

# A robust factor analysis model using the restricted skew- $t$ distribution

Tsung-I Lin · Pal H. Wu ·  
Geoffrey J. McLachlan · Sharon X. Lee

Received: 13 November 2013 / Accepted: 24 November 2014 / Published online: 7 December 2014  
© Sociedad de Estadística e Investigación Operativa 2014

**Abstract** Factor analysis is a classical data-reduction technique that seeks a potentially lower number of unobserved variables that can account for the correlations among the observed variables. This paper presents an extension of the factor analysis model, called the skew- $t$  factor analysis model, constructed by assuming a restricted version of the multivariate skew- $t$  distribution for the latent factors and a symmetric  $t$ -distribution for the unobservable errors jointly. The proposed model shows robustness to violations of normality assumptions of the underlying latent factors and provides flexibility in capturing extra skewness as well as heavier tails of the observed data. A computationally feasible expectation conditional maximization algorithm is developed for computing maximum likelihood estimates of model parameters. The usefulness of the proposed methodology is illustrated using both simulated and real data.

**Keywords** ECM algorithm · ML estimation · SNFA model · STFA model · rMSN distribution · rMST distribution

**Mathematics Subject Classification** 62H12 · 62H25

---

**Electronic supplementary material** The online version of this article (doi:[10.1007/s11749-014-0422-2](https://doi.org/10.1007/s11749-014-0422-2)) contains supplementary material, which is available to authorized users.

---

T.-I. Lin (✉) · P. H. Wu  
Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan  
e-mail: [tilin@nchu.edu.tw](mailto:tilin@nchu.edu.tw)

T.-I. Lin  
Department of Public Health, China Medical University, Taichung 404, Taiwan

G. J. McLachlan · S. X. Lee  
Department of Mathematics, University of Queensland, St. Lucia 4072, Australia  
e-mail: [g.mclachlan@uq.edu.au](mailto:g.mclachlan@uq.edu.au)

## 1 Introduction

Factor analysis (FA), which originated from the work of (1904), is concerned with a way of summarizing the variability between a number of correlated variables; see, for example, (Lawley and Maxwell 1971). The correlations between the variables under consideration are explained by their linear dependence on a usually much smaller number of unobservable (latent) factors. In particular, FA can be considered as an extension of principal component analysis (PCA), both of which are widely used statistical tools for reducing dimensionality by constructing linear combinations of the variables. Unlike the PCA model, the FA model enjoys a powerful invariance property: changes in the scales of the variables in  $Y$  appear only as scale changes in the appropriate rows of the matrix of factor loadings.

FA has been successfully applied to numerous problems that arise naturally in many areas, see Basilevsky (2008) for a literature survey. In the FA framework, errors and factors are routinely assumed to have a Gaussian distribution because of their mathematical and computational tractability. However, the traditional FA approach has often been criticized for the lack of stability and robustness against non-normal characteristics such as skewness and heavy tails. Statistical methods which ignore the departure of normality may cause biased or misleading inference. To remedy this weakness, authors such as McLachlan et al. (2007), Wang and Lin (2013), and Zhang et al. (2013) considered the use of the multivariate  $t$  (MT) distribution for robust estimation of FA models, known as the tFA model.

When the data have longer than normal tails or contain atypical observations (the so-called outliers), the MT distribution has been shown to be a natural extension of the normal for making robust statistical inference (Lange et al. 1989; Kotz and Nadarajah 2004) as it has an extra tuning parameter, the degrees of freedom ( $df$ ), to regulate the thickness of tails. In many biological applications (cf. Pyne et al. 2009; Rossin et al. 2011; Ho et al. 2012) and other applied problems, however, the data often involve observations whose distributions are highly asymmetric as well as having fat tails.

Over the past two decades, there has been a growing interest in proposing more flexible parametric families that can accommodate skewness and other non-normal features. In particular, the family of multivariate skew- $t$  (MST) distributions (Azzalini and Capitanio 2003; Jones and Faddy 2003; Sahu et al. 2003; Azzalini and Genton 2008) have received considerable attention. This family contains additional skewness parameters for modeling asymmetry and includes the MT family as a special case.

This paper presents a robust version of the standard FA model by considering the joint distribution of the factors and the error vector to have a joint restricted skew- $t$  distribution in which the skewness parameters are zero for the error vector; that is, the latent factors have a rMST distribution and the error vector has a symmetric MT distribution. Henceforth, we refer to this skew- $t$  factor analysis as STFA. Notably, the practical use of STFA would be more widely applicable as it includes the classical FA as a limiting case and the tFA as a special case. The rMST distribution denotes the skew distribution of (Sahu et al. 2003) with the restriction that the skewing latent variables in its convolution formulation are all equal; that is, the rMST distribution has a univariate skewing function. The rMST distribution is the same after an appropriate transformation as the skew- $t$  distribution proposed by Azzalini and Capitanio

(2003) and has been widely studied in the literature and used in practice. When the df approaches infinity, the limiting distribution of rMST is the restricted multivariate skew-normal (rMSN) distribution. A comprehensive overview of their characterizations together with their conditioning-type and convolution-type representations can be found in Lee and McLachlan (2013, 2014).

Recently, several different skew factor-analytic models have been proposed in the literature, for example, Montanari and Viroli (2010) and Wall et al. (2012). More recently, Murray et al. (2014a) proposed a skew factor analysis model in which the error vector is taken to have a generalized hyperbolic skew  $t$  (GHST) distribution (Barndorff-Nielsen and Shephard 2001), while the factor vector is assumed to have a MT distribution. For brevity, we call this approach the “generalized hyperbolic skew- $t$  factor analysis (GHSTFA)” model.

It is important to note that the GHST distribution is quite different from the rMST distribution as pointed out in Aas and Haff (2006). Firstly, as the degrees of freedom parameter in the rMST distribution approaches infinity, the rMST distribution is reduced to the restricted multivariate skew-normal (rMSN) distribution, whereas the GHST distribution tends to an elliptically symmetric distribution, namely, an ordinary multivariate normal (MN) distribution (Lee and Poon 2011). Secondly, the rMST distribution has heavy tails (polynomial) in all directions, whereas the GHST distribution has some tails that are semi-heavy (exponential).

To further reduce the number of free parameters, Murray et al. (2014b) have put forward an alternative to the GHSTFA model that assumes skew common factors. This new approach is called the “generalized hyperbolic common skew- $t$  factor analysis (GHCSTFA)” model, constructed by taking the latent factor vector rather than the error vector to have the GHST distribution. It is also important to note that except in Lin et al. (2013) and Murray et al. (2014b), in all previous works the factor vector is taken to have a symmetric distribution with the asymmetric distribution being assumed for the error vector as, for example, in Murray et al. (2013) and Tortora et al. (2013).

The paper is structured as follows. In Sect. 2, we establish the notation and briefly outline some preliminary properties of the rMSN and rMST distributions. Section 3 discusses the specification of the STFA model and presents the development of an ECM algorithm for obtaining the ML estimates of model parameters. In Sect. 4, we describe two simple ways of computing the standard errors of the STFA model parameters based on the information-based method and the parametric bootstrap procedure. In Sect. 5, we illustrate the usefulness of the proposed method with a real-life data set. A simulation study is undertaken to compare the performance of the STFA, GHSTFA and GHCSTFA methods. Some concluding remarks are given in Sect. 6 and technical derivations are sketched in Supplementary Appendices.

## 2 Preliminaries

We begin with a brief review of the rMST distribution and a study of some essential properties. To establish notation, we let  $\phi_p(\cdot; \mu, \Sigma)$  be the probability density function of  $N_p(\mu, \Sigma)$  (a  $p$ -variate MN distribution with mean  $\mu$  and covariance matrix  $\Sigma$ );  $\Phi(\cdot)$  be the cumulative distribution function (cdf) of the standard normal distribution;

$t_p(\cdot; \mu, \Sigma, \nu)$  be the pdf of  $t_p(\cdot; \mu, \Sigma, \nu)$  (a  $p$ -variate MT with location  $\mu$  and scale covariance matrix  $\Sigma$  and degrees of freedom  $\nu$ );  $T(\cdot; \nu)$  be the cdf of the Student's  $t$  distribution with df  $\nu$ ;  $TN(\mu, \sigma^2; (a, b))$  be the truncated normal distribution for  $N(\mu, \sigma^2)$  lying within a truncated interval  $(a, b)$ ;  $M^{1/2}$  denote the square root of a symmetric matrix  $M$ ;  $1_p$  denote a  $p \times 1$  vector of ones;  $I_p$  be the  $p \times p$  identity matrix;  $\text{Diag}\{\cdot\}$  be a diagonal matrix created by extracting the main diagonal elements of a square matrix or the diagonalization of a vector and  $\text{vec}(\cdot)$  for an operator that vectorizes a matrix by stacking its columns vertically.

