9-2011

Multiplicative Noise for Masking Numerical Microdata Data with Constraints

Anna Oganian
Georgia Southern University, aoganyan@georgiasouthern.edu

SORT Special issue: Privacy in statistical databases, 2011, 99-112

Multiplicative noise for masking numerical microdata with constraints

Anna Oganian*

Georgia Southern University

Abstract

Before releasing databases which contain sensitive information about individuals, statistical agencies have to apply Statistical Disclosure Limitation (SDL) methods to such data. The goal of these methods is to minimize the risk of disclosure of the confidential information and at the same time provide legitimate data users with accurate information about the population of interest. SDL methods applicable to the microdata (i.e. collection of individual records) are often called masking methods. In this paper, several multiplicative noise masking schemes are presented. These schemes are designed to preserve positivity and inequality constraints in the data together with the vector of means and covariance matrix.

MSC: 62Pxx, 62Hxx

Keywords: Statistical disclosure limitation (SDL), SDL method, multiplicative noise, positivity and inequality constraints.

1. Introduction

When statistical offices release information about individuals, they face two conflicting goals: preserve confidentiality of the data—identities of the data subjects and values of sensitive attributes— and at the same time, release useful information for policy, research or other purposes.

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

* Georgia Southern University, Department of Mathematical Sciences, Department of Computer Engineering and Mathematics, P.O.Box 8093, GA, 30460-8093. aoganyan@georgiasouthern.edu
Received: November 2010
Accepted: March 2011

This paper focuses entirely on the microdata releases. Multiple means of access to microdata records exist, including restricted data centers (*e.g.*, ANES; MEPS; SSDS), licensing [NCES] and remote access servers [Gomatam *et al.*]. These are effective, but they do not meet all needs, and many agencies also release deliberately altered microdata publicly.

For public microdata releases, the role of statistical disclosure limitation (SDL) is to alter the data in a way that maintains the utility but limits disclosure risk.

Many Statistical Disclosure Limitation (SDL) methods can be used to prepare microdata releases. Of course, the initial step is to remove explicit identifiers for individuals – names, addresses and social security numbers.

Almost always, removal of identifiers alone is inadequate. Rare attribute combinations (for example, a 17-year old widow) can lead to re-identification. Moreover, in high-dimensional data, virtually every subject may have a unique set of attributes. Therefore, almost invariably, released data attributes must be modified. Some SDL techniques coarsen the resolution of the data; for example, date of birth can be replaced by age, and age may be reported in five-year intervals. Extreme attribute values can be top- or bottom-coded.

Another approach is to generate synthetic records, which are draws from a distribution (typically, a posterior predictive distribution) representing the original data.

Other methods actually change attribute values. Examples are addition of noise, data swapping and microaggregation [Karr *et al.* (2006); Oganian and Karr (2006)]. We term methods whose output is a perturbed version of the original data *the perturbation methods*. This paper focuses on one of these – a perturbation by means of externally generated "noise." Each specific perturbation method has consequences on both disclosure risk and data utility. Some limit risk effectively but are poor at preserving utility, while others yield high utility, but at the price of high risk. No method is superior with respect to both. Oganian and Karr (2006) show how to combine two methods with the goal of capturing the good aspects of each.

From a data utility perspective, it is important to preserve qualitative characteristics of data, for example, positivity constraints of the form $X \geq 0$ for some variables and inter-attribute relationships such as linear inequalities. Age, many economic variables (gross income, taxes) and many demographic variables (number of employees, number of students in the sixth grade) obey positivity constraints; examples of inequality constraints are "Federal taxes $\leq$ gross income", "number of salaried employees $\leq$ number of employees" and "year of birth $\leq$ year of death."

There is also a risk aspect. Because such characteristics are derived from domain knowledge available to both legitimate data users and intruders, failure to preserve them poses a disclosure risk: the extent to which constraints are violated can be informative about the nature and intensity of the SDL applied to the data.

Some SDL methods preserve such characteristics more by coincidence than by design, and only partially. For instance, data swapping preserves positivity, but not multi-attribute constraints. Microaggregation preserves positivity, but whether it preserves linear inequalities depends on specifics of the implementation.

In this paper, we present several SDL methods applicable to numerical data that *preserves positivity constraints, inequality constraints and the first two moments* – the vector of means and covariance matrix.

For the purposes of this paper, the original and released (which we hereafter term masked) databases are flat files in which rows represent data subjects (individuals, households, business establishments, ...) and columns numerical attributes of those subjects. We denote the original data by X_o and the released (masked) data by X_m. We assume that some variables in X_o are nonnegative, others can take positive and negative values. The goal is to obtain $X_m(j) \geq 0$ for those variables j which are nonnegative in the original data, also X_m should have the same mean and covariance matrix as X_o.

