# Robust statistical modelling using the multivariate skew $t$ distribution with complete and incomplete data

Lin, Tsung-I
*Department of Applied Mathematics and Institute of Statistics*
*National Chung Hsing University*
*Taichung 402, Taiwan*
*E-mail: tilin@amath.nchu.edu.tw*

Lin, Tzy-Chy
*Center for Drug Evaluation*
*Division of Clinical Science*
*Taipei 100, Taiwan*
*E-mail: tclin323@cde.org.tw*

## 1. Introduction

In the statistical modelling of multivariate data, sometimes, not all designed measurements are fully collected. The occurrence of missing values in multivariate analysis is a common problem that might lead to biased estimates of parameters or inefficient inferences. Learning multivariate normal (MVN) models from incomplete data has been well-developed and systematically studied in the literature; see, e.g., Anderson (1957), Hocking and Smith (1968), Rubin (1987), and Liu (1999). For analyzing data with incomplete observations, modern imputation methods such as the Expectation Maximization (EM; Dempster et al., 1977) and data augmentation (DA; Tanner and Wong, 1987) algorithms can be easily implemented by using the statistical packages: `proc MI` in SAS (SAS, 2001) and the `Missing Data Analysis Library` in S-PLUS (Schimert et al., 2000).

Typically, the normality assumption for incomplete multivariate data is usually hard to be justified. In some situations, the input data may be asymmetrically distributed or contain influential outliers. The multivariate $t$ (MVT) distribution has become the most frequently used analytical tool for robust inference (Lange et al., 1989; Kotz and Nadarajah, 2004). As pointed out by Liu (1995), to fit data having longer than normal tails, multiple imputations under MVT models allow to yield more valid statistical inferences than using the normal distribution. Lange et al. (1989) remarked the MVT model is not a panacea for modelling data with highly asymmetric observations.

Over the past decades, there has been a growing interest in proposing a parametric family of multivariate skew $t$ (MST) distributions (Jones and Faddy, 2003; Azzalini and Capitaino, 2003; Azzalini and Genton, 2008), which is an extension of the MVT family with additional shape parameters to regulate skewness. In this paper, we consider the missing imputation problems under a new class of MST distributions, defined by Sahu et al. (2003), which can accommodate a wider range of distributional features. Sahu et al. (2003) remarked that the correlation structure within this family is not affected by the introduction of skewness parameters.

To simplify the notation throughout this paper, we let $t_p(\boldsymbol{Y} \mid \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) \propto \big(1+\nu^{-1}(\boldsymbol{Y}-\boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{Y}-\boldsymbol{\xi})\big)^{-(\nu+p)/2}$ denote the probability density function (pdf) of a $p$-dimensional MVT distribution with location vector $\boldsymbol{\xi}$, scale covariance matrix $\boldsymbol{\Omega}$ and degrees of freedom (df) $\nu$, and let $T_p(\cdot \mid \boldsymbol{\Delta}; \nu)$ denote the cumulative distribution function (cdf) of $t_p(\boldsymbol{0}, \boldsymbol{\Delta}, \nu)$. As considered in Sahu et al. (2003), a $p$-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ follows a MST distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^p$, scale covariance matrix $\boldsymbol{\Sigma}$, skewness matrix $\boldsymbol{\Lambda} = \mathrm{Diag}\{\lambda_1, \ldots, \lambda_p\}$ and df $\nu$, denoted as

$\boldsymbol{Y} \sim St_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$, if its pdf is given by

$$(1) \quad f(\boldsymbol{Y} \mid \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu) = 2^p t_p(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) T_p \left( \boldsymbol{q} \sqrt{\frac{\nu+p}{U+\nu}} \,\Big|\boldsymbol{\Delta}; \nu+p \right),$$

where $\boldsymbol{\Delta} = (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$, $\boldsymbol{q} = \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{Y}-\boldsymbol{\xi})$ , $U = (\boldsymbol{Y}-\boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{Y}-\boldsymbol{\xi})$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$.

Specifically, the MST distribution (1) has the following convenient stochastic representation:

$$(2) \quad \boldsymbol{Y} = \boldsymbol{\xi} + \tau^{-1/2}\boldsymbol{\Lambda}|\boldsymbol{Z}_1| + \tau^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}_2, \quad \tau \sim \Gamma(\nu/2, \nu/2), \quad \tau \perp (\boldsymbol{Z}_1, \boldsymbol{Z}_2),$$

where $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are two independent $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ and '$\perp$' indicates independence.