Based on Pyne et al. (2009), a  $p$ -dimensional random vector  $Y$  is said to follow a rMST distribution with location vector  $\mu \in \mathbb{R}^p$ , scale covariance matrix  $\Sigma$ , skewness vector  $\lambda \in \mathbb{R}^p$  and df  $\nu \in (0, \infty)$ , denoted as  $rSt_p(\mu, \Sigma, \lambda, \nu)$ , if it can be represented by

$$\begin{aligned} Y &= \mu + W^{-1/2}X, \quad X \sim rSN_p(0, \Sigma, \lambda), \\ W &\sim \text{gamma}(\nu/2, \nu/2), \quad X \perp W, \end{aligned} \tag{1}$$

where  $\text{gamma}(\alpha, \beta)$  stands for a gamma distribution with mean  $\alpha/\beta$ . If  $\lambda = 0$ , the distribution of  $Y$  reduces to  $t_p(\mu, \Sigma, \nu)$  and to  $rSN_p(\mu, \Sigma, \lambda)$  as  $\nu \rightarrow \infty$ . In addition, this class of distributions also includes the MN distribution, recovered by setting  $\lambda = 0$  and  $\nu \rightarrow \infty$ . Combining the strengths of the MT and rMSN distributions, the rMST distribution offers a robustness mechanism against both asymmetry and outliers observed in the data.

From (1), it is clear that the rMST distribution corresponds to a two-level hierarchical representation

$$Y \mid (W = w) \sim rSN_p(\mu, w^{-1}\Sigma, w^{-1/2}\lambda) \quad \text{and} \quad W \sim \text{gamma}(\nu/2, \nu/2). \tag{2}$$

Integrating  $W$  from the joint density of  $(Y, W)$  yields the marginal density of  $Y$

$$f(y) = 2t_p(y; \mu, \Omega, \nu) T\left(A\left(\frac{\nu + p}{\nu + M}\right)^{1/2}; \nu + p\right), \tag{3}$$

where  $\Omega = \Sigma + \lambda\lambda^T$ ,  $A = (1 - \lambda^T\Omega^{-1}\lambda)^{-1/2}\lambda^T\Omega^{-1}(y - \mu)$  and  $M = (y - \mu)^T\Omega^{-1}(y - \mu)$ .

### 3 Skew- $t$ factor analysis model

#### 3.1 Model formulation

Suppose that  $Y = \{Y_1, \dots, Y_n\}$  constitutes a random sample of  $n$   $p$ -dimensional observations. To improve the robustness for modeling correlation in the presence of asymmetric levels of sources, we consider a generalization of the  $t$ FA model in which the latent factor is described by the rMST distribution defined in (3). The model considered here is

$$Y_j = \mu + BU_j + \varepsilon_j \quad \text{with} \\ \begin{bmatrix} U_j \\ \varepsilon_j \end{bmatrix} \sim rSt_{q+p} \left( \begin{bmatrix} -a_v \Lambda^{-1/2} \lambda \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & 0 \\ 0 & D \end{bmatrix}, \begin{bmatrix} \Lambda^{-1/2} \lambda \\ 0 \end{bmatrix}, \nu \right), \quad (4)$$

for  $j = 1, \dots, n$ , where  $\mu$  is a  $p$ -dimensional location vector,  $B$  is a  $p \times q$  matrix of factor loadings,  $U_j$  is a  $q$ -dimensional vector ( $q < p$ ) of latent variables called *factors*,  $\varepsilon_j$  is a  $p$ -dimensional vector of errors called *specific factors*,  $D$  is a positive diagonal matrix,  $\Lambda = I_q + (1 - a_v^2(\nu - 2)/\nu) \lambda \lambda^T$  with

$$a_v = (\nu/\pi)^{1/2} \frac{\Gamma((\nu - 1)/2)}{\Gamma(\nu/2)} \quad (5)$$

being a scaling coefficient. Marginally, the latent factors in (4) follow an asymmetric rMST distribution, while the errors follow a (symmetric) MT distribution. Moreover, one appealing feature of (4) is that

$$E(U_j) = 0 \quad \text{and} \quad \text{cov}(U_j) = \{\nu/(\nu - 2)\} I_q,$$

which coincide with the conditions under the tFA model. According to (2), the STFA model has a two-level hierarchical representation:

$$Y_j | w_j \sim rSN_p \left( \mu - a_v \alpha, w_j^{-1} \Sigma, w_j^{-1/2} \alpha \right) \quad \text{and} \quad W_j \sim \text{gamma}(\nu/2, \nu/2). \quad (6)$$

Derivation of the marginal distribution of  $Y$  can be accomplished by direct calculation which leads to

$$Y_j \sim rSt_p(\mu - a_v \alpha, \Sigma, \alpha, \nu),$$

where  $\Sigma = B \Lambda^{-1} B^T + D$  and  $\alpha = B \Lambda^{-1/2} \lambda$ . The marginal density of  $Y_j$  is

$$f(y_j; \theta) = 2t_p(y_j; \mu - a_v \alpha, \Omega, \nu) T \left( A_j \left( \frac{\nu + p}{\nu + M_j} \right)^{1/2}; \nu + p \right), \quad (7)$$

where  $\Omega = \Sigma + \alpha \alpha^T$ ,  $M_j = (y_j - \mu + a_v \alpha)^T \Omega^{-1} (y_j - \mu + a_v \alpha)$  and  $A_j = h_j / \sigma$  with  $h_j = \alpha^T \Omega^{-1} (y_j - \mu + a_v \alpha)$  and  $\sigma^2 = 1 - \alpha^T \Omega^{-1} \alpha$ .

The mean and covariance matrix of  $Y_j$  can be obtained as

$$E(Y_j) = \mu \quad \text{and} \quad \text{cov}(Y_j) = \frac{\nu}{\nu - 2} (B B^T + D).$$

It appears that both tFA and STFA models share the same first two moments for the marginal distribution of  $Y_j$ .

For a hidden dimensionality  $q > 1$ , the STFA model also suffers from an identifiability problem associated with the rotation invariance of the loading matrix  $B$ , since model (4) still satisfies when  $B$  is replaced by  $BR$ , where  $R$  is any orthogonal rotation

matrix of order  $q$ . To remedy the situation of rotational indeterminacy, there are several different ways of placing rotational identifiability constraints. The most popular method is to choose  $R$  such that  $B^T D^{-1} B$  is a diagonal matrix (Lawley and Maxwell 1971) with its diagonal elements arranged in a descending order. The other commonly used technique is to constrain the loading matrix  $B$  so that the upper-right triangle is zero and the diagonal entries are strictly positive (e.g., Fokoué and Titterington 2003; Lopes and West 2004). Both methods impose  $q(q - 1)/2$  constraints on  $B$ . Therefore, the number of free parameters to be estimated is  $m = p(q + 2) + q - q(q - 1)/2 + 1$ .

### 3.2 Maximum likelihood estimation via the ECM algorithm

To help the derivation of the algorithm, we adopt the following scaling transformation:

$$\tilde{B} \triangleq B \Lambda^{-1/2} \quad \text{and} \quad \tilde{U}_j \triangleq \Lambda^{1/2} U_j.$$

Clearly, the model remains invariant under the above transformation. It follows from (6) that the STFA model can be formulated in a flexible hierarchical representation as follows:

$$\begin{aligned} Y_j \mid (\tilde{U}_j, v_j, w_j) &\sim N_p(\mu + \tilde{B} \tilde{U}_j, w_j^{-1} D), \\ \tilde{U}_j \mid (v_j, w_j) &\sim N_q((v_j - a_v)\lambda, w_j^{-1} I_q), \\ V_j \mid w_j &\sim TN(0, w_j^{-1}; (0, \infty)), \\ W_j &\sim \text{gamma}(v/2, v/2). \end{aligned} \tag{8}$$

Consequently, applying Bayes' rule, it suffices to show

$$\begin{aligned} \tilde{U}_j \mid (y_j, v_j, w_j) &\sim N_q(q_j, w_j^{-1} C), \\ V_j \mid (y_j, w_j) &\sim TN(h_j, w_j^{-1} \sigma^2; (0, \infty)), \\ f(w_j; y_j) &= \frac{\Phi(w_j^{1/2} A_j)}{T(A_j \left(\frac{v+p}{v+M_j}\right)^{1/2}; v+p)} f_G\left(w_j; \frac{v+p}{2}, \frac{v+M_j}{2}\right), \end{aligned} \tag{9}$$

where  $q_j = C\{d_j + \lambda(v_j - a_v)\}$ ,  $d_j = \tilde{B}^T D^{-1}(Y_j - \mu)$  and  $C = (I_q + \tilde{B}^T D^{-1} \tilde{B})^{-1}$ .

