Journal of Multivariate Analysis 145 (2016) 162–177

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Skewed factor models using selection mechanisms

Hyoung-Moon Kim [a], Mehdi Maadooliat [b], Reinaldo B. Arellano-Valle [c], Marc G. Genton [d,*]

[a] Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea
[b] Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI, USA
[c] Departamento de Estadistica, Pontificia Universidad Católica de Chile, Santiago, Chile
[d] CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Article history:
Received 20 July 2014
Available online 21 December 2015

AMS subject classifications:
62H25
62E15

Keywords:
Mixing
Selection
Skew-normal
Skew-*t*
SUN distribution

ABSTRACT

Traditional factor models explicitly or implicitly assume that the factors follow a multivariate normal distribution; that is, only moments up to order two are involved. However, it may happen in real data problems that the first two moments cannot explain the factors. Based on this motivation, here we devise three new skewed factor models, the skew-normal, the skew-*t*, and the generalized skew-normal factor models depending on a selection mechanism on the factors. The ECME algorithms are adopted to estimate related parameters for statistical inference. Monte Carlo simulations validate our new models and we demonstrate the need for skewed factor models using the classic open/closed book exam scores dataset.

1. Introduction

Factor models are among the most used and useful statistical techniques. They date back to the seminal paper of Spearman [31] and are mainly applied in two major situations: (i) data dimension reduction; and (ii) identification of underlying structures. Since then, the potential of factor models has been discovered and continually rediscovered, even after more than a century.

Traditional factor models explicitly or implicitly assume that the factors follow a multivariate normal distribution. Therefore, only moments up to the second order are involved, although it may happen in real problems that the first two moments cannot explain the factors. Based on this motivation, here we devise three new skewed factor models, namely the skew-normal, the skew-*t*, and the generalized skew-normal factor models using selection mechanisms [4,2].

In recent years, there has been a growing interest in constructing parametric classes of non-normal distributions. For instance, the univariate skew-normal distribution has been developed by Azzalini [6] and then further extended to the multivariate setting by Azzalini and Dalla Valle [10] and Azzalini and Capitanio [7]. The multivariate skew-*t* distribution was developed by Branco and Dey [13,14], Azzalini and Capitanio [8], and Gupta [19]. Skew-normal distributions have been used in many robust analyses, see, e.g., [11]. Scale mixtures of skew-normal distributions were studied by Branco and Dey [13] and include (skew-) normal distributions as special cases. These distributions have been further extended to skew-elliptical distributions by many authors, see for example the books by Genton [17] and Azzalini and Capitanio [9] and

* Corresponding author.
E-mail address: marc.genton@kaust.edu.sa (M.G. Genton).

http://dx.doi.org/10.1016/j.jmva.2015.12.007
0047-259X/

references therein. All these skewed distributions can be cast in the framework of selection distributions that arise under various selection mechanisms; see [4].

Relaxing the normality assumption of the factors is not new. Pison et al. [27] proposed a principal factor analysis method to estimate a factor analysis model that is highly robust to the effect of outliers. Yung [34] developed a confirmatory factor analysis model to handle data such that observations are drawn by several sub-populations. Obviously in this case, data are not normally distributed. This method can thus be applied to multimodal or asymmetric data. Mooijaart [26] proposed an asymptotic distribution-free method using all the cross-product moments up to the third order. However, this approach is computationally demanding with many variables. Montanari and Viroli [25] devised a skew-normal factor model for the analysis of student satisfaction in university courses. They assumed that the factors follow a skew-normal distribution and the error term follows a normal distribution. However, this approach requires more parameters be estimated than in normal-based factor analysis because of the shape parameters in the skew-normal distribution. Furthermore, skew-normality should be tested after applying the method on the factors. Recently, Bagnato and Minozzo [12] proposed a spatial latent factor model to deal with multivariate geostatistical skew-normal data. In this model they assume that the unobserved latent structure, responsible for the correlation among different variables as well as for the spatial autocorrelation among different sites is normal, and that the observed variables are skew-normal.

We, instead, use a selection mechanism [4] approach to build skewness in the factor model. The work of Montanari and Viroli [25] was motivated by examples that involve various forms of selection mechanisms and lead to skewed distributions. In this direction, we assume that there is independent normality between the factors and the error term, and then the skew-normal distribution appears in a natural way by a selection mechanism that chooses positivity of the factors. The resulting marginal distributions of the observed variables are the unified skew-normal (SUN) [2] and the unified skew-t (SUT) [5] distributions. We can show that the skewed factor models obtained by selection mechanisms contain the model of Montanari and Viroli [25] as a special case. The proof is given in Section 2.4.

This paper is organized as follows. In Section 2, we develop three new skewed factor models based on selection mechanisms. They are the skew-normal, skew-t, and generalized skew-normal factor models depending on a selection mechanism on the factors. Statistical aspects are considered in Section 3. Some simulation results are presented in Section 4. To illustrate the performance of the proposed methods on a real dataset, we use the classic open/closed book exam scores dataset in Section 5. Finally, Section 6 provides conclusions.

2. Skewed factor models

2.1. Motivation

The traditional k-factor model is defined as follows:

$$y = \mu + \Lambda f + \epsilon, \tag{1}$$

where μ is a $p \times 1$ vector of constants, Λ is a $p \times k$ matrix of constants, f and ϵ are $k \times 1$ and $p \times 1$ random vectors, and $k \leq p$. The elements of f are called common factors and the elements of ϵ are called specific or unique factors. Usually, f follows a multivariate normal distribution, $\mathcal{N}_k(0, I_k)$, and, independent of f, ϵ is $\mathcal{N}_p(0, \Psi)$, where I_k is the $k \times k$ identity matrix and $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$. Thus, it follows that $\mathrm{E}(y) = \mu$, $\mathrm{var}(y) = \Lambda\Lambda^\top + \Psi$ and $\mathrm{cov}(y, f) = \Lambda$.

