NIASRA

NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

10-17

Mining the Statistical Information of Confidential Data from Noise-Multiplied Data

Yan-Xia Lin

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

Mining the Statistical Information of Confidential Data from Noise-Multiplied Data

Yan-Xia Lin
School of Mathematics and Applied Statistics
University of Wollongong
Wollongong, Australia
yanxia@uow.edu.au

***Abstract*—Protecting data privacy and mining statistical information from protected data are the essential issues in big data.**

Protecting data privacy through noise-multiplied data is one of approaches studied in the literature. This paper introduces the B-M L2014 Approach for estimating the density function of the original data based on micro noise-multiplied data.

We show an application of the B-M L2014 Approach and demonstrates that the statistical information of the original data can be retrieved from their noise-multiplied data reasonably. The approach provides a new data mining technique for big data when data privacy is concerned.

***Keywords*-data mining; data anonymization; privacy-preserving;**

I. Introduction

In this big data era, an enormous amount of personal and company information can be easily collected by third parties such as national statistical agencies, survey organizations, hospitals, credit card companies and social media. Releasing the data and sharing the statistical information of the data can brings many benefits to policy makers, national economy, and society.

In general, data may be released in two formats: microdata (i.e. collected of individual records) and tabular data. In terms of data mining, it is desirable that the data released to the public is in the format of microdata. Regarding the issue of data privacy, the data "ID" or data identifier (for example, name, credit card number, and address.) has to be removed before the micro data are released to the public. However, this simple way cannot successfully prevent privacy disclosure in general, particularly for a data set consisting of data collected from various sources. Furthermore, due to the sensitivity of the values of some attributes and the obligation between the data agency and the data owner, some data organizations might be not willing or not allowed to make the actual values of some data publicly available. One of the solutions to the privacy disclosure is that the micro data are anonymized before being released to the public.

Micro data can be anonymized by using different manners, including swapping, shuffling, replacing by average value, masking, partially masking, *etc*. (See [3] [5] [11] [12]).

Two commonly used data masking schemes are Additive Noise Data Masking Scheme and Multiplicative Noise Data Masking Scheme.[1]

Let X be a sensitive random variable and $\{x_i\}_{i=1}^n$ be a data set of X. Let C be a random variable, independent of X and $\{c_i\}_{i=1}^n$ be a sample of C. In general, the random variable C can be used as a multiplicative noise or an additive noise. If the data of X are protected by the multiplicative noise C, the noise-multiplied data of $\{x_i\}_{i=1}^n$ is defined as $\{x_i^* = x_i c_i\}_{i=1}^n$; if the data of X are protected by the additive noise C, the noise-added data of $\{x_i\}_{i=1}^n$ is defined as $\{x_i^* = x_i + c_i\}_{i=1}^n$.

The probability density function of a continuous random variable can uniquely determine the statistical information of the random variable. This paper focuses on a technique of estimating the density function of the original data based on noise-multiplied data.

To the best of my knowledge, six papers in the literature considered the techniques for estimating the density function of the original data based on noise-added data or noise-multiplied data in data mining. Among them, [1], [2], [4] and [3] are associated with noise-added data; [8] and [9] are associate with noise-multiplied data. We call [8] the L2014 Approach in this paper hereafter.

[9] proposed a computational method for implementing the L2014 Approach and built an R package *MaskDensity14.R* for the purpose. The software works well in practice. However, there is a drawback of an algorithm adopted in the software. We explain it in Section 2. In this paper, we introduce a new algorithm to replace the algorithm adopted in *MaskDensity14* and give a better approach to implement the L2014 Approach.

This paper is organized as follows. In Section 2, we briefly introduce the L2014 Approach ([8]) and point out the drawback of the algorithm adopted in *MaskDensity14.R* for implementing the L2014 Approach. A new algorithm, used to replace the old one, is proposed in Section 3. A real-life data application is carried out in Section 4. The last section is the conclusion. The proof of Lemma 1 is presented in the Appendix.

