility in the released data. Very often some popular statistical disclosure limitation (SDL) methods such as data swapping, multiple imputation, top/bottom coding (especially for income data), and perturbations with random noise are applied before releasing the data. Rubin (1993) proposed the use of the multiple imputation method to create synthetic microdata which would protect confidentiality by replacing actual microdata by random draws from a predictive distribution. Since then, rigorous statistical methods to use synthetic data for drawing valid inferences on relevant population parameters have been developed and used in many contexts (Little 1993;

[1] Martin Klein (Email: martin.klein@census.gov) is Research Mathematical Statistician in the Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, U.S.A.
[2] Bimal Sinha (Email: sinha@umbc.edu) is Research Mathematical Statistician in the Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, U.S.A. and Professor in the Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, U.S.A.
Acknowledgments: The authors thank Eric Slud for carefully reviewing the manuscript; Jerry Reiter for some valuable discussions; four anonymous referees and an associate editor for many helpful comments; and Joseph Schafer, Yves Thibaudeau, Tommy Wright and Laura Zayatz for encouragement. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Raghunathan et al. 2003; Reiter 2003, 2005; Reiter and Raghunathan 2007). An and Little (2007) also suggested multiple imputation methods as an alternative to top coding of extreme values and proposed two methods of data analysis with examples.

Noise perturbation of original microdata by addition or multiplication has also been advocated by some statisticians as a possible data confidentiality protection mechanism (Kim 1986; Kim and Winkler 1995, 2003; Little 1993), and recently there has been a renewed interest in this topic (Nayak et al. 2011; Sinha et al. 2012). In fact, Klein, Mathew, and Sinha (2013), hereafter referred to as Klein et al. (2013), developed likelihood-based data analysis methods under noise multiplication for drawing inference in several parametric models. They provided a comprehensive comparison of the above two methods, namely, multiple imputation and noise multiplication. Klein et al. (2013) commented that while standard and often *optimum* parametric inference based on the original data can be easily drawn for simple probability models, such an analysis is far from being close to optimum or even simple when noise multiplication is used. Hence their statistical analysis is essentially based on the asymptotic theory, requiring computational details of maximum likelihood estimation and calculations of the observed Fisher information matrices. Klein et al. (2013) also developed a similar analysis for top-coded data, which arise in many instances such as income and profit data, where values above a certain threshold C are coded and only the number m of values in the data set above C are reported along with all the original values below C. These authors considered statistical analysis based on unperturbed (i.e., original) data below C and noise-multiplied data above C instead of completely ignoring the data above C, and again provided a comparison with the statistical analysis reported in An and Little (2007), who carried out the analysis based on multiple imputation of the data above C in combination with the original values below C. In this article, we use the term *mixture* data, to refer to a data set in which values below a cut-off C are unperturbed, and values above C are perturbed via noise multiplication.

In the context of data analysis under noise perturbation, if the distribution generating the noise variables has low to moderate variance, then noise-multiplied data are expected to yield accurate inferences in some commonly used parametric models under a formal likelihood-based analysis (Klein et al. 2013). However, as noted by Klein et al. (2013), the likelihood-based analysis is generally complicated due to the nonstandard and often complex nature of the distribution of the noise-perturbed sample even when the parent distribution is simple (a striking example is analysis of noise-multiplied data under a *Pareto* distribution, typically used for income data, which we hope to address in a future communication). This complexity places a burden on data users who must either develop the required statistical methods or implement these methods if already available or have access to specialized software perhaps yet to be developed. Circumventing this difficulty is essentially the motivation behind this current research, where we propose an alternate simpler analysis of noise-multiplied data based on the familiar notion of multiple imputation. We believe that a proper blend of the two statistical methods as advocated here, namely, noise perturbation to protect confidentiality and multiple imputation for ease of subsequent statistical analysis of noise-multiplied data, will prove to be quite useful to both statistical agencies and data users. Some advantages of this approach are that (1) the data user can analyze the released data as if it were never perturbed (in conjunction with

the appropriate multiple imputation combining rules), and (2) the distribution of the noise variables does not need to be disclosed to the data user.

The article is organized as follows. An overview of our proposed approach based on a general framework of fully noise-multiplied data is given in Section 2. Techniques of noise imputation from noise-multiplied data, which are essential for the proposed statistical analysis, are also presented in Section 2. This section also includes different methods of estimation of variance of the proposed parameter estimates. Section 3 contains our statistical analysis for *mixture* data. Details of computations for the normal and lognormal models are outlined in Section 4. An evaluation and comparison of the results with those under a formal likelihood-based analysis of noise-multiplied data (Klein et al. 2013) is presented in Section 5 through simulation. It turns out that the inferences obtained using the methodology of this article are comparable with, and just slightly less accurate than, those obtained in Klein et al. (2013). Section 6 presents a disclosure risk evaluation of the proposed method, discusses the benefits of the proposed method in comparison with synthetic data, and outlines how to extend this approach to multivariate data. Section 7 provides some concluding remarks, and the Appendices A, B and C contain proofs of some technical results.

2. Methodology for Fully Noise-Multiplied Data

2.1. General Framework

Suppose $y_1, \ldots, y_n \sim iid \sim f(y|\boldsymbol{\theta})$, independent of $r_1, \ldots, r_n \sim iid \sim h(r)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is an unknown $p \times 1$ parameter vector, and $h(r)$ is a known density (free of $\boldsymbol{\theta}$) such that $h(r) = 0$ if $r < 0$. It is assumed that $f(y|\boldsymbol{\theta})$ and $h(r)$ are the densities of continuous probability distributions. Define $z_i = y_i \times r_i$ for $i = 1, \ldots, n$. Let us write $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{r} = (r_1, \ldots, r_n)$, and $\boldsymbol{z} = (z_1, \ldots, z_n)$.

We note that the joint density of (z_i, r_i) is

$$g(z_i, r_i|\boldsymbol{\theta}) = f\left(\frac{z_i}{r_i}\,\middle|\,\boldsymbol{\theta}\right) h(r_i) r_i^{-1},$$

and the marginal density of z_i is

$$g(z_i|\boldsymbol{\theta}) = \int_0^\infty f\left(\frac{z_i}{\omega}\,\middle|\,\boldsymbol{\theta}\right) h(\omega)\omega^{-1} d\omega. \tag{1}$$

As clearly demonstrated in Klein et al. (2013), standard likelihood-based analysis of the noise-multiplied sample $\boldsymbol{z}$ in order to draw suitable inference about a scalar quantity $Q = Q(\boldsymbol{\theta})$ can be extremely complicated due to the form of $g(z_i|\boldsymbol{\theta})$, and the analysis also must be customized to the noise distribution $h(r)$. Instead, what we propose here is a procedure to *reconstruct* the original data $\boldsymbol{y}$ from reported sample $\boldsymbol{z}$ via *suitable* generation and division by noise terms, and enough replications of the recovered $\boldsymbol{y}$ data by applying multiple imputation method. Once this is accomplished, a data user can apply a simple and standard likelihood procedure to draw inference about $Q(\boldsymbol{\theta})$ based on each reconstructed $\boldsymbol{y}$ data as if it were never perturbed, and finally an application of some known combination rules would complete the task.

The advantages of the suggested approach, blending noise multiplication with multiple imputation, are the following:

1. to protect confidentiality through noise multiplication – satisfying data producer's desire,
2. to allow the data user to analyze the data as if it were never perturbed – satisfying data user's desire (the complexity of the analysis lies in the generation of the imputed values of the noise variables; and the burden of this task will fall on the data producer, not the user), and
3. to allow the data producer to hide information about the underlying noise distribution from data users.

The basic idea behind our procedure is to set it up as a missing data problem; we define the complete, observed, and missing data, respectively, as follows:

$$\boldsymbol{x}_c = \{(z_1, r_1), \ldots, (z_n, r_n)\}, \quad \boldsymbol{x}_{\text{obs}} = \{z_1, \ldots, z_n\}, \quad \boldsymbol{x}_{\text{mis}} = \{r_1, \ldots, r_n\}.$$

Obviously, if the complete data $\boldsymbol{x}_c$ were observed, one would simply recover the original data $y_i = z_i/r_i$, $i = 1, \ldots, n$, and proceed with the analysis in a straightforward manner under the parametric model $f(y|\boldsymbol{\theta})$. Treating the noise variables $r_1, \ldots, r_n$ as missing data, we impute these variables m times to obtain

$$\boldsymbol{x}_c^{*(j)} = \left\{ \left(z_1, r_1^{*(j)}\right), \ldots, \left(z_n, r_n^{*(j)}\right) \right\}, \quad j = 1, \ldots, m. \tag{2}$$

From $\boldsymbol{x}^{*(j)}$ we compute

$$\boldsymbol{y}^{*(j)} = \left\{ y_1^{*(j)}, \ldots, y_n^{*(j)} \right\} = \left\{ \frac{z_1}{r_1^{*(j)}}, \ldots, \frac{z_n}{r_n^{*(j)}} \right\}, \quad j = 1, \ldots, m. \tag{3}$$

The statistical agency would then release the m imputed data sets $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(m)}$, and each data set $\boldsymbol{y}^{*(j)}$ would be analyzed as if it were a random sample from $f(y|\boldsymbol{\theta})$. Thus, suppose that $\eta(\boldsymbol{y})$ is an estimator of $Q(\boldsymbol{\theta})$ based on the unperturbed data $\boldsymbol{y}$ and suppose that $v = v(\boldsymbol{y})$ is an estimator of the variance of $\eta(\boldsymbol{y})$, also computed based on $\boldsymbol{y}$. Often $\eta(\boldsymbol{y})$ will be the maximum likelihood estimator (MLE) of $Q(\boldsymbol{\theta})$, and $v(\boldsymbol{y})$ will be derived from the observed Fisher information matrix. One would then compute $\eta_j = \eta(\boldsymbol{y}^{*(j)})$ and $v_j = v(\boldsymbol{y}^{*(j)})$, the analogs of η and v, obtained from $\boldsymbol{y}^{*(j)}$, and apply a suitable combination rule to pool the information across the m simulations.

At this point two vital pieces of the proposed methodology need to be put together: (1) imputation of $\boldsymbol{r}$ from $\boldsymbol{z}$, which would be the responsibility of the statistical agency; and (2) combination rules for η_j and v_j from several imputations, which the data user would apply in order to analyze the released data. We discuss these two crucial points in Subsections 2.2 and 2.3, respectively.

2.2. *Imputation of the Noise Variables*

In this subsection we describe two procedures that a statistical agency can use to impute $\boldsymbol{r}$ from $\boldsymbol{z}$. Following Wang and Robins (1998), we refer to these two methods as the Type A and Type B imputation procedures.

Type A Imputation Procedure. Under the Type A procedure, the imputed values of $r_1, \ldots, r_n$ are obtained as draws from a posterior predictive distribution. We place a noninformative prior distribution $p(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$. In principle, sampling from the posterior predictive distribution of $r_1, \ldots, r_n$ can be done as follows:

1. Draw $\boldsymbol{\theta}^*$ from the posterior distribution of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$.
2. Draw $r_1^*, \ldots, r_n^*$ from the conditional distribution of $r_1, \ldots, r_n$ given $z_1, \ldots, z_n$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

The above steps are then repeated independently m times to get $\left(r_1^{*(j)}, \ldots, r_n^{*(j)}\right)$, $j = 1, \ldots, m$.

Notice that in step (1) above we use the posterior distribution of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$ as opposed to the posterior distribution of $\boldsymbol{\theta}$ given $y_1, \ldots, y_n$. Such a choice implies that we do not infuse any additional information into the imputes beyond what is provided by the noise-multiplied sample $\boldsymbol{z}$ and the knowledge of the noise-generating distribution $h(r)$. Step (2) above is equivalent to sampling each r_i from the conditional distribution of r_i given z_i and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. The *pdf* of this distribution is

$$h(r_i|z_i, \boldsymbol{\theta}) = \frac{f\left((z_i/r_i)|\boldsymbol{\theta}\right)h(r_i)r_i^{-1}}{\int_0^\infty f\left((z_i/\omega)|\boldsymbol{\theta}\right)h(\omega)\omega^{-1}d\omega}. \tag{4}$$

The sampling required in step (1) can be complicated due to the complex form of the joint density of $z_1, \ldots, z_n$. Certainly, in some cases, the sampling required in step (1) can be performed directly; for instance, if $\boldsymbol{\theta}$ is univariate then we can obtain a direct algorithm by inversion of the cumulative distribution function (numerically or otherwise). More generally, the data augmentation algorithm (Little and Rubin 2002; Tanner and Wong 1987) allows us to bypass the direct sampling from the posterior distribution of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$. Under the data augmentation method, we proceed as follows. Given a value $\boldsymbol{\theta}^{(t)}$ of $\boldsymbol{\theta}$ drawn at step t:

I. Draw $r_i^{(t+1)} \sim h(r|z_i, \boldsymbol{\theta}^{(t)})$ for $i = 1, \ldots, n$;
II. Draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|\boldsymbol{y}^{(t+1)})$ where $\boldsymbol{y}^{(t+1)} = \left(\left(z_1/r_1^{(t+1)}\right), \ldots, \left(z_n/r_n^{(t+1)}\right)\right)$, and $p(\boldsymbol{\theta}|y)$ is the posterior density of $\boldsymbol{\theta}$ given the original unperturbed data $\boldsymbol{y}$ (it is the functional form of $p(\boldsymbol{\theta}|\boldsymbol{y})$ which is relevant here).

The above process is run until t is large and one must, of course, select an initial value $\boldsymbol{\theta}^{(0)}$ to start the iterations. The final generations $\left(r_1^{(t)}, \ldots, r_n^{(t)}\right)$ and $\boldsymbol{\theta}^{(t)}$ form an approximate draw from the joint posterior distribution of $(r_1, \ldots, r_n)$ and $\boldsymbol{\theta}$ given $(z_1, \ldots, z_n)$. Thus, marginally, the final generation $\left(r_1^{(t)}, \ldots, r_n^{(t)}\right)$ is an approximate draw from the posterior predictive distribution of $(r_1, \ldots, r_n)$ given $(z_1, \ldots, z_n)$. This entire iterative process can be repeated independently m times to get the multiply imputed values of the noise variables. The data augmentation algorithm presented here is equivalent to Gibbs sampling. The goal here is to sample from $p(\boldsymbol{r}, \boldsymbol{\theta}|\boldsymbol{z})$, the joint posterior distribution of $(\boldsymbol{r}, \boldsymbol{\theta})$ given $\boldsymbol{z}$. Letting $p(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{\theta})$ denote the conditional density of $\boldsymbol{r}$ given $\boldsymbol{z}$ and $\boldsymbol{\theta}$, and letting $p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{r})$ denote the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{z}$ and $\boldsymbol{r}$, we note that the $(t+1)$th step of a Gibbs sampler would sample from the full conditionals such that $r^{(t+1)} \sim p(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{\theta}^{(t)})$ and $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|\boldsymbol{z}, \mathbf{r}^{(t+1)})$, and would continue until convergence. Alternate sampling from

these two full conditional distributions is equivalent to steps I and II of the data augmentation algorithm.