In this class of MST models, we are devoted to developing tactical learning tools to handle missing data problems. Throughout, the mechanism of missingness is assumed to be missing at random (MAR), introduced by Rubin (1976), meaning that the missingness of input data depends only on the observed values. A computationally flexible Monte Carlo Expectation Conditional Maximization (MCECM) algorithm that incorporates two auxiliary matrices is provided to carry out maximum likelihood (ML) estimation under a general pattern of missingness.

Typically, the inference from Bayesian perspective may result in richer inferences than does the likelihood-based approach in the case of small samples. With the application of the well-established Markov chain Monte Carlo (MCMC) method, the Bayesian sampling-based approach has become an attractive alternative to classical single-imputation techniques. To pose the MST model in a Bayesian framework, a Gibbs sampling scheme coupled with the Metropolis and Hastings (M-H) algorithm (Hastings, 1970) and the DA method are developed to take account uncertainties of all parameter and allows for creating multiple imputations of missing data. The priors are chosen to be weakly informative to avoid improper posterior distributions. In addition, the posterior and predictive inferences can be accurately extracted from a large converged Monte Carlo dependent samples.

## 2. The MST model with missing information

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a random sample of size $n$ taken from $St_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$ with $n > p$. According to (2), a three-level hierarchical representation of MST models can be expressed as

$$\begin{aligned} \boldsymbol{Y}_j \mid (\boldsymbol{\gamma}_j, \tau_j) &\sim N_p(\boldsymbol{\xi} + \boldsymbol{\Lambda}\boldsymbol{\gamma}_j, \boldsymbol{\Sigma}/\tau_j), \\ \boldsymbol{\gamma}_j \mid \tau_j &\sim TN_p(\boldsymbol{0}, \boldsymbol{I}_p/\tau_j;\ \mathbb{R}_+^p), \\ (3) \qquad \tau_j &\sim \Gamma(\nu/2, \nu/2), \end{aligned}$$

where $TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}\,; \mathbb{R}_+^p)$ denotes a $p$-variate truncated normal distribution for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ lying within the hyperplane $\mathbb{R}_+^p$, which stands for the Euclidean vector space of all $p$-tuples of positive real numbers. Some properties and moments concerning multivariate truncated normal distribution can be found in Lin (2009, Sec 2.1) and Lin et al. (2009, Theorem 1).

To set up estimating equations for multivariate data allowing for missing values, we partition $\boldsymbol{Y}_j$ into two components $(\boldsymbol{Y}_j^{\mathrm{o}\top}, \boldsymbol{Y}_j^{\mathrm{m}\top})^\top$, where $\boldsymbol{Y}_j^{\mathrm{o}}$ $(p_j^{\mathrm{o}} \times 1)$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ $((p-p_j^{\mathrm{o}}) \times 1)$ denote the observed and missing components of $\boldsymbol{Y}_j$, respectively. To facilitate computation, we introduce two types of the binary indicator matrices, denoted by $\boldsymbol{O}_j$ $(p_j^{\mathrm{o}} \times p)$ and $\boldsymbol{M}_j$ $((p-p_j^{\mathrm{o}}) \times p)$, satisfying $\boldsymbol{Y}_j^{\mathrm{o}} = \boldsymbol{O}_j\boldsymbol{Y}_j$ and $\boldsymbol{Y}_j^{\mathrm{m}} = \boldsymbol{M}_j\boldsymbol{Y}_j$, which can be extracted from a $p$-dimensional identity matrix $\boldsymbol{I}_p$ corresponding to row positions of $\boldsymbol{Y}_j^{\mathrm{o}}$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ in $\boldsymbol{Y}_j$, respectively. It can be easily verified that (a) $\boldsymbol{Y}_j = \boldsymbol{O}_j^\top\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^\top\boldsymbol{Y}_j^{\mathrm{m}}$; (b) $\boldsymbol{O}_j^\top\boldsymbol{O}_j + \boldsymbol{M}_j^\top\boldsymbol{M}_j = \boldsymbol{I}_p$. Accordingly, some important consequences are summarized in the following theorem.