In this paper, we denote X a continuous sensitive random variable and $\{x_i\}_{i=1}^N$ a data set of X, which is not available to the public. The multiplicative noise C is a positive random

[1] People might use different names for the schemes in the literature.

variable with a low boundary $c_0 > 0$. Denote the random variable $X^* = XC$. The noise-multiplied data of $\{x_i\}_{i=1}^N$ is $\{x_i^*\}_{i=1}^N = \{x_i c_i\}_{i=1}^N$, where $\{c_i\}_{i=1}^N$ is a sample from C. We also assume that $\{x_i^*\}_{i=1}^N$ and $\{\tilde{c}_i\}_{i=1}^{N'}$ are available to the public, where $\{\tilde{c}_i\}_{i=1}^{N'}$ is another sample of C and the size N' is much larger that N, say $10 \times N$. Since $\{\tilde{c}_i\}_{i=1}^{N'}$ is a sample different from $\{c_i\}_{i=1}^N$, the values of the individuals of $\{x_i\}_{i=1}^N$ cannot be obtained from $\{x_i^*\}_{i=1}^N$ and $\{\tilde{c}_i\}_{i=1}^{N'}$ by dividing. Given the fact that $\{x_i^*\}_{i=1}^N$ and $\{\tilde{c}_i\}_{i=1}^{N'}$ are available to the public, data users can conduct the kernel-smoothed density functions of X^* and C, respectively. To simplify the study, we assume that the density functions of X^* and C are available to the public in this paper.

II. The L2014 Approach and the R package MaskDensity14

[8] introduced the L2014 Approach for estimating the density function of X based on noise multiplied data. The estimated density function of X has the following expression

$$f_{X,K|\{\{x_i^*\}_1^N,\{\tilde{c}_i\}_1^{N'}\}}(x) = \sum_{k=0}^{K} a_k(x) \frac{\overline{(X^*)^k}}{\overline{C^k}}, \qquad x \in R \quad (1)$$

where $\overline{(X^*)^k} = \sum_{i=1}^N (x_i^*)^k / N$; $\overline{C^k} = \sum_{i=1}^{N'} c_i^k / N'$; $a_k(x)$ is a continuous function of x; K is the optimal upper order of moment such that $f_{X,K|\{\{x_i^*\},\{\tilde{c}_i\}\}}$ is the best estimated function of the density function of X, subject to the information of $\{x_i^*\}$ and $\{\tilde{c}_i\}$ and a pre-set criterion.

Determining the value of K is a challenge in implementing the L2014 Approach. [9] developed an R package *MaskDensity14* for implementing the L2014 Approach. In the software, they proposed an algorithm for determining the value of K only based on the public information $\{x_i^*\}$ and $\{\tilde{c}_i\}$. We briefly introduce the algorithm below. The algorithm suggests the following steps for checking if a value of K is optimum. The checking procedure starts from $K = 1$. For a given K, generate a new sample $\{c_j'\}_{i=1}^N$ from C. Use this new sample together with the sample $\{x_j'\}_{i=1}^N$, which are drawn from the density function given by (1), to generate a new set of noise-multiplied data. Record the correlation coefficient of the two sets of sorted data, the set of sorted new noise-multiplied data and the set of sorted publicly available noise multiplied data. Then, update K to $K+1$, and repeat the above process. The procedure will stop if some pre-set criterion meets. The optimal upper order of moment K is determined by the largest value among the sequence of the correlation coefficients. Following the algorithm, the impact of the randomness of the samples $\{c_j'\}_{i=1}^N$ is inevitable, especially, when a "bad" sample $\{c_j'\}_{i=1}^N$, which is a small probability event, is used in the evaluation. To minimize the impact of the randomness on the final inference result, [9] gave an advice in using *MaskDensity14* if the randomness becomes an issue. Firstly, obtain a sequence of estimated density functions of the original data by independently applying *MaskDensity14* to the same noise-multiplied data repeatedly. Then, use the functional mean of the sequence of density functions as the final estimated density function of the original data. The reason used to support the suggestion can be found from [8]. The advice works well in practice.

It is the fact that the impact of the randomness is not very significant when the size of the underlying data is large, say more than 200 or 300. However, the data user might not feel comfortable with the drawback.

III. The new algorithm for determining the optimum upper order of moment K

Assume that $a \le X \le b$ is a bounded continuous random variable and the multiplicative noise $0 < c_0 \le C \le c_1$ is a bounded positive continuous random variable, where a, b, c_0 and c_1 are real numbers.

