RESEARCH REPORT SERIES
(Statistics #2013-01)

Statistical Analysis of Noise Multiplied Data Using Multiple Imputation

Martin Klein
Bimal Sinha

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: January 23, 2013

Statistical Analysis of Noise Multiplied Data Using Multiple Imputation

Martin Klein and Bimal Sinha

Abstract

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 non-standard 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 paper 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 code data.

Martin Klein (E-mail: *martin.klein@census.gov*) is Research Mathematical Statistician in the Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233. Bimal Sinha (E-mail: *sinha@umbc.edu*) is Research Mathematical Statistician in the Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, and Professor in the Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250. The authors are thankful to Eric Slud for carefully reviewing the manuscript and to 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.

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 code (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; Raghunathan, Reiter, Rubin 2003; Reiter 2003, 2005; Reiter, 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, Sinha, and Zayatz 2011; Sinha, Nayak, Zayatz 2012). In fact, Klein, Mathew, and Sinha (2012), hereafter referred to as Klein et al. (2012), developed likelihood based data analysis methods under noise multiplication for drawing inference in several parametric models; and they provided a comprehensive comparison of the above two methods, namely, multiple imputation and noise multiplication. Klein et al. (2012) 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. (2012) also developed similar analysis for top code 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 paper we will refer to both these data setups as *mixture* data rather than top code data which is strictly reserved for the case when values above C are completely ignored.

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. 2012). However, as noted by Klein et al. (2012), the likelihood based analysis is generally complicated due to the non-standard 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. This obviously provides an extra layer of confidentiality protection against data intruders!

The paper 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 three common parametric 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. 2012) is presented in Section 5 through simulation. It turns out that the inferences obtained using the methodology of this paper are comparable with, and just slightly less accurate than, those obtained in Klein et al. (2012). Section 6 provides some concluding remarks, and the Appendices A, B and C contain proofs of some technical results.

We end this section with an important observation that a direct application of multiple imputation procedures along the lines of Reiter (2003) based on the induced distribution of the noise perturbed data, which would naturally provide a highly desirable *double privacy protection*, is also possible. However, since such induced distributions are generally complicated in nature, the resulting data analysis based on multiple imputations may be involved. We will return to this approach along with some other relevant issues (see Section 6) in a future communication.

2 Overview of the method for full noise multiplication

In this section we first provide an overview of the proposed data analysis approach in a general framework, including a *crucial* method for imputing noise variables from noise multiplied data. We also describe in details two general methods of variance estimation of the parameter estimates, those of Rubin (1993) and Wang and Robins (1998).

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

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

and the marginal density of z_i is

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

As clearly demonstrated in Klein et al. (2012), 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 also the analysis must be customized to the noise distribution $h(r)$. A direct use of the familiar synthetic data method (Raghunathan, Reiter, and Rubin 2003; Reiter 2003) based on the noise multiplied sample $z_1, \ldots, z_n$, which would naturally provide *double privacy protection*, can also be quite complicated due to the same reason. Instead what we propose here is a procedure to *recover* 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 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:

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

Obviously, if the complete data $\boldsymbol{x}_c$ were observed, one would simply recover the original data $y_i = \frac{z_i}{r_i}$, $i = 1, \ldots, n$, 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

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

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

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

Each data set $\boldsymbol{y}^{*(j)}$ is now analyzed as if it were an original 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 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 proposed data analysis need to be put together: imputation of r^* from $\boldsymbol{z}$ and combination rules for η_j and v_j from several imputations. We discuss below these two crucial points.

2.2 Imputation of r from z and Rubin's (1987) combination rule

The imputed values of $r_1, \ldots, r_n$ here 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 $(r_1^{*(j)}, \ldots, r_n^{*(j)})$, $j = 1, \ldots, m$. 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, namely, $\boldsymbol{z}$. 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

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

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 $\boldsymbol{y}^{(t+1)} = (\frac{z_1}{r_1^{(t+1)}}, \ldots, \frac{z_n}{r_n^{(t+1)}})$, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ 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 $(r_1^{(t)}, \ldots, r_n^{(t)})$ 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 $(r_1^{(t)}, \ldots, r_n^{(t)})$ 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. Note that sampling from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ in step (II) above will typically be straightforward, either directly or via appropriate MCMC algorithms. 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$(1-\epsilon, 1+\epsilon)$, i.e., when

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

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*

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

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

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

(I) Generate U, V as independent Uniform$(0,1)$ and let $W = \gamma^V/(1-\epsilon)^{V-1}$.

(II) Accept $R = W$ if $U \leq M^{-1}q(\frac{z}{W}|\boldsymbol{\theta})$, otherwise reject W and return to step (I).

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

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

The proof of Proposition 1 appears in Appendix A.

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

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

Drawing r_i^* from the conditional density $h(r_i|z_i,\boldsymbol{\theta}^*)$ defined in (4) and setting $y_i^* = \frac{z_i}{r_i^*}$ 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 $\frac{z_i}{r_i^*}$ 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).

Rubin's (1987) Rule. Based on Rubin's (1987) combining rules, we obtain the MI estimator of Q:

$$\bar{\eta}_m = \frac{1}{m}\sum_{j=1}^{m}\eta_j, \tag{7}$$

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

$$T_m = (1 + 1/m)b_m + \bar{v}_m, \tag{8}$$

where $b_m = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \bar{\eta}_m)^2$ and $\bar{v}_m = \frac{1}{m}\sum_{j=1}^{m} v_j$. $\bar{\eta}_m$ and 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 $(m-1)(1+a_m^{-1})^2$ where $a_m = \frac{(1+m^{-1})b_m}{\bar{v}_m}$.

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. For example, taking $n = 2$, when $f(y|\theta) = \frac{1}{\theta}e^{-\frac{y}{\theta}}$, $\theta > 0$, $y > 0$, the posterior distribution of θ, given $\boldsymbol{y}$, under the noninformative prior $p(\theta) \propto [\frac{1}{\theta}]^\delta$ will be proper whenever $1 + \delta > 0$. But the same posterior, given $\boldsymbol{z}$, will be proper only if

$$A_\delta(z_1, z_2) = \int_0^\infty \int_0^\infty \frac{h(r_1)h(r_2)dr_1dr_2}{[\frac{z_1}{r_1} + \frac{z_2}{r_2}]^{1+\delta}[r_1r_2]} \tag{9}$$

is finite. Taking $h(r) = \frac{(r\beta)^\beta e^{-r\beta}}{r\Gamma(\beta)}$ with $E(R) = 1$ and $\text{Var}(R) = \frac{1}{\beta} > 0$, and $z_1 = z_2$, this amounts to the finiteness of the integral

$$I_\delta = \int_0^\infty \int_0^\infty \frac{e^{-\beta(r_1+r_2)} r_1^{\beta+\delta-1} r_2^{\beta+\delta-1} dr_1dr_2}{(r_2 + r_1)^{1+\delta}}. \tag{10}$$

Upon making the transformation from (r_1, r_2) to $u = r_1 + r_2$ and $v = \frac{r_1}{r_1+r_2}$, I_δ simplifies to

$$I_\delta = [\int_0^1 v^{\beta+\delta-1}(1-v)^{\beta+\delta-1}dv] \times [\int_0^\infty e^{-u\beta} u^{2\beta+\delta-2}du] \tag{11}$$

which is not finite when either $\beta + \delta \leq 0$ or $2\beta + \delta \leq 1$! One can choose $\beta = 0.5$ and $\delta = 0$ or $\delta = -0.5$ (recall the condition $1 + \delta > 0$). 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.

2.3 Wang and Robins's (1998) combination rules

Wang and Robins (1998) described variance estimators in the context of two types of multiple imputation: Type A and Type B. We discuss below these two approaches.

Type A. Here the procedure to generate r^* and hence $y^* = \frac{z}{r^*}$ is the same as just described in the preceding subsection. However the variance estimators use different formulas as described below.

