Application of the Truncated Distributions and Copulas in Masking Data

Rahul A. Parsa[1], Jay J. Kim[2] and Myron Katzoff[2]

[1] College of Business and Public Administration, Drake University, Des Moines, IA 50321

[2] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD. 20782

Abstract[3]: In masking microdata, two approaches – adding independent noise and multiplying by independent noise have been used. The truncated distribution has been used for masking microdata. The random variable which follows the truncated distribution serves as the noise factor. For multiplicative noise method, the natural candidate distribution is the one which is centered at 1 and for additive noise method, it would be one centered at 0. Kim (2007) investigated triangular distribution truncated around 1 as noise distribution for multiplicative noise. In this paper we generalize his idea using copulas and correlated noise. We show that by using correlated noise, we can protect the moments, that is, the moments of the perturbed variable will have the same values as the original variable. We present two examples, one using correlated noise from a triangular distribution and second, from a truncated uniform distribution.

Key words: truncated triangular distribution, truncated uniform distribution, masking, confidentiality

1. Introduction

Researchers have been investigating methods for releasing data for public use while protecting the confidentiality for a long time. Additive noise scheme (adding noise to the original data) has been investigated by many researchers [e.g., Spruill (1983), Kim (1986), Kim and Winkler (1995), Muralidhar, et al. (1999)] to protect the confidentiality of the records. Most common approach is to add noise to the original data. However, this approach results in the variance which is higher than the original data. To make the variance of the masked data the same as that of the original data, Kim (1986) suggested moving the masked data points toward the mean of the masked data using the linear transformation.

Instead of adding noise to the original data, multiplying the original data by noise (multiplicative noise) has also been investigated. Evans, et al [1998] proposed the use of multiplicative noise to mask economic data. They considered noise which follows distributions such as the normal and truncated normal distributions. Kim and Winkler (2001) considered a noise distribution following the truncated normal distribution. The

[3] **Disclaimer**: The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Bureau of the Census uses a truncated triangular distribution for masking the Commodity Flow Survey data. Kim (2007) developed the probability density function of the truncated triangular distribution and showed that the estimate from the data masked by the distribution is unbiased, if the triangular distribution is symmetric about 1 and truncated symmetrically also about 1.

In this paper we extend Kim's method of using truncated distributions for masking the data. We generate correlated noise from the appropriate truncated distribution using Copulas and then use the resulting noise to mask the data. This method guarantees that every observation has some minimum noise added to it and the researcher gets to control the minimum and maximum noise. We show that this method approximately protects first two moments in additive case but only the first moment in the multiplicative case.

We briefly summarize below the two methods.

Additive Noise Method The additive-noise methodology (Kim, 1986, Muralidhar, 1999) for masking multivariate normal data that preserves confidentiality and preserves means can be described as follows. Let *X* be the original variable and Y the masked variable. Then,

$$Y = X + e,$$

where *e* is noise chosen independent of *X*. The distribution of e is usually chosen to be that of *X*. Since, one would like to have *E(Y) = E(X*), it implies that *E(e) = 0*.

Also,

$$Var\ (Y) = Var\ (X) + Var(e).$$

It means that $Var\ (Y) \geq Var(X)$.

Multiplicative Noise Method In this case, *Y* is defined as follows.

$$Y = Xe,$$

where e once again is chosen to be independent of *X*. Since we want, *E(Y) = E(X)*, it implies that *E(e)* = 1 and

$$Var(y) = Var(x)Var(e) + \mu_e^2 Var(x) + \mu_x^2 Var(e)$$

Once again, $Var(Y) \geq Var(X)$.

Note that in both methods, there is a possibility that some of the observations could be unchanged leading to full disclosure. In the additive noise method, this happens when the values of e are equal to zero and in the multiplicative noise method, when the values of e are equal to1. To avoid these values, truncated distributions have been used for masking data (Kim, 2007). So, values of e close to zero are truncated from the distribution of e for the additive noise method and values close to 1 are truncated from the distribution of e in the multiplicative noise method. The variance of the masked variable will still be higher than the variance of original variable, that is, $Var(Y) \geq Var(X)$.

If one assumes that noise is correlated with the original variable (instead of being statistically independent of it), equality of variances can be achieved with the additive noise method. That is, as we will show below, it is possible to obtain both the equality of means and variances if you assume correlated noise in certain situations. Perhaps it is most important that, when the noise is correlated with the original variable, the noise that gets added (or multiplied) depends on the value of the original variable, which seems to be a very desirable property.