For notational convenience, let  $y = (y_1^T, \dots, y_n^T)^T$  be the observed data. Moreover, we define  $U = (U_1^T, \dots, U_n^T)^T$ ,  $V = (V_1, \dots, V_n)^T$ , and  $W = (W_1, \dots, W_n)^T$ , which are treated as missing values in the complete data framework. In light of (8), the complete data log-likelihood function for  $\theta = (\mu, B, D, \lambda, v)$  given  $y_c = (y^T, U^T, V^T, W^T)^T$ , aside from additive constants, is

$$\begin{aligned} \ell_c(\theta; y_c) &= -\frac{n}{2} \log |D| - \frac{1}{2} \text{tr} \left( D^{-1} \sum_{j=1}^n \tilde{Y}_j \right) \\ &\quad - \frac{1}{2} \sum_{j=1}^n \left[ W_j \left\{ (V_j - a_v)^2 \lambda^T \lambda - 2(V_j - a_v) \lambda^T \tilde{U}_j + \tilde{U}_j \tilde{U}_j^T \right\} \right] \\ &\quad + \frac{nv}{2} \log \left( \frac{\nu}{2} \right) - n \log \Gamma \left( \frac{\nu}{2} \right) + \frac{\nu}{2} \sum_{j=1}^n (\log W_j - W_j), \end{aligned} \tag{10}$$

where  $\tilde{Y}_j = W_j(y_j - \mu - \tilde{B}\tilde{U}_j)(y_j - \mu - \tilde{B}\tilde{U}_j)^T$ .

The expectation–maximization (EM) algorithm (Dempster et al. 1977) is a popular iterative method to compute the ML estimates when the data are incomplete. Given an initial solution  $\theta^{(0)}$ , the implementation of the EM algorithm consists of alternating repeatedly the Expectation (E)- and Maximization (M)-steps until convergence has been reached. Often in many practical problems, the solution to the M-step may encounter some difficulties such that no closed-form expressions exist for updating parameters. For ML estimation of the STFA model, we resort to the ECM algorithm (Meng and Rubin 1993) in which the M-step is replaced by a sequence of computationally simpler conditional maximization (CM) steps while sharing all appealing advantages of the standard EM algorithm.

To calculate the expectation of the complete data log-likelihood, called the  $Q$ -function, we require the following conditional expectations:

$$\begin{aligned} \hat{w}_j^{(k)} &= E(W_j | y_j, \hat{\theta}^{(k)}), \quad \hat{\kappa}_j^{(k)} = E(\log W_j | y_j, \hat{\theta}^{(k)}), \\ \hat{s}_{1j}^{(k)} &= E(W_j V_j | y_j, \hat{\theta}^{(k)}), \quad \hat{s}_{2j}^{(k)} = E(W_j V_j^2 | y_j, \hat{\theta}^{(k)}), \\ \hat{\Omega}_j^{(k)} &= E(W_j \tilde{U}_j \tilde{U}_j^T | y_j, \hat{\theta}^{(k)}), \\ \hat{\eta}_j^{(k)} &= E(W_j \tilde{U}_j | y_j, \hat{\theta}^{(k)}) \text{ and } \hat{\zeta}_j^{(k)} = E(W_j V_j \tilde{U}_j | y_j, \hat{\theta}^{(k)}), \end{aligned} \tag{11}$$

which are directly obtainable from using (A.1)–(A.7) given in Supplementary Proposition 4. As a result, the  $Q$ -function can be written as

$$\begin{aligned} Q(\theta; \hat{\theta}^{(k)}) &= -\frac{n}{2} \log |D| - \frac{1}{2} \text{tr} \left( D^{-1} \sum_{j=1}^n \hat{\Upsilon}_j^{(k)} \right) \\ &\quad - \frac{1}{2} \sum_{j=1}^n \left\{ \left( \hat{s}_{2j}^{(k)} - 2a_v \hat{s}_{1j}^{(k)} + a_v^2 \hat{w}_j^{(k)} \right) \lambda^T \lambda - 2\lambda^T \left( \hat{\zeta}_j^{(k)} - a_v \hat{\eta}_j^{(k)} \right) + \hat{\Omega}_j^{(k)} \right\} \\ &\quad + \frac{nv}{2} \log \left( \frac{\nu}{2} \right) - n \log \Gamma \left( \frac{\nu}{2} \right) + \frac{\nu}{2} \sum_{j=1}^n (\hat{\kappa}_j^{(k)} - \hat{w}_j^{(k)}), \end{aligned} \tag{12}$$

where

$$\hat{\Upsilon}_j^{(k)} = \hat{w}_j^{(k)}(y_j - \mu)(y_j - \mu)^T - \tilde{B}\hat{\eta}_j^{(k)}(y_j - \mu)^T - (y_j - \mu)\hat{\eta}_j^{(k)T}\tilde{B}^T + \tilde{B}\hat{\Omega}_j^{(k)}\tilde{B}^T, \tag{13}$$

which contains free parameters  $\mu$  and  $\tilde{B}$ . In summary, the implementation of the ECM algorithm proceeds as follows:

E-step: Given  $\theta = \hat{\theta}^{(k)}$ , compute  $\hat{w}_j^{(k)}, \hat{\kappa}_j^{(k)}, \hat{s}_{1j}^{(k)}, \hat{s}_{2j}^{(k)}, \hat{\eta}_j^{(k)}, \hat{\zeta}_j^{(k)}$  and  $\hat{\Omega}_j^{(k)}$  in (11), for  $j = 1, \dots, n$ .

CM-step 1: Update  $\hat{\mu}^{(k)}$  by maximizing (12) over  $\mu$ , which leads to

$$\hat{\mu}^{(k+1)} = \frac{\sum_{j=1}^n (\hat{w}_j^{(k)} y_j - \hat{B}^{(k)} \hat{\eta}_j^{(k)})}{\sum_{j=1}^n \hat{w}_j^{(k)}}.$$

CM-step 2: Given  $\mu = \hat{\mu}^{(k+1)}$ , update  $\hat{B}^{(k)}$  by maximizing (12) over  $\tilde{B}$ , which gives

$$\hat{B}^{(k+1)} = \left\{ \sum_{j=1}^n (y_j - \hat{\mu}^{(k+1)}) \hat{\eta}_j^{(k)T} \right\} \left( \sum_{j=1}^n \hat{\Omega}_j^{(k)} \right)^{-1}.$$

CM-step 3: Given  $\mu = \hat{\mu}^{(k+1)}$  and  $\tilde{B} = \hat{B}^{(k+1)}$ , update  $\hat{D}^{(k)}$  by maximizing (12) over  $D$ , which leads to

$$\hat{D}^{(k+1)} = \frac{1}{n} \text{Diag} \left( \sum_{j=1}^n \hat{\Upsilon}_j^{(k)} \right).$$

where  $\hat{\Upsilon}_j^{(k)}$  is  $\tilde{\Upsilon}_j^{(k)}$  in (13) with  $\mu$  and  $\tilde{B}$  replaced by  $\hat{\mu}^{(k+1)}$  and  $\hat{B}^{(k+1)}$ , respectively.