1. Compute the multiple imputation (MI) estimator of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\theta}}_A = \frac{1}{m}\sum_{j=1}^m \hat{\boldsymbol{\theta}}_j$, where $\hat{\boldsymbol{\theta}}_j$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ computed on jth imputed dataset. Recall that the jth imputed dataset $[y_1^{*(j)}, \cdots, y_n^{*(j)}]$ is obtained by first drawing $\boldsymbol{\theta}_j^*$ from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{z}$, and then drawing $r_i^{*(j)}$ the conditional distribution of r_i given z_i and $\boldsymbol{\theta} = \boldsymbol{\theta}_j^*$, and finally substituting $y_i^{*(j)} = \frac{z_i}{r_i^{*(j)}}$.

2. Compute $S_{ij}(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j)$, the $p \times 1$ score vector, with its ℓth element as $S_{ij\ell}(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j) =$

$\frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \theta_\ell}\Big|_{y=y_i^{*(j)},\,\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j}$, $\ell = 1, \cdots, p$, $i = 1, \cdots, n$, $j = 1, \cdots, m$. Obviously the above quantity also depends on $\boldsymbol{\theta}_j^*$ through $y_i^{*(j)}$.

3. Also compute the $p \times p$ information matrix $S_{ij}^*(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j)$ whose (ℓ, ℓ')th element is computed as $S_{ij\ell\ell'}^*(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j) = \frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \theta_\ell \partial \theta_{\ell'}}\Big|_{y=y_i^{*(j)},\,\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j}$, $\ell, \ell' = 1, \cdots, p$, $i = 1, \cdots, n$, $j = 1, \cdots, m$.

4. By Wang and Robins (1998): $\sqrt{n}(\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_p[\mathbf{0}, V_A]$, where $V_A = I_{\text{obs}}^{-1} + \frac{1}{m} I_c^{-1} J + \frac{1}{m} J' I_{\text{obs}}^{-1} J$ with $J = I_{\text{mis}} I_c^{-1} = (I_c - I_{\text{obs}}) I_c^{-1}$, and $I_c = E[-((\frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \theta_\ell \partial \theta_{\ell'}}))]$ and $I_{\text{obs}} = E[-((\frac{\partial^2 \log g(z|\boldsymbol{\theta})}{\partial \theta_\ell \partial \theta_{\ell'}}))]$.

5. A consistent variance estimator $\hat{V}_A$ is obtained by estimating I_c by $\hat{I}_c = \frac{1}{m}\sum_{j=1}^m \hat{I}_{c,j}$ with $\hat{I}_{c,j} = -\frac{1}{n}\sum_{i=1}^n S_{ij}^*(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j)$ and estimating I_{obs} by

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

6. For any given $Q(\boldsymbol{\theta})$, the variance of the 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.

Type B. 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 is usually computed via the EM algorithm (Klein et al. 2012). Here are the essential steps.

1. Draw $r_i^* \sim h(r|z_i, \hat{\boldsymbol{\theta}}_{\text{mle}}(\boldsymbol{z}))$, $i = 1, \cdots, n$.

2. Having obtained r_i^*'s, perform multiple imputation and obtain the MLE on each completed dataset to get $\hat{\boldsymbol{\theta}}_1, \cdots, \hat{\boldsymbol{\theta}}_m$.

3. Compute MI estimate of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\theta}}_B = \frac{1}{m}\sum_{j=1}^m \hat{\boldsymbol{\theta}}_j$.

4. Compute $S_{ij}(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j)$, the $p \times 1$ score vector, with its ℓth element as $S_{ij\ell}(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j) = \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \theta_\ell}\Big|_{y=y_i^{*(j)},\,\theta=\hat{\theta}_j}$, $\ell = 1, \cdots, p$, $i = 1, \cdots, n$, $j = 1, \cdots, m$.

5. Also compute the $p \times p$ information matrix $S_{ij}^*(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j)$ with its (ℓ, ℓ')th element computed as $S_{ij\ell\ell'}^*(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j) = \frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \theta_\ell \partial \theta_{\ell'}}\Big|_{y=y_i^{*(j)},\,\theta=\hat{\boldsymbol{\theta}}_j}$, $\ell, \ell' = 1, \cdots, p$, $i = 1, \cdots, n$, $j = 1, \cdots, m$.

6. By Wang and Robins (1998): $\sqrt{n}(\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_p[\mathbf{0}, V_B]$, where $V_B = I_{\text{obs}}^{-1} + \frac{1}{m} I_c^{-1} J = I_{\text{obs}}^{-1} + \frac{1}{m} I_c^{-1}(I_c - I_{\text{obs}}) I_c^{-1}$.

7. A consistent variance estimator $\hat{V}_A$ is obtained by estimating I_c by $\hat{I}_c = \frac{1}{m}\sum_{j=1}^{m} \hat{I}_{c,j}$ with $\hat{I}_{c,j} = -\frac{1}{n}\sum_{i=1}^{n} S_{ij}^*(y_i^{*(j)}, \hat{\theta}_j)$, and estimating I_{obs} by

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

8. 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 $\hat{\theta}_B$ will generally lead to more accurate inferences than $\hat{\theta}_A$, 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 which 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 Analysis of mixture data

Recall that a mixture data in our context consist of unperturbed values below C and a masked version of values above C, obtained by either an imputation method or by noise multiplication.

Analysis of mixture data can be carried out in several different ways (An and Little 2007; Klein et al. 2012). In this section we discuss the analysis of such data following the procedure outlined earlier, namely, by (i) suitably recovering the top code 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 top code y-values and methods to appropriately combine the original y-values and *synthetic* top code 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 $\boldsymbol{y} = (y_1, \cdots, y_n)$, $\boldsymbol{r} = (r_1, \cdots, r_n)$, $\boldsymbol{z} = (z_1, \cdots, z_n)$ where $z_i = y_i \times r_i$, we define $\boldsymbol{x} = (x_1, \cdots, x_n)$ and $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_n)$ with $\Delta_i = I(y_i \leq C)$ and $x_i = y_i$ if $y_i \leq C$, and $= z_i$ if $y_i > C$. Inference for $\boldsymbol{\theta}$ will be based on either (i) $[(x_1, \Delta_1), \cdots, (x_n, \Delta_n)]$ or (ii) just $(x_1, \cdots, x_n)$. Under both the scenarios, which each guarantee that the sensitive y-values are protected, several data sets of the type $(y_1^*, \cdots, y_n^*)$ will be released along with a data analysis plan. Naturally, in case (i) when information on the indicator variables $\boldsymbol{\Delta}$ 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 Census Bureau) has a choice depending upon to what extent information about the released data should be provided to the data users. We describe below the 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 $y_i^* = \frac{x_i}{r_i^*}$. 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 (Klein et al. 2012)

$$h(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta}) = \frac{f(\frac{x_i}{r_i}|\boldsymbol{\theta})\frac{h(r_i)}{r_i}}{\int_0^{\frac{x_i}{C}} f(\frac{x_i}{\omega}|\boldsymbol{\theta})\frac{h(\omega)}{\omega}d\omega}, \text{ for } 0 \leq r_i \leq \frac{x_i}{C}. \tag{12}$$

When the noise distribution is the uniform density (5), then (12) becomes

$$h_U(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta}) = \frac{f(\frac{x_i}{r_i}|\boldsymbol{\theta})r_i^{-1}}{\int_{1-\epsilon}^{\min\{\frac{x_i}{C}, 1+\epsilon\}} f(\frac{x_i}{\omega}|\boldsymbol{\theta})\omega^{-1}d\omega}, \text{ for } 1-\epsilon \leq r_i \leq \min\{\frac{x_i}{C}, 1+\epsilon\}, \tag{13}$$

and Proposition 1 provides an algorithm for sampling from the above density (13).

