NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

11-14

Density Approximant Based on 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

Density Approximant Based on Noise Multiplied Data

Yan-Xia Lin

National Institute for Applied Statistics Research Australia
School of Mathematics and Applied Statistics
University of Wollongong, Australia

Abstract. Using noise multiplied data to protect confidential data has recently drawn some attention. Understanding the probability property of the underlying confidential data based on their masked data is of interest in confidential data analysis. This paper proposes the approach of sample-moment-based density approximant based on noise multiplied data and provides a new manner for approximating the density function of the underlying confidential data without accessing the original data. The approach of sample-moment-based density approximant is an extension of the approach of moment-based density approximant, which is mathematically equivalent to traditional orthogonal polynomials approaches to the probability density function (Provost, 2005). This paper shows that, regardless of a negligible probability, a moment-based density approximant can be well approximated by its sample-moment-based approximant if the size of the sample used in the evaluation is reasonable large. Consequently, a density function can be reasonably approximated by its sample-moment-based density approximant.
This paper focuses on the properties and the performance of the approach of the sample-moment-based density approximant based on noise multiplied data. Due to the restriction on the number of pages, some technical issues on implementing the approach proposed in practice will be discussed in another paper.

Keywords: Confidential data, Masked data, Multiplicative noise, Moment-based density approximant.

1 Introduction

Many government institutions and statistical agencies collect survey data from individuals and businesses. Publishing these data with certain level of protection is necessary. Many different protection methods, including microaggregation of sensitive data, local suppression of unique data cells, top and bottom coding of continuous variables, rank swapping, rounding, adding noise, imputation and multiplicative noise, have been introduced and used in practice. More information on data protection can be found in Duncan and Lambert (1986 and 1989), Willenborg and De Waal (2001), Oganian (2010), Shlomo (2010), and the references therein.

The aim of government institutions and data agencies publishing the masked data sets is to provide end-users an opportunity to work out the statistical information on the underlying data without breaching confidentiality. As mentioned in Nayak *et al.* (2011), data perturbation may destroy unbiasedness and other properties of estimators. Methods and formulas for analysing an original data set may not be appropriate for analysing a masked version of it.

Describing and estimating the probability density function of a random variable are the basic tenets in statistical data analysis.

Provost (2005) introduced the moment-based density approximant method for probability density approximation. He proved and demonstrated that using the moment-based density approximant to approach the density function is mathematically equivalent to using those orthogonal polynomials, such as the Legendre, Laguerre, Jacobi, and Hermite polynomials.

The multiplicative noise method is one type of noise addition used to perturb and protect confidential data. Kim and Jeong (2008) classified the multiplicative noise scheme into two schemes, Multiplicative Noise Scheme I and Multiplicative Noise Scheme II. The multiplicative noise scheme considered in this paper is Multiplicative Noise Scheme I. It is briefly defined as follows. Let Y be a sensitive random variable with observations $y_1, y_2, \cdots, y_N$ (original data). Let C be a positive random variable, independent of Y. When we say the original data $y_1, y_2, \cdots, y_N$ are masked by C, it means the masked data have the form $y_i^* = y_i \times c_i$ where $\{c_i\}$ is a sample from C. In literature, sometimes it imposes $E(C) = 1$. With this restriction, y^* is an unbiased estimator of y given y. This restriction does not apply to the method proposed in this paper. Therefore, the unbiased estimator of y will be $y^*/E(C)$, given y. Without further explanation, the term "masked data" used in this paper is for "noise multiplied data".

For noise multiplied data, developing appropriate data analysis methods and formulas for different inference purposes is necessary (see Kim and Jeong (2008) for domain estimation, Sinha, *et al.* (2011) for quantile estimation, and Lin and Wise (2012) for linear regression parameters estimation). This paper proposes a method to obtain the density approximant of a sensitive random variable Y based on its masked data.

Many properties of the multiplicative noise method, including evaluation of disclosure risk, confidential protection, moment estimation, linear regression parameter estimation, properties of balanced noise distribution and effects on data quality and privacy protection in context of tabular magnitude data, have been deeply discussed and investigated in literature (Evans, 1996; Evans *et al.*, 1998; Hwang, 1998; Kim and Winkler, 2003; Kim and Jeong, 2008; Oganian, 2010; Krsinich and Piesse, 2002; Nayak, et al., 2011; Sinha, et al., 2011; Lin and Wise, 2012 and Klein and Sinha, 2013). One of the important properties is the moments of Y can be evaluated through the moments of its masked variable Y^* and the moments of the noise C used to mask Y.

With the well developed numerical result of the density approximant provided by Provost (2005) and the nice relationship among the moments of Y, masked variable Y^* and noise C, respectively, the density function of Y can be

theoretically well approximated by the density approximant based on the moments given by Y^* and C. By noting that, only masked data $\{y_i^*\}$ and, in the best scenario, noise information are available in practice, the motivation of this paper is to investigate the properties of the moment-based density approximant of Y if the $E[(Y^*)^k]$ and $E(C^k)$ in the moment-based density approximant are replaced by their corresponding sample moments estimators. The moment-based density approximant with moments replaced by sample moments is called the sample-moment-based density approximant.

