December 2001

Issues in Perturbing Non-Normal, Confidential Attributes

Rathindra Sarathy
Oklahoma State University

Krish Muralidhar
University of Kentucky

Rahul Parsa
Drake University

ISSUES IN PERTURBING NON-NORMAL, CONFIDENTIAL ATTRIBUTES

Rathindra Sarathy
Oklahoma State University
sarathy@okstate.edu

Krish Muralidhar
University of Kentucky
kmura0@pop.uky.edu

Rahul Parsa
Drake University
rahul.parsa@drake.edu

Abstract

Muralidhar et al. (1999) recently provided a new method of perturbation that, for databases whose numerical attributes can be described by a multivariate normal distribution, provided maximum data utility and minimum disclosure risk. For non-normal databases however, the method resulted in bias and provided less than maximum data utility. In this study, we identify the specific issues relating to perturbing non-normal databases, and provide results of using a new approach that eliminates problems with existing approaches.

Introduction

A variety of techniques have been developed for providing data access to legitimate users of databases while preserving confidentiality of sensitive information from snoopers (or legitimate users attempting to gather confidential information). A comprehensive discussion of these techniques can be found in Adam and Wortmann (1989). These techniques can be broadly classified into two types, namely, access restriction techniques and perturbation techniques. As their name implies, access restriction techniques attempt to provide security by restricting access (to the number or type of queries issued) to the database. In order for these techniques to be successful, it is necessary to impose stringent restrictions both on the number and types of queries issued (Palley and Simonoff 1987). Imposing such restrictions may also make a large segment of the database unattainable to the legitimate user, thereby reducing the usefulness of the database. In addition, even if such stringent restrictions are imposed, there is no guarantee that a snooper, through *inferences*, will not be able to gain access to the *exact* value (*exact value disclosure*) or gain an accurate *estimate* of the value (*partial value disclosure*) of a numerical, confidential attribute (Adam and Wortmann 1989, Palley and Simonoff 1987). Hence, the applicability of access restriction techniques for preserving confidentiality of numerical data residing in large databases is limited.

Perturbation methods are a second class of techniques used to protect the confidential, numerical data in databases. In simple terms, perturbation involves replacing the values of the original, confidential attribute by a new set of perturbed values. Users are provided complete access to the perturbed attributes (and no access to the original attributes). Since every value in the database is "perturbed" by random noise, perturbation methods guarantee that exact disclosure will not occur. However, perturbation methods may result in partial disclosure (Adam and Wortmann 1989, Muralidhar, et. al. 1999, Muralidhar and Sarathy 1999). In addition, when perturbation methods are employed to preserve confidentiality, it is possible that the response to a given query using the perturbed data may be different from the response using the original data. Muralidhar et al. (1999) provided a new method of perturbation, called the General Additive Data Perturbation (GADP) method, that is capable of eliminating all types of bias, *if the continuous, numerical attributes in the database can be described by a multivariate normal distribution*. In this study, we identify and illustrate the problems associated with using the GADP method for perturbing non-normal data, namely, the inability maintain the marginal distribution of the perturbed attributes to be the same before and after perturbation. We also identify a new methodology for overcoming these problems.

General Additive Data Perturbation Method

Consider a database consisting of a set of L continuous, numerical, confidential attributes **X**. Assume that the database also consists of M non-confidential attributes **S**. These attributes are considered non-confidential since information regarding these

attributes can be accessed freely either through the database or through other sources. In addition, it is assumed that the non-confidential attributes are either numerical or are meaningful categorical attributes that can be converted to numerical form. Let $\mathbf{\Sigma}_{XX}$ represent the covariance matrix of **X**, let $\mathbf{\Sigma}_{SS}$ represent the covariance matrix of **S**, and let $\mathbf{\Sigma}_{XS}$ represent the covariance between **X** and **S**. In simple terms, perturbation is essentially replacing the values in **X** by a new set of (perturbed) values **Y**. Users may be provided complete access to **Y** but denied all access to **X**. As with any other disclosure limitation technique, perturbation has two major (conflicting) objectives, namely, data utility and disclosure risk.