One connection between the skewed factor models and the normal factor model is that the factor loadings, Λ, are determined only up to an orthogonal random sign matrix, P, if we relax the possible change of signs in the factor loadings. It is well known that factor loadings are determined only up to an orthogonal matrix, P. Under model (1), let $P = \mathrm{diag}\{\mathrm{sgn}(f_i)\}$, where the signum function of a real number, x, is defined as:

$$\mathrm{sgn}(x) = \begin{cases} -1, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

Since $P = P^\top$ is orthogonal and $f = P|f|$, then model (1) becomes (2) with the new factor loadings, ΛP, which are different only up to a possible change of sign in each row of the factor loadings. That is,

$$y = \mu + \Lambda P P^\top f + \epsilon = \mu + \Lambda P|f| + \epsilon. \tag{2}$$

This model is called the skew-normal factor model discussed in the next section. A similar approach can be taken for the skew-t factor model, (4), and the generalized skew-normal factor model, (6). This is the reason why we adopted the skewed models (3), (4) and (6). By doing so, we can handle skewed and/or heavy tailed data.

2.2. The skew-normal factor model

Under model (1), suppose that $f = (f_1, \ldots, f_k)^\top > 0$; that is, $f_i > 0$, $i = 1, \ldots, k$. Then,

$$x \overset{d}{=} [y|f > 0] \overset{d}{=} \mu + \Lambda(f|f > 0) + \epsilon \overset{d}{=} \mu + \Lambda|f| + \epsilon, \tag{3}$$

where $|f| = (|f_1|, \ldots, |f_k|)^\top$ is equal in distribution to $f|f > 0$. Hence, we have the following theorem based on some well-known properties of the multivariate normal distribution. In the sequel, all proofs are relegated to the Appendix.

Theorem 1. *Under model (1) with $f > 0$, we have that*

(i) $x|\,|f| \sim \mathcal{N}_p(\mu + \Lambda|f|, \Psi)$,
(ii) $|f| \sim \mathcal{TN}_k(0, I_k, 0)$,
(iii) $|f|\,|x \sim \mathcal{TN}_k(\xi, \Omega, 0)$,

where $\xi = \Lambda^\top\left(\Lambda\Lambda^\top + \Psi\right)^{-1}(x-\mu)$, and $\Omega = I_k - \Lambda^\top\left(\Lambda\Lambda^\top + \Psi\right)^{-1}\Lambda$. Here, $\mathcal{TN}_k(\xi, \Omega, 0)$ denotes the truncated k-dimensional normal distribution with the left truncation point at 0.

From now on, to obtain the inverse of $\Lambda\Lambda^\top + \Psi$ whenever it is needed, we use the Woodbury formula [18]; that is,

$$\left(\Lambda\Lambda^\top + \Psi\right)^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda\left(I + \Lambda^\top\Psi^{-1}\Lambda\right)^{-1}\Lambda^\top\Psi^{-1},$$

which is easier to evaluate than the original matrix inversion since $k \le p$, usually $k < p$, and Ψ is a diagonal matrix whose inverse is easy to calculate. This formula is particularly helpful for the calculation of the marginal distributions of x when the ECME algorithms [23] are applied to estimate the parameters. Based on model (3), we can find the marginal distribution of x in the following theorem, where we use the same notation as Arellano-Valle and Azzalini [2].

Theorem 2. *The random vector x defined at (3) follows a unified skew-normal distribution, $\mathcal{SUN}_{p,k}\left(\mu, 0, 1_p, \Omega^*\right)$, where*

$$\Omega^* = \begin{pmatrix} I_k & \Lambda^\top \\ \Lambda & \Lambda\Lambda^\top + \Psi \end{pmatrix}.$$

The marginal density of x is thus given by

$$f(x) = 2^k\phi_p\left(x; \mu, \Lambda\Lambda^\top + \Psi\right)\Phi_k(\xi; \Omega), \quad x \in \mathbb{R}^p,$$

where ξ and Ω are given in Theorem 1. Here, $\Phi_k(\alpha; \Sigma)$ is defined as the usual normal cumulative distribution function (cdf) with mean 0 and covariance Σ, i.e., $\Phi_k(\alpha; \Sigma) = \Pr(X < \alpha)$, where $X \sim \mathcal{N}_k(0, \Sigma)$. See [2] for more general classes of SUN distributions.

Compared to the mean and covariance structures of the traditional factor model, we give those of the skew-normal factor model. Note that $\mathrm{E}(|f|) = \sqrt{2/\pi}\,1_k$ and $\mathrm{var}(|f|) = (1 - 2/\pi)\,I_k$, where 1_k denotes a k-dimensional column vector with the element of 1. Using the laws of total expectation and total variance, we have that

$$\mathrm{E}(x) = \mu + \sqrt{\frac{2}{\pi}}\Lambda 1_k, \qquad \mathrm{var}(x) = \left(1 - \frac{2}{\pi}\right)\Lambda\Lambda^\top + \Psi, \qquad \mathrm{cov}(x, |f|) = \left(1 - \frac{2}{\pi}\right)\Lambda.$$

Note that, unlike the standard factor model, in our case, the loading matrix Λ affects also the means of the observed vector, x. These coincide with the results of Arellano-Valle and Azzalini [2], and can be written in componentwise expressions as follows:

$$\mathrm{E}(x_i) = \mu_i + \sqrt{\frac{2}{\pi}}\sum_{j=1}^{k}\lambda_{ij}, \qquad \mathrm{var}(x_i) = \left(1 - \frac{2}{\pi}\right)\sum_{j=1}^{k}\lambda_{ij}^2 + \psi_i,$$

$$\mathrm{cov}(x_i, x_j) = \left(1 - \frac{2}{\pi}\right)\sum_{l=1}^{k}\lambda_{il}\lambda_{jl}, \qquad \mathrm{cov}(x_i, |f_j|) = \left(1 - \frac{2}{\pi}\right)\lambda_{ij}.$$

To estimate parameters, we need to find the moments of the truncated k-dimensional normal distribution; that is, $\mathrm{E}(g|x)$ and $\mathrm{cov}(g|x)$, where $g = |f|$. We further define $h = g - \xi$. Then, $h|x \sim \mathcal{TN}_k(0, \Omega, -\xi)$. Hence, $\mathrm{E}(g|x) = \mathrm{E}(h|x) + \xi = \{\mathrm{E}(g_i|x)\}$ and $\mathrm{cov}(g|x) = \mathrm{cov}(h|x) = \{\mathrm{E}(g_ig_j|x) - \mathrm{E}(g_i|x)\mathrm{E}(g_j|x)\}$, where $g = (g_1, \ldots, g_k)^\top$, $h = (h_1, \ldots, h_k)^\top$, and $h_i = g_i - \xi_i$, $i = 1, \ldots, k$. We therefore have the moments using the method of Tallis [32].