Sampling from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ in step (II) above will typically be straightforward, either directly or via an embedded Markov chain Monte Carlo step. Under the data augmentation algorithm, we still must sample from the conditional density $h(r|z, \boldsymbol{\theta})$ as defined in (4). The level of complexity here will depend on the form of $f(y|\boldsymbol{\theta})$ and $h(r)$. Usually, sampling from this conditional density will not be too difficult. The following result provides a general rejection algorithm (Devroye 1986; Robert and Casella 2005) to sample from $h(r|z, \boldsymbol{\theta})$ for any continuous $f(y|\boldsymbol{\theta})$, when the noise distribution is Uniform $(1 - \epsilon, 1 + \epsilon)$, that is, when

$$h(r) = \frac{1}{2\epsilon}, \quad 1 - \epsilon \le r \le 1 + \epsilon, \tag{5}$$

where $0 < \epsilon < 1$.

Proposition 1 *Suppose that $f(y|\boldsymbol{\theta})$ is a continuous probability density function, and let us write $f(y|\boldsymbol{\theta}) = c(\boldsymbol{\theta})q(y|\boldsymbol{\theta})$ where $c(\boldsymbol{\theta}) > 0$ is a normalizing constant. Let $M \equiv M(\boldsymbol{\theta}, \epsilon, z)$ be such that*

$$q\left(\frac{z}{r}\middle|\boldsymbol{\theta}\right) \le M \text{ for all } r \in [1 - \epsilon, \gamma]$$

where $\gamma \equiv \gamma(z, \epsilon) > 1 - \epsilon$. Then the following algorithm produces a random variable R having the density

$$h_U(r|z, \boldsymbol{\theta}) = \frac{q((z/r)|\boldsymbol{\theta})r^{-1}}{\int_{1-\epsilon}^{\gamma} q((z/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega}, \quad 1 - \epsilon \le r \le \gamma.$$

I. *Generate U, V as independent Uniform*$(0, 1)$ *and let $W = \gamma^V/(1 - \epsilon)^{V-1}$.*
II. *Accept $R = W$ if $U \le M^{-1}q((z/W)|\boldsymbol{\theta})$, otherwise reject W and return to step (I).*

The expected number of iterations of steps (I) and (II) required to obtain R is

$$\frac{M[\log(\gamma) - \log(1 - \epsilon)]}{\int_{1-\epsilon}^{\gamma} q((z/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega}.$$

The proof of Proposition 1 appears in Appendix A.

Remark 1. The conditional density of y_i given z_i and $\boldsymbol{\theta}$ is

$$f(y_i|z_i, \boldsymbol{\theta}) = \begin{cases} \dfrac{f(y_i|\boldsymbol{\theta})h(z_i/y_i)y_i^{-1}}{\int_0^{\infty} f((z_i/\omega)|\boldsymbol{\theta})h(\omega)\omega^{-1}d\omega}, & \text{if } 0 < z_i < \infty, \quad 0 < y_i < \infty, \\ \dfrac{f(y_i|\boldsymbol{\theta})h(z_i/y_i)(-y_i^{-1})}{\int_0^{\infty} f((z_i/\omega)|\boldsymbol{\theta})h(\omega)\omega^{-1}d\omega}, & \text{if } -\infty < z_i < 0, \quad -\infty < y_i < 0. \end{cases} \tag{6}$$

Drawing r_i^* from the conditional density $h(r_i|z_i, \boldsymbol{\theta}^*)$ defined in (4) and setting $y_i^* = z_i/r_i^*$ is equivalent to drawing y_i^* directly from the conditional density $f(y_i|z_i, \boldsymbol{\theta}^*)$ in the sense that given z_i and $\boldsymbol{\theta}^*$, the variable z_i/r_i^* has the density $f(y_i|z_i, \boldsymbol{\theta}^*)$.

Remark 2. As to the choice of $\boldsymbol{\theta}^{(0)}$, one can choose moment-based estimates (Nayak et al. 2011).

Remark 3. We have tacitly assumed in the above analysis that the posterior distribution of the parameter $\boldsymbol{\theta}$, given noise-multiplied data $\boldsymbol{z}$, is proper. In applications, this needs to be verified on a case by case basis because the posterior propriety under the original data $\boldsymbol{y}$, which may routinely hold under many parametric models, may *not* guarantee the same under $\boldsymbol{z}$ when an improper prior distribution for $\boldsymbol{\theta}$ is used. We refer to the technical report Klein and Sinha (2013) for an example. The same remark holds in the case of the posterior distribution of $\boldsymbol{\theta}$, given the mixture data. We have verified the posterior propriety in our specific applications for fully noise-multiplied data and mixture data in Appendices B and C, respectively.

Type B Imputation Procedure. In this procedure there is no Bayesian model specification. Instead, the unknown parameter $\boldsymbol{\theta}$ is set equal to $\hat{\boldsymbol{\theta}}_{\text{mle}}(\boldsymbol{z})$, the MLE based on the noise-multiplied data $\boldsymbol{z}$, which can often be computed via the EM algorithm (Klein et al. 2013). The imputed values of the noise variables are then randomly drawn such that

$$r_i^* \sim h\big(r|z_i, \hat{\boldsymbol{\theta}}_{\text{mle}}(\boldsymbol{z})\big), \quad \text{for } i = 1, \ldots, n. \tag{7}$$

The above sampling is repeated, independently, m times to obtain $\big(r_1^{*(j)}, \ldots, r_n^{*(j)}\big)$, $j = 1, \ldots, m$. If $h(r)$ is the uniform density (5), then Proposition 1 can be used to implement the sampling in (7).

2.3. Combination Rules for Analyzing the Released Data

We now present methods for analyzing the released data $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(m)}$. Naturally, under the proposed methodology, analysis of the released data would usually be the responsibility of the data user. The analysis involves first analyzing each $\boldsymbol{y}^{*(j)}$ as if it were a random sample from $f(y|\boldsymbol{\theta})$, and then suitably combining the results across $j = 1, \ldots, m$ to obtain final inference. We first present the combination rules of Rubin (1987), which should yield valid inferences when the agency uses the Type A method to impute the noise variables. Rubin's (1987) combination rules often work well, and are simple to apply; however, they may not be optimal, and hence we also consider alternative methods of Wang and Robins (1998).

Rubin's (1987) Rule for Type A Imputation. We assume here that the released data (3) are obtained using the Type A imputation procedure. The multiple imputation estimator of Q is

$$\bar{\eta}_m = \frac{1}{m}\sum_{j=1}^{m} \eta_j, \tag{8}$$

and the estimator of the variance of $\bar{\eta}_m$ is

$$T_m = \left(1 + \frac{1}{m}\right) b_m + \bar{v}_m, \tag{9}$$

where $b_m = (1/(m-1))\sum_{j=1}^{m}(\eta_j - \bar{\eta}_m)^2$ and $\bar{v}_m = (1/m)\sum_{j=1}^{m} v_j$. The point estimator $\bar{\eta}_m$ and its variance estimator T_m can now be used along with a normal cut-off

point to construct a confidence interval for Q. We can also use a t cut-off point based on setting the degrees of freedom equal to $(m-1)\left(1+a_m^{-1}\right)^2$ where $a_m = (1+m^{-1})b_m/\bar{v}_m$.

Wang and Robins's (1998) Rule for Type A Imputation. Once again we assume that the released data (3) are obtained using the Type A imputation procedure. Let

$$\hat{\boldsymbol{\theta}}_j = \arg\max_{\boldsymbol{\theta}}\left\{\prod_{i=1}^{n} f\left(y_i^{*(j)}|\boldsymbol{\theta}\right)\right\}, j = 1, \ldots, m, \tag{10}$$

denote the MLE of $\boldsymbol{\theta}$ computed on the jth imputed data set $\boldsymbol{y}^{*(j)}$ under the model $f(y|\boldsymbol{\theta})$. The multiple imputation estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_A = (1/m)\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j$. By Wang and Robins (1998),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_p[\mathbf{0}, V_A], \text{ as } n \to \infty,$$

where $V_A = I_{\text{obs}}^{-1} + (1/m)I_c^{-1}J + (1/m)J'I_{\text{obs}}^{-1}J$, $J = I_{\text{mis}}I_c^{-1} = (I_c - I_{\text{obs}})I_c^{-1}$, and where I_c and I_{obs} are the $p \times p$ matrices defined by

$$I_c = E\left[-\left(\left(\frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial\,\theta_l \partial\,\theta_{l'}}\right)\right)\right] \text{ and } I_{obs} = E\left[-\left(\left(\frac{\partial^2 \log g(z|\boldsymbol{\theta})}{\partial\,\theta_l \partial\,\theta_{l'}}\right)\right)\right]. \tag{11}$$

Let $S_{ij}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)$ denote the $p \times 1$ score vector, with its lth element defined as

$$S_{ijl}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right) = \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial\,\theta_l}\Big|_{y=y_i^{*(j)}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j}, \; l = 1, \ldots, p, \; i = 1, \ldots, n, \; j = 1, \ldots, m;$$

and let $S_{ij}^*\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)$ denote the $p \times p$ matrix whose (l, l')th element is defined as

$$S_{ijll'}^*\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right) = \frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial\,\theta_l \partial\,\theta_{l'}}\Big|_{y=y_i^{*(j)}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j},$$

$$l, l' = 1, \ldots, p, \; i = 1, \ldots, n, \; j = 1, \ldots, m.$$

A consistent variance estimator $\hat{V}_A$ is obtained by estimating I_c *by*

$$\hat{I}_c = \frac{1}{m}\sum_{j=1}^{m}\hat{I}_{c,j}, \quad \hat{I}_{c,j} = -\frac{1}{n}\sum_{i=1}^{n} S_{ij}^*\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right), \tag{12}$$

and estimating I_{obs} by

$$\hat{I}_{\text{obs}} = \frac{1}{2nm(m-1)}\sum_{i=1}^{n}\sum_{j\neq j'=1}^{m}\left[S_{ij}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)S_{ij'}\left(y_i^{*(j')}, \hat{\boldsymbol{\theta}}_{j'}\right)' + S_{ij'}\left(y_i^{*(j')}, \hat{\boldsymbol{\theta}}_{j'}\right)S_{ij}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)'\right]. \tag{13}$$

For any given $Q(\boldsymbol{\theta})$, the variance of the multiple imputation estimator $Q(\hat{\boldsymbol{\theta}}_A)$ is obtained by applying the familiar δ-method, and Wald-type inferences can be directly applied to obtain confidence intervals.

Wang and Robins's (1998) Rule for Type B Imputation. We now assume that the released data (3) are obtained using the Type B imputation procedure. Let $\hat{\boldsymbol{\theta}}_j$ be defined

by (10). The multiple imputation estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_B = (1/m)\sum_{j=1}^{m} \hat{\boldsymbol{\theta}}_j$. By Wang and Robins (1998),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_p[0, V_B], \text{ as } n \to \infty,$$

where $V_B = I_{\text{obs}}^{-1} + (1/m)I_c^{-1}J = I_{\text{obs}}^{-1} + (1/m)I_c^{-1}(I_c - I_{\text{obs}})I_c^{-1}$ with I_c and I_{obs} defined in (11). A consistent variance estimator $\hat{V}_B$ is obtained by estimating I_c using (12) and estimating I_{obs} using (13). For any given $Q(\boldsymbol{\theta})$, the variance of the estimator $Q(\hat{\boldsymbol{\theta}}_B)$ is obtained by applying the familiar δ-method, and Wald-type inferences can be directly applied to obtain confidence intervals.

Remark 4. Wang and Robins (1998) provide a comparison between the Type A and Type B imputation procedures, and compare the corresponding variance estimators with Rubin's (1987) variance estimator T_m. Their observation is that the estimators $\hat{V}_A$ and $\hat{V}_B$ are consistent for V_A and V_B, respectively; and the Type B estimator $\hat{\boldsymbol{\theta}}_B$ will generally lead to more accurate inferences than $\hat{\boldsymbol{\theta}}_A$, because for finite m, $V_B < V_A$ (meaning $V_A - V_B$ is positive definite). Under the Type A procedure and for finite m, Rubin's (1987) variance estimator has a nondegenerate limiting distribution; however, the asymptotic mean is V_A, and thus T_m is also an appropriate estimator of variance (in defining Rubin's (1987) variance estimator, Wang and Robins (1998) multiply the quantity b_m by the sample size n to obtain a random variable that is bounded in probability). The variance estimator T_m would appear to underestimate the variance if applied in the Type B procedure because under the Type B procedure, if $m = \infty$, then T_m has a probability limit that is smaller than the asymptotic variance V_B (when $m = \infty$, $V_A = V_B = I_{\text{obs}}^{-1}$). However, under the Type A procedure, if $m = \infty$ then T_m is consistent for the asymptotic variance V_A. We refer to Rubin (1987) and Wang and Robins (1998) for further details.

3. Methodology for Mixture Data

Recall that the term mixture data in our context refers to a data set in which values below C are unperturbed and values above C are perturbed using noise multiplication. In this section we discuss the analysis of such data following the procedure outlined earlier, namely, by (*i*) suitably recovering the y-values above C via use of *reconstructed* noise terms and the noise-multiplied z-values along with or without their identities (below or above C), and (*ii*) providing multiple imputations of such y-values and methods to appropriately combine the original y-values and *reconstructed* y-values to draw inference on Q.

Let $C > 0$ denote the prescribed top code so that y-values above C are sensitive and hence cannot be reported/released. Given $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{r} = (r_1, \ldots, r_n)$, $\boldsymbol{z} = (z_1, \ldots, z_n)$ where $z_i = y_i \times r_i$, we define $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_n)$ with $\Delta_i = I(y_i \le C)$ and $x_i = y_i$ if $y_i \le C$, and $= z_i$ if $y_i > C$. Inference for $\boldsymbol{\theta}$ will be based on either (i) $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ or (ii) just $\{x_1, \ldots, x_n\}$. Under both the scenarios, which each guarantee that the sensitive y-values are protected, several data sets of the type $(y_1^*, \ldots, y_n^*)$ will be released along with a data analysis plan. We describe below the imputation and data analysis plans under both the scenarios.

Case (i). Here we generate r_i^* from the reported values of $(x_i, \Delta_i = 0)$ and compute $y_i^* = x_i / r_i^*$. Of course, if $\Delta_i = 1$ then we set $y_i^* = y_i$. Generation of r_i^* is done by sampling from the conditional distribution $h(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta})$ of r_i, given x_i, $\boldsymbol{\theta}$, and $\Delta_i = 0$, where

$$h(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta}) = \frac{f\left((x_i/r_i)|\boldsymbol{\theta}\right)h(r_i)r_i^{-1}}{\int_0^{(x_i/C)} f\left((x_i/\omega)|\boldsymbol{\theta}\right)h(\omega)\omega^{-1}d\omega}, \quad \text{for } 0 < r_i < \frac{x_i}{C} \tag{14}$$

(Klein et al. 2013) Note that the support of the above conditional distribution is such that $r_i^* \in (0, (x_i/C))$, and thus, if $\Delta_i = 0$, then $y_i^* = \left(x_i/r_i^*\right) > C$. That is, when $y_i > C$, the privacy-protected data point y_i^* has the desirable property that it will also be greater than C. When the noise distribution is the uniform density (5), then (14) can be written as

$$h_U(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta}) = \frac{f((x_i/r_i)|\boldsymbol{\theta})r_i^{-1}}{\int_{1-\epsilon}^{\min\{(x_i/C), 1+\epsilon\}} f((x_i/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega}, \quad \text{for } 1-\epsilon \le r_i \le \min\left\{\frac{x_i}{C}, 1+\epsilon\right\}, \tag{15}$$

and Proposition 1 provides an algorithm for sampling from the above density (15).

