TRANSACTIONS ON DATA PRIVACY 11 (2018) 23–45

Statistical Information Recovery from Multivariate Noise-Multiplied Data, a Computational Approach

Yan-Xia Lin*, **Luke Mazur***, **Rathin Sarathy****, **Krishnamurty Muralidhar*****

*National Institution for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia.

**Spears School of Business, Oklahoma State University, Stillwater OK 74078, USA.

***Division of Marketing & Supply Chain Management, Price College of Business, The University of Oklahoma, USA.

E-mail: yanxia@uow.edu.au, lm810@uowmail.edu.au, rathin.sarathy@okstate.edu, krishm@ou.edu

Received 23 Feb. 2017; received in revised form 13th Dec. 2017; accepted 24th Dec. 2017

Abstract. This paper proposes a computational statistical method for multivariate confidential numerical microdata. The method can be employed for recovering some commonly interesting statistical information present in the microdata from noise-multiplied data. Estimating the parameters in linear regression without using the original data directly becomes feasible. This paper demonstrates that some statistical information can be recovered reasonably well for certain types of original data while the level of disclosure risk is under control if the multiplicative noises used to mask the data are appropriate.

This paper presents an alternative approach for sharing the statistical information of multivariate confidential data and carrying out data mining with multidimensional sensitive data, an area of growing interest.

An R package *MaskJointDensity* is built for implementing the method [1].

Keywords. Data Privacy and Confidentiality; Data Masking; Masked Data; Sample-Moment-Based Density Approximent; the Nataf Transformation; the Noise Multiplicative Method; Noise-Multiplied Data

1 Introduction and Motivation

Data collected by government organisations, survey organisations, statistical agencies and health care organisations contain useful statistical information. Releasing the data to the public will bring benefit to policymakers, national economy, and society. Data may be

[1] The software can be downloaded from CRAN or http://www.uow.edu.au/~yanxia/Confidential_data_analysis/

released in two formats: microdata (i.e. a collection of individual records) and tabular data. The release of microdata is often considered more dangerous from the point of view of disclosure risk, but at the same time, the range of statistical analyses may be wider for the microdata compared to tabular data [32].

Statistical disclosure control (SDC) is a balancing act between mandatory data protection, and demand from researchers for access to original data. Statistical disclosure limitation (SDL) methods are used to alter the data while achieving statistical disclosure control. Many SDL methods can be utilized for the purpose of releasing microdata ([10] [13]). The public-use data released by statistical agencies are often in the form of statistical databases, generally transformed to some extent, omitting sensitive information such as personally identifying information, and changing the values of some sensitive data. Data masking is one of the methods used in practice [32]. Data masking techniques include substitution, shuffling, encryption, truncating, noise addition, etc. ([11], [38], [40], [28]).

There are two commonly used/investigated noise masking schemes introduced in the literature: the additive noise (data) masking scheme and the multiplicative noise (data) masking scheme. Multiplicative noise masking method has the advantage that the size of perturbation is proportional to the size of the original value ([41]) and hence it is conjectured that multiplicative noise may better protect the original data than additive noise ([17] and [30]). Using appropriate multiplicative noise to mask a set of original data might optimize the level of protection provided to the original data. For instance, a multiplicative noise which has mixture distributions and takes fewer values around its mean tends to provide a better protection on the values of the original data (See examples in Section4.). While it is not clear whether multiplicative noise has actually been used by a government agency, it has a long and strong history of research by both academic and agency statisticians. Many good sources for the exploration of multiplicative noise in disclosure limitation exist. A couple of recent references include `https://www.census.gov/srd/CDAR/rrs2013-02_Comparison_of_Methods.pdfandwww.census.gov/srd/papers/pdf/rrs2013-01.pdf`. In this paper, we consider the scenario where the microdata are protected through a multiplicative noise masking scheme.

Two types of approaches for retrieving statistical information of the original data based on their respective noise multiplied data are studied in the literature. One of the approaches is to use the masked data directly in estimating the basic statistics of the underlying population, including the mean, variance, moments, covariance, etc., or regression parameters (see [32][17][14][21][35][30] and reference therein).

The other approach is to use the masked data indirectly in estimating the statistical properties of the original data (see [37] [22] [23] and [24]). For instance, under the assumption of a (parametric) structural model $f(y|\theta)$ of the distribution of the original variable y, [37] derived the induced distribution $p(z|\theta)$) of the perturbed variable $Z = YR$ for a specific noise distribution, where R is the multiplicative noise. The z-data can then be used to (efficiently) estimate θ, which in turn can be employed to infer information about the quantiles of the y-distribution. [22] introduced the sample-moment-based density approximant for estimating the density function of the original data based on noise-multiplied data. They then use the synthetic data to retrieve the statistical properties of the original data.

Ideally we prefer to have a technique that can retrieve all the exact statistical information of the original data from the masked data. In practice, there is no such technique at this stage. To obtain the same inference of the original data for specific statistics from the masked data, usually one has to accept a compromise by losing the chance to retrieve the other statistical information of the original data from the masked data. For instance, the techniques proposed by [3],[32] and [29], respectively, can preserve the sample mean, sam-

ple variance, sample covariance or the estimations of linear regression parameters. But those techniques cannot preserve other statistics and cannot maintain the same inference for the subset of the original data. Thereby, the techniques might be less helpful if the data user wants to explore an extensive range of statistical information of the original data and subsets of the original data. Techniques such as [37] [22] [23], [24] and the method proposed in this paper provide the estimation of the statistical information of the original data rather than preserving the exact statistical information of the original data. [2] These techniques give the data user more freedom in exploring the statistical information of the original data based on masked data. In other words, different approaches for retrieving the statistical information of the original data from the masked data have their advantages and limitations. Various purposes of data analysis require using different approaches to mask the original data, thereby, using different methods to retrieve the statistical information of the data.

Regarding estimating the density function of the original data based on the masked data, one might wonder whether a data intruder can use the estimated density function to attack the original data. For original numerical data, we define that the disclosure of the value of a datum occurs if the relative difference between the estimated value of the datum and the actual value of the datum is less than a pre-set criterion (see [21] and [19]). Therefore, a possible way to attack the actual value of a datum by using the information of the estimated density function could be that the data intruder guesses or identifies the value of the datum based on the probability of the disclosure occurred, conditional on the masked datum observed. The larger the probability is, the more likely the data intruder accepts the estimated value as the actual value of the original datum. When the original data are masked by the multiplicative noise method, $(\text{masked datum})/(\text{mean of the multiplicative noise})$ is the unbiased estimator of the value of the original datum. If we use this estimator to estimate the value of the original datum, our experience shows that the risk of the attack can be significantly reduced if the density function of the multiplicative noise which is used to mask the datum takes fewer values around the mean position of the noise. That is why we use this type of multiplicative noise in the simulation studies in Section 4. Our experience also shows that skewed multiplicative noise can provide better protection for the original skewed data. For a set of original data, the probability distribution of a multiplicative noise can have a significant impact on the level of disclosure risk of the dataset. It is essential for the data provider to identify an appropriate set of multiplicative noise for masking the original data in terms of reducing the level of disclosure risk and reducing the amount of data utility loss. The discussion on the topic of the appropriateness is beyond the scope of this paper; we do not discuss it herein.

Regarding estimating the density function of the original data based on masked data, we also have to mention the work of [1], [2], [9] and [15]. Unlike from [22], the techniques introduced by [1], [2], [9] and [15] are for noise-added data. A set of noise-added data (i.e., Original data + additive noise) can be converted to a set of noise-multiplied data by the exponential transformation, (i.e., $\exp$(Original data) $\times$ $\exp$(additive noise)). Conversely, a set of noise-multiplied data (i.e., Original data $\times$ multiplicative noise) can be converted to a set of noise-added data by the logarithm transformation in certain circumstance (i.e., $\log$(Original data) + $\log$ (multiplicative noise)). However, the techniques for noise-multiplied data cannot be replaced by the techniques for noise-added data in general. The techniques developed for different types of masked data rely on different theories and approaches. Thereby, different methods have different advantages. This paper only

[2] Using the techniques of synthetic data to protect the data privacy also cannot preserve the exact statistical information of the original data.

considers techniques for noise-multiplied data.

The weakness of the methods suggested by [37] and [22] is that both approaches are developed for retrieving the statistical information from masked numerical **univariate** data. In this paper, we focus on the technique of retrieving the statistical information of the original data from noise-multiplied multivariate data.

In this paper, we only consider numerical multivariate data. Developing a method for a general type of data is still an open question. The method proposed in this paper can be used to recover the statistical information of marginal distributions of the underlying original data from the noise-multiplied data and to recover the outputs of linear regression analysis of the original data from masked data.