Lemma 1. *Let $g|x \sim \mathcal{TN}_k(\xi, \Omega, 0)$. Then, the moments of $g|x$ can be obtained, for $i, j = 1, \ldots, k$, as follows:*

$$E(g_i|x) = c^{-1}\sum_{l=1}^{k}\omega_{il}F_l(-\xi_l) + \xi_i,$$

$$E(g_ig_j|x) = \omega_{ij} - c^{-1}\left[\sum_{l=1}^{k}\frac{\omega_{il}\omega_{jl}}{\omega_{ll}}\{\xi_lF_l(-\xi_l)\} - \sum_{l=1}^{k}\omega_{il}\sum_{q\neq l}\left(\omega_{jq} - \frac{\omega_{lq}\omega_{jl}}{\omega_{ll}}\right)F_{lq}(-\xi_l, -\xi_q)\right] + \xi_iE(g_j|x) + \xi_jE(g_i|x) - \xi_i\xi_j,$$

where

$$c = \Phi_k(\xi; \Omega), \qquad \Omega = \{\omega_{ij}\},$$
$$F_l(h) = \phi_1(h; 0, \omega_{ll})\Phi_{k-1}\left\{\xi_{(l)|l}(h) + \xi_{(l)}; \Omega_{(l)|l}\right\},$$

and (l) *denotes* $\{1, \ldots, l-1, l+1, \ldots, k\}$. *Note that* $\xi_{(l)|l}(h) = \Omega_{(l)l}\omega_{ll}^{-1}h$, $\Omega_{(l)|l} = \Omega_{(l)(l)} - \Omega_{(l)l}\omega_{ll}^{-1}\Omega_{l(l)}$,

$$F_{lq}(h_l, h_q) = \phi_2(h_l, h_q; 0, \Omega_{lq,lq})\Phi_{k-2}\left\{\xi_{(lq)|lq}(h_l, h_q) + \xi_{(lq)}; \Omega_{(lq)|lq}\right\},$$
$$\xi_{(lq)|lq}(h_l, h_q) = \Omega_{(lq)lq}\Omega_{lq,lq}^{-1}h, \quad \text{and} \quad \Omega_{(lq)|lq} = \Omega_{(lq)(lq)} - \Omega_{(lq)lq}\Omega_{lq,lq}^{-1}\Omega_{lq(lq)}.$$

Here, $c^{-1}F_l(h)$ and $c^{-1}F_{l,q}(h_l, h_q)$ are respectively univariate and bivariate marginal distributions of h_l and (h_l, h_q). To evaluate the mean and covariance that appear in Lemma 1, Leppard and Tallis [22] suggested a numerical approach.

From now on, we demonstrate how to use the EM-type algorithm for maximum likelihood (ML) estimation of the skew-normal factor model. For n independent observations of x, we have, given $g_1, \ldots, g_n$ and μ, Λ, Ψ, that $x_1, \ldots, x_n$ are independent with the normal distribution given in Theorem 1. The complete data are (x, g), and g is fully missing. Related parameters are $\Theta = (\mu, \Lambda, \Psi)$. We remark that $g = |f|$ follows a $\mathcal{TN}_k(0, I_k, 0)$ distribution given in Theorem 1. Hence, the log-likelihood function of Θ based on complete data, aside from additive constant terms, can be written as

$$l_c(\Theta|x, g) = -\frac{n}{2}\ln|\Psi| - \frac{1}{2}\mathrm{tr}\left(\Psi^{-1}\sum_{i=1}^{n}M_i\right),$$

where $M_i = (x_i - \mu - \Lambda g_i)(x_i - \mu - \Lambda g_i)^\top$.

At the jth iteration of the E-step, we need to calculate the Q-function, defined by

$$Q\left(\Theta|\widehat{\Theta}^{(j)}\right) = \mathrm{E}\left\{l_c(\Theta|x, g)|\widehat{\Theta}^{(j)}, x\right\}$$
$$= -\frac{n}{2}\ln|\Psi| - \frac{1}{2}\mathrm{tr}\left\{\Psi^{-1}\sum_{i=1}^{n}\mathrm{E}(M_i|\widehat{\Theta}^{(j)}, x)\right\},$$

which is the conditional distribution given by Theorem 1 and the mean and covariance are given in Lemma 1. The M-step consists in the maximization of $Q(\Theta|\widehat{\Theta}^{(j)})$ with respect to Θ. To do this, we use a faster extension of the original EM [15], called the ECME algorithm [23], by replacing the M step with a sequence of conditional maximization steps.

CM-steps:

1. Update $\mu^{(j)}$ by $\mu^{(j+1)} = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\Lambda^{(j)}\sum_{i=1}^{n}\mathrm{E}(g_i|\widehat{\Theta}^{(j)}, x)$;
2. Update $\Lambda^{(j)}$ by

$$\Lambda^{(j+1)} = \left\{\sum_{i=1}^{n}(x_i - \mu^{(j+1)})\,\mathrm{E}\left(g_i^\top|\widehat{\Theta}^{(j)}, x\right)\right\}\left\{\sum_{i=1}^{n}\mathrm{E}\left(g_ig_i^\top|\widehat{\Theta}^{(j)}, x\right)\right\}^{-1};$$

3. Find $\Psi^{(j+1)}$ to maximize the actual constrained log-likelihood given $\mu^{(j+1)}$ and $\Lambda^{(j+1)}$; that is,

$$\Psi^{(j+1)} = \arg\max_{\Psi}\sum_{i=1}^{n}\ln\{f(x_i)\},$$

which can be done, for example, using the Newton–Raphson method.

