MANAGEMENT SCIENCE
Vol. 56, No. 2, February 2010, pp. 318–333
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5602 | 0318

informs

DOI 10.1287/mnsc.1090.1104

Perturbation of Numerical Confidential Data via Skew-t Distributions

Seokho Lee
Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115,
seokhol@hsph.harvard.edu

Marc G. Genton
Department of Statistics, Texas A&M University, College Station, Texas 77843,
genton@stat.tamu.edu

Reinaldo B. Arellano-Valle
Departamento de Estadística, Facultad de Matemática, Pontificia Universidad Católica de Chile,
Santiago 22, Chile, reivalle@mat.puc.cl

We propose a new data perturbation method for numerical database security problems based on skew-t distributions. Unlike the normal distribution, the more general class of skew-t distributions is a flexible parametric multivariate family that can model skewness and heavy tails in the data. Because databases having a normal distribution are seldom encountered in practice, the newly proposed approach, coined the skew-t data perturbation (STDP) method, is of great interest for database managers. We also discuss how to preserve the sample mean vector and sample covariance matrix exactly for any data perturbation method. We investigate the performance of the STDP method by means of a Monte Carlo simulation study and compare it with other existing perturbation methods. Of particular importance is the ability of STDP to reproduce characteristics of the joint tails of the distribution in order for database users to answer higher-level questions. We apply the STDP method to a medical database related to breast cancer.

Key words: confidentiality; database management; kurtosis; multivariate; security; simulation; skewness
History: Received January 28, 2008; accepted September 24, 2009, by Ramayya Krishnan, information systems.
Published online in *Articles in Advance* December 14, 2009.

1. Introduction

This work is concerned with database privacy, which, according to Domingo-Ferrer (2007), can be classified into three independent categories: respondent, owner, and user privacy. We are mainly concerned with respondent privacy, that is, preventing reidentification of the respondents to which the records of a database correspond, when the database is made available to third parties. In recent years, new definitions of privacy such as k-anonymity and l-diversity have gained popularity (see, for instance, Machanavajjhala et al. 2007), although their limitations have also been recognized (see Barak et al. 2007).

The protection of confidential variables in governmental, commercial, and medical databases has become a very important problem. Database managers should provide as much relevant information contained in databases as possible to legitimate database users without revealing confidential contents. One approach to maintain confidentiality is to use data perturbation methods, that is, to construct a perturbed version of the database and to answer queries on the perturbed data. Specifically, it consists in perturbing, or masking, the original variables by replacing the confidential variables with new perturbed variables. Instead of releasing confidential variables, only perturbed variables are released to the public or to legitimate database users. To provide accurate information, also called utility, distributional properties of the perturbed variables should remain as close as possible to those of the original confidential variables. Moreover, to maintain confidentiality, data perturbation methods must minimize disclosure risk, that is, providing users with microdata access should not result in any additional information. Obviously, there is a trade-off between data utility and disclosure risk. Literature on database security and data perturbation methods is abundant; see Muralidhar and Sarathy (2003) for a recent theoretical basis and comprehensive overview. It is important to note that those data perturbation methods differ from multiple imputation procedures (see Rubin 1993) in that for the former the original data are assumed to be the population, whereas for the latter the original data are assumed to be a sample from an unknown population.

A variety of data perturbation methods have been developed for database security problems. Muralidhar et al. (1999, 2001) introduced the general additive

data perturbation (GADP) method. For databases that can be described by the multivariate normal distribution, they showed that the GADP method is superior to all previously proposed data perturbation methods. Thus, in that setting, GADP should be the preferred method of perturbation in terms of bias (the fact that the response to a query based on the perturbed data may be different from the response based on the original data) and of security (prevention of disclosure). In particular, GADP maintains the mean vector and covariance matrix asymptotically. However, it is obvious that GADP cannot be successfully applied to databases that cannot be described by the multivariate normal distribution. This is a serious problem because databases following the normal law are more often the exception rather than the rule. Indeed, it is common to have highly skewed or heavy-tailed variables, or both. Muralidhar et al. (1999) explained that the difficulty in constructing new data perturbation methods for nonnormal distributions lies in the lack of available multivariate distributions with simple characterization (other than the multivariate normal) and the lack of associated procedures for nonnormal random variates generation. As shown in the following, recent advances in the construction of parametric families of multivariate nonnormal distributions allow us to bypass this difficulty.

One attempt for addressing the issue of nonnormality was proposed by Muralidhar et al. (1995) using the log-normal distribution. Unfortunately, that method was limited to the univariate case (only one variable) and to a distribution that is too restrictive for many practical applications. Another more promising attempt is the copula-based GADP (C-GADP) proposed by Sarathy et al. (2002). It is based on the multivariate normal copula (see Clemen and Reilly 1999 for its use in decision and risk analysis) and on rank-based correlation to explain dependency between variables. C-GADP maintains the marginal distributions asymptotically and the rank order correlations. It satisfies the disclosure risk requirement but not the utility requirements in all situations. The implementation of C-GADP requires to identify the marginal distribution of confidential and nonconfidential variables. Unfortunately, the identification of marginal distributions is not an easy task, and misspecifying marginal distributions results in incorrect inferences because marginal distributions of the perturbed variables are the same as the assumed marginal distributions. In spite of such danger, C-GADP does not provide any guideline for choosing the marginal distributions, which could be practically unattractive to potential database managers. Another drawback of the C-GADP method is that it is based on the multivariate normal copula, that is, the dependency structure between variables is still of normal type. Hence the normal copula may not preserve tail dependence, a restriction noted by Sarathy et al. (2002). In principle, nonnormal copulas could be implemented in the C-GADP method, but their choice remains an open problem, possibly quite difficult.

Besides data perturbation methods, data swapping approaches can be applied to database security problems. Instead of modifying variables, data swapping methods swap confidential variables in a systematic fashion. From the literature on data swapping, however, it is well known that data swapping methods do not address data utility and disclosure risk simultaneously because they are not based on conditional distributions as in perturbation methods (for instance, GADP). Muralidhar and Sarathy (2006) developed a data shuffling method (DSP) based on conditional distributions, although it involves C-GADP (with its limitations described above) in one step of the procedure. It can be implemented by using only rank order data and therefore provides a nonparametric method for masking variables (Muralidhar and Sarathy 2006). DSP maintains the marginal distributions exactly, the rank order correlations, and monotonic relationships among the variables. Finally, because DSP is based on a conditional distribution approach, it provides a high level of security and does not suffer from the security issues of other data swapping procedures.

Recently, there has been a burgeoning of interest in multivariate skew-elliptical distributions, a flexible parametric family that can model skewness and heavy tails in the data; see Genton (2004) and Azzalini (2005) for an overview. A particular member of this family, the multivariate skew-t distribution, has been advocated by Azzalini and Genton (2008) as a general-purpose compromise between flexibility and simplicity. They have shown that the skew-t family can be used to model the distribution of many different types of data. Moreover, it includes the multivariate normal distribution as a special case. Because databases having a multivariate normal distribution are seldom encountered in practice, we expect that the skew-t family is an excellent candidate for solving database security problems. Our objective in this paper is to develop a new data perturbation method of numerical confidential variables for nonnormal databases that will reduce to the GADP method in case of a normal database.

The structure of this paper is as follows. In §2, we describe the skew-t distribution, its conditional distribution, and how to simulate from the latter. In §3, we recall the general requirements of the conditional distribution approach and consider the skew-t case in detail. We propose a skew-t data perturbation (STDP) method and describe its algorithm. We also discuss how to preserve the sample mean vector and sample covariance matrix exactly for any data

perturbation method. In §4, performances of our new STDP method are investigated and compared with the GADP, C-GADP, and DSP approaches via a Monte Carlo simulation study. In §5, the STDP and other existing methods are applied to a medical database. In §6, we describe some limitations of the proposed STDP method. In §7, we end with conclusions. Some technical results are provided in the appendix.

2. Skew-t Distribution

In this section, we present the definition and some properties of the skew-t distribution, in particular its conditional distribution with associated simulation procedure. The latter will be the key ingredient to introduce the STDP method.

2.1. Skew-t Distribution and Its Conditional Distribution

The multivariate skew-t distribution has been proposed as a simple yet flexible parametric nonnormal family for modeling skewness and heavy tails in data. The specific form that we shall consider is the one introduced by Branco and Dey (2001, 2002) and in an equivalent form by Azzalini and Capitanio (2003) and Arellano-Valle and Azzalini (2006). Specifically, a k-dimensional random vector $\mathbf{U}$ has a multivariate skew-t distribution with location vector $\boldsymbol{\xi}$, positive definite scale matrix $\boldsymbol{\Omega}$, shape (skewness) vector $\boldsymbol{\alpha}$, and degrees of freedom ν, if its density is

$$\begin{aligned} f_{\mathbf{U}}(\mathbf{u};\,\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\alpha},\nu) &= 2t_k(\mathbf{u}-\boldsymbol{\xi};\,\boldsymbol{\Omega},\nu) \\ &\quad\times T\left\{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\omega}^{-1}(\mathbf{u}-\boldsymbol{\xi})\left(\frac{\nu+k}{\nu+Q(\mathbf{u})}\right)^{1/2};\,\nu+k\right\}, \end{aligned} \tag{1}$$

where

$$t_k(\mathbf{u};\boldsymbol{\Omega},\nu)=\frac{\Gamma((\nu+k)/2)}{|\boldsymbol{\Omega}|^{1/2}(\pi\nu)^{k/2}\Gamma(\nu/2)}\left(1+\frac{Q(\mathbf{u})}{\nu}\right)^{-(\nu+k)/2}$$

denotes a k-dimensional Student's t density with location $\mathbf{0}$, scale matrix $\boldsymbol{\Omega}$, and degrees of freedom ν; $T\{\cdot;\,\nu+k\}$ denotes a univariate standard Student's t cumulative distribution function with $\nu+k$ degrees of freedom; $Q(\mathbf{u})=(\mathbf{u}-\boldsymbol{\xi})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\mathbf{u}-\boldsymbol{\xi})$, and $\boldsymbol{\omega}$ is the diagonal matrix formed by the square root of the diagonal elements of $\boldsymbol{\Omega}$. There is a single parameter ν to regulate the tail thickness of all components of t_k, hence also of $f_{\mathbf{U}}$. In that case, we use the notation $\mathbf{U}\sim \mathrm{ST}_k(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\alpha},\nu)$. The vector $\boldsymbol{\alpha}$ regulates skewness in the distribution with $\boldsymbol{\alpha}=\mathbf{0}$ resulting in symmetry, whereas ν controls the heaviness of the tails of the distribution. Three very interesting particular cases of the multivariate skew-t distribution are (i) the multivariate normal distribution $\mathrm{N}_k(\boldsymbol{\xi},\boldsymbol{\Omega})$ obtained when $\boldsymbol{\alpha}=\mathbf{0}$ and $\nu\to\infty$; (ii) the multivariate Student's t distribution obtained when $\boldsymbol{\alpha}=\mathbf{0}$; and (iii) the so-called multivariate skew-normal distribution obtained when $\nu\to\infty$ and denoted by $\mathrm{SN}_k(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\alpha})$; see, for instance, Azzalini (2005) for an overview.

The conditional distribution of the skew-t random vector is our prime interest for data perturbation problems. However, from the general theory of skew-elliptical distributions, the skew-t distribution is not closed under conditioning. This means that the conditional distribution of the skew-t distribution is not in the skew-t distribution family anymore, see (2) and (3) below. Fortunately, the conditional distribution of the skew-t distribution has a stochastic representation that can be exploited for random number generation. To illustrate this, suppose the k-dimensional skew-t random vector $\mathbf{U}$ is split into two subvectors, $\mathbf{U}_1$ and $\mathbf{U}_2$, with dimensions k_1 and k_2, respectively, $(k_1+k_2=k)$, that is, $\mathbf{U}=(\mathbf{U}_1^{\mathrm{T}},\mathbf{U}_2^{\mathrm{T}})^{\mathrm{T}}$. After some algebra described in the appendix, we obtain that the marginal distribution of $\mathbf{U}_2$ is $\mathbf{U}_2\sim \mathrm{ST}_{k_2}(\boldsymbol{\xi}_2,\boldsymbol{\Omega}_{22},\boldsymbol{\alpha}_{2(1)},\nu)$ and the conditional density of $\mathbf{U}_1$ given $\mathbf{U}_2=\mathbf{u}_2$ is

$$\begin{aligned} &f_{\mathbf{U}_1\mid\mathbf{U}_2=\mathbf{u}_2}(\mathbf{u}_1) \\ &= t_{k_1}(\mathbf{u}_1-\boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2);\,\boldsymbol{\Omega}_{11\cdot2}(\mathbf{u}_2),\nu_{1\cdot2}) \\ &\quad\times T\Big(\sqrt{(\nu_{1\cdot2}+k_1)/(\nu_{1\cdot2}+Q_{1\cdot2}(\mathbf{u}_1;\mathbf{u}_2))} \\ &\quad\cdot(\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}[\boldsymbol{\omega}_{1\cdot2}(\mathbf{u}_2)]^{-1}(\mathbf{u}_1-\boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2))+\tau_{1\cdot2}(\mathbf{u}_2));\,\nu_{1\cdot2}+k_1\Big) \\ &\quad/T\Big(\tau_{1\cdot2}(\mathbf{u}_2)/\sqrt{1+\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}};\,\nu_{1\cdot2}\Big), \end{aligned} \tag{2}$$

where

$$\nu_{1\cdot2}=\nu+k_2,\qquad \boldsymbol{\alpha}_{2(1)}=\frac{\boldsymbol{\alpha}_2+\bar{\boldsymbol{\Omega}}_{22}^{-1}\bar{\boldsymbol{\Omega}}_{21}\boldsymbol{\alpha}_1}{\sqrt{1+\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}}},$$

$$\begin{aligned} \boldsymbol{\alpha}_{1\cdot2} &= \boldsymbol{\omega}_{1\cdot2}\boldsymbol{\omega}_1^{-1}\boldsymbol{\alpha}_1, \\ \boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2) &= \boldsymbol{\xi}_1+\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}(\mathbf{u}_2-\boldsymbol{\xi}_2), \\ \boldsymbol{\Omega}_{11\cdot2}(\mathbf{u}_2) &= \left(\frac{\nu+Q(\mathbf{u}_2)}{\nu+k_2}\right)\boldsymbol{\Omega}_{11\cdot2}, \end{aligned}$$

with

$$\begin{aligned} \boldsymbol{\Omega}_{11\cdot2} &= \boldsymbol{\Omega}_{11}-\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}, \\ \boldsymbol{\omega}_{1\cdot2}(\mathbf{u}_2) &= \sqrt{\frac{\nu+Q(\mathbf{u}_2)}{\nu+k_2}}\boldsymbol{\omega}_{1\cdot2}, \\ Q_{1\cdot2}(\mathbf{u}_1;\mathbf{u}_2) &= (\mathbf{u}_1-\boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2))^{\mathrm{T}}[\boldsymbol{\Omega}_{11\cdot2}(\mathbf{u}_2)]^{-1}(\mathbf{u}_1-\boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2)), \end{aligned}$$

$$\begin{aligned} Q(\mathbf{u}_2) &= (\mathbf{u}_2-\boldsymbol{\xi}_2)^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-1}(\mathbf{u}_2-\boldsymbol{\xi}_2),\quad \text{and} \\ \tau_{1\cdot2}(\mathbf{u}_2) &= \sqrt{\frac{\nu+k_2}{\nu+Q(\mathbf{u}_2)}}\bar{\boldsymbol{\alpha}}_2^{\mathrm{T}}\boldsymbol{\omega}_2^{-1}(\mathbf{u}_2-\boldsymbol{\xi}_2), \\ \bar{\boldsymbol{\alpha}}_2 &= \boldsymbol{\alpha}_2+\bar{\boldsymbol{\Omega}}_{22}^{-1}\bar{\boldsymbol{\Omega}}_{21}\boldsymbol{\alpha}_1, \\ \bar{\boldsymbol{\Omega}} &= \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}. \end{aligned}$$