2. Masking Using Correlated Noise

Consider the additive noise method. Let *X* be the original variable and *Y* the masked variable. The correlated additive noise method implies that $Y = X + e$, and *X* and e are correlated. We want the first two moments of *X* and *Y* to be identical. That is, *E(Y) = E(X))* and *Var(Y) = Var (X)*. *E(Y) = E(X))* implies that

$$E(e) = 0. \tag{1}$$

Also

$$Var\ (Y) = Var(X) + Var(e) + 2\ Cov(X,\mathrm{e}).$$

Since we want *Var(Y) = Var (X),*

$$Var(Y) = Var(X) = Var(X) + Var\ (e) + 2{*}Cov(X,e)$$

or

$$Cov(X,e) = -\ Var(e)\ /\ 2. \tag{2}$$

This implies that to mask X and protect first two moments, we need to generate *e* so that *e* has mean zero and correlated with X with covariance given by equation (2). Since correlation between any two variables is bound by 1, applying this to equation (2), we get

$$|\,Corr(X,e)\,| = \left|\frac{Cov(X,e)}{\sigma_x\,\sigma_e}\right| \leq 1$$

Substituting for *Cov(X, e)* in the above equation, we get

$$\sigma_e \leq 2\sigma_x$$

So, there is a bound on the variance of e, it cannot exceed four times the variance of *X*.

Consider the multiplicative noise method. Here, $Y = Xe$ and once again, *X* is correlated with *e*. Since we want the first moments to be the same, it implies that

$$E(e) = 0 \qquad (3)$$

and we want this to be equal to *E(X)*.

Note that

$$Cov(X,e) = E(Xe) – E(X)*E(e)$$

or

$$E(Xe) = Cov(X,e) + E(X)*E(e), \qquad (4)$$

Substituting this in equation (3) for *E(Xe)* and also substituting *E(X)* for *E(Y)*, we get

$$E(X) = Cov(X,e) + E(X)*E(e)$$

Solving for *Cov(X,e)*, we get

$$Cov(X,e) = E(X) – E(X)*E(e) = E(X)* (1 – E(e)) \qquad (5)$$

This implies that the *E(e)* cannot be equal to 1 if we want correlated noise. It is not easy to achieve equality of variances in this case.

We will use copulas to generate correlated noise from a specified distribution for masking the data. We will first briefly describe copulas and in particular the Gaussian copula and then the algorithm for generating correlated noise.

3. Copula

Copulas are useful for describing multivariate non-normal distributions. They describe the dependence structure between the variables. Marginal distribution functions are used as inputs to the copula and these can be any set of disparate distributions. Thus, a copula is very realistic way of describing the multivariate distributions as one normally has a good idea on the distribution of the marginals and seldom on the joint distribution of these variables. The concept of a copula is to divide the multivariate distribution into two parts: (1) one that describes the dependence structure, and (2) one that describes the marginal distributions. This concept of defining the multivariate dependence structure (and hence the copula) is based on Sklar theorem which states:

Sklar's Theorem (1959) Let *F* be a n-dimensional joint cumulative function for random variables $X_1, X_2, \ldots\ldots X_n$ with marginal distribution functions $F_1(X_1), \ldots\ldots.,F_n(X_n)$, then there exists a Copula such that

$$F(x_1,x_2,\ldots\ldots.x_n) = C[F_1(x_1), \ldots\ldots.,F_n(x_n)].$$

where *C(..)* is a copula. If the marginal distributions, $F_{1,}\ F_2, \ldots F_{n,}$ are continuous, then *C* is unique. If they are discrete, then *C* is uniquely defined on the $Ran(F_1)\times Ran(F_2)\times\ldots..Ran(F_n)$, where $Ran(F_i)$ is the range of F_i.

Sklar's theorem states that a copula of the random variables $X_1, X_2, \ldots\ldots X_n$ is the joint distribution function C() of the marginal cdf's. It produces a new multivariate distribution based on what distributions are used to describe marginal distributions. The dependence structure is defined by the copula. So, in the case of continuous distributions, the multivariate dependence structure and univariate marginal distributions can be separated and the copula can be considered 'independent' of the univariate margins (Joe 1997, page 12-13). The copula thus allows one to combine arbitrary continuous marginal distributions and describe their dependence structure by forming a multivariate non-normal distribution.