This paper derives the formula for the approximant of a density function based on masked data and demonstrates that the density function of a random variable can be well approximated by its sample-moment-based density approximant. Due to the restriction on the number of pages, this paper only shows how the sample-moment-based density approximant $f_{Y,K|\{y_i^*,c_i\}_1^N}$ is built based on masked data $\{y_i^*\}$ and noise sample $\{c_i\}$. Then, carries out relevant simulation studies and a real data application of the approach of the sample-moment-based density approximant. The details of the technique treatment to implement the approach proposed in practice and the issue of risk of disclosure related to the approach will appear in another paper along with a built R package.

The remainder of this paper is organized as follows. From Section 2 to Section 4, we step by step extend the approach of moment-based density approximant with bounded domain to the approach of sample-moment-based density approaximant based on noise multiplied data for a general situation. The formula and properties of the sample-moment-based density approaximant are presented. Simulation studies and a real life data application are given in Sections 5 and 6.

2 Moment-based density approximant: density function with a finite domain $[a, b]$

Provost (2005) provided useful formulas of moment-based density approximant. The formulas and notation are adopted in this paper.

The probability density function of a continuous random variable X, taking values on interval $[-1, 1]$, can be expressed as follows:

$$f_X(x) = \sum_{k=0}^{\infty} \lambda_k P_k(x), \tag{1}$$

where

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

with the $(k-2i)$th moment of X

$$\mu_X(k-2i) = E(X^{k-2i}) = \int_{-1}^{1} x^{k-2i} f_X(x) dx;$$

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

is a Legendre polynomial of degree k in x and $Floor[k/2]$ denotes the largest integer less than or equal to $k/2$.

Denote

$$f_{X,K}(x) = \sum_{k=0}^{K} \lambda_k P_k(x) \tag{3}$$

the polynomial approximation of $f_X(x)$ with order K.

Let Y be a random variable with density function defined on a finite interval $[a, b]$. Its density function and kth moment are denoted by $f_Y(y)$ and

$$\mu_Y(k) = E(Y^k) = \int_a^b y^k f_Y(y) dy, \quad k = 0, 1, \cdots,$$

respectively. Let

$$X = \frac{2Y - (a+b)}{b-a}.$$

The domain of the density function of X is bounded by $[-1, 1]$ and the jth moment of X is

$$\mu_X(j) = \frac{1}{(b-a)^j} \sum_{k=0}^{j} \binom{j}{k} 2^k \mu_Y(k) (-1)^{j-k} (a+b)^{j-k}, \quad j = 0, 1, \cdots.$$

After transformation, by using (3), the approximant of f_Y with order K is given by

$$f_{Y,K}(y) = \frac{2}{b-a} \sum_{k=0}^{K} \lambda_k P_k(\frac{2y - (a+b)}{b-a}). \tag{4}$$

Let Y be masked by a noise C and yield Y^*. By noting that, for $k = 1, 2, \cdots$,

$$\mu_Y(k) = \frac{E[(Y^*)^k]}{E(C^k)} = \frac{\mu_{Y^*}(k)}{\mu_C(k)}$$

and λ_k is a linear functions of $E(Y)^i$, $i \leq k$, the approximant of f_Y with order K can be expressed in terms of the moments of Y^* and C as follows

$$f_{Y,K}(y) = \frac{2}{b-a} \sum_{k=0}^{K} \lambda_k P_k \left(\frac{2y - (a+b)}{b-a} \right) = \sum_{k=0}^{K} a_k(y) \frac{\mu_{Y^*}(k)}{\mu_C(k)}, \tag{5}$$

where $a_k(y)$ is a continuous function of y, $k = 0, 1, \cdots, K$.

Provost (2005) pointed out that "the density approximants so obtained may be negative on certain subranges of the support of their distributions having low density. This will likely occur if an insufficient number of moments are being used. However, by mere inspection of the approximate density plot, we should be able to determine whether a higher degree polynomial ought to be used." It means that it is possible to determine an appropriate order K such that the plot of $f_{Y,K}$ is close to or mimics to the plot of the density function of Y by inspecting the plot of the density function of of Y.

3 Sample-moment-based density approximant: density function with a finite domain $[a, b]$

Assume that Y is bounded between the real numbers a and b. Let $\{y_i\}_1^N$ be a sample of size N drawn from Y and $\{y_i^*\}_1^N$ be the masked data of $\{y_i\}_1^N$, masked by the noise C. Let $\{c_j\}_1^N$ be an independent sample drawn from C, which is not the same sample used to yield $\{y_i^*\}_1^N$.

In this section, the relationship between

$$f_{Y,K|\{y_i^*,c_i\}_1^N}(y) = \sum_{k=0}^{K} a_k(y) \frac{\overline{(Y^*)^k}}{\overline{C^k}} \tag{6}$$

and

$$f_{Y,K}(y) = \sum_{k=0}^{K} a_k(y) \frac{\mu_{Y^*}(k)}{\mu_C(k)}, \tag{7}$$

is evaluated, where $\overline{(Y^*)^k} = \sum_{i=1}^N (y_i^*)^k/N$ and $\overline{C^k} = \sum_{i=1}^N c_i^k/N$, $k = 0, 1, 2, \cdots, K$.

