DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 671–694, http://dx.doi.org/10.2478/JOS-2018-0032

Estimation of True Quantiles from Quantitative Data Obfuscated with Additive Noise

Debolina Ghatak[1] *and Bimal Roy*[1]

Privacy protection and data security have recently received a substantial amount of attention due to the increasing need to protect various sensitive information like credit card data and medical data. There are various ways to protect data. Here, we address ways that may as well retain its statistical uses to some extent. One such way is to mask a data with additive or multiplicative noise and revert to certain desired parameters of the original distribution from the knowledge of the noise distribution and masked data. In this article, we discuss the estimation of any desired quantile of a quantitative data set masked with additive noise. We also propose a method to choose appropriate parameters for the noise distribution and discuss advantages of this method over some existing methods.

Key words: Data obfuscation; quantile estimation; additive noise.

1. Introduction

In official statistics, the main goal of most studies is to analyze a data set to extract different statistics like mean, median, variance and so on, which may help in various statistical analyses. However, in case the data is sensitive (e.g., income data, medical data, marksheet data, etc.), it may be completely impossible to publish it in its raw form. In such cases, statistical agencies often release a masked version of the original data, sacrificing some information. Data obfuscation refers to the type of data masking where some useful information about the complete data set remains even after hiding the individual piece of sensitive information. Therefore, the main objectives of data obfuscation are (i) to minimize the risk of disclosure resulting from providing access to the data, and (ii) to maximize the analytic usefulness of the data.

There are various ways of obfuscating data, such as “Top-coding”, “Grouping”, “Adding Noise”, “Rank Swapping”, and so on. A detailed discussion on various ways of obfuscating sensitive data may be found in the papers by Fuller (1993) and Kim and Karr (2013). Here, we deal with the obfuscation of data using multiplicative or additive noise. A typical problem involves a true quantitative data set $X_1, X_2, \ldots, X_n$; $Y_1, Y_2, \ldots, Y_n$ is a random sample from some known continuous distribution $F(\cdot)$, drawn independent of $\{X_i,\ 1 \leq i \leq n\}$. Then the noised data looks as follows:

$$Z_i = X_i + Y_i,\quad i = 1, 2, \ldots, n \quad \text{(Additive Noise Model), or} \tag{1}$$

$$Z_i = X_i Y_i,\quad i = 1, 2, \ldots, n \quad \text{(Multiplicative Noise Model)} \tag{2}$$

[1] Indian Statistical Institute, Applied Statistics Unit, p. 8. Basudebpur Sarsuna Main Road, Kolkata 700108, India. Emails: deboghatak@gmail.com and bimal@isical.ac.in

In case $\{X_i, 1 \leq i \leq n\}$ is known or assumed to follow a certain distribution, it is enough to estimate the parameters of the distribution as discussed in the papers by Fuller (1993), Mukherjee and Duncan (1997), and Kim and Karr (2013). If there is no distributional assumption on $\{X_i, 1 \leq i \leq n\}$, except that it is continuous, estimating statistics like mean, variance or raw moments from a multiplicative noise model were studied by Zayatz et al. (2011). However, the estimation of nonpolynomial statistics like quantiles may be a problem of concern. Some Bayesian methods to do the same were discussed in the article by Sinha et al. (2011). In the article by Poole (1974), he discussed the estimation procedure of the Distribution Curve of the true population from the data collected through randomized response, randomized with multiplicative noise of a particular form.

However, in all the above cases, authors have mainly concentrated on estimating the quantiles from data, obfuscated with multiplicative noise. In our problem, we work on estimating the quantiles in case the noise is additive instead of multiplicative. The goal of our study is to suggest a procedure with "reasonable" masking of the data set that may return a "good" guess of the quantiles, (one would prefer if estimation procedures of other statistics like mean, variance and so on, are also not harmed by the suggested method). We find an estimate of the distribution function for Normal, Laplace and Uniform errors that may be equated to $0 < \alpha < 1$ to find the required quantiles. A similar problem was discussed by Fan (1991) on a more general basis, popularly known as the deconvolution problem. However, we present an alternative way to look at the problem. We also propose (see subsec. 2.5) a technique for choosing the parameter for the noise distribution (statement may be found in Proposition 2.4). This is a modest attempt at solving the problem stated in the first paragraph of the introduction.

In Section 2, we describe our procedure with required proofs in the Appendix section, and in Section 3, we give some simulation results in support of our procedure. In Section 4, we give a real life example for further illustration. Finally, in Section 5, we conclude with some discussions on the whole procedure.

2. Additive Noise Model: Obfuscation and Estimation

We have a data set $\{X_i, 1 \leq i \leq n\}$ that is sensitive and hence cannot be released. We add an error $\{Y_i, 1 \leq i \leq n\}$ to each value in the data set that comes from some known distribution with a cumulative distribution function $F(\cdot)$. $Z_i = X_i + Y_i$ is the released data known as obfuscated or masked data. $F(\cdot)$ is the obfuscating distribution.

Let $G(\cdot)$, $H(\cdot)$ be the cumulative distribution functions of X and Z, respectively. We assume that (i) X and Y are independent, and (ii) X and Y (and hence Z) are continuous random variables.

Our aim is to find the quantiles of X from the knowledge of Z and $F(\cdot)$. Since we are interested in all the quantiles, we may try estimating the whole distribution curve $G(\cdot)$ of X, which can be used to find the required quantiles.

2.1. *Basic Problem*

Since the problem is to estimate the distribution function of X, one may first think of writing the cumulative distribution function of X, $G(\cdot)$ in terms of $H(\cdot)$ and $F(\cdot)$. But that will not be convenient, since Z and Y are not independent. Instead, we try writing $H(\cdot)$ in terms of the others. For any real number z,

$$\begin{aligned} H(z) &= P(Z \leq z) \\ &= P(X + Y \leq z) \\ &= \int_{-\infty}^{\infty} P(X + Y \leq z | Y = y) f(y) dy \end{aligned}$$

where $f(\cdot)$ denotes the probability density function of Y. Since X and Y are independent, we may write

$$\begin{aligned} H(z) &= \int_{-\infty}^{\infty} P(X \leq z - y) f(y) dy \\ &= \int_{-\infty}^{\infty} G(z - y) f(y) dy \end{aligned}$$

Thus our main equation is,

$$H(z) = \int_{-\infty}^{+\infty} G(z - y) f(y) dy. \tag{3}$$

This is an integral equation with an infinite range, where $G(\cdot)$ is the unknown function to be solved, for f is known and $H(\cdot)$ is to be estimated from the data. Note that our equation says that H is a convolution of f and G. It can alternatively be written as

$$H(z) = \int_{-\infty}^{+\infty} f(z - y)\, G(y) dy \tag{4}$$

Various methods are known to solve integral equations of different kinds. In the following subsections, we will deal with some special cases that arise in practical life. Forms of estimated $G(x)$ are given for Uniform, Normal and Laplace Error (all assumed to have zero mean). Gaussian Kernel and Silverman's Rule of Thumb bandwidth were used to estimate the densities. Then these forms of $\hat{G}(x)$ are equated to $0 < \alpha < 1$, to find the αth quantile of X. Moreover, we discuss (see subsec. 2.5) the choice of appropriate parameters of the Error Distributions, which minimize the risk of disclosure and error in estimation. As far as we know, this is a novel approach to the stated purpose.