Step 3 in the above CM-steps can be replaced by

3a. Update $\Psi^{(j)}$ by

$$\Psi^{(j+1)} = \mathrm{diag}\left[\frac{1}{n}\left\{\sum_{i=1}^{n}\left(x_i - \mu^{(j+1)}\right)\left(x_i - \mu^{(j+1)}\right)^\top - \Lambda^{(j+1)}\sum_{i=1}^{n}\mathrm{E}\left(g_ig_i^\top|\widehat{\Theta}^{(j)}, x\right)(\Lambda^{(j+1)})^\top\right\}\right].$$

The marginal distribution of x is given in Theorem 1 and the probability density function (pdf) of step 3 of the CM-steps is given by

$$f(x_i) = 2^k\phi_p\left\{x_i; \mu^{(j+1)}, \Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi\right\}\Phi_k\left(\xi_i^{(j+1)}; \Omega^{(j+1)}\right), \quad x_i \in \mathbb{R}^p,$$

where $\xi_i^{(j+1)} = (\Lambda^{(j+1)})^\top\{\Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi\}^{-1}(x_i - \mu^{(j+1)})$, and $\Omega^{(j+1)} = I_k - (\Lambda^{(j+1)})^\top\{\Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi\}^{-1}\Lambda^{(j+1)}$. For the choice of the starting values of the parameters, we perform an ordinary factor analysis to obtain the starting values for the factor loading matrix, Λ, and the covariance matrix, Ψ. We use the sample mean, $\bar{x}$, as the starting value for μ.

2.3. The skew-t factor model

We adopt a similar approach as in Section 2.2 in developing the skew-t factor model. A point of difference is to allow extra variation in the common factors and unique factors using scale mixture models for incorporating the heavy tail. Hence,

under model (1), let $f > 0$; that is, $f_i > 0$, where $i = 1, \ldots, k$, $f|\eta \sim \mathcal{N}_k(0, W(\eta)I_k)$, $\epsilon|\eta \sim \mathcal{N}_p(0, W(\eta)\Psi)$ independent of $f|\eta$, and η is a mixing variable with cdf $H(\eta)$ and a weight function $W(\cdot)$. Then,

$$[x|\eta] \stackrel{d}{=} [y|f > 0, \eta] \stackrel{d}{=} \mu + \Lambda[f|f > 0, \eta] + \epsilon|\eta \stackrel{d}{=} \mu + \Lambda|f|\,|\eta + \epsilon|\eta. \tag{4}$$

The distributions of f and ϵ are called scale mixtures of normal distributions [1] with a mixing variable, η.

This family of distributions contains many distributions, for example such as the Student's t, logistic, stable, and exponential power distributions depending on $W(\cdot)$ and $H(\eta)$, but we concentrate on the t distribution for practical purposes. Even though we concentrate on the t distribution, the theory will be general. Specifically, for the t distribution, $W(\eta) = 1/\eta$ and $\eta \sim \mathcal{G}(\nu/2, \nu/2)$, where $\mathcal{G}(\alpha, \beta)$ refers to a gamma random variable with mean α/β and variance α/β^2. By standard results of scale mixtures of normal distributions [3,21], we know that $\epsilon \sim \mathrm{t}_p(0, \Psi, \nu)$ and $f \sim \mathrm{t}_k(0, I_k, \nu)$.

Similar to Theorem 2, we find that $[x|\eta] \stackrel{d}{=} [y|f > 0, \eta]$ follows a unified skew-normal distribution, $\mathcal{SUN}_{p,k}(\mu, 0, 1_p, W(\eta)\Omega^*)$, where Ω^* is given in Theorem 2. Furthermore, if $W(\eta) = 1/\eta$ and $\eta \sim \mathcal{G}(\nu/2, \nu/2)$, then we can find the marginal distribution of x in the following theorem.

Theorem 3. *The density of the marginal distribution of x defined by the random vector $x|\eta$ at (4) with $W(\eta) = 1/\eta$ and $\eta \sim \mathcal{G}(\nu/2, \nu/2)$ is given by*

$$f(x) = 2^k t_p\left(x - \mu; \Lambda\Lambda^\top + \Psi, \nu\right) T_k\left\{\xi; \frac{\nu + q(x-\mu)}{\nu + p}\Omega, \nu + p\right\}, \quad x \in \mathbb{R}^p, \tag{5}$$

where ξ and Ω are given in Theorem 1, $q(x-\mu) = (x-\mu)^\top(\Lambda\Lambda^\top + \Psi)^{-1}(x-\mu)$ and $t_l(x; \Theta, \nu)$, $T_l(x; \Theta, \nu)$ denote the density and distribution functions, respectively, of an l-dimensional t distribution with location vector 0, scale matrix Θ and ν degrees of freedom.

The corresponding distribution of Theorem 3 belongs to the unified skew-t (SUT) family defined by Arellano-Valle and Genton [5]. When $k = 1$, the density (5) reduces to the skew-t density function proposed by Azzalini and Capitanio [8], Gupta [19], and Branco and Dey [13,14]. To develop a skew-t factor model, we have the following theorem.

Theorem 4. *Under model (1) with $f > 0$, $f|\eta \sim \mathcal{N}_k(0, W(\eta)I_k)$, $\epsilon|\eta \sim \mathcal{N}_p(0, W(\eta)\Psi)$ independent of $f|\eta$, and η is a mixing variable with cdf $H(\eta)$ and a weight function $W(\cdot)$, we have that*

(i) $x|\,|f|, \eta \sim \mathcal{N}_p(\mu + \Lambda|f|, W(\eta)\Psi)$,
(ii) $|f|\,|\eta \sim \mathcal{TN}_k(0, W(\eta)I_k, 0)$,
(iii) $|f|\,|x, \eta \sim \mathcal{TN}_k(\xi, W(\eta)\Omega, 0)$,

where $\xi = \Lambda^\top(\Lambda\Lambda^\top + \Psi)^{-1}(x-\mu)$, and $\Omega = I_k - \Lambda^\top(\Lambda\Lambda^\top + \Psi)^{-1}\Lambda$.