In the context of perturbation data utility can be defined as follows. A legitimate user will be provided access to the perturbed attributes (**Y**) in place of the original attributes (**X**). Hence, it is desirable that the response to *any* query using **Y** is the same as that using **X**. If the responses to queries using **Y** are different from those using **X**, it results in perturbation bias. Maximizing data utility implies that perturbation bias must be eliminated. Disclosure risk deals with the ability of a snooper to infer information regarding a confidential attribute. Minimizing disclosure risk implies that providing access to **Y** should provide the snooper with *no additional information* regarding **X**.

The GADP method proposed by Muralidhar et al. (1999) represents the most recent and the most generalized form of additive data perturbation methods. Since this method was shown to be the general form of additive data perturbation methods and all other methods were only special cases of this method, we will limit our discussion this method. The GADP considers the characteristics of the entire set of attributes, namely, the confidential attributes (**X**), the non-confidential attributes (**S**), and the perturbed attributes (**Y**). One critical assumption underlying the GADP method is that **X**, **S**, and **Y** have a joint multivariate normal distribution. Since any multivariate normal distribution is completely defined by the mean vector and the covariance matrix, the general data utility requirements can be written as:

1. The mean and variance of a given perturbed attribute Y_i should be the same as that of the corresponding confidential attribute X_i,
2. The covariance matrix of **Y** should be the same as that of **X**, and
3. The covariance matrix of (**Y** and **S**) should be the same as that of (**X** and **S**).

Further, in the case of the multivariate normal distribution, since the best predictor of a given attribute is a linear function of other attributes, the level of security provided can be evaluated by considering linear functions of the attributes. Muralidhar et al. (2000) also showed that by specifying:

$$\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{XS}\mathbf{\Sigma}_{SS}^{-1}\mathbf{\Sigma}_{SX} \quad (1)$$

the proportion of variability explained (R^2) in any linear combination of **X** and **S** is the same using (**Y** and **S**) as that using **S** alone. Using the above specifications, the condition distribution of (**Y** | **V**) where **V** = {**X**, **S**} can be written as:

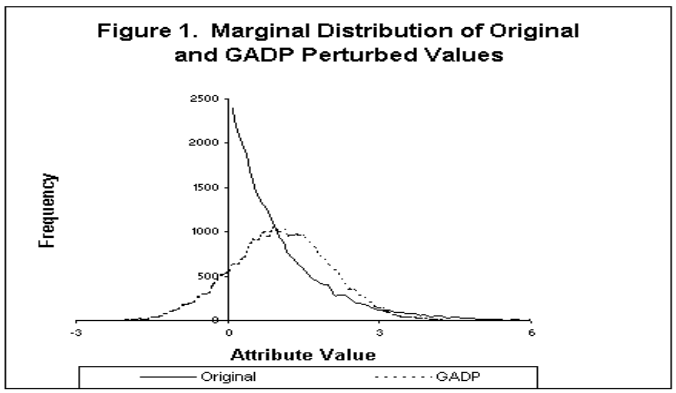
$$(\mathbf{Y} \mid \mathbf{V} = \mathbf{v}_i) \sim N(\mu_x + \mathbf{\Sigma}_{YV}\mathbf{\Sigma}_{VV}^{-1}(\mathbf{v}_i - \mu_v), \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YV}\mathbf{\Sigma}_{VV}^{-1}\mathbf{\Sigma}_{VY}) \quad (2)$$

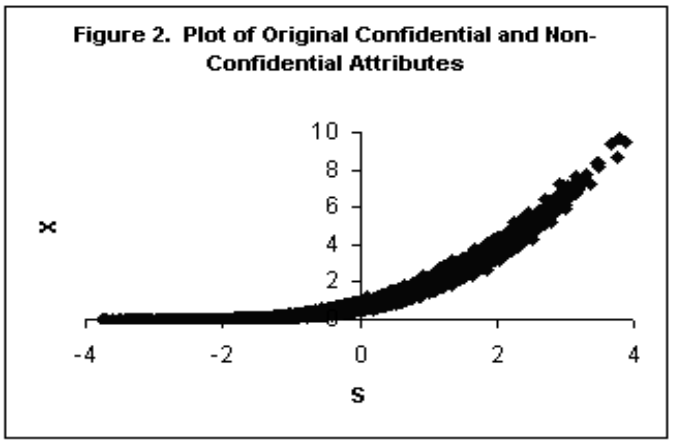
The GADP method simply generates the perturbed values using the conditional distribution function in (2) above. The collection of values generated has the same characteristics that were specified for **Y**. Thus, for the multivariate normal distribution, the GADP method provides maximum data utility (by eliminating bias) and minimum data disclosure (by maximizing security).

