WP. 42
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (vii): Risk/benefit analysis and new directions for statistical disclosure limitation

A COMMON INDEX OF SIMILARITY FOR NUMERICAL DATA MASKING TECHNIQUES

Invited Paper

Prepared by Rathindra Sarathy (Oklahoma State University) and
Krish Muralidhar (University of Kentucky), United States of America

A Common Index of Similarity for Numerical Data Masking Techniques

Rathindra Sarathy[*] and Krish Muralidhar[**]

[*]Spears College of Business, Oklahoma State University, Stillwater, OK 74078, USA
[**]Gatton College of Business & Economics, University of Kentucky, Lexington KY 40506, USA

Abstract: Numerical data masking techniques have developed from basic noise addition approaches to new approaches based on correlated noise, general additive data perturbation, multiple imputation, micro-aggregation, data swapping, data shuffling, copula based perturbation methods, etc. With this many competing approaches based on different underlying models, it has become very difficult (if not impossible) to compare the similarity between the masked data and the original data using existing measures. Hence, there is a need to develop a new measure for evaluating different methods on the same basis. In this study, we develop a new Common Index of Similarity (CIS) for all data masking techniques. The measure ranges from 0 (no similarity between the original and masked data) to 1 (the original and masked values are the same) for unmasked data allowing for comparison between different methods used to mask numeric data.

1. Introduction

Early approaches for masking numerical data focused almost exclusively on the addition of noise to the original values and releasing the noise added values as the masked value (Traub et al 1984). With such an approach, it was easy to compare different masked data values since the variance of the noise added represented a simple but effective measure of the similarity between the original and masked values. The similarity between the masked and original data was inversely related to the variance of the noise added. Greater noise variance resulted in lower similarity and vice versa. Even in this case, if for some reason, the data provider decided to add different levels of noise to different variables, comparison the similarity of entire data sets became a problem. The correlated noise addition procedure suggested by Kim (1986) alleviated this problem since the variance of noise added, relative to the original variance, was a constant across all variables.

Starting in the early 1990's, there has been a renewed interest in statistical disclosure limitation. This has led to the development of new techniques for masking numerical data. These techniques vary widely in the underlying model used to generate the masked data. Some techniques such as multiple imputation (Rubin 1993), general additive data perturbation (Muralidhar et al 1999), and sufficiency based approach (Burridge 2003, Muralidhar and Sarathy 2008) generate the masked data using linear models. Other techniques such as the copula based methods (Sarathy et al 2003) adopt a more complicated model to generate perturbed data. Even with linear models based approaches, it is not always easy to determine the "noise added" part of the model. With copula based approaches, it becomes even more difficult to isolate the noise added.

In addition to the above, we also have other approaches that do not directly involve noise addition or statistical modeling. These approaches include data swapping (Moore 1996) and micro-aggregation (Domingo-Ferrer and Mateo-Sanz 2002). In data swapping, values within a specified proximity are randomly swapped and released as the masked data. In micro-aggregation, the original values of the variables are replaced by the average value of the same variables in close proximity of the original values. The interesting aspect is that both these

techniques involve values within close proximity which is usually determined by the ranked values of the variables. There are other methods such as data shuffling (Muralidhar and Sarathy 2006) where the copula based perturbation and then performing a reverse swap so that the original values of the variables are used to generate the masked data. Data shuffling represents a hybrid of perturbation and swapping. The above methods are rank based and do not directly involve noise addition in their approaches.

When a data provider is comparing different approaches, it becomes very difficult to evaluate their relative performance without some measure that provides a common basis for such a comparison. For example, if we are to compare masked data generated using traditional noise addition method to micro-aggregated data, we have to have some common measure upon which we can compare the masked data resulting from these two techniques. Since the parameters used for the two approaches (variance for noise addition and how many values are to be aggregated for micro-aggregation), it is very difficult to directly compare the two methods. What is needed is a common benchmark that would allow for a reasonable comparison of the two methods. The objective of this manuscript is to develop a Common Index of Similarity (CIS) that will serve as such a benchmark between disparate techniques used to mask numeric data.

2. Key Characteristics of CIS

In developing CIS we attempted to capture some specific characteristics that we believe are important for such a measure. The first characteristic is that the measure should be standardized for all methods. In the case of simple noise addition, the variance of the noise represents a good measure of similarity. However, with simple noise addition, the variance of the noise is not standardized since it could vary from 0 to practically any value. It is possible for instance, to add noise whose variance is twice as much as the variance of the original variable. In addition, it is also possible that some variables are perturbed more than others. In such a case, there is no way to standardize the extent of perturbation between the two variables. The CIS measure developed in this study, by contrast, is standardized to be between 0 and 1 where a value of zero reflects no similarity with the original variables and a value of one represents unmasked data.