We have the following results:

1. $f_{Y,K|\{y_i^*,c_i\}_1^N}$ *uniformly converges to* $f_{Y,K}$ *almost surely.*
   From the Strong Law of Large Numbers (SLLN), $\overline{(Y^*)^k} \overset{a.s.}{\to} E[(Y^*)^k]$ and $\overline{C^k} \overset{a.s.}{\to} E(C^k)$, as $N \to \infty$. Since $a_k(y)$ is a continuous function of y on $[a, b]$, for each $k = 1, 2, \cdots$, $a_k(y)$ is uniformly continuous on $[a, b]$. Thus, given K fixed,

   $$f_{Y,K|\{y_i^*,c_i\}_1^N}(y) = \sum_{k=0}^{K} a_k(y) \frac{\overline{(Y^*)^k}}{\overline{C^k}} \overset{a.s.}{\to} \sum_{k=0}^{K} a_k(y) \frac{\mu_{Y^*}(k)}{\mu_C(k)} = f_{Y,K}(y),$$

   uniformly for $y \in [a, b]$ as $N \to \infty$.
   Since $f_{Y,K|\{y_i^*,c_i\}_1^N}$ uniformly converges to $f_{Y,K}$, the curve of the function $f_{Y,K|\{y_i^*,c_i\}_1^N}$ will be close to the curve of $f_{Y,K}$, so is to the curve of the density function of f_Y, subject to appropriate K and sample size N.
2. $f_{Y,K|\{y_i^*,c_i\}_1^N}(y)$ *is an approximately unbiased estimator of* $f_{Y,K}(y)$ *for each* $y \in [a, b]$.
   Mood *et al.* (1963) showed that an approximate expression for the expectation of a function $g(W_1, W_2)$ of random variables W_1 and W_2 using a Taylor's series expansion around their means (μ_{W_1}, μ_{W_2}) is given by

   $$\begin{aligned} E[g(W_1, W_2)] \approx\ & g(\mu_{W_1}, \mu_{W_2}) + \frac{1}{2} \frac{\partial^2}{\partial w_2^2} g(W_1, W_2)|_{\mu_{W_1}, \mu_{W_2}} Var(W_2) \\ & + \frac{1}{2} \frac{\partial^2}{\partial w_1^2} g(W_1, W_2)|_{\mu_{W_1}, \mu_{W_2}} Var(W_1) \\ & + \frac{\partial^2}{\partial w_1 \partial w_2} g(W_1, W_2)|_{\mu_{W_1}, \mu_{W_2}} cov(W_1, W_2). \end{aligned} \tag{8}$$

Applying (8) to $E(\overline{(Y^*)^k}/\overline{C^k})$ and noting that $\overline{(Y^*)^k}$ and $\overline{C^k}$ are independent, we obtain

$$E\left[\frac{\overline{(Y^*)^k}}{\overline{C^k}}\right] \approx \frac{E[\overline{(Y^*)^k}]}{E(\overline{C^k})} + \frac{E[\overline{(Y^*)^k}]}{(E(\overline{C^k}))^3} Var(\overline{C^k})$$

$$= E(Y^k) + \frac{1}{N} var(C^k) \frac{E(Y^k)}{[E(C^k)]^3} = E(Y^k) + o(1).$$

Therefore, when N is sufficiently large, we will have

$$E\left[\frac{\overline{(Y^*)^k}}{\overline{C^k}}\right] \approx E(Y^k), \quad k = 1, 2, \cdots, K,$$

and $f_{Y,K|\{y_i^*,c_i\}_1^N}(y)$ is an approximately unbiased estimator of $f_{Y,K}(y)$ for each $y \in [a, b]$.

4 Sample-moment-based density approximant: non-restriction on the domain of the density function

Let $\{y_i\}_{0<i\leq N}$ be a sample drawn from a random variable Y. In this section, we point out the facts that (i) the probability of Y taking values beyond the interval $(\min_{1\leq i\leq N}\{y_i\}, \max_{1\leq i\leq N}\{y_i\})$ can be negligible if the sample size N is reasonable large; (ii) $f_{Y,K|\{y_i^*,c_i\}_1^N}$ could be a good candidate for $f_{Y,K}$ regardless of Y bounded or not, as long as the sample size N is reasonable large.

Lemma 1. *Let $Y_1, \cdots, Y_N$ be i.i.d. random variables, defined on some probability space $(\Omega, \mathcal{F}, P)$, and have the probability distribution of Y. For $\omega \in \Omega$, define*

$$g_{min}^{(N)}(\omega) = g_{min}(Y_1(\omega), \cdots, Y_N(\omega)) = P(Y \leq min_{1\leq i\leq N}\{Y_i(\omega)\})$$

and

$$g_{max}^{(N)}(\omega) = g_{max}(Y_1(\omega), \cdots, Y_N(\omega)) = P(Y \leq max_{1\leq i\leq N}\{Y_i(\omega)\}).$$

Then, for any real number $0 \leq a \leq 1$,

$$P(g_{min}^{(N)} \leq a) = 1 - (1-a)^N \ \text{ and } \ P(g_{max}^{(N)} \leq a) = a^N.$$

Lemma 2. *For $p \in (0, 1)$ and $0 < \alpha < 1$, if $N \geq \log(1-p)/\log(1-\alpha/2)$, then*

$$P(g_{min}^{(N)} \leq \alpha/2) > p \quad \text{and} \quad P(g_{max}^{(N)} > 1 - \alpha/2) > p.$$

The proofs of Lemma 1 and 2 are in the Appendix.