There are several bi-variate and multivariate copula distributions discussed in the literature. Hutchinson and Lai (1990), Nelson (1999) and Joe (1997) are excellent texts that discuss various properties of bi-variate copula distributions. Paul Embrechts, et. al (2002) uses copulas to show how Pearson correlation coefficients can be misleading. Klugman, S. A. and Parsa, R. (1999), Frees, E.W., and Valdez, E. (1998) use copulas in modeling insurance data.

Multivariate Normal or Gaussian Copula

We will use a Gaussian copula for generating the correlated noise as it is fairly easy to implement. We describe below its distribution and density.

Let G be a k-dimensional distribution function with margins $G_1, \ldots..G_k$, then the copula is of the form

$$C_G(u_{1,} \quad . \quad . \quad . \quad .u_k) = G\left\{G_1^{-1}(u_1), \ldots\ldots\ldots, G_k^{-1}(u_k)\right\} \qquad (6)$$

assuming that G^{-1} exists, and $u = (u_1, \ldots..u_k)$ is the uniform vector. A special case of this distribution is multivariate normal copula which is of interest in this paper. It is based on k-variate normal distribution $N(0, \rho)$, with unit variances (ρ_{ii} = 1 for all i) and obtained by substituting $G_i = \Phi_i$. Its distribution function is thus given by (Song 2000)

$$C_\Phi(u_1,. \quad . \quad . \quad . \quad u_k) = \Phi\{\Phi_1^{-1}(u_1), \ldots\ldots\ldots.\Phi_k^{-1}(u_k)\} \qquad (7)$$

and its density function is given by

$$f(u_1, u_2, \ldots\ldots.u_k/\rho) = c_\Phi(u_1, \ldots\ldots., u_k / \rho) = \frac{1}{|\rho|^{1/2}} \exp\{\frac{-y^t(\rho^{-1} - I)y}{2}\} \qquad (8)$$

where y is a vector with elements $y_{i,}$ and $y_i = \Phi^{-1}(u_i)$, $\rho = [\rho_{ij}]$, $\rho_{ij} = corr[\Phi^{-1}(u_i), \Phi^{-1}(u_j)]$, Φ is the usual distribution function of a normal, $N(0, \rho)$. Here, ρ determines the level of dependence.

If the marginal distributions are continuous then the density function of a vector ***x*** = $(x_1,x_2,\ldots\ldots.x_k)$ with arbitrary marginals, $F_i(x_i)$, is obtained by substituting $u_i = F_i(x_i)$ in equation (8) and is given by (Clemen and Riley 1999, Song 2000)

$$\begin{aligned} f(x_1,\ldots\ldots..,x_k) &= c_\Phi\{F_1(x_1,\ldots\ldots...,F_k(x_k)/\rho\}*f_1(x_1)\times f_2(x_2)\times\ldots\times f_k(x_k) \\ &= f_1(x_1)\times f_2(x_2)\times\ldots\times f_k(x_k)\times\exp\left\{\frac{-y^t(R^{-1}-I)y}{2}\right\}\times|R|^{-0..5} \qquad (9)\end{aligned}$$

where $y_i = \Phi^{-1}[F_i\{x_i)]$ and $R_{ij} = corr[\Phi^{-1}(F_i(x_i)), \Phi^{-1}(F_j(x_j))]$.

Generating random numbers from a multivariate normal copula is fairly straight forward. Let R^* denote the matrix of relationships among the variables measured with Kendall's τ or Spearman's rank correlation, ρ. Note that Pearson correlation no longer adequately measures dependence between non-normal random variables. Then, for each element of R *, calculate the corresponding product-moment correlation r_{ij} of R using the relationship $r_{ij} = sin(\pi\ \tau_{ij}/2)$ or $r_{ij} = 2*sin(\pi\ \rho_{ij}/6)$, where r_{ij} is the Pearson correlation coefficient. Then, generate random k-vectors $z \sim N(0, R)$ and secondly, obtain random variables $x_1,x_2,\ldots\ldots.x_k$ by $x_j = F_j^{-1}(\Phi(z_j))$. The resulting $x_1,x_2,\ldots\ldots.x_k$ will have the dependencies given by R^*.

Algorithm for generating correlated noise:

In our case, we have to generate only one variable that is correlated with the given variable. Let X denote the original variable with distribution function F_x and let r be the desired product-moment correlation between X and e.

1. $V_1 = \Phi^{-1}[F_x(x)]$ ~ Normal (0,1) distribution
2. Generate V_2 that is correlated with V_1 with specified correlation.
3. e = $F_e^{-1}[\Phi(V_2)]$
4. Resulting e will have dependence with X given by r^* corresponding to r.