As background, the analogous procedure for addition of noise to unconstrained numerical data is as follows. Let Σ_o be the covariance matrix of X_o – in practice, one can use either the usual empirical estimator or a shrinkage-based estimator. Let $k > 0$ be a parameter chosen by the agency; then

$$X_m = E[X_o] + \frac{(X_o - E[X_o]) + E}{\sqrt{1+k}}, \tag{1}$$

where the noise E has distribution $N(\mathbf{0}, k\Sigma_o)$, has the requisite properties [Oganian and Karr (2006)]. Note that the value of k need not be released, even if it were made known that the method of SDL is addition of noise. As $k \to \infty$, X_m becomes a very simplistic form of synthetic data [Reiter (2002)], and any non-normal distributional characteristics of X_o are lost.

The structure of this paper is the following: several multivariate noise protocols that preserve the first two moments are presented in Section 2, close forms for higher order moments are given in Section 3, the extension of these protocols to satisfy inequality constraints is described in Section 4 and the results of the numerical experiments are reported in Section 5.

2. Multiplicative noise protocols

Suppose that X_o contain n records, each with d numerical attributes. Some of the attributes are nonnegative, denote them ${X_o}^p$. We wish to construct and release a masked data set X_m with these characteristics:

$${X_m}^p \geq 0 \tag{2}$$

$$E[X_m] = E[X_o] \tag{3}$$

$$\Sigma(X_m) = \Sigma(X_o), \tag{4}$$

where ${X_m}^p$ are the masked values of ${X_o}^p$ and $\Sigma(\cdot)$ means "covariance matrix of $(\cdot)$."

Oganian and Karr (2011) proposed a masking scheme which preserves the positivity, means and covariance matrix. The basis of this scheme is to use multiplicative noise, implemented by taking logarithms, applying additive, normally distributed noise and exponentiating. This scheme works only if all the variables in the data set are nonnegative. Below are the details.

Let E be noise that is conditionally independent of X_o given $E[X_o]$ and $\Sigma(X_o)$, and satisfies

$$E[X_o \circ \exp(E)] = E[X_o] \tag{5}$$

$$\Sigma(X_o \circ \exp(E)) = (1+k)\Sigma(X_o), \tag{6}$$

where $k > 0$ is an agency-chosen parameter and $\circ$ denotes elementwise matrix multiplication (Schur or Hadamard product). That is, the exponentiation in (5), (6) and elsewhere below also takes place componentwise. Then

$$X_m = \frac{(\sqrt{1+k}-1)E[X_o] + [X_o \circ \exp(E)]}{\sqrt{1+k}} \tag{7}$$

satisfies (2)–(4).

For normally distributed noise E, Oganian and Karr (2011) showed that the following vector of means μ_E and the covariance matrix Σ_E should be chosen for E to satisfy (5) and (6):

$$\Sigma_E(i,j) = \log\left(1 + \frac{k\Sigma_o(i,j)}{E[X_o(i)X_o(j)]}\right), \qquad i,j = 1,\ldots,d \tag{8}$$

$$\mu_E(i) = -\sigma_E(i)/2, \qquad i = 1,\ldots,d. \tag{9}$$

Here, d is the number of the dimensions in the data.

Note that the fact that the original data are multiplied by the lognormal noise does not mean that such a noise introduce a significant skewness to the data. In fact, because of the specific choice of the parameters of the lognormal distribution, the introduced skewness is minimal. In particular, from (8), the variance of the lognormal noise is less than k, where k is a parameter of the method and typically small, *e.g.* 0.15. The lognormal noise with such a small variance is practically symmetrical with very slight skew to the right, to the point that its distribution is almost indistinguishable from a normal distribution.