This paper is constructed as follows. In Section 2, the method used to build the framework proposed in this paper is introduced. Section 3 briefly discusses the issues of information loss and disclosure risk. Simulation studies on the method proposed are presented in Section 4. Finally, conclusion and discussion follow.

2 The Method Proposed

In many practical situations, often only limited information such as the marginal distribution and covariance can be practically obtained. The modeling of joint probability distributions of correlated variables based on this limited information remains a challenge. The copula is one of the popular techniques for approximating the joint outcome of variables. Copulas have been used for many applications in the real-world including quantitative finance, civil engineering, reliability engineering, warranty data analysis, turbulent combustion, medicine, hydrology, weather research and random vector generation among others. Copula-based methods for modeling the joint probability distributions of multiple correlated variables have been investigated in the literature and adopted in the practice of data analysis ([39] [4]), including the applications in data privacy protection ([36]).

Based on copula functions, the Nataf transformation is used to handle the dependence of correlated predictor variables and marginal distributions. [20] showed that the Nataf distribution of a random vector could be accurately estimated if marginal cumulative distribution functions and correlation matrix of the random vector is available. [22] introduced the method of estimating the density function of univariate distributions based on noise-multiplied data. Combining these two pieces of work, we propose a computational approach to modeling the joint density function of a random vector based on noise-multiplied data.

Let $X = (X_1, \cdots, X_M)^T$ be a sensitive random vector with continuous marginal probability cumulative functions $\{F_i\}_{i=1}^M$, where the superscript "T" denotes the transpose of a vector or a matrix. The entries of X are also called attributes in the literature. Denote $\{x_{(k)} = (x_{k,1}, \cdots, x_{k,M})^T\}_{k=1}^N$ a set of (original) multivariate data of $X = (X_1, \cdots, X_M)^T$, where $x_{(k)} = (x_{k,1}, \cdots, x_{k,M})^T$ is the kth observation of X.

Multiplicative noise (data) masking scheme: Let $C = (C_1, \cdots, C_M)^T$ be a continuous random vector, independent of X. The entries $C_1, \cdots, C_M$ are **mutually independent** and their probability distributions are not necessarily the same. Using C to mask the dataset $\{x_{(k)} = (x_{k,1}, \cdots, x_{k,M})^T\}_{k=1}^N$ involves two steps: (i) generating a set of sample $\{\tilde{c}_{(k)} = (\tilde{c}_{k,1}, \cdots, \tilde{c}_{k,M})^T\}_{k=1}^N$ with size N from C; (ii) yielding a set of noise-multiplied data $\{x^*_{(k)} = (x^*_{k,1}, \cdots, x^*_{k,M})^T = (x_{k,1}\tilde{c}_{k,1}, \cdots, x_{k,M}\tilde{c}_{k,M})^T\}_{k=1}^N$.

In this paper, only the dataset $\{x^*_{(k)} = (x^*_{k,1}, \cdots, x^*_{k,M})^T\}_{k=1}^N$, together with $\{c_{(k)} = (c_{k,1},$

$\cdots, c_{k,M})^T\}_{k=1}^{N'}$, is available to the public [3], where $\{c_{(k)}\}_{k=1}^{N'}$ is another sample from C with size $N' \gg N$. Since $\{c_{(k)}\}$ is different from $\{\tilde{c}_{(k)}\}$, the actual values of $\{x_{(k)}\}$ cannot be obtained by dividing. We call $X^* = (X_1^*, \cdots, X_M^*)^T = (X_1C_1, \cdots, X_MC_M)^T$ the masked random vector.

2.1 The Nataf Transformation

In this subsection, we briefly introduce how to use the computational method to obtain the Nataf density function of a random vector $X = (X_1, \cdots, X_M)$ when the marginal density functions and the correlation matrix of the random vector are known. Let

$$Z_i = \Phi^{-1}\left[F_i(X_i)\right], \qquad i = 1, 2, \cdots, M,$$

where $\Phi(\cdot)$ is the probability cumulative function of a standard normal distribution. Thus, Z_i is normally distributed with zero mean and unit variance, $i = 1, 2, \cdots, M$. Denote $Z = (Z_1, \cdots, Z_M)^T$ a new random vector.

Using the technique for finding the distribution of a transformation of a random vector, the joint probability density function (pdf) of the random vector X can be modeled by the density function

$$f_{X,Nataf}(x) = \phi_M(z, \rho_0)\frac{f_1(x_1)f_2(x_2)\cdots f_M(x_M)}{\phi(z_1)\phi(z_2)\cdots\phi(z_M)}, \qquad x = (x_1, \cdots, x_M)^T \in R^M \tag{1}$$

where the covariance matrix given by $f_{X,Nataf}$ is forced to be equal to the covariance matrix of X; $z = (z_1, \cdots, z_M)^T \in R^M$; $z_i = \Phi^{-1}\left[F_i(x_i)\right]$, $i = 1, \cdots, M$; $\phi(\cdot)$ is the standard normal pdf; f_i is the pdf of X_i and

$$\phi_M(z, \rho_0) = \frac{1}{\sqrt{(2\pi)^M det(\rho_0)}}\exp(-\frac{1}{2}z^T\rho_0^{-1}z)$$

is the M-dimensional normal pdf with zero mean, unit variance and correlation matrix $\rho_0 = (\rho_{0,ij})_{M\times M}$. This distribution model is referred to Nataf distribution (see [20]).

Denote the correlation matrix of X by ρ_X. The (i, j) entry of ρ_X, i.e. $\rho_{X,ij}$, is the correlation of the ith attribute X_i and the jth attribute X_j. If ρ_X is available, [20] introduced a numerical approach for estimating the (i, j) entry of ρ_0, i.e. $\rho_{0,ij}$, from

$$\rho_{X,ij} = \sum_{l=1}^{m}\sum_{k=1}^{m} P_lP_k(\frac{\tilde{x}_{il} - \mu_i}{\sigma_i})(\frac{\tilde{x}_{jk} - \mu_j}{\sigma_j}), \tag{2}$$

where m is the number of Gaussian points, and P_l and P_k are the corresponding weights (See Appendix A);

$$\begin{aligned}\tilde{x}_{il} &= F_i^{-1}(\Phi(u_{il}^*))\\ \tilde{x}_{jk} &= F_j^{-1}(\Phi(\rho_{0ij}u_{il}^* + \sqrt{1-\rho_{0ij}^2}u_{jk}^*))\end{aligned}$$

and $(u_{il}^*, u_{jk}^*)^T = \sqrt{2}(u_{il}, u_{jk})^T$ defined in Appendix A. [20] also pointed out that it is sufficient to ensure the accuracy of the estimation of ρ_0 if the number of Gaussian points used in (2) is $m = 7$. Therefore, the R package built in this paper uses $m = 7$.

[3] Following the same treatment used in *MaskDensity14* (see [24]), the noise information is released to the public in a binary file.

For convenience, we drop the subscript "Nataf" from the Nataf density function $f_{X,Nataf}$ from now onwards. Recall that the Nataf density function of X will preserve the marginal density information of X and covariance of X, that is, the information of the correlation relationship between the entries of the random vector X.

2.2 Estimating the Joint Probability Density Function based on Masked Data

To estimate the Nataf joint density function f_X in (1), two pieces of information are necessary. One is the information on ρ_X, which is used to estimate ρ_0, and the other is the information of the marginal density functions of X.

The **matrix** ρ_X can be estimated by two methods:

Method 1 Using the sample correlation matrix of $\{x_{(k)} = (x_{k,1}, \cdots, x_{k,M})^T\}_{k=1}^N$ to estimate ρ_X. This method is feasible if the data provider is willing to make the sample correlation matrix of $\{x_{(k)} = (x_{k,1}, \cdots, x_{k,M})^T\}_{k=1}^N$ available to the public. To save space, simulation studies presented in Section 4 do not consider the scenario where ρ_X is estimated by the sample correlation matrix.

Method 2 Using masked sample $\{x^*_{(k)} = (x^*_{k,1}, \cdots, x^*_{k,M})^T\}$ and independent noise sample $\{c_{(k)} = (c_{k,1}, \cdots, c_{k,M})^T\}$ to estimate ρ_X.[4] It is equivalent to estimate $\rho_{X,ij}$, for all $i, j = 1, \cdots, M$.