**Theorem 1** *Given the specification of (3), we have*

**(a)** *The conditional distribution of* $\boldsymbol{Y}_j^{\rm o}$ *given* $\boldsymbol{\gamma}_j$ *and* $\tau_j$ *is*

$$\boldsymbol{Y}_j^{\rm o} \mid (\boldsymbol{\gamma}_j, \tau_j) \sim N_{p_j^{\rm o}}(\boldsymbol{\zeta}_j^{\rm o}, \tau_j^{-1}\boldsymbol{\Sigma}_j^{\rm oo}),$$

*where* $\boldsymbol{\zeta}_j^{\rm o} = \boldsymbol{O}_j(\boldsymbol{\xi} + \boldsymbol{\Lambda}\boldsymbol{\gamma}_j)$ *and* $\boldsymbol{\Sigma}_j^{\rm oo} = \boldsymbol{O}_j\boldsymbol{\Sigma}\boldsymbol{O}_j^\top$.

**(b)** *The conditional distribution of* $\boldsymbol{Y}_j^{\rm m}$ *given* $\boldsymbol{Y}_j^{\rm o}$, $\boldsymbol{\gamma}_j$, *and* $\tau_j$ *is*

$$\boldsymbol{Y}_j^{\rm m} \mid (\boldsymbol{Y}_j^{\rm o}, \boldsymbol{\gamma}_j, \tau_j) \sim N_{p-p_j^{\rm o}}(\boldsymbol{\zeta}_j^{\rm m\cdot o}, \tau_j^{-1}\boldsymbol{\Sigma}_j^{\rm mm\cdot o}),$$

*where* $\boldsymbol{\zeta}_j^{\rm m\cdot o} = \boldsymbol{M}_j\big(\boldsymbol{\xi} + \boldsymbol{\Lambda}\boldsymbol{\gamma}_j + \boldsymbol{\Sigma}\boldsymbol{S}_j^{\rm oo}(\boldsymbol{Y}_j - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\gamma}_j)\big)$, $\boldsymbol{\Sigma}_j^{\rm mm\cdot o} = \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}\boldsymbol{S}_j^{\rm oo})\boldsymbol{\Sigma}\boldsymbol{M}_j^\top$ *and* $\boldsymbol{S}_j^{\rm oo} = \boldsymbol{O}_j^\top(\boldsymbol{O}_j\boldsymbol{\Sigma}\boldsymbol{O}_j^\top)^{-1}\boldsymbol{O}_j$.

**(c)** *The marginal distribution of* $\boldsymbol{Y}_j^{\rm o}$ *is* $\mathcal{S}t_{p_j^{\rm o}}(\boldsymbol{\xi}_j^{\rm o}, \boldsymbol{\Sigma}_j^{\rm oo}, \boldsymbol{\Lambda}_j^{\rm oo}, \nu)$ *with density*

$$f(\boldsymbol{Y}_j^{\rm o}) = 2^{p_j^{\rm o}} t_{p_j^{\rm o}}(\boldsymbol{Y}_j^{\rm o}|\boldsymbol{\xi}_j^{\rm o}, \boldsymbol{\Omega}_j^{\rm oo}, \nu) T_{p_j^{\rm o}}\Big(\boldsymbol{q}_j^{\rm o}\sqrt{\frac{\nu + p_j^{\rm o}}{\nu + U_j^{\rm o}}} \mid \boldsymbol{\Delta}_j^{\rm oo}; \nu + p_j^{\rm o}\Big),$$

*where* $\boldsymbol{\xi}_j^{\rm o} = \boldsymbol{O}_j\boldsymbol{\xi}$, $\boldsymbol{\Lambda}_j^{\rm oo} = \boldsymbol{O}_j\boldsymbol{\Lambda}\boldsymbol{O}_j^\top$, $\boldsymbol{\Omega}_j^{\rm oo} = \boldsymbol{O}_j\boldsymbol{\Omega}\boldsymbol{O}_j^\top$, $\boldsymbol{q}_j^{\rm o} = \boldsymbol{\Lambda}_j^{\rm oo}\boldsymbol{\Omega}_j^{\rm oo^{-1}}(\boldsymbol{Y}_j^{\rm o} - \boldsymbol{\xi}_j^{\rm o})$, $U_j^{\rm o} = (\boldsymbol{Y}_j - \boldsymbol{\xi})^\top\boldsymbol{C}_j^{\rm oo}(\boldsymbol{Y}_j - \boldsymbol{\xi})$, $\boldsymbol{C}_j^{\rm oo} = \boldsymbol{O}_j^\top\boldsymbol{\Omega}_j^{\rm oo^{-1}}\boldsymbol{O}_j$ *and* $\boldsymbol{\Delta}_j^{\rm oo} = \boldsymbol{I}_{p_j^{\rm o}} - \boldsymbol{\Lambda}_j^{\rm oo}\boldsymbol{\Omega}_j^{\rm oo^{-1}}\boldsymbol{\Lambda}_j^{\rm oo}$.