CM-step 4: Update  $\hat{\lambda}^{(k)}$  by maximizing (12) over  $\lambda$ , which gives

$$\hat{\lambda}^{(k+1)} = \frac{\sum_{j=1}^n (\hat{\zeta}_j^{(k)} - a_v \hat{\eta}_j^{(k)})}{\sum_{j=1}^n (\hat{s}_{2j}^{(k)} - 2a_v \hat{s}_{1j}^{(k)} + a_v^2 \hat{w}_j^{(k)})}.$$



CM-step 5: Calculate  $\hat{\nu}^{(k+1)}$  by maximizing (12) over  $\nu$ , which is equivalent to solving the root of the following equation:

$$-\frac{1}{n} \sum_{j=1}^n \left\{ \left( -2a'_\nu \hat{s}_{1j}^{(k)} + 2a'_\nu a_\nu \hat{w}_j^{(k)} \right) \lambda^T \lambda + 2a'_\nu \lambda^T \hat{\eta}_j^{(k)} \right\} + \log \left( \frac{\nu}{2} \right) - \text{DG} \left( \frac{\nu}{2} \right) + 1 + \frac{1}{n} \sum_{j=1}^n \left( \hat{\kappa}_j^{(k)} - \hat{w}_j^{(k)} \right) = 0,$$

where  $\text{DG}(\cdot)$  denotes the digamma function and

$$a'_\nu = \frac{da_\nu}{d\nu} = \frac{1}{2} \left( \frac{1}{\pi\nu} \right)^{1/2} \frac{\Gamma \left( \frac{\nu-1}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right)} + 2 \left( \frac{\nu}{\pi} \right)^{1/2} \times \frac{\Gamma \left( \frac{\nu-1}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right)} \left\{ \text{DG} \left( \frac{\nu-1}{2} \right) - \text{DG} \left( \frac{\nu}{2} \right) \right\}.$$

In the above CM-step 5, the R function ‘uniroot’ is employed to obtain the solution of  $\nu$ . To facilitate faster convergence, the range of  $\nu$  is restricted to have a maximum of 200, which does not affect the inference when the underlying distribution of factor scores has a near skew-normal or normal shape. Upon convergence, the ML estimate of  $\theta$  is denoted by  $\hat{\theta} = (\hat{\mu}, \hat{B}, \hat{D}, \hat{\lambda})$ , where  $\hat{B} = \hat{B} \hat{\Lambda}^{1/2}$  and  $\hat{\Lambda} = I_q + \left( 1 - \frac{\hat{\nu}-2}{\hat{\nu}} \hat{a}_\nu^2 \right) \hat{\lambda} \hat{\lambda}^T$ . Consequently, the estimation of factor scores through *conditional prediction* is obtained by

$$\hat{U}_j = E(U_j | y_j, \hat{\theta}) = \hat{\Lambda}^{-1/2} \hat{C} \left\{ \hat{d}_j + \hat{\lambda} (\hat{\nu}_j - \hat{a}_\nu) \right\},$$

where  $\hat{\nu}_j = E(V_j | y_j, \hat{\theta})$  can be evaluated via (A.2) with  $\theta$  replaced by  $\hat{\theta}$ , and  $\hat{a}_\nu$  is  $a_\nu$  in (5) with  $\nu$  replaced by  $\hat{\nu}$ .

We further make some remarks on the implementation of the proposed ECM algorithm.

*Remark 1* To assess the convergence based on the monotonicity property of the algorithm, we adopt the Aitken’s acceleration method (cf. Aitken 1926; Böhning et al. 1994), which outperforms the lack of progress criterion and allows to avoid the premature convergence (McNicholas et al. 2010). Denote by  $l^{(k)}$  the log-likelihood value evaluated at  $\hat{\theta}^{(k)}$ . The asymptotic estimate of the log-likelihood at iteration  $k$  can be calculated as

$$l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1 - a^{(k)}} (l^{(k+1)} - l^{(k)}),$$

where  $a^{(k)} = (l^{(k+1)} - l^{(k)}) / (l^{(k)} - l^{(k-1)})$  is called the Aikten acceleration factor. Lindsay (1995) proposed that the algorithm can be considered to have converged when  $\ell_{\infty}^{(k)} - \ell^{(k)} < \epsilon$ , where  $\epsilon$  is the desired tolerance.

*Remark 2* Analogous to other iterative optimization procedures, one needs to search for appropriate initial values to avoid divergence or time-consuming computation. A direct way of deriving the initial estimate for mean vector, factor loading and error covariance matrix can be obtained by performing a simple FA fit using the *factanal* command in the R package. The resulting estimates are taken as initial values, namely  $\hat{\mu}^{(0)}$ ,  $\hat{B}^{(0)}$  and  $\hat{D}^{(0)}$ , respectively. Next, compute the factor scores via the conditional prediction method. The initial skewness vector  $\hat{\lambda}^{(0)}$  and df  $\hat{\nu}^{(0)}$  are obtained by fitting the rMST distribution to the sample of factor scores via the R package EmSkew (Wang et al. 2009).

*Remark 3* A number of information criteria taking the form of a penalized log-likelihood  $-2\ell_{\max} + C(n)m$  are used for model selection and determination of  $q$ , where  $\ell_{\max}$  is the maximized log-likelihood and  $m$  is the number of free parameters in the considered model. Five popular criteria are considered in later analysis, including the Akaike information criterion (AIC; Akaike 1973) with  $C(n) = 2$ , the consistent version of AIC (CAIC; Bozdogan 1987) with  $C(n) = \log(n) + 1$ , the Bayesian information criterion (BIC; Schwarz 1978) with  $C(n) = \log(n)$ , the sample-size adjusted BIC (SABIC; Sclove 1987) with  $C(n) = \log((n + 2)/24)$ , and the Hannan–Quinn criterion (HQC, Hannan and Quinn 1979) with  $C(n) = 2 \log(\log(n))$ . When several competing models are compared, the models with smaller values of these criteria are favored on the basis of fit and parsimony.

#### 4 Provision of standard errors

Under regularity conditions (Zacks 1971), the asymptotic covariance matrix of  $\hat{\theta}$  can be approximated by the inverse of the observed information matrix; see also Efron and Hinkley (1978). Specifically, the observed information matrix is defined as

$$I(\hat{\theta}; y) = - \frac{\partial^2 \ell(\theta; y)}{\partial \theta \partial \theta^T} \Big|_{\theta = \hat{\theta}}$$

To obtain  $I(\hat{\theta}; y)$  numerically, Jamshidian (1997) suggested using the central difference method. Let  $G = [g_1; \dots; g_m]$  be a  $m \times m$  matrix with the  $c$ th column being

$$g_c = \frac{s(\theta + h_c e_c; y) - s(\theta - h_c e_c; y)}{2h_c}, \quad c = 1, \dots, m,$$

where  $s(\theta; y) = \partial \ell(\theta; y) / \partial \theta$  is the score vector of  $\ell(\theta; y)$ ,  $e_c$  is a unit vector with all of its elements equal to zero except for its  $c$ th element which is equal to 1,  $h_c$  is a small number, and  $m$  is the number of parameters in  $\theta$ . Explicit expressions for the elements of  $s(\theta; y)$  are summarized in Supplementary Appendix B.

Since  $G$  may not be symmetric, we suggest using

$$\tilde{I}(\hat{\theta}; y) = -\frac{G + G^T}{2}. \quad (14)$$

to approximate  $I(\hat{\theta}; y)$ . The asymptotic standard errors of  $\hat{\theta}$  can be calculated by taking the square roots of the diagonal elements of  $[\tilde{I}(\hat{\theta}; y)]^{-1}$ .

Notably, the inverse of (14) is not always guaranteed to yield proper (positive) standard errors. The parametric bootstrap method (Efron and Tibshirani 1986), although computationally expensive, is often used instead to obtain estimates of the standard errors. Let  $f(y; \hat{\theta})$  be the estimated density function of (7) obtained from fitting the STFA model to the original data. The calculation of bootstrap standard error estimates consists of the following four steps.

1. Drawing a bootstrap sample  $y_1^*, \dots, y_n^*$  from the fitted distribution  $f(y; \hat{\theta})$ .
2. Compute the ML estimates  $\hat{\theta}^*$  from fitting the STFA model to the generated bootstrap samples  $y_1^*, \dots, y_n^*$ .
3. Repeat Steps 1 and 2 a large number of times, say  $B$ , thereby obtaining bootstrap replications, namely  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .
4. Estimate the bootstrap standard errors of  $\hat{\theta}$  via the sample standard errors of  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .

## 5 Numerical examples

### 5.1 A simulation study

We conduct a simulation study to examine the performance of the STFA, GHSTFA and GHCSTFA approaches. We implement the alternating expectation conditional maximization (AECM) algorithms described in Murray et al. (2014a, b) for fitting the latter two models. A comparison of some characterizations among the three considered models is summarized in Table 1.

We generate artificial data from the basic FA model with  $p = 10$  and 50 variables and  $q = 2, 3$  and 4 factors, while the underlying distribution for the latent factors is non-normal. The presumed parameter values are  $\mu = 0$ ,  $B = \text{Unif}(p, q)$  and  $D = \text{Diag}\{\text{Unif}(p, q)\}$ , where  $\text{Unif}(p, q)$  denotes a  $p \times q$  matrix of random numbers drawn from a uniform distribution on the unit interval (0,1). Moreover, the latent factors  $U$  are assumed to be one of two standardized distributions with varying degrees of skewness and kurtosis, including Beta(0.1, 30) and Chi-square distribution with one df ( $\chi_1^2$ ). Indeed, the population skewness/kurtosis for Beta(0.1, 30) is 6/52 (high) and that for  $\chi_1^2$  is 2.8/12 (mild). Errors are generated from the MT distribution with zero mean, scale covariance  $D$  and  $v = 5$ .