Recall that the entries of the random vector C are mutually independent. Therefore, the covariance of the ith entry of X and the jth entry of X can be evaluated by using the characteristics of the masked random vector X^* and the characteristics of the multiplicative noise vector C. (i) if $i \neq j$,

$$Cov(X_i, X_j) = Cov(X_i^*, X_j^*)/[E(C_i)E(C_j)]; \tag{3}$$

(ii) if $i = j$,

$$Var(X_i) = \left\{Var(X_i^*) - Var(C_i)[E(X_i^*)/E(C_i)]^2]\right\}/E(C_i^2). \tag{4}$$

The entry $\rho_{X,ij}$ of ρ_X has expression $Cov(X_i, X_j)/\sqrt{Var(X_i)Var(X_j)}$, $i, j = 1, \cdots, M$.

Therefore, ρ_X can be estimated by replacing $Var(X_i^*)$, $Cov(X_i^*, X_j^*)$, $E(C_i)$ and $Var(C_i)$, $i, j = 1, \cdots, M$, with their sample estimators, respectively. In fact, the high-order mixed moments of the attributes can also be estimated by using the sample high-order mixed moments of masked variables and the sample high-order moments of multiplicative noise.

The **estimated marginal density functions** of X can be obtained by applying *MaskDensity14* to $\{x^*_{k,i}\}_{k=1}^N$, $i = 1, \cdots, M$, respectively (see [24]). Use $\hat{f}_{X_i}$ to denote the estimated marginal density function of X_i, $i = 1, \cdots, M$. Thus, the Nataf joint density function f_X can be estimated as follows:

$$\hat{f}_X(x) = \phi_M(z, \rho_0)\frac{\hat{f}_{X_1}(x_1)\hat{f}_{X_2}(x_2)\cdots\hat{f}_{X_M}(x_M)}{\phi(z_1)\phi(z_2)\cdots\phi(z_M)}. \tag{5}$$

[4] Compared to Method 1, Method 2 is more practical and convenient for both data users and data providers, particularly, in the process of exploring the statistical information for subsets of data (see Section 4).

2.3 Simulating Samples from the Estimated Joint Density Function $\hat{f}_X$

We can simulate a sample from the estimated joint density function $\hat{f}_X$. Use the sample to represent the original data of X, and obtain the statistical properties of the original data from the sample.

Many techniques for simulating samples from a (joint) density function can be found from the literature, including Rejection sampling [31], Metropolis-Hastings algorithm ([26], [12]) and the naive technique described in [6].

Given the estimated density function in (5), there is an easy and efficient way to simulate sample data with size N from $\hat{f}_X$. We describe the process below and adopt it in the R package built in this paper:

(1) Simulate a set of multivariate sample data $\{(z_{k,1}, \cdots, z_{k,M})\}_{k=1}^{N}$ from a M-dimensional normal distribution. The distribution has zero mean, unit variance and correlation matrix ρ_o.

(2) Transform $z_{k,i}$ to

$$\tilde{x}_{k,i} = \hat{F}_{X_i}^{-1}(\Phi(z_{k,i})), \quad k = 1, \cdots, N$$

where $\hat{F}_{X_i}$ is the estimated marginal cumulated distribution, $i = 1, \cdots, M$. Then, the multivariate data $\{(\tilde{x}_{k,1}, \cdots, \tilde{x}_{k,M})\}_{k=1}^{N}$ is a sample from $\hat{f}_X$.

We built an R package, named *MaskJointDensity* to implement the method and the process of data simulation described in Sections 2.1-2.3 [5].

3 Information Loss and Disclosure Risk

There are no universal quantitative criteria for evaluating disclosure risk and information loss. [7] and [8] evaluate them by aggregating the results from different measures of disclosure risk and information loss. While this may be appropriate for comparing across different masking procedures, the data provider might have the different focus on disclosure risk and information loss for the different type of data. Given that the primary focus of the paper is not a comparison of different masking methods or assessing the disclosure risk and the information loss, we take a broader, more flexible approach. We merely employ the commonly used criteria for evaluating the value disclose risk. Depending on the particular context, the data provider may choose some or all measures for performing this assessment.

Many studies on the measurements of disclosure risk can be found in the literature (see [33] [34] [42] and reference therein). We use two ways to check the appropriateness of a multiplicative noise for masking the value of an attribute. (i) **Examining the plot of the original data vs. its masked counterpart and evaluating the (sample) correlation coefficient between the original data and its masked data**. The plot of the original data vs. the masked data is not available to the public as the original data is not accessible to the public. However, the data provider can use it to visually check the proportion of the values of the original data which can be accurately identified or estimated from the values of their corresponding masked data. The correlation coefficient is a quantitative measurement on the linear relation between the original data and masked data. Our experience shows that

[5] The software can be downloaded from CRAN or `http://www.uow.edu.au/~yanxia/Confidential_data_analysis/`

the value of the correlation coefficient needs to be controlled under 0.9 ([25]). (ii) **Checking the probability measure of the disclosure risk** The probability measure of disclosure risk is defined as the probability $P(|C/E(C) - 1| < \delta)$, where C is the multiplicative noise under consideration ([19] and [21]). This measure is the conditional probability of the relative distance given the value of the original datum. The δ is decided by the data provider and used to control the relative difference between the value of the original datum and its unbiased estimator. Having a small probability $P(|C/E(C) - 1| < \delta)$ is a necessary condition for an appropriate multiplicative noise C. In this paper, we use $\delta = 0.05$. We focus on values disclosure risk in this paper. As showed in [25], none of the above single criterion can be used alone or dominate the others in terms of identifying an inappropriate multiplicative noise set. Making a balanced judgment using these criteria is necessary.

There are also many discussions on the measures of data utility for masked microdata ([44], [43], [27] and reference therein). [27] suggested that, with noise addition, transformed data (i.e., masked data) has to keep the same statistical properties as the original data. [27] explained that keeping the same statistical properties as the original data in practice means making statistics such as the marginal distribution, mean, variance, standard deviation, covariances, and correlation coefficient the same for both original and perturbed data sets. [44] suggested measures including the measure of the similarity between the distributions of the original data and released data.

As explained in the introduction, ideally we would prefer that there is no information loss, that is all results from analyzing the masked data are identical to the same analyses performed on the original data. However, in practice this is not possible. In this paper we use the method proposed to retrieve the statistical information of the original data, including summary statistics, skewness, and kurtosis of the marginal distribution, and the estimates of regression parameters. We evaluate the information loss by comparing the values of the statistic and its estimate in absolute difference. We also visually compare the plots of the estimated marginal density functions with the marginal density functions of the original data and see the similarity between the plots. This is a more flexible approach that allows the data provider to assess and evaluate information loss in a particular context.

4 Simulation Studies

If X is multivariate normally distributed, the Nataf density function of X will be close to the actual joint density function of X. In this section, we use simulation studies to demonstrate the performance of the method proposed in this paper. Therefore, all X considered in this section are not multivariate normally distributed. While potentially an interesting area for future study, we do not consider the impact of noise based on different measurement criteria of disclosure risk, or for confidential data with different probability distributions. We adopt only the basic measures of disclosure risk described in Section 3.

A set of multiplicative noise with different probability distributions can provide different levels of protection for the original data. The amount of statistical information of the original retrieved from the masked data can be affected by the probability distribution of the multiplicative noise used to mask the original data and the technique used to extract the information from the masked data. The purpose of this section is to use examples to demonstrate the computational method proposed for recovering the statistical information of multivariate data based on their noise-multiplied data. In practice, a set of original data can be masked by different types of multiplicative noise. However, we only focus on, for the original data studied in this section, whether there is an associated multiplicative noise

vector such that (i) the original data can be well protected; (ii) the statistical information of the original data can be reasonably retrieved from noise-multiplied data by using the method proposed. The multiplicative noise met the requirement is considered as an appropriate multiplicative noise for the underlying original data. We do not touch on the issue of whether the noise applied to the original data in this section is the best noise in terms of minimising the information loss and disclosure risk. In practice, the data provider can do his best to search for the best appropriate multiplicative noise for the underlying original data if he/she wants.

The multiplicative noise masking scheme does not protect zero-valued observations. One of the manners for protecting the values of a data set involved zero-valued observations is to transform all the values of the dataset by shifting a constant before data masking (see an example in [21]). We can estimate the density function of the shifted data, then obtained the synthetic data of the shifted original data. Finally, get the synthetic data of the original data by shifting back the synthetic data of the shifted original data. Based on the method of [22] and the method proposed in this paper, the (joint) density function is approximated by the function of moments. Theoretically, shifting data will not cause any issues in estimating the (joint) density function. Practically, replacing theoretical moments by sample moments accordingly might cause some extra error in the estimation. Therefore for different types of data, there are various strategies for selecting an appropriate constant for the shifting. For this paper, we will not pay attention to this issue and we do not apply the shifting strategy to the data studied in this section.