Compared with the mean and covariance structures of the traditional factor model, here we give those of the skew-t factor model. Using the laws of total expectation, total variance, and total covariance, we have that

$$\mathrm{E}(x) = \mu + 2c_\nu \Lambda 1_k, \quad \nu > 1,$$
$$\mathrm{var}(x) = \Lambda(aI_k + bJ_k)\Lambda^\top + \frac{\nu}{\nu-2}\Psi, \quad \nu > 2,$$
$$\mathrm{cov}(x, |f|) = \Lambda(aI_k + bJ_k), \quad \nu > 2,$$

where

$$a = \frac{\nu}{\nu-2}\left(1 - \frac{2}{\pi}\right), \qquad b = \frac{2}{\pi}\frac{\nu}{\nu-2} - 4c_\nu^2, \qquad c_\nu = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}},$$

and $J_k = 1_k 1_k^\top$. These coincide with the results of the skew-normal factor model when ν goes to infinity from an extension of Stirling's formula. That is, $\lim_{\nu\to\infty} c_\nu = 1/\sqrt{2\pi}$. These can be written in componentwise expressions as follows:

$$\mathrm{E}(x_i) = \mu_i + 2c_\nu \sum_{j=1}^k \lambda_{ij},$$
$$\mathrm{var}(x_i) = b\sum_{q=1}^k \lambda_{iq}\left(\sum_{l=1}^k \lambda_{il}\right) + a\sum_{q=1}^k \lambda_{iq}^2 + \frac{\nu}{\nu-2}\psi_i,$$
$$\mathrm{cov}(x_i, x_j) = b\sum_{q=1}^k \lambda_{jq}\left(\sum_{l=1}^k \lambda_{il}\right) + a\sum_{q=1}^k \lambda_{iq}\lambda_{jq},$$
$$\mathrm{cov}(x_i, |f_j|) = b\sum_{l=1}^k \lambda_{il} + a\lambda_{ij}.$$

To estimate parameters, we need to find the moments $\mathrm{E}(\eta|x)$, $\mathrm{E}(\eta g|x)$, and $\mathrm{E}(\eta gg^\top|x)$. To derive them, the following lemma is helpful.

Lemma 2. *If $\eta \sim \mathcal{G}(\alpha, \beta)$, then for any $a \in \mathbb{R}^k$,*

$$E\left\{\Phi_k\left(a-\mu; \frac{1}{\eta}\Sigma\right)\right\} = T_k\left\{\sqrt{\frac{\alpha}{\beta}}(a-\mu); \Sigma, 2\alpha\right\}.$$

Using Lemma 2, we derive the conditional moments as follows. These results are useful in the implementation of the EM-algorithm.

Theorem 5. *Let $X|\eta \sim \mathcal{SUN}_{p,k}\left(\mu, 0, 1_p, W(\eta)\Omega^*\right)$, where Ω^* is defined in Theorem 2. Suppose that η has a density $h = H'$. Then, for any integrable function, $S(\eta)$, we have*

$$E\{S(\eta)|x\} = \frac{2^k f_0(x)}{f(x)} E\left[S(\eta)\Phi_k\{\xi; W(\eta)\Omega\}|X_0 = x\right],$$

where f and f_0 are the marginal densities of X and X_0, respectively, and X_0 is such that $X_0|\eta \sim \mathcal{N}_p(\mu, W(\eta)(\Lambda\Lambda^\top + \Psi))$.

The following corollary is also helpful in constructing an EM-algorithm.

Corollary 1. *Let $X|\eta \sim \mathcal{SUN}_{p,k}(\mu, 0, 1_p, W(\eta)\Omega^*)$, $W(\eta) = 1/\eta$, and $\eta \sim \mathcal{G}(\nu/2, \nu/2)$. We have the following moments:*

$$\text{(i) } E(\eta|X=x) = \frac{(\nu+p)\,\tilde{c}^{-1}}{\nu+q(x-\mu)} T_k\left\{\sqrt{\frac{\nu+p+2}{\nu+q(x-\mu)}}\,\xi; \Omega, \nu+p+2\right\},$$

$$\text{(ii) } E(\eta g|X=x) = \{E(\eta g_i|X=x)\} = \{\tau_i\}, \quad \text{where } g^\top = (g_1, \ldots, g_k),$$

$$\tau_i = \tilde{c}^{-1}\sum_{l=1}^{k}\omega_{il}\tilde{F}_l(-\xi_l; \nu+p+1) + \xi_i E(\eta|X=x),$$

$$\text{(iii) } E(\eta gg^\top|X=x) = \{E(\eta g_i g_j|X=x)\} = \{\varsigma_{ij}\}, \quad \text{for } i,j = 1, \ldots, k,$$

$$\begin{aligned}\varsigma_{ij} &= \omega_{ij} - \tilde{c}^{-1}\left[\sum_{l=1}^{k}\frac{\omega_{il}\omega_{jl}}{\omega_{ll}}\{\xi_l\tilde{F}_l(-\xi_l; \nu+p+1)\} - \sum_{l=1}^{k}\omega_{il}\sum_{q\neq l}\left(\omega_{jq} - \frac{\omega_{lq}\omega_{jl}}{\omega_{ll}}\right)\tilde{F}_{lq}(-\xi_l, -\xi_q; \nu+p)\right] \\ &\quad + \xi_i\tau_j + \xi_j\tau_i - \xi_i\xi_j E(\eta|X=x),\end{aligned}$$