Regarding choice of $\boldsymbol{\theta}$, we can proceed following the Type B method (see Section 2) and use the MLE of $\boldsymbol{\theta}$ ($\hat{\boldsymbol{\theta}}_{\text{mle}}$) based on the data $[(x_1, \Delta_1), \cdots, (x_n, \Delta_n)]$. This will often be direct (via EM algorithm) in view of the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\Delta})$ reported in Klein et al. (2012) and reproduced below:

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\Delta}) = \prod_{i=1}^{n} [f(x_i|\boldsymbol{\theta})]^{\Delta_i} [\int_0^{\frac{x_i}{C}} f(\frac{x_i}{r}|\boldsymbol{\theta}) \frac{h(r)}{r} dr]^{1-\Delta_i}. \tag{14}$$

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 $[(x_1, \Delta_1), \cdots, (x_n, \Delta_n)]$ using the likelihood $L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\Delta})$ given above.

2. Draw r_i^* for those $i = 1, \cdots, n$ for which $\Delta_i = 0$, from the conditional distribution (12) 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 $L(\theta|\boldsymbol{x}, \boldsymbol{\Delta})$. 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 $[(x_1, \Delta_1), \cdots, (x_n, \Delta_n)]$.

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 $i = 1, \cdots, n$ for which $\Delta_i = 0$.

II. Draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|y_1^{(t+1)}, \cdots, y_n^{(t+1)})$ where $y_i^{(t+1)} = \frac{x_i}{r_i^{(t+1)}}$ 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 (r_i^{**}, Δ_i^*) from the reported values of $(x_1, \cdots, x_n)$ and compute $y_i^{**} = \frac{x_i}{r_i^{**}}$ if $\Delta_i^* = 0$, and $y_i^{**} = x_i$, otherwise, $i = 1, \cdots, n$. 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 (Klein et al. 2012)

$$\psi(\delta = 1|x, \boldsymbol{\theta}) = P[\Delta = 1|x, \boldsymbol{\theta}] = \frac{f(x|\boldsymbol{\theta})I(x < C)}{f(x|\boldsymbol{\theta})I(x < C) + I(x > 0)\int_0^{\frac{x}{C}} f(\frac{x}{r}|\boldsymbol{\theta})\frac{h(r)}{r}dr}, \tag{15}$$

drawing of (r_i^{**}, Δ_i^*), 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 (12).

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 (Klein et al. 2012)

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n}[f(x_i|\boldsymbol{\theta})I(x_i < C) + I(x_i > 0)\int_0^{\frac{x_i}{C}} f(\frac{x_i}{r}|\boldsymbol{\theta})\frac{h(r)}{r}dr]. \tag{16}$$

Alternatively, one can proceed as in Type A method (sampling $r_1^{**}, \ldots, r_n^{**}$ from the posterior predictive distribution) by plugging in $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ which 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 $(r_i^{(t+1)}, \Delta_i^{(t+1)})$ from $h(r, \delta|x_i, \boldsymbol{\theta}^{(t)})$. 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 $p(\boldsymbol{\theta}|y_1^{(t+1)}, \cdots, y_n^{(t+1)})$ of $\boldsymbol{\theta}$, where $y_i^{(t+1)} = x_i$ if $\Delta_i^{(t+1)} = 1$, and $y_i^{(t+1)} = \frac{x_i}{r_i^{(t+1)}}$ 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 $\boldsymbol{y}^{*(1)}$, ..., $\boldsymbol{y}^{*(m)}$ 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.

4 Details for normal, exponential, and lognormal

4.1 Normal data

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

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}, \quad -\infty < \mu < \infty, 0 < \sigma^2 < \infty. \tag{17}$$

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

$$(\sigma^2|\boldsymbol{y}) \sim \frac{(n-1)s^2}{\chi^2_{n-1}}, \quad (\mu|\sigma^2, \boldsymbol{y}) \sim N(\bar{y}, \sigma^2/n), \tag{18}$$

with $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ (Gelman et al. 2004). The conditional density $h(r|z, \boldsymbol{\theta})$ as defined in (4) now takes the form

$$h(r|z, \boldsymbol{\theta}) = \frac{\exp[-\frac{1}{2\sigma^2}(z/r - \mu)^2]r^{-1}}{\int_{1-\epsilon}^{1+\epsilon} \exp[-\frac{1}{2\sigma^2}(z/\omega - \mu)^2]\omega^{-1}d\omega}, 1 - \epsilon \leq r \leq 1 + \epsilon. \tag{19}$$

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 (19).*

(I) Generate U, V as independent Uniform$(0, 1)$ and let $W = (1+\epsilon)^V/(1-\epsilon)^{V-1}$.

(II) Accept $R = W$ if $U \leq \exp[-\frac{1}{2\sigma^2}(z/W - \mu)^2]/M$, otherwise reject W and return to step (I).

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

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

and if $z < 0$ then

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

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

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

In the case of mixture data, the conditional density (12) now becomes

$$h(r|x, \Delta = 0, \boldsymbol{\theta}) = \frac{\exp[-\frac{1}{2\sigma^2}(x/r - \mu)^2]r^{-1}}{\int_{1-\epsilon}^{\min\{\frac{x}{C}, 1+\epsilon\}} \exp[-\frac{1}{2\sigma^2}(x/\omega - \mu)^2]\omega^{-1}d\omega}, 1 - \epsilon \le r \le \min\{\frac{x}{C}, 1+\epsilon\}, \qquad (20)$$

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

4.2 Exponential data

In this section we consider the case of an exponential population, and thus we let $f(y|\theta) = \frac{1}{\theta}e^{-y/\theta}$, $0 \le y < \infty$. We place the following improper prior on θ: $p(\theta) \propto 1$, $0 < \theta < \infty$. The posterior distribution of θ given $\boldsymbol{y}$ is

$$p(\theta|\boldsymbol{y}) = \frac{(\sum_{i=1}^n y_i)^{n-1}}{\Gamma(n-1)}\theta^{-(n-1)-1}e^{-(\sum_{i=1}^n y_i)/\theta}, \ 0 < \theta < \infty,$$

which has the form of an inverse gamma distribution, i.e., $(\theta^{-1}|\boldsymbol{y}) \sim \text{Gamma}(n-1, \frac{1}{\sum_{i=1}^n y_i})$.

Customized noise distribution for fully perturbed data. Suppose that the noise distribution

is given by

$$h(r) = \frac{\delta^{\delta+1}}{\Gamma(\delta+1)} r^{-(\delta+1)-1} e^{-\delta/r}, \;\; 0 < r < \infty, \tag{21}$$

where $\delta > 1$, and $E(R) = 1$ and $\mathrm{Var}(R) = (\delta-1)^{-1}$. We note that $h(r)$ is a form of the inverse gamma distribution such that $R \sim h(r) \Leftrightarrow R^{-1} \sim \mathrm{Gamma}(\delta+1, 1/\delta)$. This choice of the noise distribution is customized to the exponential distribution in the sense that it permits closed form evaluation of the integral in (1). The *pdf* $g(z|\boldsymbol{\theta})$ defined in (1) now takes the form $g(z|\theta) = \frac{\delta^{\delta+1}(\delta+1)}{\theta(\frac{z}{\theta}+\delta)^{\delta+2}}$, $0 < z < \infty$, and hence the conditional *pdf* $h(r|z, \boldsymbol{\theta})$ defined in (4) is now

$$h(r|z,\theta) = \frac{(\frac{z}{\theta}+\delta)^{\delta+2}}{\Gamma(\delta+2)} \exp[-\frac{1}{r}(\frac{z}{\theta}+\delta)] r^{-(\delta+2)-1}, \;\; 0 < r < \infty. \tag{22}$$

We note that (22) is an inverse gamma density, more specifically, $(r_i^{-1}|z_i, \theta) \sim \mathrm{Gamma}(\delta+2, \frac{1}{\frac{z_i}{\theta}+\delta})$, and thus samples from the conditional distribution of r_i given z_i and θ are easily extracted.

