Generating Sufficiency-based Non-Synthetic Perturbed Data

Krishnamurty Muralidhar* and Rathindra Sarathy**
* School of Management
Gatton College of Business and Economics
University of Kentucky Lexington KY 40506 USA
krishm@uky.edu
** Department of Management Science & Information Systems
Spears School of Business
Oklahoma State University, Stillwater OK 74073 USA
Sarathy@okstate.edu

Abstract. The mean vector and covariance matrix are sufficient statistics when the underlying distribution is multivariate normal. Many type of statistical analyses used in practice rely on the assumption of multivariate normality (Gaussian model). For these analyses, maintaining the mean vector and covariance matrix of the masked data to be the same as that of the original data implies that if the masked data is analyzed using these techniques, the results of such analysis will be the same as that using the original data. For numerical confidential data, a recently proposed perturbation method makes it possible to maintain the mean vector and covariance matrix of the masked data to be exactly the same as the original data. However, as it is currently proposed, the perturbed values from this method are considered synthetic because they are generated without considering the values of the confidential variables (and are based only on the non-confidential variables). Some researchers argue that synthetic data results in information loss. In this study, we provide a new methodology for generating non-synthetic perturbed data that maintains the mean vector and covariance matrix of the masked data to be exactly the same as the original data while offering a selectable degree of similarity between original and perturbed data.

1 Introduction

Many organizations, both private and public, gather and store data about individuals and other entities in their databases. In some cases, this data contains sensitive information regarding the entities. The organizations are required either ethically or legally to prevent disclosure of the sensitive information contained in their databases. Organizations use several different approaches to prevent disclosure of the sensitive information. These include restricting access to specific individuals, using database control techniques, masking the data prior to providing access, releasing interval data rather than individual data points, etc. Among masking approaches, data perturbation has gained considerable attention in the literature.

As its name implies, data perturbation involves the perturbation of the original values of the sensitive data by using random noise terms. Data perturbation is applicable both when the sensitive data is categorical and numerical. However, this study, as in most applications of data perturbation, deals with numerical confidential variables. Data perturbation is non-reversible (unless the specific random numbers used to generate the noise terms are known). Hence, the user can be told exactly what type of modification was performed on the data. There has been considerable research in the area of data perturbation relating to both categorical and numerical confidential data. Important contributions to this topic have come from a variety of researchers including Fienberg [5], Fuller [6], and Rubin [11].

For numerical data, Burridge [1] recently proposed a perturbation method (called Information Preserving Statistical Obfuscation) that makes it possible to maintain the mean vector and covariance matrix of the masked data to be exactly the same as the original data. A related approach has also been suggested by Ting et al. [12] for the case where the entire data set is confidential. A modification to ensure maximum security was proposed by Muralidhar and Sarathy [10]. It is well known that the mean vector and covariance matrix are sufficient statistics for the multivariate normal distribution. In addition, the assumption of multivariate normality also underlies many statistical analyses that are commonly used in practice including hypothesis tests and confidence intervals relating to one mean, multiple means, regression analysis, and even some advanced analyses such as principal components analysis and canonical correlation analysis. Hence, even if the original data is not multivariate normal, maintaining the mean vector and covariance matrix of the masked data to be the same as the original data has the advantage that, for many parametric statistical analyses, the results of the analysis using the masked data are identical to that using the original data. Thus, generating perturbed data with the mean vector and covariance matrix identical to the original data enhances its statistical usefulness.

One perceived disadvantage of Burridge's approach [1] is that the data generated in this manner is considered "synthetic" or artificial, since the perturbed values are not generated as a direct function of the original, sensitive values. While it is true that the perturbed data is generated to have some of the same characteristics (mean vector and covariance matrix) of the original data, the perturbed values are not "based" on the original values. Some researchers have argued that synthetic data results in information loss since the original and perturbed values may differ considerably in form and value [2]. This perceived loss of information may induce reluctance among some researchers and practitioners to adopt this otherwise useful approach. Other researchers have criticized the use of synthetic data because it may change the marginal distribution of the individual confidential variables both in the complete data set as well as in subdomains. This means that results of analyses that do not rely on the normal (Gaussian) model will be different when masked data is used in place of the original data.

Thus, there is a need to develop alternative techniques that will satisfy two important requirements for the general case in which the data set consists of both confidential and non-confidential variables: (1) maintaining the mean vector and covariance matrix of the masked data to be the same as that of the original data, and (2) for reasons mentioned above, allow data providers to generate perturbed data that has a selectable level of similarity with the original data. Currently, such a methodology is not available. In this study, we develop new perturbation models which exactly preserve the mean vector and covariance matrix of the original attributes while offering a selectable degree of similarity between original and perturbed data.

