Journal of Privacy and Confidentiality (2014) **6**, Number 1, 77–125

Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples

Martin Klein*, Thomas Mathew†, and Bimal Sinha‡

1 Introduction

Statistical agencies must control disclosure risk when releasing data to the public. If income data on individuals or businesses are released, it could be possible to match extremely large values to specific individuals or businesses that are known to be wealthy, especially if some additional information is available on the same units in the dataset. The purpose of the present investigation is to explore noise multiplication as a strategy to protect large values in a dataset from disclosure, and to develop methodology for analyzing the resulting data under the assumption of a log-normal distribution on the sensitive variable. We assume that the log-scale mean of the sensitive variable is described by a linear regression on a set of non-sensitive covariates, and that the goal of the data analysis is to draw inference on parameters in the regression. We focus on the log-normal distribution because it is well known to be appropriate for modeling income data ([6]; [8]; [11]; [20]), and for income data, the extreme values usually need disclosure protection.

In such situations where the extreme values in the dataset require protection, the method of top coding is often used for statistical disclosure control. Under top coding, a threshold C is determined, and any value that exceeds C is not reported, instead what may be reported is the value C itself. Top coding is straightforward to apply, and usually provides sufficient protection, since very limited information is revealed about the data values above the threshold C. There are, however, some drawbacks to top coding, such as: (1) the information contained in the top part of the data is essentially lost, and (2) there is no explicit tuning mechanism to allow the statistical agency to balance the level of privacy protection with the accuracy of inferences derived from the

*Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, mailto:martin.klein@census.gov.

†Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, and Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, mailto:mathew@umbc.edu.

‡Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC, USA, and Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, mailto:sinha@umbc.edu.

Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

protected data.

An and Little [5] proposed synthetic data methods as alternatives to top coding for statistical disclosure control. Under synthetic data methodology, sensitive values are not released, instead sensitive values are imputed multiple times by draws from an appropriate statistical model. The multiply imputed values are combined with any non-sensitive (original) values to produce multiple synthetic copies of the original confidential data file, which are then released. For a data user, each of the multiple datasets is analyzed as if it were a true sample from the population, and the results are combined using simple formulas [27] to obtain the final inference. Multiple imputation (MI) has been in existence for some time as a viable methodology to handle missing data [35]. Rubin [36] proposed using MI as a method of sensitive data protection. The rigorous foundations for MI as a method for sensitive data protection were further developed (e.g., [10]; [9]; [26]; [27, 28, 29, 30]) and this still continues to be an active area of research. We refer to the recent monograph by Drechsler [10] for a detailed and general discussion of multiple imputation as a tool for disclosure control. An and Little [5] and Jenkins et al. [15] argue that synthetic data methodology provides an attractive alternative to top coding, and a remedy to some of the drawbacks of top coding, because analysis of the synthetic data is straightforward for a data user, and some information contained in the top part of the data is retained. The MI method also appears to have a desirable robustness property to certain model misspecification as discussed by An and Little [5].

Noise perturbation by addition or multiplication has also been advocated by some statisticians as a possible data confidentiality protection mechanism [14], [22], [16]. Noise multiplication continues to be investigated for its potential role in statistical disclosure limitation; see the recent articles by Nayak et al. [24], Sinha et al. [37], and Lin and Wise [21]. The first public use microdata sample (PUMS) produced from the Survey of Business Owners (SBO) was released in August 2012 (`http://www.census.gov/econ/sbo/`), and noise multiplication was employed for confidentiality protection of some variables. Here each record corresponds to a business surveyed in the 2007 SBO, and a number of variables are provided relating to firm size, business characteristics, and business owner characteristics. In this data product, a number of steps are taken to protect confidentiality of businesses, and the variables relating to receipts, payroll, and employment are rounded and multiplied by random noise prior to release [1].

Under the scenario of noise multiplication of values only above the threshold C, the released information consists of the original observations below C and the observations above C after noise multiplication. The noise distribution used and the value of the threshold C will also be publicly available. Furthermore, we shall consider two types of data releases referred to as cases (I) and (II). In case (I) data release, each released value includes an indicator of whether or not it has been perturbed (i.e., noise multiplied), while in case (II), no such indicator is provided. Naturally, case (I) data appear to carry more information than case (II) data, and so one would expect that case (I) would lead to more accurate inference than case (II), but at a potentially increased disclosure risk. We point out that a likelihood-based analysis under noise multiplication of the entire sample for disclosure control has also been carried out under several parametric models

[18].

While data analysis and the accuracy of inferences based on perturbed data is important, it is equally important to evaluate the level of disclosure risk incurred by releasing the perturbed data. To address the issue of accuracy of inference, we present a simulation study (Section 4), and a data analysis example (Section 5.1). To address the issue of disclosure risk, we present an evaluation in the context of a real data example (Section 5.2). Our methodology for evaluating disclosure risk is based on a criterion similar to one used by Lin and Wise [21] for the same purpose. We also evaluate the disclosure risk of synthetic data based on this same criterion, and make some comparisons with perturbation under noise multiplication.

Here is the organization of the paper. Section 2 presents methodology for the statistical analysis of the data resulting from noise multiplication of the observations above a threshold C, keeping the rest of the data (below C) undisturbed. The main technical results are presented here. This section also contains details about EM algorithms used to compute the maximum likelihood estimates of the model parameters. The formulas for observed Fisher information of the noise multiplied data appear in Appendices 1.1 and 1.2. In Section 3.1, we review methodology for the formal analysis of top coded log-normal data based on Tobit regression models. Appendix 1.3 provides expressions for observed Fisher information contained in the top coded data. Following the work of An and Little [5], we review synthetic data methods in Section 3.2. Section 4 reports results of a simulation study designed to assess the accuracy of inference under the proposed noise multiplication method and also to compare accuracy of inference of the proposed method with that of top coding and synthetic data. In Section 5.1 we present data analysis results under the proposed noise multiplication method using public use data from the 2000 U.S. Current Population Survey, and the results are compared with those obtained from top coded and synthetic data. Section 5.2 presents a disclosure risk evaluation of the proposed noise multiplication methods in the context of the 2000 U.S. Current Population Survey data, and compares the disclosure risk of the proposed methods with that of synthetic data. We conclude the paper in Section 6 with a discussion of some advantages, drawbacks, and possible extensions of the proposed method.

We end this section with the following general observations. A very appealing feature of noise multiplication is the presence of an explicit tuning mechanism, namely, the noise generating distribution, which allows the statistical agency that generates the data to precisely balance accuracy of the data with the desired level of disclosure control. Such a tuning mechanism is not available under top coding. Obviously, a noise generating distribution with large dispersion should provide a high level of disclosure control with perhaps less accurate inference, and the situation would be reversed if the noise generating distribution has small dispersion. Compared to top coding, the noise multiplication method can retain more information about the top part of the data; therefore, inferences derived from the noise multiplied data can be more accurate than those derived from top coded data. Another appealing feature of noise multiplication (which is shared with top coding) is that noise multiplied data are easy for the data producer to create. The data producer does not need to consider which regressor variables (or functions of regressor variables) may be important predictors of the response

variable when creating the noise multiplied data, and yet the methodology in Section 2 can be used to study the nature of this regression, using only a noise multiplied version of the response variable.

Numerical results presented in Section 4 provide specific guidance on how the dispersion of the noise generating distribution affects the accuracy of inference, while the results in Section 5.2 provide guidance on how the dispersion of the noise generating distribution affects the level of privacy protection. It should also be noted that as in the case of top coding, the likelihood-based data analysis under noise multiplication is complicated, though software can be made available to ease this burden. In fact, an R code for implementing the proposed methodology is available from us upon request.

2 Methodology For Noise Multiplication of Extreme Values

Consider a set of n independent and log-normally distributed random variables $y_1, \ldots, y_n$, along with a set of n vectors of regressor variables $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, each having dimension $p \times 1$ with $n > p$. We treat the regression variables as fixed (non-random), and we assume that $\ln y_i \sim N(\boldsymbol{u}_i'\boldsymbol{\beta}, \sigma^2)$, that is,

$$y_i \sim f_{\boldsymbol{\theta}}(y_i|\boldsymbol{u}_i) = \frac{1}{y_i\sigma\sqrt{2\pi}} \exp\left[\frac{-(\ln y_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right],\ y_i > 0,\ \text{for } i = 1, \ldots, n, \qquad (1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ and $0 < \sigma^2 < \infty$ are both unknown. Let $r_1, \ldots, r_n$ be a set of n independent and identically distributed noise random variables each having the density $h(r)$, where $h(r)$ is a known density of a continuous distribution having nonnegative support. Then the noise multiplied version of y_i is $z_i = y_i \times r_i$, and the density of z_i is $g_{\boldsymbol{\theta}}(z_i|\boldsymbol{u}_i) = \int f_{\boldsymbol{\theta}}(\frac{z}{r}|\boldsymbol{u}_i)h(r)r^{-1}dr$. Now, any observation y_i which exceeds a specified threshold $C > 0$ is considered sensitive and, under top coding such values are simply not reported. We now consider the option of reporting z_i, the noise perturbed version of y_i. More precisely, for $i = 1, ..., n$, let us define

$$\Delta_i = I(y_i \le C) \ \text{ and } \ x_i = \begin{cases} y_i, & \text{if } y_i \le C, \\ z_i, & \text{if } y_i > C, \end{cases}$$

where $I(A)$ is the indicator function of the event A.

Inference for $\boldsymbol{\theta}$ will be based on $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ along with the regressor variables $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, or based on $(x_1, \ldots, x_n)$ and $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$. We note that the latter data do not directly identify which observations have been perturbed, and hence appear to provide more disclosure control than the former. We shall refer to the former data type, namely $\{(x_1, \Delta_1, \boldsymbol{u}_1), \ldots, (x_n, \Delta_n, \boldsymbol{u}_n)\}$, as case (I), and the latter, namely $\{(x_1, \boldsymbol{u}_1), \ldots, (x_n, \boldsymbol{u}_n)\}$, as case (II). For either the case (I) or case (II) data type, the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ of the unknown parameter $\boldsymbol{\theta}$ can be computed using the EM algorithm [23], and the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be estimated using the

inverse of the observed Fisher information about $\boldsymbol{\theta}$. The MLE of a (scalar) parametric function $\psi(\boldsymbol{\theta})$ is then readily obtained as $\psi(\hat{\boldsymbol{\theta}})$, and an estimator of the variance of $\psi(\hat{\boldsymbol{\theta}})$ is obtained by applying the delta method. Wald-type inference can be used to obtain an approximate level $(1-\alpha)$ confidence interval for $\psi(\boldsymbol{\theta})$ of the form $\psi(\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2}[\text{Estimated variance of } \psi(\hat{\boldsymbol{\theta}})]^{1/2}$ where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function (*cdf*).

We now explain the details of our proposed methods of inference. To begin, the expression for the joint distribution of (x_i, r_i, Δ_i) is given by the following proposition, which is general, and does not depend on the log-normality assumption on $f_{\boldsymbol{\theta}}(y|\boldsymbol{u})$. The joint densities of $\{(x_i, \Delta_i), i = 1, \cdots, n\}$ and $\{x_1, \cdots, x_n\}$ follow as corollaries. We use these results in the sequel to derive EM algorithms, and to derive the expressions in Appendices 1.1 and 1.2 for the observed Fisher information about $\boldsymbol{\theta}$.

Proposition 1. Let $Y \sim f_{\boldsymbol{\theta}}(y|\boldsymbol{u})$, independent of $R \sim h(r)$, and let $Z = Y \times R$, $\Delta = I(Y \leq C)$, and $X = \begin{cases} Y \text{ if } Y \leq C, \\ Z \text{ if } Y > C. \end{cases}$ Let $\boldsymbol{u}$ denote a fixed vector of regressor variables. Then the joint probability distribution of (X, R, Δ) is given by

$$k_{\boldsymbol{\theta}}(x, r, \delta|\boldsymbol{u}) = \begin{cases} f_{\boldsymbol{\theta}}(x|\boldsymbol{u})h(r), & \text{if } x < C,\ 0 < r < \infty,\ \delta = 1, \\ f_{\boldsymbol{\theta}}\left(\frac{x}{r}|\boldsymbol{u}\right) h(r) r^{-1}, & \text{if } 0 < r < \frac{x}{C},\ \delta = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. We let $Y \sim f_{\boldsymbol{\theta}}(y|\boldsymbol{u})$, independent of $R \sim h(r)$, and we let $\Delta = I(Y \leq C)$, $X = Y$ if $Y \leq C$ and $X = Y \times R$ if $Y > C$, where $C > 0$ is a constant. Then letting $F_{\boldsymbol{\theta}}(y|\boldsymbol{u}) = \int_{-\infty}^{y} f_{\boldsymbol{\theta}}(t|\boldsymbol{u})dt$ and $H(r) = \int_0^r h(t)dt$, we have, for $r > 0$,

$$\begin{aligned} \Pr(X \leq x, R \leq r, \Delta = 1) &= \Pr(X \leq x, R \leq r \,|\, \Delta = 1) \Pr(\Delta = 1) \\ &= \Pr(Y \leq x, R \leq r \,|\, Y \leq C) P(Y \leq C) \\ &= \Pr(Y \leq x, Y \leq C, R \leq r) \\ &= \begin{cases} F_{\boldsymbol{\theta}}(x|\boldsymbol{u})H(r), & \text{if } x \leq C, \\ F_{\boldsymbol{\theta}}(C|\boldsymbol{u})H(r), & \text{if } x > C, \end{cases} \end{aligned} \tag{2}$$

and, for $r > 0$, $C > 0$,

$$\begin{aligned}
\Pr(X \le x, R \le r, \Delta = 0) &= \Pr(X \le x, R \le r \,|\, \Delta = 0)\Pr(\Delta = 0) \\
&= \Pr(YR \le x, R \le r \,|\, Y > C)\Pr(Y > C) \\
&= \Pr(YR \le x, Y > C, R \le r) \\
&= \Pr\left(C < Y \le \frac{x}{R}, R \le r\right) \\
&= \begin{cases} \int_0^r \int_C^{x/\omega} f_{\boldsymbol{\theta}}(y|\boldsymbol{u})h(\omega)dyd\omega, & \text{if } r \le \frac{x}{C}, \\ \int_0^{x/C} \int_C^{x/\omega} f_{\boldsymbol{\theta}}(y|\boldsymbol{u})h(\omega)dyd\omega, & \text{if } r > \frac{x}{C} > 0, \\ 0, & \text{if } x \le 0, \end{cases} \\
&= \begin{cases} \int_0^r \left[F_{\boldsymbol{\theta}}\left(\frac{x}{\omega}|\boldsymbol{u}\right) - F_{\boldsymbol{\theta}}(C|\boldsymbol{u})\right] h(\omega)d\omega, & \text{if } r \le \frac{x}{C}, \\ \int_0^{x/C} \left[F_{\boldsymbol{\theta}}\left(\frac{x}{\omega}|\boldsymbol{u}\right) - F_{\boldsymbol{\theta}}(C|\boldsymbol{u})\right] h(\omega)d\omega, & \text{if } 0 < \frac{x}{C} < r, \\ 0, & \text{if } x \le 0. \end{cases}
\end{aligned} \tag{3}$$

To obtain the joint probability density function (*pdf*) of (X, R, Δ), we differentiate (2) and (3) with respect to x and r to obtain:

$$\frac{\partial^2}{\partial x \partial r}\Pr(X \le x, R \le r, \Delta = 1) = \begin{cases} f_{\boldsymbol{\theta}}(x|\boldsymbol{u})h(r), & \text{if } x < C, \\ 0, & \text{if } x > C, \end{cases}$$

$$\frac{\partial^2}{\partial x \partial r}\Pr(X \le x, R \le r, \Delta = 0) = \begin{cases} f_{\boldsymbol{\theta}}\left(\frac{x}{r}|\boldsymbol{u}\right) h(r)r^{-1}, & \text{if } r < \frac{x}{C}, \\ 0, & \text{if } 0 < \frac{x}{C} < r, \\ 0, & \text{if } x < 0, \end{cases}$$

which completes the proof of Proposition 1. □

The following results follow from Proposition 1.