In the discussion below, we use the criteria mentioned in Section 3 to check if the values of the underlying data are protected to a reasonable level. Due to limitations of space, we do not present all the plots of the original data vs. its masked counterpart. It is worth recalling that the motivation of the method proposed is to estimate the (joint) density function of the original data based on the masked data. Then, to simulate the synthetic data from the estimated (joint) density function and to use the synthetic data to retrieve the statistical characteristics of the original data. **Since the synthetic data are different from the original data, we should not expect that the statistical inference given by the synthetic data is the same as those given by the original data.** What we wish to see is the values of a statistic and its estimate are relatively close. The acceptable levels of closeness are varied subject to the content of the underlying data and the expectation of data providers.

An R package *maskJointDensity* is built in this paper. The data provider can use the R function "maskBatch" to produce masked multivariate data and noise.bin binary files for the data user. Inputting the masked multivariate data and noise.bin files to R function "unmaskAndGetSampleBatch", the data user can obtain the synthetic data of individual marginal distributions of the original multivariate data and the synthetic data of the joint distribution of the original multivariate data.

Two simulation studies are presented in this section [6].

Example 1. The original bivariate data $\{(x_{i,1}, x_{i,2})\}_{i=1}^{1000}$ is simulated from a random vector $X = (X_1, X_2)$. Six different random vectors, listed in Table 1, are studied in this example. The random error $e \sim N(0,1)$ is independent of X_1. Except for Model 1, at least one of marginal variables is not normally distributed. The literature studies showed that a multiplicative noise with mixture distributions tends to provide more protection on the original data (see [19] [16] [18] [21]). We use $MixUnif(L = (a_1, a_2), U = (b_1, b_2), p = (p_1, 1-p_1))$ to denote the probability distribution of a random variable

$$V = I_{(W_2=0)}Unif(a_1, b_1) + I_{(W_2=1)}Unif(a_2, b_2)$$

[6] The R code for the simulation study will be available on request

Table 1: Models studied in Example 1

Model 1	$X_1 \sim N(9,2)$ and $X_2 = 0.8X_1 + e$
Model 2	$X_1 \sim N(9,2)$ and $X_2 = 0.8X_1 + 0.2X_1^2 + 8e$
Model 3	$X_1 \sim Gamma(shape = 9, scale = 0.5) + 2$ and $X_2 = 1 + 0.8X_1 + e$
Model 4	$X_1 \sim Gamma(shape = 7.5, scale = 1) + 2$ and $X_2 = 0.8X_1 + e$
Model 5	$X_1 \sim Gamma(shape = 9, scale = 0.5) + 2$ and $X_2 = 0.8X_1 + 0.2X_1^2 + 8e$
Model 6	$X_1 \sim Gamma(shape = 7.5, scale = 1) + 2$ and $X_2 = 1 + 0.8X_1 + e$

Table 2: Correlation coefficient of the masked data and the original Data

	Model1	Model2	Model3	Model4	Model5	Model6
$cor(X_1, X_1^*)$	0.3614051	0.3614051	0.4267313	0.4681106	0.4268818	0.4681106
$cor(X_2, X_2^*)$	0.3967303	0.6224688	0.5357683	0.4365898	0.6878778	0.5632015

where $P(W_2 = 0) = p_1 = 1 - P(W_2 = 1)$. Three independent multiplicative noises are used in this example:

$$\begin{aligned} C_1 &\sim MixUnif(L = (10, 45), U = (30, 80), p = (0.5, 0.5)) \\ C_2 &\sim MixUnif(L = (10, 45), U = (30, 80), p = (0.5, 0.5)) \\ C_3 &\sim MixUnif(L = (10, 45), U = (30, 80), p = (0.7, 0.3)). \end{aligned}$$

We use C_1 to mask X_1 in all Models, C_2 to mask X_2 in Models 1 - 3, and C_3 to mask X_2 in Models 4 - 6.

The probability measures of the disclosure risk of the three noises are approximately equal to 0. The correlation coefficients of X_1 and X_1^*, and X_2 and X_2^* given by the six models are listed in Table 2. All the values of the correlation coefficients are reasonably small and less than 0.9. We only present the plots of the masked data vs. the original of Model 5 (in Figure 1) and virtually check the protection level because the correlation coefficients provided by Model 5 are slightly bigger than others. It turns out that the values of the original data are well protected at a reasonable level, except for the values close to 0. Under the assumption that the data provider considers the data protected to an acceptable standard then, the data provider can release masked multivariate data and associated noise.bin files to the data user.

The data user can use "unmaskAndGetSampleBatch" to obtain the synthetic data for each attribute X_1 and X_2, and the synthetic joint data of (X_1, X_2). The statistical inference of the original data can be estimated by applying the conventional statistical inference methods to the synthetic data.

All the statistical information of the marginal distribution of the original data, including the estimated summary statistics, skewness, and kurtosis, are reasonably retrieved from noise-multiplied data in this example. To save space we do not report them, but report the regression analysis outputs in Tables 3 - 5. We fit the synthetic bivariate data by the four different models listed in Tables 3 - 5 and check whether the regression parameters can be reasonably estimated and whether the data user can correctly identify the actual model of the data when the actual joint distribution is not multivariate normal.

To use *MaskJointDensity* to estimate the marginal density function, it involves a sequence of sampling (see *MaskDensity14* in [24]). To reduce the impact of the randomness on inference results, we independently apply *MaskJointDensity* to the same sets of multivariate noise-multiplied data 50 times and obtain 50 sets of synthetic bivariate data of the origi-

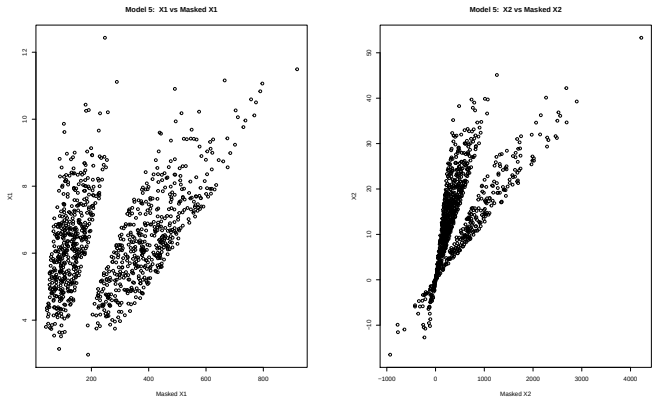

Figure 1: Model 5: the plots of X_1 vs masked X_1 (left), and X_2 vs masked X_2 (right).

nal data then fit the data to the four models, respectively. Tables 3 - 5 report the means of estimations.

Tables 3 - 5 show that, if the actual relationship between X_1 and X_2 is linear, the linear relationship can be revealed by the synthetic data, regardless if the marginal distributions are normal or not. If the actual relationship between X_1 and X_2 is not linear, it will become complicated. If the model building is based on the values R^2 and F-statistic, sometimes synthetic data and the original data might give the same result, although the fitted model is not necessarily the actual model. See the outputs of the data fitted to $X_2 = \beta_1 X_1 + e$ in Model 2, Model 5, and Model 6. The plots of residuals are also beneficial in model building, but are not shown here. Therefore, caution is necessary in interpreting the regression outputs if the actual model is nonlinear and the actual joint distribution of the original multivariate data differs greatly from the multivariate normal distribution.

Example 2. Consider the data given by [28]. The data set with $50,000$ observations was created using the procedure suggested by [5] for generating a multivariate data set with nonnormal marginal distributions. The original data consist of six variables (three nonconfidential and three confidential variables). The three nonconfidential variables (Gender, Marital Status, and Age) are discrete variables. The current version of *MaskJointDensity* is not available for discrete variables. We only consider the data analytic for the three confidential variables in this example. The three continuous variables present Home Value (lognormal), Mortgage Balance (Gamma), and Total Net Asset Value (Normal).