If the data set contains not only nonnegative variables but variables with negative values as well, the scheme described above cannot be applied directly. The variables with negative and positive values may lead to

$$1 + \frac{k\Sigma_o(i,j)}{E[X_o(i)X_o(j)]} < 0 \tag{10}$$

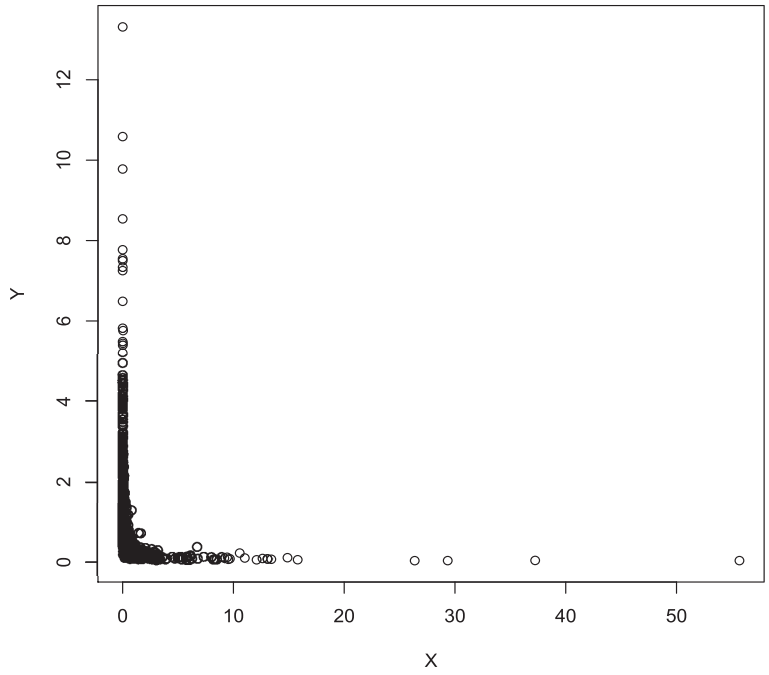

Figure 1: *Example of a data set when covariance matrix for noise cannot be computed.*

so, the covariance matrix (8) cannot be computed. After experimentation with different data sets, it was noticed that for some very rare distributions of values of ${X_o}^p$, (10) may still hold. This will happen when positive variables are negatively correlated and when

$$E[X_o(i)X_o(j)] < \frac{k}{1+k}E[X_o(i)]E[X_o(j)]$$

For positive variables this may happen only when the values of the variables are strongly aligned along the axes. Example of such distribution is shown in Figure 1.

One possible solution to this problem is described in [Oganian (2010)] which consist of converting all the variables to z-scores and making these z-scores nonnegative by adding some value (or vector – for multivariate data) lag, such that $lag \geq |min(Z)|$. Denote these nonnegative z-scores by Z_p. Then masking scheme described by (7), (8) and (9) can be applied to Z_p and after that the resulting data are returned to the original scale:

$$Z_m = \frac{(\sqrt{1+k}-1)lag + [Z_p \circ \exp(E^{z_p})]}{\sqrt{1+k}} \tag{11}$$

$$X_m = (Z_m - lag) \circ \sigma_o + E(X_o) \tag{12}$$

where σ_o is the main diagonal of Σ_o and the noise E^{z_p} has the following mean and covariance matrix:

$$\Sigma_{E_{z_p}}(i,j) = \log\left(1 + \frac{k\Sigma_{z_p}(i,j)}{E[Z_p(i)Z_p(j)]}\right), \qquad i,j = 1,\ldots,d \tag{13}$$

$$\mu_{E_{z_p}}(i) = -\sigma_{E_{z_p}}(i)/2, \qquad i = 1,\ldots,d. \tag{14}$$

where $\Sigma_{z_p}(i,j)$ is the (i,j) element of the covariance matrix of Z_p.

Masked data X_m in this case can be represented as

$$\begin{aligned} X_m &= \Big(\frac{(Z_p \circ \exp(E^{z_p}) + (\sqrt{1+k}-1)lag}{\sqrt{1+k}} - lag\Big) \circ \sigma_o + E(X_o) = \\ &= \frac{[X_o \circ \exp(E^{z_p})] - E(X_o) \circ exp(E^{z_p}) + \sigma_o \circ lag \circ exp(E^{z_p})}{\sqrt{1+k}} + \\ &+ \frac{E(X_o)\sqrt{1+k} - lag \circ \sigma_o}{\sqrt{1+k}} \end{aligned} \tag{15}$$

It is easy to see that such scheme preserves the means and covariance matrix:

$$\begin{aligned} E(X_m) &= \frac{1}{\sqrt{k+1}}[E(X_o) - E(X_o) + \sigma_o \circ lag - \sigma_o \circ lag + \\ &+ E(X_o)\sqrt{1+k}] = E(X_o) \end{aligned} \tag{16}$$

The equality in the formula above follows from the fact that the noise is independent from X_o and $E(exp(E^{z_p})) = 1$.

$$\begin{aligned} \Sigma_m(i,j) &= \Sigma\Big(\frac{Z_p(i)\exp(E^{z_p}(i))\sigma_o(i)}{\sqrt{1+k}}, \frac{Z_p(j)\exp(E^{z_p}(j))\sigma_o(j)}{\sqrt{1+k}}\Big) = \\ &= \frac{\sigma_o(i)\sigma_o(j)}{1+k}(1+k)cov(Z_p(i), Z_p(j)) = \\ &= \sigma_o(i)\sigma_o(j)cor(X_o(i), X_o(j)) = \Sigma_o(i,j) \end{aligned} \tag{17}$$

where $cov(\cdot)$ and $cor(\cdot)$ denote covariance and correlation of $(\cdot)$ respectively. Note that the second equality in the formula above follows from the property (6).