2 Burridge's Data Perturbation Method

To describe Burridge's data perturbation method, consider a data set of size n consisting of a single confidential attribute X and a single non-confidential attribute S. In the first step, a linear regression model is constructed to predict the values of X using S. From this model, the intercept ($\hat{\beta}_0$) and the slope ($\hat{\beta}_1$) are estimated, and the predicted values of x_i are computed as

$$\hat{x}_i = \hat{\beta}_0 + \left(\hat{\beta}_1 \times s_i\right).$$

(1)

Next, a vector of noise terms (A) of size n is generated from a standard normal distribution. The variable A is then regressed on (S and X) and the residuals from this regression are computed [1, 10]. Let B represent the residuals. B has mean 0 and is orthogonal to both S and X. Let σ_{BB}^2 be the variance of B. B is then transformed as follows to a new variable C,

$$c_i = \frac{b_i}{\sigma_{BB}} \sigma_{ee}, \ i = 1, 2, ..., n$$

(2)

where $\sigma_{ee}^2 = \sigma_{XX}^2 - \frac{(\sigma_{XS})^2}{\sigma_{SS}^2}$, $\sigma_{XX}^2, \sigma_{SS}^2, \text{and}\, \sigma_{XS}$ are the variance of X, S, and covariance between X and S, respectively. Finally, the perturbed values Y are generated as

$$y_i = \hat{x}_i + c_i, i = 1, 2, \ldots, n. \quad (3)$$

The resulting perturbed variable Y has *exactly* the same mean and variance as X. In addition, the covariance between Y and S is *exactly* the same as the covariance between X and S. The procedure can be easily extended to the multivariate case where there are multiple confidential variables or non-confidential variables, or both [1, page 322]. For the sake of brevity, we do not reproduce the exact procedure here.

It is important to note that, when the underlying distribution of X is not normal, the perturbed values Y resulting from Burridge's procedure will have a distribution that is different from X. Descriptive analyses performed on Y will yield different results from descriptive analyses performed on X. In addition, while the results of the regression analysis performed using Y will be identical to that using X, the residuals resulting from performing regression analysis using Y will be different from that using X. Hence, the results of residual analysis performed using Y will be different from that using X.

To illustrate Burridge's procedure, consider the data set provided in Table 1, consisting of 25 observations for X and S. For the sake of simplicity and without loss of generality, we assume that both X and S have zero mean and unit variance. The correlation between X and S was specified as 0.40. Regressing X on S results in an estimate of $\hat{\beta}_1$ = 0.40. Using this estimate, the predicted values $\hat{X}$ were computed for each observation (see Table 1). Further, we can also compute σ_{ee}^2 = (1 – 0.4^2) = 0.84. The next column represents the set of random variables generated from a univariate normal distribution with mean 0 and variance 1. The values of A were then regressed on both X and S. The residuals from this regression (B) are provided in the next column. The variance of B can be computed as 1.14986. For each observation b_i we compute (b_i × 1.14986$^{-0.5}$ × 0.84$^{0.5}$) to result in C. Finally, for each observation, we compute $y_i = \hat{x}_i + c_i$ to result in Y. It is easily verified that Y has mean 0 and variance 1, exactly the same as that of X. Further, the correlation between Y and S is 0.40 which is exactly the same as that between X and S. Hence, the results of any analysis for which the mean and covariance are sufficient statistics (such as regression analysis) will be exactly the same when using Y in place of X.

Table 1. An Illustration of Burridge's Approach

S	X	$\hat{X}$	A	B	C	Y
-2.2314	0.6972	-0.8925	1.6781	0.8870	0.7581	-0.1344
-0.6940	-1.7339	-0.2776	-0.2585	0.6289	0.5375	0.2600
1.2790	1.1636	0.5116	-0.8054	-0.5939	-0.5076	0.0040
0.4442	-0.4836	0.1777	0.0599	0.7554	0.6457	0.8233
-1.5884	0.1432	-0.6354	-1.4330	-1.7398	-1.4870	-2.1224
1.0069	0.8774	0.4028	1.4600	1.7114	1.4627	1.8655
-0.2114	0.7093	-0.0846	0.4867	0.3930	0.3359	0.2514
0.8827	0.9078	0.3531	1.6928	1.8866	1.6125	1.9655
0.4523	0.9337	0.1809	0.9563	0.9882	0.8446	1.0255
-1.0557	-2.7687	-0.4223	-1.5491	-0.3010	-0.2573	-0.6796
0.0808	0.1185	0.0323	-1.3363	-1.0504	-0.8978	-0.8654
0.0729	0.2172	0.0292	-0.7465	-0.5098	-0.4357	-0.4065
-0.3407	-1.7221	-0.1363	-0.5045	0.5002	0.4276	0.2913
0.7820	0.3549	0.3128	0.4106	0.8294	0.7089	1.0217
0.4765	1.5159	0.1906	-0.8556	-1.0892	-0.9309	-0.7403
0.8657	-0.2492	0.3463	-1.1230	-0.3910	-0.3342	0.0121
-0.0043	0.5429	-0.0017	1.6322	1.6888	1.4435	1.4417
0.5420	-0.0771	0.2168	-0.8659	-0.3275	-0.2799	-0.0631
-1.1997	-0.3667	-0.4799	-0.4529	-0.3845	-0.3287	-0.8085
1.8372	0.6342	0.7349	-1.5737	-0.9189	-0.7854	-0.0505
-1.0015	-1.3335	-0.4006	0.5212	1.1133	0.9515	0.5509
0.7178	-0.2504	0.2871	-1.3421	-0.6609	-0.5649	-0.2778
0.3491	0.2160	0.1397	-2.3514	-2.0179	-1.7248	-1.5851
0.1329	-0.1370	0.0531	-1.5908	-1.1666	-0.9971	-0.9440
-1.5950	0.0904	-0.6380	0.0534	-0.2308	-0.1973	-0.8353
	0.0000			Mean		0.0000
	1.0000			Variance		1.0000
	0.4000			Correlation with S		0.4000

By maintaining the mean vector and covariance matrix exactly, the procedure described above maximizes its analytical value of the perturbed data for many common types of statistical analyses based on the normal model. In addition, generating the perturbed values as a function of S and a noise term also maximizes security [9]. In practice, perturbed data generated in this fashion are referred to as *synthetic* data. Thus, from a statistical perspective, the procedure described above is an extremely versatile procedure resulting in analytical value *and* low disclosure risk.