From Lemma 2, given $\alpha = 0.05$, if we wish to have at least $p = 0.975$ probability to ensure $g_{min}^{(N)} \leq \alpha/2$ and $1 - g_{max}^{(N)} < \alpha/2$, the sufficient condition

for N will be $N \geq \log(0.025)/\log(0.975) = 145.703$; for $\alpha = 0.05$ and $p = 0.9975$, the sufficient condition will be $N \geq 237$; for $\alpha = 0.005$ and $p = 0.975$, the sufficient condition will be $N \geq 1474$.

No matter Y is bounded or not, once the sample $\{y_i\}_{i\leq N}$ was drawn from Y, $\{y_i\}_{i\leq N}$ will be bounded. It is of interest that, for a pre-set real number $0 < \alpha < 1$, what size N will ensure that we have a sufficient confidence to claim that the probability $P(\min_{1\leq i\leq N}\{y_i\} \leq Y \leq \max_{1\leq i\leq N}\{y_i\})$ is at least $1-\alpha$.

From Lemma 2, if $N \geq \log(1-p)/\log(1-\alpha/2)$,

$$P[(g_{min}^{(N)} \leq \alpha/2)\bigcap(g_{max}^{(N)} > 1-\alpha/2)] \geq 1-(1-p)-(1-p) = 1-2p.$$

For $\omega \in (g_{min}^{(N)} \leq \alpha/2)\bigcap(g_{max}^{(N)} > 1-\alpha/2)$, we have

$$P(min_{1\leq i\leq N}\{Y_i(\omega)\} \leq Y \leq max_{1\leq i\leq N}\{Y_i(\omega)\}) = 1-g_{min}^{(N)}(\omega)-g_{max}^{(N)}(\omega) \geq 1-\alpha.$$

Therefore, we have at least $1-2p$ confidence to claim that, for sample $\{y_i\}_1^N$,

$$P(min_{1\leq i\leq N}\{y_i\} \leq Y \leq max_{1\leq i\leq N}\{y_i\}) \geq 1-\alpha,$$

if $N \geq \log(1-p)/\log(1-\alpha/2)$.

Example 1. If we wish to have $0.95 = 1-2\times 0.975$ confidence to claim that more than $0.995 = 1-0.005$ chance the values of Y will drop between $\min_{1\leq i\leq N}\{y_i\}$ and $\max_{1\leq i\leq N}\{y_i\}$, the sufficient condition for N is $N \geq 1474 \geq log(1-0.975)/log(1-0.0025)$.

Now, we are at the position of extending the result in Section 3 to sample-moment-based density approximant without the restriction on the domain of the density function.

Assume that Y is a random variable on some probability space $(\Omega, \mathcal{F}, P)$ and $\{y_i\}_{1\leq i\leq N}$ is a sample from Y. Define a random variable $\tilde{Y}$ from Y. Let $\tilde{Y}(\omega) = Y(\omega)$ if $\min_{i\leq i\leq N}\{y_i\} \leq Y(\omega) \leq \max_{i\leq i\leq N}\{y_i\}$; $= 0$ otherwise, where $\omega \in \Omega$. $\tilde{Y}$ is called a truncated random variable of Y.

Following Example 1, if $N > 1474$, with odds of 0.95, the difference between the cumulative distribution functions of Y and $\tilde{Y}$ can be evaluated as following: if $y \leq \min_{i\leq i\leq N}\{y_i\}$,

$$|F_Y(y) - F_{\tilde{Y}}(y)| = F_Y(y) \leq 0.0025;$$

if $\min_{i\leq i\leq N}\{y_i\} < y < \max_{i\leq i\leq N}\{y_i\}$,

$$|F_Y(y) - F_{\tilde{Y}}(y)| = \Big| P(\min_{i\leq i\leq N}\{y_i\} < Y < y) + P(Y \leq \min_{i\leq i\leq N}\{y_i\}) - \frac{P(\min_{i\leq i\leq N}\{y_i\} < Y < y)}{P(\min_{i\leq i\leq N}\{y_i\} < Y < \max_{i\leq i\leq N}\{y_i\})} \Big|$$

$$\leq P(\min_{i\leq i\leq N}\{y_i\} < Y < y)\frac{1-P(\min_{i\leq i\leq N}\{y_i\} < Y < \max_{i\leq i\leq N}\{y_i\})}{P(\min_{i\leq i\leq N}\{y_i\} < Y < \max_{i\leq i\leq N}\{y_i\})} + 0.0025$$

$$\leq (1 - 0.995) + 0.0025 = 0.0075;$$

if $y > \max_{i \leq i \leq N}\{y_i\}$,

$$|F_Y(y) - F_{\tilde{Y}}(y)| = |1 - P(Y > y) - 1| < 0.0025.$$

Thus, with odds of 0.95, $\max_y\{|F_Y(y) - F_{\tilde{Y}}(y)|\} < 0.0075$ if $N \geq 1474$.