2.2. *Uniform Error*

The following result holds if Y is *Uniform(0,a)*; that is, if the density function of Y is of the following form,

$$f(y) = \begin{cases} 1/a, 0 < y < a \\ 0, \ otherwise. \end{cases}$$

Lemma 2.1. *If $h(\cdot)$ is the density function of the obfuscated variable Z, then $\forall x \in R$*

$$G(x) = ah(x) + ah(x - a) + ah(x - 2a) + \cdots$$

In our problem, $h(\cdot)$ is unknown; instead, we can use the kernel density estimate of $h(\cdot)$ to get an estimate $\hat{G}(x)$ of $G(x)$ for all $x \in R$. Then, equating $\hat{G}(x) = \alpha$ for $0 < \alpha < 1$ we get the αth quantile of X.

Note: If Y has 0 mean, i.e., $Y \sim Uniform\left(-\frac{a}{2}, \frac{a}{2}\right)$, the form of $G(x)$ becomes

$$G(x) = ah\left(x - \frac{a}{2}\right) + ah\left(x - \frac{3a}{2}\right) + ah\left(x - \frac{5a}{2}\right) + \cdots$$

in a similar way.

2.3. Normal Error

Here $f(x) = \phi_\sigma(x) = \phi\ (x, 0, \sigma^2)$ for $x \in R$, where $\phi(x, \mu, \sigma^2)$ is the normal density at point x with mean μ and variance σ^2.

Note that if the mean is $\mu \neq 0$ then

$$Z = X + Y \Rightarrow Z - \mu = X + (Y - \mu),\ Y - \mu \text{ has mean } 0,\ Z - \mu \text{ is known.}$$

So without loss of generality, the mean can be assumed to be zero. The following Lemma 2.2 gives an estimated form of the distribution function of X.

Before stating the next Lemma, we introduce the following assumption

(A1) The probability densities of X and Y are bounded.

We also let $\Phi(x, \mu, \sigma^2)$ denote the cumulative distribution function of the normal distribution with mean μ and variance σ^2, evaluated at the point x.

Lemma 2.2. *Assume that assumption (A1) holds. Then if $Y \sim N(0, \sigma^2)$, an estimate of $G(x)$ is,*

$$\hat{G}(x) = \frac{1}{n}\sum_{j=1}^{n} \Phi\left(x - Z_j, 0, \sqrt{b^2 - \sigma^2}\right), \quad \forall x \in R,\ b > \sigma$$

where $b = 1.06n^{-1/5}A$,

$$A = Min\left(\sqrt{\widehat{Var}(Z)}, \frac{IQR(Z)}{1.34}\right)$$

$$\widehat{Var}(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2, \quad \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

and,

$$IQR(Z) = \textit{Interquantile range of } Z = \textit{Third quartile of } Z - \textit{First quartile of } Z.$$

Note: The restriction on σ makes the result very weak since in most cases $b > \sigma$ is not likely to happen. However if one uses a different Kernel to estimate the density, the restriction may not hold in such cases. In the next subsection, we would like to suggest an alternative way to deal with this problem such that there is no bound on the choice of σ.

2.4. Laplace Error

The main reason behind the choice of such Error distribution is because the Laplace distribution has an "ordinary smooth density" (as defined by Fan (1991)), unlike the Normal or Cauchy distributions that possess the supersmooth density, which results in an easy solution to the problem of estimating $G(x)$ with Gaussian Kernel without any restriction on the choice of parameter.

Lemma 2.3. *An estimate of $G(x)$, under assumption (A1) if $Y \sim$ Laplace $(0, \sigma^2)$, i.e.,*

$$f(x) = \frac{1}{2\sigma} e^{-\left|\frac{x}{\sigma}\right|} \ \forall x \in R,$$

is given by,

$$\hat{G}(x) = \frac{1}{n}\sum_{j=1}^{n}\left\{\left(1+\frac{\sigma^2}{b^2}\right)\Phi(x, Z_j, b) - \frac{\sigma^2}{b^2}\int_{-\infty}^{(x-Z_j)/b} u^2\Phi(u)du\right\} \tag{5}$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left\{\left(1+\frac{\sigma^2}{b^2}\right)\Phi(x, Z_j, b) - \frac{\sigma^2}{b^2}0.5\left(1+sign(x-Z_j)\mathcal{G}_{\left(\frac{3}{2},1\right)}\left(\frac{(x-Z_j)^2}{2b^2}\right)\right)\right\} \tag{6}$$

where $\mathcal{G}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{x})$ is the cumulative distribution function of Gamma distribution with parameters (α, β) at x.

Note: The density function of a Gamma distribution with parameters (α, β) is given below:

$$g_{(\alpha,\beta)}(y) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\lambda^{\alpha}} y^{\alpha-1} e^{-\lambda y}, y > 0 \\ 0, otherwise. \end{cases}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

2.5. Choice of Parameters of Error Distribution

It is to be noted that if the variance of the Error distribution is very small compared to the range of X, then the error behaves like a known constant which can be easily subtracted from Z_j to get a value very close to corresponding X_i. Hence a very small variance means no obfuscation at all. On the other hand, a very large variance may increase the error in estimation to a large extent. Hence, we need a perfect choice of the parameters of the Error Distribution to efficiently deal with the whole problem. Towards that, we make the following observation.

After obfuscating a particular value X_i we cannot get it back from $Z_i = X_i + Y_i$, but since we know the distribution of Y_i, we will get a confidence interval for each X_i. Assuming the mean of Y_i is zero, that is, Z_i and X_i has same mean, suppose for each X_i we want a minimum spread of ε with confidence $100(1 - \delta)\%$.

Proposition 2.4. *For fixed* $\delta > 0$ *and* $\varepsilon > 0$ *suppose we want a* $100(1-\delta)\%$ *Confidence Interval to be* ($Z_i - \varepsilon$, $Z_i + \varepsilon$) (ε moderately large), *then the parameter* σ *of the Error distribution can be taken as the solution of the equation*

$$F_\sigma(\varepsilon) = 1 - \frac{\delta}{2}$$

under the condition that $F_\sigma(\cdot)$ *is the cumulative distribution function of a random variable symmetric about* 0.