Some researchers have argued, however, that synthetic data results in information loss, because the perturbed data may be very "different" from the original values. In fact, one measure of this "difference" or information loss proposed in the literature is based on the distance between the original and perturbed values. If the distance is high, then information loss is deemed to be high [2]. By this definition of information loss, procedures that generate synthetic data fare poorly since X and Y are conditionally independent (given S, mean vector, and covariance matrix). Others may argue that this approach is simply the equivalent of releasing the non-confidential microdata, the mean of confidential variables and the covariance matrix of the entire data set. Thus, regardless of the merits of Burridge's procedure, it is of little consequence if it is not used in practice. Since the primary resistance to the procedure is the *synthetic* nature of the perturbed data, we propose an alternative to Burridge's procedure by creating perturbed attributes which exactly preserve the mean vector and covariance matrix of the original attributes while offering a selectable degree of similarity between original and perturbed data. We describe such a procedure in the following sections.

3 Generating Non-Synthetic Perturbed Data that Maintain the Mean Vector and Covariance Matrix

In this section, we show that it is possible to generate non-synthetic data sets that satisfy the same statistical requirements as Burridge [1]. We begin with the simple case of a single confidential variable X and a single non-confidential variable S. We will extend this to the more general case of multiple confidential and non-confidential variables in the following section. For simplicity and without loss of generality, we will assume that mean of both X and S equal 0. Let σ^2_{XX}, σ^2_{SS}, and σ_{XS} represent the variance of X, variance of S, and the covariance between X and S, respectively. Assume that the perturbed values of Y are generated as follows:

$$y_i = \gamma + \alpha x_i + \beta s_i + e_i,\ i = 1, 2, \ldots, n, \quad (4)$$

where $\gamma = (1-\alpha)\overline{X} - \beta\overline{S}$, $\overline{X}$ is the mean of X, $\overline{S}$ is the mean of S, n represents the number of observations in the data set, and e is a noise term generated from a normal distribution with mean (exactly) equal to 0, variance (exactly) equal to σ^2_{ee}. It is also assumed that e is generated in such a manner that it is orthogonal to both X and S. The parameter α is the "similarity" parameter. When α = 1, X and Y are identical while when α = 0, X and Y are most dissimilar. Thus, the parameter α allows the data provider to control the level of similarity between the original and masked data.

In the above model we can obtain the necessary conditions under which the perturbed values will satisfy the same statistical requirements as Burridge's procedure. For convenience, we shall refer to these as "sufficiency requirements", even though they represent sufficiency requirements only under multivariate normality assumptions. First it is easy to see that if we take expectation of the expression in equation (4), the expected value of Y is $\overline{X}$. For simplicity and without loss of generality, we will assume that $\overline{X}$ and $\overline{S}$ = 0. In order to satisfy the sufficiency requirements, it is necessary that the covariance between Y and S be σ_{XS}. To satisfy this condition,

$$\mathrm{Cov}(Y,S) = \mathrm{Cov}[(\alpha X + \beta S + e)(S)] = \alpha E(XS) + \beta\ E(S^2) = \alpha\sigma_{XS} + \beta\sigma^2_{SS}$$

(5)

since E(eS) = 0. Setting this expression equal to σ_{XS} yields

$$\mathrm{Cov}(Y,S) = \sigma_{XS} = \alpha\sigma_{XS} + \beta\sigma^2_{SS}.$$

(6)

Hence, in order to maintain the covariance between Y and S to be σ_{XS}, it is necessary that:

$$\beta = (1-\alpha)\frac{\sigma_{XS}}{\sigma^2_{SS}}.$$

(7)

The third and final condition is that the variance of Y equals the variance of X. We can derive the variance of Y as

$$\sigma^2_{YY} = \sigma^2_{XX} = E[(\alpha X + \beta S + e)(\alpha X + \beta S + e)] =$$
$$\alpha^2\sigma^2_{XX} + \beta^2\sigma^2_{SS} + \sigma^2_{ee} + 2(\alpha\beta)(\sigma_{XS}). \quad (8)$$

Now substituting the value of β from equation (7) yields the following:

$$\sigma^2_{ee} = (1-\alpha^2)\left[\sigma^2_{XX} - \frac{(\sigma_{XS})^2}{\sigma^2_{SS}}\right].$$

(9)

In the above expression, the expression $\sigma^2_{XX} - \frac{(\sigma_{XS})^2}{\sigma^2_{SS}}$ is always greater than or equal to zero. Hence, in order for σ^2_{ee} to be positive, it is necessary that $-1 \leq \alpha \leq +1$. When α is negative, the perturbed values are negatively correlated to the original values. This type of model is almost never used in practice and in the remainder of the paper, we will limit our discussion the range $0 \leq \alpha \leq 1$.