Here, $\boldsymbol{\omega}_i = \text{diag}(\boldsymbol{\Omega}_{ii})^{1/2}$, $\boldsymbol{\omega}_{1\cdot2} = \text{diag}(\boldsymbol{\Omega}_{11\cdot2})^{1/2}$, $\bar{\boldsymbol{\Omega}}_{ij} = \boldsymbol{\omega}_i^{-1}\boldsymbol{\Omega}_{ij}\boldsymbol{\omega}_j^{-1}$, and $\bar{\boldsymbol{\Omega}}_{11\cdot2} = \boldsymbol{\omega}_{1\cdot2}^{-1}\boldsymbol{\Omega}_{11\cdot2}\boldsymbol{\omega}_{1\cdot2}^{-1} = \boldsymbol{\omega}_{1\cdot2}^{-1}\boldsymbol{\omega}_1(\bar{\boldsymbol{\Omega}}_{11} - \bar{\boldsymbol{\Omega}}_{12}\bar{\boldsymbol{\Omega}}_{22}^{-1}\bar{\boldsymbol{\Omega}}_{21})\boldsymbol{\omega}_1\boldsymbol{\omega}_{1\cdot2}^{-1}$.

Although the exact distributional form (2) of the conditional skew-*t* density is difficult to use directly, we detail a procedure to simulate random numbers from that distribution in the next subsection. To this end, we note first that the conditional density (2) has the form of the so-called *extended skew-t* (EST) density (see Arellano-Valle and Genton 2010), which is given by

$$f_{\text{EST}}(\mathbf{u}) = t_k(\mathbf{u}-\boldsymbol{\xi};\ \boldsymbol{\Omega},\ \nu) \times \frac{T\left(\sqrt{(\nu+k)/(\nu+Q(\mathbf{u}))}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\omega}^{-1}(\mathbf{u}-\boldsymbol{\xi})+\tau);\ \nu+k\right)}{T\left(\tau/\sqrt{1+\boldsymbol{\alpha}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}\boldsymbol{\alpha}};\ \nu\right)}. \tag{3}$$

A model with this EST density shall be denoted by $\text{EST}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \tau, \nu)$. When $\tau = 0$, the EST density (3) reduces to the ST density (1), whereas when $\boldsymbol{\alpha} = \mathbf{0}$ it reduces to the symmetric density

$$f_{\text{ET}}(\mathbf{u}) = t_k(\mathbf{u}-\boldsymbol{\xi};\ \boldsymbol{\Omega},\ \nu)\frac{T(\sqrt{(\nu+k)/(\nu+Q(\mathbf{u}))}\tau;\ \nu+k)}{T(\tau;\ \nu)},$$

which is called an *extended Student's t* distribution and denoted by $\text{ET}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \tau, \nu)$. It corresponds to the distribution of a random vector $\mathbf{U}$ such that $\mathbf{U} \overset{d}{=} (\mathbf{T} \mid T_0 + \tau > 0)$, where $(T_0, \mathbf{T}^{\mathrm{T}})^{\mathrm{T}} \sim t_{1+k}(\mathbf{0}, \mathbf{I}_{1+d}, \nu)$.

2.2. Simulation From the Conditional of the Skew-*t* Distribution

We derive the stochastic representation of the conditional of the skew-*t* distribution, that is, of the EST distribution, which will be exploited in random number generation for data perturbation purpose. Let $\mathbf{U} \sim \text{ST}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu)$, and consider the partition $\mathbf{U} = (\mathbf{U}_1^{\mathrm{T}}, \mathbf{U}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are of dimensions k_1 and k_2, respectively, and the induced partition on $(\boldsymbol{\xi}, \boldsymbol{\Omega})$. Denote by $\mathbf{U}_{1\cdot2}$ the conditional random vector $\mathbf{U}_1$ given $\mathbf{U}_2$. As we show in the appendix, the conditional random vector $\mathbf{U}_{1\cdot2}$ has a convenient stochastic representation given by (7)–(9). Therefore, we can easily generate a random sample from the conditional distribution of $\mathbf{U}_1$ given $\mathbf{U}_2 = \mathbf{u}_2$ by means of the following steps:

(a) For a given $\mathbf{u}_2$, compute

$$\boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2) = \boldsymbol{\xi}_1 + \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}(\mathbf{u}_2 - \boldsymbol{\xi}_2),$$
$$Q(\mathbf{u}_2) = (\mathbf{u}_2 - \boldsymbol{\xi}_2)^{\mathrm{T}}\boldsymbol{\Omega}_{22}^{-1}(\mathbf{u}_2 - \boldsymbol{\xi}_2),$$
$$\tilde{\tau}_{1\cdot2}(\mathbf{u}_2) = \sqrt{\frac{\nu+k_2}{\nu+Q(\mathbf{u}_2)}}\boldsymbol{\alpha}_{2(1)}^{\mathrm{T}}\boldsymbol{\omega}_2^{-1}(\mathbf{u}_2 - \boldsymbol{\xi}_2),$$

where

$$\boldsymbol{\alpha}_{2(1)} = \frac{\boldsymbol{\alpha}_2 + \bar{\boldsymbol{\Omega}}_{22}^{-1}\bar{\boldsymbol{\Omega}}_{21}\boldsymbol{\alpha}_1}{\sqrt{1+\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}}}, \qquad \boldsymbol{\alpha}_{1\cdot2} = \boldsymbol{\omega}_{1\cdot2}\boldsymbol{\omega}_1^{-1}\boldsymbol{\alpha}_1.$$

(b) Generate t_0 from the first Student's t distribution in (9), for instance by means of the corresponding equations in (11) and (12). If $t_0 + \tilde{\tau}_{1\cdot2}(\mathbf{u}_2) > 0$, let $\tilde{t}_0 = t_0$ and go to the next step. Otherwise, discard t_0 and repeat (b) until $t_0 + \tilde{\tau}_{1\cdot2}(\mathbf{u}_2) > 0$.

(c) Generate $\mathbf{t}_1$ from the second Student's t distribution in (9), for instance, by means of the corresponding equations in (11) and (12). Take $\mathbf{t}_{1\cdot2}$ as

$$\sqrt{\frac{\nu_{1\cdot2} + \tilde{t}_0^2}{\nu_{1\cdot2}+1}}\mathbf{t}_1 + \boldsymbol{\delta}_{1\cdot2}\tilde{t}_0,$$

where

$$\nu_{1\cdot2} = \nu + k_2 \quad \text{and} \quad \boldsymbol{\delta}_{1\cdot2} = \frac{\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}}{\sqrt{1+\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}}},$$

and then compute

$$\mathbf{u}_{1\cdot2} = \boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2) + \sqrt{\frac{\nu+Q(\mathbf{u}_2)}{\nu+k_2}}\boldsymbol{\omega}_{1\cdot2}\mathbf{t}_{1\cdot2}.$$

The vector $\mathbf{u}_{1\cdot2}$ is a realization from the conditional distribution of $\mathbf{U}_1$ given $\mathbf{U}_2 = \mathbf{u}_2$.

3. Skew-*t* Data Perturbation Method

3.1. Conditional Distribution Approach

The conditional distribution approach is a general additive data perturbation method developed by Muralidhar et al. (1999) and further discussed by Muralidhar and Sarathy (2003), who formulated a theoretical basis for this technique. They suggested data utility and disclosure risk requirements for successful data perturbation methods.

Specifically, let $(\mathbf{X}^{\mathrm{T}}, \mathbf{S}^{\mathrm{T}})^{\mathrm{T}}$ be a $(k_{\mathbf{X}} + k_{\mathbf{S}})$-dimensional numerical vector, where $\mathbf{X}$ represents a $k_{\mathbf{X}}$-dimensional confidential data vector and $\mathbf{S}$ denotes a $k_{\mathbf{S}}$-dimensional nonconfidential data vector. Let $\mathbf{Y}$ denote the perturbed version of the confidential vector $\mathbf{X}$. Muralidhar and Sarathy (2003) precisely specified the theoretical basis for statistical perturbation methods as the following:

1. *Data utility or accuracy requirements*: The statistical characteristics of $\mathbf{Y}$ are the same as those of $\mathbf{X}$ (i.e., the same marginal densities $f_{\mathbf{X}} \equiv f_{\mathbf{Y}}$), and the relationship between $\mathbf{Y}$ and $\mathbf{S}$ is the same as that between $\mathbf{X}$ and $\mathbf{S}$ (i.e., the same joint densities $f_{\mathbf{X},\mathbf{S}} \equiv f_{\mathbf{Y},\mathbf{S}}$).

2. *Disclosure risk requirement*: The confidentiality of $\mathbf{X}$ is maintained and the released microdata $(\mathbf{Y}, \mathbf{S})$ does not increase disclosure risk (i.e., the same conditional densities $f_{\mathbf{X}|\mathbf{Y},\mathbf{S}} \equiv f_{\mathbf{X}|\mathbf{S}}$).

The disclosure risk requirement implies that the database users already know the conditional distribution of $\mathbf{X}$ given $\mathbf{S}$, otherwise releasing $\mathbf{Y}$ provides information about $\mathbf{X}$. Muralidhar and Sarathy (2003) suggested a general procedure for creating perturbed data values that satisfy the aforementioned data utility and disclosure risk requirements: generate observations from the conditional distribution of $\mathbf{X}$ given $\mathbf{S}$, such that given $\mathbf{S}$, $\mathbf{Y}$ is independent of $\mathbf{X}$.

It is important to realize that, unlike the Gaussian case, assuming a joint skew-t distribution for the whole vector $(\mathbf{X}^{\mathrm{T}}, \mathbf{Y}^{\mathrm{T}}, \mathbf{S}^{\mathrm{T}})^{\mathrm{T}}$ is unnecessarily restrictive for data perturbation purpose. Effectively, in that case the disclosure risk requirement cannot be satisfied and only weaker conditions of the form $\mathrm{E}(\mathbf{X} \mid \mathbf{Y}, \mathbf{S}) = \mathrm{E}(\mathbf{X} \mid \mathbf{S})$ or $\boldsymbol{\xi}_{\mathbf{X}|\mathbf{Y},\mathbf{S}}(\mathbf{Y}, \mathbf{S}) = \boldsymbol{\xi}_{\mathbf{X}|\mathbf{S}}(\mathbf{S})$ can be achieved.

3.2. STDP Algorithm

In this section, we propose a new data perturbation method, coined skew-t data perturbation (STDP). We expect that a better perturbation method can be obtained when we impose a flexible distribution to reflect the distributional properties of the observed variables $\mathbf{X}$ and $\mathbf{S}$ very well. We also need that the assumed distribution gives a conditional distribution of $\mathbf{X}$ given $\mathbf{S}$, which can be easily exploited in random number generation. Muralidhar et al. (1999) applied the multivariate normal distribution to perturbation methods that led to the GADP method. However, the normal distribution cannot characterize the properties of data very well because many real data sets can be skewed and heavy tailed. We propose to use the skew-t distribution for the joint distribution of $\mathbf{X}$ and $\mathbf{S}$ and exploit its flexibility and simplicity.

In the STDP approach, we assume a joint skew-t distribution for the confidential and nonconfidential variables $\mathbf{V} = (\mathbf{X}^{\mathrm{T}}, \mathbf{S}^{\mathrm{T}})^{\mathrm{T}}$ and estimate its parameters $\boldsymbol{\xi}_{\mathbf{V}}$, $\boldsymbol{\Omega}_{\mathbf{VV}}$, $\boldsymbol{\alpha}_{\mathbf{V}}$, and ν. Then we use them to specify the conditional distribution of $\mathbf{X}$ given $\mathbf{S}$. By means of the procedure described in §2.2, we generate realizations of random vectors $\mathbf{Y}$ with this conditional distribution. Based on the previous arguments, the algorithm of data generation for the STDP method is the following:

Step 1. Estimate the skew-t parameters $(\boldsymbol{\xi}_{\mathbf{V}}, \boldsymbol{\Omega}_{\mathbf{VV}}, \boldsymbol{\alpha}_{\mathbf{V}}, \nu)$ from the observed data $\mathbf{v}_i^{\mathrm{T}} = (\mathbf{x}_i^{\mathrm{T}}, \mathbf{s}_i^{\mathrm{T}})$, $i = 1, \ldots, n$, of $\mathbf{V}^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}, \mathbf{S}^{\mathrm{T}})$.