Proof. Since ($Z_i - \varepsilon$, $Z_i + \varepsilon$) is $100(1-\delta)$ % *Confidence Interval* for X_i,

$$P[X_i \in (Z_i - \varepsilon, Z_i + \varepsilon)] = 1 - \delta$$

$$\Rightarrow P(|Z_i - X_i| < \varepsilon) = 1 - \delta$$

$$\Rightarrow P(|Y_i| < \varepsilon) = 1 - \delta$$

Since $F(\cdot)$ *is symmetric around 0, we can write*

$$2F_\sigma(\varepsilon) - 1 = 1 - \delta$$

$$\Rightarrow 2F_\sigma(\varepsilon) = 2 - \delta$$

$$\Rightarrow F_\sigma(\varepsilon) = 1 - \frac{\delta}{2}.$$

Hence given ε and δ, we can find a value of σ from the equation

$$F_\sigma(\varepsilon) = 1 - \frac{\delta}{2}.$$

Special Cases

Laplace(0, $\boldsymbol{\sigma^2}$) The c.d.f. is given by,

$$F_\sigma(x) = 0.5 + 0.5sign(x)\left(1 - e^{-\frac{|x|}{\sigma}}\right)$$

Hence from Proposition 2.4 the solution of σ is

$$\sigma = -\frac{\varepsilon}{\log \delta}.$$

Uniform($\boldsymbol{-\frac{\sigma}{2}, \frac{\sigma}{2}}$) The c.d.f. is given by $F_\sigma(x) = \frac{x+\frac{\sigma}{2}}{\sigma}$. Hence from Proposition 2.4 the solution of σ is

$$\sigma = \frac{2\varepsilon}{1 - \delta}$$

Note. For Normal Error the process only works if the solution is less than the bandwith of Z. With 95% confidence, a choice of σ is approximately $\varepsilon/1.65$.

3. Some Simulation Results

In order to apply the above problem, we simulate a non-normal sample of size $n = 2{,}000$, with $IQR/1.34 \approx 1{,}000$, and then add an error Y_i to each sample unit X_i, such that $(Z_i - \varepsilon, Z_i + \varepsilon)$ is a 95% C.I. for X_i. The parameter for the error distribution is chosen by the formula in Proposition 2.4. For small ε, we apply Uniform, Normal and Laplace Errors to the sample, while for larger ε, Normal is not applicable. We therefore check results for Uniform and Laplace only. First, we check if the obfuscation is good enough. It is obvious that obfuscation improves as ε increases. In addition, for increasing ε, we also check how the estimation procedure works.

A sample of ten data points is taken from the data set and the corresponding obfuscated values are given for different errors. In the following Table 1, ε is assumed to be 200 (which is very small, since it is much smaller compared to the measure of dispersion of X).

Figure 1 shows the graph of the true distribution curve $\{G(x), x \in R\}$ along with the ones estimated from obfuscated data. Table 2 will show estimates of the true quantile values which is computed from the knowledge of $G(x)$ (Here, $G(x)$ is $Laplace(\mu = 10; \sigma = 1{,}000)$ using the function *qlaplace* under package {*rmutil*} of *R 3.3.2*. The quantile values are calculated from data $X_1, X_2; \ldots, X_n$ using function *quantile*. Also, estimated values of the quantiles are shown which we get by equating $\hat{G}(x)$ with (α: $0 < \alpha < 1$) by an iterative search method using the function *uniroot*; found in the package {*stats*} of *R 3.3.2*.

Note that the true and obfuscated values in Table 1 are quite close, which makes it easier for an intruder to guess the original value based on the obfuscated one. However, the estimation works quite well.

Now, we try increasing the value of ε. However, as the value increases, the Normal distribution is no longer an option; larger ε will make σ larger than the bandwidth of the corresponding Z.

The following Table 3 shows the true and obfuscated values of the same data points from Table 1 for increasing ε. Figure 2 will show how the estimated curve of $G(x)$ deteriorates with increasing ε. Table 4 gives the estimated and true quantiles for increasing ε.

Note that as ε increases the obfuscation improves but the estimation deteriorates. This is quite intuitive, since small ε implies no masking at all. As ε increases, both Uniform and Laplace gives result unlike Normal, but from the graph (Fig. 2), we can clearly see that for

*Table 1. Showing true and obfuscated values for ten data points selected from the 2,000 data points, **ε = 200**.*

No.	Data point	Uniform	Laplace	Normal
1	606.768	671.915	651.491	678.75
2	3139.892	3078.08	3166.548	3230.559
3	987.809	891.076	990.928	1023.493
4	2912.623	3120.068	2864.294	2714.819
5	- 1425.763	- 1369.556	- 1470.395	- 1518.552
6	- 185.086	- 305.841	- 68.098	- 205.403
7	- 940.958	- 1097.012	- 897.075	- 804.884
8	- 955.503	- 964.716	- 979.366	- 1005.702
9	- 224.565	- 46.007	- 228.214	- 304.326
10	- 511.614	- 470.031	- 469.044	- 597.995

Table 2. Estimated quantiles from obfuscated data, $\varepsilon = 200$.

α	TRUE	Original	Uniform	Laplace	Normal
"0.1"	- 1599.438	- 1476.929	- 1525.415	- 1534.134	- 1512.133
"0.2"	- 906.291	- 847.771	- 895.061	- 900.945	- 893.431
"0.3"	- 500.826	- 491.793	- 521.976	- 522.429	- 525.321
"0.4"	- 213.144	- 224.8	- 240.816	- 243.329	- 245.115
"0.5"	10	- 9.7	3.925	2.659	6.166
"0.6"	233.144	242.808	257.094	260.244	267.592
"0.7"	520.826	533.289	552.537	559.502	564.615
"0.8"	926.291	922.478	954.164	966.336	963.852
"0.9"	1619.438	1655.947	1687.02	1697.753	1698.098

larger quantiles the Uniform distribution gives very bad estimates, since the estimate of $G(x)$ at times even becomes decreasing, which is not at all desirable. However, Laplace seems to give comparatively better results compared to the Uniform. A theoretical explanation of the drawback of using Uniform Error is discussed in Section 5. Hence, we here prefer the use of Laplace Error over Uniform and Normal for reasonably large ε.

Hence, to investigate deeper into the statistical properties of such estimates, we note that the estimate is consistent, as is the estimate by Fan (1991). To evaluate other properties, such as the bias and mean square error in estimation, we find the Monte Carlo estimates of the bias and root-mean-squared-error (RMSE) over a simulation of S error samples (We take $S = 500$, 800 and 1,000). The Tables 5-8 present estimates of bias and RMSE for growing ε.

Compared to the dispersion of the data set (IQR $= 1{:}34 \approx 1{,}000$), the RMSE does not seem to be very large for $\varepsilon = 200$, 500 or 1,000. $\varepsilon = 2{,}000$ gives very large bias and RMSE but that large ε is rarely needed.

It can be easily observed that the bias and RMSE were consistent in the sense that 500, 800, and 1,000 simulations resulted in approximately similar values for all the cells in the above tables 5-8.

Observing the tables 5-8, we note that the main error in estimation comes from the bias of the estimate. Hence, an estimation of bias for the above problem can be a very interesting problem and a useful result for future research work.

But from this scenario, it is not clear whether the estimator is consistent, that is, with increasing n whether the bias decreases, although from Fan (1991) we can easily see that theoretically the estimate of $G(x)$ is consistent for all $x \in R$. So, to investigate, we simulate some other samples $X_1, X_2, \ldots, X_n$ using the same distribution as before, but larger n (we take $n = 5{,}000$, 10,000), and obfuscate using Laplace error similarly to find the Monte Carlo estimates of bias and RMSE, using $S = 1{,}000$.

One may easily observe from the tables (Table 7 and Table 8) that there is a decrease in the value of absolute bias and RMSE with larger n. Hence, with increasing n, ideally, the error tends to vanish.

