Acta Appl Math (2008) 102: 99–118
DOI 10.1007/s10440-008-9212-8

Estimation Methods for the Multivariate *t* Distribution

Saralees Nadarajah · Samuel Kotz

Received: 5 January 2007 / Accepted: 18 January 2008 / Published online: 1 February 2008

Abstract The known estimation and simulation methods for multivariate t distributions are reviewed. A review of selected applications is also provided. We believe that this review will serve as an important reference and encourage further research activities in the area.

Keywords Multivariate t distribution · Multivariate normal distribution · Estimation methods

1 Introduction

A p-dimensional random vector $\mathbf{X}^T = (X_1, \ldots, X_p)$ is said to have the t distribution with degrees of freedom ν, mean vector $\boldsymbol{\mu}$ and correlation matrix $\mathbf{R}$ if its joint pdf is given by:

$$f(\mathbf{x}) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\mathbf{R}|^{1/2}}\left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}. \tag{1}$$

The degrees of freedom parameter ν is also referred to as the shape parameter, as the peakedness of (1) may be diminished, preserved or increased by varying ν [14]. The distribution is said to be central if $\boldsymbol{\mu} = \mathbf{0}$. Note that if $p = 1$, $\boldsymbol{\mu} = 0$ and $\mathbf{R} = 1$ then (1) reduces to the univariate Student's t distribution. If $p = 2$, then (1) is a slight modification of the bivariate surface of Pearson [43]. If $\nu = 1$, then (1) is the p-variate Cauchy distribution. If $(\nu + p)/2 = m$, an integer, then (1) is the p-variate Pearson type VII distribution. The limiting form of (1) as $\nu \to \infty$ is the joint pdf of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. The particular case of (1) for $\boldsymbol{\mu} = 0$ and $\mathbf{R} = \mathbf{I}_p$ is a mixture of the normal density with zero means and covariance matrix $v\mathbf{I}_p$—in the scale parameter v.

S. Nadarajah (✉)
School of Mathematics, University of Manchester, Manchester M60 1QD, UK
e-mail: saralees.nadarajah@manchester.ac.uk

S. Kotz
Department of Engineering Management and Systems Engineering, George Washington University,
Washington, 20052, USA

Multivariate t distributions are of increasing importance in classical as well as in Bayesian statistical modeling; however, relatively little is known by means of mathematical properties or statistical methods. These distributions have been perhaps unjustly overshadowed by the multivariate normal distribution. Both the multivariate t and the multivariate normal are members of the general family of elliptically symmetric distributions. However, we feel that it is desirable to focus on these distributions separately for several reasons:

- Multivariate t distributions are generalizations of the classical univariate Student t distribution, which is of central importance in statistical inference. The possible structures are numerous, and each one possesses special characteristics as far as potential and current applications are concerned.
- Application of multivariate t distributions is a very promising approach in multivariate analysis. Classical multivariate analysis is soundly and rigidly tilted toward the multivariate normal distribution while multivariate t distributions offer a more viable alternative with respect to real-world data, particularly because its tails are more realistic. We have seen recently some unexpected applications in novel areas such as cluster analysis, discriminant analysis, multiple regression, robust projection indices, and missing data imputation.
- Multivariate t distributions for the past 20 to 30 years have played a crucial role in Bayesian analysis of multivariate data. They serve by now as the most popular prior distribution (because elicitation of prior information in various physical, engineering, and financial phenomena is closely associated with multivariate t distributions) and generate meaningful posterior distributions.

In this paper, we provide a comprehensive review of estimation and simulation methods for multivariate t distributions. A review of selected applications is also provided. We believe that this review will serve as an important reference and encourage further research activities in the area.

2 Tiku and Kambo's Estimation Procedure

Suppose (X_1, X_2) has the bivariate normal distribution with means (μ_1, μ_2), variances (σ_1^2, σ_2^2), and correlation coefficient ρ. Its joint pdf can be factorized as

$$f(x_1, x_2) = f(x_1 \mid x_2) f(x_2),$$

where

$$f(x_1 \mid x_2) = \frac{1}{\sigma_1 \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_1^2 (1-\rho^2)} \left\{ x_1 - \mu_1 - \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2)^2 \right\} \right] \tag{2}$$

and

$$f(x_2) \propto \frac{1}{\sigma_2} \exp\left\{ -\frac{1}{2\sigma_2^2} (x_2 - \mu_2)^2 \right\}. \tag{3}$$

Numerous nonnormal distributions can be generated by replacing either $f(x_1 \mid x_2)$ and/or $f(x_2)$ by nonnormal distributions. Tiku and Kambo [56] studied the family of symmetric bivariate distributions obtained by replacing (3) by the Student's t density

$$f(x_2) \propto \frac{1}{\sqrt{k}\sigma_2} \left\{ 1 + \frac{(x_2 - \mu_2)^2}{k\sigma_2^2} \right\}^{-\nu}, \tag{4}$$

which yields the joint pdf

$$f(x_1,x_2)=\frac{1}{\sigma_1\sigma_2\sqrt{k(1-\rho^2)}}\left\{1+\frac{(x_2-\mu_2)^2}{k\sigma_2^2}\right\}^{-\nu}$$
$$\times\exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\left\{x_1-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)^2\right\}\right], \tag{5}$$

where $k=2\nu-3$ and $\nu\geq 2$. This is motivated by the fact that in many applications it is reasonable to assume that the difference $Y_1-\mu_1-\rho(\sigma_1/\sigma_2)(Y_2-\mu_2)$ is normally distributed and the regression of Y_1 on Y_2 is linear. Moreover, in numerous applications Y_2 represents time-to-failure with a distribution [11, 56], which might be symmetric but is not normally distributed. Besides, most types of time-to-failure data are such that a transformation cannot be performed to impose normality on the underlying distribution.

Here, we discuss estimation of the parameters μ_1, μ_2, σ_1, σ_2, and ρ when ν is known. The method for estimating the location and scale parameters developed by Tiku and Suresh [57] is used for this problem. For a random sample $\{(X_{1i},X_{2i}),i=1,\dots,n\}$ from (5), the likelihood function is

$$L\propto\left\{\sigma_1^2\sigma_2^2(1-\rho^2)\right\}^{-n/2}\prod_{i=1}^{n}\left\{1+\frac{X_{(2:i)}-\mu_2}{k\sigma_2^2}\right\}^{-\nu}$$
$$\times\exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\sum_{i=1}^{n}\left\{X_{[1:i]}-\mu_1-\frac{\rho\sigma_1}{\sigma_2}(X_{(2:i)}-\mu_2)\right\}^2\right],$$

where $k=2\nu-3$, $X_{(2:i)}$, $i=1,\dots,n$ are the order statistics of X_{2i} and $X_{[1:i]}$, $i=1,\dots,n$ are the corresponding concomitant X_1 observations. Consider the following three situations:

(i) Complete samples are available and ν is not too small ($\nu>3$).

(ii) Complete samples are available but ν is small ($\nu\leq 3$).

(iii) A few smallest or a few largest X_{2i} observations and the corresponding concomitant $X_{[1:i]}$ are censored due to the constraints of an experiment. This situation arises in numerous practical situations. In a time mortality experiment, for example, n mice are inoculated with a uniform culture of human tuberculosis. What is recorded is X_{2i}: the time to death of the first $A(<n)$ mice, and X_{1i}: the corresponding weights at the time of death.