Step 2. For each $\mathbf{v}_i$, $i = 1, \ldots, n$:

(a) Calculate:

(i) $\boldsymbol{\xi}_{\mathbf{X}\cdot\mathbf{S}}(\mathbf{s}_i) = \boldsymbol{\xi}_{\mathbf{X}} + \boldsymbol{\Omega}_{\mathbf{XS}}\boldsymbol{\Omega}_{\mathbf{SS}}^{-1}(\mathbf{s}_i - \boldsymbol{\xi}_{\mathbf{S}})$;

(ii) $Q(\mathbf{s}_i) = (\mathbf{s}_i - \boldsymbol{\xi}_{\mathbf{S}})^{\mathrm{T}}\boldsymbol{\Omega}_{\mathbf{SS}}^{-1}(\mathbf{s}_i - \boldsymbol{\xi}_{\mathbf{S}})$;

(iii) $\boldsymbol{\Omega}_{\mathbf{XX}\cdot\mathbf{S}} = \boldsymbol{\Omega}_{\mathbf{XX}} - \boldsymbol{\Omega}_{\mathbf{XS}}\boldsymbol{\Omega}_{\mathbf{SS}}^{-1}\boldsymbol{\Omega}_{\mathbf{SX}}$,

$\boldsymbol{\omega}_{\mathbf{X}\cdot\mathbf{S}} = \mathrm{diag}(\boldsymbol{\Omega}_{\mathbf{XX}\cdot\mathbf{S}})^{1/2}$,

$\bar{\boldsymbol{\Omega}}_{\mathbf{XX}\cdot\mathbf{S}} = \boldsymbol{\omega}_{\mathbf{X}\cdot\mathbf{S}}^{-1}\boldsymbol{\Omega}_{\mathbf{XX}\cdot\mathbf{S}}\boldsymbol{\omega}_{\mathbf{X}\cdot\mathbf{S}}^{-1}$;

(iv) $\boldsymbol{\alpha}_{\mathbf{X}\cdot\mathbf{S}} = \boldsymbol{\omega}_{\mathbf{X}\cdot\mathbf{S}}\boldsymbol{\omega}_{\mathbf{X}}^{-1}\boldsymbol{\alpha}_{\mathbf{X}}$,

$$\boldsymbol{\alpha}_{\mathbf{S}(\mathbf{X})} = \frac{\boldsymbol{\alpha}_{\mathbf{S}} + \bar{\boldsymbol{\Omega}}_{\mathbf{SS}}^{-1}\bar{\boldsymbol{\Omega}}_{\mathbf{SX}}\boldsymbol{\alpha}_{\mathbf{X}}}{\sqrt{1 + \boldsymbol{\alpha}_{\mathbf{X}\cdot\mathbf{S}}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{\mathbf{XX}\cdot\mathbf{S}}\boldsymbol{\alpha}_{\mathbf{X}\cdot\mathbf{S}}}},$$

$$\boldsymbol{\delta}_{\mathbf{X}\cdot\mathbf{S}} = \frac{\bar{\boldsymbol{\Omega}}_{\mathbf{XX}\cdot\mathbf{S}}\boldsymbol{\alpha}_{\mathbf{X}\cdot\mathbf{S}}}{\sqrt{1 + \boldsymbol{\alpha}_{\mathbf{X}\cdot\mathbf{S}}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{\mathbf{XX}\cdot\mathbf{S}}\boldsymbol{\alpha}_{\mathbf{X}\cdot\mathbf{S}}}};$$

(v) $\tilde{\tau}_{\mathbf{X}\cdot\mathbf{S}}(\mathbf{s}_i) = \sqrt{\dfrac{\nu + k_{\mathbf{S}}}{\nu + Q(\mathbf{s}_i)}}\,\boldsymbol{\alpha}_{\mathbf{S}(\mathbf{X})}^{\mathrm{T}}\boldsymbol{\omega}_{\mathbf{S}}^{-1}(\mathbf{s}_i - \boldsymbol{\xi}_{\mathbf{S}})$.

(b) Draw v_0 from $\chi^2_{\nu+k_{\mathbf{S}}}$, z_0 from $\mathrm{N}(0, 1)$, and calculate $t_0 = \sqrt{(\nu + k_{\mathbf{S}})/v_0}\,z_0$. If $t_0 + \tilde{\tau}_{\mathbf{X}\cdot\mathbf{S}}(\mathbf{s}_i) > 0$, set $\tilde{t}_0 = t_0$; otherwise repeat (b) until $t_0 + \tilde{\tau}_{\mathbf{X}\cdot\mathbf{S}}(\mathbf{s}_i) > 0$ is satisfied.

(c) Draw v_1 from $\chi^2_{\nu+k_{\mathbf{S}}+1}$, $\mathbf{z}_1$ from $\mathrm{N}_{k_{\mathbf{S}}}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{\mathbf{XX}\cdot\mathbf{S}} - \boldsymbol{\delta}_{\mathbf{X}\cdot\mathbf{S}}\boldsymbol{\delta}_{\mathbf{X}\cdot\mathbf{S}}^{\mathrm{T}})$, and calculate $\mathbf{t}_1 = \sqrt{(\nu + k_{\mathbf{S}} + 1)/v_1}\,\mathbf{z}_1$.

(d) Calculate

$$\mathbf{t}_{\mathbf{X}\cdot\mathbf{S}} = \sqrt{\frac{\nu + k_{\mathbf{S}} + \tilde{t}_0^2}{\nu + k_{\mathbf{S}} + 1}}\,\mathbf{t}_1 + \boldsymbol{\delta}_{\mathbf{X}\cdot\mathbf{S}}\tilde{t}_0$$

and set

$$\mathbf{y}_i = \boldsymbol{\xi}_{\mathbf{X}\cdot\mathbf{S}}(\mathbf{s}_i) + \sqrt{\frac{\nu + Q(\mathbf{s}_i)}{\nu + k_{\mathbf{S}}}}\,\boldsymbol{\omega}_{\mathbf{X}\cdot\mathbf{S}}\mathbf{t}_{\mathbf{X}\cdot\mathbf{S}}.$$

Step 3. Report $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ as the perturbed variables of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

To estimate the skew-t parameters from the data in Step 1, we can use the R (R Development Core Team 2008) statistical package, named sn.[1] Note that GADP is a particular case of the STDP method as the multivariate skew-t distribution includes the normal distribution.

Because the conditional distribution approach assumes that the database user already knows the conditional distribution of $\mathbf{X}$ given $\mathbf{S}$, we release this information along with the fitted parameters to the user. In addition, we suggest to provide two measures of goodness-of-fit of the skew-t distribution to the original data. The first one is simply a p-value for the likelihood ratio test (LRT) that the underlying distribution is actually normal, a special case of the skew-t distribution. Typically, if that p-value is smaller than 5%, the hypothesis of normality is rejected. Then the smaller the p-value, the stronger the evidence against normality, which suggests moving away from GADP. The second measure is a graphical representation due

[1] This package, which is developed and maintained by Adelchi Azzalini, can be freely downloaded from the main webpage of the R project or at http://azzalini.stat.unipd.it/SN/.

to Healy (1968) (see also Azzalini and Capitanio 2003). It compares the fits of multivariate normal and skew-t distributions. The probabilities of a quadratic form in the data vector based on the assumed distribution and on the sample distribution are plotted. Thus, when the data are similar to the assumed distribution, the points in the plots should be located on a straight diagonal line. Those two measures are secure in the sense that they cannot be used to disclose confidential information.

3.3. Preserving Sample Information

Confidential data perturbation methods, for instance, such as GADP, are generally able to preserve the mean vector and the covariance matrix asymptotically, i.e., in large samples. Burridge (2003) proposed a method similar to GADP, coined information preserving statistical obfuscation, that can preserve the mean vector and the covariance matrix exactly in the sample. This implies that the sample mean vector and the sample covariance matrix of the perturbed data are exactly the same as those of the original confidential data. Recently, Muralidhar and Sarathy (2008) have proposed an extension of the aforementioned method that generates nonsynthetic perturbed data while preserving mean and covariance exactly. A proximity parameter between the original and the perturbed data must be selected, and this makes that approach somewhat ad hoc.

We believe that preserving the sample mean vector and the sample covariance matrix exactly can be achieved for any data perturbation method, for example, such as C-GADP or STDP, and it is not restricted to GADP based on the multivariate normal distribution. The method is very simple and we describe it next. Note also that an important consequence is that any method of multivariate data analysis based on the sample mean vector and the sample covariance matrix will yield the same results on the perturbed data as on the original confidential data. Such methods include principal component analysis and canonical correlation analysis, among many other techniques.

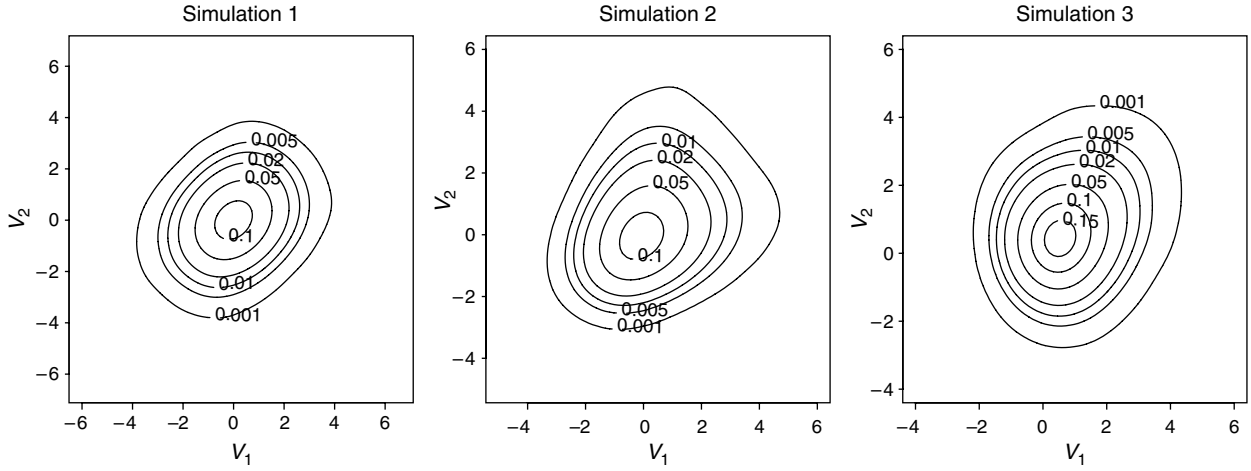
Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sample of confidential variables with sample mean vector $\hat{\boldsymbol{\mu}}_{\mathbf{x}}$ and sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. Suppose that some data perturbation method has produce a sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ with sample mean vector $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$ and sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}$. Define $\mathbf{y}_i = \hat{\boldsymbol{\mu}}_{\mathbf{x}} + \hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{1/2}\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}^{-1/2}(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\mathbf{z}})$ for $i = 1, \ldots, n$. Then, straightforward algebra shows that $\hat{\boldsymbol{\mu}}_{\mathbf{y}} = \hat{\boldsymbol{\mu}}_{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. Because $\mathbf{y}$ is obtained through a linear transformation of $\mathbf{z}$, many equivariant estimators such as sample skewness and kurtosis will be unaffected. This means that they will remain the same on the $\mathbf{y}$ data and the $\mathbf{z}$ data, but of course not on the $\mathbf{x}$ data in general. Also, the distribution of the $\mathbf{y}$ data will remain the same as the distribution of the $\mathbf{z}$ data as long as that distribution is closed under linear transformations. For example, this is the case for the multivariate normal, elliptical (Fang et al. 1990), and skew-elliptical distributions, among others. Sample quantiles will be unaffected asymptotically because in that case, the linear transformation is the identity operation.

4. Performance of STDP

In this section, we present Monte Carlo simulation results to compare the performance of the STDP method with existing methods when they are applied to nonnormal databases. We consider three different situations: the first two scenarios (simulations 1 and 2) consider the multivariate **g**-and-**h** distribution (Field and Genton 2006). This is a different kind of multivariate distribution than the skew-t. We can control the skewness and heaviness of its tails, but it is not in the skew-elliptical distribution class. The detail of the simulation setup is given in the next subsection. The reason for considering the **g**-and-**h** distribution in this simulation is that we can show how well STDP performs compared to other perturbation methods in the case where the actual data distribution is not the skew-t distribution under which the STDP method was constructed. The third scenario (simulation 3) uses the multivariate skew-t distribution and is expected to show better performance of STDP over other perturbation methods.

We consider the STDP method and its algorithm described in §3. For the C-GADP method, we use the empirical distribution as the marginal distribution because there is no general guide for selecting a parametric marginal distribution. Besides GADP, C-GADP, and DSP, we also include in the comparison an interesting variant of the C-GADP method, denoted by C-GADPST. It uses the univariate skew-t distribution as the marginal distribution in its construction. Throughout the simulation study, we apply the sample mean and covariance preserving (MCP) technique described in §3.3 on all perturbation methods, except DSP. The reason we did not apply MCP to DSP is that MCP would destroy the data shuffling mechanism, which is a key factor in DSP. We use the postfix "MCP" (mean-covariance preserving) to distinguish these perturbation techniques from their original versions. In the following sections, we present the results for the MCP versions only, because we observed that the original versions performed similarly to the MCP versions (except for preserving mean and covariance). In addition, we also report the results of C-GADP without MCP because of its ability to preserve the marginal distributions.

Figure 1 **Contour Plots of the First Two (Confidential) Variables of Four-Dimensional g-and-h (Simulations 1 and 2) and Skew-*t* (Simulation 3) Distributions Used in the Simulation Study**



4.1. Simulation Setting

In the simulation, we consider four-dimensional random vectors $\mathbf{V} = (\mathbf{X}^{\mathrm{T}}, \mathbf{S}^{\mathrm{T}})^{\mathrm{T}}$, with two-dimensional vectors $\mathbf{X}$ and $\mathbf{S}$, from the aforementioned two distributions. The **g**-and-**h** distributed random vector is defined by $\mathbf{V} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are arbitrary location vector and covariance matrix, respectively, and $\mathbf{Z} \sim \mathrm{N}_4(\mathbf{0}, \mathbf{I}_4)$, a standard multivariate normal distribution. The vector-valued function $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ is defined by $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) = (\tau_{g_1,h_1}(Z_1), \tau_{g_2,h_2}(Z_2), \tau_{g_3,h_3}(Z_3), \tau_{g_4,h_4}(Z_4))^{\mathrm{T}}$, with

$$\tau_{g,h}(Z) = \left(\frac{\exp(gZ)-1}{g}\right)\exp\left(\frac{h}{2}Z^2\right), \quad h > 0,$$

where $\mathbf{g} = (g_1, g_2, g_3, g_4)^{\mathrm{T}}$ controls the skewness and $\mathbf{h} = (h_1, h_2, h_3, h_4)^{\mathrm{T}}$ controls the kurtosis of each variable separately. In this simulation we set $\boldsymbol{\mu} = (0, 0, 0, 0)^{\mathrm{T}}$ and the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.3 & 0.5 & 0.2 \\ 0.3 & 1.0 & 0.7 & 0.5 \\ 0.5 & 0.7 & 1.0 & 0.1 \\ 0.2 & 0.5 & 0.1 & 1.0 \end{bmatrix}. \tag{4}$$

We use $\mathbf{g} = (0, 0, 0, 0)^{\mathrm{T}}$ and $\mathbf{h} = (0.05, 0.05, 0.05, 0.05)^{\mathrm{T}}$ for simulation 1, that is, there is no skewness at all and the same amount of kurtosis in all directions before introducing scale. The resulting distribution is somewhat similar to an elliptical shape. Next, $\mathbf{g} = (0.2, 0.3, 0.2, 0.1)^{\mathrm{T}}$ and $\mathbf{h} = (0.05, 0.03, 0.02, 0.01)^{\mathrm{T}}$ are used for simulation 2, which leads to a shape far from elliptical and with different tail behavior in each direction. For simulation 3, we use the skew-*t* distribution with location $\boldsymbol{\xi} = (0, 0, 0, 0)^{\mathrm{T}}$, shape $\boldsymbol{\alpha} = (1, 2, 3, 1)^{\mathrm{T}}$, degrees of freedom $\nu = 9$, and scale matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}$ as defined in (4). Because of space limitation, we only present contour plots of the first two (confidential) variables of the distributions used in the simulation study in Figure 1.