The summary statistics of "Home Value" ("Home" briefly afterward) show that more than 80% observations have the value less than 20 and the maximum value is 32380.00. The plots of the smoothing density function of the three variables are presented in Figure 2. The plot of the density function of "Home" is highly skewed. It is very difficult to estimate the joint density function of the three variables even for the original data. Therefore, in this example we transform the variable "Home" and create a new variable "LHomeNew" which is the logarithm of "Home." No transformation is applied to the other two variables, "Mortgage" and "Value." The new data set has four attributes "LHomeNew", "MortgageNew" and "ValueNew", and the nonconfidential variable "Marital Status". The entries for each attribute are in a column while the observations in the same row belong to the same ID. The observations of those IDs taking non-positive value in the attribute "Home" are dropped from this new data set; therefore the new data set has 49937 observations. The variable

Table 3: Outputs of Model 1 and Model 2

Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + e$		$X_2 = \beta_1 X_1 + e$	
$X_1 \sim N(9,2)$		Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0,1)$	β_0	1.555687 (0.6022119)	-0.35517*	0.7799612(0.01436671)	0.80651***
X_1 masked by C_1	β_1	0.6178469(0.0645241)	0.84402***	0.9570224(0.01294209)	0.9821
X_2 masked by C_2	R^2	0.5085655(0.06952767)	0.7347	23648.14(5471.043)	5.466e+04
$X_2 = 0.8X_1 + e$	F-statistic	1072.717(295.0479)	2764		
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$		$X_2 = \beta_1 X_1 + \beta_2 X_1^2 + e$	
$X_1 \sim N(9,2)$		Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0,1)$	β_0	1.682734(1.115814)	-0.536297	0.9658108(0.07120174)	0.768363***
X_1 masked by C_1	β_1	0.5853288(0.2310711)	0.885833**	-0.0182869(0.006798344)	0.003865*
X_2 masked by C_2	β_2	0.00191177(0.0122461)	-0.002298	0.960108(0.01235704)	0.9821
$X_2 = 0.8X_1 + e$	R^2	0.5100675(0.006946062)	0.7348	12704.11(2679.878)	2.743e+04
	F-statistic	539.1822(148.9154)	1381		
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + e$		$X_2 = \beta_1 X_1 + e$	
$X_1 \sim N(9,2)$		Masked Data	Original Data	Masked Data	Original Data
$e \sim 8N(0,1)$	β_0	-8.363363(3.360144)	-18.6055***		
X_1 masked by C_1	β_1	3.680228(0.3737002)	4.7910	2.804626(0.06382428)	2.8258***
X_2 masked by C_2	R^2	0.4880324(0.0625858)	0.5787	0.88627(0.01512251)	0.8933
$X_2 = 0.8X_1 + 0.2X_1^2 + e$	F-statistic	979.5087(240.3645)	1371	7924.903(1108.053)	8362
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$		$X_2 = \beta_1 X_1 + \beta_2 X_1^2 + e$	
$X_1 \sim N(9,2)$		Masked Data	Original Data	Masked Data	Original Data
$e \sim 8N(0,1)$	β_0	-2.666681(4.340804)	-4.29038	1.689241(0.4096214)	0.54691***
X_1 masked by C_1	β_1	2.293687(0.9008286)	1.48666	0.1106661(0.04188498)	0.23092***
X_2 masked by C_2	β_2	0.07831112(0.05139867)	0.18162***	0.8933439(0.01716335)	0.9153
$X_2 = 0.8X_1 + 0.2X_1^2 + e$	R^2	0.4905648(0.06273843)	0.585	4292.197(743.9691)	5391
	F-statistic	494.3548(121.3622)	702.8		

Table 4: Outputs of Model 3 and Model 4

Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + e$		$X_2 = \beta_1 X_1 + e$	
$X_1 \sim Gamma(k = 9, \theta = 0.5) + 2$		Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0, 1)$	β_0	1.34777(0.7673133)	0.81746***		
X_1 masked by C_1	β_1	0.7172592(0.1079233)	0.82637***	0.9066756(0.0205471)	0.948044***
X_2 masked by C_2	R^2	0.4798417(0.1287141)	0.5788	0.9532772(0.01735666)	0.9733
$X_2 = 1 + 0.8X_1 + e$	F-statistic	1042.822(515.5402)	1371	23433.16(8865.175)	3.646e+04
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$		$X_2 = \beta_1 X_1 + \beta_2 X_1^2 + e$	
$X_1 \sim Gamma(k = 9, \theta = 0.5) + 2$		Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0, 1)$	β_0	-1.460278(1.633087)	0.7871741		
X_1 masked by C_1	β_1	1.51863(0.3692051)	0.8357071***	1.132753(0.1026366)	1.067556***
X_2 masked by C_2	β_2	-0.05346552(0.01988797)	-0.0006851	-0.02965726(0.01291028)	-0.016951
$X_2 = 1 + 0.8X_1 + e$	R^2	0.5019369(0.1376542)	0.5788	0.9582279(0.01452455)	0.9741
	F-statistic	585.9885(318.9057)	684.9	12895.48(4504.874)	1.875e+04
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + e$		$X_2 = \beta_1 X_1 + e$	
$X_1 \sim Gamma(k = 7.5, \theta = 1) + 2$		Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0, 1)$	β_0	2.046339(0.7118936)	-0.02014		
X_1 masked by C_1	β_1	0.573385(0.07828797)	0.80194***	0.7630425(0.03418099)	0.79994***
X_2 masked by C_3	R^2	0.3462(0.08263176)	0.8227	0.9045951(0.021586)	0.9842
$X_2 = 0.8X_1 + e$	F-statistic	552.9797(199.9216)	4631	10035.28(2606.254)	6.212e+04
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$		$X_2 = \beta_1 X_1 + \beta_2 X_1^2 + e$	
$X_1 \sim Gamma(k = 7.5, \theta = 1) + 2$		Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0, 1)$	β_0	1.283613(0.8956606)	-0.0500266		
X_1 masked by C_1	β_1	0.7296791(0.1586336)	0.8082246***	0.9683963(0.06788973)	0.798453***
X_2 masked by C_3	β_2	-0.007267686(0.007644142)	-0.0003059	-0.01747713(0.005490354)	0.000137
$X_2 = 0.8X_1 + e$	R^2	0.3491676(0.08282318)	0.8227	0.910924(0.0197617)	0.9842
	F-statistic	279.9134(100.9762)	2313	5387.15(1336.655)	3.103e+04

Table 5: Outputs of Model 5 and Model 6

Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + e$		$X_2 = \beta_1 X_1 + e$	
$X_1 \sim Gamma(k = 9, \theta = 0.5) + 2$		Masked Data	Original Data	Masked Data	Original Data
$e \sim 8N(0, 1)$	β_0	-8.316816(4.676478)	-10.301***	2.22236(0.1303766)	2.2037***
X_1 masked by C_1	β_1	3.399294(0.6988804)	3.737***	0.6726517(0.0566061 5)	0.7458
X_2 masked by C_3	R^2	0.2443125(0.0714611)	0.304	2116.031(384.5923)	2931
$X_2 = 0.8X_1 + 0.2X_1^2 + e$	F-statistic	333.6984(120.241)	435.9	$X_2 = \beta_1 X_1 + \beta_2 X_1^2 + e$	
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$		Masked Data	Original Data
$X_1 \sim Gamma(k = 9, \theta = 0.5) + 2$	β_0	Masked Data	Original Data	1.196944(0.6024541)	0.5842**
$e \sim 8N(0, 1)$	β_1	-13.4177(8.344308)	-1.70261	0.1369704(0.08285079)	0.2297***
X_1 masked by C_1	β_2	4.836672(2.041151)	1.08566	0.6844251(0.05758757)	0.7652
X_2 masked by C_3	R^2	-0.09435038(0.1280057)	0.19452*	1119.735(216.5102)	1626
$X_2 = 0.8X_1 + 0.2X_1^2 + e$	F-statistic	0.2481659(0.07275868)	0.3076		
		170.3906(62.25458)	221.4		
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + e$		$X_2 = \beta_1 X_1 + e$	
$X_1 \sim Gamma(k = 7.5, \theta = 1) + 2$	β_0	Masked Data	Original Data	Masked Data	Original Data
$e \sim N(0, 1)$	β_1	-4.294402(4.037657)	-18.55777***	3.065216(0.150749)	3.07175***
X_1 masked by C_1	R^2	3.472599(0.4687219)	4.91161***	0.8588557(0.03152121)	0.9663
X_2 masked by C_3	F-statistic	0.4174612(0.1002963)	0.9691	6458.538(1828.002)	2.868e+04
$X_2 = 1 + 0.8X_1 + 0.2X_1^2 + e$		772.002(345.5833)	3.127e+04	$X_2 = \beta_1 X_1 + \beta_2 X_1^2 + e$	
Model fitted		$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e$		Masked Data	Original Data
$X_1 \sim Gamma(k = 7.5, \theta = 1) + 2$	β_0	Masked Data	Original Data	2.488065(0.4012823)	0.993773***
$e \sim N(0, 1)$	β_1	3.901281(4.527405)	0.949973**	0.05046049(0.03988176)	0.191284***
X_1 masked by C_1	β_2	1.754096(0.9919218)	0.808225***	0.8626416(0.03270248)	0.9989
X_2 masked by C_3	R^2	0.08234199(0.0601304)	0.199694***	3363.331(1044.016)	4.686e+05
$X_2 = 1 + 0.8X_1 + 0.2X_1^2 + e$	F-statistic	0.4244269(0.1018562)	0.9944		
		397.7724(178.9396)	8.904e+04		