The sample sizes evaluated range from small ( $n = 100$ ) to moderately large ( $n = 300$ ). The objective of such settings is to see how the performance of the three models varies with respect to the degree of non-normality of the latent factors across different numbers of  $p$  and  $q$ , and the sample size  $n$ . Assuming the number of latent factors is

**Table 1** Comparison of some characterizations among three different skew- $t$  factor analysis models

Model	STFA	GHSTFA	GHCSTFA
Formulation	$Y = \mu + BU + \varepsilon$ with $U w \sim rSN_p(-a_v \Lambda^{-1} \lambda, \Lambda^{-1}/w, \Lambda^{-1} \lambda)$ , $\varepsilon w \sim N_p(0, D/w)$ , $W \sim \text{gamma}(v/2, v/2)$ , $U w \perp \varepsilon w$ .	$Y = \mu + W\alpha + B\tilde{U} + \tilde{\varepsilon}$ with $\tilde{U} = W^{-1/2}U, \tilde{\varepsilon} = W^{1/2}\varepsilon$ , $U \sim N_q(0, I_q), \varepsilon \sim N_p(0, D)$ , $W^{-1} \sim \text{gamma}(v/2, v/2)$ , $\tilde{U} w \perp \tilde{\varepsilon} w$ .	$Y = B\tilde{U} + \tilde{\varepsilon}$ with $\tilde{U} = \xi + W\zeta + WU, \tilde{\varepsilon} = W^{1/2}\varepsilon$ , $U \sim N_q(0, \Omega), \varepsilon \sim N_p(0, D)$ $W^{-1} \sim \text{gamma}(v/2, v/2)$ , $\tilde{U} w \perp \tilde{\varepsilon} w$ .
Conditional distribution	$Y   (\tilde{U}, w) \sim N_p(\mu + B\tilde{U}, w^{-1}D)$ ,	$Y   (\tilde{U}, w) \sim N_p(\mu + w\alpha + B\tilde{U}, wD)$ ,	$Y   (\tilde{U}, w) \sim N_p(B\tilde{U}, wD)$ ,
Marginal distribution	$\tilde{B} = B\Lambda^{-1/2}$ and $\tilde{U} = \Lambda^{1/2}U$ . $Y \sim rSt_p(\mu - a_v \alpha, \Sigma, \alpha, v)$ ,	$\tilde{U} = W^{-1/2}U$ . $Y \sim GST_p(\mu, \Sigma, \alpha, v)$ ,	$\tilde{U} w \sim N_p(\xi + w\zeta + wU, w\Omega)$ $Y \sim GST_p(B\xi, \Sigma, B\zeta, v)$ ,
Mean	$\Sigma = B\Lambda^{-1}B^T + D$ and $\alpha = B\Lambda^{-1/2}\lambda$ . $E(Y) = \mu$	$\Sigma = BB^T + D$ $E(Y) = \mu + \frac{v}{v-2}\alpha$	$\Sigma = B\Omega B^T + D$ . $E(Y) = B\xi + \frac{v}{v-2}\zeta$
Covariance matrix	$\text{Cov}(Y) = \frac{v}{v-2}(BB^T + D)$	$\text{Cov}(Y) = \frac{v}{v-2}(BB^T + D)$ $+ \frac{2v^2}{(v-2)^2(v-4)}\alpha\alpha^T$	$\text{Cov}(Y) = \frac{v}{v-2}(B\Omega B^T + D)$ $+ \frac{2v^2}{(v-2)^2(v-4)}B\zeta\zeta^T B^T$
No. parameters	$m = p(q+2) + q - q(q-1)/2 + 1$	$m + p - q$	$p + q(p+2) + q(q+1)/2 - q^2 + 1$

**Table 2** Numbers of free parameters involved in three skew-*t* factor analysis models

<i>q</i>	<i>p</i> = 10			<i>p</i> = 50		
	STFA	GHSTFA	GHCSTFA	STFA	GHSTFA	GHCSTFA
1	32	41	23	152	201	103
2	42	50	34	202	250	154
3	51	58	44	251	298	204

**Table 3** Simulation results based on 100 replications

<i>n</i>	<i>p</i>	<i>q</i>		High skew/kurtosis Beta (0.1,30)			Mild skew/kurtosis $\chi^2_1$		
				STFA	GHSTFA	GHCSTFA	STFA	GHSTFA	GHCSTFA
100	10	2	Mean	2,603	2,652	2,615	2,675	2,774	2,760
			Freq.	86	0	14	80	0	20
100	10	3	Mean	2,749	2,793	2,773	2,846	2,892	2,864
			Freq.	95	0	5	91	0	9
100	10	4	Mean	2,859	2,896	2,892	2,976	3,013	29,96
			Freq.	100	0	0	95	0	5
100	50	2	Mean	11,970	12,218	11,945	12,039	12,266	12,022
			Freq.	24	0	76	21	0	79
100	50	3	Mean	12,448	12,690	12,494	12,529	12,750	12,594
			Freq.	73	0	27	67	0	33
100	50	4	Mean	12,819	13,054	12,903	12,923	13,136	13,033
			Freq.	89	0	11	84	0	16
300	10	2	Mean	7,606	7,770	7,670	7,843	7,951	7,902
			Freq.	100	0	0	100	0	0
300	10	3	Mean	7,952	8,050	8,047	8,274	8,370	8,346
			Freq.	100	0	0	100	0	0
300	10	4	Mean	8,426	8,512	8,532	8,779	8,855	8,854
			Freq.	100	0	0	100	0	0
300	50	2	Mean	34,792	35,255	34,875	35,017	35,407	35,048
			Freq.	100	0	0	75	0	25
300	50	3	Mean	35,794	36,242	35,947	36,074	36,450	36,181
			Freq.	100	0	0	95	0	5
300	50	4	Mean	36,737	37,166	36,983	37,078	37,442	37,291
			Freq.	100	0	0	100	0	0

known, each simulated datum is fitted using the three considered models. Simulation results are based on the 100 repeated Monte Carlo samples.

The numbers of parameters contained in the three models with various *p* and *q* are listed in Table 2. As can be seen, the model complexity of STFA falls between GHSTFA and CHCSTFA. For ease of exposition, comparisons made in Table 3 are

based on the average BIC values together with the frequencies of the particular model chosen based on the smallest BIC value. In 23 out of a total of 24 scenarios, the STFA model provides a better fit than the other two GH-based approaches. The performance of STFA can be improved as the degree of non-normality and the sample size increase. The GHSTFA model has the worst fit and is never chosen as it is penalized more heavily by BIC.

In this study, it is important to note that the results are limited to the extent of our simulation experiments. We are certainly not making a claim that the STFA model can replace any of the others. A comparison of our proposed model versus any of the models in Murray et al. (2013, 2014a, b) is of limited practical value. This is because each of the models applies to different situations. For example, if the data were simulated from our proposed model, we can so specify it to ensure that it will produce superior results to those based on the models in Murray et al. (2013, 2014a, b). Likewise, if the data were generated from the GHST distribution as in Murray et al. (2014b), then the configuration of the latter can be chosen to make our model have relative inferior performance. This study contributes to providing the user with a wider choice of existing skew factor models to cover distinct situations that might arise in practice.

## 5.2 The AIS data set

As an illustration, we apply the proposed technique to the Australian Institute of Sport (AIS) data, which were originally reported by Cook and Weisberg (1994) and subsequently analyzed by Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999, 2003) and Azzalini (2005), among others. The dataset consists of  $p = 11$  physical and hematological measurements on athletes in different sports which are almost equally bisected between 102 male and 100 female.