Regarding choice of $\boldsymbol{\theta}$, we can proceed following the Type B method (Section 2) and use the MLE of $\boldsymbol{\theta}$ ($\hat{\boldsymbol{\theta}}_{\text{mle}}$) based on the data $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$. This will often be direct (via EM algorithm) in view of the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta)$ reported in Klein et al. (2013) and reproduced below:

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta) = \prod_{i=1}^{n} [f(x_i|\boldsymbol{\theta})]^{\Delta_i} \left[\int_0^{(x_i/C)} f\left(\frac{x_i}{r}|\theta\right)\frac{h(r)}{r}dr\right]^{1-\Delta_i}. \tag{16}$$

Alternatively, following Type A method discussed in Section 2, r^*-values can also be obtained as draws from a posterior predictive distribution. We place a noninformative prior distribution $p(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, and sampling from the posterior predictive distribution of $r_1, \ldots, r_n$ can be done as follows:

1. Draw $\boldsymbol{\theta}^*$ from the posterior distribution of $\boldsymbol{\theta}$ given $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ using the likelihood $L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta)$ given above.
2. Draw r_i^* for those $i = 1, \ldots, n$ for which $\Delta_i = 0$, from the conditional distribution (14) of r_i, given x_i, $\Delta_i = 0$, and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

As mentioned in Section 2, the sampling required in step (1) above can be complicated due to the complex form of the joint density $L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta)$. The data augmentation algorithm (Little and Rubin 2002; Tanner and Wong 1987) allows us to bypass the direct sampling from the posterior distribution of $\boldsymbol{\theta}$ given $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$.

Under the data augmentation method, given a value $\boldsymbol{\theta}^{(t)}$ of $\boldsymbol{\theta}$ drawn at step t:

I. Draw $r_i^{(t+1)} \sim h(r|x_i, \Delta_i = 0, \boldsymbol{\theta}^{(t)})$ for those $i = 1, \ldots, n$ for which $\Delta_i = 0$.
II. Draw $\boldsymbol{\theta}^{(t+1)} \sim p\left(\boldsymbol{\theta}|y_1^{(t+1)}, \ldots, y_n^{(t+1)}\right)$ where $y_i^{(t+1)} = x_i/r_i^{(t+1)}$ when $\Delta_i = 0$, and $y_i^{(t+1)} = x_i$ otherwise. Here $p(\boldsymbol{\theta}|\boldsymbol{y})$ stands for the posterior *pdf* of $\boldsymbol{\theta}$, given the original data $\boldsymbol{y}$ (only its functional form is used).

The above process is run until t is large and one must, of course, select an initial value $\boldsymbol{\theta}^{(0)}$ to start the iterations.

Case (ii). Here we generate $\left(r_i^{**}, \Delta_i^*\right)$ from the reported values of $(x_1, \ldots, x_n)$ and compute $y_i^{**} = \left(x_i / r_i^{**}\right)$ if $\Delta_i^* = 0$, and $y_i^{**} = x_i$, otherwise, $i = 1, \ldots, n$. This is done by using the conditional distribution $g(r, \delta | x, \boldsymbol{\theta})$ of r and Δ, given x and $\boldsymbol{\theta}$. Since $g(r, \delta | x, \boldsymbol{\theta}) = h(r | x, \delta, \boldsymbol{\theta}) \times \psi(\delta | x, \boldsymbol{\theta})$, and the conditional Bernoulli distribution of Δ, given x and $\boldsymbol{\theta}$, is readily given by

$$\psi(\delta = 1 | x, \boldsymbol{\theta}) = \Pr\{\Delta = 1 | x, \boldsymbol{\theta}\}$$

$$= \frac{f(x|\boldsymbol{\theta}) I(x < C)}{f(x|\boldsymbol{\theta}) I(x < C) + I(x > 0) \int_0^{x/C} f((x/r)|\boldsymbol{\theta}) h(r) r^{-1} dr} \tag{17}$$

(Klein et al. 2013), drawing of $\left(r_i^{**}, \Delta_i^*\right)$, given x_i and $\boldsymbol{\theta}$, is carried out by first randomly selecting Δ_i^* according to the above Bernoulli distribution, and then randomly choosing r_i^{**} if $\Delta_i^* = 0$ from the conditional distribution given by (14).

Again, in the above computations, following Type B approach, one can use the MLE of $\boldsymbol{\theta}$ (via EM algorithm) based on the $\boldsymbol{x}$-data alone whose likelihood is given by

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} \left[f(x_i|\boldsymbol{\theta}) I(x_i < C) + I(x_i > 0) \int_0^{x_i/C} f\left(\frac{x_i}{r} | \boldsymbol{\theta}\right) h(r) r^{-1} dr \right] \tag{18}$$

(Klein et al. 2013). Alternatively, one can proceed as in Type A method (sampling $r_1^{**}, \ldots, r_n^{**}$ from the posterior predictive distribution) by plugging in $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ that are random draws from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{x}$, based on the above likelihood and choice of prior for $\boldsymbol{\theta}$. As noted in the previous case, here too a direct sampling of $\boldsymbol{\theta}$, given $\boldsymbol{x}$, can be complicated, and we can use the data augmentation algorithm suitably modified following the two steps indicated below.

1. Starting with an initial value of $\boldsymbol{\theta}$ and hence $\boldsymbol{\theta}^{(t)}$ at step t, draw $\left(r_i^{(t+1)}, \Delta_i^{(t+1)}\right)$ $h(r, \delta | x_i, \boldsymbol{\theta}^{(t)})$. This of course is accomplished by first drawing $\Delta_i^{(t+1)}$ and then $r_i^{(t+1)}$, in case $\Delta_i^{(t+1)} = 0$.
2. At step $(t + 1)$, draw $\boldsymbol{\theta}^{(t+1)}$ from the posterior distribution $p\left(\boldsymbol{\theta} | y_1^{(t+1)}, \ldots, y_n^{(t+1)}\right)$ of $\boldsymbol{\theta}$, where $y_i^{(t+1)} = x_i$ if $\Delta_i^{(t+1)} = 1$, and $y_i^{(t+1)} = x_i / r_i^{(t+1)}$ if $\Delta_i^{(t+1)} = 0$. Here, as before, the functional form of the *standard* posterior of $\boldsymbol{\theta}$, given $\boldsymbol{y}$, is used.

In both case (i) and case (ii), after recovering the multiply imputed complete data $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(m)}$ using the techniques described above, methods of parameter estimation, variance estimation, and confidence interval construction are the same as those discussed in Section 2 for fully noise-multiplied data. Naturally, in case (i) when information on the indicator variables Δ is used to generate y^*-values, data users will know exactly which y-values are original and which y-values have been noise-perturbed and de-perturbed. Of course, this need not happen in case (ii), thus providing more privacy protection with perhaps less accuracy. Thus the data producer (such as the Census Bureau) has a choice depending upon to what extent information about the released data should be provided to the data users.

4. Details for Normal and Lognormal Data

In this section we provide some details of the proposed methodology for normal and lognormal populations. Similar details for the exponential population appear in the technical report Klein and Sinha (2013).

4.1. Normal Data

We consider the case of a normal population with uniform noise, that is, we take $f(y|\boldsymbol{\theta}) = (1/(\sigma\sqrt{2\pi}))\exp[-(1/(2\sigma^2))(y-\mu)^2]$, $-\infty < y < \infty$, and we let $h(r)$ be the uniform density (5). We place a standard noninformative improper prior on (μ, σ^2):

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}, \quad -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty. \tag{19}$$

The posterior distribution of (μ, σ^2) given $\boldsymbol{y}$ is obtained as $p(\mu, \sigma^2|\boldsymbol{y}) = p(\mu|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y})$ where

$$(\sigma^2|\boldsymbol{y}) \sim \frac{(n-1)s^2}{\chi^2_{n-1}}, \quad (\mu|\sigma^2, \boldsymbol{y}) \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right), \tag{20}$$

with $\bar{y} = (1/n)\sum_{i=1}^n y_i$ and $s^2 = (1/(n-1))\sum_{i=1}^n (y_i - \bar{y})^2$ (Gelman et al. 2003). The conditional density $h(r|z, \boldsymbol{\theta})$ as defined in (4) now takes the form

$$h(r|z, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))((z/r)-\mu)^2]r^{-1}}{\int_{1-\epsilon}^{1+\epsilon} \exp[-(1/(2\sigma^2))((z/\omega)-\mu)^2]\omega^{-1}d\omega}, \ 1-\epsilon \le r \le 1+\epsilon. \tag{21}$$

We apply Proposition 1 to obtain an algorithm for sampling from this conditional density of r_i given z_i.

Corollary 1 *The following algorithm produces a random variable R whose density is (21).*

I. *Generate U, V as independent Uniform*(0, 1) *and let* $W = (1+\epsilon)^V/(1-\epsilon)^{V-1}$.
II. *Accept* $R = W$ *if* $U \le \exp[(-1/(2\sigma^2))(z/W-\mu)^2]/M$, *otherwise reject W and return to step (I).*

If $z > 0$ *then the constant M is defined as*

$$M \equiv M(\mu, \sigma^2, \epsilon, z) = \begin{cases} \exp\left[-\frac{1}{2\sigma^2}(z/(1+\epsilon)-\mu)^2\right], & \text{if } \mu \le z/(1+\epsilon), \\ 1, & \text{if } z/(1+\epsilon) < \mu < z/(1-\epsilon), \\ \exp\left[-\frac{1}{2\sigma^2}(z/(1-\epsilon)-\mu)^2\right], & \text{if } \mu \ge z/(1-\epsilon). \end{cases}$$

and if $z < 0$ *then*

$$M \equiv M(\mu, \sigma^2, \epsilon, z) = \begin{cases} \exp\left[-\frac{1}{2\sigma^2}(z/(1-\epsilon)-\mu)^2\right], & \text{if } \mu \le z/(1-\epsilon), \\ 1, & \text{if } z/(1-\epsilon) < \mu < z/(1+\epsilon), \\ \exp\left[-\frac{1}{2\sigma^2}(z/(1+\epsilon)-\mu)^2\right], & \text{if } \mu \ge z/(1+\epsilon). \end{cases}$$

The expected number of iterations of steps (I) and (II) required to obtain R is

$$\frac{M[\log(1+\epsilon) - \log(1-\epsilon)]}{\int_{1-\epsilon}^{1+\epsilon} \exp[-(1/(2\sigma^2))((z/\omega)-\mu)^2]\omega^{-1}d\omega}.$$

In the case of mixture data, the conditional density (14) now becomes

$$h(r|x, \Delta = 0, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))(x/r-\mu)^2]r^{-1}}{\int_{1-\epsilon}^{\min\{(x/C),1+\epsilon\}} \exp[-(1/(2\sigma^2))(x/\omega-\mu)^2]\omega^{-1}d\omega}, \quad 1-\epsilon \le r \le \min\left\{\frac{x}{C}, 1+\epsilon\right\}, \tag{22}$$

and a simple modification of Corollary 1 yields an algorithm to sample from this *pdf*.

4.2. *Lognormal Data*

We next consider the case of the lognormal population: $f(y|\boldsymbol{\theta}) = (1/(y\sigma\sqrt{2\pi}))\exp[(-1/(2\sigma^2))(\log y - \mu)^2]$, $0 \le y < \infty$. We define a prior distribution on (μ, σ^2) as in (19). The posterior distribution of (μ, σ^2) is then given by (20) upon replacing each y_i by $\log(y_i)$.

Customized noise distribution for fully perturbed data. Let us take the noise density as:

$$h(r) = \frac{1}{r\xi\sqrt{2\pi}} \exp\left[-\frac{1}{2\xi^2}(\log r + \xi^2/2)^2\right], \quad 0 < r < \infty, \tag{23}$$

where $0 < \xi < \infty$, and $E(R) = 1$ and $\mathrm{Var}(R) = e^{\xi^2} - 1$. We note that $h(r)$ is a lognormal density such that $R \sim h(r) \Leftrightarrow \log(R) \sim N(-\xi^2/2, \xi^2)$. It then follows that $h(r|z, \boldsymbol{\theta})$ is also a lognormal density such that

$$R \sim h(r|z, \boldsymbol{\theta}) \Leftrightarrow \log(R) \sim N\left\{-\frac{\xi^2}{2} + \frac{\xi^2}{\sigma^2+\xi^2}\left[\log(z) + \frac{\xi^2}{2} - \mu\right], \frac{\sigma^2\xi^2}{\sigma^2+\xi^2}\right\}. \tag{24}$$

Uniform noise distribution. Suppose we take the noise distribution to be uniform as defined in (5). Then the conditional *pdf* (4) takes the form

$$h(r|z, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))(\log(z/r) - \mu)^2]}{\int_{1-\epsilon}^{1+\epsilon} \exp\ -(1/(2\sigma^2))(\log(z/\omega) - \mu)^2 d\omega}, \quad 1-\epsilon \le r \le 1+\epsilon \tag{25}$$

We apply Proposition 1 to obtain an algorithm for sampling from this conditional density of r_i given z_i.

Corollary 2 *The following algorithm produces a random variable R whose density is (25).*

I. *Generate U, V as independent Uniform*(0, 1) *and let* $W = (1+\epsilon)^V/(1-\epsilon)^{V-1}$.
II. *Accept* $R = W$ *if* $U \le Wz^{-1}\exp[-(1/(2\sigma^2))(\log(z/W) - \mu)^2]/M$, *otherwise reject W and return to step (I).*

The constant M is defined as

$$M \equiv M(\mu, \sigma^2, \epsilon, z)$$

$$= \begin{cases} (1+\epsilon) z^{-1} \exp\left[-\frac{1}{2\sigma^2} (\log(z/(1+\epsilon)) - \mu)^2 \right], & \text{if } e^{\mu - \sigma^2} \le z/(1+\epsilon), \\ \exp\left[-\mu + \frac{\sigma^2}{2} \right], & \text{if } z/(1+\epsilon) < e^{\mu-\sigma^2} < z/(1-\epsilon), \\ (1-\epsilon) z^{-1} \exp\left[-\frac{1}{2\sigma^2} (\log(z/(1-\epsilon)) - \mu)^2 \right], & \text{if } e^{\mu-\sigma^2} \ge z/(1-\epsilon). \end{cases}$$

The expected number of iterations of steps (I) and (II) required to obtain R is

$$\frac{M[\log(1+\epsilon) - \log(1-\epsilon)]}{\int_{1-\epsilon}^{1+\epsilon} z^{-1} \exp[-(1/(2\sigma^2))(\log(z/\omega) - \mu)^2] d\omega}.$$

In the case of mixture data, the conditional density (14) now becomes

$$h(r|x, \Delta = 0, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))(\log(x/r) - \mu)^2]}{\int_{1-\epsilon}^{\min\{(x/C), 1+\epsilon\}} \exp[-(1/(2\sigma^2))(\log(x/\omega) - \mu)^2] d\omega}, \quad (26)$$

$$1 - \epsilon \le r \le \min\left\{ \frac{x}{C}, 1 + \epsilon \right\},$$

and a simple modification of Corollary 2 yields an algorithm to sample from this *pdf*.