Oganian (2010) shows that masking scheme (15) with the specific choice for lag will never lead to the case described by (10).

In particular, first, let us see what are the possible values for lag in this scheme. lag should be greater than $|min(Z)|$, however, a very big lag may lead to a negative masked data (this follows from equation(12)), which violates positivity constrants for the variables ${X_o}^p$.

From (11), Z_m is minimized when $E_n \to -\infty$:

$$min(Z_m) > \frac{(\sqrt{1+k}-1)lag}{\sqrt{1+k}}$$

From (12), $min(X_m)$ is larger than

$$\frac{-lag}{\sqrt{1+k}}\sigma_o + E(X_o) \tag{18}$$

To preserve positivity in the masked data, it would be enough to require positivity of (18). So, we have an upper bound for lag:

$$lag \leq \frac{E(X_o)}{\sigma_o}\sqrt{1+k}$$

where division is done componentwise.

The lower bound for lag is $|min(Z)|$. For nonnegative variables with zeros $|min(Z)| = E(X_o)/\sigma_o$. So, the lower and upper bound for lag are:

$$\frac{E(X_o)}{\sigma_o} \leq lag \leq \frac{E(X_o)}{\sigma_o}\sqrt{1+k} \tag{19}$$

Let us consider a few choices for lag in this range. If we choose $lag = E(X_o)/\sigma_o$, then the scheme with z-scores transformation (15) is equivalent to the scheme without transformation (7). In fact, it is straightforward to verify that the masked data in this case can be written as:

$$X_m = \frac{(\sqrt{1+k}-1)E[X_o] + [X_o \circ \exp(E^{z_p})]}{\sqrt{1+k}} \tag{20}$$

Expression (20) is almost identical to (7) except the second term in the nominator: $[X_o \circ \exp(E^{z_p})]$.

Below we will show that even this term is identical in both schemes. In particular, after the application of our masking scheme to the positive z-scores, noise E^{z_p} has the mean and covariance matrix defined by (14) and (13) respectively.

Note, that

$$\frac{\Sigma_{z_p}(i,j)}{E[Z_p(i)Z_p(j)]} = \frac{cor(X_o(i),X_o(j))}{E[(\frac{(X_o(i)-E(X_o(i)))}{\sigma_o(i)} + lag(i))(\frac{(X_o(j)-E(X_o(j)))}{\sigma_o(j)} + lag(j))]} =$$

$$= \frac{cor(X_o(i),X_o(j))}{E[X_o(i)/\sigma_o(i) * X_o(j)/\sigma_o(j)]} = \frac{\Sigma_o(i,j)}{E[X_o(i)X_o(j)]}$$

So, when $lag = E(X_o)/\sigma_o$, the transformation to positive $z-$scores does not make any changes in the original scheme (7).

Now let us consider another extreme for lag: $lag = \sqrt{(1+k)}E(X_o)/\sigma_o$.

It is easy to verify that masked data in this case can be written as:

$$X_m = \frac{(\sqrt{1+k}-1)E[X_o] \circ \exp(E^{z_p}) + [X_o \circ \exp(E^{z_p})]}{\sqrt{1+k}} \tag{21}$$

Covariance matrix for the noise for this scheme is:

$$\Sigma_{E_{z_p}}(i,j) = \log\left(1 + \frac{k\Sigma_{z_p}(i,j)}{E[Z_p(i)Z_p(j)]}\right) \tag{22}$$

To prove that the expression under logarithm of (22) is always positive, let's express it in terms of original data.