Corollary 1. The joint *pdf* of (X, Δ) is given by

$$k_{\boldsymbol{\theta}}(x, \delta|\boldsymbol{u}) = \begin{cases} f_{\boldsymbol{\theta}}(x|\boldsymbol{u}), & \text{if } x < C,\ \delta = 1, \\ 0, & \text{if } x_i > C,\ \delta = 1, \\ \int_0^{x/C} f_{\boldsymbol{\theta}}\left(\frac{x}{r}|\boldsymbol{u}\right) h(r)r^{-1}dr, & \text{if } x > 0,\ \delta = 0, \\ 0, & \text{if } x < 0,\ \delta = 0. \end{cases}$$

Corollary 2. The likelihood function for $\boldsymbol{\theta}$ based on $(x_1, \Delta_1), ..., (x_n, \Delta_n)$ is given by

$$\begin{aligned}
L(\boldsymbol{\theta}|x_1, ..., x_n, \Delta_1, ..., \Delta_n, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) &= \prod_{i=1}^n k_{\boldsymbol{\theta}}(x_i, \Delta_i|\boldsymbol{u}_i) \\
&= \prod_{i=1}^n \left\{ [f_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)]^{\Delta_i} \times \left[\int_0^{x_i/C} f_{\boldsymbol{\theta}}\left(\frac{x_i}{r}|\boldsymbol{u}_i\right) h(r)r^{-1}dr\right]^{1-\Delta_i} \right\}.
\end{aligned}$$

Corollary 3. The marginal *pdf* of X is given by

$$k_{\boldsymbol{\theta}}(x|\boldsymbol{u}) = f_{\boldsymbol{\theta}}(x|\boldsymbol{u})I(x<C) + \int_0^{x/C} f_{\boldsymbol{\theta}}\left(\frac{x}{r}|\boldsymbol{u}\right) h(r)r^{-1}drI(x>0)$$

$$= \begin{cases} f_{\boldsymbol{\theta}}(x|\boldsymbol{u}) + \int_0^{x/C} f_{\boldsymbol{\theta}}\left(\frac{x}{r}|\boldsymbol{u}\right) h(r)r^{-1}dr, & \text{if } 0<x<C, \\ f_{\boldsymbol{\theta}}(x|\boldsymbol{u}), & \text{if } x<0, \\ \int_0^{x/C} f_{\boldsymbol{\theta}}\left(\frac{x}{r}|\boldsymbol{u}\right) h(r)r^{-1}dr, & \text{if } x>C. \end{cases}$$

Corollary 4. The likelihood function for $\boldsymbol{\theta}$ based on $x_1, ..., x_n$ is given by

$$L(\boldsymbol{\theta}|x_1, ..., x_n, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \prod_{i=1}^n k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)$$

$$= \prod_{i=1}^n \left\{ f_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)I(x_i<C) + \int_0^{x_i/C} f_{\boldsymbol{\theta}}\left(\frac{x_i}{r}|\boldsymbol{u}_i\right) h(r)r^{-1}drI(x_i>0) \right\}.$$

EM Algorithms. We now derive EM algorithms for MLE computation under the case (I) and case (II) data types. In order to derive EM algorithms, we frame analysis of the noise perturbed data as a missing data problem. If case (I) data $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ are released, then we define the observed and missing data as $\boldsymbol{v}_{i,\text{obs}} = \{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ and $\boldsymbol{v}_{i,\text{mis}} = \{r_1, \ldots, r_n\}$, respectively. If case (II) data $\{x_1, \ldots, x_n\}$ are released, then we define the observed and missing data as $\boldsymbol{v}_{ii,\text{obs}} = \{x_1, \ldots, x_n)$ and $\boldsymbol{v}_{ii,\text{mis}} = \{(r_1, \Delta_1) \ldots, (r_n, \Delta_n)\}$, respectively. In both cases, the complete data are $\boldsymbol{v}_c = \{(x_1, r_1, \Delta_1), \ldots, (x_n, r_n, \Delta_n)\}$, and hence by Proposition 1, the complete data likelihood function is

$$\begin{aligned} & L(\boldsymbol{\theta}|\boldsymbol{v}_c) \\ &= \prod_{i=1}^n \frac{h(r_i)}{x_i\sqrt{2\pi\sigma^2}} \left\{ \exp\left[-\frac{(\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right] \right\}^{\Delta_i} \left\{ \exp\left[-\frac{(\ln(x_i/r_i) - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right] \right\}^{1-\Delta_i} \\ &\propto \prod_{i=1}^n \left\{ \frac{1}{\sqrt{\sigma^2}} \exp\left[-\frac{(\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right] \right\}^{\Delta_i} \left\{ \frac{1}{\sqrt{\sigma^2}} \exp\left[-\frac{(\ln(x_i/r_i) - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right] \right\}^{1-\Delta_i} \\ &= (\sigma^2)^{-n/2} \prod_{i=1}^n \left\{ \exp\left[-\frac{(\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right] \right\}^{\Delta_i} \left\{ \exp\left[-\frac{(\ln(x_i/r_i) - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right] \right\}^{1-\Delta_i}, \end{aligned}$$

and hence we can write the complete data log-likelihood function as

$$\ell(\boldsymbol{\theta}|\boldsymbol{v}_c) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left\{ \Delta_i \left(\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}\right)^2 + (1-\Delta_i)\left(\ln\frac{x_i}{r_i} - \boldsymbol{u}_i'\boldsymbol{\beta} \right)^2 \right\}$$

$$= -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (\boldsymbol{u}_i'\boldsymbol{\beta})^2$$

$$- \frac{1}{2\sigma^2}\sum_{i=1}^n \left\{ \Delta_i \left[(\ln x_i)^2 - 2(\ln x_i)(\boldsymbol{u}_i'\boldsymbol{\beta}) \right] + (1-\Delta_i)\left[\left(\ln\frac{x_i}{r_i} \right)^2 - 2(\boldsymbol{u}_i'\boldsymbol{\beta})\ln\frac{x_i}{r_i} \right] \right\}.$$

The specific details for cases (I) and (II) are given below.

EM Algorithm For Case (I) Type Data Release. The EM algorithm for computing the MLE of $\boldsymbol{\theta}$ based on $\boldsymbol{v}_{i,\text{obs}}$ is as follows.

E-step. We have

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{\theta}^{(t)}}\left[\ell(\boldsymbol{\theta}|\boldsymbol{v}_{\text{c}})|\boldsymbol{v}_{\text{i,obs}}\right] \\ &= -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{u}_i'\boldsymbol{\beta})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Delta_i\left[(\ln x_i)^2 - 2(\ln x_i)(\boldsymbol{u}_i'\boldsymbol{\beta})\right] \\ &\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(1-\Delta_i)\left\{\psi_2^*(x_i,\boldsymbol{\theta}^{(t)}) - 2(\boldsymbol{u}_i'\boldsymbol{\beta})\psi_1^*(x_i,\boldsymbol{\theta}^{(t)})\right\}, \end{aligned}$$

where

$$\begin{aligned} \psi_1^*(x_i,\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{\theta}^{(t)}}\left[\left(\ln\frac{x_i}{r_i}\right)\Big| x_i, \Delta_i = 0\right] \\ &= \frac{\int_0^{x_i/C}\left(\ln\frac{x_i}{r}\right)\exp\left[-\frac{\{\ln(x_i/r)-\boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(r)dr}{\int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/r)-\boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(r)dr}, \\ \psi_2^*(x_i,\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{\theta}^{(t)}}\left[\left(\ln\frac{x_i}{r_i}\right)^2\Big| x_i, \Delta_i = 0\right] \\ &= \frac{\int_0^{x_i/C}\left(\ln\frac{x_i}{r}\right)^2\exp\left[-\frac{\{\ln(x_i/r)-\boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(r)dr}{\int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/r)-\boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(r)dr}. \end{aligned}$$

To compute the conditional expectations above, we used Proposition 1 and Corollary 1 to obtain the conditional density of r_i given x_i and $\Delta_i = 0$.

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

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \left(\sum_{i=1}^{n}\boldsymbol{u}_i\boldsymbol{u}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{u}_i\left[\Delta_i\ln x_i + (1-\Delta_i)\psi_1^*(x_i,\boldsymbol{\theta}^{(t)})\right]\right), \\ (\sigma^{(t+1)})^2 &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i\left(\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t+1)}\right)^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\Delta_i)\left[\psi_2^*(x_i,\boldsymbol{\theta}^{(t)}) - 2(\boldsymbol{u}_i'\boldsymbol{\beta}^{(t+1)})\psi_1^*(x_i,\boldsymbol{\theta}^{(t)}) + (\boldsymbol{u}_i'\boldsymbol{\beta}^{(t+1)})^2\right], \end{aligned}$$

where the expressions for $\psi_1^*(x_i,\boldsymbol{\theta}^{(t)})$ and $\psi_2^*(x_i,\boldsymbol{\theta}^{(t)})$ are given above.

EM Algorithm For Case (II) Type Data Release. The EM algorithm for computing the MLE of $\boldsymbol{\theta}$ based on $\boldsymbol{v}_{ii,\text{obs}}$ is as follows.

E-step. We have

$$
\begin{aligned}
&Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta}|\boldsymbol{v}_{\text{c}})|\boldsymbol{v}_{\text{ii,obs}}] \\
&= -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{u}_i'\boldsymbol{\beta})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\left[(\ln x_i)^2 - 2(\ln x_i)(\boldsymbol{u}_i'\boldsymbol{\beta})\right]\psi_0^{**}(x_i,\boldsymbol{\theta}^{(t)})\right\} \\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\psi_2^{**}(x_i,\boldsymbol{\theta}^{(t)}) - 2(\boldsymbol{u}_i'\boldsymbol{\beta})\psi_1^{**}(x_i,\boldsymbol{\theta}^{(t)})\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
&\psi_0^{**}(x_i,\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}\left[\Delta_i|x_i\right] \\
&= \frac{I(x_i<C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]}{I(x_i<C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right] + \int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/\omega) - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(\omega)d\omega}, \qquad (4)
\end{aligned}
$$

$$
\begin{aligned}
&\psi_1^{**}(x_i,\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}\left[(1-\Delta_i)\left(\ln\frac{x_i}{r_i}\right)\Big|x_i\right] \\
&= \frac{\int_0^{x_i/C}\left(\ln\frac{x_i}{r}\right)\exp\left[-\frac{\{\ln(x_i/r) - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(r)dr}{I(x_i<C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right] + \int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/\omega) - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(\omega)d\omega},
\end{aligned}
$$

$$
\begin{aligned}
&\psi_2^{**}(x_i,\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}\left[(1-\Delta_i)\left(\ln\frac{x_i}{r_i}\right)^2\Big|x_i\right] \\
&= \frac{\int_0^{x_i/C}\left(\ln\frac{x_i}{r}\right)^2\exp\left[-\frac{\{\ln(x_i/r) - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(r)dr}{I(x_i<C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right] + \int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/\omega) - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t)}\}^2}{2(\sigma^{(t)})^2}\right]h(\omega)d\omega}.
\end{aligned}
$$

To compute the conditional expectations above, we used Proposition 1 and Corollary 3 to obtain the conditional density of Δ_i and r_i given x_i; and we used Corollaries 1 and 3 to get the conditional density of of Δ_i given x_i.

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

$$
\begin{aligned}
&\boldsymbol{\beta}^{(t+1)} = \left(\sum_{i=1}^{n}\boldsymbol{u}_i\boldsymbol{u}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{u}_i\left[(\ln x_i)\psi_0^{**}(x_i,\boldsymbol{\theta}^{(t)}) + \psi_1^{**}(x_i,\boldsymbol{\theta}^{(t)})\right]\right), \\
&(\sigma^{(t+1)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}^{(t+1)}\right)^2\psi_0^{**}(x_i,\boldsymbol{\theta}^{(t)}) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{\psi_2^{**}(x_i,\boldsymbol{\theta}^{(t)}) - 2(\boldsymbol{u}_i'\boldsymbol{\beta}^{(t+1)})\psi_1^{**}(x_i,\boldsymbol{\theta}^{(t)}) + (\boldsymbol{u}_i'\boldsymbol{\beta}^{(t+1)})^2[1-\psi_0^{**}(x_i,\boldsymbol{\theta}^{(t)})]\right\},
\end{aligned}
$$

where formulas for $\psi_0^{**}(x_i, \boldsymbol{\theta}^{(t)})$, $\psi_1^{**}(x_i, \boldsymbol{\theta}^{(t)})$, and $\psi_2^{**}(x_i, \boldsymbol{\theta}^{(t)})$ are given above.

Remark 1. The results in this section are presented for a general noise generating distribution $h(r)$. We only assume that $h(r)$ is the density of a continuous distribution having nonnegative support. In particular, we do not assume that $h(r)$ has mean 1, which is often required in order to get valid inferences in other noise multiplication methodologies that rely on moment based estimators of population parameters. Thus, under our proposed methodology, the choice of $h(r)$ is quite flexible, and in fact, $h(r)$ is the tuning mechanism that allows the data producer to control the accuracy of inference and level of privacy of protection. Numerical results presented later in Sections 4 and 5 are designed to give insight about the accuracy of inference and level of privacy protections of the proposed methods, using the noise density defined in (8), which is flexible and has no mass in an interval that contains 1.

Remark 2. The methodology outlined in this section provides tools for performing model selection in the regression context. For example, a test of significance for an individual regression coefficient can be obtained by checking if its confidence interval contains zero. Similarly, since we can compute MLEs of the parameters, and we have an expression for the likelihood function, likelihood ratio test statistics can be constructed and used to test a variety of hypotheses (e.g., comparing two competing regression models, testing significance of a subset of regression coefficients, etc.). This remark points to an advantage of releasing noise multiplied data over releasing a collection of summary statistics on the original data. By having the noise multiplied data available, a data user can explore any linear regression model that may be of interest, which may include quadratic terms, higher order interaction terms, power transformations of regressor variables, etc. On the contrary, if only summary statistics are provided, then the set of linear regression models that could be explored would be much more limited.

3 Review of Top Coding and Synthetic Data Methodology

As in the previous section, let $y_1, \ldots, y_n$ be a set of independent and log-normally distributed random variables such that $y_i \sim f_{\boldsymbol{\theta}}(y_i|\boldsymbol{u}_i) = \frac{1}{y_i\sigma\sqrt{2\pi}} \exp\left[\frac{-(\ln y_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]$, $y_i > 0$, where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are fixed regressor variables, each having dimension $p \times 1$ with $n > p$. The unknown parameter is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $0 < \sigma^2 < \infty$. Let $C > 0$ be a fixed constant such that any $y_i > C$ requires protection, and let $\psi(\boldsymbol{\theta})$ be the (scalar) population quantity that we want to draw inference on.

3.1 Top Coding

In a pioneering paper by Tobin [38] and in a relatively recent paper by Amemiya [4], the basics of data analysis under top code scenario have been nicely and extensively

discussed (except that in the classical Tobit model, $\ln(C)$ is taken as 0 and in fact it is the bottom code which is dealt with). Because of its similarities with the *probit* models, Goldberger [13] coined the phrase *Tobit* models and their analysis based on the normal distribution of $\ln(y)$ is indeed quite straightforward.

Let $x_i = \min\{y_i, C\}$ and $\Delta_i = I(y_i \leq C)$, $i = 1, \ldots, n$, and thus the top coded data that are released are $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$. Furthermore, let us define $\tilde{y}_i = \ln y_i$, $\tilde{C} = \ln C$, and $\tilde{x}_i = \ln x_i = \min\{\tilde{y}_i, \tilde{C}\}$. Following Amemiya [4], the likelihood function of the Tobit model under our setup is given by

$$L(\boldsymbol{\theta}) = \prod_{\{i\,:\,\Delta_i=0\}} \left[1 - \Phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\right] \prod_{\{i\,:\,\Delta_i=1\}} \left[\frac{1}{\sigma}\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\right],$$

where Φ and ϕ are the standard normal *cdf* and *pdf*, respectively. Obviously, the derivatives of the logarithm of the likelihood function are given by

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma} \sum_{\{i\,:\,\Delta_i=0\}} \frac{\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{u}_i}{1 - \Phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)} + \frac{1}{\sigma^2} \sum_{\{i\,:\,\Delta_i=1\}} (\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta})\boldsymbol{u}_i$$

and

$$\frac{\partial \ln L}{\partial(\sigma^2)} = \frac{1}{2\sigma^3} \sum_{\{i\,:\,\Delta_i=0\}} \left[\frac{(\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta})\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{1 - \Phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right] - \frac{\sum_{i=1}^n \Delta_i}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{\{i\,:\,\Delta_i=1\}} (\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2 .$$

The maximum likelihood estimators obtained by equating the above derivatives to 0, leading to equations which are nonlinear in the parameters, are usually based on an iterative scheme such as Newton-Raphson or the method of scoring [3], and it is proved in Amemiya [2] that the Tobit MLE is strongly consistent and asymptotically normal with the asymptotic variance-covariance matrix equal to $\left[-E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)\right]^{-1}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. The observed Fisher information matrix can then be readily obtained by replacing the unknown parameters $\boldsymbol{\theta}$ by the MLEs in $\left(-\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1}$. The expressions for the observed Fisher information about $\boldsymbol{\theta}$ contained in the top coded data are provided in Appendix 1.3. As usual, if $\hat{\boldsymbol{\theta}}$ denotes the MLE of $\boldsymbol{\theta}$, then the MLE of the (scalar) parametric function $\psi(\boldsymbol{\theta})$ is obtained as $\psi(\hat{\boldsymbol{\theta}})$, and an estimator of the variance of $\psi(\hat{\boldsymbol{\theta}})$ is obtained by applying the delta method. Wald-type inference can be used to obtain an approximate level $(1 - \alpha)$ confidence interval for $\psi(\boldsymbol{\theta})$ of the form $\psi(\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2}[\text{Estimated variance of } \psi(\hat{\boldsymbol{\theta}})]^{1/2}$, where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Routines for fitting Tobit regression models are available in several standard statistical software packages.

3.2 Synthetic Data

As already noted, MI methods are developed in An and Little [5], for the disclosure protection of the extreme values. In this scenario, we consider two MI methods presented

by An and Little [5], namely, parametric MI based on complete data (PMIC) and parametric MI based on deleted data (PMID). We now briefly describe these methods. Although it is only the values above C that are considered sensitive, a cut-point $C_I < C$ is selected, and any value $y_i > C_I$ is imputed. As discussed by An and Little [5], by choosing the cut-point C_I to be less than C, a mixing of sensitive and non-sensitive values is achieved, which should enhance the level of protection against disclosure. Let $\boldsymbol{y} = \{\boldsymbol{y}_{\text{ret}}, \boldsymbol{y}_{\text{del}}\}$ where $\boldsymbol{y}_{\text{ret}} = \{y_i : y_i \leq C_I\}$ denotes the values of $y_1, \ldots, y_n$ that will be retained, and $\boldsymbol{y}_{\text{del}} = \{y_i : y_i > C_I\}$ denotes the values $y_1, \ldots, y_n$ that will be deleted. Similarly, let $\boldsymbol{U} = (\boldsymbol{U}_{\text{ret}}, \boldsymbol{U}_{\text{del}})$ where $\boldsymbol{U}_{\text{ret}} = \{\boldsymbol{u}_i : y_i \leq C_I\}$ denotes the values of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ that correspond to values in $\boldsymbol{y}_{\text{ret}}$, and $\boldsymbol{U}_{\text{del}} = \{\boldsymbol{u}_i : y_i > C_I\}$ denotes the values of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ that correspond to values in $\boldsymbol{y}_{\text{del}}$, and hence $\boldsymbol{U}$ denotes the set of all regressor variables. Synthetic data methods are motivated from a Bayesian perspective under a diffuse prior distribution, hence, let $p(\boldsymbol{\theta}|\boldsymbol{U}) = p(\boldsymbol{\theta})$ be a diffuse prior density on $\boldsymbol{\theta}$.