In summary, the larger the N is, the more confidence we can ignore the difference between F_Y and $F_{\tilde{Y}}$. Therefore, with a sufficiently large N, the probability density function f_Y can be well approximated by the probability density function of $f_{\tilde{Y}}$, where $\tilde{Y}$ is bounded subject to $\{y_i\}_1^N$.

By ignoring the difference between F_Y and $F_{\tilde{Y}}$, the sample $\{y_i\}_1^N$ can be considered as a sample from $\tilde{Y}$. Regardless of whether or not Y is bounded, its truncated random variable $\tilde{Y}$ is always bounded. Following the discussion in Section 3, given the masked data $\{y_i^*\}$ of $\{y_i\}$ and an independent sample $\{c_i\}$ from C, where $\{y_i\}$ were masked by C, the probability density function $f_{\tilde{Y}}$ of $\tilde{Y}$ can be well approximated by $f_{\tilde{Y},K|\{y_i^*,c_i\}_1^N}$ subject to appropriate K and N.

Therefore, the **normalized $f_{\tilde{Y},K|\{y_i^*,c_i\}_1^N}$ can be used to approximate the density function of Y subject to appropriate K and N, regardless of Y bounded or not**. From now on, without further explanation, $f_{Y,K|\{y_i^*,c_i\}_1^N}$ means $f_{\tilde{Y},K|\{y_i^*,c_i\}_1^N}$ and "density approximant" means "sample-moment-based density approximant".

5 Simulation studies on density approximant based on noise multiplied data

In this section, we use simulation examples to demonstrate the performance of the density approximant based on noise multiplied data.

Example 2. Let $Y = I_{(w=0)}Y_1 + I_{(w=1)}Y_2$, where $Y_1 \sim N(30, 4^2)$, $Y_2 \sim N(50, 2^2)$ and w is a Bernoulli distributed random variable with $P(w = 0) = 0.3$. Let $C = I_{(v=0)}C_1 + I_{(v=1)}C_2$ be the multiplicative noise used to mask Y, where v has Bernoulli distribution with $P(v = 0) = 0.6$; $C_1 \sim N(80, 5^2)$ and $C_2 \sim N(100, 3^2)$.

In this example, three issues are investigated/demonstrated. The first issue is the determination of K, such that $f_{Y,K|\{y_i^*,c_i\}_1^N}$ best presents f_Y. For the sake of convenience, this K is called the (optimal) upper order. The second issue is about the fact that the upper order K is related to the sample $\{y_i\}$ and the sample of noise used to yield $\{y_i^*\}$. The last issue is about the impact of the variance of noise on the density approximant. .

For the first issue, a sample $\{y_i\}_1^{10000}$ were simulated from Y. Then, use an independent sample from C to mask $\{y_i\}_1^{10000}$ and yield $\{y_i^*\}_1^{10000}$. In Figure 1, to save space, we only report the plots of $f_{Y,K|\{y_i^*,c_i\}_1^N}$ for $K = 5, 10, 11$ and 15. With the reference of the true density function of Y (in solid line), it shows that the plot of the density approximant is improved as K increases up to 10 or 11, then gradually run away from the plot of f_Y. We also evaluated the correlation

between f_Y and $f_{Y,K|\{y_i^*,c_i\}_1^N}$ for each K. The correlations corresponding to $K = 5, 10, 11$ and 15 are reported in Table 1. The first raw of summary statistics in the table is given by Y. When $K = 10$ or 11 was used in $f_{Y,K|\{y_i^*,c_i\}_1^N}$, the correlation of between $f_{Y,K|\{y_i^*,c_i\}_1^N}$ and f_Y is higher than those when other Ks were used.

The upper order K can be determined by inspecting the plot of f_Y or the correlation between $f_{Y,K|\{y_i^*,c_i\}_1^N}$ and f_Y, given f_Y is available. Using correlation to determine K is more convenient and easy to program.

Although for $K = 5$ and 15, the performance of $f_{Y,K|\{y_i^*,c_i\}_1^N}$ is not as good as those with $K = 10$ and 11, interestedly, the summaries statistics given by those $f_{Y,K|\{y_i^*,c_i\}_1^N}$ in this example are not different too much from the summary statistics given by Y.

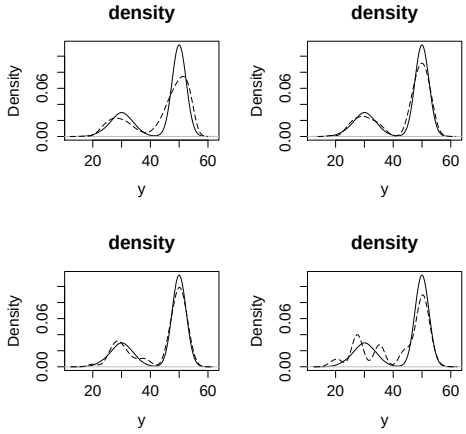

Fig. 1. Left top is for $K = 5$. Right top is for $K = 10$. Left bottom is for $K = 11$ and right bottom is for $K = 15$. The plot of the true density function is in solid line.