We generate artificial databases of multiple sample sizes ($n = 100$, 500, and 1,000 observations) from the above **g**-and-**h** and skew-*t* distributions. We regard the first two variables ($\mathbf{X}$) as confidential variables that will be perturbed by the perturbation methods and the remaining two variables ($\mathbf{S}$) as nonconfidential variables that will remain the same after perturbation. To investigate the performance of the perturbation methods, we calculate the first four sample moments (mean, covariance, skewness, and kurtosis), rank order correlation, and some quantiles, of the perturbed variables. The error of those estimators is assessed by computing the average and standard deviation of biases over 1,000 simulated data sets. In this simulation study, we assume that each simulated microdata itself is a finite population. Therefore, the perturbation methods are compared through assessing how well they replicate the characteristics of a known population represented by the simulated data set.

4.2. Moments

As we described in §3.3, all perturbation methods with sample information preserving technique (MCP) lead to zero biases in the first two moments (mean and covariance). DSP also preserves exactly the mean and variance, but the covariance will only be maintained asymptotically. After perturbation over 1,000 simulated data sets, we calculate the average biases (AB) and standard deviations (SD) of sample rank order correlations. In Table 1, we report the results for simulation 2, the most nonnormal setting, only (those for simulations 1 and 3 were qualitatively similar) for sample sizes $n = 100$, 500, and 1,000. All average biases are small as expected, because the sample covariance matrix is maintained by the MCP technique. The DSP

Table 1 Rank Order Correlations for the Perturbed Data Sets in Simulation 2

Simulation 2	n	V_1, V_2	V_1, V_3	V_1, V_4	V_2, V_3	V_2, V_4	V_3, V_4
STDP	100	−0.0120 (0.0932)	−0.0178 (0.0834)	−0.0006 (0.0910)	−0.0095 (0.0478)	−0.0054 (0.0550)	0.0000 (0.0000)
	500	−0.0134 (0.0405)	−0.0168 (0.0364)	−0.0034 (0.0393)	−0.0115 (0.0213)	−0.0031 (0.0252)	0.0000 (0.0000)
	1,000	−0.0138 (0.0287)	−0.0174 (0.0260)	−0.0022 (0.0286)	−0.0111 (0.0152)	−0.0023 (0.0174)	0.0000 (0.0000)
STDP-MCP	100	−0.0229 (0.0582)	−0.0202 (0.0537)	−0.0045 (0.0501)	−0.0165 (0.0355)	−0.0022 (0.0375)	0.0002 (0.0078)
	500	−0.0218 (0.0268)	−0.0190 (0.0242)	−0.0040 (0.0226)	−0.0180 (0.0161)	−0.0006 (0.0181)	0.0000 (0.0013)
	1,000	−0.0208 (0.0180)	−0.0187 (0.0176)	−0.0042 (0.0156)	−0.0180 (0.0115)	0.0008 (0.0128)	0.0000 (0.0007)
GADP-MCP	100	−0.0262 (0.0527)	−0.0209 (0.0527)	−0.0062 (0.0472)	−0.0197 (0.0355)	−0.0050 (0.0357)	−0.0001 (0.0068)
	500	−0.0279 (0.0227)	−0.0213 (0.0220)	−0.0067 (0.0216)	−0.0207 (0.0148)	−0.0076 (0.0163)	−0.0001 (0.0011)
	1,000	−0.0283 (0.0163)	−0.0220 (0.0168)	−0.0055 (0.0148)	−0.0210 (0.0107)	−0.0067 (0.0114)	0.0000 (0.0005)
C-GADP	100	0.0215 (0.0786)	0.0006 (0.0715)	0.0161 (0.0810)	−0.0047 (0.0427)	0.0018 (0.0505)	0.0000 (0.0000)
	500	0.0035 (0.0352)	−0.0009 (0.0325)	0.0023 (0.0365)	−0.0006 (0.0185)	−0.0008 (0.0229)	0.0000 (0.0000)
	1,000	0.0019 (0.0260)	−0.0006 (0.0228)	0.0012 (0.0265)	−0.0002 (0.0130)	−0.0003 (0.0163)	0.0000 (0.0000)
C-GADP-MCP	100	−0.0858 (0.0695)	−0.0230 (0.0562)	−0.0297 (0.0565)	0.0039 (0.0389)	−0.0071 (0.0388)	0.0002 (0.0089)
	500	−0.0561 (0.0306)	−0.0136 (0.0252)	−0.0184 (0.0245)	−0.0012 (0.0174)	−0.0049 (0.0181)	0.0003 (0.0019)
	1,000	−0.0437 (0.0211)	−0.0130 (0.0185)	−0.0137 (0.0169)	−0.0023 (0.0131)	−0.0028 (0.0132)	0.0003 (0.0009)
C-GADPST-MCP	100	−0.0140 (0.0584)	−0.0077 (0.0561)	0.0001 (0.0531)	−0.0029 (0.0360)	0.0016 (0.0388)	0.0007 (0.0084)
	500	−0.0160 (0.0271)	−0.0091 (0.0256)	−0.0031 (0.0229)	−0.0076 (0.0167)	−0.0024 (0.0172)	0.0001 (0.0014)
	1,000	−0.0164 (0.0189)	−0.0096 (0.0181)	−0.0032 (0.0158)	−0.0069 (0.0122)	−0.0022 (0.0124)	0.0001 (0.0007)
DSP	100	−0.0017 (0.0812)	−0.0041 (0.0735)	0.0017 (0.0853)	−0.0021 (0.0422)	−0.0034 (0.0534)	0.0000 (0.0000)
	500	−0.0023 (0.0352)	−0.0017 (0.0310)	−0.0001 (0.0362)	−0.0005 (0.0184)	−0.0004 (0.0222)	0.0000 (0.0000)
	1,000	−0.0018 (0.0260)	−0.0002 (0.0229)	−0.0007 (0.0259)	−0.0002 (0.0131)	−0.0017 (0.0160)	0.0000 (0.0000)

Note. AB with SD in parentheses.

approach also yields small average biases of the rank order correlations. In addition, we included the results of STDP without MCP to see the effect of mean and covariance preserving on the rank order correlations. As can be seen, this effect is fairly small.

Next, we consider higher-order moments (skewness and kurtosis) from 1,000 pairs of the simulated and perturbed data sets under the three scenarios. We use the measures of multivariate skewness $b_{1,k}$ and kurtosis $b_{2,k}$ defined by Mardia et al. (1979, p. 21) as

$$b_{1,k} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\left\{(\mathbf{u}_r - \hat{\boldsymbol{\mu}}_{\mathbf{u}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{u}}^{-1}(\mathbf{u}_s - \hat{\boldsymbol{\mu}}_{\mathbf{u}})\right\}^3,$$

$$b_{2,k} = \frac{1}{n}\sum_{r=1}^{n}\left\{(\mathbf{u}_r - \hat{\boldsymbol{\mu}}_{\mathbf{u}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{u}}^{-1}(\mathbf{u}_r - \hat{\boldsymbol{\mu}}_{\mathbf{u}})\right\}^2,$$

where k denotes the dimension of the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ under consideration, and $\hat{\boldsymbol{\mu}}_{\mathbf{u}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{u}}$ are the sample mean vector and sample covariance matrix, respectively. These quantities are frequently used to measure the skewness and kurtosis for multivariate situations. For example, the theoretical values of $b_{1,k}$ and $b_{2,k}$ for the multivariate normal distribution are 0 and $k(k+2)$, respectively. These measures of multivariate skewness and kurtosis are invariant under linear transformation so that the MCP technique does not affect their value. Therefore, we do not report the C-GADP results in this experiment, only C-GADP-MCP.

The results are summarized in Table 2 with three different sample sizes ($n = 100$, 500, and 1,000). In simulation 3, with the skew-t distribution, STDP-MCP shows excellent performance in both skewness and kurtosis experiments, as we expected. The more interesting situations are simulations 1 and 2 using the multivariate **g**-and-**h** distribution where we are able to compare all methods on a fair ground. Table 2 shows that STDP-MCP is best in simulation 3, which uses data sets whose distribution is close to an elliptical or skew-elliptical shape, whereas DSP performs best in simulation 2. In simulation 1, DSP is best with respect to skewness, whereas STDP-MCP is best with respect to kurtosis. Overall, although the skewness and kurtosis performance of STDP-MCP, DSP, and C-GADPST-MCP are comparable, performance is determined by the underlying distribution of the data set.

A different point of view consists in computing the empirical coverage (in %) of skewness and kurtosis for the perturbed data sets, corresponding to the results presented in Table 2. Specifically, this means that for each of the 1,000 simulated biases, we add ± 1.96 times the overall SD and check whether it contains the value 0, that is, zero bias. Table 3 reports the proportion of such intervals containing zero over the 1,000 replicates. By construction, the nominal coverage should be approximately 95%. The results are given for samples of size $n = 100$, 500, and 1,000. Focusing on the sample size $n = 1,000$, we can make the following interesting comments. GADP-MCP has very poor coverage in terms of skewness in simulations 2 and 3, and generally poor coverage in terms of kurtosis for all three simulation settings. This is not surprising when comparing normal elliptical contours with the nonnormal contours in Figure 1. The

Table 2 Skewness and Kurtosis Experiment for the Perturbed Data Sets

	n	Simulation 1	Simulation 2	Simulation 3
Skewness				
STDP-MCP	100	0.2850 (0.9654)	−1.1251 (2.0971)	−0.1208 (2.3969)
	500	0.0755 (0.2549)	−1.2796 (1.0526)	0.1245 (1.8246)
	1,000	0.0430 (0.1144)	−1.2465 (0.6717)	−0.0224 (0.8998)
GADP-MCP	100	−0.3258 (0.7037)	−2.1061 (2.1139)	−2.0503 (2.0973)
	500	−0.0937 (0.2143)	−2.0904 (1.0832)	−1.6060 (1.3371)
	1,000	−0.0471 (0.0938)	−2.0561 (0.6725)	−1.5210 (0.7376)
C-GADP-MCP	100	1.3587 (1.7172)	2.2500 (2.8172)	0.2454 (1.9304)
	500	0.5783 (0.9148)	2.5431 (3.5223)	0.2848 (1.8073)
	1,000	0.2916 (0.4112)	1.9486 (2.7534)	0.0684 (1.1084)
C-GADPST-MCP	100	0.4552 (1.9150)	0.0760 (3.2779)	−0.2534 (3.8928)
	500	0.0714 (0.2809)	−0.5973 (1.1750)	−0.5191 (1.4436)
	1,000	0.0260 (0.1178)	−0.6474 (0.6546)	−0.5611 (0.7724)
DSP	100	0.2026 (0.7928)	0.1720 (1.4506)	−0.5992 (1.6445)
	500	0.0604 (0.2134)	−0.0919 (1.0400)	−0.5325 (1.0200)
	1,000	0.0235 (0.0911)	−0.1834 (0.4764)	−0.5412 (0.4230)
Kurtosis				
STDP-MCP	100	−0.2691 (2.1620)	−1.5534 (3.3489)	−0.4729 (4.2070)
	500	−0.0482 (1.4168)	−1.6502 (2.9638)	0.2255 (5.1733)
	1,000	0.0603 (1.0533)	−1.6819 (2.2915)	−0.0614 (3.7739)
GADP-MCP	100	−1.5673 (2.0294)	−3.4686 (3.6081)	−4.8823 (3.8604)
	500	−1.9556 (1.3160)	−4.7348 (3.0974)	−6.8787 (3.8129)
	1,000	−1.9410 (0.9678)	−4.9478 (2.3381)	−7.4150 (3.0480)
C-GADP-MCP	100	2.6433 (3.5576)	4.7261 (4.9461)	0.3147 (3.6915)
	500	2.7379 (3.3597)	9.2453 (7.8680)	1.1338 (4.0723)
	1,000	2.1649 (2.6135)	9.4031 (7.8953)	0.6704 (3.9870)
C-GADPST-MCP	100	0.1946 (3.3575)	0.0920 (4.7732)	−1.6465 (5.3702)
	500	−0.5541 (1.4422)	−0.9990 (3.2672)	−3.0238 (3.9674)
	1,000	−0.6303 (1.0018)	−1.1548 (2.3933)	−3.5774 (3.0276)
DSP	100	0.2701 (1.9209)	0.7891 (2.5809)	−1.5962 (3.1554)
	500	−0.0155 (1.2495)	0.8682 (2.8676)	−2.5942 (2.6901)
	1,000	−0.1792 (0.8331)	0.7951 (1.8612)	−3.0098 (2.0517)

Note. AB with SD in parentheses.

Table 3 Empirical Coverage of Skewness and Kurtosis for the Perturbed Data Sets

		Coverage (%)					
		Skewness			Kurtosis		
Methods	n	Sim. 1	Sim. 2	Sim. 3	Sim. 1	Sim. 2	Sim. 3
STDP-MCP	100	94.5	93.6	95.1	95.3	94.0	93.4
	500	96.5	86.7	97.0	95.4	93.8	93.4
	1,000	95.8	60.6	96.1	96.0	92.0	96.0
GADP-MCP	100	93.2	90.0	89.2	89.0	88.0	82.1
	500	97.0	63.3	91.2	72.9	78.3	67.7
	1,000	95.6	6.2	57.4	52.4	52.3	32.9
C-GADP-MCP	100	89.1	87.7	94.8	89.5	84.2	93.9
	500	94.7	91.9	97.5	90.1	84.3	94.6
	1,000	93.1	91.7	96.0	90.1	84.3	95.3
C-GADPST-MCP	100	97.7	96.1	97.1	97.2	93.1	94.8
	500	96.7	95.9	97.0	94.4	96.1	94.2
	1,000	95.9	88.8	92.8	93.7	94.2	86.0
DSP	100	93.9	94.4	94.7	94.9	94.7	93.2
	500	96.1	99.3	97.7	95.5	97.5	91.0
	1,000	95.8	97.3	83.5	95.4	95.4	76.5

C-GADP-based methods have sometimes a somewhat poor coverage, probably because of their fixed normal copula dependence structure. Except for simulation 2, the STDP-MCP method has a fairly good coverage in terms of skewness. This was to be expected because the distribution in simulation 2 is quite nonnormal and its shape is rather different from the skew-t distributional shape. In addition, its has different tail behaviors in each of its components, a challenging issue for the skew-t distribution that has only one parameter, ν, to control the tails (see also §§6 and 7 for further discussions of this issue). DSP has also a fairly good coverage although not so much in simulation 3.