Uniform noise distribution. Suppose that we take the noise distribution to be uniform as defined in (5). Then the conditional *pdf* $h(r|z, \boldsymbol{\theta})$ as defined in (4) now has the form

$$h(r|z,\theta) = \frac{\exp(-\frac{z}{r\theta}) r^{-1}}{\int_{1-\epsilon}^{1+\epsilon} \exp(-\frac{z}{\omega\theta}) \omega^{-1} d\omega}, \;\; 1-\epsilon \le r \le 1+\epsilon. \tag{23}$$

We apply Proposition 1 to obtain an algorithm for sampling from this conditional density.

Corollary 2 *The following algorithm produces a random variable R whose density is (23).*

(I) Generate U, V as independent Uniform$(0,1)$ and let $W = (1+\epsilon)^V/(1-\epsilon)^{V-1}$.

(II) Accept $R = W$ if $U \le \exp(-\frac{z}{W\theta})/M$, otherwise reject W and return to step (I).

The constant M is defined as $M \equiv M(\theta, \epsilon, z) = \exp(-\frac{z}{\theta(1+\epsilon)})$. The expected number of iterations of steps (I) and (II) required to obtain R is

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

In the case of mixture data, the conditional density (12) now becomes

$$h(r|x,\Delta=0,\boldsymbol{\theta})=\frac{\exp(-\frac{x}{r\theta})r^{-1}}{\int_{1-\epsilon}^{\min\{\frac{x}{C},1+\epsilon\}}\exp(-\frac{x}{\omega\theta})\omega^{-1}d\omega},1-\epsilon\le r\le\min\{\frac{x}{C},1+\epsilon\},\tag{24}$$

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

4.3 Lognormal data

We next consider the case of the lognormal population: $f(y|\boldsymbol{\theta})=\frac{1}{y\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(\log y-\mu)^2]$, $0\le y<\infty$. We define a prior distribution on (μ,σ^2) as in (17). The posterior distribution of (μ,σ^2) is then given by (18) upon replacing each y_i by $\log(y_i)$.

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

$$h(r)=\frac{1}{r\xi\sqrt{2\pi}}\exp[-\frac{1}{2\xi^2}(\log r+\xi^2/2)^2],\ 0<r<\infty,\tag{25}$$

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

$$R\sim h(r|z,\boldsymbol{\theta})\Leftrightarrow\log(R)\sim N\left\{-\frac{\xi^2}{2}+\frac{\xi^2}{\sigma^2+\xi^2}[\log(z)+\frac{\xi^2}{2}-\mu],\ \frac{\sigma^2\xi^2}{\sigma^2+\xi^2}\right\}.\tag{26}$$

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

$$h(r|z,\boldsymbol{\theta})=\frac{\exp[-\frac{1}{2\sigma^2}(\log(z/r)-\mu)^2]}{\int_{1-\epsilon}^{1+\epsilon}\exp[-\frac{1}{2\sigma^2}(\log(z/\omega)-\mu)^2]d\omega},1-\epsilon\le r\le1+\epsilon.\tag{27}$$

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

Corollary 3 *The following algorithm produces a random variable R whose density is (27).*

(I) Generate U, V as independent Uniform$(0,1)$ and let $W=(1+\epsilon)^V/(1-\epsilon)^{V-1}$.

(II) Accept $R = W$ *if* $U \leq Wz^{-1}\exp[-\frac{1}{2\sigma^2}(\log(z/W)-\mu)^2]/M$*, otherwise reject* W *and return to step (I).*

The constant M *is defined as*

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

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

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

In the case of mixture data, the conditional density (12) now becomes

$$h(r|x, \Delta=0, \boldsymbol{\theta}) = \frac{\exp[-\frac{1}{2\sigma^2}(\log(x/r)-\mu)^2]}{\int_{1-\epsilon}^{\min\{\frac{x}{C}, 1+\epsilon\}}\exp[-\frac{1}{2\sigma^2}(\log(x/\omega)-\mu)^2]d\omega}, 1-\epsilon \leq r \leq \min\{\frac{x}{C}, 1+\epsilon\}, \quad (28)$$

and a simple modification of Corollary 3 yields an algorithm to sample from this *pdf*.

5 Simulation study

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, exponential, and lognormal populations in conjunction with uniform and customized noise distributions as far as possible, as outlined in Section 4. 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. (2012), and (2) to assess and compare the finite sample performance of Rubin's (1987) esti-

mation methods with those of Wang and Robins (1998) under our settings of fully noise multiplied and mixture data.

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; Table 2 provides results for the case of an exponential population when the parameter of interest is the mean θ; and Table 3 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; and the results in each table are based on a simulation with 5000 iterations and $m = 5$ imputations of the noise variables generated at each iteration. 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 method of Wang and Robins (1998) described in Section 2.3.

NM10UIA1: Analysis based on noise multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the method of Section 2.2 with Rubin's (1987) variance formula and 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 method of Section 2.2 with Rubin's (1987) variance formula and 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 method of Wang and Robins (1998) described in Section 2.3.

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. (2012).

NM10CIB, NM10CIA1, NM10CIA2, NM10CIA3, NM10CL: These methods are defined analogously to the methods above, but $h(r)$ is now the customized noise distribution (21) (exponential data) or (25) (lognormal data); the parameters δ and ξ appearing in $h(r)$ are chosen so that $\mathrm{Var}(R) = \frac{\epsilon^2}{3}$, 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), expected value of the estimated standard deviation of the estimator ($\widehat{\mathrm{SD}}$), coverage probability of the associated confidence interval (Cvg.), and expected length of the corresponding confidence interval relative to the expected 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 Section 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 - 3.

1. In terms of RMSE, bias, and SD of point estimators, as well as expected 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. (2012) 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 Section 2.2 with that of Wang and Robins's (1998) standard deviation estimators of Section 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 and exponential cases. With the sample size of $n = 100$ we notice in Tables 1 and 2 that the mean of Rubin's (1987) estimator is slightly less than the true SD while both of Wang and Robins (1998) estimators have mean slightly larger than the true SD. Interestingly, in the lognormal case, for the sample size $n = 100$ of Table 3, 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 (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 4 provides results for the case of a normal population, Table 5 provides results for the case of an exponential population, and Table 6 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 4 - 6 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. In the case of fully noise perturbed data we noticed a tendency for Rubin's (1987) standard deviation estimator to exhibit a slight negative bias. In the case of mixture data we no longer observe this effect; in fact, Rubin's (1987) estimator now 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 $(\boldsymbol{x}, \boldsymbol{\Delta})$) 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 Concluding remarks

There are two primary ways of rigorous data analysis under privacy protection: multiple imputation and noise perturbation. Klein et al. (2012) 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 methods of Reiter (2003) and An and Little (2007), based on the original data. 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 paper 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 (12), 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 (12) for general continuous $f(y|\boldsymbol{\theta})$, when $h(r)$ is the uniform distribution (5). Moreover, we have seen that in the exponential and lognormal cases 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 paper, 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.

In conclusion, we observe that, from a data user's perspective, our method does require a complete knowledge of the underlying parametric model of the original data so that efficient model based estimates can be used while using the (reconstructed) y-values. In the absence of such a knowledge, likely misspecification of the population model may lead to incorrect conclusions (Robins

and Wang 2000). We also wonder if reporting both z-values (one observed set) and reconstructed y-values (multiple sets) would lead to an enhanced inference! It would also be beneficial to develop appropriate data analysis methods based on a direct application of multiple imputation on the noise multiplied data itself, thus providing double privacy protection. Lastly, it seems that, as a general principle, some sort of homogeneity tests should be carried out across the multiply imputed data sets before they are routinely combined. We will address these issues in a future communication.