**(d)** *The posterior distribution of* $\boldsymbol{\gamma}_j$ *given* $\boldsymbol{Y}_j^{\rm o}$ *follows a multivariate truncated* $t$ *distribution. That is,*

$$\boldsymbol{\gamma}_j \mid \boldsymbol{Y}_j^{\rm o} \sim Tt_p\Big(\boldsymbol{q}_j^*, \frac{U_j^{\rm o} + \nu}{p_j^{\rm o} + \nu}\boldsymbol{\Delta}_j^*, \nu + p_j^{\rm o};\ \mathbb{R}_+^p\Big),$$

*where* $\boldsymbol{q}_j^* = \boldsymbol{\Lambda}\boldsymbol{C}_j^{\rm oo}(\boldsymbol{Y}_j - \boldsymbol{\xi})$ *and* $\boldsymbol{\Delta}_j^* = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{C}_j^{\rm oo}\boldsymbol{\Lambda}$.

**(e)** *The conditional distribution of* $\tau_j$ *given* $\boldsymbol{Y}_j^{\rm o}$ *and* $\boldsymbol{\gamma}_j$ *is*

$$\tau_j \mid (\boldsymbol{Y}_j^{\rm o}, \boldsymbol{\gamma}_j) \sim \Gamma\Big(\frac{p + p_j^{\rm o} + \nu}{2}, \frac{(\boldsymbol{\gamma}_j - \boldsymbol{q}_j^*)^\top\boldsymbol{\Delta}_j^{*^{-1}}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j^*) + U_j^{\rm o} + \nu}{2}\Big).$$

## 3. A fully Bayesian approach

Due to the growing advances of modern computing technology, the Bayesian method is frequently considered as an alternative approach to dealing with missing data problems. Tanner and Wong (1987) proposed the DA algorithm, which has been shown to be an effective procedure for multiple imputations of missing data.

In MST models, it is difficult to draw samples from a posterior distribution $\boldsymbol{\theta} \mid \boldsymbol{Y}^{\rm o}$ directly. With strategies similar to the DA method, we employ a M-H within Gibbs sampler that combines the latent variables $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$, and unobserved data $\boldsymbol{Y}^{\rm m}$ to form an augmented posterior likelihood. The procedure generates posterior variates by recursively drawing from the posterior distributions $p(\boldsymbol{\gamma}_j|\boldsymbol{Y}^{\rm o}, \boldsymbol{\theta}^{(k)})$, $p(\tau_j|\boldsymbol{\gamma}_j^{(k+1)}, \boldsymbol{Y}^{\rm o}, \boldsymbol{\theta}^{(k)})$, $p(\boldsymbol{Y}_j^{\rm m}|\boldsymbol{Y}^{\rm o}, \boldsymbol{\gamma}_j^{(k+1)}, \tau_j^{(k+1)}, \boldsymbol{\theta}^{(k)})$, and $p(\boldsymbol{\theta}|\boldsymbol{Y}^{\rm o}, \boldsymbol{Y}^{{\rm m}(k+1)}, \boldsymbol{\gamma}_j^{(k+1)}, \tau_j^{(k+1)})$. Following Gelfand and Smith (1990), the simulations $\boldsymbol{\gamma}_j^{(k)}$, $\tau_j^{(k)}$, $\boldsymbol{Y}_j^{{\rm m}(k)}$ and $\boldsymbol{\theta}^{(k)}$ will converge to their associated target distributions after a sufficiently long burn-in period.