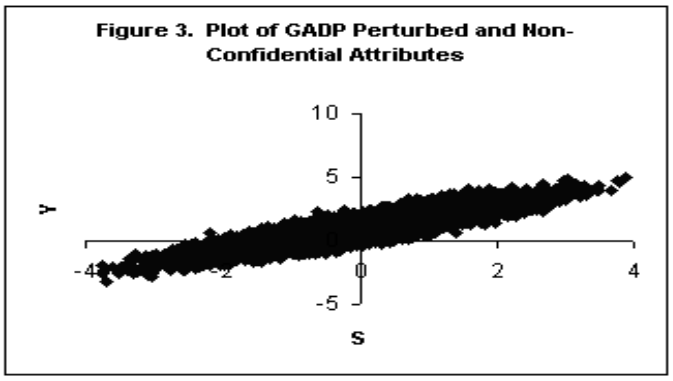
Issues in Perturbing Non-Normal Databases

As Muralidhar et al. (1999) themselves have identified, the GADP method is not suited for perturbing non-normal databases. To further illustrate this case, consider a database consisting of two attributes, one confidential (X_1) and one non-confidential (S_1). Let the marginal distribution of X_1 be exponential and let the marginal distribution of S_1 be normal. A database consisting of 25,000 observations was generated with the characteristics above. The GADP method was applied to this database. The frequency distribution of the values of the confidential and perturbed attributes is provided in Figure 1. Figure 1 clearly shows that the frequency distribution of the perturbed attribute is very different from that of the original attribute. In this case, using the perturbed values in place of the original values will result in bias in responding to queries such as percentiles, conditional sums, conditional means, etc.

Now consider the dependence between the attributes. The correlation between X_1 and S_1 (the original attributes) was 0.90. The correlation between Y_1 and S_1 is also 0.90. This is one of the strengths of the GADP procedure that guarantees that the product







moment correlation between the attributes will be maintained, *irrespective of the underlying distribution of the attributes*. This alone, however, does not guarantee that the dependence between the attributes is maintained.

Figure 2 provides a scatter plot of the original values of the confidential and non-confidential attributes. Figure 3 provides a scatter plot of the non-confidential and perturbed values of the confidential attribute. Comparing Figure 3 to Figure 2, it is clear that using the GADP method in this case has clearly distorted the relationship between the attributes. While the original relationship in Figure 2 is clearly non-linear, the relationship between the perturbed and non-confidential attribute (shown in Figure 3) is almost linear. This can be directly attributed to the fact that the GADP method focuses on product moment correlation. This focus is adequate and appropriate for the multivariate normal distribution, but as shown in the above example, may be neither adequate nor appropriate for non-normal distributions.

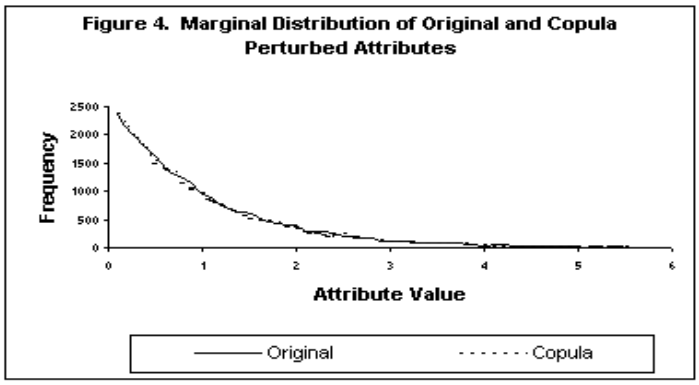
There are a variety of ways to measure and quantify dependence. Pearson's product moment correlation is probably the measure that is most often used to measure dependence or relationships between variables. The attractiveness of the product moment correlation measure stems from the fact that most statistical analysis is performed for linear models. In addition, when variables have a multivariate normal distribution, the relationship between the variables is linear (Kotz, Balakrishnan, and Kotz, 2000). Because of this characteristic, if the assumption of multivariate normality is satisfied, the product moment correlation provides the best estimate of the relationship between variables.