$$Z_p(i) = \frac{X_o(i) + E(X_o(i))(\sqrt{1+k}-1))}{\sigma_o(i)}$$

It is easy to see that

$$E[Z_p(i)Z_p(j)] = \frac{E[X_o(i)X_o(j)] + kE(X_o(i))E(X_o(j))}{\sigma_o(i)\sigma_o(j)}$$

$$\Sigma_{E_{z_p}}(i,j) = \log\left(1 + \frac{k\sigma_o(i)\sigma_o(j)cor(X_o(i),X_o(j))}{E[X_o(i)X_o(j)] + kE(X_o(i))E(X_o(j))}\right) =$$

$$= \log\left(\frac{(1+k)E[X_o(i)X_o(j)]}{E[X_o(i)X_o(j)] + kE(X_o(i))E(X_o(j))}\right) \tag{23}$$

The expression under the logarithm in (23) is always positive for the nonnegative X_o, so we can always compute $\Sigma_{E_{z_p}}$. In the same way, it is possible to show that no other value for lag (in the range of its possible values) can guarantee positivity of (10) for all possible data sets.

When the data set contains variables which can take positive and negative values together with nonnegative variables, the scheme with z-scores transformations will work too. First the data should be made nonnegative by adding $|min(X_o)|$ and then scheme (21) is applied to this data. Last, to return the data to the original location, we have to substract $|min(X_o)|$ from the result of the previous step.

3. Preservation of higher moments

Multivariate noise protocols described in Section 2 maintain positivity and the first two moments. Exact preservation of higher-order moments is not guaranteed. Here we consider the extent to which higher-order moments can be distorted by the scheme with z-scores transformation, which has a wider range of applicability than the scheme without z-scores transformation

Consider the P-th (mixed) moment, $E(X_{m1}^{p_1} X_{m2}^{p_2} \cdots X_{md}^{p_d})$, where $P = \sum_{i=1}^{d} p_i$:

$$E[\prod_{j=1}^{d} X_m(j)^{p_j}] = E\Big[\prod_{j=1}^{d} (\frac{(\sqrt{1+k}-1)E(X_o(j))\exp(E^{z_p}(j))+}{\sqrt{1+k}}$$

$$+\frac{X_o(j)\exp(E^{z_p}(j))}{\sqrt{1+k}})^{p_j}\Big] = \frac{1}{(\sqrt{1+k})^{\sum_{j=1}^{d} p_j}} E\Big[\prod_{j=1}^{d}\Big(\sum_{i_j=0}^{p_j}\binom{p_j}{i_j}\times$$

$$\times(\sqrt{1+k}-1)^{p_j-i_j}E^{p_j-i_j}(X_o(j))\exp((p_j-i_j)E^{z_p}(j))X_o(j)\times$$

$$\times\exp(i_j E^{z_p}(j))\Big)\Big] = \frac{1}{(\sqrt{1+k})^{\sum_{j=1}^{d} p_j}}\sum_{i_1=0}^{p_1}\cdots\sum_{i_d=0}^{p_d} E\Big[\prod_{j=1}^{d} X_o(j)^{i_j}\Big]\times$$

$$\times E\Big[\exp(\sum_{j=1}^{d} p_j E^{z_p}(j))\Big]\times(\sqrt{1+k}-1)^{\sum_{j=1}^{d}(p_j-i_j)}\prod_{j=1}^{d}\binom{p_j}{i_j}E^{p_j-i_j}[X_o(j)]\,.$$

Note that $W = \sum_{j=1}^{d} p_j E^{z_p}(j)$ is a weighted sum of d normal variables that are not independents but are jointly normal. So, W is a normal variable too. Thus, $\exp(W)$ is log-normal with mean equal to $\exp(\mu_W + 0.5 Var_W)$. Then, since

$$E[\exp(W)] =$$

$$= \exp\left[\sum_{j=1}^{d} p_j\mu_E(j) + 0.5\left(\sum_{j=1}^{d} p_j^2\Sigma_{E_{z_p}}(jj) + \sum_{j<l} 2p_j p_l\Sigma_{E_{z_p}}(jl)\right)\right]$$

$$= \exp\left[-0.5\sum_{j=1}^{d} i_j\Sigma_{E_{z_p}}(jj) + 0.5\left(\sum_{j=1}^{d} i_j^2\Sigma_{E_{z_p}}(jj) + \sum_{j<l} 2i_j i_l\Sigma_{E_{z_p}}(jl)\right)\right]$$

$$= \prod_{j=1}^{d}\left(\frac{(1+k)E[X_o(j)^2]}{E[X_o(j)^2]+kE^2[X_o(j)]}\right)^{\frac{p_j(p_j-1)}{2}}\prod_{j<l}\left(\frac{(1+k)E[X_o(j)X_o(l)]}{E[X_o(j)X_o(l)]+kE[X_o(j)X_o(l)]}\right)^{p_j p_l} \tag{24}$$