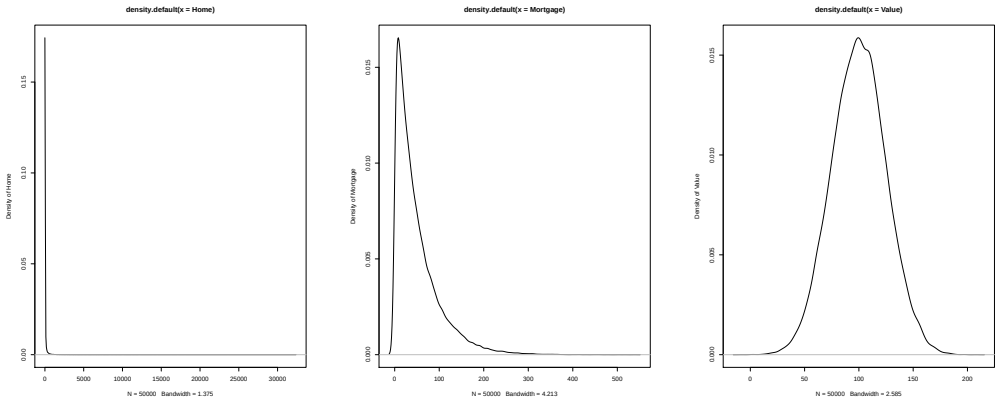

Figure 2: The plots of the density functions of Home Value (Home) (left), Mortgage Balance (Mortgage)(center) and Total Net Asset Value (Value) (right).

Table 6: Three types of Multiplicative noises for data masking

multiplicative noise
CLH: MixUnif(L = (10,50); U = (35, 80); p = (0.5, 0.5))
CM:MixUnif(L = (10,45); U = (30, 80); p = (0.5, 0.5))
CV:MixUnif(L = (10,45); U = (30, 80); p = (0.5, 0.5))

"Marital Status" takes values 1 or 0, indicating married or single.

We use three independent multiplicative noise variables to mask "LHomeNew", "MortgageNew" and "ValueNew", respectively. The distributions of the noise variables are listed in Table 6. The data of "Marital Status" remains the same. Denote the masked variables as "starLH", "starM" and "starV", respectively.

The correlation coefficients of the original data vs. their masked data are 0.8511242, 0.7893962 and 0.397864, respectively. All the probability measures given by the multiplicative noises are approximately equal to 0. The scatter plots of the original data vs. their unbiased estimator (i.e. masked data divided by the mean of noise) are presented by Figure 3. Except for the values near 0, the values of the original data are protected at a certain level. The noise-multiplied data sets could be released to the public. These noise-multiplied data and the data of "Marital Status" make up a multivariate data set. The dataset has four columns with the input names "starLH", "starM" , "starV" and "Marital Status", respectively. The number of rows is 49937.

In this example we want to demonstrate two types of application of *MaskJointDensity*. The first one is about retrieving the statistical information of **the full set of original data** based on **the full set of noise-multiplied data**. The second one is about retrieving the statistical information of **a subset of the original data** based on the data provided by **the full set of noise-multiplied data**. The approach of data masking is a non-interactive approach, and the data user only receives the set of noise-multiplied data of the entire underlying original data. The probability structure of a subset of the original data is not necessarily the same as the probability structure of the whole set of data. Therefore, data mining can receive benefit from the second type of application. In the following study, **the full set**

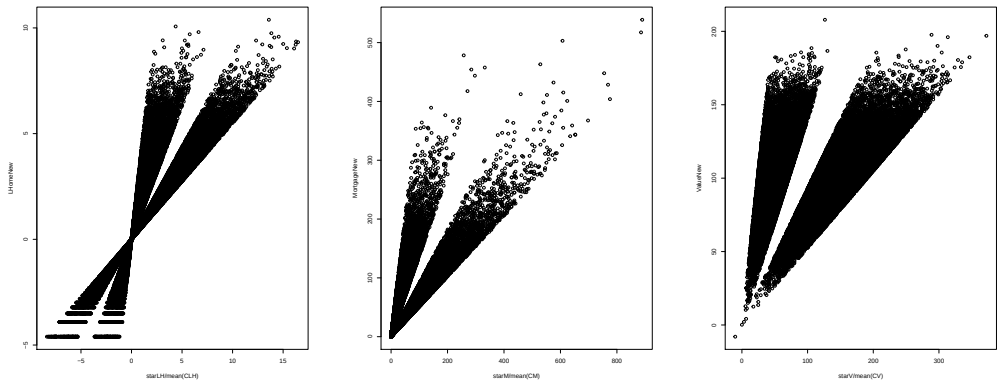

Figure 3: The plots of LHomeNew vs. starLH (left), MortgageNew vs. starM (center), and ValueNew vs. starV (right).

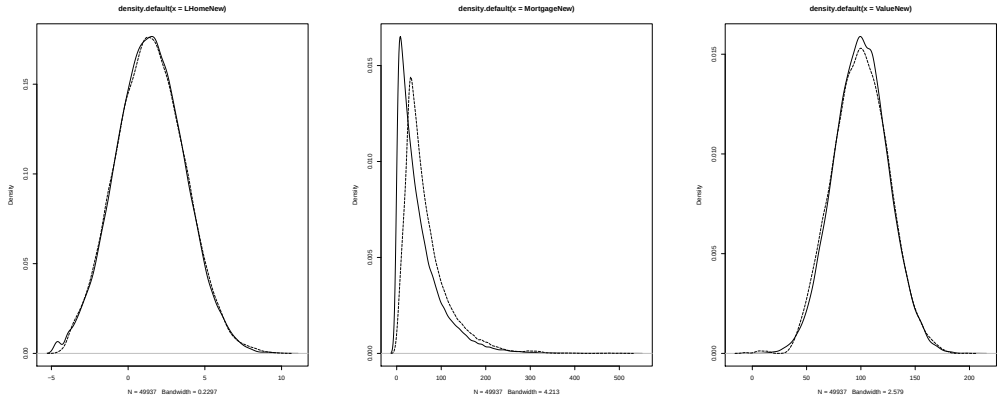

Figure 4: The plots of the density functions of LHomeNew (left), MortgageNew (center), ValueNew (right) and their estimated density functions.

of original data is the dataset of "LHomeNew", "MortgageNew" and "ValueNew"; **the subset of original data** is the dataset of "LHomeNew", "MortgageNew" and "ValueNew" with "Marital Status" taking value 1.

Figure 4 shows the plots of the marginal density functions given by the full set of original data and the estimated marginal density function of the full set of original data. The density function of "MortgageNew" is skewed. The synthetic data captured the characteristic feature, but not very accurately. It is interest to see the impact of the inaccuracy on statistical inference discussed in this example. The basic statistics for the marginal distributions of the full set of data are reported in Table 7.

We independently simulated 100 sets of synthetic data from the estimated joint density function of ($LHmoeNew$, $MortgageNew$, $ValueNew$). The means of the sample correlation coefficients are reported in Table 8.

We fit the full set of original data and each set of independent synthetic joint data to the

Table 7: Summary statistics

	based on the full set of original data			based on the subset of original data		
	LHomeNew	MortgageNew	ValueNew	LHomeNew	MortgageNew	ValueNew
Min.	-4.60500	0.00	-7.97	-4.60500	0.00	-7.97
1st Qu.	-0.09431	14.44	83.05	-0.09431	14.37	84.14
Median	1.41300	34.64	100.10	1.40900	34.70	101.00
Mean	1.42900	50.09	100.10	1.42800	50.07	101.00
3rd Qu.	2.93300	69.03	116.80	2.92800	68.97	117.60
Max.	10.39000	538.80	207.90	10.39000	538.80	197.60
skewness	0.05696396	1.979897	0.01974775	0.06036983	1.986101	0.02110218
kurtosis	2.886455	8.74088	2.951897	2.843275	8.759893	2.958549
	based on the synthetic data for the full set data			based on the synthetic data for subset data		
	LHomeNew	MortgageNew	ValueNew			
Min.	-4.5460	0.00	3.012	-4.60500	0.00	-7.548
1st Qu.	-0.1168	31.63	82.000	-0.05817	31.63	83.270
Median	1.4090	51.67	100.200	1.40900	51.67	101.000
Mean	1.4590	66.96	100.000	1.47000	66.91	100.700
3rd Qu.	2.9630	86.47	117.500	2.96300	86.47	117.900
Max.	9.7990	528.30	207.000	9.68100	503.00	204.100
skewness	0.1281217	2.023997	0.03779607	0.1322291	2.053022	-0.03815748
kurtosis	2.796491	9.48355	2.820405	2.843275	9.770844	3.067751

Table 8: Correlation coefficients of ($LHmoeNew$, $MortgageNew$, $ValueNew$) and the means of the correlation coefficients of synthetic joint data.