For simplicity and without loss of generality, consider the situation where both X and S have mean 0 and unit variance (resulting in $\gamma = 0$). Using the results in equation (5) through equation (9), we can rewrite equation (4) as follows:

$$y_i = \alpha x_i + (1-\alpha)\rho s_i + \sqrt{(1-\alpha^2)(1-\rho^2)}u_i$$

(10)

where ρ is the correlation between X and S and u_i is normally distributed with mean 0, unit variance, and is orthogonal to both X and S. When $\alpha = 1$, we get $\gamma = 0$, $\beta = 0$, and the coefficient of u_i is also zero, resulting in $y_i = x_i$, which is the equivalent of releasing the unmodified values of X. When $0 < \alpha < 1$, the value of α represents the extent to which the perturbed value is a function of the original value. Large (small) values of α indicate that the original values are a significant (non-significant) component of the perturbed value. Conversely, as the value of α approaches zero (one), the level of perturbation increases (decreases). One additional advantage of this procedure over Burridge's procedure is that, when the underlying distribution of X is not normal, the extent to which the distribution of Y resembles the distribution of X will be a direct function of the specification of α. When α is close to 1, the distribution of Y will be very similar to that of X.

Finally, when $\alpha = 0$, the perturbed values y_i are *not* a function of the confidential value x_i, but is a function of only S and u, implying that the perturbed data is synthetic. In this case, we can verify that the values of Y are generated from the conditional distribution of X|S. In other words, when $\alpha = 0$, this model reduces to Burridge's model.

It is important to note that, like Burridge's results, these results are theoretical and require no empirical validation. In addition, the results are also distribution free and are valid regardless of the distribution of X, S, and the noise terms. An additional advantage of this procedure is that the results are valid for a data set of any size. While the results need not be verified empirically, in the following section, we provide a simple illustration of the proposed approach.

Consider the data set considered earlier consisting of 25 observations for X and S. As before, let X and S have zero mean and unit variance. Let ϱ, the correlation between X and S, equal 0.4. A simulated data set with these characteristics is provided in Table 2 along with e and perturbed values for several specifications of α. When $\alpha = 0.999$, the coefficients of s_i and u_i are very small, indicating that the perturbed values Y consist mostly of the original values X, with very little additional noise. As α decreases, the value of β and the coefficient of the error term increases. In each case, it can be easily verified that sufficiency is maintained. The mean and variance of Y are exactly the same as that of X in every case. Further, in every case, the covariance between Y and S is exactly the same as that of X and S.

We can also derive the information loss arising from this model by considering the difference between the perturbed (Y) and original (X) values. As indicated earlier, Domingo-Ferrer and Torra [2] contend that the closer the values of X and Y, the greater the data utility arising from the perturbed data. We can assess the similarity between X and Y by considering the variance of (X – Y). Large values for the variance of (X – Y) would indicate greater information loss.

Table 2. Original and perturbed data for the univariate example

	α =	0.9990	0.8000	0.6000	0.4000	0.2000	0.0000
S	X	Y	Y	Y	Y	Y	Y
-2.2314	0.6972	0.7295	0.8341	0.6678	0.4382	0.1682	-0.1344
-0.6940	-1.7339	-1.7084	-1.1201	-0.7213	-0.3674	-0.0422	0.2600
1.2790	1.1636	1.1402	0.7286	0.4967	0.3072	0.1446	0.0040
0.4442	-0.4836	-0.4540	0.0361	0.2974	0.5049	0.6780	0.8233
-1.5884	0.1432	0.0759	-0.9047	-1.3578	-1.6868	-1.9366	-2.1224
1.0069	0.8774	0.9423	1.6601	1.8577	1.9332	1.9309	1.8655
-0.2114	0.7093	0.7235	0.7521	0.6605	0.5409	0.4033	0.2514
0.8827	0.9078	0.9794	1.7644	1.9759	2.0528	2.0439	1.9655
0.4523	0.9337	0.9708	1.2899	1.3083	1.2561	1.1590	1.0255
-1.0557	-2.7687	-2.7779	-2.4538	-2.0360	-1.5967	-1.1437	-0.6796
0.0808	0.1185	0.0782	-0.4374	-0.6342	-0.7560	-0.8301	-0.8654
0.0729	0.2172	0.1975	-0.0818	-0.2066	-0.2950	-0.3601	-0.4065
-0.3407	-1.7221	-1.7014	-1.1484	-0.7457	-0.3787	-0.0345	0.2913
0.7820	0.3549	0.3866	0.7718	0.9052	0.9793	1.0158	1.0217
0.4765	1.5159	1.4730	0.6923	0.2410	-0.1325	-0.4565	-0.7403
0.8657	-0.2492	-0.2635	-0.3306	-0.2783	-0.1982	-0.1002	0.0121
-0.0043	0.5429	0.6069	1.3001	1.4798	1.5391	1.5215	1.4417
0.5420	-0.0771	-0.0894	-0.1863	-0.1835	-0.1573	-0.1162	-0.0631
-1.1997	-0.3667	-0.3816	-0.5866	-0.6749	-0.7359	-0.7793	-0.8085
1.8372	0.6342	0.5992	0.1831	0.0462	-0.0252	-0.0548	-0.0505
-1.0015	-1.3335	-1.2900	-0.5760	-0.1991	0.0983	0.3451	0.5509
0.7178	-0.2504	-0.2751	-0.4818	-0.4873	-0.4456	-0.3739	-0.2778
0.3491	0.2160	0.1388	-0.8341	-1.1943	-1.4106	-1.5350	-1.5851
0.1329	-0.1370	-0.1814	-0.6972	-0.8586	-0.9368	-0.9618	-0.9440
-1.5950	0.0904	0.0808	-0.1737	-0.3588	-0.5274	-0.6856	-0.8353
Variance of X	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
ρ	0.4000	0.4000	0.4000	0.4000	0.4000	0.4000	0.4000
Variance of (Original – Perturbed) value		0.0017	0.3360	0.6720	1.0080	1.3440	1.6800