5. Simulation Study to Assess Accuracy of Inference

We use simulation to study the finite sample properties of point estimators, variance estimators, and confidence intervals obtained from noise-multiplied data. We consider the cases of normal and lognormal populations in conjunction with uniform and customized noise distributions as far as possible, as outlined in Section 4. The results for the exponential population are similar to the normal and lognormal, and appear in the technical report Klein and Sinha (2013). One may expect that the simpler method of data analysis proposed in this paper may lead to less accurate inferences than a formal likelihood-based analysis of fully noise-multiplied and mixture data. However, if the inferences derived using the proposed methodology are not substantially less accurate, then the proposed method may be preferable, in some cases, because of its simplicity. Thus the primary goals of this section are essentially to (1) compare the proposed methods with the likelihood-based method reported in Klein et al. (2013), and (2) to assess and compare the finite sample performance of Rubin's (1987) estimation methods with those of Wang and Robins (1998) under our settings of fully noise-multiplied and mixture data.

Each of the tables discussed below is based on a simulation with 5,000 iterations and $m = 5$ imputations of the noise variables generated at each iteration. We choose $m = 5$ because this is a fairly small number of imputations which may be conveniently used in practice. In each of the 5,000 iterations, five independent runs of the data augmentation algorithm, each having 50 iterations, are used to obtain the Type A imputations. Some

exploratory analysis indicated that 50 iterations of the data augmentation algorithm provided an adequate approximation in the chosen simulation settings. All results are obtained using the statistical computing software R (R Development Core Team 2011).

5.1. Fully Noise-Multiplied Data

Table 1 provides results for the case of a normal population when the parameter of interest is either the mean μ or the variance σ^2; and Table 2 provides results for the case of a lognormal population when the parameter of interest is either the mean $e^{\mu+\sigma^2/2}$ or the .95 quantile $e^{\mu+1.645\sigma}$. For each distribution we consider samples sizes $n = 100$ and $n = 500$, but we only display results for the former sample size. Each table displays results for several different methods which are summarized below.

UD: Analysis based on the unperturbed data $\boldsymbol{y}$.

NM10UIB: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type B imputation method and the associated combining rules of Wang and Robins (1998).

NM10UIA1: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type A imputation method and Rubin's (1987) combining rules with the normal cut-off point for confidence interval construction.

NM10UIA2: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type A imputation method and Rubin's (1987) combining rules with the t cut-off point for confidence interval construction.

NM10UIA3: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type A imputation method and the associated combining rules of Wang and Robins (1998).

NM10UL: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the formal likelihood based method of analysis of Klein et al. (2013).

NM10CIB, NM10CIA1, NM10CIA2, NM10CIA3, NM10CL: These methods are defined analogously to the methods above, but $h(r)$ is now the customized noise distribution (23) (for lognormal data); the parameters δ and ξ appearing in $h(r)$ are chosen so that if $R \sim h(r)$, then $\text{Var}(R) = (\epsilon^2)/3$, the variance of the Uniform $(1 - \epsilon, 1 + \epsilon)$ distribution with $\epsilon = 0.10$.

The remaining methods appearing in these tables are similar to the corresponding methods mentioned above after making the appropriate change to the parameter ϵ in the referenced Uniform$(1 - \epsilon, 1 + \epsilon)$ distribution. For each method and each parameter of interest, we display the root mean squared error of the estimator (RMSE), bias of the estimator, standard deviation of the estimator (SD), average over simulation runs of the estimated standard deviation of the estimator ($\widehat{\text{SD}}$), empirical coverage probability of the associated confidence interval (Cvg.), and average length (over simulation iterations) of the corresponding confidence interval relative to the average length of the confidence interval computed from the unperturbed data (Rel. Len.). In each case the nominal coverage probability of the confidence interval is 0.95. For computing an estimate of the standard deviation of an estimator, we simply compute the square root of the appropriate variance estimator. For computing the estimator $\eta(\boldsymbol{y})$ and variance estimator $v(\boldsymbol{y})$ of

Table 1. Inference under fully perturbed N ($\mu = 0$, $\sigma^2 = 1$) data with $n = 100$

	Parameter of interest is the mean μ						Parameter of interest is the variance σ^2					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	99.99	3.24	99.94	99.40	94.66	1.0000	143.89	−6.70	143.73	140.47	92.92	1.0000
NM10UIB	100.11	3.18	100.06	100.99	95.16	1.0160	145.92	−6.02	145.79	148.67	93.06	1.0584
NM10UIA1	100.10	3.12	100.05	99.62	94.80	1.0021	145.87	−6.34	145.74	142.14	92.66	1.0119
NM10UIA2	100.10	3.12	100.05	99.62	94.80	1.0021	145.87	−6.34	145.74	142.14	92.66	1.0119
NM10UIA3	100.10	3.12	100.05	101.24	95.12	1.0185	145.87	−6.34	145.74	149.76	93.36	1.0661
NM10UL	100.10	3.13	100.05	99.58	94.74	1.0018	145.59	−6.32	145.46	141.87	92.62	1.0100
NM20UIB	100.92	3.27	100.87	101.48	95.20	1.0209	150.15	−4.91	150.07	152.80	93.56	1.0878
NM20UIA1	100.83	3.10	100.79	100.26	94.92	1.0086	150.45	−4.17	150.39	146.89	93.02	1.0457
NM20UIA2	100.83	3.10	100.79	100.26	94.92	1.0087	150.45	−4.17	150.39	146.89	93.02	1.0458
NM20UIA3	100.83	3.10	100.79	101.78	95.38	1.0238	150.45	−4.17	150.39	154.09	93.70	1.0969
NM20UL	100.74	3.09	100.69	100.11	94.94	1.0071	149.43	−4.84	149.35	145.74	93.10	1.0375
NM50UIB	103.96	3.39	103.90	104.18	94.80	1.0480	170.21	−4.83	170.15	173.55	93.26	1.2355
NM50UIA1	104.11	3.46	104.06	103.53	94.40	1.0415	171.79	1.78	171.78	169.74	93.12	1.2083
NM50UIA2	104.11	3.46	104.06	103.53	94.52	1.0438	171.79	1.78	171.78	169.74	93.16	1.2109
NM50UIA3	104.11	3.46	104.06	104.79	94.56	1.0541	171.79	1.78	171.78	176.64	93.78	1.2575
NM50UL	103.31	3.29	103.26	102.64	94.52	1.0326	167.38	−4.24	167.32	164.29	93.16	1.1695

Table 2. Inference under fully perturbed LN ($\mu = 0$, $\sigma^2 = 1$) data with $n = 100$

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{SD}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{SD}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	202.26	1.56	202.26	201.82	93.88	1.0000	799.81	−11.83	799.72	793.27	93.16	1.0000
NM10UIB	202.80	1.69	202.79	208.31	94.10	1.0321	802.34	−11.16	802.26	826.38	93.16	1.0417
NM10UIA1	203.18	1.83	203.17	202.46	93.64	1.0032	803.98	−10.57	803.91	796.22	92.88	1.0037
NM10UIA2	203.18	1.83	203.17	202.46	93.64	1.0032	803.98	−10.57	803.91	796.22	92.88	1.0037
NM10UIA3	203.18	1.83	203.17	208.34	94.16	1.0323	803.98	−10.57	803.91	826.50	93.30	1.0419
NM10UL	202.91	1.70	202.90	202.31	93.62	1.0025	802.80	−11.16	802.72	795.55	92.78	1.0029
NM10CIB	202.72	1.48	202.71	208.30	93.92	1.0321	801.97	−12.25	801.88	826.34	93.34	1.0417
NM10CIA1	202.81	1.52	202.80	202.38	93.80	1.0028	802.40	−11.87	802.31	795.89	93.04	1.0033
NM10CIA2	202.81	1.52	202.80	202.38	93.80	1.0028	802.40	−11.87	802.31	795.89	93.04	1.0033
NM10CIA3	202.81	1.52	202.80	208.29	94.02	1.0320	802.40	−11.87	802.31	826.26	93.38	1.0416
NM10CL	202.68	1.41	202.68	202.25	93.84	1.0021	801.56	−12.39	801.47	795.26	93.20	1.0025
NM20UIB	204.60	2.55	204.59	210.24	94.16	1.0417	811.20	−7.89	811.16	835.35	93.26	1.0530
NM20UIA1	204.76	2.21	204.75	204.24	93.84	1.0120	811.69	−9.16	811.63	804.47	93.02	1.0141
NM20UIA2	204.76	2.21	204.75	204.24	93.84	1.0122	811.69	−9.16	811.63	804.47	93.02	1.0144
NM20UIA3	204.76	2.21	204.75	210.16	94.02	1.0413	811.69	−9.16	811.63	834.97	93.34	1.0526
NM20UL	204.33	2.29	204.32	203.83	93.94	1.0099	810.06	−8.76	810.02	802.52	93.34	1.0117
NM20CIB	204.59	2.05	204.58	209.93	94.18	1.0402	810.41	−11.38	810.33	834.00	93.22	1.0513
NM20CIA1	204.41	1.72	204.40	204.04	94.02	1.0110	809.98	−12.28	809.89	803.51	92.98	1.0129
NM20CIA2	204.41	1.72	204.40	204.04	94.04	1.0112	809.98	−12.28	809.89	803.51	93.00	1.0132
NM20CIA3	204.41	1.72	204.40	209.88	94.08	1.0399	809.98	−12.28	809.89	833.77	93.28	1.0511
NM20CL	204.06	1.62	204.05	203.56	93.98	1.0086	808.43	−12.73	808.33	801.31	92.92	1.0101
NM50UIB	217.16	1.62	217.16	221.96	94.06	1.0998	866.70	−16.33	866.55	890.55	93.30	1.1226
NM50UIA1	217.31	2.95	217.29	216.77	93.44	1.0741	867.67	−9.31	867.62	862.13	92.64	1.0868

Table 2. Continued

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
NM50UIA2	217.31	2.95	217.29	216.77	93.56	1.0810	867.67	−9.31	867.62	862.13	92.78	1.0960
NM50UIA3	217.31	2.95	217.29	222.23	93.62	1.1012	867.67	−9.31	867.62	891.63	92.78	1.1240
NM50UL	214.82	0.82	214.81	213.53	93.52	1.0580	855.59	−17.25	855.41	847.91	92.86	1.0689
NM50CIB	214.35	3.42	214.32	220.94	93.96	1.0948	854.98	−7.29	854.95	885.62	93.58	1.1164
NM50CIA1	215.22	4.67	215.17	215.77	93.84	1.0691	857.50	−1.24	857.50	857.56	93.16	1.0810
NM50CIA2	215.22	4.67	215.17	215.77	93.94	1.0749	857.50	−1.24	857.50	857.56	93.32	1.0888
NM50CIA3	215.22	4.67	215.17	221.25	94.02	1.0963	857.50	−1.24	857.50	886.83	93.50	1.1179
NM50CL	212.48	2.53	212.46	212.80	93.96	1.0544	845.95	−9.46	845.90	844.25	93.00	1.0643

Subsection 2.2, we use the maximum likelihood estimator and inverse of observed Fisher information, respectively. All results shown for unperturbed data use Wald-type inferences based on the maximum likelihood estimator and observed Fisher information. The following is a summary of the simulation results of Tables 1–2.

1. In terms of RMSE, bias, and SD of point estimators, as well as average confidence interval length, the proposed methods of analysis are generally only slightly less accurate than the corresponding likelihood-based analysis.
2. In terms of coverage probability of confidence intervals, the multiple imputation-based and formal likelihood-based methods of analysis yield similar results.
3. We consider Uniform$(1 - \epsilon, 1 + \epsilon)$ noise distributions with $\epsilon = 0.1$, 0.2, and 0.5, or equivalent (in terms of variance) customized noise distributions. Generally, for noise distributions with $\epsilon = 0.1$ and 0.2, the proposed analysis based on the noise-multiplied data results only in a slight loss of accuracy in comparison with that based on unperturbed data. When the noise distribution has a larger variance (i.e., when $\epsilon = 0.5$) we notice that the bias of the resulting estimators generally remains small, while the SD clearly increases. When the parameter of interest is the mean, the noise-multiplied data with $\epsilon = 0.5$ still appear to provide inferences with only a slight loss of accuracy compared with the unperturbed data. In contrast, when the parameter of interest is the normal variance as in the right-hand panel of Table 1, the loss of accuracy in terms of SD and hence RMSE appears to be more substantial when ϵ increases to 0.5. We refer to Klein et al. (2013) for a detailed study of the properties of noise-multiplied data.
4. We observe very little difference in the bias, SD, and RMSE of estimators derived under the Type A imputation procedure versus those derived under the Type B imputation procedure.
5. In each table, the column $\widehat{\text{SD}}$ provides the finite sample mean of each of the multiple imputation standard deviation estimators (square root of variance estimators) presented in Section 2. Thus we can compare the finite sample bias of Rubin's (1987) standard deviation estimator of Subsection 2.2 with that of Wang and Robins's (1998) standard deviation estimators of Subsection 2.3 under our setting of noise multiplication. We find that the mean of both of Wang and Robins's (1998) standard deviation estimators is generally larger than the mean of Rubin's (1987) standard deviation estimator. From these numerical results it appears that we cannot make any general statement about which estimators possess the smallest bias, because none of these estimators uniformly dominates the other in terms of minimization of bias. With a larger sample size of $n = 500$ (results not displayed here), we find that all standard deviation estimators have similar expectation; this statement is especially true for the normal case. With the sample size of $n = 100$ we notice in Table 1 that the mean of Rubin's (1987) SD estimator is slightly less than the true SD while both of Wang and Robins's (1998) estimators have a mean slightly larger than the true SD. We should point out that this slight negative bias of Rubin's (1987) SD estimator is most likely due to the fact that the SD estimator based on the original data is itself slightly downward-biased. In the lognormal case, for the sample size $n = 100$ of

Table 2, we notice that Rubin's (1987) estimator is nearly unbiased for the true SD while Wang and Robins's (1998) estimators tend to overestimate the true SD more substantially.

6. When the customized noise distribution is available (e.g., exponential and lognormal cases), the results obtained under the customized noise distribution are quite similar to those obtained under the equivalent (in terms of variance) uniform noise distribution.
7. For confidence interval construction based on Rubin's (1987) variance estimator, the interval based on the normal cut-off point performs very similarly to the interval based on the t cut-off point.
8. The data augmentation algorithm, used by the Type A methods to sample from the posterior predictive distribution of $\boldsymbol{r}$, given the noise-multiplied data, appears to provide an adequate approximation.

5.2. *Mixture Data*

We now study the properties of estimators derived from mixture data as presented in Section 3. Table 3 provides results for the case of a normal population, and Table 4 provides results for the case of a lognormal population. The parameters of interest in each case are the same as in the previous subsection, and the top-coding threshold value C is set equal to the 0.90 quantile of the population. The methods in the rows of Tables 3–4 are as described in the previous subsection, except that each ends with either .i or .ii to indicate either case (i) or case (ii) of Section 3, respectively. The conclusions here are generally in line with those of the previous subsection. Below are some additional findings.