These situations also arise in the context of ranking and selection [5]. We provide some details of the inference for situation (i) as described in Tiku and Kambo [56]. Using a linear approximation of the likelihood based on the expected values of order statistics, it is shown that the maximum likelihood estimators are

$$\widehat{\mu}_1=\bar{x}_1-\frac{\widehat{\rho}\,\widehat{\sigma}_1}{\widehat{\sigma}_2}(\bar{x}_2-\mu_2),$$

$$\widehat{\sigma}_1=\sqrt{s_1^2+\frac{s_{12}^2}{s_2^2}\left(\frac{\widehat{\sigma}_2^2}{s_2^2}-1\right)},$$

$$\widehat{\mu}_2=\bar{x}_2-\frac{\widehat{\rho}\,\widehat{\sigma}_2}{\widehat{\sigma}_1}(\bar{x}_1-\mu_1),$$

$$\widehat{\sigma}_2 = \sqrt{s_2^2 + \frac{s_{12}^2}{s_1^2}\left(\frac{\widehat{\sigma}_1^2}{s_1^2} - 1\right)},$$

and

$$\widehat{\rho} = \frac{s_{12}}{s_2^2}\frac{\widehat{\sigma}_2}{\widehat{\sigma}_1},$$

where $(\bar{x}_1, \bar{x}_2)$ are the usual sample means, (s_1^2, s_2^2) are the usual sample variances, and s_{12} is the sample covariance. The estimators $\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\sigma}_1$, $\widehat{\sigma}_2$, and $\widehat{\rho}$ are asymptotically unbiased and minimum variance bound estimators. The estimator $\widehat{\sigma}_1^2$ is always real and positive while the estimator $\widehat{\rho}$ always assumes values between -1 and 1. The asymptotic variances and covariances of the estimators can be written as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{pmatrix},$$

where $\mathbf{V}_1$ is positive definite and is the asymptotic variance-covariance matrix of $(\widehat{\mu}_1, \widehat{\mu}_2)$ while $\mathbf{V}_2$ is positive definite and is the asymptotic variance-covariance matrix of $(\widehat{\sigma}_1, \widehat{\sigma}_2, \widehat{\rho})$. The asymptotic distribution of $\sqrt{n}(\widehat{\mu}_1 - \mu_1, \widehat{\mu}_2 - \mu_2)$ is bivariate normal with zero means and variance-covariance matrix $n\mathbf{V}_1$. For testing $H_0 : (\mu_1, \mu_2) = (0, 0)$ versus $H_1 : (\mu_1, \mu_2) \neq (0, 0)$, a useful statistic is $T_p^2 = (\widehat{\mu}_1, \widehat{\mu}_2)^T \widehat{\mathbf{V}}_1^{-1}(\widehat{\mu}_1, \widehat{\mu}_2)$, the asymptotic null distribution of which is chi-squared with degrees of freedom 2. Tiku and Kambo [56] also provided evidence to the fact that the use of T_p^2 in place of the Hotelling's T^2 statistic can result in a substantial gain in power.

3 ML Estimation via EM Algorithm

Consider fitting a p-variate t distribution to data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with the log-likelihood function

$$L(\boldsymbol{\mu}, \mathbf{R}, \nu) = -\frac{n}{2} \log |\mathbf{R}| - \frac{\nu + p}{2} \sum_{i=1}^{n} \log(\nu + s_i), \tag{6}$$

where $s_i = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ and ν is assumed to be fixed. Differentiating (6) with respect to $\boldsymbol{\mu}$ and $\mathbf{R}$ leads to the estimating equations

$$\boldsymbol{\mu} = \text{ave}\{w_i \mathbf{x}_i\} / \text{ave}\{w_i\} \tag{7}$$

and

$$\mathbf{R} = \text{ave}\left\{w_i (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right\}, \tag{8}$$

where $w_i = (\nu + p)/(\nu + s_i)$ and "ave" stands for the arithmetic average over $i = 1, 2, \ldots, n$. Note that equations (7)–(8) can be viewed as an adaptively weighted sample mean and sample covariance matrix where the weights depend on the Mahalanobis distance between $\mathbf{x}_i$ and $\boldsymbol{\mu}$. The weight function $w(s) = (\nu + p)/(\nu + s)$, where $s = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{x} - \boldsymbol{\mu})$, is a decreasing function of s, so that the outlying observations are downweighted. Maronna [38] proved, under general assumptions, the existence, uniqueness, consistency, and asymptotic normality of the solutions of (7)–(8). For instance, if there exists $a > 0$ such that, for every

Table 1 Primary references for EM type algorithms

Algorithm	Primary references
Extended EM	Kent et al. (1994), Arsian et al. (1995)
Restricted EM	Arsian et al. (1995)
MC-ECM1	Liu and Rubin (1995)
MC-ECM2	Liu and Rubin (1995), Meng and van Dyk (1997)
ECME1	Liu and Rubin (1995), Liu (1997)
ECME2	Liu and Rubin (1995)
ECME3	Liu and Rubin (1995), Meng and van Dyk (1997)
ECME4	Liu and Rubin (1995), Liu (1997)
ECME5	Liu (1997)
PXEM	Liu et al. (1998)

hyperplane H, $\Pr(H) \leq p/(\nu + p) - a$, then (7)–(8) has a unique solution. Also, every solution satisfies the consistency property that $\lim_{n\to\infty}(\widehat{\boldsymbol{\mu}}, \widehat{\mathbf{R}}) = (\boldsymbol{\mu}, \mathbf{R})$ with probability 1.

The standard approach for solving (7)–(8) for $\boldsymbol{\mu}$ and $\mathbf{R}$ is the popular EM algorithm because of its simplicity and stable convergence [6, 59]. The EM algorithm takes the form of iterative updates of (7)–(8), using the current estimates of $\boldsymbol{\mu}$ and $\mathbf{R}$ to generate the weights. The iterations take the form

$$\boldsymbol{\mu}^{(m+1)} = \text{ave}\left\{w_i^{(m)}\mathbf{x}_i\right\}\Big/\text{ave}\left\{w_i^{(m)}\right\}$$

and

$$\mathbf{R}^{(m+1)} = \text{ave}\left\{w_i^{(m)}\left(\mathbf{x}_i - \boldsymbol{\mu}^{(m+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}^{(m+1)}\right)^T\right\},$$

where

$$w_i^{(m)} = (\nu + p)\Big/\left\{\nu + \left(\mathbf{x}_i - \boldsymbol{\mu}^{(m)}\right)^T\left(\mathbf{R}^{(m)}\right)^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}^{(m)}\right)\right\}.$$

This is known as the direct EM algorithm and is valid for any $\nu > 0$. For details of this algorithm see the pioneering papers of Dempster et al. [6, 7], Rubin [46], and Little and Rubin [32]. Several variants of the above have been proposed in the literature, as summarized in Table 1.

Consider the maximum likelihood (ML) estimation for a g-component mixture of t distributions given by

$$f(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\mathbf{x}; \boldsymbol{\mu}_i, \mathbf{R}_i, \nu_i),$$

where

$$f(\mathbf{x}; \boldsymbol{\mu}_i, \mathbf{R}_i, \nu_i) = \frac{\Gamma((\nu_i + p)/2)}{(\pi\nu_i)^{p/2}\Gamma(\nu_i/2)|\mathbf{R}_i|^{1/2}}\left[1 + \frac{(\mathbf{x} - \boldsymbol{\mu}_i)^T\mathbf{R}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)}{\nu_i}\right]^{-(\nu_i+p)/2},$$

$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}^T, \boldsymbol{\nu}^T)^T$, $\boldsymbol{\theta} = ((\boldsymbol{\mu}_1, \mathbf{R}_1)^T, \ldots, (\boldsymbol{\mu}_g, \mathbf{R}_g)^T)^T$, and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_g)^T$. The application of the EM algorithm for this model in a clustering context has been considered by McLachlan and Peel [39] and Peel and McLachlan [44]. The iteration updates now take the form

$$\boldsymbol{\mu}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ij}^{(m)} u_{ij}^{(m)} \mathbf{x}_j \Big/ \sum_{j=1}^{n} \tau_{ij}^{(m)} u_{ij}^{(m)}$$

and

$$\mathbf{R}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ij}^{(m)} u_{ij}^{(m)} \left(\mathbf{x}_j - \boldsymbol{\mu}_i^{(m+1)}\right)\left(\mathbf{x}_j - \boldsymbol{\mu}_i^{(m+1)}\right)^T \Big/ \sum_{j=1}^{n} \tau_{ij}^{(m)},$$

where

$$u_{ij}^{(m)} = \frac{\nu_i^{(m)} + p}{\nu_i^{(m)} + (\mathbf{x}_j - \boldsymbol{\mu}_i^{(m)})^T \mathbf{R}_i^{(m)^{-1}} (\mathbf{x}_j - \boldsymbol{\mu}_i^{(m)})}$$

and

$$\tau_{ij}^{(m)} = \frac{\pi_i^{(m)} f(\mathbf{x}_j; \boldsymbol{\mu}_i^{(m)}, \mathbf{R}_i^{(m)}, \nu_i^{(m)})}{f(\mathbf{x}_j; \boldsymbol{\Psi}^{(m)})}.$$

The EMMIX program of McLachlan et al. [40] for the fitting of normal mixture models has an option that implements the above procedure for the fitting of mixtures of t-components. The program automatically generates a selection of starting values for the fitting if they are not provided by the user. The user only has to provide the data set, the restrictions on the component-covariance matrices (equal, unequal, diagonal), the extent of the selection of the initial groupings to be used to determine the starting values, and the number of components that are to be fitted. The program is available from the software archive StatLib or from Professor Peel's homepage at the Web site address http://www.maths.uq.edu.au/~gjm/.