where

$$q(x-\mu) = (x-\mu)^\top(\Lambda\Lambda^\top + \Psi)^{-1}(x-\mu),$$

$$\tilde{c} = T_k\left\{\sqrt{\frac{\nu+p}{\nu+q(x-\mu)}}\,\xi; \Omega, \nu+p\right\},$$

$$\begin{aligned}\tilde{F}_l(h; \tilde{\nu}) &= \sqrt{\frac{\tilde{\nu}-1}{\nu+q(x-\mu)}}\,t_1\left\{\sqrt{\frac{\tilde{\nu}-1}{\nu+q(x-\mu)}}\,h; \omega_{ll}, \tilde{\nu}-1\right\} \\ &\quad \times T_{k-1}\left[\sqrt{\frac{\tilde{\nu}}{\nu+q(x-\mu)+h^2/\omega_{ll}}}\{\xi_{(l)|l}(h) + \xi_{(l)}\}; \Omega_{(l)|l}, \tilde{\nu}\right],\end{aligned}$$

$$\begin{aligned}\tilde{F}_{lq}(h_l, h_q; \tilde{\nu}) &= t_2\left\{\sqrt{\frac{\tilde{\nu}-2}{\nu+q(x-\mu)}}\,(h_l, h_q); \Omega_{lq,lq}, \tilde{\nu}-2\right\} \\ &\quad \times T_{k-2}\left[\sqrt{\frac{\tilde{\nu}}{\nu+q(x-\mu)+h^\top\Omega_{lq,lq}^{-1}h}}\,\{\xi_{(lq)|lq}(h_l, h_q) + \xi_{(lq)}\}; \Omega_{(lq)|lq}, \tilde{\nu}\right].\end{aligned}$$

Next, we demonstrate how to use the EM-type algorithm for ML estimation of the skew-t factor model. For n independent observations of x, we have, given $g_1, \ldots, g_n$ and μ, Λ, Ψ, that $x_1, \ldots, x_n$ are independent with a normal distribution given in Theorem 4. The log-likelihood function of $\Theta = (\mu, \Lambda, \Psi, \nu)$ based on complete data (x, g), aside from additive constant terms, can be written as

$$l_c(\Theta|x, g) \propto -\frac{n}{2}\ln|\Psi| - \frac{1}{2}\mathrm{tr}\left\{\Psi^{-1}\sum_{i=1}^{n}W^{-1}(\eta_i)M_i\right\} + \sum_{i=1}^{n}\ln h(\eta_i; \nu),$$

where $M_i = (x_i - \mu - \Lambda g_i)(x_i - \mu - \Lambda g_i)^\top$ as was defined before.

At the jth iteration of the E-step, we need to calculate the Q-function, defined by

$$Q(\Theta|\widehat{\Theta}^{(j)}) = \mathrm{E}\left\{l_c(\Theta|x,g)|\widehat{\Theta}^{(j)}, x\right\} = -\frac{n}{2}\ln|\Psi| - \frac{1}{2}\mathrm{tr}\left[\Psi^{-1}\sum_{i=1}^{n}\mathrm{E}\left\{W^{-1}(\eta_i)M_i|\widehat{\Theta}^{(j)}, x_i\right\}\right].$$

The following conditional expectations are necessary to obtain the Q-function:

$$\kappa^{(j)} = \mathrm{E}\left\{W^{-1}(\eta_i)|\widehat{\Theta}^{(j)}, x_i\right\}, \qquad \tau^{(j)} = \mathrm{E}\left\{W^{-1}(\eta_i)g_i|\widehat{\Theta}^{(j)}, x_i\right\},$$
$$\varsigma^{(j)} = \mathrm{E}\left\{W^{-1}(\eta_i)g_ig_i^\top|\widehat{\Theta}^{(j)}, x_i\right\}.$$

We do not consider the computation of $\mathrm{E}\{\ln h(\eta_i;\nu)|\widehat{\Theta}^{(j)}, x_i\}$ because the ECME algorithm [23] used here does not employ it. For the skew-t factor model, $\kappa^{(j)}$, $\tau^{(j)}$, and $\varsigma^{(j)}$ can be obtained in closed forms by Corollary 1.

The M-step consists in the maximization of $Q(\Theta|\widehat{\Theta}^{(j)})$ with respect to Θ. To do this, we use the faster extension of the original EM, the ECME algorithm [23], by replacing the M-step with a sequence of conditional maximization steps.
CM-steps:

1. Update $\mu^{(j)}$ by $\mu^{(j+1)} = \left(\sum_{i=1}^{n}\kappa^{(j)}\right)^{-1}\left(\sum_{i=1}^{n}\kappa^{(j)}x_i - \Lambda^{(j)}\sum_{i=1}^{n}\tau^{(j)}\right)$;
2. Update $\Lambda^{(j)}$ by

$$\Lambda^{(j+1)} = \left\{\sum_{i=1}^{n}\left(x_i - \mu^{(j+1)}\right)\left(\tau^{(j)}\right)^\top\right\}\left(\sum_{i=1}^{n}\varsigma^{(j)}\right)^{-1};$$

3. Find $\Psi^{(j+1)}$ to maximize the actual constrained log-likelihood given $\mu^{(j+1)}$ and $\Lambda^{(j+1)}$; that is,

$$\Psi^{(j+1)} = \arg\max_{\Psi}\sum_{i=1}^{n}\ln\{f(x_i)\},$$

where the marginal density of x is given in Theorem 3. This distribution is a scale mixture of SUN distributions;
4. Find $\nu^{(j+1)}$, which maximizes the actual constrained log-likelihood given $\mu^{(j+1)}$, $\Lambda^{(j+1)}$ and $\Psi^{(j+1)}$; that is,

$$\nu^{(j+1)} = \arg\max_{\nu}\sum_{i=1}^{n}\ln\{f(x_i)\}.$$

Step 3 in the above CM-steps can be replaced by:

3a. Update $\Psi^{(j)}$ by

$$\Psi^{(j+1)} = \mathrm{diag}\left[\frac{1}{n}\left\{\sum_{i=1}^{n}\kappa^{(j)}\left(x_i - \mu^{(j+1)}\right)\left(x_i - \mu^{(j+1)}\right)^\top - \Lambda^{(j+1)}\sum_{i=1}^{n}\varsigma^{(j)}\left(\Lambda^{(j+1)}\right)^\top\right\}\right].$$