we obtain

$$E\left[\prod_{j=1}^{d} X_m(j)^{p_j}\right] = \sum_{i_1=0}^{p_1} \cdots \sum_{i_d=0}^{p_d} E\left[\prod_{j=1}^{d} X_o(j)^{i_j}\right] \frac{(\sqrt{1+k}-1)^{\sum_{j=1}^{d}(p_j-i_j)}}{(\sqrt{1+k})^{\sum_{j=0}^{d} p_j}} \times$$

$$\prod_{j=1}^{d} \binom{p_j}{i_j} \times \prod_{j=1}^{d} \left(\frac{(1+k)E[X_o(j)^2]}{E[X_o(j)^2]+kE^2[X_o(j)]}\right)^{\frac{p_j(p_j-1)}{2}} \times \tag{25}$$

$$\times \prod_{j<l} \left(\frac{(1+k)E[X_o(j)X_o(l)]}{E[X_o(j)X_o(l)]+kE[X_o(j)X_o(l)]}\right)^{p_j p_l} \prod_{j=1}^{d} E^{p_j-i_j}[X_o(j)]\,.$$

Now, (25) can be written as

$$E\left[\prod_{j=1}^{d} X_m(j)^{p_j}\right] = E\left[\prod_{j=1}^{d} X_o(j)^{p_j}\right] \frac{A}{(\sqrt{1+k})^{\sum_{j=1}^{d} p_j}} + \sum_{i_1=0}^{u_1} \cdots \sum_{i_d=0}^{u_d} E\left[\prod_{j=1}^{d} X_o(j)^{i_j}\right] \times$$

$$\times \prod_{j=1}^{d} \binom{p_j}{i_j} E^{p_j-i_j}[X_o(j)] \frac{A(\sqrt{1+k}-1)^{\sum_{j=1}^{d}(p_j-i_j)}}{(\sqrt{1+c})^{\sum_{j=1}^{d} p_j}}, \tag{26}$$

where $u_1 \in \{(p_1-1), p_1\}$, $u_2 \in \{(p_2-1), p_2\} \cdots$ $u_d \in \{(p_d-1), p_d\}$, such that $u_1, u_2 \cdots u_d \neq \{p_1, p_2 \cdots p_d\}$ and

$$A = \prod_{j=1}^{d} \left(\frac{(1+k)E[X_o(j)^2}{E[X_o(j)^2]+kE^2[X_o(j)]}\right)^{\frac{p_j(p_j-1)}{2}} \times$$

$$\times \prod_{j<l} \left(\frac{(1+k)E[X_o(j)X_o(l)]}{E[X_o(j)X_o(l)]+kE[X_o(j)X_o(l)]}\right)^{p_j p_l}$$

From (26) we see how the moments of the original and masked data are related. If the agency decides to release information about masking algorithm – in particular the value of k, then this formula can be reported to data users, allowing them to adjust their analyses and to calculate the original moments. To compute the original moments users would employ expression (26) recursively: first and second order moments of the original data in (26) can be substituted by the corresponding moments computed on the masked data. All higher order original moments can be computed recursively using formula (26). However, the safety of the releasing k is problematic in some scenarios, because doing so might lead to attribute disclosure risk for some records.

A question of practical interest is how large the expression (26) can be, compared to the corresponding original moments. Because the masked data are scaled to have the

same covariance matrix as the original data, higher-order moments seem unlikely to be grossly inflated, but it is possible. In most our experiments with different data sets, third-order moments were only 2.5% larger than the original moments on average for skewed original data with outliers, such as the lognormal data sets described in Section 5). For the symmetrical data sets with the same covariance matrix as the lognormal ones, they were only .15% larger than the corresponding original ones. Fourth-order moments were about 15% larger on average for the lognormal original data and only .8% larger for the symmetrical data. In general, the discrepancy increases with the order of the moment, but only slowly.

4. Inequality constraints preservation

Suppose our original data in addition to positivity constraints also have inequality constraints of the form $X > Y$. For example, masking an income data with the variables "Gross income" and "Federal taxes" should produce a masked data such that "Gross income $>$ Federal taxes". The protocols described above can be used as building blocks of a new scheme which would guarantee the preservation of inequality constraints. This scheme is the following:

- Apply the multiplicative noise scheme to $(Y_o, [X_o - Y_o])$. Denote the result by $(Y^*, [X_o - Y_o]^*)$
- The masked data corresponding to (X_o, Y_o) are $(X_m, Y_m) = (Y^* + [X_o - Y_o]^*, Y^*)$