Theoretically, we can derive the variance of (X – Y) as

$$\text{Variance}(X-Y)=\left(\sigma_{XX}^2+\sigma_{YY}^2-2\sigma_{XY}\right)=2\left(\sigma_{XX}^2-\sigma_{XY}\right), \tag{11}$$

where σ_{XY} is the covariance between X and Y. We can derive σ_{XY} as

$$\sigma_{XY}=E\left[(\alpha X+\beta S+e)X\right]=\alpha\sigma_{XX}^2+\beta\sigma_{XS}=\alpha\sigma_{XX}^2+(1-\alpha)\frac{(\sigma_{XS})^2}{\sigma_{SS}^2}. \tag{12}$$

Replacing the above result in equation (11) results in

$$\text{Variance}(X-Y)=2\left[(1-\alpha)\left(\sigma_{XX}^2-\frac{(\sigma_{XS})^2}{\sigma_{SS}^2}\right)\right]. \tag{13}$$

For a given data set, the expression $\left(\sigma_{XX}^2-\frac{(\sigma_{XS})^2}{\sigma_{SS}^2}\right)$ is fixed. Hence, the utility resulting from the perturbation is a direct function of *α*. When *α* is close to 1, the variance of (X – Y) is very small, resulting in very little information loss. When *α* approaches 0, the resulting variance of (X – Y) is higher, resulting in higher information loss. The variance of X and Y is the highest for Burridge's model (*α* = 0). Thus, when *α* is small, the results of analyses that rely on the normal (Gaussian) model using the masked data will be the same as that using the original data; but not the results of analyses for all other models. As α approaches 1, the masked data resembles the original data to a large extent. In this case, all analyses that are appropriate for the original data are also appropriate for the masked data. The extent to which the results differ will depend upon the selection of α. We believe that this represents a reasonable compromise between releasing unmasked data (and risking complete disclosure) and synthetic data. An important aspect of this compromise is the level of security that results from the selection of a specific value of α.

As the variance of (X – Y) decreases, the reduction in information loss is also accompanied by an increase in disclosure risk. When $\alpha = 1$, the variance of (X – Y) = 0 resulting in complete disclosure of confidential information. This is consistent since when $\alpha = 1$, the original values are released unmodified. When $\alpha = 0$, the variance of (X – Y) is maximum resulting in the lowest possible level of disclosure risk. However, this risk is non-zero since an intruder will use the values of the non-confidential variables to predict values of the confidential variables [8, 9]. The results in Table 2 indicate that as α approaches 1 (0), disclosure risk increases (decreases) while information loss decreases (increases). Thus, the results in Table 2 clearly indicate the explicit trade-off between information loss as measured by the variance of (X – Y) and disclosure risk. It is important to note that we do not, in general, suggest the use of the variance of (X – Y) as a measure of disclosure risk. Many sophisticated approaches for assessing disclosure risk have been proposed [3, 4, 13]. In this study however, we are only interested in assessing the *relative* level of disclosure risk resulting from the specification of α. The variance of (X – Y) serves this narrow purpose.

4 Generating Perturbed Values for the Multivariate Case

The results for the multivariate case are presented in this section. Let **X** (= $X_1, \ldots, X_K$) represent a set of K confidential variables, let **S** (= $S_1, \ldots, S_L$) represent a set of L non-confidential variables, and let **Y** (= $Y_1, \ldots, Y_k$) represent the set of K masked variables. Let n represent the number of records in the data set. Let $\mathbf{\Sigma_{XX}}$, $\mathbf{\Sigma_{SS}}$, and $\mathbf{\Sigma_{YY}}$ represent the covariance matrix of **X**, **S**, and **Y**, respectively. Let $\mathbf{\Sigma_{XS}}$ and $\mathbf{\Sigma_{YS}}$ represent the covariance between (**X** and **S**) and (**Y** and **S**), respectively. Let $\overline{\mathbf{X}}$, $\overline{\mathbf{S}}$, and $\overline{\mathbf{Y}}$ be the mean vector of **X**, **S** and **Y**, respectively. Let $\boldsymbol{\alpha}$ be a matrix of size (K × K) representing the multipliers of **X** and let $\boldsymbol{\beta}$ be a matrix of size (K × L) representing the multipliers of **S**.

Using the above definition, the perturbed values $\mathbf{y}_i$ are generated as follows:

$$\mathbf{y}_i = \boldsymbol{\gamma} + \mathbf{x}_i\boldsymbol{\alpha}^T + \mathbf{s}_i\boldsymbol{\beta}^T + \mathbf{e}_i, \; i = 1, \ldots, n.$$

(14)

As in the univariate case, we require that

(1) $\boldsymbol{\Sigma}_{YY} = \boldsymbol{\Sigma}_{XX}$,

(2) $\boldsymbol{\Sigma}_{YS} = \boldsymbol{\Sigma}_{XS}$, and

(3) $\overline{\mathbf{Y}} = \overline{\mathbf{X}}$.