4. A Real Life Example

For further illustration, we consider a real life application of the problem. We collect a data set of grades achieved by 445 students in the second year of the Masters of Statistics

Table 3. *True and obfuscated values for ten data points selected from the 2,000 data points with increasing ε.*

		$\varepsilon = 500$		$\varepsilon = 1{,}000$		$\varepsilon = 2{,}000$	
No.	Data point	Uniform	Laplace	Uniform	Laplace	Uniform	Laplace
1	606.768	777.005	697.307	1549.425	- 54.751	- 866.566	326.243
2	3139.892	3414.243	3134.548	4152.921	3718.174	3635.376	2679.141
3	987.809	1210.838	936.253	52.861	1216.174	3055.399	1140.865
4	2912.623	2760.988	2985.984	2376.626	2442.173	1178.522	3182.984
5	- 1425.763	- 1521.242	- 1451.637	- 1908.008	- 1720.502	- 530.379	- 1017.015
6	- 185.086	330.237	- 245.401	676.281	796.201	- 1985.132	- 163.254
7	- 940.958	- 420.662	- 960.868	- 1199.835	- 1051.968	- 1665.925	- 936.007
8	- 955.503	- 948.84	- 1040.34	- 1429.07	- 1083.252	- 1027.686	- 860.724
9	- 224.565	146.065	- 299.93	- 901.975	- 786.876	- 1592.798	- 1222.795
10	- 511.614	- 216.055	- 532.046	- 568.219	381.672	404.65	- 1145.256

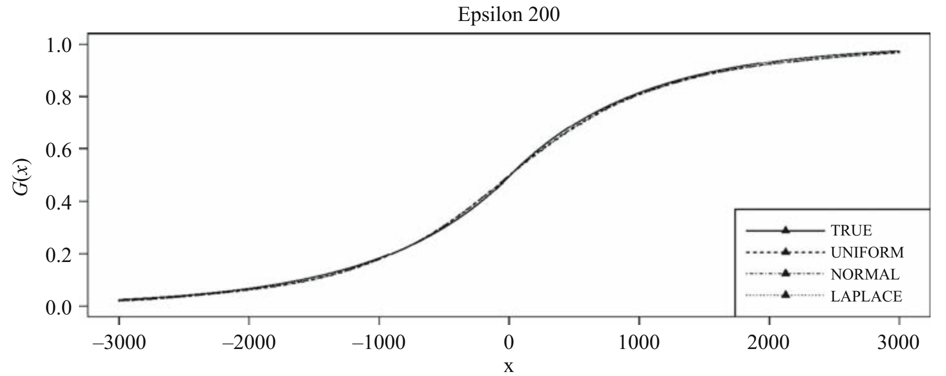

Fig. 1. True and estimated distribution curve with ε = 200.

program at the Indian Statistical Institute Kolkata over ten years, from 2006–2015. As grades are sensitive data, they cannot be released in raw form. We therefore apply the above problem to this data and try to find the results. Standard variation of the data was checked to be approximately 100, so we assumed an $\varepsilon = 200$. The bandwidth values from Uniform and Laplace data was found to be 48.68 and 41.15. The following Table 9 represents true and obfuscated values of ten data points to show how the values are masked with Uniform and Laplace Errors. From the obfuscated values, the true distribution and quantiles are then estimated as shown in Figure 3 and Table 10 respectively.

In this problem, σ was chosen according to Proposition 2.4 with $\varepsilon = 200$. Without access of the obfuscated data, all one knew about the marks of an individual was that it ranged between 0 to 1,000. Consider the first individual in Table 9. Its masked value after masking with $Laplace(0,\sigma^2)$ is 733.93. Now, we can say $X_i \in (533.93; 933.93)$ with 95% confidence. Hence, a disclosure takes place here. Note that, as per our knowledge, Z_i is the best estimator of X_i, based on the available information. However, if the intruder has an algorithm that can be used to find a better estimator of X_i using the knowledge of the obfuscating distribution and obfuscated data, this disclosure risk may not be valid (it can easily be shown that if true variance of Y is greater than $\frac{n}{n-1}$ times the true variance of X, then $\hat{Z}$ is a better estimator of X_i than Z_i; that is, the mean squared error of $\hat{Z}$ about X_i is less than that of Z_i about X_i but such a case is rare as σ usually does not need to be so large).

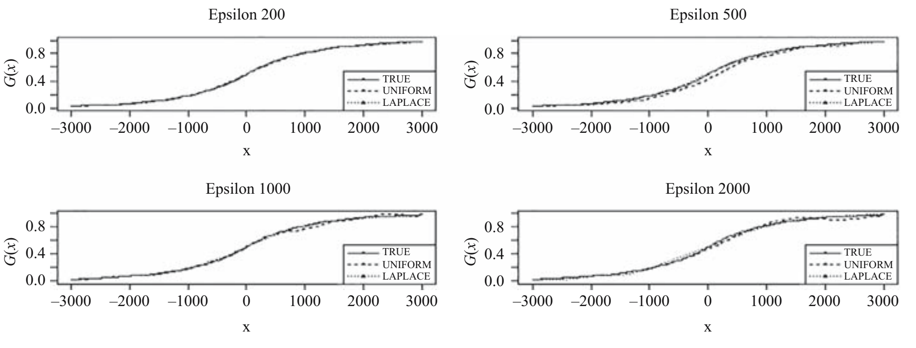

Fig. 2. True and estimated distribution curves with increasing ε.

Table 4. Estimated quantiles from obfuscated data with increasing ε.

			$\varepsilon = 500$		$\varepsilon = 1{,}000$		$\varepsilon = 2{,}000$	
α	TRUE	No error	Uniform	Laplace	Uniform	Laplace	Uniform	Laplace
“0.1”	- 1599.438	- 1476.929	- 1270.597	- 1540.032	- 1464.955	- 1695.023	- 1633.896	- 1892.266
“0.2”	- 906.291	- 847.771	- 742.253	- 925.49	- 880.811	- 1014.747	- 917.164	- 1090.942
“0.3”	- 500.826	- 491.793	- 360.159	- 554.813	- 487.193	- 615.94	- 496.031	- 730.971
“0.4”	- 213.144	- 224.8	- 55.626	- 264.023	- 221.973	- 266.593	- 179.127	- 337.419
“0.5”	10	- 9.7	178.56	- 12.464	- 5.545	- 1.624	91.225	6.49
“0.6”	233.144	242.808	389.331	257.945	213.649	296.902	339.212	349.937
“0.7”	520.826	533.289	644.638	580.443	540.851	610.137	590.126	717.073
“0.8”	926.291	922.478	1168.204	989.891	1179.53	1065.751	876.444	1155.097
“0.9”	1619.438	1655.947	1679.645	1745.236	1730.618	1827.97	1284.765	1902.892

*Table 5. Showing true values of quantiles of a data set and the corresponding **bias** of the estimate for **Laplace error** with increasing ε.*