4 Missing Data Imputation

When a data set contains missing values, multiple imputation for missing data [47] appears to be an ideal technique. Most importantly, it allows for valid statistical inferences. In contrast, any single imputation method, such as filling in the missing values with either their marginal means or their predicted values from linear regression, typically leads to biased estimates of parameters and thereby often to an invalid inference [47, pp. 11–15].

The multivariate normal distribution has been a popular statistical model in practice for rectangular continuous data sets. To impute the missing values in an incomplete normal data set, Rubin and Schafer [48] (see also [49], and [33]) proposed an efficient method, called monotone data augmentation (MDA), and implemented it using the factorized likelihood approach. A more efficient technique to implement the MDA than the factorized likelihood approach is provided by Liu [33] using Bartlett's decomposition, which is the extension of the Bayesian version of Bartlett's decomposition of the Wishart distribution with complete rectangular normal data to the case with monotone ignorable missing data.

When a rectangular continuous data set appears to have longer tails than the normal distribution, or it contains some values that are influential for statistical inferences with the normal distribution, the multivariate t distribution becomes useful for multiple imputation as an alternative to the multivariate normal distribution. First, when the data have longer tails than the normal distribution, the multiply imputed data sets using the t distribution allow more valid statistical inferences than those using the normal distribution with some "influential" observations deleted. Second, it is well known that the t distribution is widely used in applied statistics for robust statistical inferences. Therefore, when an incomplete data set contains some influential values or outliers, the t distribution allows for a robust multiple imputation method. Furthermore, the multiple imputation appears to be more useful than the asymptotic method of inference since the likelihood functions of the parameters of the t distribution given the observed data can have multiple modes. For a complete description of the MDA using the multivariate t distribution, see [34]. See also [35] for extensions in two aspects, including covariates in the multivariate t models (as in [36]), and replacing the multivariate t distribution with a more general class of distributions, that is, the class of normal/independent distributions (as in [27]). These extensions provide a flexible class of models for robust multivariate linear regression and multiple imputation. Liu [35] described methods to implement the MDA for these models with fully observed predictor variables and possible missing values from outcome variables.

5 Laplacian T-Approximation

The Laplacian T-approximation [51] is a useful tool for Bayesian inferences for variance component models. Let $p(\boldsymbol{\theta} \mid \mathbf{y})$ be the posterior pdf of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ given data $\mathbf{y}$, and let $\eta = g(\boldsymbol{\theta})$ be the parameter of interest. Leonard et al. [30] introduced a Laplacian T-approximation for the marginal posterior of η of the form

$$p^*(\eta \mid \mathbf{y}) \propto |\mathbf{T}_\eta|^{-1/2} p(\boldsymbol{\theta}_\eta \mid \mathbf{y}) \lambda_\eta^{-w/2} f(\eta \mid w, \boldsymbol{\theta}_\eta^*, \mathbf{T}_\eta) \tag{9}$$

to be the marginal posterior pdf of η, where $f(\eta \mid w, \boldsymbol{\theta}_\eta^*, \mathbf{T}_\eta)$ denotes the pdf of $\eta = g(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ possesses a multivariate t distribution with mean vector $\boldsymbol{\theta}_\eta^*$, covariance matrix $\mathbf{T}_\eta$, and degrees of freedom w. Here, $\boldsymbol{\theta}_\eta$ represents some convenient approximation to the conditional posterior mean vector of $\boldsymbol{\theta}$, given η, and w should be taken to roughly approximate the degrees of freedom of a generalized multivariate T-approximation to the conditional distribution of $\boldsymbol{\theta}$ given η.

When $\boldsymbol{\theta}_\eta$ is the conditional posterior mode vector of $\boldsymbol{\theta}$, given η, (9) reduces to the Laplacian approximation introduced by Leonard [29] and shown by Tierney and Kadane [55] and Leonard et al. [31] to possess saddlepoint accuracy as well as an excellent finite-sample accuracy, in many special cases. It was previously used for hierarchical models by Kass and Steffey [24].

6 Sutradhar's Score Test

Consider a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a p-variate t distribution with the pdf

$$f(\mathbf{x}_j) = \frac{(\nu-2)^{\nu/2}\Gamma((\nu+p)/2)}{\pi^{p/2}\Gamma(\nu/2)|\mathbf{R}|^{1/2}} \left[\nu - 2 + (\mathbf{x}_j - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})\right]^{-(\nu+p)/2}.$$

Note this is a slight reparameterization of the usual t pdf. The log-likelihood

$$G = \sum_{j=1}^{n} \log f(\mathbf{x}_j)$$

is a function of the parameters $\mathbf{R}$, $\boldsymbol{\mu}$, and ν.

Frequently in social sciences, and particularly in factor analysis, one of the main inference problems is to test the null hypothesis $\mathbf{R} = \mathbf{R}_0$ when $\boldsymbol{\mu}$ and ν are known. Sutradhar [53] proposed Neyman's [42] score test for this test for large n. Le $\mathbf{r} = (r_{11}, \dots, r_{hl}, \dots, r_{pp})^T$ be the $p(p+1)/2 \times 1$ vector formed by stacking the distinct elements of $\mathbf{R}$, with r_{hl} being the (h, l)th element of the $p \times p$ matrix $\mathbf{R}$. Also let

$$(\lambda_1, \dots, \lambda_i, \dots, \lambda_{p(p+1)/2})^T \equiv b(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu})$$

and

$$(\gamma_1, \dots, \gamma_j, \dots, \lambda_{p+1})^T \equiv \begin{bmatrix} \xi(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}) \\ \eta(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}) y \end{bmatrix},$$

where $b(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu})$, $\xi(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu})$, and $\eta(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu})$ are the score functions obtained under the null hypothesis $\mathbf{r} = \mathbf{r}_0$, by replacing $\boldsymbol{\mu}$ and ν with their consistent estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\nu}$ in

$$b(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}) = \frac{\partial G}{\partial \mathbf{r}}, \tag{10}$$

$$\xi(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}) = \frac{\partial G}{\partial \boldsymbol{\mu}}, \tag{11}$$

and

$$\eta(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}) = \frac{\partial G}{\partial \nu}, \tag{12}$$

respectively. Furthermore, let $T_i(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}) = \lambda_i - \sum_{j=1}^{p+1} \beta_{ij} \gamma_j$, where β_{ij} is the partial regression coefficient of λ_i on γ_j. Then, Neyman's partial score test statistic is given by

$$W(\widehat{\boldsymbol{\mu}}, \widehat{\nu}) = \mathbf{T}^T \left[\widehat{\mathbf{M}}_{11} - \left(\widehat{\mathbf{M}}_{12} \widehat{\mathbf{M}}_{13} \right) \begin{pmatrix} \widehat{\mathbf{M}}_{22} & \widehat{\mathbf{M}}_{23} \\ & \widehat{\mathbf{M}}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\mathbf{M}}_{12}^T \\ \widehat{\mathbf{M}}_{13}^T \end{pmatrix} \right]^{-1} \mathbf{T}, \tag{13}$$

where $\mathbf{T} \equiv [T_1(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu}), \dots, T_{p(p+1)/2}(\mathbf{r}_0, \widehat{\boldsymbol{\mu}}, \widehat{\nu})]^T$ for $i, r = 1, 2, 3$; $\widehat{\mathbf{M}}_{ir}$ are obtained from $\mathbf{M}_{ir} = E(-D_{ir})$ by replacing $\boldsymbol{\mu}$ and ν with their consistent estimates; and $\mathbf{D}_{ir}$ for i, $r = 1, 2, 3$ are the derivatives given by

$$\mathbf{D}_{11} = \frac{\partial^2 G}{\partial \mathbf{r} \partial \mathbf{r}'},$$

$$\mathbf{D}_{12} = \frac{\partial^2 G}{\partial \mathbf{r} \partial \boldsymbol{\mu}'},$$

$$\mathbf{D}_{13} = \frac{\partial^2 G}{\partial \mathbf{r} \partial \nu},$$

$$\mathbf{D}_{22} = \frac{\partial^2 G}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'},$$

$$\mathbf{D}_{23} = \frac{\partial^2 G}{\partial \boldsymbol{\mu} \partial \nu},$$

and

$$\mathbf{D}_{33} = \frac{\partial^2 G}{\partial \nu^2}.$$

Under the null hypothesis $\mathbf{r} = \mathbf{r}_0$, the test statistic $W(\widehat{\boldsymbol{\mu}}, \widehat{\nu})$ has an approximate chi-squared distribution with degrees of freedom $p(p+1)/2$. The test based on (13) is asymptotically locally most powerful. Clearly the implementation of this test requires consistent estimates of $\widehat{\boldsymbol{\mu}}, \widehat{\nu}$ as well as expressions for the score functions and the information matrix. The maximum likelihood estimates of $\boldsymbol{\mu}$ and ν are obtained by simultaneously solving

$$\widehat{\boldsymbol{\mu}} = \sum_{j=1}^{n} q_j^{-1} \mathbf{X}_j \Big/ \sum_{j=1}^{n} q_j^{-1}$$

and

$$\eta(\widehat{\boldsymbol{\mu}}, \mathbf{r}_0, \widehat{\nu}) = 0,$$

where $q_j = \widehat{\nu} - 2 + (\mathbf{X}_j - \widehat{\boldsymbol{\mu}})^T \mathbf{R}_0 (\mathbf{X}_j - \widehat{\boldsymbol{\mu}})$ and $\mathbf{R}_0$ is specified by the null hypothesis. The moment estimates of $\boldsymbol{\mu}$ and ν (which also turn out to be consistent) are