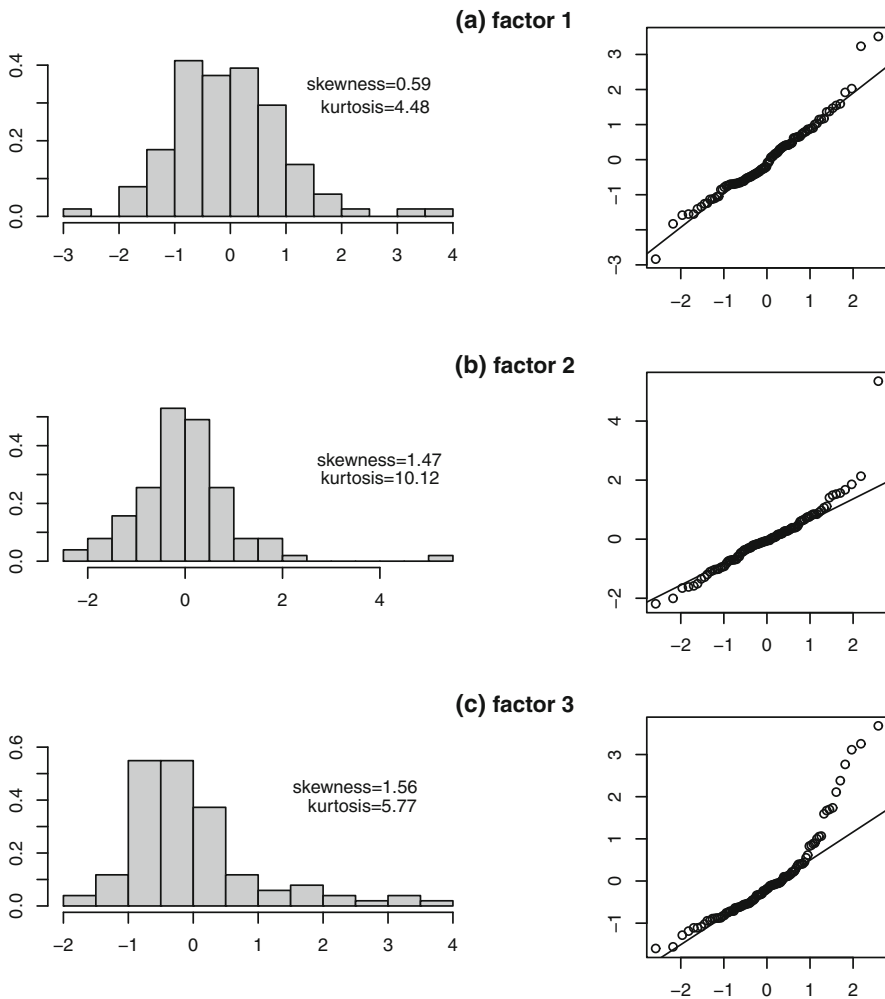
For simplicity of illustration, we focus solely on  $n = 102$  observations of male. A summary of the 11 attributes along with their sample skewness and kurtosis is given in Table 4. It is readily seen that most of the attributes are moderately to strongly skewed with a heavy tail.

**Table 4** An overview of 11 attributes of 102 male athletes of the AIS data

Attribute	Variable	Description	Skewness	Kurtosis
$x_1$	rcc	Red cell count	0.924	7.730
$x_2$	wcc	White cell count	0.859	4.579
$x_3$	Hc	Hematocrit	1.489	10.374
$x_4$	Hg	Hemoglobin	0.974	5.312
$x_5$	Fe	Plasma ferritin concentration	0.877	3.133
$x_6$	bmi	Body mass index	1.411	5.986
$x_7$	ssf	Sum of skin folds	1.386	4.789
$x_8$	Bfat	Body fat percentage	1.528	5.080
$x_9$	lbm	Lean body mass	0.274	3.621
$x_{10}$	Ht	Height (cm)	0.072	3.001
$x_{11}$	Wt	Weight (kg)	0.390	3.410

Figure 1 depicts the histograms and corresponding normal quantile plots of the first three factor score estimates obtained from the classical FA with  $q = 4$ . The factor score estimates are obtained using the “regression” method, see Chapter 9.5 of Johnson and Wichern (2007). The histograms in the left panels indicate that the distributions of factor scores deviate from normality due to positive skewness and high excess kurtosis. This feature can also be demonstrated through the normal quantile–quantile plots shown in the right panels. The result motivates us to advocate the use of STFA model as a proper tool for the analysis of this data set.

Next, we are interested in comparing the ML results of STFA with three of its nested models, namely the FA, tFA and SNFA models. To assess the performance



**Fig. 1** Histograms and corresponding normal quantile plots of the estimated factor scores obtained from fitting FA to 102 male athletes of the AIS data

**Table 5** Comparison of ML estimation results on 102 male athletes.

Model	$q$	$\ell_{\max}$	$m$	AIC	CAIC	BIC	SABIC	HQC
FA	3	-788.9	52	1,681.7	1,870.2	1,818.2	1,654.0	1737.0
	4	-639.7	60	1,399.3	1,616.8	1,556.8	1,367.3	1,463.1
	5	-637.2	67	1,408.4	1,651.3	1,584.3	1,372.6	1,479.6
	6	-633.8	73	1,413.6	1,678.3	1,605.3	1,374.7	1,491.2
tFA	3	-709.7	53	1,525.4	1,717.5	1,664.5	1,497.1	1581.7
	4	-590.7	61	1,303.4	1,524.5	1,463.5	1,270.8	1,368.2
	5	-588.5	68	1,313.0	1,559.5	1,491.5	1,276.7	1,385.3
	6	-586.1	74	1,320.2	1,588.4	1,514.4	1,280.7	1,398.9
SNFA	3	-761.5	55	1,632.8	1,832.2	1,777.2	1,603.45	1691.26
	4	-609.6	64	1,347.1	1,579.1	1,515.1	1,313.0	1,415.2
	5	-606.9	72	1,357.8	1,618.8	1,546.8	1,319.4	1,434.3
	6	-615.1	79	1,388.1	1,674.5	1,595.5	1,346.0	1,472.1
STFA	3	-682.1	56	1,476.3	1,679.3	1,623.1	1,446.4	1535.8
	<b>4</b>	<b>-560.9</b>	<b>65</b>	<b>1,251.8</b>	<b>1,487.5</b>	<b>1,422.5</b>	<b>1,217.2</b>	<b>1,320.9</b>
	5	-557.8	73	1,261.5	1,526.2	1,453.2	1,222.6	1,339.1
	6	-557.1	80	1274.1	1,564.1	1,484.1	1,231.5	1,359.2
GHSTFA	3	-693.6	64	1,515.2	1,747.2	1,683.2	1,481.0	1,583.2
	4	-574.4	72	1,292.9	1,553.9	1,481.9	1,254.5	1,369.4
	5	-572.4	79	1,302.7	1,589.1	1,510.1	1,260.6	1,386.7
	6	-568.4	85	1,306.8	1,614.9	1,529.9	1,261.4	1,397.1
GHCSTFA	3	-714.0	48	1,523.9	1,697.9	1,649.9	1,498.3	1,575.0
	4	-593.1	58	1,302.3	1,512.5	1,454.5	1,271.3	1,363.9
	5	-585.5	67	1,305.1	1,547.9	1,480.9	1,269.3	1,376.3
	6	-583.1	75	1,316.1	1,588.0	1,513.0	1,276.1	1,395.8

The best model chosen by the five information criteria was shown in bold

of non-nested models, comparisons are also made on the GHSTFA and CHCSTFA approaches. The data have been standardized to have zero mean and unit standard deviation to avoid variables having a greater impact due to different scales. We fit these models with  $q$  ranging from 3 to 6 using the ECM algorithm developed in Sect. 3. Notice that the choice of maximum  $q = 6$  satisfies the restriction  $(p - q)^2 \geq (p + q)$  as suggested by Eq. (8.5) of [McLachlan and Peel \(2000\)](#).

A summary of ML fitting results, including the maximized log-likelihood values, the number of parameters together with five model selection indices, is reported in Table 5. From this table, the model selected by the five information criteria is the STFA model with  $q = 4$ . Table 6 reports the ML solutions of the best chosen model along with the standard errors in parentheses obtained using 500 bootstrap replications. We found that the estimated skewness parameters are moderately to highly significant, revealing that the joint distribution of latent factors is skewed. Moreover, the estimated  $df$  ( $\hat{\nu} = 6.15$ ) is quite small, confirming the presence of thick tails.