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Org. data	$S = 500$	7.172	3.874	1.122	0.457	-0.235	-0.145	0.912	0.853	0.637
	$S = 800$	5.025	1.564	-0.239	0.137	0.111	0.187	0.217	0.258	-0.881
	$S = 1{,}000$	5.224	1.278	0.083	0.507	0.293	0.73	0.642	0.971	-0.164
$\varepsilon = 200$	$S = 500$	-23.802	-26.068	-27.564	-23.289	0.19	24.065	29.521	30.645	31.316
	$S = 800$	-26.253	-28.116	-28.747	-23.852	0.102	24.06	29.041	29.885	30.019
	$S = 1{,}000$	-26.096	-28.036	-28.432	-23.574	0.46	24.571	29.69	30.612	30.632
$\varepsilon = 500$	$S = 500$	-25.19	-27.786	-30.065	-24.923	-0.009	24.816	30.874	32.841	32.347
	$S = 800$	-27.962	-30.432	-30.921	-25.012	0.249	24.973	30.592	32.277	31.562
	$S = 1{,}000$	-28.165	-30.218	-30.566	-24.798	0.601	25.499	31.235	32.769	32.494
$\varepsilon = 1{,}000$	$S = 500$	-27.934	-35.081	-35.918	-28.76	0.431	30.062	38.386	40.433	40.181
	$S = 800$	-31.278	-36.546	-37.2	-29.646	-0.157	29.498	37.29	39.21	39.93
	$S = 1{,}000$	-32.331	-36.816	-37.039	-29.157	0.419	29.86	37.521	39.493	40.519
$\varepsilon = 2{,}000$	$S = 500$	-52.89	-54.773	-52.252	-39.383	-2.804	35.897	52.904	58.526	59.112
	$S = 800$	-56.447	-54.972	-53.402	-38.78	-1.55	36.939	52.953	58.96	57.323
	$S = 1{,}000$	-53.661	-56.074	-53.609	-38.725	-0.888	37.88	54.206	59.954	59.896

Table 6. *Showing true values of quantiles of a dataset and the corresponding* ***root mean square error*** *of the estimate for* ***Laplace error*** *with increasing ε.*

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Org. data	$S = 500$	68.292	44.792	33.593	27.865	23.739	28.55	34.165	44.095	64.891
	$S = 800$	67.933	44.249	33.581	27.893	23.425	28.543	34.946	45.653	65.1
	$S = 1{,}000$	67.585	43.817	33.424	27.402	23.145	28.412	34.5	45.605	65.465
$\varepsilon = 200$	$S = 500$	67.882	48.976	41.846	35.056	24.495	35.696	43.637	52.063	69.307
	$S = 800$	68.329	49.882	42.652	35.395	24.516	35.976	43.876	52.454	68.891
	$S = 1{,}000$	67.585	43.817	33.424	27.402	23.145	28.412	34.5	45.605	65.465
$\varepsilon = 500$	$S = 500$	69.828	50.778	44.136	36.893	25.412	36.825	45.225	53.997	71.167
	$S = 800$	71.076	52.533	44.756	36.962	25.6	37.374	45.691	54.721	71.555
	$S = 1{,}000$	71.256	52.176	44.357	36.542	25.266	37.605	46.008	54.802	72.471
$\varepsilon = 1{,}000$	$S = 500$	81.559	61.189	52.276	42.909	30.259	44.654	55.829	65.916	84.804
	$S = 800$	81.106	61.823	53.471	43.672	30.466	44.809	55.617	65.361	83.706
	$S = 1{,}000$	80.892	61.798	53.174	43.091	29.979	44.58	55.214	64.933	84.427
$\varepsilon = 2{,}000$	$S = 500$	124.104	93.579	77.468	62.005	45.251	60.263	78.433	94.74	128.227
	$S = 800$	126.806	93.281	78.115	62.302	46.944	61.471	78.53	95.605	127.758
	$S = 1{,}000$	125.164	93.145	77.559	62.169	46.741	61.918	79.275	96.219	128.16

*Table 7. Showing true values of quantiles of three data sets with sample size 2,000, 5,000, 10,000 and the corresponding estimated **bias** of the quantile estimate with $S = 1,000$ simulations for Laplace error with increasing ε.*

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Original	$n = 2,000$	5.224	1.278	0.083	0.507	0.293	0.73	0.642	0.971	-0.164
	$n = 5,000$	-0.277	-1.503	-1.164	-0.523	-0.183	-0.437	-1.164	-2.326	-0.914
	$n = 10,000$	-0.584	-0.231	-0.688	-0.671	-0.313	-0.635	-0.802	-0.698	-2.466
$\varepsilon = 200$	$n = 2,000$	-26.096	-28.036	-28.432	-23.574	0.46	24.571	29.69	30.612	30.632
	$n = 5,000$	-21.416	-21.935	-21.277	-18.278	-0.361	17.343	19.126	18.568	19.985
	$n = 10,000$	-16.184	-15.692	-15.939	-14.558	-0.467	13.341	14.505	14.66	12.957
$\varepsilon = 500$	$n = 2,000$	-28.165	-30.218	-30.566	-24.798	0.601	25.499	31.235	32.769	32.494
	$n = 5,000$	-23.077	-22.767	-22.223	-19.07	-0.291	18.139	20.025	19.7	21.558
	$n = 10,000$	-16.73	-16.634	-17.02	-15.612	-0.783	14.054	15.741	15.824	14.79
$\varepsilon = 1,000$	$n = 2,000$	-32.331	-36.816	-37.039	-29.157	0.419	29.86	37.521	39.493	40.519
	$n = 5,000$	-26.49	-26.753	-26.414	-22.493	-0.72	21.409	24.189	23.618	26.626
	$n = 10,000$	-21.879	-20.424	-19.784	-17.714	-0.925	16.116	18.12	18.263	18.515
$\varepsilon = 2,000$	$n = 2,000$	-53.661	-56.074	-53.609	-38.725	-0.888	37.88	54.206	59.954	59.896
	$n = 5,000$	-45.808	-40.622	-39.889	-29.43	0.823	30.261	39.033	37.867	38.36
	$n = 10,000$	-28.213	-30.479	-29.453	-24.909	-1.367	22.811	28.866	29.327	26.131

Table 8. Showing true values of quantiles of three data sets with sample size 2,000, 5,000, 10,000 and the corresponding estimated ***root mean square error*** *of the quantile estimate with $S = 1{,}000$ simulations for* ***Laplace error*** *with increasing ε.*