To avoid yielding improper posterior distributions, we need to choose proper prior distributions for $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$. Let $W_p(a, \mathbf{V})$ denote the $p$-dimensional Wishart distribution with degrees of freedom $a$ and $p \times p$ scale matrix $\mathbf{V}$. When the prior information is not available, a convenient strategy of avoiding improper posterior distribution is to use diffuse proper priors. The prior distributions adopted are as follows:

$$\begin{aligned}\boldsymbol{\xi} &\sim N_p(\boldsymbol{a}, \boldsymbol{\kappa}^{-1}), \quad \mathbf{B} \sim W_p(2\gamma, (2\mathbf{H})^{-1}),\\ \boldsymbol{\Sigma}^{-1}|\mathbf{B} &\sim W_p(2\alpha, (2\mathbf{B})^{-1}), \quad \boldsymbol{\lambda} \sim N_p(\mathbf{0}, \boldsymbol{\Gamma}), \quad \log(1/\nu) \sim U(-10, 10),\end{aligned}$$

where $(\boldsymbol{a}, \boldsymbol{\kappa}, \alpha, \gamma, \mathbf{H}, \boldsymbol{\Gamma})$ are fixed as appropriate quantities to yield the proper posterior distributions. Thus, the joint prior density function of $\boldsymbol{\theta}$ and $\mathbf{B}$ is

$$\begin{aligned}\pi(\boldsymbol{\theta}, \mathbf{B}) &\propto |\mathbf{B}|^{\alpha+(2\gamma-p-1)/2}|\boldsymbol{\Sigma}|^{-(2\alpha-p-1)/2} \exp\Big\{-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{a})^\top \boldsymbol{\kappa}(\boldsymbol{\xi}-\boldsymbol{a})\Big\}\\ &\quad\times \exp\Big\{-\mathrm{tr}\big((\boldsymbol{\Sigma}^{-1}+\mathbf{H})\mathbf{B}\big) - \frac{1}{2}\boldsymbol{\lambda}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}\Big\} J_\nu, \qquad (4)\end{aligned}$$

where $J_\nu = \nu^{-1}$ $(0 < \nu < \infty)$ is the Jacobian of transforming $\log(1/\nu)$ to $\nu$.

Multiplying (4) with the complete data likelihood function, the joint posterior density can be obtained by

$$\begin{aligned}p(\boldsymbol{\theta}, \mathbf{B}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{\tau}, \gamma | \boldsymbol{Y}^{\mathrm{o}}) &\propto \pi(\boldsymbol{\theta}, \mathbf{B}) \prod_{j=1}^n f(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}}, \tau_j, \boldsymbol{\gamma}_j, \boldsymbol{\theta})\, f(\tau_j|\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\gamma}_j, \boldsymbol{\theta})\\ &\quad\times f(\boldsymbol{\gamma}_j|\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\theta})\, f(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\theta}). \qquad (5)\end{aligned}$$

To implement the Gibbs sampler, we are now in a position to present the full conditional posterior densities.

**Theorem 2** *The full conditional posteriors of $\boldsymbol{\theta}$, $\mathbf{B}$, $\boldsymbol{\gamma}_j$, $\tau_j$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ are as follows (the symbol "$|\cdots$" denotes conditioning on all other variables):*

$$\begin{aligned}p(\boldsymbol{\gamma}_j|\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\theta}) &\sim Tt_p\Big(\boldsymbol{q}_j^*, \frac{\nu+U_j^{\mathrm{o}}}{\nu+p_j^{\mathrm{o}}}\boldsymbol{\Delta}_j^*, \nu+p_j^{\mathrm{o}};\ \mathbb{R}_+^p\Big),\\ p(\tau_j|\boldsymbol{\gamma}_j, \boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\theta}) &\sim \Gamma\Big(\frac{p+p_j^{\mathrm{o}}+\nu}{2}, \frac{(\boldsymbol{\gamma}_j-\boldsymbol{q}_j^*)^\top\boldsymbol{\Delta}_j^{*-1}(\boldsymbol{\gamma}_j-\boldsymbol{q}_j^*)+U_j^{\mathrm{o}}+\nu}{2}\Big),\\ p(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\gamma}_j, \tau_j, \boldsymbol{\theta}) &\sim N_{p-p_j^{\mathrm{o}}}(\boldsymbol{\zeta}_j^{\mathrm{m\cdot o}}, \tau_j^{-1}\boldsymbol{\Sigma}_j^{\mathrm{mm\cdot o}}),\\ p(\boldsymbol{\xi}|\cdots) &\sim N_p(\boldsymbol{\mu}^*, \boldsymbol{\kappa}^*),\\ p(\mathbf{B}|\cdots) &\sim W_p(2\gamma^*, (2\mathbf{H}^*)^{-1}),\\ p(\boldsymbol{\Sigma}^{-1}|\cdots) &\sim W_p(\alpha^*, \mathbf{B}^{*-1}),\\ p(\boldsymbol{\lambda}|\cdots) &\sim N_p(\boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*),\end{aligned}$$