The PMIC method proceeds as follows. One obtains the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{U}) \propto p(\boldsymbol{\theta}) \times \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i|\boldsymbol{u}_i)$, and then synthetic data are generated as follows:

1. Draw $\boldsymbol{\theta}^*$ from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{U})$.
2. For each i such that $y_i \in \boldsymbol{y}_{\text{del}}$, obtain the corresponding imputed value y_i^* as a draw from the truncated version of $f_{\boldsymbol{\theta}^*}(y_i|\boldsymbol{u}_i)$, defined as $f_{\boldsymbol{\theta}^*}^{(C_I,\infty)}(y_i|\boldsymbol{u}_i) = \frac{f_{\boldsymbol{\theta}^*}(y_i|\boldsymbol{u}_i) \times I_{(C_I,\infty)}(y_i)}{\int_{C_I}^{\infty} f_{\boldsymbol{\theta}^*}(\omega|\boldsymbol{u}_i)d\omega}$. Thus the imputed data $\boldsymbol{y}_{\text{del}}^*$ is the set of y_i^* values for each i such that $y_i \in \boldsymbol{y}_{\text{del}}$.

Steps (1) and (2) above are repeated independently a total of m times to obtain $\boldsymbol{y}_{\text{del}}^{*(1)}, ..., \boldsymbol{y}_{\text{del}}^{*(m)}$. Finally, we obtain the synthetic data as $\boldsymbol{y}^{*(j)} = (\boldsymbol{y}_{\text{ret}}, \boldsymbol{y}_{\text{del}}^{*(j)})$, $j = 1, ..., m$. Note that $f_{\boldsymbol{\theta}^*}(y_i|\boldsymbol{u}_i)$ in step (2) denotes the density $f_{\boldsymbol{\theta}}(y_i|\boldsymbol{u}_i)$ with the unknown $\boldsymbol{\theta}$ set equal to $\boldsymbol{\theta}^*$, the posterior draw from step (1).

Now we describe the PMID method. In the PMID method, the model is fit to the deleted data $\boldsymbol{y}_{\text{del}}$ instead of the complete data. Thus the posterior distribution of $\boldsymbol{\theta}$ is computed as $p(\boldsymbol{\theta}|\boldsymbol{y}_{\text{del}}, \boldsymbol{U}_{\text{del}}) \propto p(\boldsymbol{\theta}) \times \prod_{\{i:y_i \in \boldsymbol{y}_{\text{del}}\}} f_{\boldsymbol{\theta}}(y_i|\boldsymbol{u}_i)$, and synthetic data are generated as follows:

1. Draw $\boldsymbol{\theta}^*$ from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}_{\text{del}}, \boldsymbol{U}_{\text{del}})$.
2. For each i such that $y_i \in \boldsymbol{y}_{\text{del}}$, obtain the corresponding imputed value y_i^* as a draw from $f_{\boldsymbol{\theta}^*}(y_i|\boldsymbol{u}_i)$. Note that here we do not draw from the truncated distribution as in PMIC. Thus the imputed data $\boldsymbol{y}_{\text{del}}^*$ is the set of y_i^* values for each i such that $y_i \in \boldsymbol{y}_{\text{del}}$.

As usual, steps (1) and (2) above are repeated independently m times to get $\boldsymbol{y}_{\text{del}}^{*(1)}, ..., \boldsymbol{y}_{\text{del}}^{*(m)}$, and the synthetic data are then obtained as $\boldsymbol{y}^{*(j)} = (\boldsymbol{y}_{\text{ret}}, \boldsymbol{y}_{\text{del}}^{*(j)})$, $j = 1, ..., m$.

Using the synthetic data, inference is drawn on the (scalar) population quantity $\psi(\boldsymbol{\theta})$ as follows. Let $\eta = \eta(\boldsymbol{y})$ denote an estimator of $\psi(\boldsymbol{\theta})$ that is computed from the original data, and let $v = v(\boldsymbol{y})$ denote an estimator of the variance of η, also computed from the original data. For instance, η may be the maximum likelihood estimator of $\psi(\boldsymbol{\theta})$, and v may be the estimated asymptotic variance obtained from the inverse of the observed Fisher information matrix. Let $\boldsymbol{y}^{*(1)}, ..., \boldsymbol{y}^{*(m)}$ denote the m sets of multiply imputed data, based on PMID or PMIC, as described above. Given the synthetic data, one then proceeds to compute $\eta_j = \eta(\boldsymbol{y}^{*(j)})$ and $v_j = v(\boldsymbol{y}^{*(j)})$, the analogs of η and v, on the jth synthetic dataset, for $j = 1, ..., m$. As discussed by An and Little [5], the MI estimator of $\psi(\boldsymbol{\theta})$ is $\bar{\eta}_m = \frac{1}{m}\sum_{j=1}^{m} \eta_j$, and the estimator of the variance of $\bar{\eta}_m$ is $T_m = b_m/m + \bar{v}_m$, where $b_m = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \bar{\eta}_m)^2$ and $\bar{v}_m = \frac{1}{m}\sum_{j=1}^{m} v_j$. An approximate level $(1-\alpha)$ confidence interval for $\psi(\boldsymbol{\theta})$ is $\bar{\eta}_m \pm z_{\alpha/2} T_m^{1/2}$ where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the standard normal *cdf*.

We have described above the synthetic data methodology in the framework of a general density $f_\theta(y_i|\boldsymbol{u}_i)$. For our log-normal scenario, we have

$$f_{\boldsymbol{\theta}}(y_i|\boldsymbol{u}_i) = \frac{1}{y_i\sigma\sqrt{2\pi}} \exp\left[\frac{-(\ln y_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right], y_i > 0, \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2),$$

and we specify a standard diffuse prior on $\boldsymbol{\theta}$ as $p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}, \boldsymbol{\beta} \in \mathbb{R}^p$, $0 < \sigma^2 < \infty$. Under the PMIC method, the posterior distribution of $\boldsymbol{\theta}$ is obtained as $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{U}) = p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{U})p(\sigma^2|\boldsymbol{y}, \boldsymbol{U})$ where the distributions $p(\sigma^2|\boldsymbol{y}, \boldsymbol{U})$ and $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{U})$ are defined by [12]:

$$(\sigma^2|\boldsymbol{y}) \sim \frac{(n-p)\tilde{s}^2}{\chi^2_{n-p}} \text{ and } (\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}) \sim N_p\left(\tilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{V}\right), \tag{5}$$

where

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{u}_i\boldsymbol{u}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{u}_i \ln y_i\right), \qquad \boldsymbol{V} = \left(\sum_{i=1}^{n} \boldsymbol{u}_i\boldsymbol{u}_i'\right)^{-1},$$
$$\tilde{s}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(\ln y_i - \boldsymbol{u}_i'\tilde{\boldsymbol{\beta}}\right)^2.$$

Under the PMID method, the posterior distribution of $\boldsymbol{\theta}$ is $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}_{\text{del}}, \boldsymbol{U}_{\text{del}}) = p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}_{\text{del}}, \boldsymbol{U}_{\text{del}})p(\sigma^2|\boldsymbol{y}_{\text{del}}, \boldsymbol{U}_{\text{del}})$, where $p(\sigma^2|\boldsymbol{y}_{\text{del}}, \boldsymbol{U}_{\text{del}})$ and $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{U}_{\text{del}})$ are the distributions defined by

$$(\sigma^2|\boldsymbol{y}_{\text{del}}) \sim \frac{(n_{\text{del}}-p)\tilde{s}^2_{\text{del}}}{\chi^2_{n_{\text{del}}-p}} \text{ and } (\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}_{\text{del}}) \sim N_p\left(\tilde{\boldsymbol{\beta}}_{\text{del}}, \sigma^2\boldsymbol{V}_{\text{del}}\right), \tag{6}$$

where

$$\tilde{\boldsymbol{\beta}}_{\text{del}} = \left(\sum_{\{i\,:\,y_i \in \boldsymbol{y}_{\text{del}}\}} \boldsymbol{u}_i \boldsymbol{u}_i' \right)^{-1} \left(\sum_{\{i\,:\,y_i \in \boldsymbol{y}_{\text{del}}\}} \boldsymbol{u}_i \ln y_i \right), \quad \boldsymbol{V}_{\text{del}} = \left(\sum_{\{i\,:\,y_i \in \boldsymbol{y}_{\text{del}}\}} \boldsymbol{u}_i \boldsymbol{u}_i' \right)^{-1},$$

$$\tilde{s}^2_{\text{del}} = \frac{1}{n_{\text{del}} - p} \sum_{\{i\,:\,y_i \in \boldsymbol{y}_{\text{del}}\}} \left(\ln y_i - \boldsymbol{u}_i' \tilde{\boldsymbol{\beta}}_{\text{del}} \right)^2, \qquad n_{\text{del}} = \sum_{i=1}^{n} I(y_i > C_I).$$

4 Simulation Study to Assess Accuracy of Inference

In this section we report results of a simulation study. The purposes of the simulation study are to (1) evaluate the performance of the noise multiplication methods in finite samples, (2) evaluate how much the accuracy of inference is reduced as the dispersion of the noise generating distribution increases, (3) compare accuracy of inference of case (I) noise multiplied data with accuracy of inference of case (II) noise multiplied data, and (4) compare the proposed noise multiplication methods with top coding and synthetic data methods. The statistical computing software R [25] was used for all computations.

To conduct the simulation, we generate data under the following special case of model (1):

$$\ln y_i \sim N(\beta_0 + \beta_1 u_{i1},\ \sigma^2). \tag{7}$$

To generate data in the simulation, we set $\beta_0 = 1$, $\beta_1 = 1.5$, and $\sigma^2 = 1$. The covariates u_{i1} are generated independently (across i) from a $N(\delta = 0, \tau^2 = 1)$ distribution one time at the beginning of the simulation, and then held constant from one iteration to the next. The top coding threshold C is taken to be the 90th percentile of a log-normal distribution having log-scale mean $\beta_0 + \beta_1\delta$ and log-scale variance $\sigma^2 + \tau^2\beta_1^2$. That is, C is the 90th percentile of the marginal distribution of y_i, obtained by integrating out the covariate u_{1i}; hence in any particular iteration of the simulation, approximately 10% of the y_i values will exceed C.

The noise generating distribution is taken as

$$h(r) = \begin{cases} \gamma(\xi_2 - \xi_1)^{-1}, & \text{if } \xi_1 \le r \le \xi_2, \\ (1-\gamma)(\xi_4 - \xi_3)^{-1}, & \text{if } \xi_3 \le r \le \xi_4, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $0 < \xi_1 < \xi_2 < 1 < \xi_3 < \xi_4 < \infty$ and $0 \le \gamma \le 1$. Notice that this $h(r)$ is simply a mixture of the Uniform(ξ_1, ξ_2) and Uniform(ξ_3, ξ_4) distributions, with mixing probability γ, and its mean is

$$\frac{1}{2}(\xi_1 + \xi_2)\gamma + \frac{1}{2}(\xi_3 + \xi_4)(1 - \gamma),$$

and variance is

$$\frac{1}{12}(\xi_2 - \xi_1)^2\gamma + \frac{1}{12}(\xi_3 - \xi_4)^2(1 - \gamma) + \frac{1}{4}[(\xi_1 + \xi_2) - (\xi_3 + \xi_4)]^2\gamma(1 - \gamma).$$

The noise distribution (8) is quite flexible, and has the desirable feature that it is a nonnegative distribution, but has no mass in the interval (ξ_2, ξ_3) which contains 1. Hence, the noise multiplier r_i generated from this distribution is always bounded away from 1, guaranteeing that the relative distance between the noise multiplied value $z_i = r_i y_i$ and the original value y_i, defined by $\left|\frac{z_i - y_i}{y_i}\right| = |r_i - 1|$, is bounded away from zero. Thus by choosing the parameters of this noise generating distribution appropriately, the data producer can guarantee a specified relative distance between the noise perturbed value and the original value. Furthermore, note that the density (8) does not necessarily have mean 1, and the lengths of the two intervals (ξ_1, ξ_2) and (ξ_3, ξ_4) could be quite different, depending on how the data producer wants to perturb the data. The mixing parameter γ also gives the data producer control over how often noise multiplication will deflate or inflate the original value. For example, by taking γ close (equal) to 1, the data producer would ensure that the perturbed values are usually (always) deflated versions of the original confidential value. For the simulation study, we consider the following four settings for the parameters in (8). (In the following, the numbers in parentheses are the mean and variance, respectively, of the corresponding noise generating distribution.)

$$\begin{aligned}
&\text{Setting h1}: \ \xi_1 = 0.8,\ \xi_2 = 0.9,\ \xi_3 = 1.1,\ \xi_4 = 1.2,\ \gamma = 0.5,\ (1.000, 0.023).\\
&\text{Setting h2}: \ \xi_1 = 0.5,\ \xi_2 = 0.9,\ \xi_3 = 1.1,\ \xi_4 = 1.5,\ \gamma = 0.8,\ (0.820, 0.071).\\
&\text{Setting h3}: \ \xi_1 = 0.5,\ \xi_2 = 0.9,\ \xi_3 = 1.1,\ \xi_4 = 1.5,\ \gamma = 0.5,\ (1.000, 0.103).\\
&\text{Setting h4}: \ \xi_1 = 0.1,\ \xi_2 = 0.8,\ \xi_3 = 1.2,\ \xi_4 = 1.5,\ \gamma = 0.8,\ (0.630, 0.164).
\end{aligned} \tag{9}$$

The simulation results are reported in Tables 1–4. We report simulation results in the case that the parameter of interest is β_1, the slope of the regression line, and the case that the parameter of interest is σ^2, the residual variance. Tables 1 and 2 show results for inference on β_1, and Tables 3 and 4 show results for inference on σ^2. Tables 1 and 3 show results for sample sizes $n = 200$ and 500, and Tables 2 and 4 show results for sample sizes $n = 1000$ and 1500. These tables display results for the following methods.

UD: Analysis based on unperturbed data, without any masking.

TC: Analysis based on top coded data using the Tobit model, as described in Section 3.1. For fitting the Tobit model to the top-coded data, we use the R function `tobit` from the R package `AER` [17], which provides the parameter estimates along with their estimated covariance matrix.

PMIC2, PMIC4, PMID2, and PMID4: Analysis based on the PMIC and PMID methods described in Section 3.2. Let n_S denote the number of values in the sample that exceed the top coding threshold C. For PMIC2 and PMID2, C_I is chosen so that $2n_S$ values in the sample are larger than C_I. For PMIC4 and PMID4, C_I is chosen so that $4n_S$ values in the sample are larger than C_I.