Alpha		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRUE		-1599.438	-906.291	-500.826	-213.144	10	233.144	520.826	926.291	1619.438
Original	$n = 2{,}000$	67.585	43.817	33.424	27.402	23.145	28.412	34.5	45.605	65.465
	$n = 5{,}000$	41.639	27.444	21.334	16.825	14.143	17.589	21.532	28.819	42.719
	$n = 10{,}000$	29.708	20.233	15.28	12.255	9.894	11.977	15.526	19.61	29.455
$\varepsilon = 200$	$n = 2{,}000$	68.099	49.741	42.235	34.963	24.266	36.193	44.147	52.711	69.264
	$n = 5{,}000$	44.931	34.362	29.217	24.229	14.738	23.881	28.026	32.84	44.692
	$n = 10{,}000$	32.599	24.863	21.596	18.638	10.382	17.568	20.496	23.802	30.9
$\varepsilon = 500$	$n = 2{,}000$	71.256	52.176	44.357	36.542	25.266	37.605	46.008	54.802	72.471
	$n = 5{,}000$	47.129	36.027	30.832	25.65	15.838	25.245	29.45	34.327	46.771
	$n = 10{,}000$	34.199	26.524	23.124	20.063	11.156	18.565	21.985	25.269	33.25
$\varepsilon = 1{,}000$	$n = 2{,}000$	80.892	61.798	53.174	43.091	29.979	44.58	55.214	64.933	84.427
	$n = 5{,}000$	57.132	43.227	36.534	30.642	18.875	30.116	36.236	41.937	56.043
	$n = 10{,}000$	43.107	32.83	27.969	24.14	14.94	22.751	26.719	31.432	41.413
$\varepsilon = 2{,}000$	$n = 2{,}000$	125.164	93.145	77.559	62.169	46.741	61.918	79.275	96.219	128.16
	$n = 5{,}000$	98.15	66.757	57.042	44.217	31.295	45.695	56.853	67.185	93.56
	$n = 10{,}000$	71.966	52.769	44.693	36.178	25.85	35.557	43.169	50.703	71.999

Table 9. True and obfuscated values for ten data points selected from the 445 data points, ε = 200.

No.	TRUE	Uniform	Laplace
"1"	814	960.562	733.931
"2"	750	695.214	829.526
"3"	764	656.395	591.158
"4"	574	704.041	599.055
"5"	614	670.67	586.944
"6"	669	595.926	670.136
"7"	616	553.873	533.097
"8"	674	748.607	677.74
"9"	714	595.295	658.648
"10"	740	883.885	764.591

In this case, Y_i is the error in estimation, and there is no risk of disclosure. However, there is a probability that the error is very small. Hence, the risk of disclosure with error less than d, is given by,

$$P[|Z_i - X_i| < d] = P[|Y_i| < d]$$

For $S = 1{,}000$ simulations, an estimate of this risk is

$$\frac{\sum_{s=1}^{S} I_{[Z_{si} \in (X_i - d, X_i + d)]}}{S}$$

where Z_{si} is the masked value of X_i for sth simulation and $I_{[A\ ]} = 1$, if event A occurs and zero otherwise. The following Table 11 shows estimates of disclosure risk for growing error values at ten selected points (the points in Table 9), and also a column giving the true risk value. We see the estimated risks are quite close to the theoretically determined risk at all the selected points.

5. Conclusion

Given the simulation results and also the real life example one can easily see that an increase in the value of ε, that is, an increase in obfuscation, results in weakly reliable

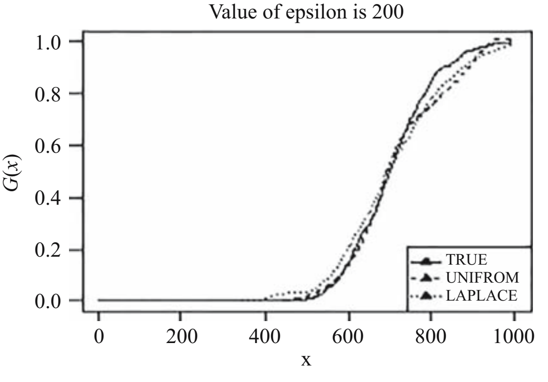

Fig. 3. Showing estimated distribution curve from TRUE and obfuscated data sets.

Table 10. Showing estimation of quantiles from original and obfuscated data.

No.	Original	Uniform	Laplace
"0.1"	580.8	578.394	555.663
"0.2"	612.8	622.305	596.067
"0.3"	645.2	650.741	633.059
"0.4"	675.6	673.521	664.011
"0.5"	700	695.237	693.693
"0.6"	727	720.346	734.52
"0.7"	750	762.636	770.46
"0.8"	786	831.202	809.933
"0.9"	826.6	888.999	879.513

estimates for both Laplace and Uniform Errors. However, we would prefer the use of Laplace over Uniform Error, since the Uniform has a serious drawback, as explained in the next paragraph.

In the case of Uniform Error, the estimate of $G(x)$ is given by the expression

$$\hat{G}(x) = \frac{a}{n}\sum_{j=1}^{n}\sum_{m=0}^{\infty}\phi\left(x, Z_j + \left(m+\frac{1}{2}\right)a, b\right)$$

which is nondecreasing if,

$$\hat{g}(x) = \frac{a}{n}\sum_{j=1}^{n}\sum_{m=0}^{\infty}\phi'\left(x, Z_j + \left(m+\frac{1}{2}\right)a, b\right) \geq 0,$$

that is if,

$$-\frac{c}{n}\sum_{j=1}^{n}\sum_{m=0}^{\infty}\left(x - Z_j - \left(m+\frac{1}{2}\right)a\right)e^{-\frac{\left(x-Z_j-\left(m+\frac{1}{2}\right)a\right)^2}{2b^2}} \geq 0$$

where c is a positive constant.

However, this term may become negative for certain cases. $\hat{G}(x)$ can therefore decrease at times, which is not at all desirable, since it is an estimate of a cumulative distribution function. In our simulations, we found that this problem arose several times, while in case of Laplace Error, this problem never arose. However, theoretically Equation (5), resulting from Laplace noise distribution, could not be proven to have a nondecreasing distribution function, either.

We have currently checked results for Uniform and Laplace distributions. However, the choice of an optimal density function for obfuscation and estimation has not yet been properly defined. It would be a challenging problem to define the optimal criterion and find a density that is capable of satisfying the criterion. The same challenge applies to finding an optimal ε (as defined in subsec. 2.5) for a given data set $(X_1, X_2, \ldots, X_n)$.

As discussed in Section 3, the error in estimation is mainly a result of the bias of the estimate. Hence, an estimation of bias and its correction should lead to a better resolution of the problem.

Table 11. Showing estimated risk of disclosure at ten selected points for increasing error value and theoretically determined risk value.

	X_i										
d	814	750	764	574	614	669	616	674	714	740	True value
10	0.154	0.114	0.141	0.127	0.14	0.125	0.141	0.145	0.143	0.163	0.139
20	0.282	0.215	0.25	0.272	0.264	0.254	0.261	0.268	0.274	0.283	0.259
30	0.393	0.33	0.343	0.368	0.369	0.344	0.369	0.385	0.386	0.38	0.362
40	0.47	0.418	0.453	0.462	0.457	0.431	0.456	0.475	0.493	0.475	0.451
50	0.542	0.491	0.533	0.539	0.522	0.494	0.528	0.549	0.553	0.559	0.527
60	0.617	0.575	0.602	0.588	0.58	0.573	0.579	0.604	0.619	0.608	0.593
70	0.666	0.636	0.659	0.649	0.631	0.637	0.627	0.651	0.677	0.649	0.65
80	0.709	0.689	0.707	0.696	0.68	0.691	0.678	0.703	0.722	0.685	0.698
90	0.742	0.724	0.752	0.736	0.723	0.728	0.729	0.745	0.754	0.734	0.74
100	0.775	0.754	0.785	0.771	0.764	0.756	0.771	0.779	0.796	0.779	0.776

Moreover, as mentioned in Section 4, if the boundary values of the original data are known, the obfuscation in the boundary region degrades. There is no known solution to this problem.