The product moment correlation, however, does not represent a universal measure of dependence. As Lancaster (1982) observes "the product moment correlation is the index of choice" for measuring dependence between variables that have a joint normal distribution, but "A general index of dependence, whereby joint distributions can be arranged in order of the degree of dependence, does not exist." Neter et al. (1996) also note that when two variables have a joint distribution that "differs considerably from the bivariate normal distribution" a non-parametric measure of association (such as Spearman's rank order correlation or Kendall's Tau) may be used for making inferences regarding the association between the two variables. The rank order correlation between the original values of the confidential and non-confidential attributes is 0.99 while that between the GADP perturbed values and the non-confidential attributes is only 0.89. In other words, while the GADP method is able to maintain product moment correlation, it does not maintain the rank order correlation between the attributes.

Irrespective of the measure of dependence used, for a data perturbation method to satisfy the data utility requirements, *all* relationships between *all* attributes must be the same before and after perturbation. The above illustration shows that the GADP method, which is the most advanced of existing methods of perturbation, does not provide this ability. Thus, there are currently no approaches that allows the DBA to perturb non-normal attributes such that (1) the marginal distribution of the perturbed attributes is the same as that of the original, confidential attributes, and (2) preserves the dependence between attributes. It is desirable to develop a perturbation method that provides the DBA with this capability.

A New Approach for Perturbing Non-Normal Databases

Copulas provide a method by which it is possible to perturb non-normal databases so as to eliminate the problems associated with the GADP method. The derivations of using the Copula approach have been deleted for brevity and can be obtained from the authors. The results of using the copula approach are summarized below.



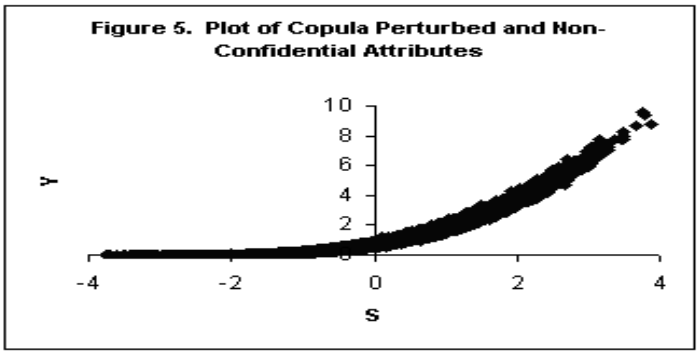


Figure 4 provides the marginal distribution of the confidential attributes before and after Copula perturbation. Figure 4 clearly indicates that the marginal distribution of the perturbed attributes almost exactly matches the marginal distribution of the original attributes. Figure 5 provides a scatter plot of the non-confidential attribute and the (copula) perturbed values. Comparing Figure 2 (scatter plot of the original values of the confidential and non-confidential attributes) and Figure 5, it is clear that the relationship between the attributes after perturbation is almost identical to the relationship before perturbation. The product moment and rank order correlation values between the copula perturbed attribute and the non-confidential attributes are 0.90 and 0.99, respectively. The product moment and rank order correlation of the original confidential and non-confidential attributes are also 0.90 and 0.99, respectively.

Experimentation with other simulated examples indicate that the results are similar to the results presented above. These results suggest that the copula method is capable of:

1. Maintaining the marginal distribution of the perturbed attributes to be the same as that of the original attributes, and
2. Maintaining the relationships between attributes in the perturbed database to be the same as that in the original database,

and provide strong evidence to suggest that the copula method of perturbation is suited for databases consisting of attributes that have non-normal distributions.

References

Adam, N.R. and J.C. Wortmann, "Security-Control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys* (21),1989, 515-556.

Kotz, S., N. Balakrishnan, and N.L. Johnson, *Continuous Multivariate Distributions: Volume 1: Models and Applications*, John Wiley & Sons, Inc., New York, 2000.

ancaster, H.O., "Dependence, Measures and Indices of," In S. Kotz and N.L. Johnson, *Encyclopedia of Statistical Sciences (Volume 2)*, John Wiley & Sons, New York, 1982.

Muralidhar, K., R. Parsa, and R. Sarathy, "A General Additive Data Perturbation Method for Database Security," *Management Science* (45), 1999, 1399-1415.

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*. Irwin, Chicago, 1996.

Palley, M.A. and J.S. Simonoff, "The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases," *ACM Transactions on Database Systems*, (12), 1987, 593-608.