The marginal distribution of x is given at Theorem 3 and the pdf of step 4 in the CM-steps with step 3a is given by

$$f(x_i) = 2^k t_p\left\{x_i - \mu^{(j+1)}; \Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi^{(j+1)}, \nu\right\}T_k\left\{\xi_i^{(j+1)}; \frac{\nu + q\left(x_i - \mu^{(j+1)}\right)}{\nu + p}\Omega^{(j+1)}, \nu + p\right\}, \quad x_i \in \mathbb{R}^p,$$

where

$$\xi_i^{(j+1)} = (\Lambda^{(j+1)})^\top\{\Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi^{(j+1)}\}^{-1}(x_i - \mu^{(j+1)}),$$
$$\Omega^{(j+1)} = I_k - (\Lambda^{(j+1)})^\top\{\Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi^{(j+1)}\}^{-1}\Lambda^{(j+1)},$$
$$q(x_i - \mu^{(j+1)}) = (x_i - \mu^{(j+1)})^\top\{\Lambda^{(j+1)}(\Lambda^{(j+1)})^\top + \Psi^{(j+1)}\}^{-1}(x_i - \mu^{(j+1)}).$$

For the choice of the starting values of the parameters, we perform an ordinary factor analysis to obtain the starting values for the factor loading matrix, Λ, and the covariance matrix, Ψ. We use the sample mean, $\bar{x}$, as the starting value for μ.

2.4. The generalized skew-normal factor model

Theoretically, the skew-normal factor model is a particular case of the model by Montanari and Viroli [25]. Here we propose an extension, the generalized skew-normal factor model, that contains both the skew-normal factor model of Section 2.2 and the model by Montanari and Viroli [25] as special cases.

Under model (1), suppose that $f_1 > 0$, where $f = (f_1^\top f_2^\top)^\top$ and f_1 is k_1-dimensional and f_2 is k_2-dimensional random vectors with $k = k_1 + k_2$. Then,

$$x \overset{d}{=} [y|f_1 > 0] \overset{d}{=} \mu + \Lambda(f|f_1 > 0) + \epsilon \overset{d}{=} \mu + \Lambda_1|f_1| + \Lambda_2 f_2 + \epsilon, \tag{6}$$

where $\Lambda = (\Lambda_1\ \Lambda_2)$. Note that model (6) can be formulated equivalently as:

$$\text{(i) } x = \mu + \Lambda_1|f_1| + \epsilon_*, \quad \text{with } \epsilon_* \sim \mathcal{N}_p(0, \Lambda_2\Lambda_2^\top + \Psi); \quad \text{or}$$
$$\text{(ii) } x = \mu + \Lambda_2 f_2 + \epsilon_{**}, \quad \text{with } \epsilon_{**} \sim \mathcal{SUN}_{p,k_1}(0, 0, 1_k, \Omega_{**}) \quad \text{and}$$
$$\Omega_{**} = \begin{pmatrix} I_{k_1} & \Lambda_1^\top \\ \Lambda_1 & \Lambda_1\Lambda_1^\top + \Psi \end{pmatrix}.$$

Furthermore, we can easily show that the generalized skew-normal factor model contains the model of Montanari and Viroli [25] as a special case. Indeed, suppose that $f \sim \mathcal{SN}_k(0, \bar{\Omega}, \alpha)$, where $\bar{\Omega}$ is a positive-definite $d \times d$ correlation matrix. Then $f \overset{d}{=} D_\delta U_0 + \delta|U_1|$, where $D_\delta = \{I_k - \text{diag}(\delta)^2\}^{1/2}$, $\delta = (\delta_1, \ldots, \delta_k)^\top$ with all elements in $(-1, 1)$, $U_0 \sim \mathcal{N}_k(0, \bar{\Psi})$ and $U_1 \sim \mathcal{N}_1(0, 1)$ are independent, and $\bar{\Psi}$ is a full-rank correlation matrix [9].

Let $y = \mu + \Lambda f + \epsilon$, where $\epsilon \sim \mathcal{N}_p(0, \Psi)$ and is independent of $f \sim \mathcal{SN}_k(0, \bar{\Omega}, \alpha)$. We then have

$$\begin{aligned} y &= \mu + \Lambda f + \epsilon \\ &\overset{d}{=} \mu + \Lambda(D_\delta U_0 + \delta|U_1|) + \epsilon \\ &= \mu + \Lambda\delta|U_1| + \Lambda\{I_k - \text{diag}(\delta)^2\}^{1/2}U_0 + \epsilon \\ &= \mu + \Lambda\delta|U_1| + \Lambda\{I_k - \text{diag}(\delta)^2\}^{1/2}\bar{\Psi}^{1/2}\bar{\Psi}^{-1/2}U_0 + \epsilon \\ &= \mu + \Lambda_1|f_1| + \Lambda_2 f_2 + \epsilon = x, \end{aligned}$$

where $\Lambda_1 = \Lambda\delta, f_1 = U_1, \Lambda_2 = \Lambda\{I_k - \text{diag}(\delta)^2\}^{1/2}\bar{\Psi}^{1/2}$ and $f_2 = \bar{\Psi}^{-1/2}U_0$. The final line is the form of (6), which means that the generalized skew-normal factor model contains the model of Montanari and Viroli [25] as a special case. Next we can find the marginal distribution of x in the following theorem.