NMh1.i, NMh2.i, NMh3.i, NMh4.i: Each of these four rows indicate analysis based on case (I) type noise multiplied data as presented in Section 2 with $h(r)$ taken as the

Table 1: Inference for the slope β_1 for $n = 200$ and 500.

n	Method	RMSE $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
200	UD	69.4	69.4	68.8	94.4	1.000
	TC	75.7	75.7	75.4	94.8	1.095
	PMIC2 ($m = 50$)	74.8	74.8	69.1	93.0	1.004
	PMIC4 ($m = 50$)	74.4	74.4	69.5	93.4	1.010
	PMID2 ($m = 50$)	69.6	69.6	69.5	94.9	1.010
	PMID4 ($m = 50$)	69.8	69.8	69.7	94.7	1.012
	NMh1.i	69.8	69.8	69.0	94.4	1.003
	NMh1.ii	69.8	69.8	69.0	94.4	1.003
	NMh2.i	70.3	70.3	69.5	94.8	1.010
	NMh2.ii	70.6	70.6	69.7	94.5	1.012
	NMh3.i	70.4	70.4	69.6	94.6	1.012
	NMh3.ii	70.6	70.6	69.8	94.5	1.014
	NMh4.i	71.5	71.5	71.2	94.5	1.034
	NMh4.ii	74.3	74.3	74.5	95.0	1.082
500	UD	43.9	43.9	43.5	94.2	1.000
	TC	47.4	47.4	47.2	94.8	1.086
	PMIC2 ($m = 50$)	47.0	47.0	43.7	92.7	1.004
	PMIC4 ($m = 50$)	47.6	47.6	43.9	92.7	1.008
	PMID2 ($m = 50$)	44.1	44.1	43.8	94.3	1.006
	PMID4 ($m = 50$)	44.2	44.2	43.9	94.4	1.009
	NMh1.i	44.1	44.1	43.6	94.0	1.003
	NMh1.ii	44.1	44.1	43.6	94.1	1.003
	NMh2.i	44.3	44.3	44.0	94.3	1.010
	NMh2.ii	44.4	44.4	44.1	94.4	1.013
	NMh3.i	44.3	44.3	44.0	94.3	1.012
	NMh3.ii	44.4	44.4	44.1	94.2	1.014
	NMh4.i	45.2	45.2	44.9	94.5	1.032
	NMh4.ii	47.6	47.6	47.2	94.2	1.085

Table 2: Inference for the slope β_1 for $n = 1000$ and 1500.

n	Method	RMSE $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
1000	UD	31.4	31.4	31.2	94.6	1.000
	TC	33.8	33.8	33.7	94.9	1.081
	PMIC2 ($m = 50$)	33.6	33.6	31.3	92.9	1.003
	PMIC4 ($m = 50$)	34.0	34.0	31.4	93.0	1.006
	PMID2 ($m = 50$)	31.6	31.6	31.3	94.5	1.005
	PMID4 ($m = 50$)	31.6	31.6	31.4	94.9	1.007
	NMh1.i	31.4	31.4	31.3	94.8	1.003
	NMh1.ii	31.5	31.5	31.3	94.7	1.003
	NMh2.i	31.7	31.7	31.5	94.9	1.009
	NMh2.ii	31.9	31.9	31.6	94.8	1.012
	NMh3.i	31.7	31.7	31.5	95.0	1.011
	NMh3.ii	31.8	31.8	31.6	94.9	1.013
	NMh4.i	32.3	32.3	32.1	94.8	1.030
	NMh4.ii	34.0	34.0	33.7	95.1	1.082
1500	UD	26.5	26.5	25.9	94.7	1.000
	TC	28.5	28.5	28.2	95.1	1.088
	PMIC2 ($m = 50$)	28.4	28.4	26.0	92.7	1.003
	PMIC4 ($m = 50$)	28.6	28.6	26.1	92.5	1.006
	PMID2 ($m = 50$)	26.6	26.6	26.1	94.4	1.004
	PMID4 ($m = 50$)	26.7	26.7	26.1	94.5	1.006
	NMh1.i	26.6	26.6	26.0	94.7	1.003
	NMh1.ii	26.6	26.6	26.0	94.7	1.003
	NMh2.i	26.6	26.6	26.2	94.6	1.010
	NMh2.ii	26.7	26.7	26.3	94.7	1.012
	NMh3.i	26.7	26.7	26.2	95.0	1.012
	NMh3.ii	26.8	26.8	26.3	94.7	1.014
	NMh4.i	27.1	27.1	26.8	95.1	1.032
	NMh4.ii	28.5	28.5	28.1	94.7	1.083

Table 3: Inference for the residual variance σ^2 for $n = 200$ and $n = 500$.

n	Method	RMSE $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
200	UD	100.2	99.5	98.8	93.5	1.000
	TC	107.5	106.9	106.5	93.4	1.079
	PMIC2 ($m = 50$)	107.8	107.4	99.5	91.6	1.007
	PMIC4 ($m = 50$)	109.4	109.3	100.4	92.0	1.017
	PMID2 ($m = 50$)	100.8	100.8	100.8	94.4	1.020
	PMID4 ($m = 50$)	101.5	101.4	101.3	94.3	1.026
	NMh1.i	100.5	99.8	99.2	93.5	1.004
	NMh1.ii	100.6	99.8	99.2	93.4	1.005
	NMh2.i	102.0	101.3	100.2	93.2	1.015
	NMh2.ii	102.5	101.8	100.6	93.0	1.019
	NMh3.i	101.8	101.1	100.3	93.2	1.015
	NMh3.ii	102.2	101.4	100.5	93.3	1.018
	NMh4.i	103.8	103.2	102.5	93.5	1.037
	NMh4.ii	108.6	107.9	106.5	93.1	1.078
500	UD	62.5	62.4	63.0	95.0	1.000
	TC	66.7	66.6	67.5	94.8	1.071
	PMIC2 ($m = 50$)	67.0	67.0	63.3	93.1	1.004
	PMIC4 ($m = 50$)	69.1	69.1	63.6	92.7	1.010
	PMID2 ($m = 50$)	63.0	62.9	63.7	95.6	1.011
	PMID4 ($m = 50$)	63.4	63.4	64.0	95.4	1.015
	NMh1.i	62.8	62.7	63.2	94.9	1.004
	NMh1.ii	62.9	62.7	63.2	95.0	1.004
	NMh2.i	63.3	63.2	63.9	95.1	1.013
	NMh2.ii	63.6	63.5	64.1	95.2	1.017
	NMh3.i	63.3	63.2	63.8	95.2	1.013
	NMh3.ii	63.4	63.3	64.0	95.2	1.016
	NMh4.i	64.6	64.5	65.1	94.8	1.033
	NMh4.ii	67.5	67.3	67.6	94.8	1.072

Table 4: Inference for the residual variance σ^2 for $n = 1000$ and 1500.

n	Method	RMSE $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
1000	UD	44.8	44.7	44.6	94.7	1.000
	TC	47.7	47.6	47.6	94.6	1.066
	PMIC2 ($m = 50$)	47.8	47.8	44.8	93.2	1.003
	PMIC4 ($m = 50$)	48.6	48.6	44.9	93.0	1.007
	PMID2 ($m = 50$)	45.1	45.1	45.0	95.1	1.007
	PMID4 ($m = 50$)	45.2	45.2	45.1	94.6	1.011
	NMh1.i	44.9	44.8	44.8	94.7	1.004
	NMh1.ii	44.9	44.9	44.8	94.8	1.004
	NMh2.i	45.3	45.2	45.2	94.7	1.013
	NMh2.ii	45.5	45.5	45.4	94.4	1.016
	NMh3.i	45.4	45.4	45.2	94.6	1.013
	NMh3.ii	45.5	45.4	45.3	94.6	1.015
	NMh4.i	46.0	45.9	46.0	94.7	1.031
	NMh4.ii	47.6	47.5	47.7	95.2	1.069
2000	UD	36.0	36.0	36.5	95.0	1.000
	TC	38.4	38.4	38.7	95.1	1.062
	PMIC2 ($m = 50$)	38.7	38.7	36.6	93.1	1.003
	PMIC4 ($m = 50$)	39.7	39.7	36.7	93.3	1.006
	PMID2 ($m = 50$)	36.2	36.2	36.7	95.4	1.006
	PMID4 ($m = 50$)	36.3	36.3	36.8	95.3	1.009
	NMh1.i	36.1	36.1	36.6	95.0	1.003
	NMh1.ii	36.1	36.1	36.6	95.0	1.004
	NMh2.i	36.5	36.5	36.9	95.1	1.012
	NMh2.ii	36.7	36.7	37.0	95.2	1.016
	NMh3.i	36.5	36.5	36.9	95.1	1.012
	NMh3.ii	36.6	36.6	37.0	95.2	1.015
	NMh4.i	37.2	37.2	37.5	95.2	1.029
	NMh4.ii	38.7	38.7	38.9	95.3	1.067

mixture density defined in (8). For NMh1.i the parameters in $h(r)$ are set to the values of Setting h1 in (9); for NMh2.i the parameters in $h(r)$ are set to the values of Setting h2 in (9), and so on.

NMh1.ii, NMh2.ii, NMh3.ii, NMh4.ii: Each of these four rows indicate analysis based on case (II) type noise multiplied data as presented in Section 2 with $h(r)$ taken as the mixture density defined in (8). For NMh1.ii the parameters in $h(r)$ are set to the values of Setting h1 in (9); for NMh2.ii the parameters in $h(r)$ are set to the values of Setting h2 in (9), and so on.

For the estimators of parameters β_1 and σ^2 under each of the methods, the following quantities were estimated by Monte Carlo simulation based on 5000 iterations: the root mean squared error (RMSE), standard deviation (SD), expected value of standard deviation estimator ($\widehat{\text{SD}}$), coverage probability of the nominal level 0.95 confidence interval (Cvg.), and expected length of the confidence interval relative to the expected length of the confidence interval computed on the unperturbed data (Rel. Len.). For estimating standard deviation, we use the square root of the appropriate variance estimator; and for confidence intervals, we always take the nominal level as 0.95. To facilitate a comparison of the methods, all results shown for unperturbed data are based on MLEs, observed Fisher information, and confidence intervals of the form (MLE) $\pm$ (1.96 $\times$ estimated standard deviation). For EM algorithms used under noise multiplication, the stopping criterion used was $\max\left\{\left|\beta_0^{(t)}-\beta_0^{(t+1)}\right|,\left|\beta_1^{(t)}-\beta_1^{(t+1)}\right|,\left|(\sigma^{(t)})^2-(\sigma^{(t+1)})^2\right|\right\} \leq 10^{-5}$. To get starting values of parameter estimates to use in the EM algorithms, we run a linear regression of $\ln x_i$ on $\boldsymbol{u}_i$. As mentioned above, the statistical computing software R [25] was used for the computations. The `integrate` function in R was used to evaluate the required univariate integrals that could not be obtained in closed form. For methods PMIC2, PMIC4, PMID2, and PMID4, we used $m = 50$ sets of imputed values, and model (7) was used to generate the imputed values. The choice $m = 50$ may be larger than what is often used in practice, but we chose a large value in order to get a clear picture of accuracy of inference of the multiple imputation method. Furthermore, we refer to Reiter [29] for a discussion of issues that arise when the model used for generating synthetic data differs from the model used for data analysis. The following is a summary of the findings from our simulation study.

1. In terms of RMSE and SD of the point estimators, the noise multiplication methods give reasonable results in all of the simulation scenarios we considered. In each of the simulation settings of Tables 1–4, the RMSE and SD under noise multiplication are similar to and just slightly larger than the RMSE and SD based on the unperturbed data. Also, in all cases of noise multiplication considered, the SD estimators are nearly unbiased for the true SD. Notice that the methods NMh2.i, NMh2.ii, NMh4.i, and NMh4.ii have a noise generating distribution with mean not equal to 1 (refer to (9)), yet all results remain valid for these methods.

2. Comparing noise settings h1 and h2, we notice that the RMSE, SD, and relative confidence interval length are larger for h2 than for h1. This finding is expected

since the noise distribution h2 has greater dispersion than h1. A similar conclusion holds if one compares noise settings h2 with h4, or h3 with h4. The noise settings h2 and h3 are very similar, the only difference being the value of the mixing probability γ, and the simulation results are nearly equivalent.

3. In Tables 1 and 2, we see that for inference on β_1 under noise multiplication, the coverage probability of confidence intervals is quite close to the nominal value of 0.95. However, for inference on σ^2, we see in Table 3 that under noise multiplication, the coverage probability is slightly below the nominal value when $n = 200$. However, in this case of $n = 200$, we see that the confidence interval for σ^2 based on the unperturbed data (which is also a Wald-type interval as described above) is also below the nominal value because the sample size is too small for the sampling distribution of the MLE of σ^2 to be well approximated by normality. Thus, even though the noise multiplication method yields a confidence interval with slightly low coverage probability here, it gives an inference which is quite similar to the inference obtained using the unperturbed data. For the larger sample sizes $n = 500$ (right hand panel of Table 3), and $n = 1000$ and 1500 (Table 4) we see that both the noise multiplication methods and the unperturbed data yield confidence intervals for σ^2 with coverage probability very close to the nominal level of 0.95.

4. If we observe noise multiplied data under case (I) of Section 2, then a top coded sample can be re-constructed from the observed data. However, a case (I) noise multiplied sample cannot be constructed from the top coded data. So in this sense, case (I) noise multiplied data carry more information than top coded data. The simulation results confirm this statement because the case (I) noise multiplication method generally leads to a shorter confidence interval and smaller SD than top coding for all noise distributions that we considered in Tables 1–4.

5. Clearly, we cannot re-construct a case (II) noise multiplied sample (as defined in Section 2) based on only the top coded data. Likewise, if we observe a case (II) noise multiplied data and the noise density is (8), the complete top coded sample cannot be deterministically re-constructed from these observed data. In fact, considering that the range of possible x_i values is $(0, \infty)$, the following cases are possible:

   (a) if $0 < x_i < C\xi_1$, then we know x_i is not noise perturbed (i.e., $x_i = y_i$);

   (b) if $C\xi_1 < x_i < C$, then x_i is either not noise perturbed or perturbed by a noise multiplier in the interval (ξ_1, ξ_2);

   (c) if $C < x_i < C\xi_3$, then x_i is definitely noise perturbed by a noise multiplier in the interval (ξ_1, ξ_2);

   (d) if $C\xi_3 < x_i$, then x_i is definitely noise perturbed and the noise multiplier may be either in the interval (ξ_1, ξ_2) or (ξ_3, ξ_4).

   Hence, in case (a) above, $\Pr(\Delta_i = 1|x_i) = 1$; in case (b), $0 < \Pr(\Delta_i = 1|x_i) < 1$; and

in cases (c) and (d), $\Pr(\Delta_i = 1|x_i) = 0$. Recall that by (4),

$$\Pr(\Delta_i = 1|x_i) = \frac{I(x_i < C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right]}{I(x_i < C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right] + \int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/\omega) - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right]h(\omega)d\omega}, \quad (10)$$

and therefore in case (b), while we would not know for sure, it may be possible to make an informed guess as to whether or not x_i is noise perturbed based on the value of $\Pr(\Delta_i = 1|x_i)$ (which could be estimated by replacing unknown parameters by their estimates). So we would usually expect a case (II) noise multiplied sample to carry more information than a top coded sample, but since there can be cases where a large number of x_i values fall in case (b) above, and $\Pr(\Delta_i = 1|x_i)$ is not close to 0 or 1, there can be scenarios where a case (II) noise multiplied sample carries less information than a top coded sample (with the trade off likely being an enhanced level of privacy protection). The simulation results confirm this statement; for most of the noise settings we considered, the case (II) noise multiplication scenario leads to a shorter confidence interval and smaller SD than top coding. But, in several of the scenarios considered, we observe that under method NMh4.ii, where a large number of x_i values can fall in case (b) because the interval $C\xi_1 < x_i < C$ is fairly large (here $\xi_1 = 0.1$), the case (II) noise multiplied sample yields RMSE and SD slightly larger than top coding.

6. As expected, the case (I) noise multiplied data always yield more or equally accurate inference than case (II) noise multiplied data. For instance, in Setting h1, where the variance of the noise generating distribution is small, the results in Tables 1–4 between methods NMh1.i and NMh1.ii are nearly identical. But in Setting h4, where the dispersion of the noise generating distribution is much larger, the method NMh4.i is noticeably more accurate than NMh4.ii.

7. In Tables 1–4, we observe that for inference on both β_1 and σ^2, the methods PMIC2 and PMIC4 yield confidence intervals with coverage probability below the nominal level, even for the larger sample sizes $n = 1000$ and 1500. On the other hand, the methods PMID2 and PMID4 tend to yield confidence intervals having coverage probability close to the nominal level in each of the chosen simulation settings. The PMID methods also tend to give accurate inference in terms of RMSE, SD, $\widehat{\text{SD}}$, and relative confidence interval length. In each case, it would be possible to select a noise generating distribution for which the noise multiplication method has similar accuracy to an MI method.

8. The EM algorithms used to compute the MLE under noise multiplication tended to be stable and to converge rapidly.

5 Data Analysis Illustration and Disclosure Risk Evaluation Using Current Population Survey Data

In this section we present an application based on public use data from the 2000 Current Population Survey (CPS) March supplement. These data are available online from `http://www.census.gov/cps/`, and have been used previously by Reiter [31, 32] and Drechsler and Reiter [9] for illustrating various aspects of multiple imputation methodology. The entire data comprise household, family, and individual records. For our illustration, we focus on the household records, as did Reiter [31, 32] and Drechsler and Reiter [9]. The data file contains records on 51,016 households, and of these households, 50,661 of them have positive household income.

In Section 5.1, we present an example of the inferences on regression parameters obtained using our proposed noise multiplication methodology. Furthermore, we compare the inferences obtained under noise multiplication with those obtained on the original data, and also with those obtained based on the statistical disclosure control methods of top coding and synthetic data. In Section 5.2, we provide a disclosure risk evaluation of our proposed noise multiplication method, and compare the disclosure risk of noise multiplication with that of synthetic data. We proceed as if the $n = 50{,}661$ households with positive income are a random sample, and as if the household income value is confidential for the high income households (of course, in reality, these are public use data). There are additional available variables on these households that can be used as covariates. After some exploratory analysis to determine which covariates may serve as good predictors of household income, we choose to include the following covariates:

P: household property tax,

N: number of people in household,

L: number of people in the household who are less than 18 years old,

A: age for the head of the household,

E: education level for the head of the household (coded to take values 31–46),

M: martial status for the head of the household (coded to take values 1–7),

R: race for the head of the household (coded to take values 1–4),

S: sex for the head of the household (coded to take values 1–2).

We refer to the Current Population Survey March 2000 technical documentation (available at `http://www.census.gov/prod/techdoc/cps/cpsmar00.pdf`) for the meaning of the coding of the variables E, M, R, and S, and for additional information about the dataset. In the notation of Section 2, the variable y is total income for the household,

and $\boldsymbol{u}$, the vector of regressors, includes the following variables:

$$\Big\{1,\ \mathrm{P},\ \mathrm{N},\ \mathrm{L},\ \mathrm{A},\ I(\mathrm{E}{=}32), I(\mathrm{E}{=}33)\ldots, I(\mathrm{E}{=}46),$$
$$I(\mathrm{M}{=}2), I(\mathrm{M}{=}3), \ldots, I(\mathrm{M}{=}7),\ I(\mathrm{R}{=}2), I(\mathrm{R}{=}3), I(\mathrm{R}{=}4),\ I(\mathrm{S}{=}2)\Big\}, \tag{11}$$

where $I(\mathrm{E}{=}32)$ is an indicator for E=32, $I(\mathrm{E}{=}33)$ is an indicator for E=33, and so on. The model matrix $\begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_n' \end{pmatrix}$ has $n = 50{,}661$ rows and $p = 30$ columns, and has full column rank.