1. Rubin's (1987) SD estimator in this case tends to exhibit very little bias.
2. Generally we find here that the noise multiplication methods yield quite accurate inferences, even more so than in the case of full noise multiplication; this finding is expected since with mixture data only a subset of the original observations are noise-perturbed.
3. As expected, the inferences derived under the case (i) data scenario (observe $(\boldsymbol{x}, \Delta)$) are generally more accurate than those derived under the case (ii) data scenario (observe only $\boldsymbol{x}$), but for the noise distributions considered, the differences in accuracy generally are not too substantial.

6. Further Evaluations and Extensions

6.1. *Disclosure Risk Evaluation*

In this section we report the results of a numerical study designed to give an indication of the amount of disclosure protection provided by the proposed methodology. To be specific, we determine how tightly the m draws $y_i^{*(1)}, \ldots, y_i^{*(m)}$ are centered around the true value y_i, and how well the average and median of these m draws approximate the true value y_i. We consider both the fully noise-multiplied data and mixture data scenarios.

Table 3. Inference for mixture N ($\mu = 0$, $\sigma^2 = 1$) data with C = .90 quantile and n = 100

	Parameter of interest is the mean μ						Parameter of interest is the variance σ^2					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	98.70	−1.21	98.69	99.30	94.50	1.0000	139.88	−9.00	139.59	140.15	93.68	1.0000
NMl0UIB.i	98.81	−1.18	98.81	101.00	94.88	1.0171	140.72	−8.81	140.45	149.18	94.10	1.0645
NMl0UIAl.i	98.79	−1.17	98.78	99.37	94.46	1.0007	140.62	−8.75	140.35	140.88	93.48	1.0053
NM10UIA2.i	98.79	−1.17	98.78	99.37	94.46	1.0007	140.62	−8.75	140.35	140.88	93.48	1.0053
NM10UIA3.i	98.79	−1.17	98.78	101.01	94.82	1.0172	140.62	−8.75	140.35	149.17	94.14	1.0644
NMl0UL.i	98.81	−1.19	98.80	99.36	94.48	1.0005	140.54	−8.87	140.26	140.74	93.56	1.0042
NMl0UIB.ii	98.83	−1.15	98.83	101.01	94.78	1.0172	140.71	−8.67	140.45	149.20	94.16	1.0646
NMl0UIAl.ii	98.81	−1.20	98.81	99.37	94.50	1.0006	140.76	−8.89	140.48	140.87	93.52	1.0052
NM10UIA2.ii	98.81	−1.20	98.81	99.37	94.50	1.0006	140.76	−8.89	140.48	140.87	93.52	1.0052
NM10UIA3.ii	98.81	−1.20	98.81	101.00	94.84	1.0171	140.76	−8.89	140.48	149.21	94.06	1.0646
NMl0UL.ii	98.81	−1.20	98.80	99.36	94.38	1.0006	140.54	−8.88	140.26	140.75	93.54	1.0043
NM20UIB.i	99.23	−1.12	99.22	101.10	94.70	1.0181	142.24	−8.52	141.99	150.74	93.92	1.0756
NM20UIAl.i	99.13	−0.97	99.13	99.55	94.48	1.0025	142.10	−7.89	141.88	142.68	93.64	1.0180
NM20UIA2.i	99.13	−0.97	99.13	99.55	94.48	1.0025	142.10	−7.89	141.88	142.68	93.64	1.0186
NM20UIA3.i	99.13	−0.97	99.13	101.13	94.90	1.0184	142.10	−7.89	141.88	150.71	94.12	1.0753
NM20UL.i	99.09	−1.06	99.09	99.51	94.42	1.0021	141.77	−8.20	141.54	142.24	93.56	1.0149
NM20UIB.ii	99.17	−1.11	99.17	101.13	94.78	1.0184	142.12	−8.37	141.88	150.76	93.90	1.0757
NM20UIAl.ii	99.13	−0.96	99.12	99.58	94.36	1.0028	142.61	−7.76	142.39	142.80	93.40	1.0189
NM20UIA2.ii	99.13	−0.96	99.12	99.58	94.36	1.0028	142.61	−7.76	142.39	142.80	93.44	1.0195
NM20UIA3.ii	99.13	−0.96	99.12	101.16	94.62	1.0187	142.61	−7.76	142.39	150.79	94.02	1.0760
NM20UL.ii	99.10	−1.07	99.09	99.52	94.40	1.0022	141.92	−8.25	141.68	142.31	93.44	1.0154
NM50UIB.i	99.67	−0.59	99.66	101.41	94.56	1.0212	148.43	−6.19	148.30	155.53	94.04	1.1098
NM50UIAl.i	99.77	−0.05	99.77	100.18	94.32	1.0088	149.25	−3.94	149.20	148.33	93.72	1.0584

Table 3. Continued

	Parameter of interest is the mean μ						Parameter of interest is the variance σ^2					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
NM50UIA2.i	99.77	−0.05	99.77	100.18	94.32	1.0089	149.25	−3.94	149.20	148.33	93.78	1.0630
NM50UIA3.i	99.77	−0.05	99.77	101.53	94.64	1.0224	149.25	−3.94	149.20	155.79	94.08	1.1116
NM50UL.i	99.55	−0.57	99.54	99.96	94.32	1.0066	147.32	−6.08	147.19	146.70	93.66	1.0467
NM50UIB.ii	99.99	−0.64	99.99	101.82	94.86	1.0254	150.46	−6.41	150.33	157.79	93.84	1.1259
NM50UIAl.ii	100.07	−0.01	100.07	100.60	94.44	1.0130	150.68	−3.90	150.63	150.30	93.64	1.0724
NM50UIA2.ii	100.07	−0.01	100.07	100.60	94.46	1.0133	150.68	−3.90	150.63	150.30	93.70	1.0791
NM50UIA3.ii	100.07	−0.01	100.07	101.98	94.76	1.0270	150.68	−3.90	150.63	158.04	94.08	1.1277
NM50UL.ii	99.74	−0.72	99.74	100.29	94.48	1.0100	148.93	−6.48	148.79	148.34	93.66	1.0584

Table 4. Inference for mixture LN ($\mu = 0$, $\sigma^2 = 1$) data with $C = .90$ quantile and $n = 100$

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	99.45	2.10	99.43	99.21	94.78	1.0000	781.78	1.65	781.78	794.90	93.48	1.0000
NMl0UIB.i	99.46	2.11	99.44	100.87	95.08	1.0167	783.15	1.95	783.15	824.70	93.62	1.0375
NMl0UIAl.i	99.46	2.11	99.43	99.23	94.78	1.0002	783.32	2.04	783.32	796.04	93.40	1.0014
NM10UIA2.i	99.46	2.11	99.43	99.23	94.78	1.0002	783.32	2.04	783.32	796.04	93.40	1.0014
NM10UIA3.i	99.46	2.11	99.43	100.87	95.10	1.0167	783.32	2.04	783.32	824.66	93.72	1.0374
NMl0UL.i	99.47	2.11	99.45	99.22	94.78	1.0002	783.09	1.97	783.09	795.82	93.38	1.0011
NMl0UIB.ii	99.48	2.11	99.46	100.87	95.06	1.0167	783.92	2.22	783.92	824.75	93.72	1.0376
NMl0UIAl.ii	99.47	2.09	99.45	99.22	94.72	1.0002	783.18	1.64	783.18	796.00	93.36	1.0014
NM10UIA2.ii	99.47	2.09	99.45	99.22	94.72	1.0002	783.18	1.64	783.18	796.00	93.36	1.0014
NM10UIA3.ii	99.47	2.09	99.45	100.86	95.10	1.0167	783.18	1.64	783.18	824.70	93.70	1.0375
NMl0UL.ii	99.47	2.11	99.45	99.23	94.72	1.0002	783.15	1.98	783.14	795.85	93.36	1.0012
NM20UIB.i	99.50	2.10	99.47	100.89	95.12	1.0169	787.17	2.26	787.17	827.71	93.60	1.0413
NM20UIAl.i	99.47	2.10	99.45	99.27	94.82	1.0006	786.76	2.44	786.76	798.97	93.30	1.0051
NM20UIA2.i	99.47	2.10	99.45	99.27	94.82	1.0006	786.76	2.44	786.76	798.97	93.30	1.0052
NM20UIA3.i	99.47	2.10	99.45	100.89	95.04	1.0170	786.76	2.44	786.76	827.52	93.82	1.0410
NM20UL.i	99.49	2.08	99.47	99.26	94.80	1.0005	785.69	1.62	785.69	798.04	93.34	1.0039
NM20UIB.ii	99.50	2.08	99.47	100.90	95.04	1.0170	786.09	1.92	786.09	827.94	93.66	1.0416
NM20UIAl.ii	99.51	2.09	99.49	99.28	94.84	1.0008	787.30	2.51	787.30	799.37	93.44	1.0056
NM20UIA2.ii	99.51	2.09	99.49	99.28	94.84	1.0008	787.30	2.51	787.30	799.37	93.44	1.0057
NM20UIA3.ii	99.51	2.09	99.49	100.91	95.06	1.0171	787.30	2.51	787.30	827.80	93.72	1.0414
NM20UL.ii	99.50	2.07	99.48	99.26	94.76	1.0006	785.97	1.54	785.97	798.27	93.36	1.0042
NM50UIB.i	99.83	2.33	99.80	101.09	95.24	1.0189	804.56	9.96	804.50	842.34	93.76	1.0597
NM50UIAl.i	99.84	2.46	99.81	99.58	94.90	1.0037	803.02	12.96	802.92	816.58	93.50	1.0273

Table 4. Continued

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
NM50UIA2.i	99.84	2.46	99.81	99.58	94.90	1.0038	803.02	12.96	802.92	816.58	93.50	1.0282
NM50UIA3.i	99.84	2.46	99.81	101.12	95.12	1.0193	803.02	12.96	802.92	842.43	93.96	1.0598
NM50UL.i	99.73	2.32	99.71	99.50	94.86	1.0029	798.51	8.40	798.47	811.47	93.56	1.0208
NM50UIB.ii	100.05	2.42	100.02	101.32	95.18	1.0213	809.84	12.40	809.75	850.03	93.74	1.0694
NM50UIAl.ii	100.07	2.55	100.04	99.78	94.78	1.0058	809.88	14.73	809.75	822.84	93.68	1.0351
NM50UIA2.ii	100.07	2.55	100.04	99.78	94.78	1.0058	809.88	14.73	809.75	822.84	93.70	1.0366
NM50UIA3.ii	100.07	2.55	100.04	101.34	95.12	1.0215	809.88	14.73	809.75	850.50	93.78	1.0699
NM50UL.ii	99.96	2.40	99.93	99.68	94.66	1.0047	803.94	10.09	803.87	817.17	93.54	1.0280

Table 5. Illustration of y, z, and y-values for fully perturbed LN ($\mu = 0$, $\sigma^2 = 1$) data with n = 100 and uniform noise*

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
0.26	0.1	0.26	5	0.25	0.25	0.28	0.27	0.29	0.29	0.26	0.27	0.27	0.27	0.28	0.28
	0.2	0.26		0.23	0.24	0.29	0.27	0.29	0.30	0.22	0.24	0.25	0.27	0.31	0.31
	0.5	0.29		0.20	0.26	0.38	0.34	0.42	0.46	0.28	0.34	0.37	0.39	0.43	0.55
	0.1	0.26	5000	0.24	0.25	0.27	0.27	0.28	0.29	0.24	0.25	0.27	0.27	0.28	0.29
	0.2	0.26		0.22	0.24	0.27	0.27	0.29	0.32	0.22	0.24	0.27	0.27	0.29	0.32
	0.5	0.29		0.19	0.27	0.35	0.36	0.45	0.58	0.19	0.27	0.35	0.36	0.45	0.58
0.96	0.1	1.05	5	1.01	1.09	1.11	1.11	1.17	1.17	0.96	1.02	1.05	1.07	1.14	1.16
	0.2	1.07		0.94	0.95	1.02	1.02	1.07	1.10	0.94	1.11	1.16	1.16	1.29	1.32
	0.5	0.76		0.64	0.67	0.95	0.96	1.10	1.46	0.63	0.64	0.86	0.93	1.22	1.29
	0.1	1.05	5000	0.96	1.00	1.05	1.05	1.10	1.17	0.96	1.00	1.05	1.05	1.10	1.17
	0.2	1.07		0.89	0.98	1.07	1.09	1.19	1.34	0.89	0.98	1.07	1.09	1.19	1.34
	0.5	0.76		0.51	0.65	0.81	0.87	1.05	1.53	0.51	0.63	0.81	0.87	1.07	1.52
2.98	0.1	3.26	5	3.08	3.09	3.14	3.17	3.27	3.29	2.98	3.01	3.10	3.20	3.31	3.63
	0.2	2.56		2.18	2.21	2.62	2.50	2.71	2.78	2.21	2.43	2.55	2.56	2.56	3.03
	0.5	2.84		1.92	1.98	2.13	3.00	4.39	4.58	2.58	3.12	3.83	3.68	4.01	4.87
	0.1	3.26	5000	2.97	3.10	3.25	3.26	3.42	3.63	2.97	3.10	3.25	3.26	3.42	3.62
	0.2	2.56		2.13	2.31	2.53	2.57	2.81	3.20	2.13	2.31	2.53	2.57	2.80	3.20
	0.5	2.84		1.89	2.16	2.58	2.87	3.35	5.68	1.89	2.17	2.57	2.86	3.30	5.68
8.95	0.1	8.13	5	8.00	8.73	8.88	8.70	8.93	8.97	7.43	7.61	8.11	8.03	8.40	8.58
	0.2	9.06		8.24	8.40	9.14	9.17	9.66	10.38	8.24	8.48	9.28	9.09	9.54	9.91
	0.5	7.22		5.04	5.27	5.38	6.55	5.88	11.18	5.01	5.13	5.78	5.66	5.99	6.40
	0.1	8.13	5000	7.39	7.70	8.06	8.11	8.48	9.03	7.39	7.70	8.07	8.11	8.48	9.04
	0.2	9.06		7.55	8.06	8.72	8.95	9.71	11.32	7.55	8.05	8.74	8.96	9.73	11.32
	0.5	7.22		4.82	5.32	6.08	6.75	7.49	14.41	4.81	5.33	6.10	6.75	7.61	14.42
18.21	0.1	19.03	5	17.59	17.62	19.01	18.89	19.14	21.06	17.42	19.13	19.58	19.28	19.66	20.62
	0.2	21.79		19.99	21.68	24.91	23.79	26.19	26.19	18.69	20.59	24.11	23.11	25.33	26.84

Table 5. Continued

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
	0.5	20.42		13.79	14.84	15.93	15.46	16.15	16.58	13.96	14.42	17.64	16.64	18.31	18.88
	0.1	19.03	5000	17.30	17.95	18.75	18.90	19.77	21.14	17.30	17.96	18.77	18.90	19.74	21.14
	0.2	21.79		18.16	19.21	20.62	21.21	22.86	27.23	18.16	19.24	20.69	21.27	22.94	27.24
	0.5	20.42		13.61	14.71	16.37	17.92	19.42	40.62	13.61	14.70	16.25	17.88	19.35	40.81

Table 6. Illustration of y, z, and y-values for fully perturbed LN ($\mu = 0$, $\sigma^2 = 1$) data with n = 100 and customized noise*