For the purpose of exposition of our methodology, we chose the truncated triangular distribution and the truncated uniform distribution for e.

4. Truncated Triangular Distribution

When multiplicative noise is used for masking data, one must avoid using a number close to one (1) for noise because multiplying by a number very close to 1 does not change the original value much, and, thus, the original value may not get sufficient protection. We also note that when use is made of the triangular distribution symmetric about the value 1, the probability density for noise (e) is the greatest when e is near 1 if that is the mode. Thus, in this latter case, one would expect that a significant proportion of the values would not get a lot of protection. For reasons like these, we surmise that Evans, et. al.

(1998) have suggested truncating the mid-section, or the section near 1, of the symmetric triangular distribution with mode 1. For similar reasons, in case of the additive noise method, we would suggest avoiding the use of numbers close to zero (0) for noise.

The truncated triangular distribution has the following shape.

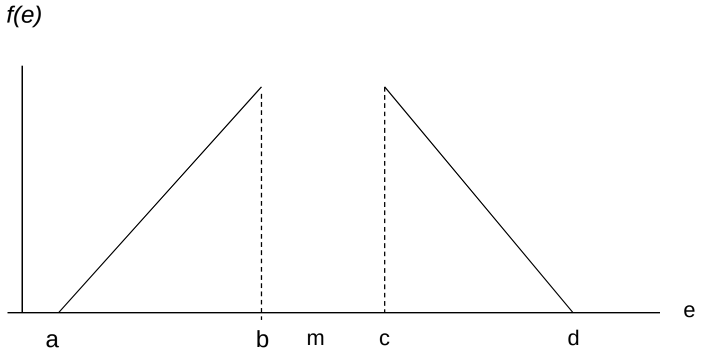

Figure 1. Truncated Triangular Distribution

Suppose the distribution is truncated at b and c, $c > b$, as shown in Figure 1. In this case, the pdf has the following form:

$$f(e) = \begin{cases} \dfrac{2(d-m)}{(b-a)^2(d-m)+(d-c)^2(m-a)}\,(e-a), & a \le e < b \\ \dfrac{2(m-a)}{(b-a)^2(d-m)+(d-c)^2(m-a)}\,(d-e), & c \le e < d\,. \end{cases} \tag{10}$$

In the above equation, let

$$k = \frac{2}{(b-a)^2(d-m)+(d-c)^2(m-a)}\ .$$

Then equation (10) can be re-expressed as

$$f(x) = \begin{cases} k\,(d-m)(x-a), & a \le x < b \\ k\ (m-a)(d-x), & c \le x < d\,. \end{cases} \tag{11}$$

The cumulative distribution function of the truncated triangular distribution after some algebra is

$$F(e)=\begin{cases}\frac{k}{2}(d-m)\left(e-a\right)^2, & a\le e<b\\ \frac{k}{2}(d-m)\left(c-a\right)^2, & b\le e<c\\ \frac{k}{2}(d-m)\left(c-a\right)^2+\frac{k}{2}(m-a)\left[\left(d-c\right)^2-\left(d-e\right)^2\right], & c\le e<d\end{cases} \tag{12}$$

We will assume that the triangular distribution is symmetric about m and the truncation is also symmetric about m. This is a reasonable assumption for a noise distribution as we want noise to be symmetrically distributed. Then,

$$f(e)=\begin{cases}\frac{(e-a)}{(b-a)^2}, & a\le e<b\\ \frac{(d-e)}{(b-a)^2}, & c\le e<d\end{cases} \tag{13}$$

and

$$F(e)=\begin{cases}\frac{(e-a)^2}{2(b-a)^2}, & a\le e<b\\ 0.5, & b\le e<c\\ 0.5+\frac{1}{2(b-a)^2}(2de-e^2-2dc+c^2), & c\le e<d\end{cases} \tag{14}$$

The mean and variance of *e* are

$$E(e)=\frac{a+2b+2c+d}{6} \tag{15}$$

and

$$.Var(e)=\frac{5b^2+2ab+2a^2+5c^2+2cd+2d^2-4ac-2ad-2bc-4bd}{36} \tag{16}$$

5. Truncated Uniform Distribution

Once again, we will consider a symmetric distribution. The density is given by

$$f(e)=\begin{cases}\dfrac{0.5}{(b-a)}\,, & -b\le e<-a\\[2ex] \dfrac{0.5}{(b-a)}\,, & a\le e<b\end{cases} \tag{17}$$

and the cumulative distribution is given by

$$F(e)=\begin{cases}0 & e<-b\\[1ex] \dfrac{0.5(e+b)}{(b-a)}\,, & -b\le e<-a\\[1ex] 0.5, & -a\le e<a\\[1ex] 0.5+\dfrac{0.5(e-a)}{(b-a)}\,, & a\le e<b\\[1ex] 1 & e\ge b\end{cases} \tag{18}$$