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{X}_j$$

and

$$\widehat{\nu} = \frac{2\{2\widehat{\beta}_2 - f(\mathbf{r}_0)\}}{\widehat{\beta}_2 - f(\mathbf{r}_0)},$$

where

$$\widehat{\beta}_2 = \frac{1}{n} \sum_{j=1}^{n} \left[(\mathbf{X}_j - \bar{\mathbf{X}})^T \mathbf{R}_0 (\mathbf{X}_j - \bar{\mathbf{X}}) \right]^2$$

is a consistent estimator of the multivariate measure of skewness (see, for example, [37]), and

$$f(\mathbf{r}_0) = 3 \sum_{h=1}^{p} \left(r_0^{hh}\right)^2 \left\{r_{hh}^{(0)}\right\}^2 + \sum_{h \neq h'}^{p} \left\{r_{hh}^{(0)}\right\}^2 \left\{r_0^{hh} r_0^{h'h'} + \left(r_0^{hh'}\right)^2\right\},$$

where $r_{hh'}^{(0)}$ and $r_0^{hh'}$ denote the (h, h')th element of $\mathbf{R}_0$ and $\mathbf{R}_0^{-1}$, respectively.

7 Multivariate t Model

Joarder and Ahmed [17] and others found it more instructive to consider dependent but uncorrelated t distributions and suggested the model:

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\Gamma((\nu+p)/2)}{(\pi^n \nu)^{p/2} \Gamma(\nu/2) |\mathbf{R}|^{n/2}} \left[1 + \frac{1}{\nu} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]^{-(\nu+np)/2}, \quad (14)$$

which they referred to as the *multivariate t model*. Among others, Zellner [60] and Sutradhar and Ali [54] considered (14) in the context of stock market problems. By successive integration, one can show that the marginal distribution of $\mathbf{X}_j$ in the multivariate t model (14) is

p-variate t. It also follows from (14) that $E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_l - \boldsymbol{\mu}) = 0$ for $j \neq l$. Thus, in (14), although $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are pairwise uncorrelated, they are not necessarily independent. More specifically, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in (14) are not independent if $\nu < \infty$, since independence would imply that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are normally distributed. The case of independent normally distributed random vectors can be included in (14) by letting $\nu \to \infty$. In the case $\nu = 1$, (14) is the multivariate Cauchy distribution for which neither the mean nor the variance exists. Kelejian and Prucha [25] proved that (14) is better able to capture heavy-tailed behavior than an independent t model.

In this section, we consider estimation issues associated with the correlation matrix $\mathbf{R}$ and its trace $\operatorname{tr}(\mathbf{R})$.

7.1 Estimation of $\mathbf{R}$

Joarder and Ali [19] developed estimators of $\mathbf{R}$ (when the mean vector $\boldsymbol{\mu}$ is unknown) under the entropy loss function

$$L(u(\mathbf{A}), \mathbf{R}) = \operatorname{tr}\big(\mathbf{R}^{-1} u(\mathbf{A})\big) - \log\big|\mathbf{R}^{-1} u(\mathbf{A})\big| - p,$$

where $u(\mathbf{A})$ is any estimator of $\mathbf{R}$ based on the Wishart matrix $\mathbf{A}$ defined by

$$\mathbf{A} = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T. \tag{15}$$

Based on the form of the likelihood function, the entropy loss function has been suggested in the literature by James and Stein [13] and is sometimes known as the Stein loss function. Some important features of the entropy loss function are that it is zero if the estimator $u(\mathbf{A})$ equals the parameter $\mathbf{R}$, positive when $u(\mathbf{A}) \neq \mathbf{R}$, and invariant under translation as well as under a natural group of transformations of covariance matrices. Moreover, the loss function approaches infinity as the estimator approaches a singular matrix or when one or more elements (or one or more latent roots) of the estimator approaches infinity. This means that gross underestimation is penalized just as heavily as gross overestimation.

In estimating $\mathbf{R}$ by $u(\mathbf{A})$, Joarder and Ali [20] considered the risk function $R(u(\mathbf{A}), \mathbf{R}) = E[L(u(\mathbf{A}), \mathbf{R})]$. An estimator $u_2(\mathbf{A})$ of $\mathbf{R}$ will be said to dominate another estimator $u_1(\mathbf{A})$ of $\mathbf{R}$ if, for all $\mathbf{R}$ belonging to the class of positive definite matrices, the inequality $R(u_2(\mathbf{A}), \mathbf{R}) \leq R(u_1(\mathbf{A}), \mathbf{R})$ holds and the inequality $R(u_2(\mathbf{A}), \mathbf{R}) < R(u_1(\mathbf{A}), \mathbf{R})$ holds for at least one $\mathbf{R}$.

Joarder and Ali [20] obtained three estimators for $\mathbf{R}$, by minimizing the risk function of the entropy loss function among three classes of estimators.

- First, it is shown that the unbiased estimator $\widetilde{\mathbf{R}} = (\nu - 2)\mathbf{A}/(\nu n)$ has the smallest risk among the class of estimators of the form $c\mathbf{A}$, where $c > 0$, and the corresponding minimum risk is given by

$$R(\widetilde{\mathbf{R}}, \mathbf{R}) = p \log n - \sum_{i=1}^{n} E\big[\log\big(\chi^2_{n+i-i}\big)\big] + p \log\left(\frac{\nu}{\nu - 2}\right) - 2pE(\log \tau),$$

where τ has the inverted gamma distribution with the pdf

$$h(\tau) = \frac{2\tau^{-(1+\nu)}}{(2/\nu)^{\nu/2}\Gamma(\nu/2)} \exp\left(-\frac{\nu}{2\tau^2}\right) \tag{16}$$

for $\tau > 0$.

- Second, the estimator $\mathbf{R}^* = \mathbf{T}\mathbf{D}^*\mathbf{T}^T$, where $\mathbf{T}$ is a lower triangular matrix such that $\mathbf{A} = \mathbf{T}\mathbf{T}^T$ and $\mathbf{D}^* = \operatorname{diag}(d_1^*, \ldots, d_p^*)$ with d_i^* defined by

  $$d_i^* = \frac{\nu - 2}{\nu} \frac{1}{n + p + 1 - 2i},$$

  has the smallest risk among the class of estimators $\mathbf{T}\boldsymbol{\Delta}\mathbf{T}^T$, where $\boldsymbol{\Delta}$ belongs to the class of all positive definite diagonal matrices and the corresponding minimum risk function of $\mathbf{R}^*$ is given by

  $$R(\mathbf{R}^*, \mathbf{R}) = \sum_{i=1}^{p} \log(n + 1 + p - 2i) - \sum_{i=1}^{p} E\left[\log\left(\chi^2_{n+1-i}\right)\right] + p \log\left(\frac{\nu}{\nu - 2}\right) \\ - 2pE(\log \tau),$$

  where τ is as defined above. Furthermore, $\mathbf{R}^*$ dominates the unbiased estimator $\widetilde{\mathbf{R}} = (\nu - 2)\mathbf{A}/(\nu n)$.
- Finally, consider the estimator $\widehat{\mathbf{R}} = \mathbf{S}\phi(\mathbf{M})\mathbf{S}$, where $\mathbf{A}$ has the spectral decomposition $\mathbf{A} = \mathbf{S}\mathbf{M}\mathbf{S}^T$, with $\phi(\mathbf{M}) = \mathbf{D}^*\mathbf{M}$. Then the estimator $\widehat{\mathbf{R}} = \mathbf{S}\mathbf{D}^*\mathbf{M}\mathbf{S}^T$ dominates the estimator $\mathbf{R}^* = \mathbf{T}\mathbf{D}^*\mathbf{T}^T$.