Table 1. The summary of statistics and the values of correlation.

data source	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	cor.
y	16.34	33.63	48.83	43.90	50.74	57.70	
$(y1, K = 5)$	16.34	35.36	47.99	44.00	51.46	55.67	0.989
$(y1, K = 10)$	17.80	34.39	48.23	43.88	50.84	57.61	0.9979
$(y1, K = 11)$	16.51	34.79	48.55	43.94	50.90	56.56	0.9978
$(y1, K = 15)$	17.07	32.99	47.42	41.72	50.41	57.70	0.966

For the second issue, two independent samples $\{y_i\}_1^{10000}$ and $\{y_i'\}_1^{10000}$ were simulated from Y. They were independently masked by the noise C. The upper

order Ks determined by the two sets of masked data are 10 and 14 , respectively (corresponding cor. are 0.9975 and 0.9989, respectively). The plots of the density approximants determined by $\{y_i^*\}_1^{10000}$ and $\{y_i'^*\}_1^{10000}$ based on their own upper order K are given in Figure 3 (in the Appendix B). Both of them well present f_Y. This study shows the upper order K is sample related.

For the third issue, a sample $\{y_i\}_1^{10000}$ were simulated from Y. Two multiplicative noises, $|R_1|$ and $|R_2|$, are considered. $R_1 = I_{(v_1=0)}R_{1,1} + I_{(v1=1)}R_{1,2}$ and $R_2 = I_{(v_2=0)}R_{2,1} + I_{(v2=1)}R_{2,2}$ where v_1 and v_2 are independent and have Bernoulli distribution Bernoulli(0.7) and Bernoulli(0.3), respectively; $R_{1,1}, R_{1,2}$, $R_{2,1}$ and $R_{2,2}$ are independent and have normal distributions $N(100, 5^2)$, $N(150, 3^2)$, $N(100, 25^2)$ and $N(150, 18^2)$, respectively. The standard errors given by the samples from $|R_1|$ and $|R_2|$ are 23.36032 and 32.54968, respectively. The summary statistics and the correlations between f_Y and the density approximant based on their own upper order K are reported in Table 2. The plots of density approximants based on their own upper order K are presented in Figure 4(in the Appendix B).

Table 2. The summary of statistics, the values of correlation and the upper order K.

data source	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	cor.	K
y	16.34	33.63	48.83	43.90	50.74	57.70		
y masked by R_1	16.67	34.07	48.23	43.84	50.98	57.61	0.9974	8
y masked by R_2	16.43	32.61	47.90	43.66	51.14	55.67	0.9950	7

From data protection point of view, the larger the variance of the multiplicative noise is, the better protection on the original data the noise will provide. In terms of having a better approximation of the density function of the original data, we might guess or expect that, the larger the variance of the multiplicative noise is, the poor the performance of the associated density approximant will have. However, Figure 4 as well as Table 2 show that, although the ratio of the standard errors of R_2 to R_1 (32.54968/23.36032 = 1.39) is much larger than 1, the difference between the performance of the density approximant given by the data masked by R_1 and R_2, respectively, is not significant. It means that, sometimes, the impact of the variance of noise on the performance of the density approximant might not be significant. It is good in terms of data protection.

6 Real data application

In this section, an example of the density approximant based on real life data is presented.

Example 3. A real life data set taken from the United States Energy Information Authority is considered. This data set can be found in the R package *sdcMicro*, and also available from the United States Energy Information Authority website http://www. eia.doe.govcneaf/electricity/page/eia826.html un-

der year 1996. The data set consists 15 variables generally concerning income and sales data and each of them has 4092 observations.

The smoothing density function given by the data of "othrevenue" is skewed to the right. The majority observations of "othrevenue" are less than 10000 and outliers on the right tail are beyond 60000. There are few observations between values 10000 and 60000. To approximate the smoothing density function of "othrevenue" by the approach of density approximant, the density approximant has to take care the outliers on the right tail as well as a few observations in the interval $[10000, 60000]$. Therefore, the density approximant will shift to the right. If those outliers are removed from the original data set (the number of observations (> 10000) is 96), the density approximant will be more close to the smooting density function given by the original data.

To see the performance of the density approximant, we use two types of noises, a mixture normal noise $C \sim 1/2N(170, 1)+1/2N(120, 1)$ and an identity noise $C \equiv 1$, respectively, to mask the observations of "othrevenue" and yield two sets of masked data for "othrevenue". The set of masked data given by $C \equiv 1$ is the same as the original data set. We evaluate the density approximants given by the two sets of masked data, respectively. Two scenarios of the sets of original data "othrevenue" are considered. One is the full set of data of "othrevenue" and the other is the subset of data with values > 10000 removed. For each scenario, the plots of density approximant given by the two sets of noise multiplied data are presented in Figure 2, respectively.