Let $X^* = XC$, the masked random variable of X. Denote the density functions of X^*, X and C as f_{X^*}, f_X and f_C, respectively. Denote

$$f_{X,K}(x) = \frac{2}{b-a} \sum_{k=0}^{K} \lambda_k P_k\left(\frac{2x-(a+b)}{b-a}\right) \; x \in [a,b] \quad (2)$$

where

$$\lambda_k = \frac{2k+1}{2} \sum_{i=0}^{Floor[k/2]} \frac{(-1)^i 2^{-k} (2k-2i)!}{i!(k-i)!(k-2i)!} \mu_X(k-2i),$$

and

$$P_k(x) = \sum_{i=0}^{Floor[k/2]} (-1)^i 2^{-k} \frac{(2k-2i)!}{i!(k-i)!(k-2i)!} x^{k-2i},$$

the Legendre polynomial of degree k, where $Floor[k/2]$ denotes the largest integer less than or equal to $k/2$ and $\mu_X(k) = E(X^k)$, the kth moment of X. The function $f_{X,K}(x)$ is the polynomial approximation of $f_X(x)$ with order K ([13]). After algebra, we have

$$f_{X,K}(x) = \sum_{k=0}^{K} a_k(x) \frac{\mu_{X^*}(k)}{\mu_C(k)}, \quad x \in [a,b]$$

where $a_k(x)$ is a continuous function of x, $\mu_{X^*}(k) = E[(X^*)^k]$, $\mu_C(k) = E(C^k)$ and $\mu_X(k) = \mu_{X^*}(k)/\mu_C(k)$.

Denote

$$f_{X,K|\{\{x_i^*\}_1^N,\{\tilde{c}_i\}_1^{N'}\}}(x) = \sum_{k=0}^{K} a_k(x) \frac{\overline{(X^*)^k}}{\overline{C^k}} \qquad (3)$$

where $\{\tilde{c}_i\}_1^{N'}$ is a sample of C, $\overline{(X^*)} = \sum_{i=1}^N x_i^*/N$ and $\overline{C^k} = \sum_{i=1}^{N'} \tilde{c}_i^k / N'$. [8] showed that $f_{X,K|\{\{x_i^*\},\{\tilde{c}_i\}\}}(x)$ almost surely uniformly converges to $f_{X,K}(x)$ for $x \in [a,b]$ as $N \to \infty$, where $[a,b]$ is the range of X.

Before introducing the new algorithm to replace the algorithm adopted in *MaskDensity14*, we give the following three facts/results.

(i) *The density function of f_{X^*} can be evaluated through f_C and f_X.*
From the Bayesian theorem, the density function f_{X^*} can be expressed as follows

$$f_{X^*}(x^*) = \int_{c_0}^{c_1} \frac{1}{c} f_C(c) f_X(\frac{x^*}{c})dc, \qquad x^* \in R. \tag{4}$$

(ii) *The density function f_{X^*} can be well approached by $f_{X^*,K}$ as K increases.*
Lemma 1 Denote

$$f_{X^*,K}(x^*) = \int_{c_0}^{c_1} \frac{1}{c} f_C(c) f_{X,K}(\frac{x^*}{c})dc,$$

where $x^* \in$ the domain of X^*, $K = 1, 2, \cdots$.
Then, the sequence of functions $f_{X^*,K}$ converges to f_{X^*} in L_2, as $K \to \infty$.
The proof of Lemma 1 is presented in the Appendix.

(iii) *The sequence of functions $f_{X^*,K|\{x_i^*\}_1^N,\{\tilde{c}_i\}_1^{N'}}$ uniformly converges to $f_{X^*,K}$ on any finite interval, as $N \to \infty$*
Denote

$$f_{X^*,K|\{x_i^*,\tilde{c}\}}(x^*) = \int_{c_0}^{c_1} \frac{1}{c} f_C(c) f_{X,K|\{x_i^*,\tilde{c}\}}(\frac{x^*}{c})dc, \tag{5}$$

where $x^* \in$ the domain of X^*. Employing Equation (3) and the result of (ii), we have $f_{X^*,K|\{x_i^*,\tilde{c}_i\}}$ uniformly converges to $f_{X^*,K}$ on any finite interval, as $N \to \infty$.