7.2 Estimation of tr($\mathbf{R}$)

Let $\delta = \operatorname{tr}(\mathbf{R})$ denote the trace of $\mathbf{R}$. Joarder [15] considered the estimation of δ for the multivariate t model under a squared error loss function following Dey [8]. The usual estimator of δ is given by $\widetilde{\delta} = c_0 \operatorname{tr}(\mathbf{A})$, where c_0 is a known positive constant and $\mathbf{A}$ is the Wishart matrix defined in (15). The estimator $\widetilde{\delta}$ defines an unbiased estimator of δ for $c_0 = (\nu - 2)/(\nu n)$ and a maximum likelihood estimator of $\widetilde{\delta}$ for $c_0 = 1/(n + 1)$ (see, for example, Anderson and Fang, [1, p. 208]). Joarder and Singh [21] proposed an improved estimator of δ—based on a power transformation—given by

$$\widehat{\delta} = c_0 tr(\mathbf{A}) + c_0 c\left\{p|\mathbf{A}|^{1/p} - \operatorname{tr}(\mathbf{A})\right\},$$

where c_0 is a known positive constant and c is a constant chosen so that the mean square error (MSE) of $\widehat{\delta}$ is minimized. Calculations show that

$$MSE\left(\widehat{\delta}\right) = MSE\left(\widetilde{\delta}\right) + c\beta_1 + c^2\beta_2,$$

where

$$\beta_1 = 2c_0^2 E\left[(c_0 \operatorname{tr}(\mathbf{A}) - \delta)\left(p|\mathbf{A}|^{1/p} - \operatorname{tr}(\mathbf{A})\right)\right] \tag{17}$$

and

$$\beta_2 = c_0^2 E\left[p\,|\mathbf{A}|^{1/p} - \operatorname{tr}(\mathbf{A})\right]. \tag{18}$$

Thus $MSE(\widehat{\delta})$ is minimized at $c = -\beta_1/(2\beta_2)$ and the minimum value is given by $MSE(\widetilde{\delta}) - \beta_1^2/(4\beta_2)$. This proves that $\widehat{\delta}$ is always better than the usual estimator in the sense of having a smaller MSE. The estimate of c is given by $\widehat{c} = -\widehat{\beta_1}/(2\widehat{\beta_2})$, where $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are obtained by calculating the expectations in (17) and (18) using the numerous properties given in [16, 19] and then replacing $\mathbf{R}$ by the usual estimator $c_0\mathbf{A}$. It can be noted from Anderson and

Table 2 Percent relative efficiencies

ν	$\mathbf{R}=\mathrm{diag}(1,1,1)$	$\mathbf{R}=\mathrm{diag}(4,2,1)$	$\mathbf{R}=\mathrm{diag}(25,1,1)$
5	105.32	130.31	153.90
10	102.13	117.56	148.76
15	101.53	112.07	127.15

Fang ([1, p. 208]) that the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the maximum likelihood estimators of β_1 and β_2, respectively, provided $\mathbf{R}=c_0\mathbf{A}$ and $c_0=1/(n+1)$.

Table 2 taken from [21] presents the percent relative efficiency of $\widehat{\delta}$ over $\widetilde{\delta}$. The numbers are from a Monte Carlo study carried out by generating 100 Wishart matrices from the multivariate t-model with $n=25$ and $p=3$.

8 Generalized Multivariate *t* Model

Joarder and Ahmed [18] considered a generalization of the multivariate t-model in (14) when the random sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ is assumed to come from a p-variate elliptical distribution with the joint pdf

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_n)=\int_0^\infty \frac{|\tau^2\mathbf{R}|^{-n/2}}{(2\pi)^{np/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol{\mu})^T\left(\tau^2\mathbf{R}\right)^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right\}h(\tau)d\tau, \tag{19}$$

where $h(\cdot)$ is the pdf of a non-discrete random variable τ. Many multivariate distributions having constant pdf on the hyper-ellipse $(\mathbf{x}-\boldsymbol{\mu})^T\ \mathbf{R}^{-1}\ (\mathbf{x}-\boldsymbol{\mu})=c^2$ may be generated by varying $h(\cdot)$. The observations $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are independent only if τ is degenerate at the point unity in which case the joint pdf (19) denotes the pdf of the product of n independent p-variate normal distributions each being $N_p(\boldsymbol{\mu},\mathbf{R})$. Further, if ν/τ^2 has the chi-squared distribution with degrees of freedom ν then (19) reduces to (14).

The usual estimator of $\mathbf{R}$ is a multiple of the Wishart matrix of the form $\widetilde{\mathbf{R}}=c_0\mathbf{A}$, where $c_0>0$. Joarder and Ahmed [18] proposed improved estimates for $\mathbf{R}$ as well as its trace and inverse under the quadratic loss function. The proposed estimators for $\mathbf{R}$ are

$$\widehat{\mathbf{R}}=c_0\mathbf{A}-c\,|\mathbf{A}|^{1/p}\,\mathbf{I}, \tag{20}$$

where $\mathbf{I}$ is an identity matrix and c is chosen such that $\widehat{\mathbf{R}}$ is positive definite. For an estimator $\mathbf{R}^*$ of $\mathbf{R}$, let $L(\mathbf{R}^*,\mathbf{R})=\mathrm{tr}[(\mathbf{R}^*-\mathbf{R})^2]$ denote the quadratic loss function and let $R(\mathbf{R}^*,\mathbf{R})=EL(\mathbf{R}^*,\mathbf{R})$ denote the corresponding risk function. The relationship between $\widehat{\mathbf{R}}$ and $\widetilde{\mathbf{R}}$ is rather involved. Defining the dominance of one estimator over another in the same manner as in Sect. 7.1, Joarder and Ahmed [18] established that $\widehat{\mathbf{R}}$ dominates $\widetilde{\mathbf{R}}$ for any c satisfying $d<c<0$, where

$$d=\left(c_0\frac{np+2}{p}-\gamma\right)\frac{\Gamma_p((n-1)/2+1/p)}{\Gamma_p((n-1)/2+2/p)} \tag{21}$$

with $c_0<p\gamma/((n-1)p+2)$ or $0<c<d$, where d is given by (21) with $c_0>p\gamma/(np+2)$ and γ by $\gamma=\gamma_2/\gamma_4$ and $\gamma_i=E(\tau^i)$, $i=1,2,3,4$. The risk functions of the two estimators are given by

$$R\big(\widehat{\mathbf{R}},\mathbf{R}\big) = 4p\gamma_4 \frac{\Gamma_p(n/2+2/p)}{\Gamma_p(n/2)}|\mathbf{R}|^{2/p} c\left(c - \frac{d\,\mathrm{tr}(\mathbf{R}/p)}{|\mathbf{R}|^{1/p}}\right) + \{1+(n-1)c_0\gamma_4(c_0 n - 2\gamma)\}\mathrm{tr}\big(\mathbf{R}^2\big) + (n-1)c_0^2\gamma_4(\mathrm{tr}\mathbf{R})^2$$

and

$$R\big(\widetilde{\mathbf{R}},\mathbf{R}\big) = \{1+(n-1)c_0\gamma_4(c_0 n - 2\gamma)\}\mathrm{tr}\big(\mathbf{R}^2\big) + (n-1)c_0^2\gamma_4(\mathrm{tr}\mathbf{R})^2,$$

respectively. Now consider estimating the trace $\delta = \mathrm{tr}\mathbf{R}$. The usual and the proposed estimators are $\widetilde{\delta} = c_0\mathrm{tr}\mathbf{A}$ and $\widehat{\delta} = c_0\mathrm{tr}\mathbf{A} - cp|\mathbf{A}|^{1/p}$, respectively, where $c_0 > 0$ and c is such that the proposed estimator is positive. Joarder and Ahmed [18] established that the corresponding risk functions are given by

$$R\big(\widetilde{\delta},\delta\big) = [(n-1)c_0\{(n-1)c_0\gamma_4 - 2\gamma_2\} + 1]\delta^2 + 2(n-1)c_0^2\gamma_4\mathrm{tr}\big(\mathbf{R}^2\big)$$