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
0.26	0.1	0.28	5	0.25	0.28	0.28	0.28	0.29	0.30	0.28	0.28	0.29	0.29	0.29	0.30
	0.2	0.22		0.19	0.20	0.22	0.22	0.24	0.25	0.18	0.19	0.22	0.22	0.22	0.27
	0.5	0.31		0.24	0.34	0.36	0.35	0.37	0.44	0.36	0.37	0.37	0.38	0.39	0.42
	0.1	0.28	5000	0.23	0.27	0.28	0.28	0.29	0.35	0.23	0.27	0.28	0.28	0.29	0.34
	0.2	0.22		0.15	0.21	0.22	0.23	0.24	0.34	0.14	0.21	0.22	0.23	0.24	0.33
	0.5	0.31		0.14	0.29	0.36	0.37	0.43	0.93	0.16	0.30	0.36	0.37	0.43	0.90
0.96	0.1	1.05	5	0.97	1.03	1.05	1.04	1.05	1.09	0.94	1.00	1.07	1.04	1.11	1.11
	0.2	1.08		0.91	0.95	0.98	1.08	1.27	1.31	0.95	1.01	1.15	1.12	1.21	1.29
	0.5	0.75		0.61	0.79	0.79	0.85	0.93	1.10	0.62	0.68	0.69	0.71	0.76	0.82
	0.1	1.05	5000	0.85	1.01	1.05	1.05	1.09	1.28	0.85	1.01	1.05	1.05	1.09	1.30
	0.2	1.08		0.74	1.01	1.09	1.10	1.18	1.83	0.71	1.01	1.09	1.10	1.18	1.58
	0.5	0.75		0.33	0.68	0.81	0.84	0.98	1.97	0.25	0.68	0.81	0.85	0.98	2.48
2.98	0.1	2.98	5	2.91	2.95	2.99	3.07	3.26	3.26	2.85	2.89	2.94	2.99	3.13	3.14
	0.2	3.08		2.46	2.76	2.93	2.82	2.95	3.02	2.85	2.93	3.30	3.26	3.45	3.79
	0.5	3.10		2.90	2.92	3.59	3.52	4.06	4.13	1.97	2.80	2.81	3.48	3.29	6.52
	0.1	2.98	5000	2.39	2.87	2.99	2.99	3.10	3.58	2.49	2.87	2.98	2.99	3.10	3.69
	0.2	3.08		1.98	2.83	3.06	3.08	3.30	4.48	2.04	2.85	3.08	3.09	3.32	4.56
	0.5	3.10		0.96	2.52	3.02	3.14	3.65	6.91	1.00	2.51	2.99	3.12	3.60	8.03
8.95	0.1	7.75	5	7.41	7.54	7.70	7.72	7.80	8.16	7.21	7.23	7.56	7.54	7.70	7.99
	0.2	7.74		5.86	7.36	7.54	7.46	7.80	8.77	6.38	6.89	7.23	7.75	8.96	9.27
	0.5	8.28		5.20	6.56	8.56	8.27	10.45	10.57	6.05	6.49	6.59	7.12	7.99	8.46
	0.1	7.75	5000	6.22	7.42	7.71	7.72	8.02	9.39	6.20	7.42	7.72	7.73	8.02	9.38
	0.2	7.74		4.66	7.05	7.63	7.68	8.24	12.45	4.80	7.05	7.60	7.66	8.22	11.55
	0.5	8.28		2.71	6.24	7.48	7.75	8.96	17.58	2.47	6.18	7.42	7.71	8.97	21.01
18.21	0.1	18.01	5	17.18	17.26	18.06	18.00	18.59	18.88	16.32	17.16	18.16	17.76	18.53	18.62
	0.2	18.00		14.41	15.40	17.19	16.83	18.17	18.96	18.92	19.12	19.98	20.63	21.22	23.90

Table 6. Continued

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
	0.5	31.32		25.48	29.49	30.15	29.78	31.63	32.16	18.39	24.32	26.99	25.35	28.17	28.88
	0.1	18.01	5000	14.43	17.18	17.86	17.89	18.56	22.38	14.57	17.18	17.87	17.90	18.56	22.77
	0.2	18.00		11.63	16.22	17.53	17.65	18.92	27.71	10.68	16.21	17.46	17.60	18.88	26.66
	0.5	31.32		10.55	21.20	25.48	26.53	30.73	71.57	9.68	21.39	25.58	26.59	30.77	68.18

Case of Fully Noise-Multiplied Data. Tables 5 and 6 report the results of the numerical study for evaluating the disclosure risk in the case of full noise multiplication. In Table 5, $f(y|\boldsymbol{\theta})$ is the lognormal density as in Subsection 4.2 with $\mu = 0$, $\sigma^2 = 1$, and the table shows, for a few selected y_i values, the corresponding z_i values, and a summary of the distribution of the associated values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. The z-values are shown for the cases of the uniform noise density (5) with $\epsilon = 0.1$, 0.2, and 0.5; and the minimum, 1st quartile, median, mean, 3rd quartile, and maximum of the associated values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ are displayed for two cases: $m = 5$ and $m = 5{,}000$. While such a large value as $m = 5{,}000$ may not be used in practice, we consider this large m in order to obtain an accurate picture of the distribution of released values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. Of course for the case $m = 5$, the minimum, 1st quartile, median, 3rd quartile, and maximum are simply the ordered values of $y_i^{*(1)}, \ldots, y_i^{*(5)}$, respectively. Furthermore, results for both the Type A and Type B imputation methods for y^*-values are shown in the table. Table 6 reports similar results for lognormal except that instead of uniform, we use the customized noise distribution for lognormal data as defined in Subsection 4.2, with variances matching those of the Uniform$(1 - \epsilon, 1 + \epsilon)$ density with $\epsilon = 0.1$, 0.2, and 0.5. The following is a summary of the results of Tables 5 and 6.

1. As the variation in the noise distribution $h(r)$ increases (i.e., as ϵ increases), the dispersion in $y_i^{*(1)}, \ldots, y_i^{*(m)}$ also increases. Therefore, as one would expect, the amount of privacy protection provided by this method increases with the variance of the noise-generating distribution.
2. Generally, even for large m, one does not recover the original y_i by averaging or computing the median of the imputed copies $y_i^{*(1)}, \ldots, y_i^{*(m)}$. Usually we find that the noise-multiplied observation z_i is contained between the 1st and 3rd quartiles of $y_i^{*(1)}, \ldots, y_i^{*(m)}$, but interestingly, the y_i value may not be contained between these quartiles. In fact, when ϵ is small, the distribution of the $y_i^{*(1)}, \ldots, y_i^{*(m)}$ values tends to be concentrated around z_i and not y_i. However, when the noise multiplication results in a large perturbation as in the bottom row of Table 6 where $y_i = 18.21$ and $z_i = 31.32$, then we find that the distribution of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ is shifted downward toward y_i, yet still the original value of $y_i = 18.21$ is not contained between the 1st and 3rd quartiles of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. This finding gives some indication that the method does provide some correction of an extreme z_i value, while at the same time does not disclose the original y_i value.
3. Comparing the results of the Type A and Type B imputation procedures, we find them to be quite similar.
4. The results for the uniform and customized noise distributions are similar, although the uniform noise does tend to give a slightly larger interquartile range of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ than the customized noise, thus providing perhaps slightly more privacy protection.

Case of Mixture Data. Table 7 reports the results of the numerical study for evaluating the disclosure risk in the case of mixture data. The population density $f(y|\boldsymbol{\theta})$ is again the lognormal density as in Subsection 4.2 with $\mu = 0$, $\sigma^2 = 1$, the top-coding threshold is $C = 3.60$ which is the 0.90 quantile of the population density (rounded to two decimal places), and the table shows, for three particular y_i values, the corresponding x_i value, and

Table 7. Illustration of y, x, and y-values for mixture LN ($\mu = 0$, $\sigma^2 = 1$) data with C = .90 quantile, n = 100 and uniform noise*

					y*-value											
					Type B method						Type A method					
y-value	ϵ	*x*-value	case	*m*	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
5.71	0.1	6.01	(i)	5	5.73	5.74	6.03	6.10	6.49	6.52	5.53	5.71	5.83	5.84	5.89	6.23
	0.2	6.38			5.32	5.47	5.55	5.63	5.79	6.02	5.40	5.50	6.09	6.17	6.34	7.52
	0.5	5.09			3.77	4.13	4.15	4.58	4.96	5.90	3.78	3.78	6.02	5.95	6.03	10.11
	0.1	6.01		5000	5.46	5.70	5.97	6.00	6.30	6.68	5.46	5.69	5.97	6.00	6.29	6.68
	0.2	6.38			5.32	5.70	6.23	6.36	6.95	7.98	5.32	5.70	6.19	6.33	6.87	7.98
	0.5	5.09			3.60	4.02	4.63	5.10	5.78	10.12	3.60	4.02	4.68	5.13	5.79	10.16
	0.1	6.01	(ii)	5	5.61	6.02	6.12	6.11	6.38	6.39	5.57	5.71	5.74	5.92	6.04	6.56
	0.2	6.38			5.57	5.81	6.07	6.14	6.42	6.84	5.39	5.50	5.64	6.05	6.32	7.38
	0.5	5.09			3.62	3.99	6.51	5.54	6.60	6.96	4.73	5.15	5.21	5.84	6.85	7.28
	0.1	6.01		5000	5.46	5.70	5.98	6.01	6.30	6.68	5.46	5.70	5.97	6.00	6.29	6.68
	0.2	6.38			5.32	5.70	6.20	6.34	6.88	7.98	5.32	5.70	6.21	6.34	6.91	7.98
	0.5	5.09			3.60	4.04	4.69	5.14	5.82	10.16	3.60	4.01	4.65	5.15	5.87	10.18
3.75	0.1	3.40	(i)	5	3.65	3.68	3.71	3.69	3.71	3.73	3.61	3.61	3.65	3.66	3.70	3.74
	0.2	3.38			3.61	3.80	3.80	3.88	4.02	4.15	3.63	3.63	3.78	3.85	4.00	4.20
	0.5	1.94			3.60	3.61	3.64	3.65	3.65	3.77	3.63	3.66	3.73	3.71	3.73	3.80
	0.1	3.40		5000	3.60	3.64	3.69	3.69	3.73	3.77	3.60	3.64	3.68	3.69	3.73	3.77
	0.2	3.38			3.60	3.73	3.88	3.89	4.04	4.22	3.60	3.73	3.87	3.89	4.04	4.22
	0.5	1.94			3.60	3.66	3.73	3.73	3.80	3.87	3.60	3.67	3.73	3.73	3.80	3.87
	0.1	3.40	(ii)	5	3.40	3.40	3.40	3.46	3.40	3.71	3.40	3.40	3.40	3.47	3.40	3.77
	0.2	3.38			3.38	3.38	3.38	3.49	3.38	3.92	3.38	3.38	3.38	3.38	3.38	3.38
	0.5	1.94			1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94
	0.1	3.40		5000	3.40	3.40	3.40	3.45	3.40	3.77	3.40	3.40	3.40	3.45	3.40	3.77
	0.2	3.38			3.38	3.38	3.38	3.50	3.38	4.22	3.38	3.38	3.38	3.50	3.38	4.22
	0.5	1.94			1.94	1.94	1.94	1.98	1.94	3.87	1.94	1.94	1.94	1.98	1.94	3.87
3.56	0.1	3.56	(i)	5	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.2	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56

Table 7. Continued

					y^*-value											
					Type B method						Type A method					
y-value	ϵ	x-value	case	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
	0.5	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.1	3.56		5000	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.2	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.5	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.1	3.56	(ii)	5	3.56	3.56	3.56	3.58	3.56	3.65	3.56	3.56	3.56	3.65	3.72	3.84
	0.2	3.56			3.56	3.56	3.56	3.64	3.56	3.95	3.56	3.56	3.56	3.56	3.56	3.56
	0.5	3.56			3.56	3.56	4.13	4.27	4.94	5.15	3.56	3.56	3.56	4.62	5.40	7.04
	0.1	3.56		5000	3.56	3.56	3.56	3.62	3.65	3.96	3.56	3.56	3.56	3.62	3.65	3.96
	0.2	3.56			3.56	3.56	3.56	3.68	3.69	4.45	3.56	3.56	3.56	3.69	3.70	4.45
	0.5	3.56			3.56	3.56	3.56	3.85	3.62	7.12	3.56	3.56	3.56	3.85	3.60	7.12

distribution of the associated values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. In this table, the x-values are shown for the cases of the uniform noise density (5) with $\epsilon = 0.1$, 0.2, and 0.5; and the minimum, 1st quartile, median, mean, 3rd quartile, and maximum of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ are displayed for the cases $m = 5$ and $m = 5{,}000$. Results are shown for both cases (i) and (ii) of Section 3 and for both the Type A and Type B imputation methods. Most of the findings here are similar to those of the case of full noise multiplication. Below is a summary of findings from Table 7 which highlights the similarities and differences in privacy protection between cases (i) and (ii) of Section 3.

1. The first part of the table shows results when the y-value is $y_i = 5.71$, which is, of course, greater than the top-coding threshold $C = 3.60$. It happens here that each of the displayed noise-multiplied values is also larger than C. Therefore, based on each of the x-values shown, we know with certainty that $\Delta_i = 0$ (that is, the conditional probability (17) equals 0), and hence the case (ii) method will always impute this particular Δ_i value correctly. Here, the properties of the replications $y_i^{*(1)}, \ldots, y_i^{*(m)}$ for both cases (i) and (ii) are similar to each other and similar to those noted for the full noise multiplication case (replications not centered at y_i, dispersion increasing with ϵ, etc.). Note that the imputations under case (i) may be of slightly higher quality, since the estimate of $\boldsymbol{\theta}$ (either posterior draw or MLE) needed to generate the imputations may be of higher quality when based on case (i) data.
2. The second part of the table shows results when $y_i = 3.75$, which is again greater than $C = 3.60$, but each of the displayed x-values happen to fall in the interval $((1 - \epsilon)C, C)$. When the x-value falls in this interval, the indicator Δ_i cannot be determined from x_i with certainty (that is, the conditional probability (17) does not equal 0 or 1). Therefore, the case (ii) method will sometimes (with a probability governed by (17)), impute Δ_i by the value one, and hence release the noise-multiplied data point as the y^*-value. Here it is interesting to look at the $\epsilon = 0.50$ case where $x_i = 1.94$ because in this case we see a large difference between the results in cases (i) and (ii). In case (i) we use the information that $\Delta_i = 0$ when generating imputations, and hence the released y^*-values are more similar to the original y-value. In case (ii) we do not have this knowledge about the true value of Δ_i. Since the noise-multiplied observation is fairly small, Δ_i is often imputed as 1 in case (ii). Therefore, under case (ii), the noise-multiplied data point is often directly released in the replications $y_i^{*(1)}, \ldots, y_i^{*(m)}$ and a user who sees these data would not immediately know if the value repeated several times in the released $y_i^{*(1)}, \ldots, y_i^{*(m)}$ was the original y_i or its noise perturbed version.
3. The third part of the table shows results with $y_i = 3.56$. In this case, the y-value is less than the top-coding threshold $C = 3.60$, while each of the x-values happen to fall in the interval $((1 - \epsilon)C, C)$. Therefore, the value of Δ_i cannot be determined with certainty from x_i (the conditional probability (17) does not equal 0 or 1). Thus, the case (ii) method sometimes imputes Δ_i by 0, and in these cases the released y^* will not be equal to the original y-value, since it will be divided by a random draw from (14). In this situation, unlike the situation described in item (2) directly above, the value repeated several times in the replications $y_i^{*(1)}, \ldots, y_i^{*(m)}$ for case (ii) is the original observation, not its noise-perturbed version. In this case, the case

(i) method, which uses knowledge of $\Delta_i = 1$, always sets the released $y*$-value to the true value of y.