In developing the CIS measure, we also wanted to ensure that the measure did not directly represent either utility or disclosure risk. Since there are so many measures of both data utility and disclosure risk, the selection of a particular measure can present a problem to some methods and favor another. For instance, it is possible that a particular method satisfies data utility defined on some measure, but not on another. Similarly, it is possible for a particular method to provide low identity disclosure risk, but high value disclosure risk. Selecting a similarity index based on any particular measure of utility or disclosure could potentially bias the results in favor of one method. The CIS measure developed in this study does not *directly* represent any particular measure of data utility or disclosure risk and hence is not biased in favor of any particular method. In the following section we provide a detailed description of the similarity index developed based on the above characteristics.

3. Description of CIS

Let $\boldsymbol{X}$ represent a set of numerical confidential variables and let $\boldsymbol{Y}$ represent the masked values for the same variables. The masked variables can be generated by any data masking approach available. The objective of this manuscript is develop an index to measure the similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the characteristics described above.

The CIS measure that we propose is based on the concept of canonical correlation between numerical variables. Canonical correlation analysis (CCA) is a statistical procedure that allows us to *identify and quantify* the relationship between *two sets of variables*. CCA identifies a linear combination of variables in one set that have the *highest* correlation with a linear combination of variables in another set (Johnson and Wichern 1992). In the case of data masking we are concerned with one set of original variables ($\boldsymbol{X}$) and another set of masked variables ($\boldsymbol{Y}$) and CCA allows us to do that.

Specifically, the CIS measure is computed as follows. Let $\boldsymbol{\Sigma_{XX}}$, $\boldsymbol{\Sigma_{YY}}$, and $\boldsymbol{\Sigma_{XY}}$ represent the covariance matrix of $\boldsymbol{X}$, $\boldsymbol{Y}$, and the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Now consider the following expression:

$$\boldsymbol{\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}} \qquad (1)$$

It is easy to verify that the resulting matrix is symmetric with dimensions (k × k) where k represents the number of (confidential) variables in $\boldsymbol{X}$. Note that the above expression can also be specified by starting with $\boldsymbol{Y}$ rather than $\boldsymbol{X}$. The results of the analysis would be the same regardless of the manner in which this expression is specified.

The matrix resulting from equation (1) has k eigen values. The primary (largest) eigen value of the above expression (λ) has the following interesting statistical property. It represents the *maximum proportion of variability in $\boldsymbol{Y}$ that can be attributed to $\boldsymbol{X}$.* In other words, consider two arbitrary linear combinations of $\boldsymbol{a^TX}$ and $\boldsymbol{b^TY}$ where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of length k. Canonical correlation analysis identifies that particular vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ that maximizes the correlation between $\boldsymbol{a^TX}$ and $\boldsymbol{b^TY}$. The square root of the primary eigen value represents the maximum absolute correlation between *any linear combination of $\boldsymbol{X}$ and $\boldsymbol{Y}$*. Note that the square of the remaining (k – 1) eigen values also represent the lower level canonical correlations.

CCA has been applied extensively in practice. Typically, in addition to identifying the highest correlation, CCA analysis is often used to identify $\boldsymbol{a}$ and $\boldsymbol{b}$ that result in the maximum correlation. CCA represents an important multivariate analytical tool that enables us to simplify the relationship between two sets of potentially complex variables. While the eigen values and the corresponding correlation can be computed directly from equation (1), most statistical analysis tools (such as SAS) provide for comprehensive CCA analysis where they identify the eigen values, correlations, vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and other detailed analysis.
For these reasons, we propose that the primary eigen value from equation (1) be used as the Common Index of Similarity (CIS). The interesting feature of this measure is that it incorporates the relationship between the variables since the covariance matrix of $\boldsymbol{X}$ plays a critical role in the assessment. For instance, for a single variable and noise variance equal to $d\sigma^2$, the expression in equation (1) reduces to $1/(1+d)$. For multiple variables, a simple reduction is not possible

because of the relationship between the variables in $\boldsymbol{X}$. As we will show in the following sections, with multiple variables, CIS is actually higher than what would be expected. In other words, because of the structure of $\boldsymbol{X}$, for a given level of noise is added, the resulting similarity is higher than expected.

By contrast, for correlated noise, for any number of variables, and variance of noise added is $d\boldsymbol{\Sigma}_{\boldsymbol{XX}}$, the expression in equation (1) reduces to 1/(1+d). Typically, when comparing simple noise addition and correlated noise addition, the measure of comparison is usually specified by the value of d. However, when we look at the similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$, there is a considerable difference between simple noise addition and correlated noise addition. The CIS of correlated noise addition will be less than that of simple noise addition when the number of variables is greater than 1. The only case where they would be equal for more than one variable is when $\boldsymbol{\Sigma}_{\boldsymbol{XX}}$ is a diagonal matrix with zero in all off-diagonal terms, that is, the variables in $\boldsymbol{X}$ are uncorrelated.