Having obtained a quantile estimate, computation of a confidence interval for the unknown population quantile could be an interesting problem for future work.

However, the problem discussed can easily be applied to many real life problems. The technique used to solve the above problem can be applied to solve the equations for other error distributions too. Unlike the historical technique to solve such problems, as given in Fan (1991), this technique can be applied to cases where the characteristic function of the Error distribution may take nonpositive value in some regions over the real line.

Appendix

Proof of Lemma 2.1

Proof. Putting the form of $f(y)$ in Equation (3), we have

$$H(z) = \frac{1}{a}\int_0^a G(z-y)dy$$

Now differentiating with respect to z we have

$$h(z) = \frac{1}{a}\{G(z) - G(z-a)\},$$

which gives

$$G(z) = ah(z) + G(z-a).$$

Now, from this relation we have

$$G(z-a) = ah(z-a) + G(z-2a).$$

Inserting this in the expression for $G(z)$ we find

$$G(z) = ah(z) + ah(z-a) + G(z-2a)$$

Repeating this by putting the values of $G(z - ma)$ for $m = 1, 2, \ldots$ in a similar way, we arrive at the given result.

Proof of Lemma 2.2

Proof. The lemma is proved using the following result from Polyanin and Manzhirov (Polyanin and Manzhirov 2008).

Result: *Consider the equation* $\int_{-\infty}^{\infty} K(x-t)y(t)dt = f(x), -\infty < x < \infty$ *where* $y(\cdot)$ *is the unknown function to be determined. Suppose,*

(i) $f(x), y(x) \in L_2(-\infty, \infty)$
(ii) $K(x) \in L_1(-\infty, \infty)$

where the function space $L_k(S)$ for some set S and integer k, is the set of all real-valued functions $\left\{f : S \to R, \int_{-\infty}^{\infty} |f(x)|^k dx < \infty\right\}$.

Then, $y(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\tilde{f}(u)}{\tilde{K}(u)} e^{iux} du$, *where $\tilde{f}$ is the Fourier Transform of f, $\tilde{K}$ is the Fourier Transform of K.*

Now to apply the given result in our problem note that our equation is

$$H(z) = \int_{-\infty}^{\infty} \mathrm{G}(y)\phi_\sigma(z-y)dy = \int_{-\infty}^{\infty} \mathrm{G}(z-y)\phi_\sigma(y)dy$$

But $H(\cdot)$ and $G(\cdot)$ are not $L_2(-\infty, \infty)$. So taking the derivative w.r.t. z, we get

$$h(z) = \frac{d}{dz}\int_{-\infty}^{\infty} \mathrm{G}(z-y)\phi_\sigma(y)dy.$$

Now, since g(·) is bounded, for some real 0 < M < ∞, we have,

$$\frac{d}{dz}(G(z-y)\phi_\sigma(y)) = g(z-y)\phi_\sigma(y) \leqslant M\phi_\sigma(y)$$

Now $\int_{-\infty}^{\infty} M\phi_\sigma(y)dy = M < \infty$. Hence we can interchange the integration and differentiation sign which gives us,

$$h(z) = \int_{-\infty}^{\infty} g(z-y)\phi_\sigma(y)dy$$

Here, we have used the Leibniz rule for infinite range.

Now, since $g(\cdot)$ and $h(\cdot)$ are bounded by assumption (A1), they are L_2 $-$ *bounded* by Lemma 2.3 of the book "Deconvolution Problems in Non-Parametric Statistics" by Meister (2009). Also, $\phi_\sigma \in L_1(-\infty, \infty)$.

Hence, applying the last result, in our problem,

$$g(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\tilde{h}(k)}{\tilde{\phi}_\sigma(k)} e^{ikx} dk$$

But h is not known. So, we replace it by $\hat{h}$, the Kernel Density Estimate of h using standard Gaussian Kernel and bandwidth selected by Silverman's "Rule of Thumb". The general form of such kind of estimators with an arbitrary kernel function $K(\cdot)$ was discussed by Fan (1991) where the kernel estimators of mixture densities were studied along with their asymptotic properties. It is given by

$$\hat{h}(x) = \frac{1}{nb}\sum_{j=1}^{n} K\left(\frac{x - Z_j}{b}\right) \tag{7}$$

where $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $b = 1.06\, n^{-\frac{1}{5}} A$ as defined in the statement of the Lemma. Plugging in, we get,

$$\tilde{\hat{h}}(k) = \int_{-\infty}^{\infty} \left\{ \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{x - Z_j}{b}\right) \right\} e^{-ikx} dx$$

$$= \frac{1}{n} \sum_{j=1}^{n} e^{-ikZ_j - \frac{k^2 b^2}{2}}$$

Since

$$\frac{1}{b} \int_{-\infty}^{\infty} e^{-ikx} K\left(\frac{x - Z_j}{b}\right) dx = \frac{1}{\sqrt{2\pi} b} \int_{-\infty}^{\infty} e^{-ikx} e^{-\frac{(x - Z_j)^2}{2b^2}} dx,$$

which is the characteristic function of a normal random variable with mean Z_j and standard deviation b at the point $(-k)$ and we know that to be equal to $e^{-ikZ_j - \frac{k^2 b^2}{2}}$.

Also, note that

$$\tilde{\phi}_\sigma(k) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} e^{-ikx} dx$$

$$= e^{\frac{-k^2\sigma^2}{2}}$$

Therefore, $\frac{\tilde{\hat{h}}(k)}{\tilde{\phi}_\sigma(k)} = \frac{\frac{1}{n}\sum_{j=1}^{n} e^{-ikx - \frac{k^2 b^2}{2}}}{e^{-\frac{k^2\sigma^2}{2}}} = \frac{1}{n}\sum_{j=1}^{n} e^{-ikx - \frac{k^2(b^2 - \sigma^2)}{2}} \in L_2(-\infty, \infty)$ if $b^2 - \sigma^2 > 0$, that is, $b > \sigma$

If $b > \sigma$, then,

$$\hat{g}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{n} \sum_{j=1}^{n} e^{-ikZ_j - \frac{k^2(b^2 - \sigma^2)}{2}} e^{ixk} dk$$

$$= \frac{1}{\sqrt{2\pi} n \sqrt{\mathrm{b}^2 - \sigma^2}} \sum_{j=1}^{n} e^{-\frac{(x - Z_j)^2}{2(\mathrm{b}^2 - \sigma^2)}}.$$

where we have changed the order of summation and integration. This is nothing but the mean of n normal p.d.f.s with mean Z_j and variance $b^2 - \sigma^2$. Hence, we get the form given in Lemma 2.2.