5.1 Data Analysis Illustration

For the data analysis, we report results based on unperturbed data, top coded data, synthetic data, and noise multiplied data. Specifically, we analyze the data using the methods UD, TC, PMIC2, PMIC4, PMID2, PMID4, NMh1.i, NMh1.ii, NMh2.i, NMh2.ii, NMh3.i, NMh3.ii, NMh4.i, and NMh4.ii, as defined in Section 4. Results of the data analysis appear in Tables 5 and 6. The top coding threshold C is taken as the empirical 90th percentile of the household income variable. For the synthetic data methods, the regression model used for imputation includes all the variables in (11), and hence is the same as the model used for data analysis. Table 5 displays the parameter estimates in the regression of logarithm of household income on the variables in (11). The rows of this table give the estimates of the various parameters (row 1 gives the estimate of the intercept, row 2 gives the estimate of the regression coefficient for P, row 3 gives the estimate of the regression coefficient for N, and so on), and the columns correspond to the different methods defined previously. Similarly, Table 6 displays the estimated standard deviation of each parameter estimator under each method. The data analysis is conducted in R [25] using the computational methods discussed in Section 4. Furthermore, the convergence criterion and starting values for EM algorithms, and the number of multiple imputations used are the same as in Section 4. The following is a summary of the findings of the data analysis.

1. We notice in Tables 5 and 6, that when the variance of the noise generating distribution is fairly small such as in Setting h1, the noise multiplication methods give almost identical results as the unperturbed data. For all noise generating distributions considered, the results under noise multiplication tend to be in line with those obtained for the unperturbed data. Naturally, the standard deviation estimates based on noise multiplied data tend to increase when dispersion in the noise generating distribution increases.

2. In almost all cases, the noise multiplied data yield smaller standard deviation estimates than top coded data. In fact, in all cases, the case (I) noise multiplied data

Table 5: Estimates of parameters in the CPS data example.

			$m = 50$											
			PMI	PMI	PMI	PMI	NM	NM	NM	NM	NM	NM	NM	NM
Parameter	UD	TC	C2	C4	D2	D4	h1.i	h1.ii	h2.i	h2.ii	h3.i	h3.ii	h4.i	h4.ii
Intercept	9.53	9.47	9.48	9.54	9.53	9.53	9.53	9.53	9.52	9.53	9.52	9.53	9.50	9.53
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.36	0.39	0.40	0.42	0.36	0.36	0.36	0.36	0.37	0.37	0.37	0.37	0.38	0.40
L	-0.37	-0.40	-0.42	-0.43	-0.37	-0.37	-0.37	-0.37	-0.38	-0.38	-0.38	-0.38	-0.39	-0.42
A	-0.00	-0.00	-0.01	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01
I(E=32)	-0.05	-0.06	-0.06	-0.08	-0.05	-0.05	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.08
I(E=33)	0.02	0.02	0.01	-0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
I(E=34)	0.10	0.11	0.11	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
I(E=35)	0.15	0.15	0.15	0.13	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
I(E=36)	0.19	0.20	0.20	0.18	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.18
I(E=37)	0.21	0.22	0.22	0.21	0.21	0.21	0.21	0.22	0.22	0.22	0.22	0.22	0.22	0.21
I(E=38)	0.31	0.31	0.31	0.30	0.31	0.31	0.31	0.31	0.31	0.32	0.31	0.31	0.31	0.31
I(E=39)	0.56	0.57	0.58	0.57	0.56	0.56	0.56	0.57	0.57	0.57	0.57	0.57	0.57	0.58
I(E=40)	0.75	0.76	0.77	0.77	0.75	0.75	0.75	0.75	0.76	0.76	0.76	0.76	0.76	0.77
I(E=41)	0.80	0.81	0.83	0.82	0.79	0.80	0.80	0.80	0.80	0.81	0.80	0.80	0.81	0.83
I(E=42)	0.91	0.93	0.95	0.94	0.91	0.90	0.91	0.91	0.91	0.92	0.91	0.92	0.92	0.93
I(E=43)	1.10	1.14	1.15	1.13	1.10	1.10	1.11	1.11	1.11	1.12	1.11	1.12	1.12	1.11
I(E=44)	1.29	1.35	1.36	1.35	1.29	1.29	1.29	1.30	1.31	1.31	1.31	1.31	1.33	1.30
I(E=45)	1.50	1.57	1.53	1.44	1.50	1.50	1.50	1.50	1.51	1.49	1.51	1.51	1.54	1.43
I(E=46)	1.47	1.53	1.51	1.47	1.47	1.47	1.47	1.48	1.48	1.48	1.47	1.47	1.50	1.40
I(M=2)	-0.06	-0.05	-0.04	-0.03	-0.06	-0.05	-0.06	-0.06	-0.06	-0.06	-0.05	-0.05	-0.05	0.01
I(M=3)	-0.30	-0.30	-0.31	-0.31	-0.30	-0.30	-0.31	-0.30	-0.31	-0.31	-0.31	-0.31	-0.30	-0.30
I(M=4)	-0.31	-0.29	-0.28	-0.28	-0.31	-0.31	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.29
I(M=5)	-0.21	-0.20	-0.21	-0.21	-0.21	-0.21	-0.21	-0.21	-0.21	-0.21	-0.21	-0.21	-0.21	-0.20
I(M=6)	-0.41	-0.40	-0.40	-0.42	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.40	-0.41	-0.40	-0.41
I(M=7)	-0.41	-0.40	-0.42	-0.43	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.42
I(R=2)	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13
I(R=3)	-0.24	-0.24	-0.24	-0.26	-0.23	-0.24	-0.24	-0.24	-0.24	-0.24	-0.24	-0.24	-0.24	-0.25
I(R=4)	-0.11	-0.11	-0.12	-0.14	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.13
I(S=2)	-0.12	-0.12	-0.13	-0.14	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.14
σ^2	0.62	0.65	0.67	0.71	0.63	0.63	0.63	0.63	0.63	0.65	0.63	0.64	0.64	0.69

Table 6: Estimated standard deviations, multiplied by 100, of parameter estimators in the CPS data example.

			$m = 50$											
			PMI	PMI	PMI	PMI	NM	NM	NM	NM	NM	NM	NM	NM
Parameter	UD	TC	C2	C4	D2	D4	h1.i	h1.ii	h2.i	h2.ii	h3.i	h3.ii	h4.i	h4.ii
Intercept	6.25	6.43	6.49	6.66	6.25	6.26	6.26	6.28	6.30	6.38	6.29	6.35	6.35	6.84
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.53	0.57	0.55	0.57	0.53	0.53	0.53	0.53	0.54	0.55	0.54	0.55	0.55	0.66
L	0.65	0.69	0.67	0.69	0.65	0.65	0.65	0.65	0.66	0.67	0.66	0.67	0.67	0.79
A	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
I(E=32)	6.66	6.83	6.91	7.09	6.66	6.67	6.68	6.69	6.71	6.79	6.70	6.76	6.75	7.21
I(E=33)	6.28	6.44	6.52	6.69	6.28	6.29	6.29	6.30	6.32	6.40	6.32	6.37	6.37	6.80
I(E=34)	6.07	6.22	6.30	6.46	6.07	6.08	6.08	6.09	6.11	6.18	6.11	6.16	6.15	6.56
I(E=35)	6.23	6.39	6.47	6.64	6.24	6.24	6.25	6.26	6.28	6.35	6.27	6.32	6.32	6.75
I(E=36)	6.15	6.30	6.38	6.55	6.15	6.16	6.16	6.17	6.19	6.26	6.19	6.24	6.23	6.65
I(E=37)	6.13	6.28	6.36	6.53	6.13	6.14	6.14	6.15	6.17	6.24	6.17	6.22	6.21	6.63
I(E=38)	6.64	6.81	6.90	7.08	6.65	6.65	6.66	6.67	6.69	6.77	6.68	6.74	6.74	7.22
I(E=39)	5.84	5.99	6.06	6.22	5.85	5.85	5.86	5.87	5.88	5.95	5.88	5.93	5.92	6.32
I(E=40)	5.87	6.02	6.09	6.25	5.87	5.88	5.88	5.89	5.91	5.98	5.91	5.95	5.95	6.36
I(E=41)	6.06	6.22	6.29	6.46	6.07	6.07	6.08	6.09	6.11	6.18	6.10	6.15	6.15	6.60
I(E=42)	6.12	6.28	6.35	6.51	6.12	6.12	6.13	6.14	6.16	6.24	6.15	6.21	6.20	6.67
I(E=43)	5.88	6.03	6.10	6.26	5.88	5.89	5.89	5.91	5.92	5.99	5.92	5.97	5.97	6.38
I(E=44)	6.00	6.16	6.22	6.38	6.00	6.00	6.01	6.02	6.04	6.12	6.04	6.09	6.09	6.55
I(E=45)	6.47	6.77	6.72	6.90	6.48	6.48	6.49	6.50	6.54	6.63	6.54	6.61	6.63	7.29
I(E=46)	6.61	6.91	6.86	7.05	6.61	6.62	6.63	6.64	6.68	6.78	6.68	6.75	6.77	7.46
I(M=2)	7.35	7.63	7.63	7.84	7.35	7.37	7.37	7.38	7.41	7.56	7.42	7.52	7.51	8.83
I(M=3)	2.94	3.03	3.05	3.13	2.94	2.95	2.95	2.96	2.97	3.01	2.96	2.99	2.99	3.25
I(M=4)	1.48	1.53	1.54	1.58	1.48	1.48	1.48	1.49	1.49	1.51	1.49	1.51	1.51	1.66
I(M=5)	1.19	1.23	1.24	1.27	1.19	1.19	1.19	1.20	1.20	1.22	1.20	1.21	1.21	1.36
I(M=6)	2.21	2.28	2.30	2.36	2.21	2.22	2.22	2.22	2.23	2.26	2.23	2.25	2.25	2.43
I(M=7)	1.18	1.22	1.22	1.25	1.18	1.18	1.18	1.18	1.19	1.21	1.19	1.20	1.20	1.35
I(R=2)	1.19	1.23	1.24	1.27	1.19	1.20	1.20	1.20	1.20	1.22	1.20	1.22	1.21	1.33
I(R=3)	3.31	3.41	3.44	3.53	3.31	3.32	3.32	3.32	3.34	3.38	3.33	3.37	3.36	3.66
I(R=4)	2.12	2.23	2.20	2.26	2.12	2.12	2.13	2.13	2.15	2.19	2.15	2.18	2.18	2.50
I(S=2)	0.78	0.80	0.80	0.83	0.78	0.78	0.78	0.78	0.78	0.80	0.78	0.79	0.79	0.89
σ^2	0.39	0.44	0.42	0.45	0.39	0.39	0.40	0.40	0.40	0.42	0.40	0.41	0.42	0.50

yield smaller standard deviation estimates than top coded data. but we see that the standard deviation estimates under NMh4.ii tend to be larger than those under top coding; the reasons for these phenomena are explained in Section 4 (refer to points 4 and 5 in the discussion of simulation results in Section 4).

3. We note from Table 6 that the PMIC2 and PMIC4 methods yield large standard deviations in comparison with most of the other methods including top coding. On the other hand, the PMID methods give standard deviation estimates that are similar to the estimates based on the unperturbed data. Similarly, we notice in Table 5 that the parameter estimates under the PMIC methods often differ from the estimates based on the unperturbed data, while the PMID methods give estimates more in line with the unperturbed data.

In summary, the PMID2, PMID4, and noise multiplication methods all appear to give reasonable results, and can usually offer increased accuracy compared to top-coding. The PMID2 and PMID4 methods appear to give quite accurate inferences. The noise multiplication methods with low dispersion in the noise generating distribution provide results almost identical as the unperturbed data. The case (I) noise multiplied data provide more accurate inference than top coding in all cases.

5.2 Disclosure Risk Evaluation

In this section we report the results of a disclosure risk evaluation in the context of the CPS data example. Our main purpose is to evaluate the level of disclosure protection offered by the proposed noise multiplication method, and to compare noise multiplication with the synthetic data methods PMIC and PMID on the basis of disclosure risk. We have not included top coding in this study because top coding only reports if a particular y_i value exceeds C, but provides no further information about this value. While evaluation of disclosure risk when releasing any kind of perturbed data is essential, there is no unique way to exactly quantify it or even define it. Reiter [33], Reiter and Mitra [34], Klein and Sinha [19] and others have attempted it earlier with their own approaches and understandings. In this section we take the following premise.

Having observed the released privacy protected data (e.g., case (I) noise multiplied data, case (II) noise multiplied data, or synthetic data), let $\hat{y}_i$ denote an intruder's estimate of the confidential target value y_i. We assume that $\hat{y}_i$ is some function of the data that are released after a method of statistical disclosure control has been applied to the y_i values that exceed C. We measure the level of privacy protection using the following criterion, which is similar to the one used by Lin and Wise [21]:

$$p_{i,\epsilon} = \Pr\left\{ \left| \frac{\hat{y}_i - y_i}{y_i} \right| \leq \epsilon \,\middle|\, \boldsymbol{y} \right\}. \tag{12}$$

In the above probability, $\epsilon > 0$, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the vector of original observations, and because we have conditioned on $\boldsymbol{y}$, the above probability is free of the unknown parameters $\boldsymbol{\beta}$ and σ^2. For some reasonable choice of ϵ, if the probability (12) is small,

then we would conclude that there is a high level of protection against disclosure of y_i; on the other hand, if this probability is large, then we would conclude that the amount of protection against disclosure of y_i is low. In order to compute (12) we need to fix the form of the estimator $\hat{y}_i$, and there are a variety of possibilities. For the disclosure risk evaluation presented here, the estimators $\hat{y}_i$ of y_i in (12) based on noise multiplied data and synthetic data, are as follows.

Case (I) noise multiplied data. In the notation of Section 2, for case (I) noise multiplied data, we define the estimator $\hat{y}_i$ as follows:

$$\hat{y}_i = \begin{cases} x_i, \text{ if } \Delta_i = 1, \\ E_{\hat{\boldsymbol{\theta}}_{(\mathrm{I})}}(y_i | x_i, \Delta_i = 0), \text{ if } \Delta_i = 0, \end{cases} \tag{13}$$

where $E_{\hat{\boldsymbol{\theta}}_{(\mathrm{I})}}(y_i | x_i, \Delta_i = 0)$ denotes the conditional expectation of y_i, given $\Delta_i = 0$ and x_i, with the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ set equal to their MLEs, denoted by $\hat{\boldsymbol{\theta}}_{(\mathrm{I})} = (\hat{\boldsymbol{\beta}}_{(\mathrm{I})}, \hat{\sigma}^2_{(\mathrm{I})})$. Here, the MLEs $(\hat{\boldsymbol{\beta}}_{(\mathrm{I})}, \hat{\sigma}^2_{(\mathrm{I})})$ are computed based on case (I) noise multiplied data using the EM algorithm of Section 2. An expression for the conditional expectation is obtained as:

$$\begin{aligned} E_{\boldsymbol{\theta}}(y_i | x_i, \Delta_i = 0) &= E_{\boldsymbol{\theta}}\left(\frac{x_i}{r_i} \,\middle|\, x_i, \Delta_i = 0 \right) \\ &= \frac{\int_0^{x_i/C} \left(\frac{x_i}{r}\right) \exp\left[-\frac{\{\ln(x_i/r) - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2} \right] h(r) dr}{\int_0^{x_i/C} \exp\left[-\frac{\{\ln(x_i/r) - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2} \right] h(r) dr}, \end{aligned} \tag{14}$$

where the second equality above follows by using Proposition 1 and Corollary 1 of Section 2 to obtain the conditional density of r_i given x_i and $\Delta_i = 0$. Notice that when $\Delta_i = 0$, the estimator $\hat{y}_i$ defined in (13) uses information from the released case (I) noise multiplied data, information in the covariates, and also information in the noise generating distribution $h(r)$. These three pieces of information need to be made public in order to draw valid inferences using the methods developed in Section 2, and hence we have defined a $\hat{y}_i$ that makes use of all of the above information. Obviously, when $\Delta_i = 1$, y_i is not sensitive and we simply have $\hat{y}_i = x_i = y_i$.