	$LHmoeNew$	$MortgageNew$	$ValueNew$
LHomeNew	1	0.5432643	0.6941981
MortgageNew		1	0.718935
ValueNew			1
	SimuLH	SimuM	SimuV
SimuLH	1	0.513514 (0.02240162)	0.6625484 (0.00405837)
SimuM		1	0.6453576 (0.03507253)
SimuV			1

Table 9: The estimates of linear regression parameters

	Full set of data			Subset of data		
	Original data		Synthetic data	Original data		Synthetic data
$\hat{\beta}_0$	-4.3763360	Mean ($\hat{\beta}_0$) (sd) Median($\hat{\beta}_0$)	-3.87023 (0.0678778) -3.874046	-4.4774152	Mean ($\hat{\beta}_0$) (sd) Median($\hat{\beta}_0$)	-4.003642 0.2069711 -4.025206
$\hat{\beta}_1$	0.0040528	Mean ($\hat{\beta}_1$) (sd) Median($\hat{\beta}_1$)	0.006256887 (0.0007046795) 0.006390341	0.0038934	Mean ($\hat{\beta}_1$) (sd) Median($\hat{\beta}_1$)	0.006449104 0.001468772 0.006395167
$\hat{\beta}_2$	0.0559900	Mean ($\hat{\beta}_2$) (sd) Median($\hat{\beta}_2$)	0.04914956 (0.001122921) 0.0489815	0.0565257	Mean ($\hat{\beta}_2$) (sd) Median($\hat{\beta}_2$)	0.04982898 0.002632977 0.0500252
R^2	0.486	Mean(R^2) (sd) Median(R^2)	0.4362 (0.4517501) 0.4511517	0.4883	Mean(R^2) (sd) Median(R^2)	0.4654302 0.02650493 0.4688621

following linear regression models, respectively:

$$LHmoeNew = \beta_0 + \beta_1 MortgageNew + \beta_2 ValueNew + \epsilon, \tag{6}$$

$$SimuLH = \beta_0 + \beta_1 SimuM + \beta_2 SimuV + \epsilon. \tag{7}$$

The regression analysis outputs are reported in Table 9. The 100 sets of synthetic data produced 100 least squares estimates (β_0, β_1, β_2). The mean (sd) and median of the estimates of each regression parameter are presented in Table 9.

The full set of masked data ("starLH", "starM", "starV", "Marital Status") is available to the data user. If the data user is interested in the impact of "Marital Status" on the probability structure of ("LHomeNew", "MortgageNew", "ValueNew"), he/she can conduct two subsets of masked data from the full set of masked data based on the values of "Marital Status" accordingly. We consider the subset of masked data ("starLH", "starM", "starV", "Marital Status"=1) in this example. Recall that, when the data user receives a set of masked data, he/she also receives the files of noise.bin related to the set of masked data. These files of noise.bin contain the information of the multiplicative noises used to mask the underlying data, and they are still valid when we applied *MaskJointDensity* to the subset of masked data.[7]

The basic statistics for marginal distributions of the subset of data are reported in Table 7. The regression analysis outputs for the subset data are reported in Table 9. Following the same way as for the full set of data, we also independently simulated 100 sets of synthetic data from the estimated Nataf joint density function of the subset of data. The 100 sets of synthetic data produced 100 least squares estimates (β_0, β_1, β_2). The mean (sd) and median of the estimates of each regression parameter are also presented in Table 9.

The data analysis outputs showed in Tables 7, 8 and 9 are very impressive. Most of the basic statistical characteristics of the marginal distributions of the original data, the correlation coefficients of the original data and the estimated regression parameters are recovered at a reasonable level in terms of the absolute difference between the values obtained from the original data and the values derived based masked data, respectively.

Discussion and Conclusion:

[7] A detailed discussion on this issue be found from [23].

This paper proposes a computational statistical method for estimating the Nataf joint density function based on noise-multiplied data. An R package *MaskJointDensity* was built as part of this paper. The data provider can use the package to generate the masked data of the underlying original multivariate microdata for the data user. Using the same package, the data user can obtain the estimated marginal density functions and the estimated Nataf joint density function of the underlying original data based on noise-multiplied data. The data user then can generate synthetic data from the estimated (joint) density functions. Those synthetic data are beneficial for exploring the statistical information of the original multivariate data or subsets of the original multivariate data without accessing them. Shuffling also makes use of the multivariate normal copula. However, shuffling treats the whole data set as the population and maintains the marginals as well as the rank-order correlation among the variables, before and after masking. In this sense, shuffling provides an approximation to the joint density of the variables. In the method proposed, the data is not treated as a population. Rather, we consider the data as a sample and use a sample of the population to approximate the joint.

The method proposed in this paper has the following merits:

(i) *Sharing the statistical information of confidential data:* Following the method proposed and the software *MaskJointDensity* designed, the data provider masks each attribute independently. The masked data are released to the data user by each owner of original datasets. Therefore, in preparing the masked data sets for the data user data providers do not need to share the values of the original data among them. [8]

(ii) *Analyzing data with conventional statistical methods:* The statistical information of the original data can be obtained/estimated by applying the conventional statistical methods to the synthetic data of the original data.

(iii) *Generating synthetic data for the subset of the original data:* With a full-set of noise-multiplied data, the data user can explore the joint statistical information of the full-set of original data as well as the joint statistical information of subsets of the original data (see Example 2). Most existing statistical data analysis methods developed for masked data, for instance, data shuffling, or released synthetic data which generated from models based on the full-set of data, do not have such an advantage. This property is beneficial for big data mining, particularly when data privacy issues are involved.

The method proposed in this paper provides an alternative approach for recovering the statistical information of the original data based on noise-multiplied data. Though the method has its merits, it also has its limitations in practice. Since the basic technique of the method relies on the theory of approximation, the method cannot preserve the exact statistical inference results of the original data. In the approach proposed, the theoretical values of the moments of the underlying random variables are replaced by the sample moments accordingly. Thereby, the size of the sample of the underlying original data has an impact on the performance of the method proposed. Additionally, the method proposed combined the two approaches, the Nataf transformation, and Sample-moment-based density approximants. Both of the methods have their limitations. Evidently, the statistical information of most concern such as the information of the marginal distribution and the inference in linear regression, can be reasonably obtained from the synthetic data generated by the

[8] We assume that the original multivariate data are linked. We do not discuss the technique how to link the data collected from different sources in this paper.

Table 10: The typical weights and Gaussian points for Gauss-Hermite quadrature

m	Points $u_{ik}^* = \sqrt{2}u_{ik}$	Weights P_k
1	0	1
2	$\pm$ 1	0.5
3	$\pm$ 1.73205080757	0.16666666667
	0	0.66666666667
4	$\pm$ 2.33441421834	0.045875854768
	$\pm$ 0.74196378430	0.454124145232
5	$\pm$ 2.85697001387	0.011257411328
	$\pm$ 1.3552617997	0.222075922006
	0	0.533333333333
6	$\pm$ 3.32425743359	0.00255578440233
	$\pm$ 1.88917587773	0.088615746029
	$\pm$ 0.616706590154	0.408828469542
7	$\pm$ 3.75043971768	0.000548268858737
	$\pm$ 2.36675941078	0.0307571239681
	$\pm$ 1.1544053948	0.240123178599
	0	0.457142857143

method proposed. However, inference for high order polynomial regression and nonlinear regression based on the synthetic data might not be reliable unless the actual probability distribution of the original data is close to the multivariate normal distribution. Furthermore, when the original data have extreme outliers, missing data or are highly skewed, the estimated marginal density function(s) might be less accurate. The outcomes of the estimation might consequently affect the estimated Nataf density function and the final statistical inference results. Improving the method (and software) to address these constraints will be the focus of our future research. Information loss is inevitable after data is perturbed. Our experience shows that the distribution of the multiplicative noise has some impact on the level of data protection and amount data information retrieved. An appropriate multiplicative noise can provide a good balance in the value disclosure risk and data utility information loss. Developing a general regulation for identifying an appropriate noise set will benefit the applications of the multiplicative noise masking scheme in practice.