and

$$R\big(\widehat{\delta},\delta\big) = R\big(\widetilde{\delta},\delta\big) + 4p^2\gamma_4 \frac{\Gamma_p(n/2+2/p)}{\Gamma_p(n/2)}|\mathbf{R}|^{2/p} c\left(c - \frac{\mathrm{tr}(\mathbf{R}/p)}{|\mathbf{R}|^{1/p}}d\right),$$

respectively. It is evident that $\widehat{\delta}$ dominates $\widetilde{\delta}$. Finally, consider estimating the inverse $\boldsymbol{\Psi} = \mathbf{R}^{-1}$ with the usual and the proposed estimators given by $\widetilde{\boldsymbol{\Psi}} = c_0\mathbf{A}^{-1}$ and $\widehat{\boldsymbol{\Psi}} = c_0\mathbf{A}^{-1} - c_0|\mathbf{A}|^{-1/p}\mathbf{I}$, respectively, where $c_0 > 0$ and c is such that the proposed estimator is positive definite. In this case, it turns out that $\widehat{\boldsymbol{\Psi}}$ dominates $\widetilde{\boldsymbol{\Psi}}$ for any c satisfying $d < c < 0$, where

$$d = 4\left(\frac{c_0}{n-2/p-p-2} - \frac{\gamma_{-2}}{\gamma_{-4}}\right)\frac{\Gamma_p((n-1)/2-1/p)}{\Gamma_p((n-1)/2-2/p)} \tag{22}$$

with $c_0 < (n-2/p-p-2)\gamma_{-2}/\gamma_{-4}$ or $0 < c < d$, where d is given by (22) with $c_0 > (n-2/p-p-2)\gamma_{-2}/\gamma_{-4}$ and $\gamma_i = E(\tau^i)$.

9 Simulation

Simulation is a key element in modern statistical theory and applications. In this section, we describe two known approaches for simulating from multivariate *t* distributions. Undoubtedly, many other methods will be proposed and elaborated in the near future.

9.1 Vaduva's Method

Vaduva [58] provided a general algorithm for generating from multivariate distributions and illustrated its applicability for multivariate normal, Dirichlet, and multivariate *t* distributions. Here, we present a specialized version of the algorithm for generating the p-variate *t* distribution with the joint pdf

$$f(\mathbf{x}) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\mathbf{R}|^{1/2}}\left[1 + \frac{1}{\nu}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right]^{-(\nu+p)/2}$$

over some domain D in $\Re^p$. It is as follows

1. Initialize.
2. Determine an interval $I = [v_0^0, v_0^1] \times \cdots \times [v_p^0, v_p^1]$, where

$$\begin{aligned} v_0^0 &= 0, \\ v_0^1 &= 1, \\ v_i^0 &= -\sqrt{\frac{\nu(p+1)}{\nu-1}}, \quad i = 1, \ldots, p, \end{aligned}$$

   and

$$v_i^1 = \sqrt{\frac{\nu(p+1)}{\nu-1}}, \quad i = 1, \ldots, p.$$

3. Generate the random vector $\mathbf{V}^*$ uniformly distributed over I. If RND is a uniform random number generator, then $\mathbf{V}^*$ may be generated as follows
   (a) Generate $U_0, U_1, \ldots, U_p$ uniformly distributed over $[0, 1]$ and stochastically independent.
   (b) Calculate $V_i^* = v_i^0 + (v_i^1 - v_i^0)U_i$, $i = 0, 1, \ldots, p$.
   (c) Take $\mathbf{V}^* = (V_0^*, V_1^*, \ldots, V_p^*)$.
4. If $\mathbf{V}^* \notin D$, then go to step (3).
5. Otherwise, take $\mathbf{V} = \mathbf{V}^*$.
6. Calculate $Y_i = V_i / V_0$, $i = 1, \ldots, p$.
7. Take $\mathbf{X} = (Y_1, \ldots, Y_p)^T$. Stop.

Note that the steps from (3) to (5) constitute a rejection algorithm. The performance of this algorithm is characterized by the probability to accept $\mathbf{V}^*$. This probability can be calculated in the form

$$p_a = \frac{\pi^{p/2}\Gamma(\nu/2)}{2^p(p+1)\Gamma((\nu+p)/2)|\mathbf{R}|^{1/2}} \left(\frac{\nu-1}{p+2}\right)^{p/2},$$

which yields

$$\lim_{\nu \to \infty} p_a = 0$$

and

$$\lim_{p \to \infty} p_a = 0,$$

indicating inadequate behavior of the algorithm for large values of p and/or ν.

9.2 Simulation Using BUGS

A relatively simple way to generate a multivariate t involves a sampling of z from gamma($\nu/2$, $\nu/2$) and then sampling a multivariate normal $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{R}/z)$. This mode of generation reflects the scale mixture form of the multivariate t pdf. In BUGS the multivariate normal is parameterized by the precision matrix $\mathbf{P}$; thus one programs a multivariate t pdf as follows to generate a sample of n cases (for Sigma[,], nu.2 and mu[] known)

```
for (i in 1:n)
{z[i] ~ dgamma(nu.2, nu.2)
y[i, 1:q] ~ dmnorm(mu, P.sc[,])}
for (i in 1:q) {for (j in 1:q)
{P[i, j] <- inverse(Sigma[,], i, j)
P.sc[i, j] <- z[i] * P[i, j]}}
```

If one has observed multivariate data and wishes to assume multivariate t sampling, then in BUGS the dmt() form is available

```
for (i in 1:n) {y[i, 1:q] ~ dmt(mu[], P[,], nu)}
```

where nu is assumed known.

10 Applications

In this section, we present a small number of relatively recent applications of multivariate t distributions. The treatment is by no means exhaustive.

10.1 Projection Pursuit

Exploratory projection pursuit is a technique for finding "interesting" low p-dimensional projections of high P-dimensional multivariate data; see [23] for an introduction. Typically, projection pursuit uses a projection index, a functional computed on a projected density (or data set), to measure the "interestingness" of the current projection and then uses a numerical optimizer to move the projection direction to a more interesting position. Loosely speaking, a robust projection index is one that prefers projections involving true clusters over those consisting of a cluster and an outlier. A good robust projection index should perform well even when specific assumptions required for "normal operation" fail to hold or hold only approximately. In a paper that was awarded the Royal Statistical Society Bronze Medal, Nason [41] developed five new indices based on measuring divergence from the multivariate t distribution with the joint pdf

$$f(\mathbf{x}) = \frac{\Gamma((\nu+p)/2)}{\pi^{p/2}(\nu-2)^{p/2}\Gamma(\nu/2)}\left(1+\frac{\mathbf{x}^T\mathbf{x}}{\nu-2}\right)^{-(\nu+p)/2}$$

that are intended to be especially robust. The first three indices are all weighted versions of the L^2-divergences from f for $\nu \geq 3$. They are given by

$$I_{\nu,\alpha}^{\mathrm{TL2}} = \int \{g(\mathbf{x}) - f(\mathbf{x})\}^2 f^{\alpha}(\mathbf{x}) d\mathbf{x}$$

for $\alpha = 0, 1/2, 1$. Nason [41] derived an explicit formula for the case $\alpha = 0$. The fourth index is the Student's t index defined by

$$I_{\nu}^{\mathrm{TI}} = -\int g^{1-2/(\nu+p)}(\mathbf{x}) d\mathbf{x}.$$

This index is minimized over all spherical densities by $f(\mathbf{x})$. Using the transformation $x = \tan(\theta)$, Nason further developed the orthogonal expansion index given by

$$I_{3,1/2}^{\mathrm{TL2}} = \sqrt{\frac{2}{\pi}} \int_{-\pi/2}^{\pi/2} \left\{ g_{\Theta}(\theta) - \frac{2}{\pi}\cos^4\theta \right\}^2 d\theta,$$

where g_Θ is the pdf of the transformed projected data X. Through both numerical calculation and explicit analytical formulas, Nason [41] found the Student's t indices are generally more robust and that indices based on L^2-divergences are also the most robust in their class. A detailed analytical exploration of one of the indices ($I_{\nu,0}^{\mathrm{TL2}}$) showed that it acts robustly when outliers diverge from a main cluster but behaves like a standard projection index when two clusters diverge, that is, its behavior automatically changes depending on the degree of outlier contamination. The degree of sensitivity to outliers can be reduced by increasing the degrees of freedom ν of the $I_{\nu,0}^{\mathrm{TL2}}$-index to make it behave increasingly like Hall's index [12] as $\nu \to \infty$.