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

$$\frac{s_{\text{target}}(r)}{s_{\text{instr}}(r)} = [\log(\gamma)-\log(1-\epsilon)]q(\frac{z}{r}|\boldsymbol{\theta}) \le [\log(\gamma)-\log(1-\epsilon)]M, \tag{29}$$

for all $r \in [1-\epsilon,\gamma]$. Note that the cumulative distribution function corresponding to $s_{\text{instr}}(r)$ is $\mathbb{S}_{\text{instr}}(r) = \frac{\log(r)-\log(1-\epsilon)}{\log(\gamma)-\log(1-\epsilon)}$, $1-\epsilon \le r \le \gamma$, and the inverse of this distribution function is $\mathbb{S}^{-1}_{\text{instr}}(u) = \frac{\gamma^u}{(1-\epsilon)^{u-1}}$, $0 \le u \le 1$. 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 $\frac{M^{-1}s_{\text{target}}(W)}{[\log(\gamma)-\log(1-\epsilon)]s_{\text{instr}}(W)} = \frac{q(\frac{z}{W}|\boldsymbol{\theta})}{M}$, step (II) is equivalent to accepting V if $U \le \frac{M^{-1}s_{\text{target}}(W)}{[\log(\gamma)-\log(1-\epsilon)]s_{\text{instr}}(W)}$, which is the usual rejection step based on the bound in (29). Finally, we use the well known fact that the expected number of iterations of the rejection algorithm is equal to the bounding constant in (29) times the normalizing constant for $s_{\text{target}}(r)$, i.e., $\frac{[\log(\gamma)-\log(1-\epsilon)]M}{\int_{1-\epsilon}^{\gamma} q(\frac{z}{\omega}|\boldsymbol{\theta})\omega^{-1}d\omega}$.

Appendix B

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

Exponential distribution. Here $g(z|\theta) = \int \frac{1}{\theta} e^{-\frac{z}{r\theta}} \frac{h(r)}{r} dr$. When the noise distribution is uniform over $[1-\epsilon, 1+\epsilon]$, since $e^{-\frac{z}{r\theta}}$ is monotone decreasing in z, the joint *pdf* of $\boldsymbol{z}$ can be bounded above by $K(\theta)^{-n} e^{-\frac{nz_{(1)}}{(1+\epsilon)\theta}}$ for some $K > 0$, which is integrable under a flat or noninformative prior for θ. Under the *customized* prior for θ, in the *pdf* of Z, namely $g(z|\theta) \propto \frac{1}{\theta}[\frac{z}{\theta} + \delta]^{-(\delta+2)}$, replacing any z by $z_{(1)}$, the joint *pdf* of $\boldsymbol{z}$ is dominated by $\frac{1}{\theta^n}[\frac{z_{(1)}}{\theta} + \delta]^{-n(\delta+2)}$ which is readily seen to be integrable under a flat or noninformative prior for θ.

Normal distribution. Here $g(z|\boldsymbol{\theta}) \propto \frac{1}{\sigma} \int e^{-\frac{(\frac{z}{r}-\mu)^2}{2\sigma^2}} \frac{h(r)}{r} dr$. Writing down the joint *pdf* of $z_1, \cdots, z_n$, 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

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

where $\delta \geq 0$. To prove that $U(\boldsymbol{z})$ is finite for any given $\boldsymbol{z}$, note that $[\sum_{i=1}^{n} \frac{z_i^2}{r_i^2} - \frac{(\sum_{i=1}^{n} \frac{z_i}{r_i})^2}{n}] = \frac{1}{2} \sum_{i,j=1}^{n} (\frac{z_i}{r_i} - \frac{z_j}{r_j})^2 \geq \frac{1}{2}[\frac{z_1}{r_1} - \frac{z_2}{r_2}]^2$ 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 $[\frac{z_1}{r_1} - \frac{z_2}{r_2}]^2 = [\frac{z_1}{z_2} - \frac{r_1}{r_2}]^2 \times z_2^2 r_1^{-2}$. Then under the *condition*

$$\int_r \frac{h(r)}{r} dr = K_1 < \infty, \quad \int_{r_1 \leq r_2} r_1^{2(n+\delta)-1} r_2^{-1} h(r_1) h(r_2) dr_1 dr_2 = K_2 < \infty, \tag{30}$$

$U(\boldsymbol{z})$ is bounded above by

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

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

Lognormal distribution. Here $g(z|\boldsymbol{\theta}) \propto \frac{1}{z\sigma} \int e^{-\frac{(\log(\frac{z}{r})-\mu)^2}{2\sigma^2}} h(r) dr$. Writing down the joint density of $z_1, \cdots, z_n$, and putting $u = \log(\frac{z}{r})$, 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 $U(\boldsymbol{z})$ given by

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

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

$$[\sum_{i=1}^{n} (u_i - \bar{u})^2] = \frac{1}{2} \sum_{i,j=1}^{n} (u_i - u_j)^2 \geq \frac{1}{2}(u_1 - u_2)^2 = \frac{1}{2}[\log(\frac{z_1}{z_2}) - \log(\frac{r_1}{r_2})]^2 \geq \frac{1}{2}[\log(\frac{z_1}{z_2})]^2 \text{ for } r_1 < r_2.$$

Hence $U(\boldsymbol{z})$ is always finite since $\int_{r_1<r_2} h(r_1)h(r_2)dr_1dr_2 < \infty$.

Appendix C

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

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

Exponential distribution. From (14), the likelihood function in this case is given by

$$L(\theta|\text{data}) \propto \theta^{-n} e^{-\sum_{i=1}^{n} \frac{x_i \Delta_i}{\theta}} \prod_{i=1}^{n} [\int_0^{\frac{x_i}{C}} e^{-\frac{x_i}{r\theta}} \frac{h(r)}{r} dr]^{1-\Delta_i}$$

Under a uniform noise distribution, the term $\int_0^{\frac{x_i}{C}} e^{-\frac{x_i}{r\theta}} \frac{h(r)}{r} dr$ is bounded above by $\frac{1}{2\epsilon} \times \int_{1-\epsilon}^{1+\epsilon} e^{-\frac{x_i}{r\theta}} \frac{dr}{r} \leq K_\epsilon e^{-\frac{x_i}{\theta(1+\epsilon)}}$ where $K_\epsilon > 0$ is a constant. Hence, apart from a finite constant, $L(\theta|\text{data})$ is bounded above by

$$L(\theta|\text{data}) \leq \theta^{-n} \times e^{-\frac{[\sum_{i=1}^{n} x_i \Delta_i + (1+\epsilon)^{-1} \sum_{i=1}^{n} x_i (1-\Delta_i)]}{\theta}}$$

which is integrable with respect to flat or noninformative prior for θ, irrespective of any configuration of the given data!

Normal distribution. Given the data $[(x_1, \Delta_1), \cdots, (x_n, \Delta_n)]$, 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

$$L(\boldsymbol{\theta}|\text{data}) \propto \sigma^{-n}[e^{-\sum_{i\in I_1}\frac{(x_i-\mu)^2}{2\sigma^2}}][\prod_{i\in I_0}\int_0^{\frac{x_i}{C}} e^{-\frac{(\frac{x_i}{r_i}-\mu)^2}{2\sigma^2}}\frac{h(r_i)}{r_i}I(x_i > 0)dr_i].$$

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 $U(\text{data})$ given by

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

Writing $v_i = \frac{x_i}{r_i}$ for $i \in I_0$, the expression $\Psi(\text{data}) = \sum_{i\in I_1} x_i^2 + \sum_{i\in I_0}\frac{x_i^2}{r_i^2} - \frac{(\sum_{i\in I_1} x_i + \sum_{i\in I_0}\frac{x_i}{r_i})^2}{n}$ 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 $(\bar{x}_1, S_1^2)$ and $(\bar{x}_0, S_0^2)$ 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 $\Psi(\text{data})$ is $\frac{S_1^2}{r+s}$, and if I_1 is empty, $\Psi(\text{data}) = S_0^2/n$. In the first case, $U(\text{data})$ is finite whenever $\int_0^{\frac{x_i}{C}}\frac{h(r)}{r}dr < \infty$ for $i \in I_0$. In the second case, we proceed as in the fully noise perturbed case for normal and conclude that $U(\text{data})$ is finite under the conditions stated in (30) except that the bounds of r_i in the integrals are replaced by $\frac{x_i}{C}$. In particular, for uniform noise distribution, the conditions trivially hold.