4.3. Quantiles

Because the skew-t distribution is flexible to capture the skewness and heavy tail of the original data set, we also expect that the STDP perturbed variables will have better tail properties, and this should be reflected in quantile experiments. This is important if the database user needs to answer questions such as the following: What is the value of the variable X such that 90% of the observations fall below that value? In other words, what is the 90% quantile of X? This aspect can be explored by comparing quantiles of the perturbed variables.

Tables 4–6 present the AB and SD of 90%, 95%, 99%, and 99.5% quantiles of the perturbed variables $\mathbf{Y}$ by several perturbation methods under the three different simulation setups for sample sizes $n = 100$, 500, and 1,000. Unlike for the higher-order moments, it is difficult to judge which method performs best. In simulation 1, we observe that the skew-t distribution-based methods (STDP-MCP and C-GADPST-MCP) perform better than GADP-MCP, C-GADP, and C-GADP-MCP, especially in upper quantiles (see 99% and 99.5%). All methods, except GADP-MCP, show comparable average bias in simulation 2. This is interesting in the sense that in simulation 2, where the original data sets are generated from a highly nonnormal distribution, we expect the copula-based methods (C-GADP, C-GADP-MCP, and C-GADPST-MCP) to be much better than the STDP method. Indeed, the marginal distributions of the perturbed variables will reflect well the marginal properties of the original confidential variables, and the multivariate skew-t distribution may suffer from not fitting the original data set as well. In simulation 3 from the skew-t distribution, the skew-t-based methods STDP-MCP and C-GADPST-MCP give the best performance, as we expected. Here also, we included the results of STDP without MCP to see the effect of mean and covariance preserving on the quantiles. As can be seen, this effect is again fairly small. Of course, DSP preserves the quantiles exactly, giving zero biases, because it is based on the reordering of the original values. We also considered databases simulated under a multivariate normal distribution. In that case, all perturba-

Table 4 Quantile Experiment in Simulation 1

Quantiles (%)	*n*	STDP	STDP-MCP	GADP-MCP	C-GADP	C-GADP-MCP	C-GADPST-MCP
Y_1							
90	100	0.0202 (0.2137)	0.0250 (0.1649)	0.0404 (0.1593)	0.2694 (0.2406)	0.0234 (0.1732)	0.0030 (0.1612)
	500	0.0160 (0.0896)	0.0218 (0.0727)	0.0441 (0.0751)	0.0491 (0.0915)	−0.0145 (0.0621)	0.0033 (0.0734)
	1,000	0.0164 (0.0655)	0.0164 (0.0526)	0.0441 (0.0526)	0.0257 (0.0614)	−0.0106 (0.0433)	0.0031 (0.0498)
95	100	−0.0014 (0.2797)	0.0038 (0.2175)	0.0029 (0.2187)	0.3758 (0.3761)	0.0926 (0.2648)	−0.0107 (0.2111)
	500	0.0009 (0.1228)	0.0053 (0.1039)	0.0178 (0.1049)	0.0652 (0.1280)	−0.0055 (0.0928)	−0.0003 (0.1009)
	1,000	0.0092 (0.0872)	0.0086 (0.0735)	0.0218 (0.0739)	0.0339 (0.0827)	−0.0063 (0.0612)	0.0016 (0.0720)
99	100	−0.0666 (0.4823)	−0.0658 (0.4087)	−0.1085 (0.4285)	0.5051 (0.5244)	0.1410 (0.4384)	−0.0108 (0.4396)
	500	−0.0724 (0.2657)	−0.0670 (0.2408)	−0.1161 (0.2404)	0.2562 (0.3585)	0.1635 (0.3043)	−0.0133 (0.2227)
	1,000	−0.0556 (0.1910)	−0.0571 (0.1779)	−0.1176 (0.1730)	0.1148 (0.1983)	0.0638 (0.1698)	−0.0115 (0.1655)
99.5	100	−0.1094 (0.5740)	−0.1067 (0.4872)	−0.1779 (0.4982)	0.2525 (0.2622)	−0.0599 (0.2457)	−0.0214 (0.4762)
	500	−0.1138 (0.3542)	−0.1083 (0.3277)	−0.1910 (0.3257)	0.4707 (0.5162)	0.3350 (0.4363)	−0.0200 (0.3125)
	1,000	−0.0891 (0.2622)	−0.0912 (0.2461)	−0.2085 (0.2456)	0.2434 (0.3518)	0.1832 (0.3146)	−0.0168 (0.2413)
Y_2							
90	100	0.0035 (0.1775)	0.0047 (0.1587)	0.0177 (0.1587)	0.0453 (0.1825)	−0.0095 (0.1396)	0.0001 (0.1521)
	500	0.0086 (0.0787)	0.0093 (0.0687)	0.0242 (0.0710)	0.0059 (0.0717)	−0.0140 (0.0565)	0.0025 (0.0694)
	1,000	0.0108 (0.0585)	0.0116 (0.0501)	0.0243 (0.0492)	0.0023 (0.0505)	−0.0098 (0.0400)	0.0060 (0.0497)
95	100	0.0023 (0.2293)	0.0022 (0.2032)	0.0154 (0.2033)	0.1356 (0.2762)	0.0602 (0.2124)	0.0014 (0.2059)
	500	0.0079 (0.1043)	0.0077 (0.0946)	0.0105 (0.0984)	0.0185 (0.1030)	−0.0066 (0.0852)	−0.0010 (0.0940)
	1,000	0.0065 (0.0783)	0.0075 (0.0678)	0.0106 (0.0675)	0.0098 (0.0691)	−0.0057 (0.0559)	0.0058 (0.0695)
99	100	−0.0274 (0.4163)	−0.0249 (0.3777)	−0.0430 (0.3831)	0.5136 (0.5025)	0.2862 (0.4112)	−0.0080 (0.3685)
	500	−0.0260 (0.2258)	−0.0268 (0.2066)	−0.0558 (0.2125)	0.1673 (0.2922)	0.1254 (0.2544)	−0.0003 (0.2184)
	1,000	−0.0194 (0.1759)	−0.0189 (0.1652)	−0.0711 (0.1619)	0.0714 (0.1734)	0.0499 (0.1537)	−0.0039 (0.1593)
99.5	100	−0.0650 (0.4680)	−0.0643 (0.4210)	−0.0997 (0.4366)	0.2568 (0.2513)	0.0886 (0.2189)	−0.0284 (0.4042)
	500	−0.0483 (0.3013)	−0.0484 (0.2822)	−0.1000 (0.2774)	0.3826 (0.4220)	0.3028 (0.3634)	−0.0088 (0.2796)
	1,000	−0.0459 (0.2443)	−0.0460 (0.2331)	−0.1152 (0.2262)	0.1690 (0.3291)	0.1429 (0.2979)	−0.0126 (0.2275)

Note. AB with SD in parentheses.

Table 5 Quantile Experiment in Simulation 2

Quantiles (%)	*n*	STDP	STDP-MCP	GADP-MCP	C-GADP	C-GADP-MCP	C-GADPST-MCP
Y_1							
90	100	0.0040 (0.2538)	0.0259 (0.2093)	0.0225 (0.1937)	0.3364 (0.3095)	0.0169 (0.2129)	0.0204 (0.2081)
	500	−0.0174 (0.1125)	0.0063 (0.0944)	0.0224 (0.0924)	0.0612 (0.1147)	−0.0331 (0.0799)	0.0183 (0.0969)
	1,000	−0.0259 (0.0855)	0.0031 (0.0684)	0.0241 (0.0624)	0.0326 (0.0759)	−0.0243 (0.0544)	0.0194 (0.0648)
95	100	−0.0819 (0.3300)	−0.0581 (0.2707)	−0.1094 (0.2573)	0.5262 (0.5560)	0.1125 (0.3400)	−0.0062 (0.2810)
	500	−0.1025 (0.1715)	−0.0745 (0.1413)	−0.1020 (0.1264)	0.0920 (0.1806)	−0.0227 (0.1245)	0.0151 (0.1369)
	1,000	−0.1218 (0.1238)	−0.0886 (0.1003)	−0.0991 (0.0874)	0.0482 (0.1121)	−0.0232 (0.0801)	0.0166 (0.0962)
99	100	−0.3380 (0.6510)	−0.3087 (0.5491)	−0.4855 (0.5425)	0.7716 (0.8574)	0.1449 (0.6180)	−0.0681 (0.5843)
	500	−0.3950 (0.3890)	−0.3642 (0.3335)	−0.5370 (0.3136)	0.3968 (0.5979)	0.2065 (0.4487)	−0.0494 (0.3151)
	1,000	−0.4522 (0.2893)	−0.4116 (0.2505)	−0.5514 (0.2269)	0.1763 (0.2896)	0.0698 (0.2331)	−0.0510 (0.2284)
99.5	100	−0.4660 (0.7964)	−0.4380 (0.6592)	−0.6802 (0.6630)	0.3858 (0.4287)	−0.1725 (0.3136)	−0.1192 (0.6247)
	500	−0.5641 (0.5281)	−0.5295 (0.4670)	−0.7637 (0.4460)	0.7753 (0.8834)	0.4716 (0.6514)	−0.0972 (0.4512)
	1,000	−0.6083 (0.3880)	−0.5637 (0.3443)	−0.7932 (0.3320)	0.3848 (0.5676)	0.2541 (0.4754)	−0.0859 (0.3354)
Y_2							
90	100	0.0147 (0.2133)	0.0220 (0.1926)	0.0064 (0.1908)	0.0557 (0.2321)	−0.0346 (0.1700)	0.0193 (0.1928)
	500	0.0378 (0.1187)	0.0328 (0.1014)	0.0095 (0.0894)	0.0087 (0.0926)	−0.0335 (0.0751)	0.0244 (0.0905)
	1,000	0.0511 (0.0829)	0.0384 (0.0691)	0.0101 (0.0619)	0.0046 (0.0682)	−0.0263 (0.0558)	0.0278 (0.0637)
95	100	−0.0219 (0.3051)	−0.0175 (0.2702)	−0.0827 (0.2575)	0.1853 (0.3942)	0.0522 (0.2705)	0.0132 (0.2752)
	500	0.0239 (0.1733)	0.0185 (0.1447)	−0.0982 (0.1237)	0.0309 (0.1482)	−0.0295 (0.1119)	0.0194 (0.1313)
	1,000	0.0467 (0.1205)	0.0320 (0.0998)	−0.0943 (0.0833)	0.0170 (0.0977)	−0.0241 (0.0771)	0.0312 (0.0935)
99	100	−0.1955 (0.6044)	−0.1943 (0.5269)	−0.3707 (0.5101)	0.8174 (0.8684)	0.3561 (0.6146)	−0.0632 (0.5219)
	500	−0.1181 (0.3809)	−0.1271 (0.3411)	−0.4475 (0.2940)	0.2732 (0.4696)	0.1568 (0.3726)	−0.0428 (0.3170)
	1,000	−0.0621 (0.2903)	−0.0812 (0.2535)	−0.4690 (0.2168)	0.1367 (0.2840)	0.0619 (0.2310)	−0.0393 (0.2332)
99.5	100	−0.3184 (0.6761)	−0.3200 (0.5648)	−0.5489 (0.5885)	0.4087 (0.4342)	0.0318 (0.2915)	−0.1218 (0.5302)
	500	−0.2053 (0.4894)	−0.2149 (0.4356)	−0.6327 (0.3880)	0.6904 (0.8013)	0.4578 (0.5901)	−0.0920 (0.4001)
	1,000	−0.1477 (0.4094)	−0.1689 (0.3689)	−0.6577 (0.3170)	0.3320 (0.5656)	0.2337 (0.4705)	−0.0911 (0.3328)

Note. AB with SD in parentheses.

Table 6 Quantile Experiment in Simulation 3

Quantiles (%)	n	STDP	STDP-MCP	GADP-MCP	C-GADP	C-GADP-MCP	C-GADPST-MCP
Y_1							
90	100	0.0091 (0.1926)	0.0140 (0.1575)	0.0087 (0.1422)	0.2680 (0.2656)	0.0113 (0.1679)	0.0110 (0.1607)
	500	−0.0022 (0.0886)	−0.0006 (0.0790)	0.0132 (0.0718)	0.0552 (0.0877)	−0.0289 (0.0633)	0.0042 (0.0739)
	1,000	0.0013 (0.0621)	0.0028 (0.0535)	0.0164 (0.0503)	0.0283 (0.0596)	−0.0213 (0.0439)	0.0088 (0.0541)
95	100	0.0043 (0.2752)	0.0058 (0.2168)	−0.0701 (0.1980)	0.4401 (0.5056)	0.1013 (0.2886)	−0.0023 (0.2179)
	500	−0.0027 (0.1268)	−0.0029 (0.1080)	−0.0813 (0.0948)	0.0883 (0.1337)	−0.0210 (0.0960)	0.0063 (0.1102)
	1,000	0.0015 (0.0938)	0.0030 (0.0776)	−0.0769 (0.0676)	0.0483 (0.0867)	−0.0161 (0.0627)	0.0112 (0.0754)
99	100	−0.0291 (0.5736)	−0.0408 (0.4756)	−0.3260 (0.4277)	0.6436 (0.7142)	0.1774 (0.5275)	−0.0278 (0.4359)
	500	−0.0044 (0.3198)	−0.0087 (0.2653)	−0.3839 (0.2302)	0.3665 (0.4764)	0.1891 (0.3575)	−0.0004 (0.2643)
	1,000	−0.0049 (0.2348)	−0.0054 (0.1961)	−0.4007 (0.1730)	0.1874 (0.2618)	0.0824 (0.2073)	−0.0009 (0.1995)
99.5	100	−0.0581 (0.6917)	−0.0809 (0.5204)	−0.4699 (0.4847)	0.3218 (0.3571)	−0.0684 (0.2853)	−0.0616 (0.4866)
	500	−0.0009 (0.4261)	−0.0093 (0.3627)	−0.5336 (0.3153)	0.7085 (0.8243)	0.4532 (0.5784)	0.0026 (0.3715)
	1,000	−0.0064 (0.3506)	−0.0081 (0.3059)	−0.5647 (0.2536)	0.3568 (0.4506)	0.2240 (0.3715)	0.0008 (0.2991)
Y_2							
90	100	0.0047 (0.2182)	0.0150 (0.1785)	0.0349 (0.1577)	0.4348 (0.3027)	0.0135 (0.1963)	0.0150 (0.1646)
	500	−0.0012 (0.0998)	0.0005 (0.0786)	0.0302 (0.0733)	0.0830 (0.0917)	−0.0177 (0.0626)	0.0058 (0.0737)
	1,000	−0.0017 (0.0673)	0.0009 (0.0585)	0.0327 (0.0533)	0.0448 (0.0638)	−0.0102 (0.0448)	0.0051 (0.0546)
95	100	−0.0007 (0.2976)	0.0067 (0.2317)	−0.0397 (0.2149)	0.5753 (0.5569)	0.0878 (0.3289)	0.0045 (0.2300)
	500	−0.0020 (0.1413)	−0.0019 (0.1118)	−0.0609 (0.1036)	0.1029 (0.1420)	−0.0125 (0.1007)	−0.0043 (0.1084)
	1,000	−0.0034 (0.0986)	−0.0008 (0.0793)	−0.0571 (0.0725)	0.0526 (0.0878)	−0.0131 (0.0610)	−0.0010 (0.0756)
99	100	−0.0237 (0.6195)	−0.0307 (0.4944)	−0.3086 (0.4410)	0.5795 (0.7079)	0.0342 (0.5099)	−0.0153 (0.4612)
	500	−0.0096 (0.3375)	−0.0135 (0.2901)	−0.3721 (0.2514)	0.2286 (0.3707)	0.0744 (0.2919)	−0.0264 (0.2736)
	1,000	0.0005 (0.2555)	0.0019 (0.2133)	−0.3761 (0.1815)	0.1210 (0.2530)	0.0317 (0.2051)	−0.0178 (0.2018)
99.5	100	−0.0584 (0.7776)	−0.0739 (0.5944)	−0.4540 (0.5357)	0.2895 (0.3542)	−0.1950 (0.2671)	−0.0814 (0.5231)
	500	−0.0018 (0.4480)	−0.0089 (0.3865)	−0.5430 (0.3290)	0.3824 (0.5723)	0.2050 (0.4707)	−0.0293 (0.3750)
	1,000	−0.0016 (0.3656)	−0.0036 (0.3158)	−0.5614 (0.2618)	0.1768 (0.3761)	0.0777 (0.3184)	−0.0325 (0.2950)