Case (II) noise multiplied data. Working again under the notation of Section 2, for case (II) noise multiplied data, we consider an estimator $\hat{y}_i$ that has a similar form as (13), except that in this case Δ_i, the indicator for $y_i \leq C$, is not observed. Hence, for case (II) noise multiplied data, we define $\hat{y}_i$ as

$$\hat{y}_i = E_{\hat{\boldsymbol{\theta}}_{(\mathrm{II})}}(y_i | x_i). \tag{15}$$

Here, $E_{\hat{\boldsymbol{\theta}}_{(\mathrm{II})}}(y_i | x_i)$ denotes the conditional expectation of y_i given x_i, with the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ set equal to their MLEs (denoted by $\hat{\boldsymbol{\theta}}_{(\mathrm{II})} = (\hat{\boldsymbol{\beta}}_{(\mathrm{II})}, \hat{\sigma}^2_{(\mathrm{II})})$). Here,

the MLEs are computed based on case (II) noise multiplied data using the EM algorithm of Section 2. To evaluate $E_{\boldsymbol{\theta}}(y_i|x_i)$, note that,

$$\begin{aligned} E_{\boldsymbol{\theta}}(y_i|x_i) &= E_{\boldsymbol{\theta}}[E_{\boldsymbol{\theta}}(y_i|x_i,\Delta_i)|x_i] \\ &= E_{\boldsymbol{\theta}}(y_i|x_i,\Delta_i=1)\Pr(\Delta_i=1|x_i) + E_{\boldsymbol{\theta}}(y_i|x_i,\Delta_i=0)[1-\Pr(\Delta_i=1|x_i)] \\ &= x_i \Pr(\Delta_i=1|x_i) + E_{\boldsymbol{\theta}}\left(\frac{x_i}{r_i}\,\middle|\, x_i,\Delta_i=0\right)[1-\Pr(\Delta_i=1|x_i)]. \end{aligned} \tag{16}$$

The expression for $E_{\boldsymbol{\theta}}(y_i|x_i,\Delta_i=0)$ is given by (14), and the expression for $\Pr(\Delta_i = 1|x_i)$ is given by (10), and therefore, by plugging these expressions into (16) and simplifying, we get

$$E_{\boldsymbol{\theta}}(y_i|x_i) = \frac{x_i\, I(x_i<C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right] + \int_0^{x_i/C}\left(\frac{x_i}{r}\right)\exp\left[-\frac{\{\ln(x_i/r)-\boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right]h(r)dr}{I(x_i<C)\exp\left[-\frac{\{\ln x_i - \boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right] + \int_0^{x_i/C}\exp\left[-\frac{\{\ln(x_i/r)-\boldsymbol{u}_i'\boldsymbol{\beta}\}^2}{2\sigma^2}\right]h(r)dr}.$$

Notice that the estimator $\hat{y}_i$ defined in (15) uses information from the released case (II) noise multiplied data, information in the covariates, and also information in the noise generating distribution $h(r)$.

Synthetic data. For the synthetic data methods PMIC and PMID discussed in Section 3.2, let $y_{i1}^*, \ldots, y_{im}^*$ denote the values corresponding to y_i in the m released synthetic datasets. In the notation of Section 3.2, recall that if $y_i > C_I$, then $y_{i1}^*, \ldots, y_{im}^*$ are obtained as draws from a posterior predictive distribution; otherwise, if $y_i \le C_I$, then $y_{i1}^* = \cdots = y_{im}^* = y_i$. So a natural and simple estimator of y_i based on the synthetic data is $\hat{y}_i = \frac{1}{m}\sum_{j=1}^m y_{ij}^*$.

Now that we have defined suitable estimators $\hat{y}_i$ under noise multiplied and synthetic data, we can evaluate the probabilities $p_{i,\epsilon}$, as defined by (12), for each method. As in Section 5.1, we take the threshold C as the 90th empirical percentile of total household income. Since our dataset consists of $n = 50,661$ households, there are 5,066 values of y_i that exceed C. For case (I) and case (II) noise multiplied data, we use simulation (with 5,000 iterations) to estimate the $p_{i,\epsilon}$ values for each of the methods NMh1.i, NMh2.i, NMh3.i, NMh4.i, NMh1.ii, NMh2.ii, NMh3.ii, and NMh4.ii as defined in Section 4. Within each iteration of the simulation, we use the `integrate` function in R to evaluate the integrals that appear in (13) and (15). For synthetic data, we also use simulation (with 5,000 iterations) to estimate the $p_{i,\epsilon}$ values for each of the methods PMIC2, PMIC4, PMID2, and PMID4, as defined in Section 4; and for each of these methods we consider two options for the number of imputations: $m = 5$ and $m = 50$.

The results of our computations are summarized in Table 7. The table summarizes the distribution of $p_{i,\epsilon}$ for the 5,066 y_i-values that require protection because they exceed the threshold C (the 90th empirical percentile of $y_1, \ldots, y_n$). The table shows the 1st quartile (Q_1), median, mean, and 3rd quartile (Q_3) of these 5,066 $p_{i,\epsilon}$-values under

Table 7: Distribution of $p_{i,\epsilon} = \Pr\{|\frac{\hat{y}_i - y_i}{y_i}| \leq \epsilon \,|\, \boldsymbol{y}\}$ in the CPS data example for the 5,066 y_i-values that exceed C and thus require protection.

	$\epsilon = 0.1$				$\epsilon = 0.2$			
	Q_1	Med.	Mean	Q_3	Q_1	Med.	Mean	Q_3
PMIC2 ($m = 5$)	0.10	0.22	0.20	0.29	0.21	0.42	0.39	0.55
PMIC2 ($m = 50$)	0.00	0.04	0.21	0.39	0.01	0.27	0.40	0.82
PMIC4 ($m = 5$)	0.07	0.17	0.15	0.23	0.16	0.35	0.31	0.46
PMIC4 ($m = 50$)	0.00	0.07	0.21	0.44	0.01	0.35	0.41	0.82
PMID2 ($m = 5$)	0.04	0.30	0.25	0.43	0.17	0.59	0.48	0.75
PMID2 ($m = 50$)	0.00	0.11	0.33	0.72	0.00	0.79	0.57	0.99
PMID4 ($m = 5$)	0.00	0.07	0.13	0.24	0.01	0.21	0.28	0.53
PMID4 ($m = 50$)	0.00	0.00	0.13	0.06	0.00	0.01	0.30	0.69
NMh1.i	0.14	0.49	0.35	0.51	0.93	1.00	0.95	1.00
NMh1.ii	0.14	0.23	0.25	0.31	0.93	1.00	0.95	1.00
NMh2.i	0.28	0.37	0.37	0.49	0.59	0.68	0.64	0.73
NMh2.ii	0.11	0.21	0.22	0.29	0.35	0.41	0.43	0.47
NMh3.i	0.19	0.27	0.28	0.35	0.40	0.45	0.54	0.69
NMh3.ii	0.08	0.13	0.16	0.24	0.23	0.38	0.38	0.48
NMh4.i	0.04	0.08	0.16	0.23	0.10	0.23	0.33	0.59
NMh4.ii	0.00	0.06	0.10	0.15	0.12	0.19	0.22	0.30

each of the methods. The results are provided in the table for the cases when $\epsilon = 0.1$ and $\epsilon = 0.2$, and the rows of the table correspond to the different statistical disclosure control methods, using notation defined in Section 4 (as explained above). Histograms of the $p_{i,0.1}$ values for the 5,066 y_i-values that exceed C under the noise multiplication methods are provided in Figures 1 and 2, while similar histograms under the synthetic data methods are provided in Figures 3 and 4. The following is a summary of the findings of our disclosure risk evaluation.

1. We see in Table 7, and Figures 1 and 2, that when comparing the methods NMh1.i, NMh2.i, NMh3.i, and NMh4.i, the $p_{i,\epsilon}$ values tend to be largest under NMh1.i and smallest under NMh4.i. A similar conclusion holds when one compares the $p_{i,\epsilon}$ values under NMh1.ii, NMh2.ii, NMh3.ii, and NMh4.ii. Such a finding is expected because among the noise distributions h1, h2, h3, and h4, the distribution h1 is the least dispersed, while the distribution h4 is the most dispersed (refer to (9)). We notice in Figures 1(a) and 1(c) that the noise methods NMh1.i and NMh2.i, tend to yield a value of $p_{i,0.1}$ that is larger than 0.5 for many y_i values. On the other hand, in Figures 2(b) and 2(d) we see that the noise methods NMh3.ii and NMh4.ii, yield a value of $p_{i,0.1}$ that is smaller than 0.5 for almost all y_i values. In fact, the method NMh4.ii gives a value of $p_{i,0.1} \leq 0.05$ for a large proportion of y_i values. Thus, it is clear that the choice of the noise generating distribution plays a critical role in setting the level of protection against disclosure.

2. We notice in Table 7 that for some y_i values, NMh1.ii provides more privacy protection than NMh1.i. Similarly, for some y_i values, NMh2.ii provides more privacy protection than NMh2.i, NMh3.ii provides more privacy protection than NMh3.i, and NMh4.ii provides more privacy protection than NMh4.i. The increase in privacy protection in case (II) over case (I) noise multiplied data can occur under the noise distribution (8), because for any y_i value whose corresponding x_i falls in the interval $(C\xi_1, C)$, it is not known with certainty whether this x_i is noise multiplied or not, that is, whether $\Delta_i = 0$ or $= 1$. (Among the y_i that exceed C, it is those values for which $C < y_i < \frac{C}{\xi_1}$ that can potentially have x_i in $(C\xi_1, C)$.) On the other hand, under case (I) noise multiplied data, the value of Δ_i is always known because it is explicitly released. As a result, there is an additional component of uncertainty in the estimator (15) in comparison with the estimator (13), caused by not knowing the value of Δ_i with certainty (which can be readily seen by comparing equation (16) with (13)). Furthermore, to compute $\hat{y}_i$, we plug an estimate of the unknown parameter $\boldsymbol{\theta}$ into the conditional expectations (13) and (15); and we saw in Section 4 that the estimate of $\boldsymbol{\theta}$ under case (II) noise multiplication is less efficient than that under case (I) noise multiplication. The additional uncertainty translates to an increase in privacy protection as measured by criterion (12). Figures 1 and 2 indicate that case (II) noise multiplication can provide a substantial increase in privacy protection in comparison with case (I) noise multiplication, for many of the sensitive values. The increase in privacy protection in case (II) over case (I) occurs for quite a large proportion of observations under the noise setting h4 (Figures 2(c) and 2(d)), because in this case $\xi_1 = 0.1$, and hence the interval $(C\xi_1, C)$, is quite wide. Of course, this gain in privacy protection is at the expense of a slight loss in accuracy of inference, as shown in Section 4 and Section 5.1.

3. For the synthetic data methods, we notice in Table 7 that when m increases from 5 to 50, the disclosure risk increases substantially for some observations, while for others, the disclosure risk decreases substantially. To examine the situation further, Figure 3 shows histograms of the $p_{i,0.1}$ values for all $y_i > C$ for the PMID2 and PMID4 methods, and Figure 4 shows histograms of the $p_{i,0.1}$ values for all $y_i > C$ for the PMIC2 and PMIC4 methods. We observe in Figures 3 and 4 that when m increases from 5 to 50, the disclosure risk for a portion of the observations becomes quite high, while for many observations the value of $p_{i,0.1}$ becomes very small.

   To see why the disclosure risk can become so high for some observations and small for others, consider first the PMID methods. Under the PMID methods, for fixed $\boldsymbol{y}$ we have $\hat{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}^* \xrightarrow{P} E(y_{i1}^*|\boldsymbol{y})$ as $m \to \infty$ by the Law of Large Numbers, where

$$E(y_{i1}^*|\boldsymbol{y}) = \exp[\boldsymbol{u}_i'\tilde{\boldsymbol{\beta}}_{\text{del}}]\, E\left\{\exp\left[(\sigma^*)^2(1 + \boldsymbol{u}_i'\boldsymbol{V}_{\text{del}}\boldsymbol{u}_i)/2\right] \mid \boldsymbol{y}\right\},$$

   and $(\sigma^*)^2|\boldsymbol{y} \sim (n_{\text{del}} - p)\tilde{s}_{\text{del}}^2/\chi_{n_{\text{del}}-p}^2$, and $\tilde{\boldsymbol{\beta}}_{\text{del}}$, $\tilde{s}_{\text{del}}^2$, $\boldsymbol{V}_{\text{del}}$, and n_{del} are defined in Section 3.2 just after equation (6). We ran a Monte Carlo simulation with 10,000 iterations to compute $\psi_i(\boldsymbol{y}) = E\{\exp[(\sigma^*)^2(1+\boldsymbol{u}_i'\boldsymbol{V}_{\text{del}}\boldsymbol{u}_i)/2] \mid \boldsymbol{y}\}$; we found that under PMID2,

$$1.061 \le \psi_i(\boldsymbol{y}) \le 1.077, \quad \text{for all } i \text{ such that } y_i \in \boldsymbol{y}_{\text{del}},$$

and under PMID4,

$$1.077 \leq \psi_i(\boldsymbol{y}) \leq 1.081, \quad \text{for all } i \text{ such that } y_i \in \boldsymbol{y}_{\text{del}},$$

where $\boldsymbol{y}_{\text{del}} = \{y_i : y_i > C_I\}$ as defined in Section 3.2. Thus we conclude that for large m, $\hat{y}_i \approx \psi_i(\boldsymbol{y}) \exp[\boldsymbol{u}_i' \tilde{\boldsymbol{\beta}}_{\text{del}}]$, and since $\psi_i(\boldsymbol{y})$ is always just slightly larger than 1, $\hat{y}_i$ is just slightly greater than the fitted value from the regression of $\ln y_i$ on $\boldsymbol{u}_i$ (based on all $y_i \in \boldsymbol{y}_{\text{del}}$) transformed back to the y_i scale. So for data points such that y_i is well approximated by the fitted value $\exp[\boldsymbol{u}_i' \tilde{\boldsymbol{\beta}}_{\text{del}}]$, the value of $p_{i,\epsilon}$ will be large for large m, otherwise the value of $p_{i,\epsilon}$ will be small. Figures 5(a) and 5(b) illustrate this point. Figure 5(a) plots the observed y_i versus the fitted value $\exp[\boldsymbol{u}_i' \tilde{\boldsymbol{\beta}}_{\text{del}}]$ for those y_i with $p_{i,0.1} \leq 0.5$ under PMID2 with $m = 50$, while Figure 5(b) shows a similar plot for those y_i with $p_{i,0.1} > 0.5$ under PMID2 with $m = 50$. We notice in Figure 5(b) that the y_i values with $p_{i,0.1} > 0.5$ are well approximated by their fitted values, while in Figure 5(a) we see that the y_i values with $p_{i,0.1} < 0.5$ are not approximated as well by the fitted values. A similar situation occurs under the PMIC2 method, as displayed by Figures 5(c) and 5(d). Notice in Figure 5(d) that there is a strong linear relationship between y_i and the fitted value $\exp[\boldsymbol{u}_i' \tilde{\boldsymbol{\beta}}]$ for those y_i values with $p_{i,0.1} > 0.5$ under PMIC2. However, in this case the fitted values tend to be much smaller than the original values, which occurs because, under PMIC, synthetic data are sampled from a truncated distribution. In the PMIC method, after generating parameter values from the complete data posterior distribution, synthetic data values are drawn from a truncated log-normal distribution, and as a result, for fixed $\boldsymbol{y}$, $\hat{y}_i = \frac{1}{m} \sum_{j=1}^{m} y_{ij}^* \xrightarrow{P} E(y_{i1}^* | \boldsymbol{y})$ as $m \to \infty$ where

$$E(y_{i1}^* | \boldsymbol{y}) = E\left\{ \exp\left[\boldsymbol{u}_i' \boldsymbol{\beta}^* + (\sigma^*)^2/2\right] \left[\frac{1 - \Phi\left(\frac{\ln C_I - \boldsymbol{u}_i' \boldsymbol{\beta}^*}{\sigma^*} - \sigma^*\right)}{1 - \Phi\left(\frac{\ln C_I - \boldsymbol{u}_i' \boldsymbol{\beta}^*}{\sigma^*}\right)} \right] \,\middle|\, \boldsymbol{y} \right\},$$

and $(\boldsymbol{\beta}^*, (\sigma^*)^2)$, conditional on $\boldsymbol{y}$, have the joint posterior distribution (5).

4. Looking again at Figure 3, we notice that the $p_{i,0.1}$ values tend to be concentrated closer to zero under PMID4 in comparison with the $p_{i,0.1}$ values under PMID2, when the number of imputations m is the same. So in this sense, PMID4 can provide more protection than PMID2, as expected. A similar conclusion holds when one looks at Figure 4, and compares $p_{i,0.1}$ values under PMIC2 with those under PMIC4, for the same value of m.

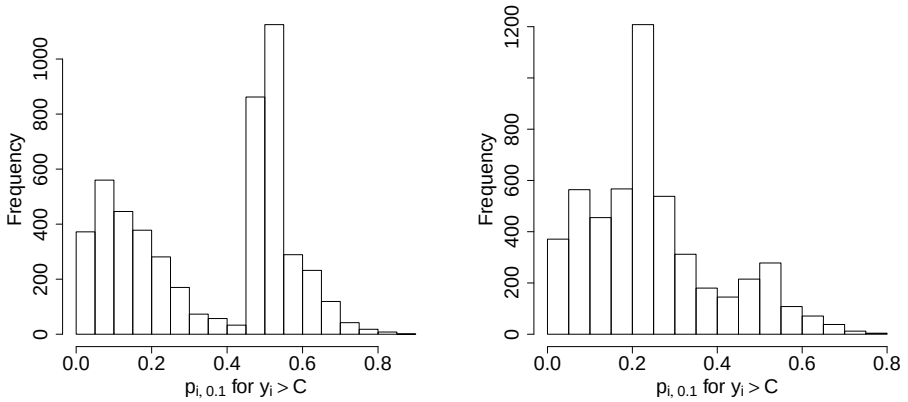
5. When comparing the privacy protection of noise multiplication with that of synthetic data, it is difficult to make any overall conclusions. Under noise multiplication, the noise generating distribution acts as a tuning mechanism; while under synthetic data, the values of C_I and m are the tuning mechanisms. It appears that through the choice of the tuning mechanism, noise multiplication and synthetic data can yield similar levels of privacy protection. For instance, by looking at Figures 2(d) and 3(c), we note how the methods NMh4.ii and PMID4 with $m = 5$ tend to give a similar distribution of $p_{i,0.1}$ values; so the level of privacy protection offered by those

two methods appears to be similar. On the other hand, looking at Figure 3(d), we see that under PMID4 with $m = 50$, most of the y_i values are well protected (the histogram has a spike near 0), but a few y_i values have large $p_{i,0.1}$, and hence are not well protected. Similarly, looking at Figure 2(c), we see that under NMh4.i, there are a few y_i values that have a large $p_{i,0.1}$ value, while most $p_{i,0.1}$ values are small. Thus, under synthetic data, when one increases m which allows for more accurate inferences [27], one also increases the potential for disclosure of some observations. Similarly, under noise multiplication, when one releases the indicators $\Delta_1, \ldots, \Delta_n$ which allows for more accurate inferences (Section 2), one also increases the potential for disclosure of some observations.