It is easy to see that this scheme preserves the means and covariance matrix.

$$
\begin{aligned}
E(X_m, Y_m) &= E(Y^* + [X_o - Y_o]^*, Y^*) = \\
&= (E(Y_o) + E[X_o - Y_o]), (E(Y_o)) = (E(X_o), E(Y_o)) \\
cov(X_m, Y_m) &= cov(Y^* + [X_o - Y_o]^*, Y^*) = var(Y_o) + \\
&+ cov([X_o - Y_o]^*, Y^*) = var(Y_o) + cov([X_o - Y_o], Y_o) = \\
&= var(Y_o) + cov(X_o, Y_o) - var(Y_o) = cov(X_o, Y_o) \\
var(X_m) &= var((Y^* + [X_o - Y_o]^*) = var(Y_o) + var([X_o - Y_o]) + \\
&+ 2cov(Y_o, [X_o - Y_o]) = var(X_o)
\end{aligned}
$$

The scheme can be readily extended for the cases when multiple variables are related by inequality constraints. For example, suppose
$X_{o_1} > X_{o_2} > X_{o_3}$, then $X_{m1} = X_3^* + [X_{o_2} - X_{o_3}]^* + [X_{o_1} - X_{o_2}]^*$, $X_{m2} = X_3^* + [X_{o2} - X_{o3}]^*$ and $X_{m3} = X_3^*$.

Or in general case if $X_{o_1} > X_{o_2} > \cdots > X_{o_{l-1}} > X_{ol}$

$$X_{mi} = X_l^* + \sum_{j=i+1}^{l} [X_{o_{j-1}} - X_{o_j}]^* \tag{27}$$

5. Numerical experiments

Both multiplicative noise schemes (with and without *z*-scores transformation) were implemented and evaluated on different data sets. These data sets have different distributional characteristics: a skewed distribution with many outliers and a symmetrical one without outliers. The symmetrical data sets had a multivariate normal distribution and the skewed sets were log-normally distributed. 500 replicates of three-dimensional normal and lognormal sets were generated. Each set had $10,000$ records. They were moderately correlated ($cor = 0.5$). The log-normal sets had means around 2 and variances ranging from 4 to 16. These sets had outliers – values close to 50 or larger.

The normal sets had means around 3.5 and variances ranging from 5 to 10. The variance inflation factor k was chosen to be 0.15 as recommended in Oganian (2003).

The experiments showed that means were very well preserved for both schemes and both types of data: the ratio of masked and original means showed only a very small variation around 1. The results on variance/covariance matrix were different for skewed and symmetrical data sets. The experiments showed that covariance matrix was preserved for the symmetrical data sets without outliers. There was slight variability in variance/covariance matrix inflation, defined as Σ_m/Σ_o, where / denotes elementwise division. Values of this ratio ranged from 0.98 to 1.02.

There was more variability in variance/covariance matrix inflation for the skewed data sets with outliers. Values of this ratio ranged approximately from 0.7 to 1.3. The scheme with *z*-transformation resulted to be slightly more stable: variance/covariance inflation ranged approximately from 0.8 to 1.2. However, the average and most frequent value were 1 in both schemes and both types of data sets, as expected.

Such variability over replications is not very surprising in light of the nature of the noise and the variation in log-normal original data, which as noted above had a number of large outlying values. Records in the original data with big values – especially outliers – can undergo significant changes when multiplied by noise, distorting the covariance matrix.

One possible solution to reduce variability in the resulting masked data when the original is skewed and/or has many outliers is to apply different levels of noise to different zones of the data, as discussed in Oganian and Karr (2011). It is illustrated in the Figure 2, where zone 1 is masked with the parameter k_1 and zone 2 with the parameter $k_2 < k_1$. Because all the protocols presented in Section 2 are designed to preserve the mean and covariance matrix of the original data, we can apply different independent noises to different zones of the data and the covariance matrix of the masked data should be the same as that of the original data.