*where*

$$\text{(6)}\quad \boldsymbol{\kappa}^* = \Big(\sum_{j=1}^n \tau_j\boldsymbol{\Sigma}^{-1}+\boldsymbol{\kappa}\Big)^{-1}, \quad \boldsymbol{\mu}^* = \boldsymbol{\kappa}^*\Big(\boldsymbol{\Sigma}^{-1}\big(\sum_{j=1}^n \tau_j(\boldsymbol{Y}_j-\boldsymbol{\Lambda}\boldsymbol{\gamma}_j)\big)+\boldsymbol{\kappa}\boldsymbol{a}\Big),$$

$$\text{(7)}\quad \gamma^* = \alpha+\gamma, \quad \mathbf{H}^* = \mathbf{H}+\boldsymbol{\Sigma}^{-1},$$

$$\text{(8)}\quad \alpha^* = 2\alpha+n, \quad \mathbf{B}^* = 2\mathbf{B}+\sum_{j=1}^n \tau_j(\boldsymbol{Y}_j-\boldsymbol{\xi}-\boldsymbol{\Lambda}\boldsymbol{\gamma}_j)(\boldsymbol{Y}_j-\boldsymbol{\xi}-\boldsymbol{\Lambda}\boldsymbol{\gamma}_j)^\top,$$

$$\text{(9)}\quad \boldsymbol{\Gamma}^* = \Big(\boldsymbol{\Gamma}^{-1}+\boldsymbol{\Sigma}^{-1}\odot\sum_{j=1}^n \tau_j\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^\top\Big)^{-1}, \boldsymbol{\delta}^* = \boldsymbol{\Gamma}^*\Big(\boldsymbol{\Sigma}^{-1}\odot\sum_{j=1}^n \tau_j\boldsymbol{\gamma}_j(\boldsymbol{Y}_j-\boldsymbol{\xi})^\top\Big)\mathbf{1}_p\ .$$

*The full conditional distribution of $\nu$ is*

$$\text{(10)}\quad p(\nu|\cdots) \propto \Big[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\Big]^n \Big(\prod_{j=1}^n \tau_j^{\nu/2}\Big)\exp\Big\{-\frac{\nu}{2}\sum_{j=1}^n \tau_j\Big\} J_\nu,$$

*which is not a standard form.*

**Proof:** See Supplementary Section 3.

In the simulation process, samples for $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$, $\boldsymbol{Y}^m$, $\mathbf{B}$, and $\boldsymbol{\theta}$ are alternately generated. In light of Theorem 2 the Gibbs sampler can be implemented as follows:

**1.** Generate $\boldsymbol{\gamma}_j$ from $Tt_p\Big(\boldsymbol{q}_j^*, \frac{\nu+U_j^{\mathrm{o}}}{\nu+p_j^{\mathrm{o}}}\boldsymbol{\Delta}_j^*, \nu+p_j^{\mathrm{o}};\ \mathbb{R}_+^p\Big)$, where $\boldsymbol{q}_j^*$, $U_j^{\mathrm{o}}$ and $\boldsymbol{\Delta}_j^*$ are given in Theorem 1.

**2.** Generate $\tau_j$ from

$$\Gamma\Big(\frac{p+p_j^{\mathrm{o}}+\nu}{2}, \frac{(\boldsymbol{\gamma}_j-\boldsymbol{q}_j^*)^\top\boldsymbol{\Delta}_j^{*^{-1}}(\boldsymbol{\gamma}_j-\boldsymbol{q}_j^*)+U_j^{\mathrm{o}}+\nu}{2}\Big).$$

**3.** Generate $\boldsymbol{Y}_j^{\mathrm{m}}$ from $N_{p-p_j^{\mathrm{o}}}(\boldsymbol{\zeta}_j^{\mathrm{m\cdot o}}, \tau_j^{-1}\boldsymbol{\Sigma}_j^{\mathrm{mm\cdot o}})$, where $\boldsymbol{\zeta}_j^{\mathrm{m\cdot o}}$ and $\boldsymbol{\Sigma}_j^{\mathrm{mm\cdot o}}$ are as in Theorem 1.