In summary, based on the criterion $p_{i,\epsilon}$, under noise multiplication (either case (I) or case (II) data release) the data producer can control the disclosure risk by choosing $h(r)$ to achieve a desired level of protection. In order to do so, the data producer could select a few candidate noise distributions (such as h1, h2, h3, and h4 of (9)), and run a simulation study based on the actual dataset that requires protection, similar to the one presented in this section for the CPS data example. The data producer would then examine the distribution of $p_{i,\epsilon}$ values under each noise setting, and choose the noise distribution that yielded satisfactory results (or if none were found to be satisfactory, run a another simulation with different noise distributions). In the case of synthetic data, obviously there is no noise distribution to set, but the quantities C_I and m can be tuned. Thus the data producer could choose a few candidate values of C_I and m, and then run a simulation to compute the corresponding $p_{i,\epsilon}$ values. As with noise multiplied data, the data producer would then examine the distribution of $p_{i,\epsilon}$ values for each combination of C_I and m, and choose the combination that yields satisfactory results.

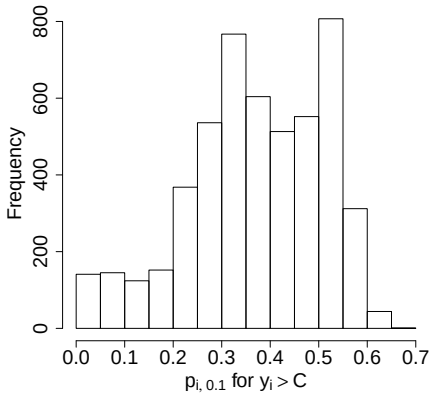
6 Discussion

Methodologies for disclosure limitation need to be balanced with accuracy of the inference before releasing data to the public in order to achieve the dual goals of disclosure avoidance and data utility. In the top code scenario, where only values above a threshold $C > 0$ require protection, we have considered noise multiplication under two data release scenarios: case (I) in which each released value includes an indicator of whether or not it was noise perturbed, and case (II) in which no such indicator is provided. We developed data analysis methods under both cases, and argued that case (I) should always provide more accurate inference than top coding at C, while case (II) can provide either more or less accurate inference than top coding at C, depending on the dispersion of the noise generating distribution. Our empirical results show that both cases provide almost equally accurate inferences when the noise variance is small, and, as expected, the difference in accuracy increases as the dispersion in the noise distribution increases. The case (II) data release may be more desirable for statistical agencies than the case (I) data release, as case (II) can provide an enhanced level of protection against disclosure, as shown in Section 5.2. The results of this article show how to obtain valid inferences

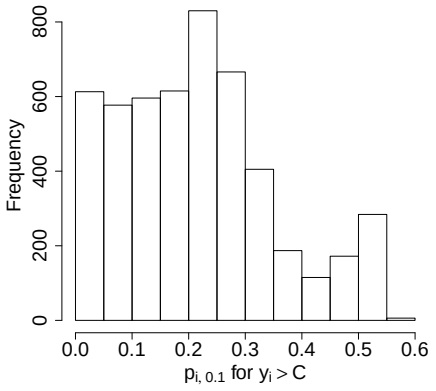


(a) NMh1.i

(b) NMh1.ii

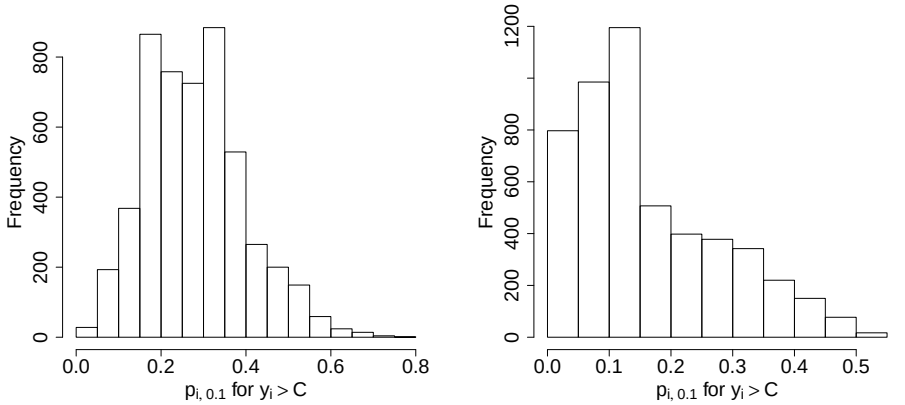


(c) NMh2.i



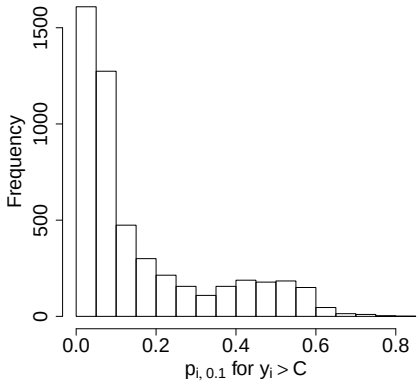
(d) NMh2.ii

Figure 1: Histograms of $p_{i,0.1}$ values for $y_i > C$ under NMh1 and NMh2 methods.

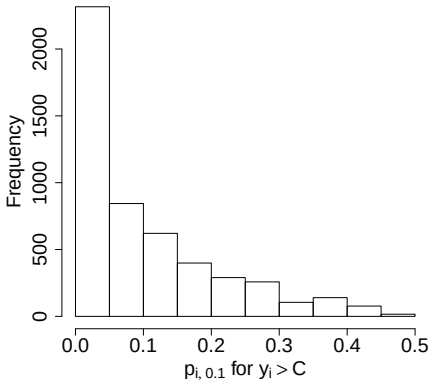


(a) NMh3.i

(b) NMh3.ii

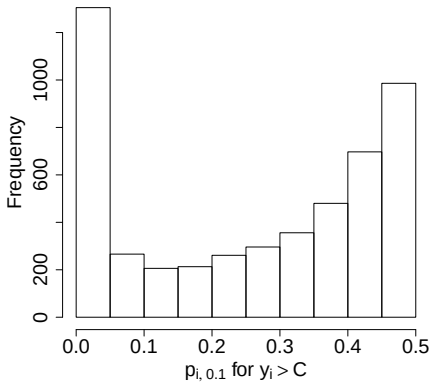


(c) NMh4.i

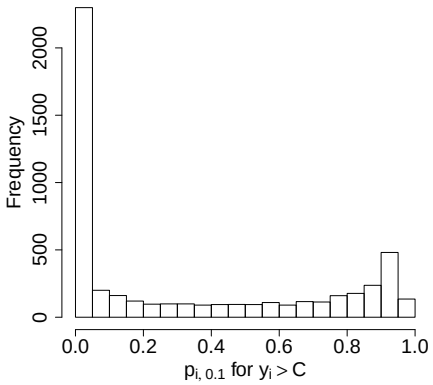


(d) NMh4.ii

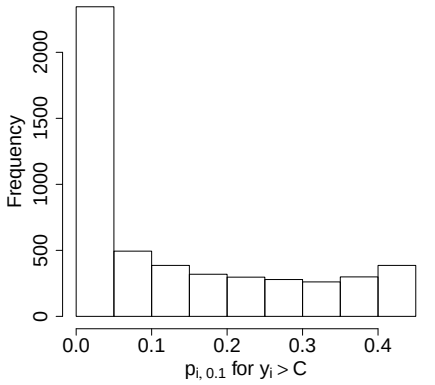
Figure 2: Histograms of $p_{i,0.1}$ values for $y_i > C$ under NMh3 and NMh4 methods.



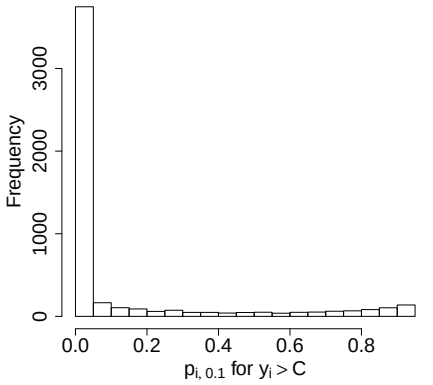
(a) PMID2 ($m = 5$)



(b) PMID2 ($m = 50$)

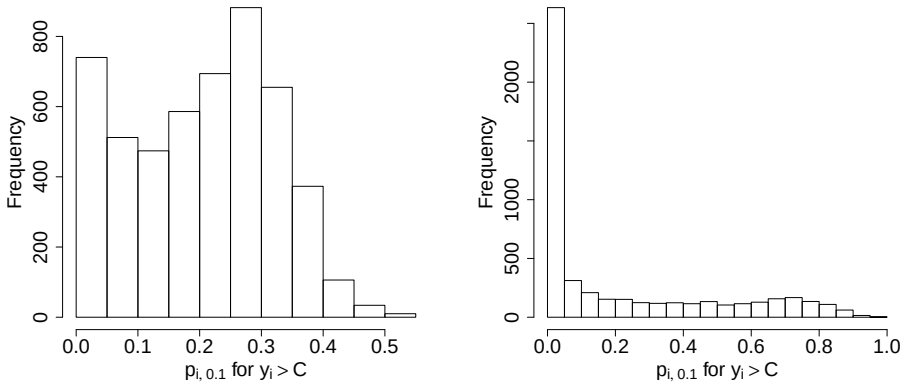


(c) PMID4 ($m = 5$)



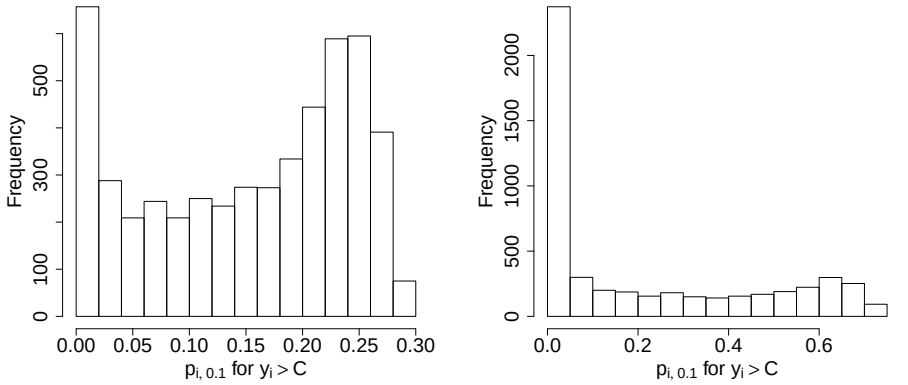
(d) PMID4 ($m = 50$)

Figure 3: Histograms of $p_{i,0.1}$ values for $y_i > C$ under PMID2 and PMID4 methods.



(a) PMIC2 ($m = 5$)

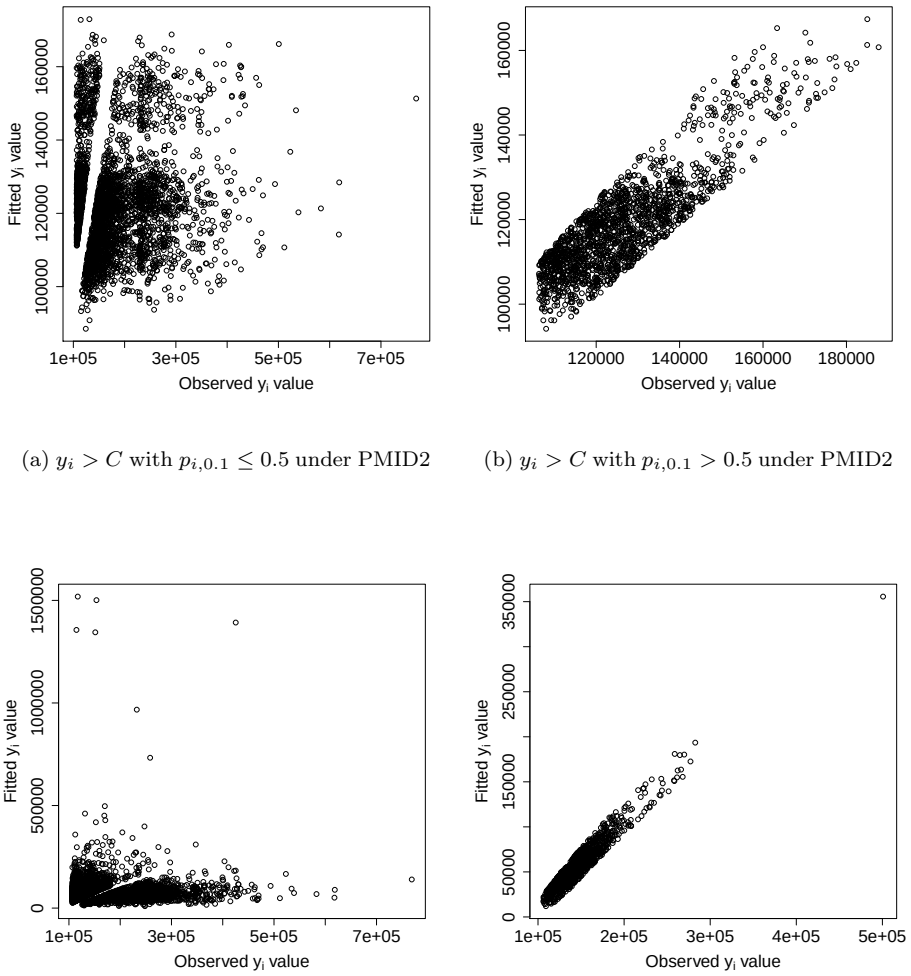
(b) PMIC2 ($m = 50$)



(c) PMIC4 ($m = 5$)

(d) PMIC4 ($m = 50$)

Figure 4: Histograms of $p_{i,0.1}$ values for $y_i > C$ under PMIC2 and PMIC4 methods.



(a) $y_i > C$ with $p_{i,0.1} \leq 0.5$ under PMID2

(b) $y_i > C$ with $p_{i,0.1} > 0.5$ under PMID2

(c) $y_i > C$ with $p_{i,0.1} \leq 0.5$ under PMIC2

(d) $y_i > C$ with $p_{i,0.1} > 0.5$ under PMIC2

Figure 5: Plots of observed y_i values versus the fitted values (defined as $e^{\boldsymbol{u}_i'\tilde{\boldsymbol{\beta}}}$ under PMIC2 and $e^{\boldsymbol{u}_i'\tilde{\boldsymbol{\beta}}_{\text{del}}}$ under PMID2) for large and small $p_{i,0.1}$ values when the number of imputations is $m = 50$.

in both cases. When the dispersion in the noise generating distribution is small, noise multiplication appears to provide inferences that are almost identical to those from the unperturbed data, but naturally, the disclosure risk may be high.

Parametric statistical procedures based on synthetic data and noise multiplied data can provide comparable inferences in many cases that we considered. The accuracy of inferences obtained from these two different methods could usually be made nearly equivalent by setting the noise variance appropriately. An appealing feature of noise multiplication is its flexibility; the noise generating distribution acts as a tuning mechanism, and its choice allows one to balance data quality with confidentiality protection. While inferences become less accurate when the noise variance is large, we note that the inferences generally are still valid, i.e., confidence interval coverage probability is generally maintained at the nominal level, bias of standard deviation estimators is small, etc. We should note that in this scenario, synthetic data also provide a tuning mechanism, through the choice of the cut-point C_I, and the number of imputations m, as discussed in Section 5.2.

As with top coding, noise multiplied data have the feature that the data are easy for the data producer to create. When creating the noise multiplied version of y_i, the data producer does not need to consider which regressor variables to include in $\boldsymbol{u}_i$, and yet the methodology presented here enables a data user to perform an analysis, using the noise multiplied version of y_i, to determine which variables would be good predictors of y_i. Nevertheless, there are drawbacks of the noise multiplication method. For instance, the proposed likelihood-based analysis is complicated for data users to apply, as it requires EM algorithms and careful expressions for observed Fisher information. However, if software is available, then complexity of the data analysis will not be an issue. In our numerical evaluations, we found that the EM algorithms were quite stable, and converged rapidly. Our methodology can be used with a noise generating distribution such as (8), which has no mass in an interval that contains 1 (as illustrated in Sections 4 and 5), and hence provides a positive lower bound for the relative distance between the original and noise multiplied values. As we have mentioned previously, if data related to income are released, then the large values often require protection. In such situations, top coding is routinely applied, and the noise multiplication methods presented in this article can serve as an alternative method of statistical disclosure control.

It should be noted that our work deals with the analysis of a univariate response variable whose log-scale mean is described by a multiple linear regression on a set of non-sensitive regressor variables. In a future communication we hope to take up the multivariate regression scenario, and the scenario of noise multiplied regressor variables. Furthermore, the methodology for noise multiplication as presented in Section 2 could be extended to other parametric models besides the log-normal. Finally, we note that our proposed methodology relies on asymptotic normality of the maximum likelihood estimator for drawing inferences; however, it would also be possible to use ideas akin to those of Charest [7] to derive a fully Bayesian approach to obtain posterior inferences for the unknown parameters.

Acknowledgments

The authors thank Paul Massell, two anonymous referees, and the Editor for providing many helpful comments that enhanced the quality of the paper. The authors also thank Jerry Reiter for some valuable discussions, and Laura McKenna, Joseph Schafer, Eric Slud, Yves Thibaudeau, and Tommy Wright for encouragement.