Note. AB with SD in parentheses.

tion methods we considered gave similar results for quantiles of the perturbed variables $\mathbf{Y}$.

5. Application to a Medical Database

We apply our new methodology to a medical database of measurements on intakes of protein (PROT), saturated fat (FAT), and carbohydrate (CARB) on $n = 3{,}145$ women during an epidemiologic cohort study related to breast cancer (see Jones et al. 1987). This medical information is usually confidential and the database cannot be released to the public in its original form. We consider the perturbation of the variable PROT while maintaining the variables FAT and CARB, i.e., $\mathbf{U} = (X, \mathbf{S}^{\mathrm{T}})^{\mathrm{T}}$ with $X = \mathrm{PROT}$ and $\mathbf{S} = (\mathrm{FAT}, \mathrm{CARB})^{\mathrm{T}}$. We also investigated the perturbation of all three variables and obtained similar results to those reported below.

We start by studying the distributional properties of this database empirically. The measures of multivariate skewness and kurtosis are $b_{1,3} = 9.2$ and $b_{2,3} = 43.2$, respectively. They seem to indicate that the distribution is not multivariate normal because they are far from the theoretical values of 0 and 15. A formal hypothesis test of normality (see, e.g., Mardia et al. 1979, p. 148) can be based on the asymptotic distribution of the measures of multivariate skewness and kurtosis given by

$$\tfrac{1}{6}nb_{1,k} \sim \chi^2_f, \quad \text{where } f = \tfrac{1}{6}k(k+1)(k+2),$$
$$\{b_{2,k} - k(k+2)\}/\{8k(k+2)/n\}^{1/2} \sim \mathrm{N}(0,1),$$

where $k = 3$ and $n = 3{,}145$ for our database. Both tests give p-values <0.0000 so we reject the hypothesis of a multivariate normal distribution.

Graphical evidence of the nonnormality is also apparent in Figure 2, depicting scatter plots for every pair of the three variables PROT, FAT, and CARB. All scatter plots are skewed and heavy tailed, and so we fit a skew-t distribution to this database. Contours of the fitted density are also plotted in Figure 2. The estimated shape parameter and degrees of freedom of the skew-t are $\hat{\boldsymbol{\alpha}} = (3.43, 0.04, -0.83)^{\mathrm{T}}$ and $\hat{\nu} = 5.26$, respectively, indicating skewness and heavy tails. The p-value of the LRT of the hypothesis of normality is essentially zero, hence we strongly reject a normal distribution for this data set. As a further diagnostic, we draw Healy plots in Figure 3 by fitting multivariate normal and skew-t distributions. Figure 3 demonstrates that the skew-t distribution is a much better fit than the normal distribution. Consequently, we expect that the STDP method will yield better results than the GADP method on this database.

Figure 2 Scatter Plots of the Medical Database of PROT, FAT, and CARB with Contours of the Fitted Skew-t Density

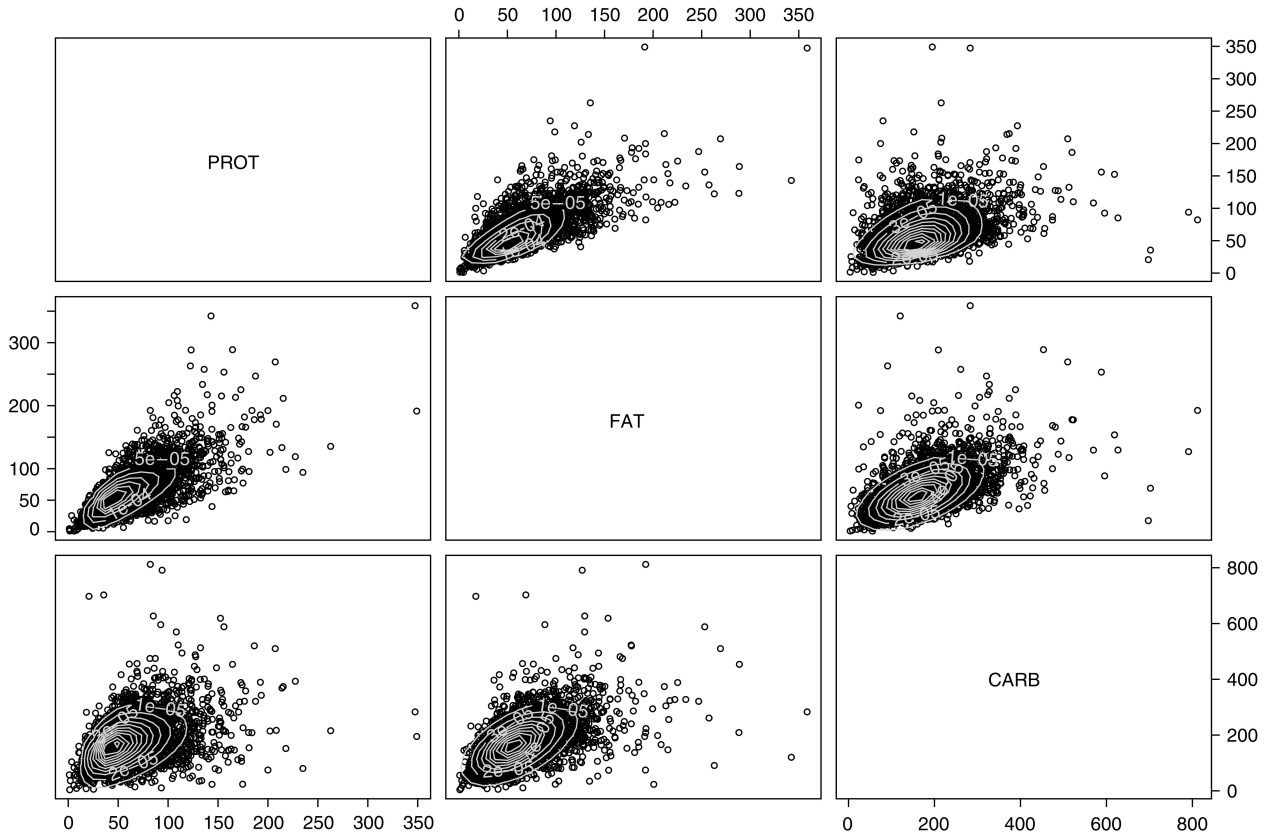


Table 7 shows that the original, DSP, and C-GADP perturbed means and covariances (correlations) on the medical database are quite close. For the GADP-MCP, C-GADP-MCP, C-GADPST-MCP, and STDP-MCP methods, the means and covariances (correlations) are exactly the same on the perturbed data as on the original data. The rank order correlations with PROT for all perturbation methods are also provided. In this example, the rank order correlations are somewhat smaller for STDP-MCP than for STDP, hence illustrating the cost for trying to preserve the mean and covariance exactly.

Figure 3 Healy Plots When Either a Normal Distribution (Left Panel) or a Skew-t Distribution (Right Panel) Is Fitted to the Medical Database

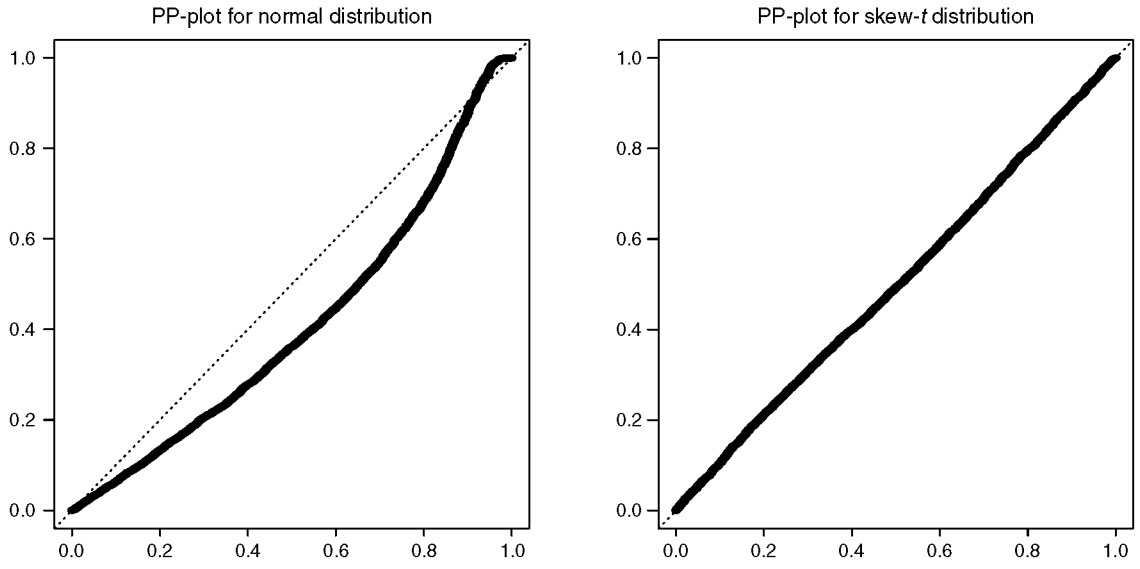


Table 7 **Panel A: Original, DSP, and C-GADP Perturbed Means (Standard Deviations) and Covariances (Correlations) with PROT on the Medical Database; Panel B: Rank Order Correlations with PROT on the Medical Database**

	PROT	FAT	CARB
		Panel A	
Original			
Mean (SD)	64.1 (32.1)	65.5 (35.0)	175.5 (81.7)
Cov (corr)	1,030.1 (1.00)	803.3 (0.72)	1,059.5 (0.40)
DSP			
Mean (SD)	64.1 (32.1)	65.5 (35.0)	175.5 (81.7)
Cov (corr)	1,030.1 (1.00)	802.7 (0.71)	1,041.3 (0.40)
C-GADP			
Mean (SD)	64.5 (33.0)	65.5 (35.0)	175.5 (81.7)
Cov (corr)	1,090.4 (1.00)	826.3 (0.72)	1,137.9 (0.42)
		Panel B	
Original			
Rank corr	1.00	0.72	0.44
DSP			
Rank corr	1.00	0.72	0.44
C-GADP			
Rank corr	1.00	0.71	0.43
STDP			
Rank corr	1.00	0.71	0.46
STDP-MCP			
Rank corr	1.00	0.65	0.41
GADP-MCP			
Rank corr	1.00	0.67	0.40
C-GADP-MCP			
Rank corr	1.00	0.71	0.42
C-GADPST-MCP			
Rank corr	1.00	0.71	0.40

The measures of multivariate skewness and kurtosis of the original and perturbed data by all previously mentioned methods are provided in Table 8. Except for GADP-MCP, all perturbation methods yield measures fairly close to those on the original data. In particular, STDP-MCP yields the kurtosis closest to the one of the original data.

The improvement in skewness and kurtosis obtained by the STDP and other nonnormal methods suggests that quantiles of the perturbed variables will be similar to those of the original variables. This property is of crucial importance because database managers are usually interested in clients, customers, or patients with high-valued attributes. For instance, in our medical database, measurements in the tail of the distribution are of great interest for the study of breast cancer. Indeed, it is a common fact that diseases are highly associated with people who are exposed to a higher level of a certain factor than a given level. Hence, data perturbation methods that can provide accurate modeling of the behavior of the tails are needed.

Table 8 **Skewness and Kurtosis of the Original and Perturbed Medical Database**

	Skewness	Kurtosis
Original	9.2	43.2
STDP-MCP	8.1	43.3
GADP-MCP	5.6	31.8
C-GADP-MCP	9.0	45.1
C-GADPST-MCP	9.3	44.1
DSP	8.8	42.2

Table 9 **Quantiles of the Original and Perturbed Medical Database**

	Quantiles (%)						
	50	60	70	80	90	95	AD
Original	58.5	65.7	74.1	85.8	104.5	123.2	—
STDP-MCP	56.9	64.4	73.4	85.3	104.1	124.1	5.4
GADP-MCP	62.2	70.6	79.3	89.6	104.9	117.8	23.4
C-GADP	58.4	66.0	74.2	86.5	104.6	123.1	1.4
C-GADP-MCP	58.2	65.5	73.5	85.4	103.1	120.9	5.2
C-GADPST-MCP	58.6	65.0	72.7	84.8	104.1	124.8	5.2
DSP	58.5	65.7	74.1	85.8	104.5	123.2	0

Table 9 provides quantiles of the original confidential variable and the perturbed variable by the previously mentioned methods on the medical database. The last column lists the sums (over the quantiles) of absolute deviations (AD) between the results from each perturbation methods and the original data. Although GADP-MCP is usually not providing reliable quantiles, it is difficult to give an overall winner among the other MCP-based perturbation methods. Of course, DSP maintains the marginal quantiles exactly and C-GADP performs quite well on this example at the cost of not preserving the mean and covariance exactly.