When $C \equiv 1$, the data used to evaluate the density approximant of "othrevenue" are the unmasked data of "othrevenue". The plot of the density approximant based on the unmasked data is used as a benchmark as it is the density approximant of the density function of the "othrevenue" without any impacts from additional noise perturbation.

From Figure 2, we find the plot of the density approximant given by the data masked by the mixture normal noise is similar to the one given by $C \equiv 1$. The plot related to the mixture normal noise catched as much information on the density function of "othrevenue" as the plot related to $C \equiv 1$ did. Both density approximants shifted to right a bit and showed a fatter tail comparing to the smoothing density function of the original data. The density approximant gives a better approximation to the smoothing density function of the true data after the 96 outliers were removed from the original data set.

The summary statistics given by the density approximants are listed in Table 3 (in the Appendix B). Bother of them, with or without noise perturbation, have successfully show the skewness, the main characteristic in the distribution of the data, though the elements of the summaries are not close to those of the summary statistics given by the data of "othrevenue".

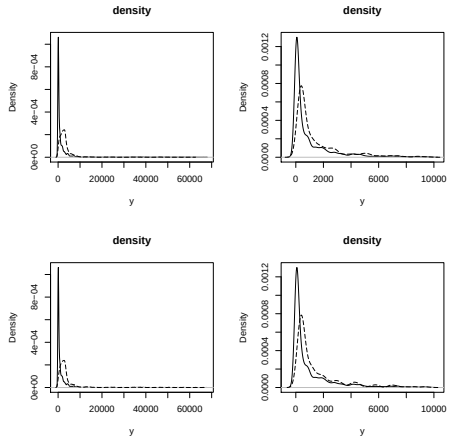

Fig. 2. The density approximants given by data masked by mixture normal distribution and $C \equiv 1$ are shown in the top panels and bottom panels, respectively. For each scenario, the left panel is given by the full data and right one is for data values ≤ 10000. The plots of the smoothing density function of "othrevenue" is in solid line.

7 Discussion

This paper extends the well developed moments-based density approximant approach to the sample-moment-based density approaximant approach based on noised multiplied data.

The motivation of this paper is to develop a fundamental framework for estimating the density function of a sensitive random variable without accessing the original observations of the variable. This work has direct applications to confidential data analysis.

The aim of this paper is to prove and demonstrate that, regardless of a negligible probability, the sample-moment-based density approximant is able to well present its associated density function if the size of the sample from the underlying variable is reasonably large.

With the density function of the underlying variable Y as a reference, we demonstrated that an (optimal) upper order K can be determined such that the sample-moment-based density approximant is close to the density function f_Y. However, if Y is a sensitive variable and it observations are confidential, the information of f_Y will be unavailable in practice. It is of interest how the upper order K can be determined. We have developed a statistical computation searching technique for determining upper order K without the reference of the true density function of the underlying confidential variable. The technique and applications will be discussed in another paper. The method proposed in this paper is developed based on and applies to continuous random variables. The technique can apply to categorical data for approximating mass function and will be presented in another paper.

Acknowledgments. I would like to thank my colleague Dr Gulati for the carefully reading and helpful comments on the early version of this manuscript.

References

1. Duncan, G. T. and Lambert, D. (1986). Disclosure limited data dissemination (with comment), *Journal of the American Statistical Association*, **81**, 1-28.
2. Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata, *Journal of Business and Economic Statistics*, **7**, 207-217.
3. Evans, T.(1996). Effects on Trend Statistics of the Use of Multiplicative Noise for Disclosure Limitation, US Bureau of the Census, *http* : *//www.census.gov/srd/sdc/papers.html*, accessed 5/12/2008.
4. Evans, T., Zayatz, L. and Slanta, J. (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data, *Journal of Official Statistics*, **14**, 537-551.
5. 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.
6. 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.
7. Kim, J.J. and Jeong, D. M. (2008) Truncated triangular distribution for multiplicative noise and domain estimation, Section on Government Statistics - JSM 2008, 1023-1030.
8. Klein, M. and Sinha, B (2013). Statistical analysis of noise-multiplied data using multiple imputation. *Journal of Official Statistics*, **29**, 425-465.
9. Krisinich, F. and Piesse, A. (2002). Multiplicative Microdata Noise for Confidentialising Tables of Business Data: Application to AES99, Data with a Comparison to Cell Suppression, Research and Analytical Report 2002 #19, Statistics New Zealand.
10. Lin, Y.-X. and Wise, P. (2012). Estimation of regression parameters from noise multiplied data, *Journal of Privacy and Confidentiality*, **4**, 55-88.
11. Mood, A.M., Graybill, F.A. and Boes, D.C. (1963). *Introduction to the Theory of Statistics*, McGraw-Hill, Singapore.
12. 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.
13. Oganian, A. (2010). Multiplicative noise protocals. In J. Domingo-Ferrer *et al.* (eds.), *Privacy in Statistical Database* (PSD 2010), *Lecture notes in Computer Science*, Springer-Verlag Berlin Heidelberg, **6344**, 107-117. UNESCO Chair in Data Privacy, International Conference, Corfu, Greece, September 22-24, 2010.
14. Provost, S. B. (2005). Moment-Based Density Approximants, *The Mathematica Journal*, **9**, 728-756
15. Shlomo, N. (2010). Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility, *Journal of Privacy and Confidentiality*, **2**, 73-91.
16. Sinha, B. Nayak, T.K. and Zayatz, L. (2011). Privacy protection and quantile estimation from noise multiplied data, Sankhya B, **73**, 297-315.
17. Willenborg, L and de Waal, T. (2001). Elements of Statistical Disclosure Control, vol. 155 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Appendix A: The proof of Lemmas 1 and 2

The proof of Lemma 1

Let $Y_1, \cdots, Y_N$ be i.i.d random variables on $(\Omega, \mathcal{F}, \mathcal{P})$ and have the same probability distribution as Y.