10.2 Portfolio Optimization

There are a number of places in finance where robust estimation has been used. For example, when a stock's returns are regressed on the market returns, the slope coefficient, called beta, is a measure of the relative riskiness of the stock in comparison to the market. Quite often, this regression will be performed using robust procedures. However, there appear to be fewer applications of robust estimation in the area of portfolio optimization. In the problem of finding a risk-minimizing portfolio subject to linear constraints, the classical approach assumes normality without exceptions. Lauprete et al. [28] addressed the problem when the return data are generated by a multivariate distribution that is elliptically symmetric but not necessarily normal. They showed that when the returns have marginal heavy tails and multivariate tail-dependence, portfolios will also have heavy tails, and the classical procedures will be susceptible to outliers. They showed theoretically, and on simulated data, that robust alternatives have lower risks. In particular, they showed that when returns have a multivariate t distribution with degrees of freedom less than 6, the least absolute deviation (LAD) estimator has an asymptotically lower risk than the one based on the classical approach. The proposed methodology is applicable when heavy tails and tail-dependence in financial markets are documented especially at high sampling frequencies.

10.3 Discriminant and Cluster Analysis

In the past, there have been many attempts to modify existing methods of discriminant and cluster analyses to provide robust procedures. Some of these have been of a rather ad hoc nature. Recently the multivariate t distribution has been employed for robust estimation. Suppose, for simplicity, that one utilizes two samples in order to assign a new observation into one of two groups, and consider the joint distribution

$$f(\mathbf{x}_1^*, \mathbf{x}_2^*) = \frac{\sqrt{\nu-2}\,\Gamma(\nu+np/2)}{\pi^{np/2}|\mathbf{R}|^{n/2}} \left[(\nu-2) + \sum_{i=1}^{2}\sum_{j=1}^{n_i} (\mathbf{x}_{ij}-\boldsymbol{\mu}_i)^T \mathbf{R}^{-1} (\mathbf{x}_{ij}-\boldsymbol{\mu}_i)\right]^{-(\nu+np)/2} \tag{23}$$

of the two samples $\mathbf{X}_1^* = (\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1})$ and $\mathbf{X}_2^* = (\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2})$ of sizes n_1 and n_2, respectively. In (23), $n = n_1 + n_2$. The $(n_1+n_2)p$-dimensional t distribution (23) was proposed by Sutradhar [52]. It is evident that the marginals are distributed according to

$$f(\mathbf{x}_{ij}) = \frac{\sqrt{\nu-2}\,\Gamma(\nu+p/2)}{\pi^{p/2}|\mathbf{R}|^{n/2}} \left[(\nu-2) + (\mathbf{x}_{ij}-\boldsymbol{\mu}_i)^T \mathbf{R}^{-1} (\mathbf{x}_{ij}-\boldsymbol{\mu}_i)\right]^{-(\nu+p)/2} \tag{24}$$

which is a slight reparameterization of the usual multivariate t pdf. Let π_1 and π_2 denote the two t-populations of the form (24) with parameters $(\boldsymbol{\mu}_1, \mathbf{R}, \nu)$ and $(\boldsymbol{\mu}_2, \mathbf{R}, \nu)$, respectively.

Fisher's optimal discrimination criterion is robust against departure from normality [52], and it assigns the new observation with measurement $\mathbf{X}$ to π_1 if

$$d(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{R}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{R}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0;$$

otherwise, it assigns the observation to π_2. But even though the classification is based on the robust criterion, the probability of misclassification depends on the degrees of freedom of the t distribution. If e_1 and e_2 are probabilities of misclassification of an individual observation from π_1 into π_2 and from π_2 into π_1, respectively, then

$$e_i = \frac{\sqrt{\nu - 2}\Gamma(\nu + 1/2)}{\sqrt{\pi}} \int_{-\infty}^{-\Delta/2} \left\{(\nu - 2) + z^2\right\}^{-(\nu+1)/2} dz$$

for $i = 1, 2$, where $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{R}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Calculations of e_1 and e_2 for selected values of Δ and ν [52] suggest that if a sample actually comes from a t-population (24) with degrees of freedom ν, then the evaluation of the classification error rates by normal-based probabilities would unnecessarily make an experimenter more suspicious. Sutradhar [52] illustrated the use of the preceding discrimination approach by fitting the t distribution to some bivariate data on two species of flea beetles.

McLachlan and Peel [39], McLachlan et al. [40], and Peel and McLachlan [44] used a mixture model of t distributions for a robust method of mixture estimation of clustering. They illustrated its usefulness by a cluster analysis of a simulated data set with added background noise and of an actual data set. For other recent methods for making cluster algorithms robust, see [4, 10, 22, 26, 45, 50, 61].

10.4 Multiple Decision Problems

The multivariate t distribution arises quite naturally in multiple decision problems. In fact, it is one of the earliest applications of this distribution in statistical inference. Suppose there are q dependent variates with means $\theta_1, \ldots, \theta_h, \ldots, \theta_q$, respectively, and that one has estimators $\hat{\theta}_h$ of θ_h, $h = 1, \ldots, q$ available, which are jointly distributed according to a q-variate normal distribution with mean θ_h, $h = 1, \ldots, q$, and covariance matrix $\sigma^2\mathbf{R}$, where $\mathbf{R}$ is a $q \times q$ positive definite matrix and σ^2 is an unknown scale parameter. Let s^2 be an unbiased estimator of σ^2 such that s^2 is independent of the $\hat{\theta}_h$'s and $\nu s^2/\sigma^2$ has the chi-squared distribution with degrees of freedom ν. Consider $p \leq q$ linearly independent linear combinations of θ_h's,

$$m_i = \sum_{h=1}^{q} c_{ih}\theta_h = \mathbf{c}_i^T \boldsymbol{\theta},$$

for $i = 1, \ldots, p$, where $\mathbf{c}_i = (c_{i1}, \ldots, c_{ih}, \ldots, c_{iq})^T$ is a $q \times 1$ vector of known constants. The unbiased estimators of the m_i's are

$$\hat{m}_i = \sum_{h=1}^{q} c_{ih}\hat{\theta}_h = \mathbf{c}_i^T \hat{\boldsymbol{\theta}},$$

each of which is a normally distributed random variable with mean m_i and variance $\mathbf{c}_i^T \mathbf{R}\mathbf{c}_i$. Then

$$Y_i = \frac{\hat{m}_i - m_i}{s\sqrt{\mathbf{c}_i^T \mathbf{R}\mathbf{c}_i}}, \quad i = 1, \ldots, p$$

is a Student's t-variate and $Y_1, \ldots, Y_p$ have the usual p-variate t distribution with degrees of freedom ν, zero means, and the correlation matrix $\{\delta_{iu}\}$ given by

$$\delta_{iu} = \frac{\mathbf{c}_i^T \mathbf{R} \mathbf{c}_u}{\sqrt{\mathbf{c}_i^T \mathbf{R} \mathbf{c}_i \mathbf{c}_u^T \mathbf{R} \mathbf{c}_u}}.$$

For multiple comparisons, one computes the one- and two-sided confidence interval estimates of m_i $(i = 1, \ldots, p)$ simultaneously with a joint confidence coefficient $1 - \alpha$, say. These estimates are given by [9]

$$\hat{m}_i \pm h_1 s \sqrt{\mathbf{c}_i^T \mathbf{R} \mathbf{c}_i}$$

and

$$\hat{m}_i \pm h_2 s \sqrt{\mathbf{c}_i^T \mathbf{R} \mathbf{c}_i},$$

respectively, where the constants h_1 and h_2 are determined so that the intervals in each case have a joint coverage probability of $1 - \alpha$.

10.5 Other Applications

Bayesian prediction approaches using the multivariate t distribution have attracted wide-ranging applications in the last several decades, and many sources are available in periodic and monographic literature. Chien [3] discusses applications in speech recognition and online environmental learning. In experiments of hands-free car speech recognition of connected Chinese digits, it was shown that the proposed approach is significantly better than conventional approaches. Blattberg and Gonedes [2] were one of the first to discuss applications to security returns data. For other applications, we refer the reader to the numerous modern books on multivariate analysis and to the *Proceedings of the Valencia International Meetings*.