Table 10 responds to an interesting query related to the joint distribution of the variables: What is the average of the PROT values when FAT $> a$ and CARB $> b$? This average is reported for the original database, as well as for the perturbed data, for various values of a and b. Here again, it is difficult to name an overall winner, but the STDP-MCP method provides overall fairly competitive answers, especially for $b = 350$. It can be seen that DSP now performs quite poorly. This was to be expected, because DSP preserves the marginal distributions exactly, but not the joint distribution.

Table 10 **Average of PROT When FAT $> a$ and CARB $> b$ of the Original and Perturbed Medical Database**

	$b = 300$				$b = 350$			
a	50	100	150	AD	50	100	150	AD
Original	87.6	109.7	152.4	—	107.2	121.1	155.9	—
STDP-MCP	93.6	109.5	169.7	23.5	100.9	105.2	157.4	23.7
GADP-MCP	92.5	113.3	135.1	25.8	103.7	116.9	140.5	23.1
C-GADP-MCP	86.4	109.5	149.9	3.9	92.4	102.3	143.4	46.1
C-GADPST-MCP	86.4	110.4	131.1	23.2	94.5	110.3	130.2	49.2
DSP	82.8	102.8	123.0	41.1	89.6	108.2	123.0	63.4

6. Limitations

Although STDP is much more flexible than GADP, it has limitations as well. For example, both the multivariate skew-t and normal distributions are unimodal. If the distribution of the database is multimodal, then these two methods will not be able to capture this feature. One possibility is to make use of even more flexible multivariate distributions, such as the flexible skew-symmetric distributions introduced by Ma and Genton (2004). Unfortunately, a procedure for simulation from the conditional distribution in that class is currently unknown and likely unavailable in closed form. Whether such extensions are possible is an open problem.

As mentioned previously, a potential disadvantage of the skew-t distribution is that there is only one parameter, ν, that regulates the tail behavior of all variables. If one variable has normal tails whereas another has lognormal tails, then the single degrees of freedom parameter has to provide a compromise between those two tail behaviors. An alternative approach would be to consider some skewed versions of other multivariate t distributions with multiple degrees of freedom parameters. Unfortunately, such distributions have neither appealing parametric forms, as discussed by Azzalini and Genton (2008), nor known conditional distributions, hence their use in data perturbation problems is not realistic. The multivariate skew-t distribution remains a reasonable compromise between flexibility and mathematical tractability, making it a particularly attractive general-purpose tool.

The number of variables in the Monte Carlo simulations and in the medical database example was intentionally small for the sake of illustration and description of the results. Nevertheless, the STDP method can be applied to databases with a large number of variables as long as the skew-t distribution can be fitted to the data. In our experience, databases with up to 200 variables and several thousands observations can be handled without problem on a desktop computer. The fit to databases with even larger numbers of variables will require the development of new procedures.

Another issue arises when some variables have a positive support, i.e., they cannot be negative. Indeed, the skew-t distribution, similarly to the normal, has the whole real space as support. Although the fit of a skew-t distribution to variables with a positive support will tend to concentrate its mass on the positive values, this can be a source of problems. An alternative is to consider multivariate log-skew-t distributions, as recently introduced by Marchenko and Genton (2010). Those are flexible extensions of the multivariate log-normal distribution.

The STDP method has been developed to deal with continuous numerical data. At this point, it is not able to handle categorical data. One option is to fit skew-t distributions to continuous numerical variables for each level of the categorical variable. Another possibility is to consider multimodal distributions as mentioned above. Both of those solutions would require further investigations.

7. Conclusions

In this study, we developed a new data perturbation procedure for database security problems based on skew-t distributions. The class of skew-t distributions is a flexible parametric multivariate family that can model skewness and heavy tails in the data. It contains the multivariate normal distribution as a special case. Therefore, the new procedure, coined skew-t data perturbation (STDP) method, includes the classical general additive data perturbation (GADP) method based on the normal distribution. The proposed method consists of two simple basic procedures: the estimation of skew-t parameters from the database and the generation of random samples from the conditional distribution of the confidential variables given the nonconfidential variables.

We investigated the performance of the STDP method by means of a Monte Carlo simulation study and demonstrated its competitive behavior with respect to other existing methods, including C-GADP, and DSP, when the distribution of the database is nonnormal. This is especially true if the underlying distribution can be well approximated by a skew-t distribution. In fact, skew-t and related distributions occur naturally in settings where the data have been selected (see Arellano-Valle et al. 2006). In particular, we have shown the ability of STDP to reproduce characteristics of the joint tails of the distribution. This is very important for database users who want to answer higher-level questions. We illustrated the STDP method on a medical database related to breast cancer.

It is also important to recognize that perturbation methods that involve random sampling from the conditional distribution will have to deal with its impact on statistical inference. This additional uncertainty can be assessed, for example, by releasing multiple perturbed data sets in the spirit of multiple imputation (see Raghunathan et al. 2003, Reiter 2003). Further research on this issue is highly desirable.

In genuine applications, the data may contain nonlinear relationships and include constraints on variables, for instance, some variables sum to other variables. We expect that the nonlinear relationship among variables can be dealt with more effectively using STDP than GADP because the multivariate normal density has an elliptical shape where only linear

relationships of variables can appear. On the contrary, the skew-t distribution with flexible skewness and kurtosis parameters can reflect the possible nonlinear relations among variables. Indeed, the conditional expectation of the skew-t distribution is nonlinear (see Appendix A.3 as well as Arellano-Valle and Genton 2010). The copula-based methods using rank order correlations can also handle such nonlinearities. In the case where some variables are constrained to be the sum of other variables, a simple remedy is to remove the constrained variables before applying a perturbation method. Then those variables are replaced by new ones, using the constraints on the perturbed variables, if the constrained variables are confidential.

Acknowledgments

The authors are grateful to the editor, associate editor, and three anonymous referees for constructive suggestions that have improved the content and presentation of this paper. Marc G. Genton's research was supported in part by NSF Grants DMS-0504896, CMG ATM-0620624, and by Award KUS-C1-016-04, made by King Abdullah University of Science and Technology. Reinaldo B. Arellano-Valle's research was supported in part by Fondecyt Grant 1085241, Chile.

Appendix

A.1. Conditional Distribution of the Skew-t

The k-dimensional random vector $\mathbf{U} \sim \mathrm{ST}_k(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu)$ has the stochastic representation

$$\mathbf{U} = \boldsymbol{\xi} + V^{-1/2}\boldsymbol{\omega}\mathbf{Z}, \tag{5}$$

where $\mathbf{Z} \sim \mathrm{SN}_k(\mathbf{0}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\alpha})$ with $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$, and $V \sim \chi^2_\nu/\nu$, independent of $\mathbf{Z}$. Moreover, in terms of the well-known reparametrization $\boldsymbol{\delta} = (1 + \boldsymbol{\alpha}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}\boldsymbol{\alpha})^{-1/2}\bar{\boldsymbol{\Omega}}\boldsymbol{\alpha}$, the stochastic representation (5) can be rewritten as $\mathbf{U} = \boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{W}$, with $\mathbf{W} \overset{d}{=} \boldsymbol{\delta}|T_0| + \mathbf{T}$ and

$$\begin{bmatrix} T_0 \\ \mathbf{T} \end{bmatrix} \sim t_{1+k}\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{\Omega}} - \boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}} \end{bmatrix}, \nu\right), \tag{6}$$

where $t_{1+k}(\cdot)$ stands for the $(1+k)$-dimensional multivariate Student's t distribution. Let $\mathbf{U} = (\mathbf{U}_1^{\mathrm{T}}, \mathbf{U}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are of dimensions k_1 and k_2, respectively. By (5), we have $\mathbf{U}_i = \boldsymbol{\xi}_i + V^{-1/2}\boldsymbol{\omega}_i\mathbf{Z}_i$, $i = 1, 2$. In particular, $\mathbf{U}_2 = \boldsymbol{\xi}_2 + V^{-1/2}\boldsymbol{\omega}_2\mathbf{Z}_2$, where $\mathbf{Z}_2 \sim \mathrm{SN}_k(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{22}, \boldsymbol{\alpha}_{2(1)})$, with $\boldsymbol{\alpha}_{2(1)} = (\boldsymbol{\alpha}_2 + \bar{\boldsymbol{\Omega}}_{22}^{-1}\bar{\boldsymbol{\Omega}}_{21}\boldsymbol{\alpha}_1)/(1 + \boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2})^{1/2}$ and $\boldsymbol{\alpha}_{1\cdot2} = \boldsymbol{\omega}_{1\cdot2}\boldsymbol{\omega}_1^{-1}\boldsymbol{\alpha}_1$ (see, e.g., Azzalini and Capitanio 1999), and is independent of V. Thus, we have $\mathbf{U}_2 \sim \mathrm{ST}_{k_2}(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_{22}, \boldsymbol{\alpha}_{2(1)}, \nu)$. Next, considering this marginal distribution and (1), it follows that the conditional density of $(\mathbf{U}_1 \mid \mathbf{U}_2 = \mathbf{u}_2)$ is given by (2). Here we used that $(\nu + k)/(\nu + Q(\mathbf{u})) = ((\nu + k_2)/(\nu + Q(\mathbf{u}_2)))((\nu_{1\cdot2} + k_1)/(\nu_{1\cdot2} + Q_{1\cdot2}(\mathbf{u}_1; \mathbf{u}_2)))$, $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\omega}^{-1}(\mathbf{u} - \boldsymbol{\xi}) = \boldsymbol{\alpha}_{1\cdot2}^{T}\boldsymbol{\omega}_{1\cdot2}^{-1}(\mathbf{u}_1 - \boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2)) + \bar{\boldsymbol{\alpha}}_2^{\mathrm{T}}\boldsymbol{\omega}_2^{-1}(\mathbf{u}_2 - \boldsymbol{\xi}_2)$, and

$$\begin{aligned} &t_k(\mathbf{u} - \boldsymbol{\xi}; \boldsymbol{\Omega}, \nu) \\ &\quad = t_{k_1}(\mathbf{u}_1 - \boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2); \boldsymbol{\Omega}_{11\cdot2}(\mathbf{u}_2), \nu_{1\cdot2})\, t_{k_2}(\mathbf{u}_2 - \boldsymbol{\xi}_2; \boldsymbol{\Omega}_{22}, \nu); \end{aligned}$$

see, e.g., Arellano-Valle and Bolfarine (1995) for this last result. The vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\alpha}_i$, $i = 1, 2$, and the matrices $\bar{\boldsymbol{\Omega}}_{ij} = \boldsymbol{\omega}_i^{-1}\boldsymbol{\Omega}_{ij}\boldsymbol{\omega}_j^{-1}$, $i, j = 1, 2$, are induced by the partition $\mathbf{U}_i$, $i = 1, 2$ of $\mathbf{U}$. The conditional distribution of $\mathbf{U}_{1\cdot2} \equiv (\mathbf{U}_1 \mid \mathbf{U}_2)$ will be denoted by $\mathrm{EST}_{k_1}(\boldsymbol{\xi}_{1\cdot2}(\mathbf{U}_2), \boldsymbol{\Omega}_{11\cdot2}(\mathbf{U}_2), \boldsymbol{\alpha}_{1\cdot2}, \tau_{1\cdot2}(\mathbf{U}_2), \nu_{1\cdot2})$, the extended skew-$t$ family of distributions with density (2) (see Arellano-Valle and Genton 2010).

A.2. Simulation from the Conditional Distribution of the Skew-t

Let $\mathbf{U}_{1\cdot2} \sim \mathrm{EST}_{k_1}(\boldsymbol{\xi}_{1\cdot2}(\mathbf{U}_2), \boldsymbol{\Omega}_{11\cdot2}(\mathbf{U}_2), \boldsymbol{\alpha}_{1\cdot2}, \tau_{1\cdot2}(\mathbf{U}_2), \nu_{1\cdot2})$ with density (2). For each value $\mathbf{u}_2$ of $\mathbf{U}_2$, we will show that $\mathbf{U}_{1\cdot2} \overset{d}{=} \widetilde{\mathbf{U}}_{1\cdot2}$, where

$$\widetilde{\mathbf{U}}_{1\cdot2} = \boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2) + \sqrt{\frac{\nu + Q(\mathbf{u}_2)}{\nu + k_2}}\,\boldsymbol{\omega}_{1\cdot2}\mathbf{T}_{1\cdot2}, \tag{7}$$

$$\mathbf{T}_{1\cdot2} = \sqrt{\frac{\nu_{1\cdot2} + |\widetilde{T}_0|^2}{\nu_{1\cdot2} + 1}}\,\mathbf{T}_1 + \boldsymbol{\delta}_{1\cdot2}\widetilde{T}_0 \quad \text{and}$$

$$\widetilde{T}_0 \overset{d}{=} (T_0 \mid T_0 + \tilde{\tau}_{1\cdot2}(\mathbf{u}_2) > 0), \tag{8}$$

with

$$T_0 \sim t_1(0, 1, \nu_{1\cdot2}) \quad \text{and}$$

$$\mathbf{T}_1 \sim t_{k_1}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{11\cdot2} - \boldsymbol{\delta}_{1\cdot2}\boldsymbol{\delta}_{1\cdot2}^{\mathrm{T}}, \nu_{1\cdot2} + 1) \tag{9}$$

being independent random quantities. Here $\nu_{1\cdot2} = \nu + k_2$, $\bar{\boldsymbol{\Omega}}_{11\cdot2} = \boldsymbol{\omega}_{1\cdot2}^{-1}\boldsymbol{\Omega}_{11\cdot2}\boldsymbol{\omega}_{1\cdot2}^{-1}$, $\boldsymbol{\delta}_{1\cdot2} = \bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}/\sqrt{1 + \boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}}$ and $\tilde{\tau}_{1\cdot2}(\mathbf{u}_2) = \tau_{1\cdot2}(\mathbf{u}_2)/\sqrt{1 + \boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}\boldsymbol{\alpha}_{1\cdot2}}$. In fact, note first that $(\mathbf{T}_{1\cdot2} \mid \widetilde{T}_0 = z) \sim t_{k_1}(\boldsymbol{\delta}_{1\cdot2}z, [(\nu_{1\cdot2} + z^2)/(\nu_{1\cdot2} + 1)][\bar{\boldsymbol{\Omega}}_{11\cdot2} - \boldsymbol{\delta}_{1\cdot2}\boldsymbol{\delta}_{1\cdot2}^{\mathrm{T}}], \nu_{1\cdot2} + 1)$, where $f_{\widetilde{T}_0}(z) = t_1(z; \nu_{1\cdot2})/T_1(\tilde{\tau}(\mathbf{u}_2); \nu_{1\cdot2})$ for $z > -\tilde{\tau}_{1\cdot2}(\mathbf{u}_2)$. It follows that $f_{\mathbf{T}_{1\cdot2} \mid \widetilde{T}_0 = z}(\mathbf{w}) = f_{\mathbf{T} \mid T_0 = z}(\mathbf{w})$ and $f_{\widetilde{T}_0}(z) = f_{T_0}(z)/T_1(\tilde{\tau}(\mathbf{u}_2); \nu_{1\cdot2})$ for $z > -\tilde{\tau}_{1\cdot2}(\mathbf{u}_2)$, where

$$\begin{bmatrix} T_0 \\ \mathbf{T} \end{bmatrix} \sim t_{1+k_1}\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{\Omega}}_{11\cdot2} - \boldsymbol{\delta}_{1\cdot2}\boldsymbol{\delta}_{1\cdot2}^{\mathrm{T}} \end{bmatrix}, \nu_{1\cdot2}\right). \tag{10}$$