Based on these requirements, we can derive the following:

$$\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}\left(\mathbf{I} - \boldsymbol{\alpha}^T\right),$$

(15)

$$\boldsymbol{\gamma} = (\mathbf{I} - \boldsymbol{\alpha})\overline{\mathbf{X}} - \boldsymbol{\beta}\overline{\mathbf{S}}, \text{and}$$

(16)

$$\boldsymbol{\Sigma}_{ee} = \left(\boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}\right) - \boldsymbol{\alpha}\left(\boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}\right)\boldsymbol{\alpha}^T$$

(17)

where **I** is the identity matrix, $\overline{\mathbf{X}}$ is the mean vector of **X**, and $\overline{\mathbf{S}}$ is the mean vector of **S** and $\boldsymbol{\Sigma}_{ee}$ is the covariance matrix of the noise terms **e**.

Thus, $\boldsymbol{\alpha}$ completely specifies the perturbation model shown in equation (14). However, $\boldsymbol{\alpha}$ must be selected carefully to ensure that the resulting covariance matrix of the error terms is positive definite. Unlike the univariate case where we were able to specify an upper and lower limit for the parameter α, the only way to verify whether a particular $\boldsymbol{\alpha}$ is appropriate is to evaluate whether $\boldsymbol{\Sigma}_{ee}$ as shown in equation (17) is positive definite.

As in the univariate case, the specification of the $\boldsymbol{\alpha}$ matrix represents the extent to which the masked data is a function of the original data. There are three possible options for specifying the $\boldsymbol{\alpha}$ matrix:

(1) $\boldsymbol{\alpha}$ is a diagonal matrix and all values in the diagonal are equal. This would represent a situation where all the confidential variables are perturbed the same level. In addition, the value of Y_i is a function only of X_i and does not depend on the value of X_j. Let α represent the value of the diagonal. In this case, it is easy to verify from equation (17) that when $0 \leq \alpha \leq 1$, $\boldsymbol{\Sigma}_{ee}$ will be positive definite.

(2) $\boldsymbol{\alpha}$ is a diagonal matrix and all values in the diagonal *are not equal.* This would represent a situation where the confidential variables are perturbed at different levels. As in the previous case, the perturbed values of a particular variable are a function of the original values of that particular confidential variable and not other confidential variables. However, in this case, after the specification of $\boldsymbol{\alpha}$, it is necessary to verify that

the resulting $\mathbf{\Sigma}_{\mathbf{ee}}$ is positive definite. If not, it may be necessary to respecify $\boldsymbol{\alpha}$ so that the resulting $\mathbf{\Sigma}_{\mathbf{ee}}$ is positive definite.

(3) $\boldsymbol{\alpha}$ is not a diagonal matrix. In this case, the perturbed values for a particular variable are a function of the original values of that confidential variable as well as the original values of other confidential variables. This is the most general of the specifications and also the most complicated. However, we do not see any advantages that arise from this specification. Hence, for the purposes of this study, we will only consider the first two specifications.

5 Empirical Examples

As noted earlier, the results derived in the previous section will hold for a data set of any size, any underlying distribution and any noise distribution. While these results require no empirical evaluation, we provide a few empirical examples to illustrate the application of the proposed approach. As a first example, consider the case where the data provider requests that all variables be perturbed the same level and specifies that α = 0.90. Table 3 provides the results of generating perturbed values for a data set consisting 25 records with 2 non-confidential variables and 2 confidential variables with the following characteristics:

$$\mathbf{\Sigma}_{\mathrm{XX}} = \begin{bmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{bmatrix},$$

(18)

$$\mathbf{\Sigma}_{\mathrm{SS}} = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}, \text{ and}$$

(19)

$$\mathbf{\Sigma}_{\mathrm{XS}} = \begin{bmatrix} 0.2 & 0.4 \\ -0.3 & -0.2 \end{bmatrix}.$$

(20)

The mean vectors $\overline{\mathbf{X}}$ and $\overline{\mathbf{S}}$ were specified as $\mathbf{0}$ resulting in γ = $\mathbf{0}$. Based on the data provider's request, $\boldsymbol{\alpha}$ is specified as:

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.900 \; 0.000 \\ 0.000 \; 0.900 \end{bmatrix}.$$

(21)

In this case, the data provider has requested that all variables be perturbed at the same level. Using equation (12), we can compute $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^{\mathbf{T}} = \begin{bmatrix} -0.006250 & 0.043750 \\ -0.028125 & -0.003125 \end{bmatrix}.$$

(22)

The resulting covariance of the noise term can be computed using equation (17) as

$$\boldsymbol{\Sigma}_{\mathbf{ee}} = \begin{bmatrix} 0.159125 & 0.089063 \\ 0.089063 & 0.172782 \end{bmatrix}.$$

(23)

It can be verified that $\boldsymbol{\Sigma}_{\mathbf{ee}}$ above is positive definite. The perturbed values for the data set above are presented in Table 3 (under Example 1). The specification of $\boldsymbol{\alpha}$ as shown in equation (23) implies that the perturbed values of the confidential variables are heavily influenced by the original values of the confidential variables. The non-confidential variables play a very small role in the perturbation. The extent of the noise term is also relatively small about 16% for the first and about 17% for the second confidential variables. The results show that the perturbed values **Y** have the same mean vector and covariance matrix as **X**. It can be easily verified that for most traditional parametric statistical analyses (such as confidence intervals and hypothesis testing for the mean, analysis of variance, regression analysis, multivariate analysis of variance, multivariate multiple regression, etc.) using (**Y** and **S)** in place of (**X** and **S)** will yield *exactly* the same results.