6.2. Comparison with Synthetic Data

The methodology developed in this article is designed to enable statistical agencies to release privacy-protected data that can be readily analyzed by data users. The methods of (partially) synthetic data developed in Reiter (2003) are designed for the same purpose, and hence a comparison of our methodology with that of Reiter (2003) is in order. A general criticism of noise multiplication is that a proper statistical analysis of noise-multiplied data is complicated for data users. The results of this article show how to remedy this criticism by making the analysis as simple (for the data user) as the analysis of synthetic data (we showed that Rubin's (1987) combining rules can be used here, and these rules are only slightly different from those of Reiter (2003)). Since the methodology of this article gives very similar results to the full likelihood-based analysis of noise-multiplied data developed in Klein et al. (2013), we believe that the pertinent comparison is that of synthetic data versus noise multiplication, assuming a valid data analysis is performed in both cases. Such a comparison, in terms of data quality, is precisely the topic of Klein et al. (2013). We note that synthetic data certainly has benefits, as it has been thoroughly studied in recent years, and successfully applied to complex multivariate data sets. At the same time, the methodology of this article can be extended to multivariate data as outlined in the subsection below. An advantage of noise multiplication over synthetic data is that noise multiplication allows the statistical agency to precisely control the quality of the released data, and also the level of privacy protection, through the choice of $h(r)$. For instance, when $h(r)$ is the uniform density (5), the extensive numerical results of Klein et al. (2013) show, for some univariate parametric models, precisely how to select ϵ so that the quality of inferences are equivalent to, less than, or greater than, the quality of inferences derived under synthetic data. Indeed, the ability to choose $h(r)$ provides the statistical agency with a very fine level of control over the data quality and privacy protection, and such an explicit tuning mechanism is not present in standard synthetic data methodology. Further privacy guarantees under noise multiplication can be made, for instance, by taking $h(r)$ to be a density such as

$$h(r) = \frac{1}{2(\epsilon - \xi)}, \text{ if } r \in (1 - \epsilon, 1 - \xi) \cup (1 + \xi, 1 + \epsilon), \tag{27}$$

where $0 < \xi < \epsilon < 1$. Notice that the noise density (27) implies that the noise multiplier r is always a distance ξ away from 1, and hence we are guaranteed that the relative distance between the original observation y and noise-multiplied observation z is $|(z - y)/y| > \xi$.

6.3. Extensions for Multivariate Data

So far in this article we assumed that the original data, $y_1, \ldots, y_n$, consist of a set of n independent random variables whose support is a subset $\mathbb{R}$. In this section, we outline an extension of our methodology to the case of multivariate and fully noise-multiplied data. In the multivariate case, we assume that the original data consist of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, a set of n independent $k \times 1$ dimensional random vectors. Thus we suppose that

$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim iid \sim f(\boldsymbol{y}|\boldsymbol{\theta})$, independent of $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n \sim iid \sim h(\boldsymbol{r})$ where $f(\,y|\boldsymbol{\theta})$ and $h(r)$ are densities of continuous probability distributions whose support is a subset of $\mathbb{R}^k$. As before, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is an unknown $p \times 1$ parameter vector, and now $h(\boldsymbol{r})$ is a known density such that $h(\boldsymbol{r}) = 0$ if any component of the vector r is less than zero. Writing $\boldsymbol{y}_i = (\,y_{i1}, \ldots, y_{ik})$ and $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{ik})$, the fully noise-multiplied data are now defined by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik}) = (\,y_{i1}r_{i1}, , \ldots, y_{ik}r_{ik})$, $i = 1, , \ldots, n$.

The joint density of $(\boldsymbol{z}_i, \boldsymbol{r}_i)$ is

$$g(\boldsymbol{z}_i, \boldsymbol{r}_i|\boldsymbol{\theta}) = f\left(\frac{z_{i1}}{r_{i1}}, \ldots, \frac{z_{ik}}{r_{ik}}\middle|\boldsymbol{\theta}\right) h(r_{i1}, \ldots, r_{ik}) \left[\prod_{l=1}^{k} r_{il}^{-1}\right],$$

the marginal density of $\boldsymbol{z}_i$ is

$$g(\boldsymbol{z}_i|\theta) = \int_0^\infty \ldots \int_0^\infty f\left(\frac{z_{i1}}{\omega_{i1}}, \ldots, \frac{z_{ik}}{\omega_{ik}}\middle|\boldsymbol{\theta}\right) h(\omega_{i1}, \ldots, \omega_{ik}) \left[\prod_{l=1}^{k} \omega_{il}^{-1}\right] d\omega_{i1} \ldots d\omega_{ik},$$

and hence the conditional density of $\boldsymbol{r}_i$ given $\boldsymbol{z}_i$ is

$$h(\boldsymbol{r}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) = \frac{f((z_{i1}/r_{i1}), \ldots, (z_{ik}/r_{ik})|\boldsymbol{\theta}) h(r_{i1}, \ldots, r_{ik}) \left[\prod_{l=1}^{k} r_{il}^{-1}\right]}{\int_0^\infty \ldots \int_0^\infty f((z_{i1}/\omega_{i1}), \ldots, (z_{ik}/\omega_{ik})|\boldsymbol{\theta}) h(\omega_{i1}, \ldots, \omega_{ik}) \left[\prod_{l=1}^{k} \omega_{il}^{-1}\right] d\omega_{i1} \ldots d\omega_{ik}}. \tag{28}$$

The complete, observed, and missing data are defined, respectively, as

$$\boldsymbol{x}_c = \{(\boldsymbol{z}_1, \boldsymbol{r}_1), \ldots, (\boldsymbol{z}_n, \boldsymbol{r}_n)\}, \quad \boldsymbol{x}_{\text{obs}} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}, \quad \boldsymbol{x}_{\text{mis}} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}.$$

The noise vectors $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ are imputed m times to obtain

$$\begin{aligned} \boldsymbol{x}_c^{*(j)} &= \left\{\left(\boldsymbol{z}_1, \boldsymbol{r}_1^{*(j)}\right), \ldots, \left(\boldsymbol{z}_n, \boldsymbol{r}_n^{*(j)}\right)\right\} \\ &= \left\{(z_{11}, \ldots, z_{1k}), \left(r_{11}^{*(j)}, \ldots, r_{1k}^{*(j)}\right), \ldots, (z_{n1}, \ldots, z_{nk}), \left(r_{n1}^{*(j)}, \ldots, r_{nk}^{*(j)}\right)\right\}, j = 1, \ldots, m, \end{aligned}$$

and the privacy-protected data are obtained as

$$\begin{aligned} \boldsymbol{y}^{*(j)} &= \left\{\boldsymbol{y}_1^{*(j)}, \ldots, \boldsymbol{y}_n^{*(j)}\right\} = \left\{\left(y_{11}^{*(j)}, \ldots, y_{1k}^{*(j)}\right), \ldots, \left(y_{n1}^{*(j)}, \ldots, y_{nk}^{*(j)}\right)\right\} \\ &= \left\{\left(\frac{z_{11}}{r_{11}^{*(j)}}, \ldots, \frac{z_{1k}}{r_{1k}^{*(j)}}\right), \ldots, \left(\frac{z_{n1}}{r_{n1}^{*(j)}}, \ldots, \frac{z_{nk}}{r_{nk}^{*(j)}}\right)\right\}, j = 1, \ldots, k. \end{aligned} \tag{29}$$

The methods of Subsection 2.2 can be used to impute the noise vectors, and the methods of Subsection 2.3 can be used to analyze the privacy-protected data given in (29). Conceptually, the methods of Subsections 2.2 and 2.3 can be readily applied to multivariate data. For instance, a data user wishing to draw inference about the correlation between y_{i1} and y_{i2} would set $Q(\boldsymbol{\theta}) = \text{Corr}(\,y_{i1}, y_{i2}|\boldsymbol{\theta})$, and apply methods of Subsection 2.3. For the statistical agency generating the imputations, there is perhaps one extension needed in applying the methods of Subsection 2.2, because when generating the imputations (either Type A or Type B), instead of sampling from the univariate

conditional density (4), we must now sample from the k-dimensional multivariate conditional density (28). In the univariate case we used Proposition 1 to extract samples from (4) when one takes the noise-generating density to be (5). In the multivariate case, a generalization of Proposition 1 can be used to sample random vectors from (28), when the noise-generating distribution is the following k-dimensional uniform density (which is a straightforward generalization of (5)):

$$h(r_1,\ldots,r_k)=\frac{1}{2^k\prod_{l=1}^{k}\epsilon_l},\ \text{for}\ (r_1,\ldots,r_k)\in[1-\epsilon_1,1+\epsilon_1]\times\cdots\times[1-\epsilon_k,1+\epsilon_k], \quad (30)$$

where $0<\epsilon_1,\ldots,\epsilon_k<1$. The generalization of Proposition 1 is stated below as Proposition 2; the proof is similar to that of Proposition 1 and hence is omitted.

Proposition 2 *Suppose that $f(\mathbf{y}|\boldsymbol{\theta})$ is a continuous probability density function of a k-dimensional distribution, and let us write $f(\mathbf{y}|\boldsymbol{\theta})=c(\boldsymbol{\theta})q(\mathbf{y}|\boldsymbol{\theta})$ where $c(\boldsymbol{\theta})>0$ is a normalizing constant. Let $M\equiv M(\boldsymbol{\theta},\epsilon_1,\ldots,\epsilon_k,\mathbf{z})$ be such that*

$$q\left(\frac{z_1}{r_1},\ldots,\frac{z_k}{r_k}\middle|\boldsymbol{\theta}\right)\le M\ \text{for all}\ (r_1,\ldots,r_k)$$

$$\in[1-\epsilon_1,1+\epsilon_1]\times\cdots\times[1-\epsilon_k,1+\epsilon_k].$$

Then the following algorithm produces a random vector $(R_1,\ldots,R_k)$ having the density

$$h_U(r_1,\ldots,r_k|z_1,\ldots,z_k,\boldsymbol{\theta})$$

$$=\frac{q((z_1/r_1),\ldots,(z_k/r_k)|\boldsymbol{\theta})\left[\prod_{l=1}^{k}r_l^{-1}\right]}{\int_{1-\epsilon_k}^{1+\epsilon_k}\cdots\int_{1+\epsilon_1}^{1+\epsilon_1}q((z_1/\omega_1),\ldots,(z_k/\omega_k)|\boldsymbol{\theta})\left[\prod_{l=1}^{k}\omega_l^{-1}\right]d\omega_1\ldots d\omega_k},$$

$$\text{for}\ (r_1,\ldots,r_k)\in[1-\epsilon_1,1+\epsilon_1]\times\cdots\times[1-\epsilon_k,1+\epsilon_k].$$

I. *Generate $U, V_1,\ldots,V_k$ as independent Uniform $(0,\ 1)$ and let $W_l=(1+\epsilon_l)^{V_l}/(1-\epsilon_l)^{V_l-1}$ for $l=1,\ldots,k$.*
II. *Accept $(R_1,\ldots,R_k)=(W_1,\ldots,W_k)$ if $U\le M^{-1}q((z_1/W_1),\ldots,(z_k/W_k)|\boldsymbol{\theta})$, otherwise reject the vector $(W_1,\ldots,W_k)$ and return to step (I).*

The expected number of iterations of steps (I) and (II) required to obtain $(R_1,\ldots,R_k)$ is

$$\frac{M\prod_{l=1}^{k}\log\left[\dfrac{1+\epsilon_l}{1-\epsilon_l}\right]}{\int_{1-\epsilon_k}^{1+\epsilon_k}\cdots\int_{1-\epsilon_1}^{1+\epsilon_1}q((z_1/\omega_1),\ldots,(z_k/\omega_k)|\boldsymbol{\theta})\left[\prod_{l=1}^{k}\omega_l^{-1}\right]d\omega_1\ldots d\omega_k}.$$

Remark 5. In this section we briefly outlined the multivariate extension for the case of fully noise-multiplied data; that is, where $\mathbf{y}_1,\ldots,\mathbf{y}_n\sim iid\sim\mathbf{Y}$ and each component of $\mathbf{Y}$ requires protection from disclosure. We note that the methodology outlined in this section allows one to use different levels of privacy protection for each component of $\mathbf{Y}$ through the choice of $\epsilon_1,\ldots,\epsilon_k$ in (30). Other scenarios are certainly possible; for instance, it

may be that $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ where the variable Y_1 must always be protected, Y_2 requires protection only if it exceeds a fixed threshold $C > 0$, and Y_3 does not require any protection. We intend to address such issues in a future communication.

7. Concluding Remarks

There are perhaps two rigorous ways of producing privacy-protected data: multiple imputation and noise perturbation. Klein et al. (2013) show that the likelihood-based method of analysis of noise-multiplied data can yield accurate inferences under several standard parametric models and compare favorably with the standard multiple imputation-based analysis methods of Reiter (2003) and An and Little (2007). Since the likelihood of the noise-multiplied data is often complex, one wonders if an alternative simpler and fairly accurate data analysis method can be developed based on such kind of privacy-protected data. With precisely this objective in mind, we have shown in this article that a proper application of multiple imputation leads to such an analysis. In implementing the proposed method under a standard parametric model $f(y|\boldsymbol{\theta})$, the most complex part is generally simulation from the conditional densities (4) or (14), and this part would be the responsibility of the data producer, not the data user. We have provided Proposition 1 which gives an exact algorithm to sample from (4) and (14) for general continuous $f(y|\boldsymbol{\theta})$, when $h(r)$ is the uniform distribution (5). Moreover, we have seen that in the lognormal case under full noise multiplication, if one uses the customized noise distribution, then the conditional density (4) takes a standard form from which sampling is straightforward. Simulation results based on sample sizes of 100 and 500 indicate that the multiple imputation-based analysis, as developed in this article, generally results in only a slight loss of accuracy in comparison to the formal likelihood-based analysis. Our simulation results also indicate that both the Rubin (1987) and Wang and Robins (1998) combining rules exhibit adequate performance in the selected sample settings. We have also reported some additional numerical results for evaluating the amount of privacy protection offered by the method. These results showed that one does not recover the original observation simply by averaging the multiply imputed copies of the original value.

In conclusion, we observe that, from a data user's perspective, our method does require a knowledge of the underlying parametric model of the original y-data so that efficient model-based estimates can be used to analyze the reconstructed y^*-values. In this article we assumed that the model used by the agency to multiply impute the original data, namely $f(y|\boldsymbol{\theta})$, is the same model adopted by the data user to analyze the released data. However, in practice this may not be the case (see Meng 1994 and Robins and Wang 2000 for a discussion of possible consequences of model misspecification). In any event, modeling by data users, if necessary, will be based on the released y^*-values, and *not* on the noise-multiplied z-values. It is expected that the sampling behaviors of y-values and y^*-values would be similar. This is in the same spirit as in the case of synthetic data usage where a data user will either be informed about the original model or try to build up a reasonable model based on the released synthetic data. We should also point out that in practice, most data sets have a complex multivariate structure. We briefly outlined how our methodology can be extended to multivariate data. In a future communication we intend to investigate the robustness of the multiple imputation-based analysis to

discrepancies between the imputation and analysis models, and to further develop the multivariate extensions of the proposed method.