The mean and variance are given by

$$\text{E(e) = 0 and } Var(e)=\frac{a^2+ab+b^2}{3}\,. \tag{19}$$

Examples:

For *X*, we generated 2000 observations from a gamma distribution with mean 8 and variance 32. We applied both additive and multiplicative noise methods to mask the 2000 observations. Also, we used both truncated triangular distribution and truncated uniform distribution for correlated noise e. We present below one set of the results.

	Additive Noise Method					
Distribution	*a*	*b*	*c*	*d*	Mean Of *Y*	Variance of *Y*
Truncated Triangular	-6.6	-1	1	6.6	8.027588	34.26451
Truncated Uniform	-10	-5	5	10	8.054353	41.47306
	Multiplicative Noise Method					
Truncated Triangular	-6.6	-1	1	6.6	8.000259	1119.277
Truncated Uniform	-2	-5.8	2	5.8	8.041039	1662.549

In the additive noise method, we were able to control both the mean and variance. We used several combinations of values for *a, b, c* and *d* and we always got similar results.

In the multiplicative noise method we were able to control the mean as expected. Unfortunately, we failed to obtain these results for all possible values of *a*, *b*, *c* and *d*. At this point it is not clear to us why it is not working. We do realize the covariance and correlation are not good measures of association for non-normal variables but we are not completely satisfied with this reasoning. We believe this problem needs further investigation.

6. **Conclusions**

If a public use microdata file is masked by additive or multiplicative noise, it is desirable to have first two moments protected. We have shown that the masking using correlated noise protects first two moments in additive noise method but only the first moment in the multiplicative noise method. We used copulas to obtain the correlated noise. We used a Gaussian copula to generate the correlated noise and we showed how to use truncated triangular distribution and truncated uniform distribution to model the correlated noise.

7. **References**

Clemen, R.T. Reilly, T (1999) Correlations and Copulas for Decision and Risk Analysis, Management Science. 45, 208-224.

Embrechts, P. Alexander McNeil, and Daniel Straumann, 2002, Correlation and dependence in Risk Management: Properties and Pitfalls, *Risk Management: Value at Risk and Beyond*, M. Dempster (Ed.), Cambridge University Press, 176-223.

Evans, T., Zayatz, L., and Slanta, J. (1998) Using Noise for Disclosure Limitation of Establishment Tabular Data, Journal of Official Statistics, 14, No. 4, pp537-551.

Frees, E.W., and Valdez, E. 1998, *Understanding Relationships Using Copulas*, North American Actuarial Journal, 2, 1-25.

Joe, J. (1997) Multivariate Models and Dependence Concepts. Chapman & Hall, London, U.K.

Hutchinson, T.P., and C.D. Lai, 1990, *Continuous Bivariate Distributions, Emphasizing Applications*, Rumsby Scientific Publishing, Adelaide.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, Proceedings of the Section on Survey Research Methods, 303-308.

Kim, J. J. and Winkler, W. E. (1995), “Masking Microdata Files,” American Statistical Association, Proceedings of the Section on Survey Research Methods, 114-119.

Kim, J..J. and Winkler, W. E. (2001) Multiplicative Noise for Masking Continuous Data, Proceedings of the Survey Methods Research Section, American Statistical Association, CD Rom.

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

Klugman, S. and Parsa, R.A., 1999, Fitting Bivariate Loss distributions with Copulas, *Insurance: Mathematics and Economics*, 24, 139-148.

Muralidhar, K., Parsa, R. and Sarathy, R. (1999) A General Additive Data Perturbation Method for Database Security, Management Science, 45, No. 10, 1399-1415.

Nelson, R.B., 1999, *An Introduction to Copulas, Springer*, New York.

Sklar, A. (1959) Fonctions de répartition à *n* dimensions et leurs marges. Publ Inst Statist Univ Paris 8 :229-231.

Song, P. X.-K., 2000, Multivariate Dispersion Models Generated from Gaussian Copula, Scandinavian Journal of Statistics, 27, 305-320.

Spruill, N.L. (1983) Confidentiality and Analytic Usefulness of Masked Business Microdata, the Public Research Institute, Alexandria, Va.