Thus, using this measure provides the following advantages:

(1) It provides a standardized measure of similarity (on a scale of 0 to 1),
(2) It incorporates the structure of the confidential variables in measuring similarity,
(3) It provides a meaningful interpretation of the "variance added" by the data masking procedure,
(4) It is consistent with a variance measure that is used in noise addition,
(5) It can be applied to any numerical data set, and
(6) It is simple to compute using standard statistical analysis tools.

Note that this measure is completely consistent with the variance measure used in noise addition. In fact, we can show that when **X** and **Y** are vectors, the value of CIS, λ, reduces to 1/(1 + noise variance).

4. Empirical Example of the Application of CIS

In this section, we provide a few simple examples to illustrate the application of CIS. For the purpose of this analysis, we used the Census Data Set that has been used extensively in practice. The data set consists of 1080 observations with 13 variables. For the first example, we used only 2 variables (Adjusted gross income (AGI) and Employee contribution for health insurance (EMCONTRB)) in columns 2-3 of the data set. In the first example, we added independent noise to each of the variables ranging from 1% to 100% of the variance of each variable. The results of this analysis are presented in Table 1.

From Table 1, it is evident that when the noise added to one of the variable is 10% of the variance or less, CSI stays at above 90%. For noise levels between 10% and 25% for both variables, CSI stays close to 80%. With 50% noise added for both variables, CSI is close to 70%. It is only when the noise added for both variables is 100% that CSI drops to around 60%. Typically, 100% noise added would be considered too high and we would expect there to be little or no similarity between the original and masked variables. However, the CSI clearly shows

that even with such high level of noise, the data maintains a good "similarity" of approximately 60%.

		Noise Level for EMCONTRB					
		1%	5%	10%	25%	50%	100%
Noise level for AGI	1%	0.9930	0.9900	0.9897	0.9896	0.9895	0.9895
	5%	0.9903	0.9658	0.9566	0.9518	0.9505	0.9499
	10%	0.9901	0.9576	0.9337	0.9135	0.9079	0.9056
	25%	0.9899	0.9533	0.9158	0.8490	0.8145	0.7998
	50%	0.9899	0.9522	0.9108	0.8185	0.7372	0.6864
	100%	0.9899	0.9516	0.9086	0.8055	0.6925	0.5830

Table 1. CSI for different levels of simple noise addition

In the next experiment, we investigated the impact of the number of variables on the measure. In this experiment, we used 8 variables (columns 2 to 9) from the Census data set. In each case, we perturbed all variables the same level ranging from 1% to 100% as in the prior experiment. Table 2 provides the results.

	Number of variables							
Noise level (d)	1	2	3	4	5	6	7	8
1%	0.9895	0.9930	0.9952	0.9966	0.9973	0.9978	0.9978	0.9979
5%	0.9494	0.9658	0.9763	0.9834	0.9865	0.9890	0.9891	0.9895
10%	0.9037	0.9337	0.9538	0.9673	0.9733	0.9782	0.9785	0.9792
25%	0.7896	0.8490	0.8921	0.9223	0.9360	0.9472	0.9480	0.9496
50%	0.6522	0.7372	0.8056	0.8560	0.8798	0.8998	0.9011	0.9041
100%	0.4838	0.5830	0.6755	0.7488	0.7859	0.8178	0.8203	0.8252

Table 2. CIS for different number of variables with simple additive noise

The above results provide some interesting insights. When the number of variables equals 1, the results indicate that CIS provides results that extremely close to the theoretically expected value of 1/(1 + d) where d represents the level of noise added. The small differences can be attributed to random number generation. This not the case when there are multiple variables. For example, when the noise added is 100% for each variable, we expect the similarity to be 1/(1+d) = 0.5. However, with 8 variables from the Census data set, we observe that the actual CIS is only 0.8252. As the number of variables increases, the actual CIS is considerably different from the expected level of 1/(1 + d). This can be attributed directly to the structure of $\boldsymbol{\Sigma}_{XX}$ in the Census data set. Incorporating the inherent relationships between the variables in $\boldsymbol{X}$ is an important characteristic of the CIS measure.