Therefore, from the symmetry of f_{T_0} and the identity $f_{\mathbf{T} \mid T_0 = z}(\mathbf{w}) f_{T_0}(z) = f_{\mathbf{T}}(\mathbf{w}) f_{T_0 \mid \mathbf{T} = \mathbf{w}}(z)$, we obtain

$$\begin{aligned} f_{\mathbf{T}}(\mathbf{w}) &= \frac{1}{T_1(\tilde{\tau}_{1\cdot2}(\mathbf{u}_2); \nu_{1\cdot2})} \int_{-\tilde{\tau}_{1\cdot2}(\mathbf{u}_2)}^{\infty} f_{\mathbf{T} \mid T_0 = z}(\mathbf{w}) f_{T_0}(z)\, dz \\ &= \frac{f_{\mathbf{T}}(\mathbf{w})}{T_1(\tilde{\tau}_{1\cdot2}(\mathbf{u}_2); \nu_{1\cdot2})} \int_{-\infty}^{\tilde{\tau}_{1\cdot2}(\mathbf{u}_2)} f_{T_0 \mid \mathbf{T} = \mathbf{w}}(z)\, dz \\ &= t_{k_1}(\mathbf{w}; \bar{\Omega}_{1\cdot2}, \nu_{1\cdot2}) \\ &\quad \times \frac{T_1(\sqrt{(\nu_{1\cdot2} + k_1)/(\nu_{1\cdot2} + \|\mathbf{w}\|^2)}(\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\mathbf{w} + \tilde{\tau}_{1\cdot2}(\mathbf{u}_2)); \nu_{1\cdot2} + k_1)}{T_1(\tilde{\tau}_{1\cdot2}(\mathbf{u}_2); \nu_{1\cdot2})}, \end{aligned}$$

because, by (10), $(T_0 \mid \mathbf{T} = \mathbf{w}) \sim t(\boldsymbol{\alpha}_{1\cdot2}^{\mathrm{T}}\mathbf{w}, (\nu_{1\cdot2} + \|\mathbf{w}\|^2)/(\nu_{1\cdot2} + k_1), \nu_{1\cdot2} + k_1)$, with $\boldsymbol{\alpha}_{1\cdot2} = \bar{\boldsymbol{\Omega}}_{11\cdot2}^{-1}\boldsymbol{\delta}_{1\cdot2}/\sqrt{1 - \boldsymbol{\delta}_{1\cdot2}^{\mathrm{T}}\bar{\boldsymbol{\Omega}}_{11\cdot2}^{-1}\boldsymbol{\delta}_{1\cdot2}}$. Finally, considering (7), the density of $\widetilde{\mathbf{U}}_{1\cdot2}$ can be computed as

$$\begin{aligned} f_{\widetilde{\mathbf{U}}_{1\cdot2}}(\mathbf{u}_1) = &\left(\frac{\nu + k_2}{\nu + Q(\mathbf{u}_2)}\right)^{k_1/2} |\boldsymbol{\omega}_{1\cdot2}|^{-1} \\ &\times f_{\mathbf{T}_{1\cdot2}}\left(\sqrt{\frac{\nu + k_2}{\nu + Q(\mathbf{u}_2)}}\,\boldsymbol{\omega}_{1\cdot2}^{-1}(\mathbf{u}_1 - \boldsymbol{\xi}_{1\cdot2}(\mathbf{u}_2))\right), \end{aligned}$$

leading to the density (2).

In (8), we note also that

$$T_0 \overset{d}{=} \sqrt{\frac{\nu+k_2}{V_0}}\, Z_0 \quad \text{and} \quad \mathbf{T}_1 \overset{d}{=} \sqrt{\frac{\nu+k_2+1}{V_1}}\, \mathbf{Z}_1, \tag{11}$$

where

$$\begin{aligned} Z_0 \sim \mathrm{N}(0,1), \quad \mathbf{Z}_1 \sim \mathrm{N}_{k_1}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{11\cdot 2} - \boldsymbol{\delta}_{1\cdot 2}\boldsymbol{\delta}_{1\cdot 2}^{\mathrm{T}}), \\ V_0 \sim \chi^2_{\nu+k_2}, \quad \text{and} \quad V_1 \sim \chi^2_{\nu+k_2+1}, \end{aligned} \tag{12}$$

and all of them are independent.

With the aforementioned ingredients, we can easily generate a random sample from the conditional distribution of $\mathbf{U}_{1\cdot 2}$ following the procedure described in §2.2.

A.3. Conditional Mean Vector and Covariance Matrix

Let $\mu_*(\mathbf{u}_2) = \mathrm{E}(T_0 \mid T_0 + \tilde{\tau}_{1\cdot 2}(\mathbf{u}_2) > 0)$, $\mu_{*2}(\mathbf{u}_2) = \mathrm{E}(T_0^2 \mid T_0 + \tilde{\tau}_{1\cdot 2}(\mathbf{u}_2) > 0)$, and $\sigma_*^2(\mathbf{u}_2) = \mathrm{Var}(T_0 \mid T_0 + \tilde{\tau}_{1\cdot 2}(\mathbf{u}_2) > 0) = \mu_{2*}(\mathbf{u}_2) - \mu_*^2(\mathbf{u}_2)$. From Arellano-Valle and Genton (2010), we have

$$\mu_*(\mathbf{u}_2) = \frac{\nu}{\nu-1}\left(1 + \frac{\tilde{\tau}_{1\cdot 2}^2(\mathbf{u}_2)}{\nu}\right)\frac{t_1(\tilde{\tau}_{1\cdot 2}(\mathbf{u}_2); \nu)}{T_1(\tilde{\tau}_{1\cdot 2}(\mathbf{u}_2); \nu)}, \quad \nu > 1,$$

$$\begin{aligned} \mu_{*2}(\mathbf{u}_2) = \frac{\nu}{\nu-2}\frac{T_1\left(\sqrt{(\nu-2)/\nu}\,\tilde{\tau}_{1\cdot 2}(\mathbf{u}_2); \nu-2\right)}{T_1(\tilde{\tau}_{1\cdot 2}(\mathbf{u}_2); \nu)} \\ - \tilde{\tau}_{1\cdot 2}(\mathbf{u}_2)\mu_*(\mathbf{u}_2), \quad \nu > 2. \end{aligned}$$

From the stochastic representation in (7), we obtain easily that

$$\mathrm{E}(\mathbf{U}_{1\cdot 2}) = \boldsymbol{\xi}_{1\cdot 2}(\mathbf{u}_2) + \sqrt{\frac{\nu+Q(\mathbf{u}_2)}{\nu+k_2}}\,\boldsymbol{\omega}_{1\cdot 2}\mathrm{E}(\mathbf{T}_{1\cdot 2}), \quad \nu_{1\cdot 2} > 1,$$

$$\mathrm{Var}(\mathbf{U}_{1\cdot 2}) = \left(\frac{\nu+Q(\mathbf{u}_2)}{\nu+k_2}\right)\boldsymbol{\omega}_{1\cdot 2}\mathrm{Var}(\mathbf{T}_{1\cdot 2})\boldsymbol{\omega}_{1\cdot 2}, \quad \nu_{1\cdot 2} > 2,$$

where by (8)

$$\begin{aligned} \mathrm{E}(\mathbf{T}_{1\cdot 2}) &= \mu_*(\mathbf{u}_2)\boldsymbol{\delta}_{1\cdot 2}, \quad \nu_{1\cdot 2} > 1, \\ \mathrm{Var}(\mathbf{T}_{1\cdot 2}) &= \left(\frac{\nu_{1\cdot 2} + \mu_{*2}(\mathbf{u}_2)}{\nu_{1\cdot 2}-1}\right)(\bar{\boldsymbol{\Omega}}_{11\cdot 2} - \boldsymbol{\delta}_{1\cdot 2}\boldsymbol{\delta}_{1\cdot 2}^{\mathrm{T}}) \\ &\quad + \sigma_*^2(\mathbf{u}_2)\boldsymbol{\delta}_{1\cdot 2}\boldsymbol{\delta}_{1\cdot 2}^{\mathrm{T}}, \quad \nu_{1\cdot 2} > 2. \end{aligned}$$

References

Arellano-Valle, R. B., A. Azzalini. 2006. On the unification of families of skew-normal distributions. *Scand. J. Statist.* **33**(3) 561–574.

Arellano-Valle, R. B., H. Bolfarine. 1995. On some characterizations of the *t* distribution. *Statist. Probab. Lett.* **25**(1) 79–85.

Arellano-Valle, R. B., M. G. Genton. 2010. Multivariate extended skew-*t* distributions and related families. *Metron*. Forthcoming.

Arellano-Valle, R. B., M. D. Branco, M. G. Genton. 2006. A unified view on skewed distributions arising from selections. *Canad. J. Statist.* **34**(4) 581–601.

Azzalini, A. 2005. The skew-normal distribution and related multivariate families (with discussion by Marc G. Genton and a rejoinder by the author). *Scand. J. Statist.* **32**(2) 159–188.

Azzalini, A., A. Capitanio. 1999. Statistical applications of the multivariate skew-normal distribution. *J. Roy. Statist. Soc., Ser. B* **61**(4) 579–602.

Azzalini, A., A. Capitanio. 2003. Distributions generated by perturbation of symmetry with emphasis on an multivariate skew *t* distribution. *J. Roy. Statist. Soc., Ser. B* **65**(2) 367–389.

Azzalini, A., M. G. Genton. 2008. Robust likelihood methods based on the skew-*t* and related distributions. *Internat. Statist. Rev.* **76**(1) 106–129.

Barak, B., K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, K. Talwar. 2007. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. *Proc. Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Sympos. on Principals of Database Systems*, ACM, New York, 273–281.

Branco, M. D., D. K. Dey. 2001. A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.* **79**(1) 99–113.

Branco, M. D., D. K. Dey. 2002. Regression model under skew elliptical error distribution. *J. Math. Sci.* **1** 151–169.

Burridge, J. 2003. Information preserving statistical obfuscation. *Stat. Comp.* **13**(4) 321–327.

Clemen, R. T., T. Reilly. 1999. Correlation and copulas for decision and risk analysis. *Management Sci.* **45**(2) 208–224.

Domingo-Ferrer, J. 2007. A three-dimensional conceptual framework for database privacy. W. Jonker, M. Petković, eds. *Lecture Notes in Computer Science*, Vol. 4721. Springer, Berlin, 193–202.

Fang, K.-T., S. Kotz, K. W. Ng. 1990. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, New York.

Field, C., M. G. Genton. 2006. The multivariate **g**-and-**h** distribution. *Technometrics* **48**(1) 104–111.

Genton, M. G. 2004. *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, edited volume. Chapman and Hall/CRC, Boca Raton, FL.

Healy, M. J. R. 1968. Multivariate normal plotting. *Appl. Statist.* **17**(2) 157–161.

Jones, D. Y., A. Schatzkin, S. B. Green, G. Block, L. A. Brinton, R. G. Ziegler, R. Hoover, P. R. Taylor. 1987. Dietary fat and breast cancer in the National Health and Nutrition Survey I: Epidemiologic follow-up study. *J. National Cancer Inst.* **79**(3) 465–471.

Ma, Y., M. G. Genton. 2004. A flexible class of skew-symmetric distributions. *Scand. J. Statist.* **31**(3) 459–468.

Machanavajjhala, A., D. Kifer, J. Gehrke, M. Venkitasubramaniam. 2007. *l*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowledge Discovery from Data* **1**(1) 1–52.

Marchenko, Y. V., M. G. Genton. 2010. Multivariate log-skew-elliptical distributions with applications to precipitation data. *Environmetrics*. Forthcoming.

Mardia, K. V., J. T. Kent, J. M. Bibby. 1979. *Multivariate Analysis*. Academic Press, London.

Muralidhar, K., R. Sarathy. 2003. A theoretical basis for perturbation methods. *Stat. Comput.* **13**(4) 329–335.

Muralidhar, K., R. Sarathy. 2006. Data shuffling—A new masking approach for numerical data. *Management Sci.* **52**(5) 658–670.

Muralidhar, K., R. Sarathy. 2008. Generating sufficiency-based non-synthetic perturbed data. *Trans. Data Privacy* **1**(1) 17–33.

Muralidhar, K., D. Batra, P. J. Kirs. 1995. Accessibility, security, and accuracy in statistical databases: The case for the multiplicative fixed data perturbation approach. *Management Sci.* **41**(9) 1549–1564.

Muralidhar, K., R. Parsa, R. Sarathy. 1999. A general additive data perturbation method for database security. *Management Sci.* **45**(10) 1399–1415.

Muralidhar, K., R. Sarathy, R. Parsa. 2001. An improved security requirement for data perturbation with implications for e-commerce. *Decision Sci.* **32**(4) 683–698.

R Development Core Team. 2008. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Raghunathan, T. E., J. P. Reiter, D. B. Rubin. 2003. Multiple imputation for statistical disclosure limitation. *J. Official Statist.* **19**(1) 1–16.

Reiter, J. P. 2003. Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**(2) 181–188.

Rubin, D. B. 1993. Discussion: Statistical disclosure limitation. *J. Official Statist.* **9**(2) 461–468.

Sarathy, R., K. Muralidhar, R. Parsa. 2002. Perturbing nonnormal confidential attributes: The copula approach. *Management Sci.* **48**(12) 1613–1627.