Now consider a specification where for the same data set, the data provider wishes that the coefficient for the first variable should be 0.9 and that for the second variable should be 0.2. In this case, the data provider would like a much higher level of perturbation for variable 2 than for variable 1. From this specification,

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.900 \; 0.000 \\ 0.000 \; 0.200 \end{bmatrix}.$$

(24)

The resulting covariance matrix for the noise term is as follows:

$$\boldsymbol{\Sigma}_{\mathbf{ee}} = \begin{bmatrix} 0.1591 & 0.3844 \\ 0.3844 & 0.8730 \end{bmatrix}.$$

(25)

It can be verified that the above covariance matrix is *not* positive definite and it would be necessary for the data provider to consider alternative specifications in this case. In order to maintain the sufficiency requirements, it is necessary that some restrictions be imposed on the selection of the $\boldsymbol{\alpha}$. Extremely disparate specifications such as the one above are likely to create problems as illustrated above.

For the same example, assume that the data provider re-specifies the following $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.800 \; 0.000 \\ 0.000 \; 0.300 \end{bmatrix}.$$

(26)

In this case, the two variables are perturbed at different levels, the first variable being perturbed less than the second variable. The resulting values for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{ee}}$ can be computed as follows:

$$\boldsymbol{\beta}^{\mathrm{T}} = \begin{bmatrix} -0.01250 & -0.19687 \\ 0.08750 & -0.02187 \end{bmatrix}.$$

(27)

$$\boldsymbol{\Sigma}_{\mathbf{ee}} = \begin{bmatrix} 0.3015 & 0.3563 \\ 0.3563 & 0.8275 \end{bmatrix}.$$

(28)

We can verify that $\boldsymbol{\Sigma}_{\mathbf{ee}}$ is positive definite. The results of applying this specification on the original data set are also provided in Table 3 (under Example 2). As before, it is easy to verify that the mean vector and covariance matrix of the masked data (**Y**, **S**) are exactly the same as that of the original data (**X**, **S**). Consequently, for those types of statistical analyses for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the masked data will yield the same results as the original data.

Table 3. Original and perturbed data for multivariate examples

				Example 1		Example 2		Example 3	
			α =	0.9000	0.0000	0.8000	0.0000	0.0000	0.0000
				0.0000	0.9000	0.0000	0.3000	0.0000	0.0000
S_1	S_2	X_1	X_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
-0.3793	0.5233	-0.2353	0.5124	0.1266	1.1704	0.4558	1.7471	0.9711	1.6967
-0.1618	0.3718	2.7470	0.6557	1.8349	0.7243	1.5139	0.4749	-1.3292	0.3339
-1.4152	-0.4663	-0.4189	0.0167	-0.4062	-0.5252	-0.5693	-0.9662	-0.1560	-0.9214
-0.4077	0.3961	-1.7639	-1.3652	-1.6766	-0.7302	-1.3380	0.7144	-0.0511	1.2222
0.2739	-0.8532	-0.0608	0.8891	0.1494	0.5883	0.0881	-0.2059	0.1673	-0.5251
1.1068	2.0839	1.8246	-0.3132	1.7891	-0.5967	1.6102	-0.9549	0.9863	-1.0123
2.1818	0.9395	-0.4480	-0.8477	-0.6355	-0.9176	-0.6351	-0.9116	-0.3213	-0.8504
-1.0073	-2.3853	-0.4345	0.2968	-0.0961	0.6512	0.0273	1.1112	-0.0790	1.1571
0.9858	1.0781	-1.3258	-1.8018	-1.3418	-2.1354	-1.3572	-1.8118	-0.0249	-1.4183
0.1139	-0.2034	-0.2942	-0.5151	0.5774	0.5455	1.0672	2.0557	1.8581	2.2952
0.4142	-0.8983	-0.3442	-0.2562	-0.5779	-0.5704	-0.7306	-0.8655	-0.9379	-0.8478
-0.8580	0.3205	1.1242	1.7454	1.0228	1.3899	0.8615	0.2431	0.1746	-0.2370
0.2773	-0.4921	-0.3768	-1.4787	0.3883	-1.3504	0.5205	-0.4834	1.4896	-0.0933
0.6457	-0.1180	0.0981	0.4519	0.1197	0.8283	0.2513	0.9642	0.0013	0.8300
-0.1903	0.6411	-0.2113	0.4001	-0.5856	-0.3399	-0.8348	-1.3984	-0.6818	-1.5803
-0.8776	0.1346	-0.5433	0.3189	-0.2591	0.5750	-0.0715	0.8461	0.6149	0.8476
-0.7941	-1.0653	-0.0322	2.0761	-0.0898	2.0652	-0.0753	1.1721	-0.4604	0.6490
-1.1311	-1.7688	-0.9724	1.1834	-1.4099	0.4398	-1.6757	-0.8390	-1.7688	-1.1468
-0.7989	-0.8843	-0.1855	-0.5428	-0.8414	-0.6864	-1.0338	-0.4976	-1.8070	-0.2596
0.2944	1.3029	1.1571	1.4588	1.7095	1.5185	1.8201	0.8500	1.9579	0.3765
-0.0638	0.0855	-0.5562	-0.3392	-0.9546	-0.1490	-0.9186	0.2239	-1.0095	0.3703
2.1286	1.3500	1.6608	-0.0101	1.6619	-0.1186	1.5464	-0.5447	0.7361	-0.7451
-1.0819	-0.3042	-0.0716	-0.9457	0.1728	-0.6908	0.2553	0.2259	0.4937	0.6096
1.5504	0.3104	-0.5016	-0.6148	-0.7614	-0.6332	-0.7699	-0.5867	-0.6810	-0.5267
-0.8059	-0.0989	0.1646	-0.9752	0.0835	-1.0525	-0.0077	-0.5629	-0.1430	-0.2240
	Variance	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Correlation with S_1	0.2000	-0.3000	0.2000	-0.3000	0.2000	-0.3000	0.2000	-0.3000
	Correlation with S_2	0.4000	-0.2000	0.4000	-0.2000	0.4000	-0.2000	0.4000	-0.2000
Correlation between X_1 & X_2		0.4000		0.4000		0.4000		0.4000	