MaskJointDensity provides a tool for estimating the marginal density function and the joint density function of the original data. It raises an interesting question: can the data intruder use the estimated (joint) density function to attack the original data? Our research showed the risk of the attack can be reduced if the underlying multiplicative noise is appropriate. We will discuss this issue in another paper.

Preliminaries

Table 10 is from [20]. The number of Gaussian m used in *MaskJointDensity* is 7.

Acknowledgements

Part of the R code in the R package *MaskJointDensity16* was developed by Jordan Morris in his Winter Project in 2014. The Winter Project scholarship is supported by School of Mathematics and Applied Statistics, University of Wollongong.

References

[1] Agrawal, R. and Srikant, R. (2000). Privacy preserving data mining, in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, **29**, 439-450.

[2] Agrawal, D. and Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms, in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, ACM New York, NY, USA, 2001, 247-255.

[3] Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, **13**, 321-327.

[4] Choi, K. K. , Noh, Y. and Du, L. (2007). Reliability Based Design Optimization with Correlated Input Variables Using Copulas. *Proceedings of the ASME 2007 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference* IDETC/CIE 2007 September 4-7, 2007, Las Vegas, Nevada, USA. doi:10.1115/DETC2007-35104

[5] Clemen, R. T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, **45**, 208-224.

[6] Cochran, W. C. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

[7] Domingo-Ferrer, J. and Torra, V. (2001a). Disclosure control methods and information loss for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. V. Zayatz (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, North-Holland, 91-110.

[8] Domingo-Ferrer, J. and Torra, V. (2001b). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. V. Zayatz (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, North-Holland, 111-133.

[9] Domingo-Ferrer, J., Sebé, F. and Castellà-Roca, J. (2004). On the security of noise addition for privacy in statistical databases. In: Domingo-Ferrer J., Torra V. (eds) *Privacy in Statistical Databases*. PSD 2004. Lecture Notes in Computer Science, **3050**, 149-161. Springer, Berlin, Heidelberg

[10] Duncan, G. T., Elliot, M., Juan Jose Salazar, G. (2011). *Statistical Confidentiality - Principles and Practice*, Statistics for Social and Behavioral Sciences, Springer.

[11] Fuller, W. (1993). Masking Procedures for Microdata Disclosure Limitation, *Journal of Official Statistics*, **9**, 383-406.

[12] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. **57**, 97109. JSTOR 2334940. Zbl 0219.65008. doi:10.1093/biomet/57.1.97.

[13] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S. and Nordholt, E. S. (2012). *Statistical Disclosure Control*. John Wiley & Sons, New York.

[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 the American Statistical Association*, **81**, 680-688.

[15] Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K. (2003). On the privacy preserving properties of random data perturbation techniques, *Third IEE International Conference on Data Mining, 2003.* ICDM2003, 22-22 Nov. 2003. DOI 10.1109/ICDM.2003.1250908

[16] Kim, J. J. (2007). Application of Truncated Triangular and Trapezoidal Distributions for Developing Multiplicative Noise. *Proceedings of the Survey Methods Research Section*, American Statistical Association, CD Rom.

[17] Kim, J. J. and Winkler, W. E. (2003). Multiplicative Noise for Masking Continuous Data, Research Report Series (Statistics ♯2003-01), Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233.

[18] Kim, J. and Jeong, D. M. (2008). Truncated triangular distribution for multiplicative noise and domain estimation. Section on Government Statistics-JSM 2008, 1023-1030.

[19] Klein, M., Mathew, T. and Sinha, B. (2014). Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regressopm Samples. *Journal of Privacy and Confidentiality*, **6**, 77-125.

[20] Li, H. S., Lu, Z. Z. and Yuan, X. K. (2008). Nataf transformation based point estimate method, *Chinese Science Bulletin*, **53**, 2586-2592.

[21] Lin, Y.-X. and Wise, P. (2012). Estimation of regression parameters from noise multiplied data, *Journal of Privacy and Confidentiality*, **4**, 55-88.

[22] Lin, Y.-X. (2014). Density approximant based on noise multiplied data. In J. Domingo-Ferrer (ed.) *Privacy in Statistical Databases*. PSD 2014, Lecture Notes in Computer Science, **8744**, 89-104, Springer International Publishing Switzerland.

[23] Lin, Y.-X. and Fielding, M. (2014). Density Approximant Based on Noise Multiplied Data:MaskDensity10.R and its Applications, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, working paper.

[24] Lin, Y.-X. and Fielding, M. (2015). MaskDensity14: An R Package for the Density Approximant of a Univariate Based on Noised Multiplied Data, *SoftwareX*, **34**, 3743 doi:10.1016/j.softx.2015.11.002

[25] Ma, Y., Lin, Y.-X. and Sarathy, R. (2017). The Vulnerability of Multiplicative Noise Protection to Correlational Attacks on Continuous Microdata, working paper, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia.

[26] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-92.

[27] Mivule, K. (2012). Utilizing Noise Addition for Data Privacy, an Overview. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012)*, At Las Vegas, USA, 65-71, doi: 10.13140/2.1.4629.2482

[28] Muralidhar, K. and Sarathy, R. (2006). Data Shuffling - A new Masking Approach for Numerical Data, *Management Science*, **52**, 658-670.

[29] Muralidhar, K. and Sarathy, R. (2012). Perturbation Methods for Protecting Numerical Data: Evolution and Evaluation. *Handbook of Statistics*, **28**, 511-532. DOI10.1016/B978-0-44-451875-0.00019-1

[30] Nayak, T. K., Sinha, B. and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection, *Journal of Official Statistics*, **27**, 527-544.

[31] von Neumann, J. (1951). Various techniques used in connection with random digits. *Journal of Research of the National Bureau of Standards* , Appl. Math. Series, **3**, 36-38.

[32] Oganian, A. (2011). Multiplicative noise for masking numerical microdata with constraints, (Statistics and Operation Research Transction) SORT special issue: Privacy in statistical databases, 99-112.

[33] Reiter, J. P. and Mitra, R. (2009). Estimating Risks of Identification Disclosure in Partially Synthetic Data, *The Journal of Privacy and Confidentiality*, **1**, 99-110.

[34] Reiter, J. P. (2005). Estimating Risks of Identification Disclosure in Microdata, *Journal of the American Statistical Association*, **100**, 1103-1112.

[35] Ruiz, N. (2012). A multiplicative masking method for preserving the skewness of the original micro-records, *Journal of official statistics*, **28**, 107-120.

[36] Sarathy, R., Muralidhar, K. and Parsa, R. (2002). Perturbing non-normal confidential variable: the copula approach. *Management Science*, **48**, 1613-1627.

[37] Sinha, B., Nayak, T.K. and Zayatz, L. (2012). Privacy protection and quantile estimation from noise multiplied data, *Sankhya* B, **73**, 297-315.

[38] Shlomo, N. (2010). Releasing micro data: disclosure risk estimation, data masking and assessing utility, *Journal of Privacy and Confidentiality*, **2**, 73-91.

[39] Tang, X.-S., Li, D.-Q., Zhou, C.-B. and Zhang, L.-M. (2013) Bivariate distribution models using copulas for reliability analysis. In *Proceedings of the Institution of Mechanical Engineers, Part O: J Risk and Reliability*, **227**, 499-512. doi.org/10.1177/1748006X13481928

[40] Templ, M. and Meindl, B. (2008). Robust Statistics Meets SDC: New Disclosure Risk Measures for Continuous Microdata Masking, In: Domingo-Ferrer J., Saygn Y. (eds) *Privacy in Statistical Databases* PSD 2008. Lecture Notes in Computer Science, **5262**, 177-189. Springer, Berlin, Heidelberg

[41] Traub, J., Yemini, Y. and Wozniakowski, H. (1984). Statistical Security of a statistical database. *ACM Transactions on Database Systems*, **9**, 672-679.

[42] Winkler, W. E. (2004). Re-identification Methods for Masked Microdata, Research Report Series (Statistics ♯2004-04) Statistical Research Division U.S. Bureau of the Census)

[43] Winkler, W. E. (2006). Modeling and Quality of Masked Microdata, Research Report Series (Statistics ♯2006-01) Statistical Research Division U.S. Bureau of the Census).

[44] Woo, M.-J., Reiter, J. P., Oganian, A. and Karr, A. F. (2009). Global Measures of Data Utility for Microdata Masked for Disclosure Limitation, *The Journal of Privacy and Confidentiality*, **1** , 111-124.