[8] proved and demonstrated that $f_{X,K|\{x_i^*,\tilde{c}_i\}}$ can be a good estimation of f_X subject to K is appropriate. From Equations (4) and (5), it concludes that $f_{X^*,K|\{x_i^*,\tilde{c}_i\}}$ can be a good estimation of f_{X^*}, subject to K is appropriate. Thereby, the necessary condition that $f_{X,K|\{x_i^*,c_i\}}$ is a good approximation of f_X is that $f_{X^*,K|\{x_i^*,\tilde{c}_i\}}$ is a good approximation of f_{X^*}. Based on this logic, a new algorithm for determining the appropriate value K in the L2014 Approach is motivated. The algorithm is described below.

(1) Set an initial upper order of moment, K=1 and a maximum upper order of moment to be tested. The maximum upper order of moment set is 100;
(2) Decide n positions in the interval $[\min\{x_i^*\}, \max\{x_i^*\}]$ and denote them as $z_0 = \min\{x_i^*\} < z_1 < \cdots < z_n < z_{n+1} = \max\{x_i^*\}$ such that $z_i - z_{i-1} = \triangle z_i$ are all equal, $i = 1, 2, \cdots, (n+1)$. [2]
(3) Evaluate

$$S(K) = \sum_{j=1}^{n+1} \Big[f_{X^*}(z_j) - f_{X^*,K|\{x_i^*,\tilde{c}_i\}_1^N}(z_j) \Big]^2 \triangle z_j$$

at the positions $\{z_j\}_{j=0}^{j=n+1}$. Keep the track of the optimum upper order of moment K_{opt} such that $S(K_{opt}) = \min_{k \le K} S(k)$.
(4) Update K to $K+1$ and return to (3) if $K+1 \le 100$. Stop when $S(K)$ jumps beyond a threshold taken as $S(K) > 100 \times S(K_{opt})$ or $K+1 > 100$. [3]
(5) Report K_{opt} as the optimum upper order of moment used.

With this new algorithm, we introduce a new method of implementing the L2014 Approach. To distinguish this new method with *Maskdensity14*, we name the new method as the Bayesian-Moment (B-M) L2014 Approach. The B-M L2014 Approach does not involve the sampling process. The estimated density function given by the B-M L2014 Approach is unique determined by $\{x_i^*\}_{i=1}^N$, $\{\tilde{c}_i\}_{i=1}^{N'}$, the partition $\{z_i\}$ on $[\min\{x_i^*\}, \max\{x_i^*\}]$ and the threshold set in the algorithm described above.

Recall that the function $f_{X,K|\{x_i^*,\tilde{c}_i\}}$ becomes a valid estimated density function of f_X is derived subject to the assumption $a \le X \le b$. In real life, the underlying sensitive attribute X is not necessarily bounded. A discussion of the L2014 Approach with non-restriction on the domain of X can be found from [8]. [8] proved that, for the underlying data $\{x_i\}$, the a and b in the expression of $f_{X,K|\{x_i^*,\tilde{c}_i\}}$ can be assigned as $\min\{x_i\}$ and $\max\{x_i\}$, respectively, as long as the sample size N is reasonable large. However, sometimes, the values of $\min\{x_i\}$ and $\max\{x_i\}$ are sensitive and cannot be released directly. [9] proposed a manner for determining the values of a and b in practice and adopted it in the R package *MaskDensity14*. The treatment of determining the values a and b in *MaskDensity14* are adopted in the B-M L2014 Approach. Therefore, the program we developed for the B-M L2014 Approach is the same as the program of *MaskDensity14* except for the part of determining the value of K.[4]

IV. Real-life data application

In this section, we apply the B-M L2014 Approach to a set of real-life data.

Many measurements of the values disclosure risk have been introduced and can be found in the literature (see [14], [15], [16] and reference therein). [11] suggested that, with noise addition, transformed data (i.e. masked data) has to keep the same statistical properties as the original data. [11] 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. [17] proposed measures including the measure of the similarity between the distributions of the original data and released data. We merely employ the commonly used criteria for the evaluation.

[2] The larger the n is, the higher the computation cost will be.

[3] To save time and based on our experience, we use this criterion as threshold for determining the optimal value of K.

[4] The R code for the B-M L2014 Approach will be available on request.