Proof of Lemma 2.3

Proof: Proceeding in the same way as in Lemma 2.2, we have

$$\tilde{\hat{h}}(k) = \frac{1}{n} \sum_{j=1}^{n} e^{-ikZ_j - \frac{k^2 b^2}{2}}$$

and the Fourier transform at point k of the Laplacian error density with scale parameter σ, denoted as $\tilde{\ell}_\sigma(k)$, is given by,

$$\tilde{\ell}_\sigma(k) = \int_{-\infty}^{\infty} \frac{1}{2\sigma} e^{-|x|/\sigma} e^{-ikx} dx = (1 + \sigma^2 k^2)^{-1}$$

Hence, the ratio becomes

$$\frac{\tilde{\hat{h}}(k)}{\tilde{\ell}_\sigma(k)} = \frac{1}{n} \sum_{j=1}^{n} (1 + \sigma^2 k^2) e^{-ikZ_j - \frac{k^2 b^2}{2}}$$

This function is now in $L_2(-\infty, \infty)$ $\forall b, \sigma$. After taking the inverse Fourier transform, we have

$$\hat{g}(x) = \frac{1}{n} \sum_{j=1}^{n} I_j,$$

where

$$\begin{aligned} I_j &= \frac{1}{2\pi} \int_{-\infty}^{\infty} (1 + \mathrm{k}^2 \sigma^2) e^{ik(x - Z_j) - \frac{k^2 b^2}{2}} dk \\ &= I_{1j} + I_{2j}, \end{aligned}$$

$$I_{1j} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ik(x - Z_j) - \frac{k^2 b^2}{2}} dk, \text{ and,}$$

$$I_{2j} = \frac{1}{2\pi} \int_{-\infty}^{\infty} k^2 \sigma^2 e^{ik(x - Z_j) - \frac{k^2 b^2}{2}} dk.$$

Note that the integrand in I_{1j} is nothing but a constant multiple of the characteristic function of $N(0, 1/b)$ at $(x - Z_j)$ and hence, it can easily be shown that

$$I_{1j} = \phi(x, Z_j, b)$$

It should now be noted that

$$I_{2j} = \frac{\sigma^2}{2\pi} \int_{-\infty}^{\infty} k^2 e^{\frac{-k^2 b^2}{2}} \{\cos(k(x - Z_j)) + i\sin(k(x - Z_j))\} dk$$

Since the sine function is odd and the cosine function is even, we can write

$$I_{2j} = \frac{\sigma^2}{\pi} \int_{0}^{\infty} \cos(k(x - Z_j)) k^2 e^{\frac{-k^2 b^2}{2}} dk$$

Defining $c_j = \frac{\sqrt{2}}{b}(x - Z_j)$ and making a change of variables, we get the expression

$$I_{2j} = \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\int_0^\infty \cos\left(c_j\sqrt{y}\right)\sqrt{y}e^{-y}dy$$

Next, expanding $\cos\left(\mathrm{c_j}\sqrt{\mathrm{y}}\right)$ by a Taylor series and changing the order of summation and integration, we have

$$I_{2j} = \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2\mathrm{m}}}{(2m)!}\Gamma\left(m+\frac{3}{2}\right)$$

where $\Gamma(x)$ denotes the Gamma function evaluated at the point x. Using the properties of the Gamma function that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ we can further calculate

$$I_{2j} = \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{(2m)!}\left(m+\frac{1}{2}\right)\left(m-\frac{1}{2}\right)\ldots\frac{1}{2}\Gamma\left(\frac{1}{2}\right)$$

$$= \frac{\sigma^2}{\pi}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{(2m)!}\left(m+\frac{1}{2}\right)\left(m-\frac{1}{2}\right)\ldots\frac{1}{2}\Gamma\left(\frac{1}{2}\right)$$

$$= \frac{\sigma^2}{\sqrt{\pi}}\frac{\sqrt{2}}{b^3}\sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{2^{2m}m!}\frac{2m+1}{2}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}b^3}\left\{2\sum_{m=1}^{\infty}(-1)^m\frac{c_j^{2m}}{2^{2m}(m-1)!} + \sum_{m=0}^{\infty}(-1)^m\frac{c_j^{2m}}{2^{2m}(m)!}\right\}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}b^3}\left\{2(-1)\left(\frac{c_j}{2}\right)^2 e^{-\left(\frac{c_j}{2}\right)^2} + e^{-\left(\frac{c_j}{2}\right)^2}\right\}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}b^3}e^{-(c_j/2)^2}[1 - 2(c_j/2)^2]$$

$$= \frac{\sigma^2}{b^2}\left[1 - \left(\frac{x - Z_j}{b}\right)^2\right]\phi(x, Z_j, b)$$

where we inserted the expression $c_j = \frac{\sqrt{2}}{b}(x - Z_j)$ in the last step. Thus, we can conclude that

$$\hat{g}(x) = \left(1 + \frac{\sigma^2}{b^2}\right)\left\{\frac{1}{n}\sum_{i=1}^{n}\phi(x, Z_j, b)\right\} - \frac{\sigma^2}{b^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x - Z_j}{b}\right)^2\phi(x, Z_j, b)$$

Hence, integrating $\hat{g}(u)$ over $(-\infty, x)$ we get Equation (5). Moreover, making a simple change of variable $\frac{u^2}{2} = y$ in the term

$$\int_{-\infty}^{\frac{x-Z_j}{b}} u^2 \phi(u) du$$

one can easily check whether it is equal to

$$0.5 + 0.5^* sign(x - Z_j)\mathcal{G}_{(3/2,1)}\left(\frac{x - Z_j}{b}\right)^2$$

as stated in Equation (6).

6. References

Fan, J. 1991. "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems." *The Annals of Statistics* 19(3): 1257–1272. Available at: http://www.jstor.org/stable/2241949 (accessed December 2017).

Fuller, W.A. 1993. "Masking Procedures for Microdata Disclosure Limitation." *Journal of Official Statistics* 9(3): 383–406. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/masking-procedures-for-microdata-disclosure-limitation.pdf (accessed December 2017).

Kim, H.J. and A.F. Karr. 2013. *The Effect of Statistical Disclosure Limitation on Parameter Estimation for a Finite Population*. NISS, October.

Meister, A. 2009. *Deconvolution Problems in Nonparametric Statistics*. Berlin Heidelberg: Springer Verlag.

Mukherjee, S. and G.T. Duncan. 1997. *Disclosure Limitation through Additive Noise Data Masking: Analysis of Skewed Sensitive Data. Disclosure Limitation through Additive Noise Data Masking: Analysis of Skewed Sensitive Data*. IEEE.

Polyanin, A.D. and A.V. Manzhirov. 2008. *Handbook of Integral Equations*. Chapman and Hall/CRC.

Poole, W.K. 1974. "Estimation of the Distribution Function of a Continuous Type Random Variable Through Randomized Response." *Journal of the American Statistical Association* 69(348): 1002–1005.

Sinha, B., T.K. Nayak, and L. Zayatz. 2011. "Privacy Protection and Quantile Estimation from Noise Multiplied Data." *Sankhya B* 73: 297–315. Doi: https://doi.org/10.1007/s13571-011-0030-z.

Zayatz, L., K.T. Nayak, and B.K. Sinha. 2011. "Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection." *Journal of Official Statistics* 27(2): 527–544. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-properties-of-multiplicative-noise-masking-for-confidentiality-protection.pdf (accessed December 2017).

Received March 2016
Revised May 2017
Accepted October 2017