References

[1] 2007 Survey of Business Owners (SBO) Public Use Microdata Sample (PUMS) Data Users Guide. (2012). Available at `http://www2.census.gov/econ/sbo/07/pums/2007_sbo_pums_users_guide.pdf`.

[2] Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41:997–1016.

[3] — (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19:1483–1536.

[4] — (1984). Tobit models: A survey. *Journal of Econometrics*, 24, 3–61.

[5] An, D., and Little, R. J. A. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170:923–940.

[6] Brown, G. and Sanders, J. S. (1981). Lognormal genesis. *Journal of Applied Probability*, 18:542–547.

[7] Charest, A. -S. (2010). How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2:21–33.

[8] Cirera, X., and Masset, E. (2010). Income distribution trends and future food demand. *Philosophical Transactions of the Royal Society, Series B*, 365:2821–2834.

[9] Drechsler, J., and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105:1347–1357.

[10] Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer.

[11] Gartner, H., and Rässler, S. (2005). Analyzing the Changing Gender Wage Gap Based on Multiply Imputed Right Censored Wages, IAB Discussion Paper No. 5/2005, Nuremberg, Germany: Institute for Employment Research. Available at `http://doku.iab.de/discussionpapers/2005/dp0505.pdf`.

[12] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Second edition. New York: Chapman & Hall/CRC.

[13] Goldberger, A. S. (1964). *Econometric Theory*. New York: Wiley.

[14] Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. *Journal of American Statistical Association*, 81, 680-688.

[15] Jenkins, S. P., Burkhauser, R. V., Feng, S., and Larrimore, J. (2009). Measuring inequality using censored data: A multiple imputation approach. Available at `http://ftp.iza.org/dp4011.pdf`.

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

[17] Kleiber, C., and Zeileis, A. (2008). *Applied Econometrics with R*, New York: Springer-Verlag. Available at `http://CRAN.R-project.org/package=AER`.

[18] Klein, M., Mathew, T. and Sinha, B. (2014). Likelihood based inference under noise multiplication. *Thailand Statistician: Journal of the Thai Statistical Association*, 12:1–23.

[19] Klein, M., and Sinha, B. (2013). Statistical analysis of noise-multiplied data using multiple imputation. *Journal of Official Statistics*, 29, 425-465.

[20] Lawrence, R. J. (1988). Applications in economics and business. In E. L. Crow and K. Shimizu (eds), *Lognormal Distributions: Theory and Applications*, New York: Marcel Dekker. 229–266.

[21] Lin, Y. -X and Wise, P. (2012). Estimation of regression parameters from noise multiplied data. *Journal of Privacy and Confidentiality*, 4:61–94.

[22] Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426.

[23] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Second edition. New York: Wiley.

[24] 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.

[25] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at `http://www.R-project.org/`.

[26] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, 19:1–16.

[27] Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–188.

[28] — (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation, *Survey Methodology*, 30:235–242.

[29] — (2005a). Releasing multiply-imputed synthetic use public use microdata: An illustration and empirical study. *Journal of Royal Statistical Society, Series A*, 168:185–205.

[30] — (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131:365–377.

[31] — (2005c). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21:441–462.

[32] — (2005d). Releasing multiply-imputed synthetic public use microdata: An illustration and empirical study. *Journal of Royal Statistical Society, Series A*, 168:185–205.

[33] — (2005e). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100:1103–1112.

[34] Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1:99–110.

[35] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

[36] — (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:461–468.

[37] Sinha, B., Nayak, T., and Zayatz, L. (2011). Privacy protection and quantile estimation from noise multiplied data, *Sankhya, Series B*, 73:297–315.

[38] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36.

Appendix

1 Expressions for Observed Fisher Information

1.1 Case (I) Noise Multiplied Data as Defined in Section 2

Here we work under the notation of Section 2. Using Corollary 2 of Section 2, we have the log-likelihood function as

$$\ell(\boldsymbol{\theta}|x_1,\ldots,x_n,\Delta_1,\ldots,\Delta_n,\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)=\sum_{i=1}^{n}\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i).$$

Thus the observed Fisher information matrix with dimension $(p+1)\times(p+1)$ is

$$-\sum_{i=1}^{n}\begin{pmatrix}\frac{\partial^2\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial\boldsymbol{\beta}\partial(\sigma^2)}\\ \frac{\partial^2\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial(\sigma^2)\partial\boldsymbol{\beta}'} & \frac{\partial^2\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial(\sigma^2)^2}\end{pmatrix}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{(\mathrm{I})}},$$

where $\hat{\boldsymbol{\theta}}_{(\mathrm{I})}$ is the MLE of $\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma^2)$ based on the data $(x_1,\Delta_1,\boldsymbol{u}_1),\ldots,(x_n,\Delta_n,\boldsymbol{u}_n)$. Defining

$$\mu_i=\boldsymbol{u}_i'\boldsymbol{\beta},\ \ a_i(r)=\exp\left[\frac{-\left(\ln\left(\frac{x_i}{r}\right)-\mu_i\right)^2}{2\sigma^2}\right],\ \ b_i(r)=\ln\left(\frac{x_i}{r}\right)-\mu_i, \tag{17}$$

the expressions for the elements of the observed Fisher information matrix are:

$$\frac{\partial\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial\beta_j}=\frac{(\ln x_i-\mu_i)\Delta_i u_{ij}}{\sigma^2}+(1-\Delta_i)\frac{\int_0^{\frac{x_i}{C}}a_i(r)b_i(r)u_{ij}\frac{1}{\sigma^2}h(r)dr}{\int_0^{\frac{x_i}{C}}a_i(r)h(r)dr},$$

$$\frac{\partial^2\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial\beta_j^2}=-\frac{u_{ij}^2}{\sigma^2}+(1-\Delta_i)u_{ij}^2\left[\frac{\int_0^{\frac{x_i}{C}}a_i(r)b_i^2(r)\frac{1}{\sigma^4}h(r)dr}{\int_0^{\frac{x_i}{C}}a_i(r)h(r)dr}-\left\{\frac{\int_0^{\frac{x_i}{C}}a_i(r)b_i(r)\frac{1}{\sigma^2}h(r)dr}{\int_0^{\frac{x_i}{C}}a_i(r)h(r)dr}\right\}^2\right],$$

$$\frac{\partial^2\ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial\beta_j\partial\beta_{j'}}=-\frac{u_{ij}u_{ij'}}{\sigma^2}+(1-\Delta_i)u_{ij}u_{ij'}\left[\frac{\int_0^{\frac{x_i}{C}}a_i(r)b_i^2(r)\frac{1}{\sigma^4}h(r)dr}{\int_0^{\frac{x_i}{C}}a_i(r)h(r)dr}-\left\{\frac{\int_0^{\frac{x_i}{C}}a_i(r)b_i(r)\frac{1}{\sigma^2}h(r)dr}{\int_0^{\frac{x_i}{C}}a_i(r)h(r)dr}\right\}^2\right],$$

$$\frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial\beta_j\partial\sigma^2} = -\frac{(\ln x_i-\mu_i)\Delta_i u_{ij}}{\sigma^4} + (1-\Delta_i)u_{ij}\Bigg[-\frac{\int_0^{\frac{x_i}{C}} a_i(r)b_i(r)\frac{1}{\sigma^4}h(r)dr}{\int_0^{\frac{x_i}{C}} a_i(r)h(r)dr}$$
$$+\frac{1}{2\sigma^6}\frac{\int_0^{\frac{x_i}{C}} a_i(r)b_i^3(r)h(r)dr}{\int_0^{\frac{x_i}{C}} a_i(r)h(r)dr} - \frac{1}{2\sigma^6}\frac{\{\int_0^{\frac{x_i}{C}} a_i(r)b_i(r)h(r)dr\}\{\int_0^{\frac{x_i}{C}} a_i(r)b_i^2(r)h(r)dr\}}{\left\{\int_0^{\frac{x_i}{C}} a_i(r)h(r)dr\right\}^2}\Bigg],$$

$$\frac{\partial \ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{\partial(\sigma^2)} = -\frac{1}{2\sigma^2} + \frac{\Delta_i(\ln x_i-\mu_i)^2}{2\sigma^4} + \frac{(1-\Delta_i)\frac{1}{2\sigma^4}\int_0^{\frac{x_i}{C}} a_i(r)b_i^2(r)\frac{h(r)}{r}dr}{\int_0^{\frac{x_i}{C}} a_i(r)\frac{h(r)}{r}dr},$$

$$\frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i,\Delta_i|\boldsymbol{u}_i)}{(\partial\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{(\ln x_i-\mu_i)^2\Delta_i}{\sigma^6} + (1-\Delta_i)\Bigg[-\frac{1}{\sigma^6}\frac{\int_0^{\frac{x_i}{C}} a_i(r)b_i^2(r)h(r)dr}{\int_0^{\frac{x_i}{C}} a_i(r)h(r)dr}$$
$$+\frac{1}{4\sigma^8}\frac{\int_0^{\frac{x_i}{C}} a_i(r)b_i^4(r)h(r)dr}{\int_0^{\frac{x_i}{C}} a_i(r)h(r)dr} - \frac{1}{4\sigma^8}\left\{\frac{\int_0^{\frac{x_i}{C}} a_i(r)b_i^2(r)h(r)dr}{\int_0^{\frac{x_i}{C}} a_i(r)h(r)dr}\right\}^2\Bigg].$$

1.2 Case (II) Noise Multiplied Data as Defined in Section 2

Here we work under the notation of Section 2. Using Corollary 4 of Section 2, we have the log-likelihood function as $\ell(\boldsymbol{\theta}|x_1,\ldots,x_n,\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n) = \sum_{i=1}^n \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)$. Thus the observed Fisher information matrix with dimension $(p+1)\times(p+1)$ is

$$-\sum_{i=1}^n \begin{pmatrix} \frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial\boldsymbol{\beta}\partial(\sigma^2)} \\ \frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial(\sigma^2)\partial\boldsymbol{\beta}'} & \frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial(\sigma^2)^2} \end{pmatrix}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{(\mathrm{II})}},$$

where $\hat{\boldsymbol{\theta}}_{(\mathrm{II})}$ is the MLE of $\boldsymbol{\theta} = (\boldsymbol{\beta},\sigma^2)$ based on the data $(x_1,\boldsymbol{u}_1),\ldots,(x_n,\boldsymbol{u}_n)$. Defining μ_i, $a_i(r)$ and $b_i(r)$ as in (17), the expressions for the derivatives are:

$$\frac{\partial \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial\beta_j} = u_{ij}\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}\frac{\ln x_i-\mu_i}{\sigma^2}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i(r)\frac{1}{\sigma^2}h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}$$
$$= -\frac{\mu_i u_{ij}}{\sigma^2} + \frac{u_{ij}}{\sigma^2}\left[\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)(\ln\frac{x_i}{r})h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}\right],$$

$$\frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial \beta_j^2} = -\frac{u_{ij}^2}{\sigma^2} + \frac{u_{ij}^2}{\sigma^4}$$
$$\times \frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)(\ln x_i-\mu_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)\ln\left(\frac{x_i}{r}\right) b_i(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}} I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}$$
$$- \frac{u_{ij}^2}{\sigma^4}\left\{\left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)\ln\left(\frac{x_i}{r}\right)h(r)dr\right]\right.$$
$$\times \left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i(r)h(r)dr\right]$$
$$\left.\times \left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}} I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr\right]^{-2}\right\},$$

$$\frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial \beta_j \partial \beta_{j'}} = -\frac{u_{ij}u_{ij'}}{\sigma^2} + \frac{u_{ij}u_{ij'}}{\sigma^4}$$
$$\times \frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)(\ln x_i-\mu_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)\ln\left(\frac{x_i}{r}\right) b_i(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}} I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}$$
$$- \frac{u_{ij}u_{ij'}}{\sigma^4}\left\{\left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)\ln\left(\frac{x_i}{r}\right)h(r)dr\right]\right.$$
$$\times \left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i(r)h(r)dr\right]$$
$$\left.\times \left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}} I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr\right]^{-2}\right\},$$

$$\frac{\partial \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial \sigma^2} = -\frac{1}{2\sigma^2}$$
$$+ \frac{1}{2\sigma^4}\left[\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)^2 I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i^2(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}} I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}\right],$$

$$\frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial\beta_j\partial\sigma^2} = \frac{\mu_i u_{ij}}{\sigma^4} - \frac{u_{ij}}{\sigma^4}$$
$$\times \left[\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)\ln\left(\frac{x_i}{r}\right)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}\right] + \frac{u_{ij}}{2\sigma^6}$$
$$\times \frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)(\ln x_i-\mu_i)^2 I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)(\ln\frac{x_i}{r})b_i^2(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}$$
$$- \frac{u_{ij}}{2\sigma^6}\left\{\left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i)I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)\ln\left(\frac{x_i}{r}\right)h(r)dr\right]\right.$$
$$\times \left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)^2 I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i^2(r)h(r)dr\right]$$
$$\left.\times \left[e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr\right]^{-2}\right\},$$

$$\frac{\partial^2 \ln k_{\boldsymbol{\theta}}(x_i|\boldsymbol{u}_i)}{\partial(\sigma^2)^2} = \frac{1}{2\sigma^4}$$
$$- \frac{1}{\sigma^6}\left[\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)^2 I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i^2(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}\right]$$
$$+ \frac{1}{4\sigma^8}\left[\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)^4 I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i^4(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x>0)h(r)dr}\right]$$
$$- \frac{1}{4\sigma^8}\left[\frac{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}(\ln x_i-\mu_i)^2 I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)b_i^2(r)h(r)dr}{e^{-\frac{(\ln x_i-\mu_i)^2}{2\sigma^2}}I(x_i<C) + \int_0^{\frac{x_i}{C}} a_i(r)I(x_i>0)h(r)dr}\right]^2.$$

1.3 Top Coded Data as Defined in Section 3.1

Here we work under the notation of Section 3.1, and we let

$$\boldsymbol{w}_{\text{obs}} = \{(\tilde{x}_1, \Delta_1, \boldsymbol{u}_1), \ldots, (\tilde{x}_n, \Delta_n, \boldsymbol{u}_n)\}$$

denote the observed top coded data. The likelihood function for $\boldsymbol{\theta}$ based on $\boldsymbol{w}_{\text{obs}}$ can be expressed as

$$L(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}}) = \prod_{i=1}^{n}\left(\left[\frac{1}{\sigma}\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\right]^{\Delta_i}\left[\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\right]^{1-\Delta_i}\right),$$

and the loglikelihood as

$$\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}}) = -\ln(\sigma)\left(\sum_{i=1}^{n}\Delta_i\right)+\sum_{i=1}^{n}\Delta_i \ln\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)+\sum_{i=1}^{n}(1-\Delta_i)\ln\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right),$$

where $\phi(u)$ is the standard normal *pdf*, $\Phi(u)$ is the standard normal *cdf*, and $\bar{\Phi}(u) = 1-\Phi(u)$. Thus the observed Fisher information matrix with dimension $(p+1)\times(p+1)$ is

$$-\begin{pmatrix} \frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial\boldsymbol{\beta}\partial(\sigma^2)} \\ \frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial(\sigma^2)\partial\boldsymbol{\beta}'} & \frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial(\sigma^2)^2} \end{pmatrix}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{TC}}},$$

where $\hat{\boldsymbol{\theta}}_{\text{TC}}$ is the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$ based on the top coded data $\boldsymbol{w}_{\text{obs}}$. To compute the elements of the Fisher information matrix, we shall use the following properties: $\phi'(u) = -u\phi(u)$ and

$$\frac{\partial}{\partial\sigma^2}\left[\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma})}\right] = \frac{1}{2\sigma^2}\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\left[\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)-\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right].$$

By direct calculations, we get

$$\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\Delta_i(\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta})u_{ij}+\frac{1}{\sigma}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)},$$

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial(\sigma^2)} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Delta_i+\frac{1}{2\sigma^4}\sum_{i=1}^{n}\Delta_i(\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta})^2 \\ &\quad+\frac{1}{2\sigma^3}\sum_{i=1}^{n}(1-\Delta_i)(\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta})\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial\beta_j^2} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\Delta_i u_{ij}^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)} \\ &\quad-\frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}\left[\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right]^2,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial\beta_j\partial\beta_{j'}} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\Delta_i u_{ij}u_{ij'}+\frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}u_{ij'}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)} \\ &\quad-\frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}u_{ij'}\left[\frac{\phi\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i-\boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right]^2,\end{aligned}$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial \beta_j \partial(\sigma^2)} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}\Delta_i u_{ij}(\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}) - \frac{1}{2\sigma^3}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}\frac{\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}$$
$$+ \frac{1}{2\sigma^3}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)^2\frac{\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}$$
$$- \frac{1}{2\sigma^3}\sum_{i=1}^{n}(1-\Delta_i)u_{ij}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\left[\frac{\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right]^2,$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{w}_{\text{obs}})}{\partial(\sigma^2)^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}\Delta_i - \frac{1}{\sigma^6}\sum_{i=1}^{n}\Delta_i(\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta})^2$$
$$- \frac{3}{4\sigma^5}\sum_{i=1}^{n}(1-\Delta_i)\frac{(\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta})\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}$$
$$+ \frac{1}{4\sigma^4}\sum_{i=1}^{n}(1-\Delta_i)\left(\frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)^3\left[\frac{\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right]$$
$$- \frac{1}{4\sigma^4}\sum_{i=1}^{n}(1-\Delta_i)\left(\frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)^2\left[\frac{\phi\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}{\bar{\Phi}\left(\frac{\tilde{x}_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)}\right]^2.$$