Appendix A

Proof of Proposition 1. This is a rejection sampling algorithm where the target density $h_U(r|z, \boldsymbol{\theta})$ is proportional to $s_{\text{target}}(r) = q((z/r)|\boldsymbol{\theta})r^{-1}$, $1 - \epsilon \le r \le \gamma$, and the instrumental density is $s_{\text{instr}}(r) = r^{-1}/(\log(\gamma) - \log(1-\epsilon))$, $1 - \epsilon \le r \le \gamma$. To fill in the details, first note that since $f(y|\boldsymbol{\theta})$ is continuous in y, it follows that $q((z/r)|\boldsymbol{\theta})$ is continuous as a function of r, on the interval $[1 - \epsilon, \gamma]$, and thus the bounding constant M exists. Then we see that

$$\frac{s_{\text{target}}(r)}{s_{\text{instr}}(r)} = [\log(\gamma) - \log(1-\epsilon)]q\left(\frac{z}{r}|\boldsymbol{\theta}\right) \le [\log(\gamma) - \log(1-\epsilon)]M, \tag{31}$$

for all $r \in [1 - \epsilon, \gamma]$. Note that the cumulative distribution function corresponding to $s_{\text{instr}}(r)$ is $\mathbb{S}_{\text{instr}}(r) = (\log(r) - \log(1-\epsilon))/(\log(\gamma) - \log(1-\epsilon))$, $1 - \epsilon \le r \le \gamma$, and the inverse of this distribution function is $\mathbb{S}^{-1}_{\text{instr}}(u) = \gamma^u/(1-\epsilon)^{u-1}$, $0 \le u \le 1$. Thus, by the inversion method (Devroye 1986), step (I) is equivalent to independently drawing $U \sim \text{Uniform}(0,1)$ and W from the density $s_{\text{instr}}(r)$. Since $M^{-1}s_{\text{target}}(W)/([\log(\gamma) - \log(1-\epsilon)]s_{\text{instr}}(W)) = q(z/w|\boldsymbol{\theta})/M$, step (II) is equivalent to accepting W *if* $U \le M^{-1}s_{\text{target}}(W)/([\log(\gamma) - \log(1-\epsilon)]s_{\text{instr}}(W))$, which is the usual rejection step based on the bound in (31). Finally, we use the well-known fact that the expected number of iterations of the rejection algorithm is equal to the bounding constant in (31) times the normalizing constant for $s_{\text{target}}(r)$, i.e., $[\log(\gamma) - \log(1-\epsilon)]M/[\int_{1-\epsilon}^{\gamma} q((z/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega]$.

Appendix B

Here we provide proofs of the posterior propriety of $\boldsymbol{\theta}$, given the fully noise-multiplied data $\boldsymbol{z}$, for normal and lognormal distributions.

Normal distribution. Here $g(z|\boldsymbol{\theta}) \propto (1/\sigma)\int \exp[-((z/r) - \mu)^2/(2\sigma^2)](h(r)/r)dr$. Writing down the joint *pdf* of $z_1, \ldots, z_n$, it is obvious that upon integrating out μ with respect to (wrt) the Lebesgue measure and σ wrt the flat or noninformative prior, we end up with the expression $U(\boldsymbol{z})$ given by

$$U(\boldsymbol{z}) = \int \cdots \int \left[\sum_{i=1}^{n} \frac{z_i^2}{r_i^2} - \frac{\left(\sum_{i=1}^{n}(z_i/r_i)\right)^2}{n}\right]^{-n-\delta} \frac{h(r_1)\ldots h(r_n)}{r_1 \ldots r_n} dr_1 \ldots dr_n$$

where $\delta \ge 0$. To prove that $U(\boldsymbol{z})$ is finite for any given $\boldsymbol{z}$, note that

$$\left[\sum_{i=1}^{n} \frac{z_i^2}{r_i^2} - \frac{\sum_{i=1}^{n}(z_i/r_i)^2}{n}\right] = \frac{1}{2}\sum_{i,\,j=1}^{n}\left(\frac{z_i}{r_i} - \frac{z_j}{r_j}\right)^2 \ge \frac{1}{2}\left[\frac{z_1}{r_1} - \frac{z_2}{r_2}\right]^2$$

for any pair $(z_1, z_2; r_1, r_2)$, Assume without any loss of generality that $z_1 > z_2$, and note that

$[(z_1/r_1) - (z_2/r_2)]^2 = [(z_1/z_2) - (r_1/r_2)]^2 \times z_2^2 r_1^{-2}$. Then under the *condition*

$$\int_r \frac{h(r)}{r} dr = K_1 < \infty, \quad \int_{r_1 \le r_2} r_1^{2(n+\delta)-1} r_2^{-1} h(r_1) h(r_2) dr_1 dr_2 = K_2 < \infty, \tag{32}$$

$U(z)$ is bounded above by

$$U(z) \le 2^{n+\delta} K_1^{n-2} \left[\frac{z_1}{z_2} - 1\right]^{-2(n+\delta)} \left[\int_{r_1 \le r_2} r_1^{2(n+\delta)-1} r_2^{-1} h(r_1) h(r_2) dr_1 dr_2\right] < \infty.$$

In particular, when $R \sim$ Uniform$(1 - \epsilon, 1 + \epsilon)$, the above condition is trivially satisfied!

Lognormal distribution. Here $g(z|\boldsymbol{\theta}) \propto (1/z\sigma) \int \exp[-(\log(z/r) - \mu)^2/(2\sigma^2)] h(r) dr$. Writing down the joint density of $z_1, \ldots, z_n$, and putting $u = \log(z/r)$, it is obvious that upon integrating out μ wrt the Lebesgue measure and σ wrt the flat or noninformative prior, we end up with the expression $U(z)$ given by

$$U(z) = \int_{r_1} \cdots \int_{r_n} \left[\sum_{i=1}^n (u_i - \bar{u})^2\right]^{-2(n+\delta)} h(r_1) \cdots h(r_n) dr_1 \cdots dr_n$$

where $\delta \ge 0$. To prove that $U(z)$ is finite for any given z, note as in the normal case that when $z_1 > z_2$ (without any loss of generality),

$$\left[\sum_{i=1}^n (u_i - \bar{u})^2\right] = \frac{1}{2} \sum_{i,j=1}^n (u_i - u_j)^2 \ge \frac{1}{2}(u_1 - u_2)^2 = \frac{1}{2}\left[\log\left(\frac{z_1}{z_2}\right) - \log\left(\frac{r_1}{r_2}\right)\right]^2$$

$$\ge \frac{1}{2}\left[\log\left(\frac{z_1}{z_2}\right)\right]^2$$

for $r_1 < r_2$. Hence $U(z)$ is always finite, since $\int_{r_1 < r_2} h(r_1) h(r_2) dr_1 dr_2 < \infty$.

Appendix C

Here we provide proofs of the posterior propriety of $\boldsymbol{\theta}$, given the mixture data, for normal and lognormal distributions. We consider two cases depending on the nature of mixture data that will be released.

Case (i): Nature of data $\{(x_1, \Delta_1), \cdots, (x_n, \Delta_n)\}$.

Normal distribution. Given the data $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$, let $I_1 = \{i : \Delta_i = 1\}$ and $I_0 = \{i : \Delta_i = 0\}$. Then the normal likelihood $L(\boldsymbol{\theta}|\text{data})$, apart from a constant, can be expressed as

$$L(\boldsymbol{\theta}|\text{data}) \propto \sigma^{-n} \left[\exp\left(-\sum_{i \in I_1} \frac{(x_i - \mu)^2}{2\sigma^2}\right)\right]$$

$$\times \left[\prod_{i \in I_0} \int_0^{(x_i/c)} \exp\left(-\frac{((x_i/r_i) - \mu)^2}{2\sigma^2}\right) \frac{h(r_i)}{r_i} I(x_i > 0) dr_i\right].$$

It is then obvious that upon integrating out μ wrt the Lebesgue measure and σ wrt the flat or noninformative prior, we end up with the expression U (data) given by

$$U(\text{data}) = \prod_{i\in I_0}\int_0^{(x_i/c)} I(x_i > 0)\left[\sum_{i\in I_1} x_i^2 + \sum_{i\in I_0}\frac{x_i^2}{r_i^2} - \frac{\left(\sum_{i\in I_1} x_i + \sum_{i\in I_0}(x_i/r_i)\right)^2}{n}\right]^{-n-\delta} \frac{h(r_i)}{r_i}dr_i.$$

Writing $v_i = x_i/r_i$ for $i \in I_0$, the expression $\Psi(\text{data}) = \sum_{i\in I_1} x_i^2 + \sum_{i\in I_0} x_i^2/r_i^2 - \left(\sum_{i\in I_1} x_i + \sum_{i\in I_0}(x_i/r_i)\right)^2/n$ is readily simplified as $[S_1^2 + S_0^2 + rs(\bar{x}_1 - \bar{x}_0)^2](r+s)^{-1}$ where r and s are the cardinalities of I_1 and I_0, respectively, and $\left(\bar{x}_1, S_1^2\right)$ and $\left(\bar{x}_0, S_0^2\right)$ are the sample means and variances of the data in the two subgroups I_1 and I_0, respectively.

When I_1 is nonempty, an obvious lower bound of Ψ(data) is $S_1^2/(r+s)$, and if I_1 is empty, $\Psi(\text{data}) = S_0^2/n$. In the first case, U(data) is finite whenever $\int_0^{(x_i/c)}(h(r)/r)dr < \infty$ for $i \in I_0$. In the second case, we proceed as in the fully noise-perturbed case for normal and conclude that U (data) is finite under the conditions stated in (32) except that the bounds of r_i in the integrals are replaced by x_i/C. In particular, for uniform noise distribution, the conditions trivially hold.

Lognormal distribution. Proceeding as in the normal case with $u = \log(x/r)$, and breaking up the sum in the exponent into two parts corresponding to I_1 and I_0, we get the finiteness of corresponding U(data) under noninformative priors of μ and σ when the noise distribution is uniform.

Case (ii): Nature of data $(x_1, \ldots, x_n)$.

Normal distribution. Upon carefully examining the joint *pdf* of the data $\boldsymbol{x}$, given by (18), let us split the entire data into three mutually exclusive sets:

$$I_1 = \{i : x_i < 0\}, \quad I_2 = \{i : 0 < x_i < C\}, \quad I_3 = \{i : x_i > C\}.$$

It is now clear from standard computations under the normal distribution that whenever I_1 is non-empty, the posterior of (μ, σ) under a flat or noninformative prior of (μ, σ) will be proper. This is because the rest of the joint *pdf* arising out of I_2 and I_3 can be bounded under a uniform noise distribution or even under a general $h(.)$ under very mild conditions, and the retained part under I_1 will lead to propriety of the posterior. Likewise, if I_1 is empty but I_3 is non-empty, we can easily bound the terms in I_2, and proceed as in the fully noise-perturbed case for data in I_3 and show that the posterior is proper. Lastly, assume that the entire data fall in I_2, resulting in the joint *pdf* $L(\boldsymbol{\theta}|\text{data} \in I_2)$ as a product of terms of the type

$$f(x_i|\boldsymbol{\theta}) + \int_0^{(x_i/c)} f\left(\frac{x_i}{r_i}|\boldsymbol{\theta}\right)\frac{h(r_i)}{r_i}dr_i < \int_0^{(x_i/c)}\left[f(x_i|\boldsymbol{\theta})\frac{C}{x_{(1)}} + f\left(\frac{x_i}{r_i}|\boldsymbol{\theta}\right)\frac{h(r_i)}{r_i}\right]dr_i$$

where $x_{(1)} = \min(x_i)$. Let us now carefully check the product of the above integrands under the normal distribution, which will be first integrated wrt (μ, σ) under a flat or noninformative prior, and later wrt the noise variables which we take to be *iid* uniform. Obviously this product will be a sum of mixed terms of the following two types which are

relevant to check the propriety of the resultant posterior:

$$\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i\in J_1}(x_i-\mu)^2+\sum_{i\in J_2}\left(\frac{x_i}{r_i}-\mu\right)^2\right)\right]$$

where J_1 and J_2 form a partition of $\{1,\ldots,n\}$. It is now immediate that the terms of the first type (standard normal theory without any noise perturbation) will lead to a proper posterior of (μ,σ). Likewise, from our previous computations under the fully noise-perturbed case, it follows that the terms of the second type will also lead to propriety of the posterior of μ and σ under a uniform noise distribution.

Lognormal distribution. Proceeding as in the normal case above by replacing x/r by $u = \log(x/r)$, we get the posterior propriety of μ and σ under flat or noninformative priors when the noise is uniform. We omit the details.

8. References

An, D. and Little, R.J.A. (2007). Multiple Imputation: An Alternative to Top Coding for Statistical Disclosure Control. Journal of Royal Statistical Society, Series A, 170, 923–940.

Devroye, L. (1986). Non-Uniform Random Variate Generation: Springer.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). Bayesian Data Analysis, (second edition). Chapman & Hall/CRC.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 303–308.

Kim, J.J. and Winkler, W.E. (1995). Masking Microdata Files. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 114–119.

Kim, J.J. and Winkler, W.E. (2003). Multiplicative Noise for Masking Continuous Data. Statistical Research Division Research Report Series (Statistics #2003-01). U.S. Census Bureau. Available at: www.census.gov/srd/papers/pdf/rrs2003-01.pdf (accessed May 14, 2012).

Klein, M., Mathew, T., and Sinha, B. (2013). A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication. Center for Statistical Research & Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-02). U.S. Census Bureau. Available at: www.census.gov/srd/papers/ pdf/rrs2013-02.pdf (accessed Jan. 23, 2013).

Klein, M. and Sinha, B. (2013). Statistical Analysis of Noise Multiplied Data Using Multiple Imputation. Center for Statistical Research and Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-01). U.S. Census Bureau. Available at: www.census.gov/srd/papers/pdf/rrs2013-01.pdf (accessed Jan. 23, 2013).

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis With Missing Data, (second edition). Wiley.

Meng, X.L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9, 538–558.

Nayak, T., Sinha, B.K., and Zayatz, L. (2011). Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection. Journal of Official Statistics, 27, 527–544.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at: www.R-project.org/.

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

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

Reiter, J.P. (2005). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. Journal of Royal Statistical Society, Series A, 168, 185–205.

Reiter, J.P. and Raghunathan, T.E. (2007). The Multiple Adaptations of Multiple Imputation. Journal of American Statistical Association, 102, 1462–1471.

Robert, C.P. and Casella, G. (2005). Monte Carlo Statistical Methods, (second edition). Springer.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys: Wiley.

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

Robins, J.M. and Wang, N. (2000). Inference for Imputation Estimators. Biometrika, 87, 113–124.

Sinha, B.K., Nayak, T., and Zayatz, L. (2012). Privacy Protection and Quantile Estimation From Noise Multiplied Data. Sankhya, Series B, 73, 297–315.

Tanner, M.A. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). Journal of the American Statistical Association, 82, 528–550.

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

Received September 2012
Revised February 2013
Accepted May 2013