We use two ways to check the appropriateness of a multiplicative noise. (i) **Examining the plot of the original data vs. its masked counterpart and examining the (sample) correlation coefficient between the original data and its masked data** The plot is not available 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 correctly identified or estimated from the masked data. If the correlation coefficient is greater than 0.9, it means that the masked data do not provide sufficient protection for the original data ([10]). (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 ([6] and [7]). In this paper, we use $\delta = 0.05$.

In this paper, we evaluate the information loss by comparing the statistical information of the original data with that retrieved from noise-multiplied data. The statistical information includes the summary statistics, skewness, and kurtosis. We also visually compare the plots of the estimated density function and the original density function. The closer the plots of the two functions are, the less the data utility loss will be.

Example The "Census" dataset considered in this example was obtained on July 27, 2000 using the Data Extraction System of the U. S. Bureau of the Census (http://www.census.gov/DES/www/welcome.html). There are 54 numerical variables arising from the extraction process. In this example, we consider the variable AFNLWGT(Final weight (2 implied decimal places)). The number of observations of AFNLWGT is 1080. The values of observations of AFNLWGT are very large. If the multiplicative noise masking scheme is applied to AFNLWGT directly, the values of masked observations will become too large to be analysed. Thereby, we consider $\log(AFNLWGT)$ instead of $AFNLWGT$ in this study. The smoothed density function of $\log(AFNLWGT)$ is presented in Figure 1

The distribution of $\log(AFNLWGT)$ is skewed left. Based on our experience, skewed multiplicative noise might provide better protection for skewed data. In this example, we use the multiplicative noise with probability distribution Beta(6,1) to mask the data of $\log(AFNLWGT)$. [5]

Before we apply the B-M L2014 Approach to the noise-multiplied data of $\log(AFNLWGT)$, we need to decide the n positions on the range of the noise-multiplied data. For this example, the range of the noise multiplied data is $[3.462, 12.89]$. For such short range, we use $n = 50$.

The scatter plot (Figure 2) shows that $\log(AFNLWGT)$ is well protected by the multiplicative noise. The value of the correlation coefficient of $\log(AFNLWGT)$ and its masked data, and the value of the probability measure of the disclosure risk are 0.2727083 and 0.238889, respectively. Both of them are not large. Regarding data protection, Beta(6,1) is an appropriate multiplicative noise for protecting $\log(AFNLWGT)$.

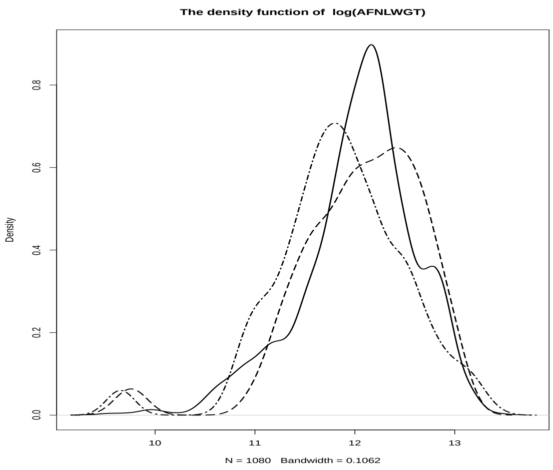

Figure 1. The plots of density function of $\log(AFNLWGT)$ (black line) and its estimated density functions recovered from noise-multiplied data. The plot of the estimated density function given by *MaskDensity14* is in twodash and the plot of the estimated density function is in dotted.

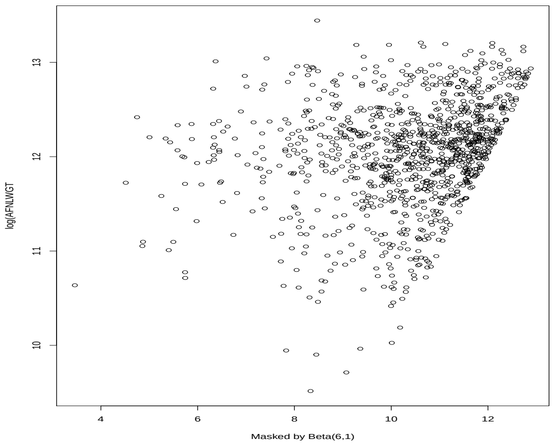

Figure 2. The plots of $\log(AFNLWGT)$ vs the data masked by Beta(6,1).