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

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

Exponential distribution. From (16), the likelihood function in this case is given by

$$L(\theta|\text{data}) = \theta^{-n}\prod_{i=1}^{n}[e^{-\frac{x_i}{\theta}}I(x_i < C) + \int_0^{\frac{x_i}{C}} e^{-\frac{x_i}{\theta r_i}}\frac{h(r_i)}{r_i}dr_i].$$

Now, for each i, the first term within [.] is bounded above by $e^{-\frac{x_i}{\theta}}$ and the second term by $e^{-\frac{C}{\theta}}\psi_C(x_i)$ where $\psi_C(x_i) = \int_0^{\frac{x_i}{C}} \frac{h(r_i)}{r_i} dr_i$ since $\frac{x_i}{r_i} > C$. Define $\psi_C = \max(\psi_C(x_1), \cdots, \psi_C(x_n))$, and assume that the noise distribution $h(r)$ satisfies: $\psi_C < \infty$. Then it is easy to check that $L(\theta|\text{data})$ is bounded above by $[\max(1, \psi_C)]\theta^{-n} e^{-\frac{2}{\theta}\sum_{i=1}^n \min(x_i, C)}$ which is integrable wrt θ under a flat or noninformative prior.

Normal distribution. Upon carefully examining the joint *pdf* of the data $\boldsymbol{x}$, given by (16), let us split the entire data into three mutually exclusive sets:

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

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* $L(\theta|\text{data} \in I_2)$ as a product of terms of the type

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

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:

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

where J_1 and J_2 form a partition of $(1, \cdots, n)$. 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 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 $\frac{x}{r}$ by $u = \log(\frac{x}{r})$, we get the posterior propriety of μ and σ under flat or noninformative priors when the noise is uniform. We omit the details.

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. Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C.

Klein, M., Mathew, T., and Sinha, B. (2013). A comparison of disclosure control methods: multiple imputation versus noise multiplication. *Technical report available from* `http://www.census.gov/srd/www/byyear.html`.

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.

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.

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.

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.

Robins, J.M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, **87**, 113-124.

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	99.95	99.40	94.66	1.0000	143.89	-6.70	143.75	140.47	92.92	1.0000
NM10UIB	100.11	3.18	100.07	100.99	95.16	1.0160	145.92	-6.02	145.81	148.67	93.06	1.0584
NM10UIA1	100.10	3.12	100.06	99.62	94.80	1.0021	145.87	-6.34	145.75	142.14	92.66	1.0119
NM10UIA2	100.10	3.12	100.06	99.62	94.80	1.0021	145.87	-6.34	145.75	142.14	92.66	1.0119
NM10UIA3	100.10	3.12	100.06	101.24	95.12	1.0185	145.87	-6.34	145.75	149.76	93.36	1.0661
NM10UL	100.10	3.13	100.06	99.58	94.74	1.0018	145.59	-6.32	145.47	141.87	92.62	1.0100
NM20UIB	100.92	3.27	100.88	101.48	95.20	1.0209	150.15	-4.91	150.09	152.80	93.56	1.0878
NM20UIA1	100.83	3.10	100.80	100.26	94.92	1.0086	150.45	-4.17	150.41	146.89	93.02	1.0457
NM20UIA2	100.83	3.10	100.80	100.26	94.92	1.0087	150.45	-4.17	150.41	146.89	93.02	1.0458
NM20UIA3	100.83	3.10	100.80	101.78	95.38	1.0238	150.45	-4.17	150.41	154.09	93.70	1.0969
NM20UL	100.74	3.09	100.70	100.11	94.94	1.0071	149.43	-4.84	149.37	145.74	93.10	1.0375
NM50UIB	103.96	3.39	103.92	104.18	94.80	1.0480	170.21	-4.83	170.16	173.55	93.26	1.2355
NM50UIA1	104.11	3.46	104.07	103.53	94.40	1.0415	171.79	1.78	171.79	169.74	93.12	1.2083
NM50UIA2	104.11	3.46	104.07	103.53	94.52	1.0438	171.79	1.78	171.79	169.74	93.16	1.2109
NM50UIA3	104.11	3.46	104.07	104.79	94.56	1.0541	171.79	1.78	171.79	176.64	93.78	1.2575
NM50UL	103.31	3.29	103.27	102.64	94.52	1.0326	167.38	-4.24	167.34	164.29	93.16	1.1695

Table 2: Inference under fully perturbed exponential data with mean $\theta = 1$ and $n = 100$

	Parameter of interest is the mean θ					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	100.69	0.84	100.70	100.08	94.40	1.0000
NM10UIB	101.20	0.95	101.20	102.72	94.12	1.0263
NM10UIA1	101.16	1.06	101.17	100.50	94.48	1.0041
NM10UIA2	101.16	1.06	101.17	100.50	94.62	1.0089
NM10UIA3	101.16	1.06	101.17	103.10	94.46	1.0301
NM10UL	101.15	0.97	101.16	100.43	94.40	1.0034
NM10CIB	100.93	0.90	100.93	102.72	94.48	1.0263
NM10CIA1	100.82	0.90	100.83	100.49	94.30	1.0040
NM10CIA2	100.82	0.90	100.83	100.49	94.38	1.0081
NM10CIA3	100.82	0.90	100.83	103.10	94.56	1.0301
NM10CL	100.78	0.91	100.79	100.42	94.34	1.0034
NM20UIB	102.19	1.12	102.20	103.84	94.66	1.0376
NM20UIA1	102.25	1.27	102.25	101.69	94.56	1.0160
NM20UIA2	102.25	1.27	102.25	101.69	95.12	1.0351
NM20UIA3	102.25	1.27	102.25	104.22	94.74	1.0413
NM20UL	101.95	0.92	101.95	101.38	94.44	1.0130
NM20CIB	102.20	0.98	102.21	103.91	94.36	1.0382
NM20CIA1	102.11	1.15	102.12	101.67	94.30	1.0158
NM20CIA2	102.11	1.15	102.12	101.67	94.66	1.0317
NM20CIA3	102.11	1.15	102.12	104.24	94.46	1.0415
NM20CL	101.82	0.94	101.83	101.37	94.42	1.0129
NM50UIB	109.22	2.05	109.21	111.05	94.38	1.1095
NM50UIA1	109.92	4.02	109.86	109.23	94.40	1.0914
NM50UIA2	109.92	4.02	109.86	109.23	96.04	1.2050
NM50UIA3	109.92	4.02	109.86	111.73	94.46	1.1164
NM50UL	108.02	2.05	108.01	107.61	94.44	1.0752
NM50CIB	109.08	1.99	109.07	110.14	94.36	1.1004
NM50CIA1	109.73	3.24	109.69	108.40	94.26	1.0830
NM50CIA2	109.73	3.24	109.69	108.40	95.72	1.1707
NM50CIA3	109.73	3.24	109.69	110.73	94.42	1.1064
NM50CL	108.08	2.02	108.08	107.07	94.18	1.0698