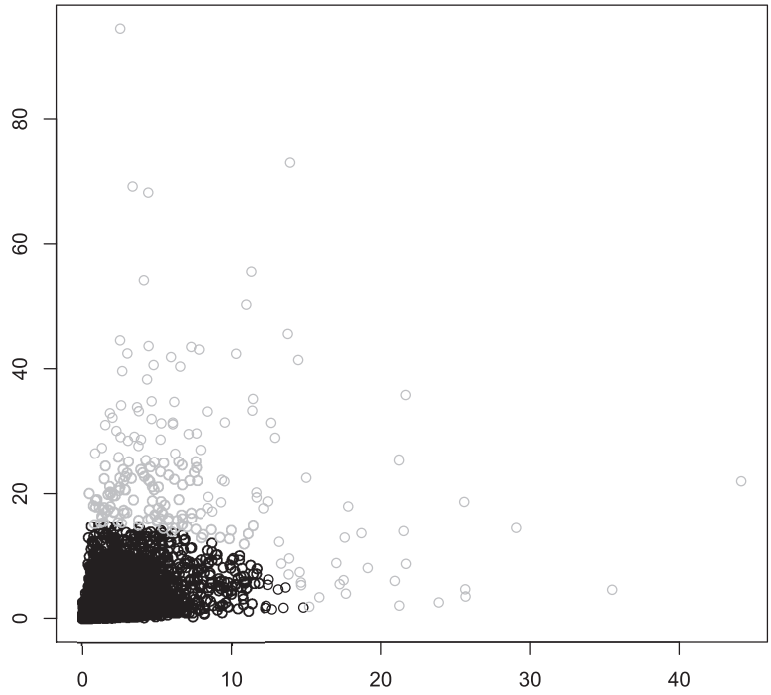

Figure 2: *Two zones of masking: black points correspond to the first zone of masking and grey points to the second zone.*

Two-zone masking was implemented with different values of k for the same three-dimensional lognormal data as in the experiment with only one zone. The first zone consisted of all the points from 0 to 15; all the other records were included in the second zone. For the second zone we chose $k_2 = 0.01$. For the first zone we chose $k_1 = 0.15$.

This approach reduced variability in the covariance matrix of the skewed data significantly: in 95% of replicates of the masked data X_m/X_o was in the interval of $[0.98, 1.02]$.

Optimal ways of variability reduction in the masked data when the original have outliers and severe skewness are the subject of our current and future research.

Note, that the multiple-zone masking may be used for other goals. For example, suppose a numerical variable in the data set has a lot of zeros, which happens often in the household data. Suppose the same numerical variable is paired with an indicator variable I, such that when $I = 0$, it is strictly positive and when $I = 1$, it is zero. Examples of I are "In the labor force" or "Income is greater than taxable min". If the agency wants to preserve such a relationship in the masked data, they can separately mask records paired with different values of the indicator variable leaving zeros in the numerical variable unchanged. Again, because our protocols preserve means and the covariance matrix, the first two moments of the overall data should be preserved.

Last, we want to discuss the disclosure risk associated with the method. Our measures of disclosure risk focus on re-identification disclosure risk. Re-identification dis-

closure is defined as an average percentage of correctly identified records when record linkage techniques [Jaro (1989)] are used to match the original and masked data. Specifically, we assume that the intruder tries to link the masked file with an external database containing a subset of the attributes present in the original data [Oganian (2003)]. The overall re-identification risk of the multiplicative noise is very small. Our experiments showed that only about 0.3% of records could be correctly identified in both schemes. So, the multiplicative noise can be successfully compared with the most protective methods, like microaggregation and rank swapping, at the same time performing significantly better than those in terms of utility.

Acknowledgments

This research was partly funded by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS). The sincerest thanks go to Alan Karr.

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

ANES. American National Election Studies Resticted Data Access, http://www.electionstudies.org/rda/anes_rda.htm

Gomatam, S., Karr, J. P. A. F., Reiter, J. P. and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers, *Statistical Science*, 20, 163–177.

Jaro, A. M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414–420.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil A. P. (2006). Framework for Evaluating the Utility of Data Altered to Protect Confidentiality, *The American Statistician*, 60, 224–232.

MEPS Medical Expenditure Panel Survey, Restricted Data Files Available at Data Centers, http://www.meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp.

NCES Confidentiality procedures, http://nces.ed.gov/StatProg/confproc.asp.

Oganian, A. (2003). *Security and Information Loss in Statistical Database Protection*, PhD thesis, Universitat Politecnica de Catalunya.

Oganian, A. (2010). Multiplicative Noise Protocols, *Privacy in Statistical Databases 2010, Lecture Notes in Computer Science*, 6344, 107–117.

Oganian, A. and Karr, A. F. (2006). Combinations of SDC Methods for Microdata Protection, *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science*, 4302, 102–113.

Oganian, A. and Karr, A. F. (2011). Masking Methods that Preserve Positivity Constraints in Microdata, *Journal of Statistical Planning and Inference*, 141, 31–41.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets, *Journal of Official Statistics*, 18, 531–544.

SSDS Social Science Data Services, http://libraries.mit.edu/guides/subjects/data/access/restricted.html.