We applied the B-M L2014 Approach and *MaskDensity14* to the noise-multiplied data of $\log(AFNLWGT)$, respectively. The plots of the estimated density functions are presented in Figure 1. Though the plots of the estimated density functions are not entirely close to the plot of the actual density function of $\log(AFNLWGT)$, the estimated density functions do catch up the information of skewness of the original data and the curve pattern of the actual density function. We simulated sample "$SimuOfX$" and $SimuMaskDensity14$ with size 1000 from the two esti-

[5] For the purpose of this paper, we do not want to spend too much time in searching for the best multiplicative noise for $\log(AFNLWGT)$ in terms of minimizing the utility loss and the level of disclose risk.

Table I
THE SUMMARY STATISTICS, SKEWNESS AND KURTOSIS

Data	Min.	1st Qu.	Median
$\log(AFNLWGT)$	9.515	11.750	12.100
$SimuOfX$(Beta(6,1))	9.577	11.690	12.140
$SimuMaskDensity14$	9.546	11.490	11.860
Data	Mean	3rd Qu.	Max.
$\log(AFNLWGT)$	12.040	12.390	13.440
$SimuOfX$ (Beta(6,1))	12.070	12.510	13.150
$SimuMaskDensity14$	11.86	12.270	13.430
Data	skewness	kurtosis	
$\log(AFNLWGT)$	-0.7130734	3.963749	
$SimutionOfX$ (Beta(6,1))	-1.185353	5.599926	
$SimuMaskDensity14$	-0.5277793	4.432838	

mated density functions, respectively. The summary statistics, skewness, and kurtosis given by $\log(AFNLWGT)$, "$SimuOfX$" and $SimuMaskDensity14$, respectively, are reported in Table I. It confirms that both estimated density functions are beneficial in retrieving the useful commonly used statistical information of $\log(AFNLWGT)$.

However, as we mentioned before, the algorithm adopted in *MaskDensity14* involves the process of random sampling. The randomness might impact on the estimated density function of $\log(AFNLWGT)$. To show the impact, we independently applied *MaskDensity14* to the noise-multiplied data of $\log(AFNLWGT)$ four times. The plots of the four estimated density functions are presented in Figure 3. Though, most of the estimated density functions are close each other. But one estimated density function behaved quick different. The B-M L2014 Approach overcome this drawback of the method of *MaskDensity14*.

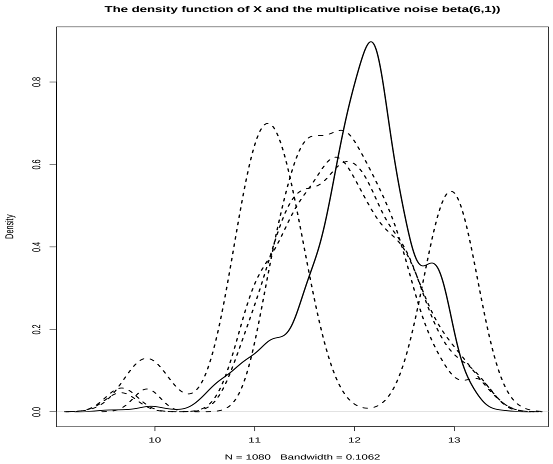

Figure 3. Independently applied *MaskDensity14* to the noise-multiplied data of $\log(AFNLWGT)$ four times. The plots of the estimated density functions.

In summary, the B-M L2014 Approach can help the data user successfully in retrieving the statistical information of $\log(AFNLWGT)$. The multiplicative noise in this example provides good protection on $\log(AFNLWGT)$ and the statistical information retrieved from the masked data is reasonable accurate. Our experience shows that multiplicative noise plays an important role in data protection and information recovery. Due to page limitation, we do not discuss them in this paper.

V. CONCLUSION

Data privacy is an attention issue in big data mining. However, not many existing techniques are available for mining the statistical information of the original data based on protected data. The L2014 Approach is one of such handful methods. The L2014 Approach provides a useful method for data mining while preserving data privacy. [9] developed an R package *MaskDensity14* for implementing the L2014 Approach successfully. This paper proposes a new method, the B-M L2014 Approach, to further improve *MaskDensity14*. The paper gives an application of the B-M L2014 Approach. It shows that method works very well. But, there are spaces to further improve the R code of the the B-M l2014 Approach in terms of computation accuracy. It is what we want to approach in the next step.