Observing the unrotated solution of factor loadings displayed in the 3–6th columns of Table 6, the first factor can be labeled *general nutritional status*, with a very high



**Table 6** Summary ML results together with the associated standard errors in parentheses for the best chosen model

Variable	Parameter					
	$\mu$	$B$			$d$	
rcc	-0.034 (0.091)	-0.057 (0.087)	0.527 (0.080)	0.414 (0.094)	-0.057 (0.080)	0.203 (0.032)
wcc	-0.015 (0.099)	-0.044 (0.087)	-0.095 (0.096)	0.346 (0.098)	-0.018 (0.100)	0.699 (0.118)
Hc	-0.030 (0.089)	-0.051 (0.083)	0.631 (0.077)	0.457 (0.091)	0.026 (0.076)	0.015 (0.017)
Hg	-0.040 (0.089)	-0.041 (0.084)	0.589 (0.078)	0.445 (0.095)	0.134 (0.081)	0.124 (0.025)
Fe	-0.042 (0.100)	-0.007 (0.093)	-0.138 (0.094)	0.148 (0.098)	0.201 (0.098)	0.723 (0.117)
bmi	-0.015 (0.064)	0.443 (0.056)	-0.091 (0.056)	0.451 (0.072)	0.395 (0.050)	0.002 (0.001)
ssf	0.021 (0.072)	0.023 (0.055)	-0.341 (0.062)	0.652 (0.058)	0.030 (0.057)	0.001 (0.007)
Bfat	0.029 (0.074)	-0.028 (0.054)	-0.360 (0.061)	0.665 (0.055)	0.072 (0.058)	0.001 (0.002)
lbm	0.004 (0.043)	0.761 (0.042)	-0.077 (0.059)	0.325 (0.086)	-0.083 (0.033)	0.001 (0.0003)
Ht	0.043 (0.052)	0.577 (0.047)	-0.147 (0.062)	0.199 (0.089)	-0.642 (0.040)	0.0001 (0.002)
Wt	0.006 (0.041)	0.654 (0.041)	-0.156 (0.055)	0.447 (0.079)	-0.047 (0.030)	0.0001 (0.0001)
			$\lambda$			$\nu$
		-3.778 (0.589)	-14.868 (2.841)	19.357 (2.102)	3.413 (0.411)	6.15 (1.610)

loading on lbm, followed by Wt, Ht and bmi. The second factor, which loads heavily on rcc, Hc and Hg, might be called a *hematological factor*. The third factor can be viewed as *overweight assessment indices* since the bmi, ssf and Bfat load highly on this factor. The fourth factor is not easily interpreted at this point.

The comparison process is also conducted for the original (non-standardized) data. Clearly, as shown in Supplementary Figure 1, the STFA still provides the best overall fit. The fit of FA is the worst, indicating a lack of adequacy of normality assumptions for this dataset. It is also noted that all criteria prefer four-factor solutions under all scenarios.

We consider diagnostics to assess the validity of the underlying distributional assumption of  $Y$ . For FA, we can use the Mahalanobis-like distance  $(y-\hat{\mu})^T \hat{\Omega}^{-1} (y-\hat{\mu})$ ,

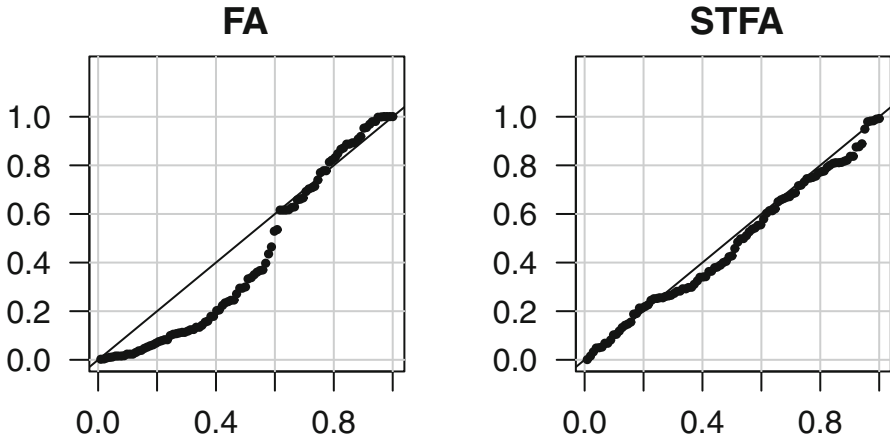


Fig. 2 Healy's plot for assessing the goodness of fit of fitted models

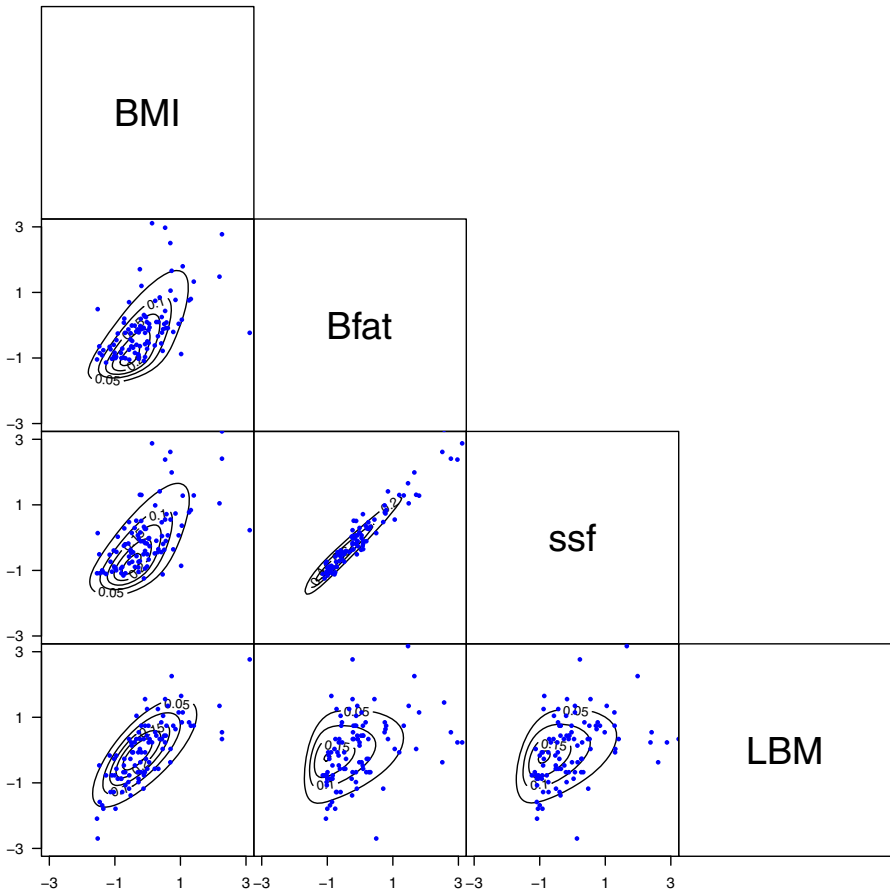


Fig. 3 Scatter plots of pairs of four selected variables of 102 male AIS athletes and coordinate projected contours

which has an asymptotic Chi-square distribution with  $p$  df. Checking the normality assumption can be achieved by constructing Healy (1968) plot. To further assess the goodness of fit of STFA, it follows from Supplementary Proposition 3 that  $f_j = (y_j - \hat{\mu})^T \hat{\Omega}^{-1} (y_j - \hat{\mu}) / p$  follows the  $F(p, \nu)$  distribution for  $j = 1, \dots, n$ . In this case, one can construct another Healy-type plot (or Snedecor's F plot) by plotting the cumulative  $F(p, \nu)$  probabilities associated with the ordered values of  $f_j$  against their nominal values  $1/n, 2/n, \dots, 1$ . As such, one can examine whether the corresponding Healy's plot resembles a straight line through the origin having unit slope. In other words, the greater the departure from the 45-degree line, the greater the evidence for concluding a poor fitting of the model. Inspecting Healy's plots shown in Fig. 2, the STFA adapts the identity more closely than does the FA, suggesting that it is appropriate to use a skew and heavy-tailed distribution.

Figure 3 depicts coordinate projected scatter plots for each pair of four selected variables superimposed with the marginal contours obtained by marginalization of the best fitted STFA model. A visual inspection reveals that the fitted contours adapt the shape of the scattering pattern satisfactorily. To summarize, the implementation of STFA tends to be more reasonable for analyzing this data set.

## 6 Conclusion

We introduce an extension of FA models obtained by replacing the normality assumption for the latent factors and errors with a joint rMST distribution, called the STFA model, as a new robust tool for dimensionality reduction. The model accommodates both asymmetry and heavy tails simultaneously and allows practitioners for analyzing data in a wide variety of situations. We have described a four-level hierarchical representation for the STFA model and presented a computationally analytical ECM algorithm for ML estimation in a flexible complete-data framework. We demonstrate our approach with a simulation study and an analysis of the AIS data set. The numerical results show that the STFA model performs relatively well for the experimental data. The computer programs used in the analyses can be downloaded from <http://amath.nchu.edu.tw/www/teacher/tilin/STFA.html>.