The impact of this incorporating the covariance structure of $\boldsymbol{X}$ can be assessed if we consider a comparison between simple noise addition and correlated noise addition. Assume that we wish to compare simple noise addition with a noise level of 100% with correlated noise addition. Traditionally, the appropriate comparison would be the same level of noise (100%) for correlated noise as well. However, when using the correlated noise method with noise equal to 100%, the

resulting value of CIS ≈ 0.50 (without accounting for the inherent variability in generating the noise) whereas for simple noise addition is 0.8252. Hence, for this data set, when comparing the simple noise for 8 variables with noise = 100% of the variance of the individual variables, the appropriate comparative level of correlated noise added would be only about 21.18%. Thus, the new CIS measure provides a better benchmark for comparing two equivalent approaches which incorporates the inherent relationship structure present in the data. *A direct comparison of simple noise addition and correlated noise addition using the level of noise added would be appropriate if and only if all variables in* ***X*** *are uncorrelated.*

The final experiment consisted of an evaluation based on different methods of data masking. For this experiment, we used three variables (AGI, EMPCONTRB, and FEDTAX, columns 2-4 of the Census data set). And for this illustration, we used four different data masking techniques, namely, simple noise addition, correlated noise addition, univariate micro-aggregation, and proximity based rank swapping. For each method, we used 6 different levels of data masking. For simple noise addition and correlated noise addition, we used the same six levels (1%, 5%, 10%, 25%, 50%, and 100%) as before. For univariate micro-aggregation, we specified the number of observations to be aggregated (k) to 5, 25, 50, 100, 200, and 500. Finally, for data swapping, we specified the rank proximity to be 5, 25, 50, 100, 200, and 500. The results are presented in Table 3.

Method							
Simple Noise Addition		Correlated Noise Addition		Univariate Micro-Aggregation		Rank Based Swapping	
Noise level (d)	CIS	Noise level (d)	CIS	Number of observations aggregated	CIS	Proximity of the observations swapped	CIS
1%	0.9952	1%	0.9907	5	1.0000	5	0.9999
5%	0.9763	5%	0.9559	25	0.9996	25	0.9978
10%	0.9538	10%	0.9167	50	0.9984	50	0.9937
25%	0.8921	25%	0.8194	100	0.9944	100	0.9753
50%	0.8056	50%	0.6872	200	0.9805	200	0.9157
100%	0.6755	100%	0.5126	500	0.8682	500	0.5464

Table 3. Comparison of different data masking techniques

As observed earlier, CIS results from correlated noise addition method are very close to the theoretically expected CIS value of 1/(1 + d). For all other approaches, it is impossible to predict the exact value of CIS since the relationship structure of ***X*** dictates the actual CIS value observed. This is clearly illustrated in the results in Table 3. Both univariate micro-aggregation and rank based swapping result in high CIS compared to noise addition methods. This could be seen both as an advantage and a disadvantage. A high CIS indicates that the masked data is likely to retain the same characteristics as the original data and hence presents an advantage in terms of data utility. Simultaneously, a high CIS could also possibly result in high disclosure risk and hence presents a disadvantage in terms of security. This is only to be expected since for all these methods, there is an inherent tradeoff between data utility and disclosure risk.

The table also provides adequate information for setting benchmarks for the different techniques. If a data provider is considering data swapping and simple noise addition, then it is easy to verify that in order to achieve the same CIS of the noise addition method with 5% noise, one would have to swap values that are within 100 ranks of the original values. In this case, both methods result in CIS of approximately 0.975. The interesting part is that is another data set, the same comparison may result in completely different results due to the structure of the confidential variables. That CIS is able to capture this important characteristic is a significant advantage of this measure.

5. Conclusions

The objective of this paper is to present a new measure to quantify the similarity between the original and masked data, regardless of the underlying data masking mechanism. Development of such a measure would assist in making effective comparisons between different data masking techniques. We present just such a measure based on canonical correlation analysis. Experimental evaluations using the Census data set indicate that the measure performs well in comparing different techniques and could be used effectively in practice.

References

1. Burridge, J. (2003). Information Preserving Statistical Obfuscation. *Statistics and Computing* 13 321-327.
2. Domingo-Ferrer, J. and J.M. Mateo-Sanz (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*. 14, 189-201.
3. Johnson, R. A. and Wichern, D.W., 1992, *Applied Multivariate Statistical Analysis*, Prentice Hall, NJ.
4. Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. *Proceedings of the American Statistical Association, Survey Research Methods Section*, ASA, Washington D.C. 370-374.
5. Moore, R.A. (1996). Controlled data swapping for masking public use micro datasets. *U.S. Census Bureau Research Report* 96/04.
6. Muralidhar, K. and R. Sarathy (2006). Data Shuffling - A New Masking Approach for Numerical Data. *Management Science* 52 658-670.
7. Muralidhar, K., Parsa, R. and Sarathy, R. (1999). A general additive data perturbation method for database security. *Management Science,* 45, 1399-1415.
8. Rubin, D.B. (1993). Discussion on 'Statistical Disclosure Limitation'. *Journal of Official Statistics*. 9, 461-468.
9. Sarathy R., K. Muralidhar, R. Parsa. (2002). Perturbing non-normal confidential variables: The copula approach. *Management Science* 48 1613-1627.
10. Traub, J.E., Yemini, Y. and Wozniakowski, H. (1984). The statistical security of a statistical database security. *ACM Trans. Database Systems,* 19, 47-63.