Table 3: 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 $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	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	202.26	1.56	202.28	201.82	93.88	1.0000	799.81	-11.83	799.80	793.27	93.16	1.0000
NM10UIB	202.80	1.69	202.81	208.31	94.10	1.0321	802.34	-11.16	802.34	826.38	93.16	1.0417
NM10UIA1	203.18	1.83	203.19	202.46	93.64	1.0032	803.98	-10.57	803.99	796.22	92.88	1.0037
NM10UIA2	203.18	1.83	203.19	202.46	93.64	1.0032	803.98	-10.57	803.99	796.22	92.88	1.0037
NM10UIA3	203.18	1.83	203.19	208.34	94.16	1.0323	803.98	-10.57	803.99	826.50	93.30	1.0419
NM10UL	202.91	1.70	202.92	202.31	93.62	1.0025	802.80	-11.16	802.80	795.55	92.78	1.0029
NM10CIB	202.72	1.48	202.73	208.30	93.92	1.0321	801.97	-12.25	801.96	826.34	93.34	1.0417
NM10CIA1	202.81	1.52	202.82	202.38	93.80	1.0028	802.40	-11.87	802.39	795.89	93.04	1.0033
NM10CIA2	202.81	1.52	202.82	202.38	93.80	1.0028	802.40	-11.87	802.39	795.89	93.04	1.0033
NM10CIA3	202.81	1.52	202.82	208.29	94.02	1.0320	802.40	-11.87	802.39	826.26	93.38	1.0416
NM10CL	202.68	1.41	202.70	202.25	93.84	1.0021	801.56	-12.39	801.55	795.26	93.20	1.0025
NM20UIB	204.60	2.55	204.61	210.24	94.16	1.0417	811.20	-7.89	811.24	835.35	93.26	1.0530
NM20UIA1	204.76	2.21	204.77	204.24	93.84	1.0120	811.69	-9.16	811.72	804.47	93.02	1.0141
NM20UIA2	204.76	2.21	204.77	204.24	93.84	1.0122	811.69	-9.16	811.72	804.47	93.02	1.0144
NM20UIA3	204.76	2.21	204.77	210.16	94.02	1.0413	811.69	-9.16	811.72	834.97	93.34	1.0526
NM20UL	204.33	2.29	204.34	203.83	93.94	1.0099	810.06	-8.76	810.10	802.52	93.34	1.0117
NM20CIB	204.59	2.05	204.60	209.93	94.18	1.0402	810.41	-11.38	810.41	834.00	93.22	1.0513
NM20CIA1	204.41	1.72	204.42	204.04	94.02	1.0110	809.98	-12.28	809.97	803.51	92.98	1.0129
NM20CIA2	204.41	1.72	204.42	204.04	94.04	1.0112	809.98	-12.28	809.97	803.51	93.00	1.0132
NM20CIA3	204.41	1.72	204.42	209.88	94.08	1.0399	809.98	-12.28	809.97	833.77	93.28	1.0511
NM20CL	204.06	1.62	204.07	203.56	93.98	1.0086	808.43	-12.73	808.41	801.31	92.92	1.0101
NM50UIB	217.16	1.62	217.18	221.96	94.06	1.0998	866.70	-16.33	866.63	890.55	93.30	1.1226
NM50UIA1	217.31	2.95	217.31	216.77	93.44	1.0741	867.67	-9.31	867.71	862.13	92.64	1.0868
NM50UIA2	217.31	2.95	217.31	216.77	93.56	1.0810	867.67	-9.31	867.71	862.13	92.78	1.0960
NM50UIA3	217.31	2.95	217.31	222.23	93.62	1.1012	867.67	-9.31	867.71	891.63	92.78	1.1240
NM50UL	214.82	0.82	214.84	213.53	93.52	1.0580	855.59	-17.25	855.50	847.91	92.86	1.0689
NM50CIB	214.35	3.42	214.35	220.94	93.96	1.0948	854.98	-7.29	855.03	885.62	93.58	1.1164
NM50CIA1	215.22	4.67	215.19	215.77	93.84	1.0691	857.50	-1.24	857.58	857.56	93.16	1.0810
NM50CIA2	215.22	4.67	215.19	215.77	93.94	1.0749	857.50	-1.24	857.58	857.56	93.32	1.0888
NM50CIA3	215.22	4.67	215.19	221.25	94.02	1.0963	857.50	-1.24	857.58	886.83	93.50	1.1179
NM50CL	212.48	2.53	212.48	212.80	93.96	1.0544	845.95	-9.46	845.98	844.25	93.00	1.0643

Table 4: 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	98.70	99.30	94.50	1.0000	139.88	-9.00	139.60	140.15	93.68	1.0000
NM10UIB.i	98.81	-1.18	98.82	101.00	94.88	1.0171	140.72	-8.81	140.46	149.18	94.10	1.0645
NM10UIA1.i	98.79	-1.17	98.79	99.37	94.46	1.0007	140.62	-8.75	140.36	140.88	93.48	1.0053
NM10UIA2.i	98.79	-1.17	98.79	99.37	94.46	1.0007	140.62	-8.75	140.36	140.88	93.48	1.0053
NM10UIA3.i	98.79	-1.17	98.79	101.01	94.82	1.0172	140.62	-8.75	140.36	149.17	94.14	1.0644
NM10UL.i	98.81	-1.19	98.81	99.36	94.48	1.0005	140.54	-8.87	140.27	140.74	93.56	1.0042
NM10UIB.ii	98.83	-1.15	98.84	101.01	94.78	1.0172	140.71	-8.67	140.46	149.20	94.16	1.0646
NM10UIA1.ii	98.81	-1.20	98.82	99.37	94.50	1.0006	140.76	-8.89	140.50	140.87	93.52	1.0052
NM10UIA2.ii	98.81	-1.20	98.82	99.37	94.50	1.0006	140.76	-8.89	140.50	140.87	93.52	1.0052
NM10UIA3.ii	98.81	-1.20	98.82	101.00	94.84	1.0171	140.76	-8.89	140.50	149.21	94.06	1.0646
NM10UL.ii	98.81	-1.20	98.81	99.36	94.38	1.0006	140.54	-8.88	140.27	140.75	93.54	1.0043
NM20UIB.i	99.23	-1.12	99.23	101.10	94.70	1.0181	142.24	-8.52	142.00	150.74	93.92	1.0756
NM20UIA1.i	99.13	-0.97	99.14	99.55	94.48	1.0025	142.10	-7.89	141.89	142.68	93.64	1.0180
NM20UIA2.i	99.13	-0.97	99.14	99.55	94.48	1.0025	142.10	-7.89	141.89	142.68	93.64	1.0186
NM20UIA3.i	99.13	-0.97	99.14	101.13	94.90	1.0184	142.10	-7.89	141.89	150.71	94.12	1.0753
NM20UL.i	99.09	-1.06	99.10	99.51	94.42	1.0021	141.77	-8.20	141.55	142.24	93.56	1.0149
NM20UIB.ii	99.17	-1.11	99.18	101.13	94.78	1.0184	142.12	-8.37	141.89	150.76	93.90	1.0757
NM20UIA1.ii	99.13	-0.96	99.13	99.58	94.36	1.0028	142.61	-7.76	142.41	142.80	93.40	1.0189
NM20UIA2.ii	99.13	-0.96	99.13	99.58	94.36	1.0028	142.61	-7.76	142.41	142.80	93.44	1.0195
NM20UIA3.ii	99.13	-0.96	99.13	101.16	94.62	1.0187	142.61	-7.76	142.41	150.79	94.02	1.0760
NM20UL.ii	99.10	-1.07	99.10	99.52	94.40	1.0022	141.92	-8.25	141.69	142.31	93.44	1.0154
NM50UIB.i	99.67	-0.59	99.67	101.41	94.56	1.0212	148.43	-6.19	148.31	155.53	94.04	1.1098
NM50UIA1.i	99.77	-0.05	99.78	100.18	94.32	1.0088	149.25	-3.94	149.22	148.33	93.72	1.0584
NM50UIA2.i	99.77	-0.05	99.78	100.18	94.32	1.0089	149.25	-3.94	149.22	148.33	93.78	1.0630
NM50UIA3.i	99.77	-0.05	99.78	101.53	94.64	1.0224	149.25	-3.94	149.22	155.79	94.08	1.1116
NM50UL.i	99.55	-0.57	99.55	99.96	94.32	1.0066	147.32	-6.08	147.20	146.70	93.66	1.0467
NM50UIB.ii	99.99	-0.64	100.00	101.82	94.86	1.0254	150.46	-6.41	150.34	157.79	93.84	1.1259
NM50UIA1.ii	100.07	-0.01	100.08	100.60	94.44	1.0130	150.68	-3.90	150.64	150.30	93.64	1.0724
NM50UIA2.ii	100.07	-0.01	100.08	100.60	94.46	1.0133	150.68	-3.90	150.64	150.30	93.70	1.0791
NM50UIA3.ii	100.07	-0.01	100.08	101.98	94.76	1.0270	150.68	-3.90	150.64	158.04	94.08	1.1277
NM50UL.ii	99.74	-0.72	99.75	100.29	94.48	1.0100	148.93	-6.48	148.80	148.34	93.66	1.0584