References

1. Anderson, T.W., Fang, K.T.: Inference in multivariate elliptically contoured distributions based on maximum likelihood. In: Fang, K.T., Anderson, T.W. (eds.) Statistical Inference in Elliptically Contoured and Related Distributions, pp. 201–216. Allerton, New York (1990)
2. Blattberg, R.C., Gonedes, N.J.: A comparison of the stable and Student distributions as statistical models for stock prices. J. Bus. **47**, 224–280 (1974)
3. Chien, J.-T.: A Bayesian prediction approach to robust speech recognition and online environmental testing. West. J. Speech Commun. **37**, 321–334 (2002)
4. Davé, R.N., Krishnapuram, R.: Robust clustering methods: A unified view. IEEE Trans. Fuzzy Syst. **5**, 270–293 (1995)
5. David, H.A.: Concomitants of order statistics: theory and applications. In: de Oliveira, T. (ed.) Some Recent Advances in Statistics, pp. 89–100. Academic Press, New York (1982)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–38 (1977)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Iteratively weighted least squares for linear regression where errors are normal independent distributed. In: Krishnaiah, P.R. (ed.) Multivariate Analysis, vol. 5, pp. 35–37. North-Holland, Amsterdam (1980)
8. Dey, D.K.: Simultaneous estimation of eigenvalues. Ann. Inst. Stat. Math. **40**, 137–147 (1988)
9. Dunnett, C.W.: A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc. **50**, 1096–1121 (1955)

10. Frigui, H., Krishnapuram, R.: A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. Pattern Recogn. Lett. **17**, 1223–1232 (1996)
11. Gill, M.L., Tiku, M.L., Vaughan, D.C.: Inference problems in life testing under multivariate normality. J. Appl. Stat. **17**, 133–147 (1990)
12. Hall, P.: On polynomial-based projection indices for exploratory projection pursuit. Ann. Stat. **17**, 589–605 (1989)
13. James, W., Stein, C.: Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 361–379. University of California Press, Berkeley (1961)
14. Jensen, D.R.: Closure of multivariate t and related distributions. Stat. Probab. Lett. **20**, 307–312 (1994)
15. Joarder, A.H.: Estimation of the trace of the scale matrix of a multivariate t-model. In: Proceedings of the Econometrics Conference, pp. 467–474. Monash University, Victoria (1995)
16. Joarder, A.H.: Some useful Wishart expectations based on the multivariate t-model. Stat. Pap. **39**, 223–229 (1998)
17. Joarder, A.H., Ahmed, S.E.: Estimation of the characteristic roots of the scale matrix. Metrika **44**, 259–267 (1996)
18. Joarder, A.H., Ahmed, S.E.: Estimation of the scale matrix of a class of elliptical distributions. Metrika **48**, 149–160 (1998)
19. Joarder, A.H., Ali, M.M.: On some generalized Wishart expectations. Commun. Stat., Theory Methods **21**, 283–294 (1992)
20. Joarder, A.H., Ali, M.M.: Estimation of the scale matrix of a multivariate t-model under entropy loss. Metrika **46**, 21–32 (1997)
21. Joarder, A.H., Singh, S.: Estimation of the trace of the scale matrix of a multivariate t-model using regression type estimator. Statistics **29**, 161–168 (1997)
22. Jolion, J.-M., Meer, P., Bataouche, S.: Robust clustering with applications in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. **13**, 791–802 (1995)
23. Jones, M.C., Sibson, R.: What is projection pursuit (with discussion)? J. R. Stat. Soc. A **150**, 1–36 (1987)
24. Kass, R.E., Steffey, D.: Approximate Bayesian in conditionally independent hierarchical models. J. Am. Stat. Assoc. **84**, 717–726 (1989)
25. Kelejian, H.H., Prucha, I.R.: Independent or uncorrelated disturbances in linear regression: An illustration of the difference. Econ. Lett. **19**, 35–38 (1985)
26. Kharin, Y.: Robustness in Statistical Pattern Recognition. Kluwer, Dordrecht (1996)
27. Lange, K., Sinsheimer, J.S.: Normal/independent distributions and their applications in robust regression. J. Comput. Graph. Stat. **2**, 175–198 (1993)
28. Lauprete, G.J., Samarov, A.M., Welsch, R.E.: Robust portfolio optimization. Metrika **55**, 139–149 (2002)
29. Leonard, T.: Comment on "A simple predictive density function" by M. Lejeune and G.D. Faukkenberry. J. Am. Stat. Assoc. **77**, 657–658 (1982)
30. Leonard, T., Hsu, J.S.J., Ritter, C.: The Laplacian T-approximation in Bayesian inference. Stat. Sin. **4**, 127–142 (1994)
31. Leonard, T., Hsu, J.S.J., Tsui, K.W.: Bayesian marginal inference. J. Am. Stat. Assoc. **84**, 1051–1058 (1989)
32. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
33. Liu, C.: Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. J. Multivar. Anal. **46**, 198–206 (1993)
34. Liu, C.: Missing data imputation using the multivariate t distribution. J. Multivar. Anal. **53**, 139–158 (1995)
35. Liu, C.: Bayesian robust multivariate linear regression with incomplete data. J. Am. Stat. Assoc. **91**, 1219–1227 (1996)
36. Liu, C., Rubin, D.B.: ML estimation of the multivariate t distribution with unknown degrees of freedom. Stat. Sin. **5**, 19–39 (1995)
37. Mardia, K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika **57**, 519–530 (1970)
38. Maronna, R.A.: Robust M-estimators of multivariate location and scatter. Ann. Stat. **4**, 51–67 (1976)
39. McLachlan, G.J., Peel, D.: Robust cluster analysis via mixtures of multivariate t-distributions. In: Amin, A., Dori, D., Pudil, P., Freeman, H. (eds.) Lecture Notes in Computer Science, vol. 1451, pp. 658–666. Springer, Berlin (1998)
40. McLachlan, G.J., Peel, D., Basford, K.E., Adams, P.: Fitting of mixtures of normal and t components, J. Stat. Software **4**, (1999)
41. Nason, G.P.: Robust projection indices. J. R. Stat. Soc. B **63**, 551–567 (2001)

42. Neyman, J.: Optimal asymptotic tests for composite hypotheses. In: Grenander, U. (ed.) Probability and Statistics, pp. 213–234. Wiley, New York (1959)
43. Pearson, K.: On non-skew frequency surfaces. Biometrika **15**, 231 (1923)
44. Peel, D., McLachlan, G.J.: Robust mixture modelling using the *t* distribution. Stat. Comput. **10**, 339–348 (2000)
45. Rousseeuw, P.J., Kaufman, L., Trauwaert, E.: Fuzzy clustering using scatter matrices. Comput. Stat. Data Anal. **23**, 135–151 (1996)
46. Rubin, D.B.: Iteratively reweighted least squares. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. 4, pp. 272–275. Wiley, New York (1983)
47. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York (1987)
48. Rubin, D.B., Schafer, J.L.: Efficiently creating multiple imputations for incomplete multivariate normal data. In: Proceedings of the Statistical Computing Section of the American Statistical Association, pp. 83–88. American Statistical Association, Washington (1990)
49. Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman and Hall, London (1997)
50. Smith, D.J., Bailey, T.C., Munford, G.: Robust classification of high-dimensional data using artificial neural networks. Stat. Comput. **3**, 71–81 (1993)
51. Sun, L., Hsu, J.S.J., Guttman, I., Leonard, T.: Bayesian methods for variance component models. J. Am. Stat. Assoc. **91**, 743–752 (1996)
52. Sutradhar, B.C.: Discrimination of observations into one of two *t* populations. Biometrics **46**, 827–835 (1990)
53. Sutradhar, B.C.: Score test for the covariance matrix of the elliptical *t*-distribution. J. Multivar. Anal. **46**, 1–12 (1993)
54. Sutradhar, B.C., Ali, M.M.: Estimation of the parameters of a regression model with a multivariate *t* error variable. Commun. Stat., Theory Methods **15**, 429–450 (1986)
55. Tierney, L., Kadane, J.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**, 82–86 (1986)
56. Tiku, M.L., Kambo, N.S.: Estimation and hypothesis testing for a new family of bivariate nonnormal distributions. Commun. Stat., Theory Methods **21**, 1683–1705 (1992)
57. Tiku, M.L., Suresh, R.P.: A new method of estimation for location and scale parameters. J. Stat. Plan. Inference **30**, 281–292 (1992)
58. Vaduva, I.: Computer generation of random vectors based on transformation if uniformly distributed vectors. In: Iosifescu, M. (ed.) Proceedings of the Seventh Conference on Probability Theory, pp. 589–598. VNU Science Press, Utrecht (1985)
59. Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)
60. Zellner, A.: Bayesian and non-Bayesian analysis of the regression model with multivariate Student-*t* error terms. J. Am. Stat. Assoc. **71**, 400–405 (1976)
61. Zhuang, X., Huang, Y., Palaniappan, K., Zhao, Y.: Gaussian density mixture modeling, decomposition and applications. IEEE Trans. Image Process. **5**, 1293–1302 (1996)