The final illustration in Table 3 (under Example 3) shows the case where the perturbed values are generated as a function of only the non-confidential variables and the coefficients of the confidential variables are set to zero ($\boldsymbol{\alpha} = \mathbf{0}$). It is easily verified this is the same as the procedure suggested by Burridge [1].

The information loss resulting from the data is also presented in the table measured by the variance of the original and perturbed values. The measure clearly shows an increase in information loss measured using variance (X – Y) as $\boldsymbol{\alpha}$ approaches $\mathbf{0}$, because when $\boldsymbol{\alpha} = \mathbf{0}$ the perturbed values are independent of the original values resulting in synthetic data.. By contrast, as $\boldsymbol{\alpha}$ approaches $\mathbf{I}$, the resulting information loss is very low. As observed earlier, the opposite is true for disclosure risk; as $\boldsymbol{\alpha}$ approaches $\mathbf{0}$, disclosure risk decreases and as $\boldsymbol{\alpha}$ approaches $\mathbf{I}$, disclosure risk approaches 100%. Thus, the implementation of this procedure needs to be evaluated by considering the trade-off between information loss and disclosure risk.

6 Conclusions

In general, when generating perturbed data it is desirable that the analysis performed on the masked data yield results that are identical to that using the original data. The only way to achieve this objective is to maintain all sufficient statistics of the masked data to be the same as the original data. In practice, it is impossible to achieve this objective. However, many statistical analyses employed in practice are based on the normal (Gaussian) model. In these cases, if we maintain the mean vector and covariance matrix of the masked data to be the same as that of the original data, the results of the analysis using the masked data will be the same as that using the original data.

In Burridge's approach, the perturbed values are generated as a function of the non-confidential values and the estimates of the mean vector and covariance matrix. The perturbed values generated in this manner are deemed *synthetic*. Some researchers and practitioners are uncomfortable with synthetic confidential data that have no "relationship" to the orginal confidential values, and prefer perturbed values to be based on the values of the confidential variables. As indicated earlier, Burridge's approach also modifies the marginal characteristics of the original data as a whole, as well as in sub-groups. Hence, results of analyses that do not rely on the normal (Gaussian) model will be different when masked data is used in place of the original data.

In this study, we provide a simple technique that allows data providers to create perturbed attributes which preserve a set of statistics (mean vector and covariance matrix) of the original attributes exactly while offering a selectable degree of similarity between original and perturbed data. The data provider can explicitly specify the degree of similarity between the original and masked data by appropriately choosing $\boldsymbol{\alpha}$. The data provider also has the flexibility to specify different levels of similarity for different confidential variables. Hence, for confidential variables that are considered more sensitive than others, the data provider can choose a higher level of perturbation (lower level of similarity). We believe that this approach offers a reasoned compromise between releasing unmasked data and synthetic data. We hope that generating perturbed values in this manner will lead to a greater acceptance of masked numerical data. The procedure described in this study will be available in μ-Argus [7] package in the near future.

References

[1] Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, 13, 321-327.

[2] Domingo-Ferrer, J. and V. Torra (2001). Disclosure control methods and information loss for microdata. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (Eds. P. Doyle, J.I. Lane, J.M. Theeuwes, L.V. Zayatz), 91-110.

[3] Duncan, G. T., D. Lambert (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*. 81, 10-18.

[4] Dwork, C. (2006). Differential privacy. In M. Bugliesi et al. (Eds.): ICALP 2006, Part II, LNCS 4052, 1–12.

[5] Fienberg, S.E (2000). Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances. *The Philippine Statistician*, 49, 1-12.

[6] Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*. 9, 383-406.

[7] Hundepool, A., A. van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P. de Wolf, J. Domingo, V. Torra, R. Brand, and S. Giessing, (2003). μ-Argus User Manual Version 4.0, *Statistics Netherlands*, *Voorburg*.

[8] Muralidhar K., R. Parsa, R. Sarathy (1999). A general additive data perturbation method for database security. *Management Science*. 45, 1399-1415.

[9] Muralidhar, K. and R. Sarathy (2003). A theoretical basis for perturbation methods. *Statistics and Computing*. 13, 329-335.

[10] Muralidhar, K. and R. Sarathy (2005). An enhanced data perturbation approach for small data sets. *Decision Sciences*. 36, 513-529.

[11] Rubin, D.B. (1993). Discussion on 'Statistical Disclosure Limitation'. *Journal of Official Statistics*. 9, 461-468.

[12] Ting, D., S.E. Fienberg, and M. Trottini (2005). ROMM methodology for microdata release. *Monograph on Official Statistics – Work session on statistical data confidentiality*, Geneva, 89-98.

[13] Winkler, W.E. (1993a). Matching and Record Linkage. U.S. Census Bureau Research Report RR93/08, Washington DC.