In the situation with the occurrence of missing data, our algorithm can be easily modified to account for missingness based on the scheme proposed in Lin et al. (2006) and Liu and Lin (2014). Due to recent advances in computer power and availability, it would be interesting to develop Markov chain Monte Carlo (MCMC) methods (Lin et al. 2009 and Lin and Lin 2011) for Bayesian estimation of the STFA model. It is also of interest to consider a finite mixture representation of STFA models. Our initial work on the latter problem has been limited to mixtures of factor analyzers with a skew-normal distribution (Lin et al. 2013).

**Acknowledgments** We are grateful to the Editor-in-Chief, the Associate Editor and two anonymous referees for their insightful comments and suggestions, which led to a much improved version of this article. This research was supported by MOST 103-2118-M-005-001-MY2 awarded by the Ministry of Science and Technology of Taiwan.

## References

- Aas K, Haff IH (2006) The generalised hyperbolic skew student's  $t$ -distribution. *J Financ Econ* 4:275–309
- Aitken AC (1926) On Bernoulli's numerical solution of algebraic equations. *Proc R Soc Edinburgh* 46:289–305
- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (Eds.) 2nd international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281
- Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, New York
- Azzalini A (1985) A class of distributions which includes the normal ones. *Scand J Stat* 12:171–178
- Azzalini A (2005) The skew-normal distribution and related multivariate families. *Scand J Stat* 32:159–188
- Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. *J R Stat Soc Ser B* 61:579–602
- Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew  $t$ -distribution. *J R Stat Soc Ser B* 65:367–389
- Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. *Biometrika* 83:715–726
- Azzalini A, Genton MG (2008) Robust likelihood methods based on the skew- $t$  and related distributions. *Int Stat Rev* 76:106–129
- Barndorff-Nielsen O, Shephard N (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *J R Stat Soc Ser B* 63:167–241
- Basilevsky A (2008) Statistical factor analysis and related methods: theory and applications. Wiley, New York
- Böhning D, Dietz E, Schaub R, Schlattmann P, Lindsay B (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann Inst Stat Math* 46:373–388
- Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52:345–370
- Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. *J Multivar Anal* 79:99–113
- Cook RD, Weisberg S (1994) An introduction to regression graphics. Wiley, New York
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc Ser B* 39:1–38
- Efron B, Hinkley DV (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information (with discussion). *Biometrika* 65:457–487
- Efron B, Tibshirani R (1986) Bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1:54–77
- Fokoué E, Titterton DM (2003) Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. *Mach Learn* 50:73–94
- Hannan EJ, Quinn BG (1979) The determination of the order of an autoregression. *J R Stat Soc Ser B* 41:190–195
- Healy MJR (1968) Multivariate normal plotting. *Appl Stat* 17:157–161
- Ho HJ, Lin TI, Chang HH, Haase HB, Huang S, Pyne S (2012) Parametric modeling of cellular state transitions as measured with flow cytometry different tissues. *BMC Bioinform* 13(Suppl 5):S5
- Jamshidian M (1997) An EM algorithm for ML factor analysis with missing data. In: Berkane M (ed) Latent variable modeling and applications to causality. Springer, New York, pp 247–258
- Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson Prentice-Hall, Upper Saddle River
- Jones MC, Faddy MJ (2003) A skew extension of the  $t$ -distribution with applications. *J R Stat Soc Ser B* 65:159–174
- Kotz S, Nadarajah S (2004) Multivariate  $t$  distributions and their applications. Cambridge University Press, Cambridge
- Lachos VH, Ghosh P, Arellano-Valle RB (2010) Likelihood based inference for skew normal independent linear mixed models. *Stat Sin* 20:303–322
- Lange KL, Little JJA, Taylor JMG (1989) Robust statistical modeling using the  $t$  distribution. *J Am Stat Assoc* 84:881–896
- Lawley DN, Maxwell AE (1971) Factor analysis as a statistical method, 2nd edn. Butterworth, London
- Lee S, McLachlan GJ (2013) On mixtures of skew normal and skew  $t$ -distributions. *Adv Data Anal Classif* 7:241–266

- Lee S, McLachlan GJ (2014) Finite mixtures of multivariate skew  $t$ -distributions: some recent and new results. *Stat Comp* 24:181–202
- Lee YW, Poon SH (2011) Systemic and systematic factors for loan portfolio loss distribution. *Econometrics and applied economics workshops*, School of Social Science, University of Manchester, pp 1–61
- Lin TI, Ho HJ, Chen CL (2009) Analysis of multivariate skew normal models with incomplete data. *J Multivari Anal* 100:2337–2351
- Lin TI, Lee JC, Ho HJ (2006) On fast supervised learning for normal mixture models with missing information. *Pattern Recog* 39:1177–1187
- Lin TI, Lee JC, Hsieh WJ (2007a) Robust mixture modeling using the skew  $t$  distribution. *Stat Comput* 17:81–92
- Lin TI, Lee JC, Yen SY (2007b) Finite mixture modelling using the skew normal distribution. *Stat Sin* 17:909–927
- Lin TI, Lin TC (2011) Robust statistical modelling using the multivariate skew  $t$  distribution with complete and incomplete data. *Stat Model* 11:253–277
- Lin TI, McLachlan GJ, Lee SX (2013) Extending mixtures of factor models using the restricted multivariate skew-normal distribution. Preprint [arXiv:1307.1748](https://arxiv.org/abs/1307.1748)
- Lindsay B (1995) Mixture models: theory. Geometry and applications. Institute of Mathematical Statistics, Hayward
- Liu M, Lin TI (2014) Skew-normal factor analysis models with incomplete data. *J Appl Statist*. doi:10.1080/02664763.2014.986437
- Lopes HF, West M (2004) Bayesian model assessment in factor analysis. *Stat Sin* 14:41–67
- Louis TA (1982) Finding the observed information when using the EM algorithm. *J R Stat Soc Ser B* 44:226–232
- McLachlan GJ, Peel D (2000) *Finite Mixture Models*. Wiley, New York
- McLachlan GJ, Bean RW, Jones LBT (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate  $t$ -distribution. *Comput Stat Data Anal* 51:5327–5338
- McNicholas PD, Murphy TB, McDaid AF, Frost D (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Comput Stat Data Anal* 54:711–723
- Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80:267–278
- Montanari A, Viroli C (2010) Heteroscedastic factor mixture analysis. *Stat Model* 10:441–460
- Murray PM, Browne RP, McNicholas PD (2013) Mixtures of ‘unrestricted’ skew- $t$  factor analyzers. Preprint [arXiv:1310.6224v1](https://arxiv.org/abs/1310.6224v1)
- Murray PM, Browne RP, McNicholas PD (2014a) Mixtures of skew- $t$  factor analyzers. *Comput Stat Data Anal* 77:326–335
- Murray PM, McNicholas PD, Browne RP (2014b) Mixtures of common skew- $t$  factor analyzers. *Stat* 3:68–82
- Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirov JP (2009) Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci USA* 106:8519–8524
- Rossin E, Lin TI, Ho HJ, Mentzer SJ, Pyne S (2011) A framework for analytical characterization of monoclonal antibodies based on reactivity profiles in different tissues. *Bioinformatics* 27:2746–2753
- Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with application to Bayesian regression models. *Can J Stat* 31:129–150
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Sclove LS (1987) Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52:333–343
- Spearman C (1904) General intelligence, objectively determined and measured. *Am J Psychol* 15:201–292
- Tortora C, McNicholas PD, Browne R (2013) A mixture of generalized hyperbolic factor analyzers. Preprint [arXiv: 1311.6530v1](https://arxiv.org/abs/1311.6530v1)
- Wall MM, Guo J, Amemiya Y (2012) Mixture factor analysis for approximating a non-normally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivar Behav Res* 47:276–313
- Wang K, McLachlan GJ, Ng SK, Peel D (2009) EMMIX-skew: EM algorithm for mixture of multivariate skew normal/ $t$  distributions. R package version 1.0-12
- Wang WL, Lin TI (2013) An efficient ECM algorithm for maximum likelihood estimation in mixtures of  $t$ -factor analyzers. *Comput Stat* 28:751–769

Zacks S (1971) The theory of statistical inference. Wiley, New York

Zhang J, Li J, Liu C (2013) Robust factor analysis using the multivariate  $t$ -distribution. unpublished manuscript