Table 5: Inference for mixture exponential data with mean $\theta = 1$, $C = .90$ quantile, and $n = 100$

	Parameter of interest is the mean θ					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	99.12	0.25	99.13	100.03	94.52	1.0000
NM10UIB.i	99.53	0.34	99.54	102.66	94.50	1.0263
NM10UIA1.i	99.54	0.44	99.55	100.26	94.66	1.0024
NM10UIA2.i	99.54	0.44	99.55	100.26	94.68	1.0050
NM10UIA3.i	99.54	0.44	99.55	103.02	94.68	1.0300
NM10UL.i	99.51	0.37	99.52	100.22	94.66	1.0019
NM10UIB.ii	99.56	0.33	99.57	103.04	94.78	1.0302
NM10UIA1.ii	99.52	0.41	99.53	100.27	94.54	1.0024
NM10UIA2.ii	99.52	0.41	99.53	100.27	94.56	1.0052
NM10UIA3.ii	99.52	0.41	99.53	103.02	94.82	1.0299
NM10UL.ii	99.50	0.34	99.51	100.22	94.56	1.0020
NM20UIB.i	100.17	0.68	100.18	103.18	94.46	1.0315
NM20UIA1.i	100.24	0.86	100.24	100.86	94.60	1.0084
NM20UIA2.i	100.24	0.86	100.24	100.86	94.76	1.0178
NM20UIA3.i	100.24	0.86	100.24	103.50	94.74	1.0348
NM20UL.i	100.09	0.69	100.10	100.72	94.60	1.0069
NM20UIB.ii	100.34	0.73	100.35	103.59	94.52	1.0357
NM20UIA1.ii	100.30	0.70	100.31	100.91	94.54	1.0088
NM20UIA2.ii	100.30	0.70	100.31	100.91	94.86	1.0190
NM20UIA3.ii	100.30	0.70	100.31	103.61	94.50	1.0359
NM20UL.ii	100.22	0.69	100.23	100.77	94.56	1.0074
NM50UIB.i	102.43	1.55	102.43	105.02	94.60	1.0499
NM50UIA1.i	102.97	2.48	102.95	103.17	94.60	1.0314
NM50UIA2.i	102.97	2.48	102.95	103.17	95.18	1.0675
NM50UIA3.i	102.97	2.48	102.95	105.26	94.58	1.0523
NM50UL.i	102.11	1.57	102.11	102.54	94.66	1.0251
NM50UIB.ii	103.85	1.63	103.85	106.67	94.56	1.0665
NM50UIA1.ii	103.73	2.45	103.71	104.41	94.54	1.0439
NM50UIA2.ii	103.73	2.45	103.71	104.41	95.32	1.0959
NM50UIA3.ii	103.73	2.45	103.71	106.64	94.56	1.0662
NM50UL.ii	102.96	1.36	102.96	103.56	94.54	1.0354

Table 6: 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	781.86	794.90	93.48	1.0000
NM10UIB.i	99.46	2.11	99.45	100.87	95.08	1.0167	783.15	1.95	783.23	824.70	93.62	1.0375
NM10UIA1.i	99.46	2.11	99.44	99.23	94.78	1.0002	783.32	2.04	783.40	796.04	93.40	1.0014
NM10UIA2.i	99.46	2.11	99.44	99.23	94.78	1.0002	783.32	2.04	783.40	796.04	93.40	1.0014
NM10UIA3.i	99.46	2.11	99.44	100.87	95.10	1.0167	783.32	2.04	783.40	824.66	93.72	1.0374
NM10UL.i	99.47	2.11	99.46	99.22	94.78	1.0002	783.09	1.97	783.16	795.82	93.38	1.0011
NM10UIB.ii	99.48	2.11	99.47	100.87	95.06	1.0167	783.92	2.22	783.99	824.75	93.72	1.0376
NM10UIA1.ii	99.47	2.09	99.46	99.22	94.72	1.0002	783.18	1.64	783.25	796.00	93.36	1.0014
NM10UIA2.ii	99.47	2.09	99.46	99.22	94.72	1.0002	783.18	1.64	783.25	796.00	93.36	1.0014
NM10UIA3.ii	99.47	2.09	99.46	100.86	95.10	1.0167	783.18	1.64	783.25	824.70	93.70	1.0375
NM10UL.ii	99.47	2.11	99.46	99.23	94.72	1.0002	783.15	1.98	783.22	795.85	93.36	1.0012
NM20UIB.i	99.50	2.10	99.48	100.89	95.12	1.0169	787.17	2.26	787.25	827.71	93.60	1.0413
NM20UIA1.i	99.47	2.10	99.46	99.27	94.82	1.0006	786.76	2.44	786.84	798.97	93.30	1.0051
NM20UIA2.i	99.47	2.10	99.46	99.27	94.82	1.0006	786.76	2.44	786.84	798.97	93.30	1.0052
NM20UIA3.i	99.47	2.10	99.46	100.89	95.04	1.0170	786.76	2.44	786.84	827.52	93.82	1.0410
NM20UL.i	99.49	2.08	99.48	99.26	94.80	1.0005	785.69	1.62	785.77	798.04	93.34	1.0039
NM20UIB.ii	99.50	2.08	99.48	100.90	95.04	1.0170	786.09	1.92	786.16	827.94	93.66	1.0416
NM20UIA1.ii	99.51	2.09	99.50	99.28	94.84	1.0008	787.30	2.51	787.37	799.37	93.44	1.0056
NM20UIA2.ii	99.51	2.09	99.50	99.28	94.84	1.0008	787.30	2.51	787.37	799.37	93.44	1.0057
NM20UIA3.ii	99.51	2.09	99.50	100.91	95.06	1.0171	787.30	2.51	787.37	827.80	93.72	1.0414
NM20UL.ii	99.50	2.07	99.49	99.26	94.76	1.0006	785.97	1.54	786.05	798.27	93.36	1.0042
NM50UIB.i	99.83	2.33	99.81	101.09	95.24	1.0189	804.56	9.96	804.58	842.34	93.76	1.0597
NM50UIA1.i	99.84	2.46	99.82	99.58	94.90	1.0037	803.02	12.96	803.00	816.58	93.50	1.0273
NM50UIA2.i	99.84	2.46	99.82	99.58	94.90	1.0038	803.02	12.96	803.00	816.58	93.50	1.0282
NM50UIA3.i	99.84	2.46	99.82	101.12	95.12	1.0193	803.02	12.96	803.00	842.43	93.96	1.0598
NM50UL.i	99.73	2.32	99.72	99.50	94.86	1.0029	798.51	8.40	798.55	811.47	93.56	1.0208
NM50UIB.ii	100.05	2.42	100.03	101.32	95.18	1.0213	809.84	12.40	809.83	850.03	93.74	1.0694
NM50UIA1.ii	100.07	2.55	100.05	99.78	94.78	1.0058	809.88	14.73	809.83	822.84	93.68	1.0351
NM50UIA2.ii	100.07	2.55	100.05	99.78	94.78	1.0058	809.88	14.73	809.83	822.84	93.70	1.0366
NM50UIA3.ii	100.07	2.55	100.05	101.34	95.12	1.0215	809.88	14.73	809.83	850.50	93.78	1.0699
NM50UL.ii	99.96	2.40	99.94	99.68	94.66	1.0047	803.94	10.09	803.95	817.17	93.54	1.0280