For each $\omega \in \Omega$, define $g_i(\omega) = P(Y \leq Y_i(\omega))$. Random variables $\{g_i\}$ are i.i.d. and have uniform distribution $U(0,1)$.

For each $\omega \in \Omega$,

$$g_{max}(Y_1(\omega), \cdots, Y_N(\omega)) = P(Y < \max_{1 \leq i \leq N}\{Y_i(\omega)\})$$

$$= \max_{1 \leq i \leq N} P(Y \leq Y_i(\omega)) = \max_{1 \leq i \leq N} g_i(\omega).$$

Therefore, $g_{max}(Y_1, Y2, \cdots, Y_N) = g_{(N)}$ the Nth order statistics of $\{g_i\}$, and

$$P(g_{max}(Y_1, \cdots, Y_N) \leq a) = P(g_{(N)} \leq a) = a^N.$$

Following the similar argument, we have

$$P(g_{min}(Y_1, \cdots, Y_N) \leq a) = P(g_{(1)} \leq a) = 1 - (1-a)^N,$$

where $g_{(1)}$ is the 1st order statistics of $\{g_i\}$.

The proof of Lemma 2

If we wish to have probability at least p to ensure $P(Y \leq \min_{1 \leq i \leq N}\{y_i\}) \leq \alpha/2$ (probability at least p to ensure $P(Y > \max_{1 \leq i \leq N}\{y_i\}) < \alpha/2)$), i.e. $P(g_{min}^{(N)} \leq \alpha/2) \geq p$ (i.e. $P(g_{max}^{(N)} > 1 - \alpha/2) > p$), the sufficient condition is that the sample size N meets the following inequality

$$\int_0^{\alpha/2} N(1-a)^{N-1} da = (1 - (1-\alpha/2)^N) \geq p,$$

$$\left(\int_{1-\alpha/2}^1 N a^{N-1} da = (1 - (1-\alpha/2)^N \geq p \right)$$

i.e.

$$N \geq \log(1-p)/\log(1-\alpha/2).$$

Appendix B: Figures

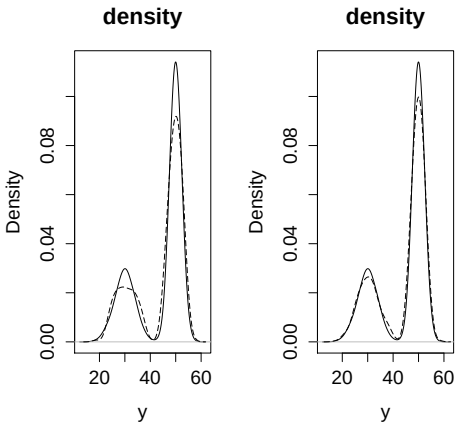

Fig. 3. The left panel is for sample one with $K = 10$ and the right one is for sample two with $K = 14$. The plots of f_Y and density approximant are in solid line and dashed lines, respectively.

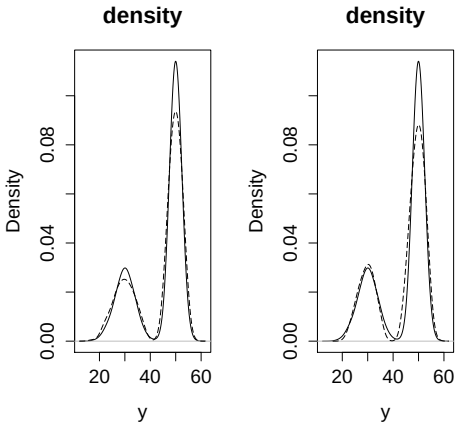

Fig. 4. The density approximant based on the data masked by R_1 is in the left panel and the other one is in the right panel. The plots of f_Y and density approximant are in solid line and dashed lines, respectively.

Table 3. Real data study: the summary of statistics, the values of correlation and the optimal order K.

data source	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	cor.	K
Full Data								
y	-190.0	55.0	255.5	1647.0	1365.0	67520.0		
y masked by mixture normal	-190	1267	2460	3789	3520	64730	0.9583	24
y masked by $C \equiv 1$	-190	1400	2460	3611	3520	67380	0.9810	17
Subset Data								
y	-190.0	53.0	239	995	1187	9853		
y masked by mixture normal	-190	360.3	674.8	1458.0	1874.0	9853	0.9941	48
y masked by $C \equiv 1$	-72.08	380	674.80	1460.0	1834.0	9853	0.9941	22