**4.** Generate $\boldsymbol{\xi}$ from $N_p(\boldsymbol{\mu}^*, \boldsymbol{\kappa}^*)$, where $\boldsymbol{\mu}^*$ and $\boldsymbol{\kappa}^*$ are given in (6).

**5.** Generate $\mathbf{B}$ from $W_p(2\gamma^*, (2\mathbf{H}^*)^{-1})$, where $\gamma^*$ and $\mathbf{H}^*$ are given in (7).

**6.** Generate $\boldsymbol{\Sigma}^{-1}$ from $W_p(\alpha^*, \mathbf{B}^{*-1})$, where $\alpha^*$ and $\mathbf{B}^*$ are given in (8).

**7.** Generate $\boldsymbol{\lambda}$ from $N_p(\boldsymbol{\delta}^*, \boldsymbol{\Gamma}^*)$, where $\boldsymbol{\delta}^*$ and $\boldsymbol{\Gamma}^*$ are given in (9).

**8.** Generate $\nu$ from (10) via the M-H algorithm.

## REFERENCES

Anderson TW (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200-203.

Azzalini A and Capitaino A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew $t$-distribution. *Journal of the Royal Statistical Society, Series B*, **65**, 367-389.

Azzalini A and Genton MG (2008) Robust likelihood methods based on the skew-$t$ and related distributions. *International Statistical Review*, **76**, 106-129.

Browne WJ and Draper D (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models (with discussion). *Bayesian Analysis*, **1**, 473-514.

Chan KS and Ledolter J (1995) Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, **90**, 242-252.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B*, **39**, 1-38.

Hocking RR and Smith WB (1968) Estimation of parameters in the multivariate normal distribution with missing observations. *Journal of the American Statistical Association*, **63**, 159-173.

Jones MC and Faddy MJ (2003) A skew extension of the $t$-distribution, with applications. *Journal of the Royal Statistical Society, B*, **65**, 159-174.

Kotz S and Nadarajah S (2004) *Multivariate t Distributions and Their Applications.* Cambridge University Press.

Lange KL, Little RJA and Taylor JMG (1989) Robust statistical modeling using the $t$ distribution. *Journal of the American Statistical Association*, **84**, 881-896.

Lin TI (2009) Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* **100**, 257-265.

Lin TI (2010) Robust mixture modeling using multivariate skew $t$ distributions. *Statistics and Computing* **20** 343-356.

Lin TI, Ho HJ and Chen CL (2009) Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis* **100**, 2337-2351.

Lin TC and Lin TI (2010) Supervised learning of multivariate skew normal mixture models with missing information. *Computational Statistics* **25** 183-201.

Liu CH (1995) Missing data imputation using the multivariate $t$ distribution. *Journal of Multivariate Analysis*, **53**, 139-158.

Liu CH (1999) Efficient ML estimation of the multivariate normal distribution from incomplete data, *Journal of Multivariate Analysis*, **69**, 206-217.

Liu, CH and Rubin DB (1995) ML estimation of the t distribution using EM and its extensions, ECM and ECME, *Statistica Sinica* **5**, 19-39.

Meng XL and Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267-78.

Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier L, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirov JP (2009) Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences (PNAS) USA,* **106** 8519-8524.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581-592.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

SAS (2001) SAS/Stat User's Guide, Version 8.2, Cary, NC: SAS Institute Inc.

Schimert J, Schafer J, Hesterberg T, Fraley C and Clarkson D (2000) Analyzing Data with Missing Values in S-Plus, Seattle, WA: Data Analysis Products Division, Insightful Corporation.

Wei GCG and Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699-704.

## RÉSUMÉ

*Missing data is inevitable in many situations that could hamper data analysis for scientific investigations. We establish flexible analytical tools for multivariate skew t models when fat-tailed, asymmetric and missing observations simultaneously occur in the input data. For the ease of computation and theoretical developments, two auxiliary indicator matrices are incorporated into the model for the determination of observed and missing components of each observation that can effectively reduce the computational complexity. Under the missing at random assumption, we present a Monte Carlo version of the ECM algorithm, which is performed to estimate the parameters and retrieve each missing observation with a single value. Additionally, a Metropolis-Hastings within Gibbs sampler with data augmentation is developed to account for the uncertainty of parameters as well as missing outcomes. The methodology is illustrated through two real data sets.*