With the B-M L2014 Approach, the density function is approximated by Legendre polynomials. The computations involved are mainly the calculation of sample moments. The methods of estimating the density function of the original data based on noise-added data, introduced in [1] and [2] include many integrations. Thereby, the computation cost of running the B-M L2014 Approach is much lower than that of executing the methods of [1] and [2]. Due to the space limit, we do not carry out the comparison studies between those methods and the B-M L2014 Approach in this paper.

APPENDIX

The proof of Lemma 1: [13] pointed that the moment-based density approximant

$$f_{X,K}(x) = \sum_{k=0}^{K} a_k(x) \frac{\mu_{X^*}(k)}{\mu_C(k)}, \quad x \in [a, b]$$

is L_2-convergent to f_X, as $K \to \infty$, that is

$$\|f_X - f_{X,K}\|^2 = \int_a^b (f_X(x) - f_{X,K}(x))^2 dx \to 0,$$

as $K \to \infty$. Denote

$$f_{X^*,K}(x^*) = \int_{c_0}^{c_1} \frac{1}{c} f_C(c) f_{X,K}(\frac{x^*}{c}) dc,$$

where $x^* \in$ the domain of X^*. Since both X and C are bounded, X^* is bounded by $0 < a^* = ac_0/c_1$ and $b^* = bc_1/c_0$. From Hölder's inequality,

$$\|f_{X^*} - f_{X^*,K}\|^2$$
$$= \int_{a^*}^{b^*} \left\{ \int_{c_0}^{c_1} \frac{1}{c} f_C(c) \left[f_X(\frac{x^*}{c}) - f_{X,K}(\frac{x^*}{c}) \right] dc \right\}^2 dx^*$$

$$\leq \int_{a^*}^{b^*} dx^* \int_{c_0}^{c_1} \frac{1}{c^2} \left[f_X(\frac{x^*}{c}) - f_{X,K}(\frac{x^*}{c}) \right]^2 f_C(c)dc.$$

Since $f_{X,K}$ converges to f_X in L_2, for any small $\epsilon > 0$, there is a K_0 such that

$$\|f_X - f_{X,K}\|^2 \leq \epsilon, \quad \text{ as } K > K_0.$$

Thus, as $K > K_0$,

$$\begin{aligned} &\|f_{X^*} - f_{X^*,K}\|^2 \\ \leq &\int_{c_0}^{c_1} \frac{1}{c^2} f_C(c)dc \int_{a^*}^{b^*} \left[f_X(\frac{x^*}{c}) - f_{X,K}(\frac{x^*}{c}) \right]^2 dx^* \\ = &\int_{c_0}^{c_1} \frac{1}{c} f_C(c)dc \int_{a^*/c}^{b^*/c} \left[f_X(t) - f_{X,K}(t) \right]^2 dt \\ \leq &\|f_X - f_{X,K}\|^2 \int_{c_0}^{c_1} \frac{1}{c} f_C(c)dc \leq \frac{\epsilon}{c_0}. \end{aligned}$$

Therefore, $f_{X^*,K}$ converges to f_{X^*} in L_2 as $K \to \infty$..

References

[1] Agrawal, R. and Srikant, R. (2000). Privacy preserving data mining, in Proceedings of the ACM SIGMOD, 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 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 2001.

[3] Domingo-Ferrer, J., Sebé, F. and Castellà-Roca, J. (2004). On the security of noise addition for privacy in statistical databases, J.Domingo-Ferrer and V. Torra (Eds.): PDS2004, LNCS 3050, 149-161.

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

[5] 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.

[6] 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.

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

[8] Lin, Y.-X. (2014). Density approximant based on noise multiplied data. In J. Domingo-Ferrer (ed.), PSD 2014, LNCS **8744**, 89-104, Springer International Publishing Switzerland.

[9] 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

[10] 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.

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

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

[13] Provost, S. B. (2005). Moment-Based Density Approximants, The Mathematica Journal, 9, 728-756.

[14] 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.

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

[16] 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)

[17] 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.