Theorem 6. *The random vector, x, defined in (6) follows a unified skew-normal distribution, $\mathcal{SUN}_{p,k_1}(\mu, 0, 1_p, \Omega^\dagger)$, where*

$$\Omega^\dagger = \begin{pmatrix} I_{k_1} & \Lambda_1^\top \\ \Lambda_1 & \Lambda\Lambda^\top + \Psi \end{pmatrix}.$$

Hence, the pdf of x is given by

$$2^{k_1}\phi_p(x - \mu; 0, \Lambda\Lambda^\top + \Psi)\Phi_{k_1}(\xi^+; \Omega^+), \quad x \in \mathbb{R}^p, \tag{7}$$

where $\xi^+ = \Lambda_1^\top(\Lambda\Lambda^\top + \Psi)^{-1}(x - \mu)$ and $\Omega^+ = I_{k_1} - \Lambda_1^\top(\Lambda\Lambda^\top + \Psi)^{-1}\Lambda_1$. Therefore, similar to Theorem 1, we have the following theorem based on some well-known properties of the multivariate normal distribution.

Theorem 7. *Under model (1) with $f_1 > 0$, we have that*

(i) $x|\,|f_1| \sim \mathcal{N}_p(\mu + \Lambda_1|f_1|, \Lambda_2\Lambda_2^\top + \Psi)$,
(ii) $|f_1| \sim \mathcal{TN}_{k_1}(0, I_{k_1}, 0)$,
(iii) $|f_1|\,|x \sim \mathcal{TN}_{k_1}(\xi^+, \Omega^+, 0)$.

Compared to the mean and covariance structures of the traditional factor model, here we give those of the generalized skew-normal factor model. Using the laws of total expectation and total variance, we have that

$$\begin{aligned} \text{E}(x) &= \mu + \sqrt{\frac{2}{\pi}}\Lambda_1 1_{k_1}, \\ \text{var}(x) &= \Lambda\Lambda^\top - \frac{2}{\pi}\Lambda_1\Lambda_1^\top + \Psi, \\ \text{cov}(x, |f_1|) &= \left(1 - \frac{2}{\pi}\right)\Lambda_1, \quad \text{and} \quad \text{cov}(x, f_2) = \Lambda_2. \end{aligned}$$

Note that $\text{E}(x)$ is not affected by f_2; i.e., it does not depend on Λ_2. This observation coincides with the results of Arellano-Valle and Azzalini [2] and can be written in componentwise expressions as follows:

$$\text{E}(x_i) = \mu_i + \sqrt{\frac{2}{\pi}}\sum_{j=1}^{k_1}\lambda_{1_{ij}}, \qquad \text{var}(x_i) = \sum_{j=1}^{k}\lambda_{ij}^2 - \frac{2}{\pi}\sum_{j=1}^{k_1}\lambda_{1_{ij}}^2 + \psi_i,$$

$$\text{cov}(x_i, x_j) = \sum_{l=1}^{k}\lambda_{il}\lambda_{jl} - \frac{2}{\pi}\sum_{l=1}^{k_1}\lambda_{1_{il}}\lambda_{1_{jl}},$$

$$\text{cov}(x_i, |f_{1_j}|) = \left(1 - \frac{2}{\pi}\right)\lambda_{1_{ij}}, \quad \text{and} \quad \text{cov}(x_i, f_{2_j}) = \lambda_{2_{ij}}.$$

To estimate the parameters, we need to find the conditional moments of the truncated k_1-dimensional normal distribution that are given in Lemma 1.

We demonstrate how to use the EM-type algorithm for ML estimation of the generalized skew-normal factor model. For n independent observations of x, we have, given $(g_{11}, f_{21}), (g_{12}, f_{22}), \ldots, (g_{1n}, f_{2n})$ and μ, Λ, Ψ, that $x_1, \ldots, x_n$ are independent with a normal distribution given in Theorem 7. The complete data are (x, g_1, f_2), and (g_1, f_2) are fully missing. The related parameters are $\Theta = (\mu, \Lambda, \Psi)$. We remark that $g_1 \overset{d}{=} |f_1|$ follows the $\mathcal{TN}_{k_1}(0, I_{k_1}, 0)$ distribution given in Theorem 7. Hence, the log-likelihood function of Θ based on complete data, aside from additive constant terms, can be written as

$$l_c(\Theta|x, g_1) \propto -\frac{n}{2}\ln|\Lambda_2\Lambda_2^\top + \Psi| - \frac{1}{2}\mathrm{tr}\left\{\left(\Lambda_2\Lambda_2^\top + \Psi\right)^{-1}\sum_{i=1}^{n} M_{1_i}\right\},$$

where $M_{1_i} = (x_i - \mu - \Lambda_1 g_{1_i})(x_i - \mu - \Lambda_1 g_{1_i})^\top$.

At the jth iteration of the E-step, we need to calculate the Q-function, defined by

$$\begin{aligned} Q(\Theta|\widehat{\Theta}^{(j)}) &= \mathrm{E}\left\{l_c(\Theta|x, g_1)|\widehat{\Theta}^{(j)}, x\right\} \\ &= -\frac{n}{2}\ln|\Lambda_2\Lambda_2^\top + \Psi| - \frac{1}{2}\mathrm{tr}\left\{\left(\Lambda_2\Lambda_2^\top + \Psi\right)^{-1}\sum_{i=1}^{n}\mathrm{E}\left(M_{1_i}|\widehat{\Theta}^{(j)}, x\right)\right\}, \end{aligned}$$

which is the conditional distribution given by Theorem 7 and the mean and covariance are given in Lemma 1. The M-step consists in the maximization of $Q(\Theta|\widehat{\Theta}^{(j)})$ with respect to Θ. To do this, we use the ECME algorithm, by replacing the M-step with a sequence of conditional maximization steps.

CM-steps:

1. Update $\mu^{(j)}$ by $\mu^{(j+1)} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\Lambda_1^{(j)}\sum_{i=1}^{n}\mathrm{E}(g_{1_i}|\widehat{\Theta}^{(j)}, x)$;
2. Update $\Lambda_1^{(j)}$ by

$$\Lambda_1^{(j+1)} = \left\{\sum_{i=1}^{n}\left(x_i - \mu^{(j+1)}\right)\ \mathrm{E}\left(g_{1_i}^\top|\widehat{\Theta}^{(j)}, x\right)\right\}\left\{\sum_{i=1}^{n}\mathrm{E}\left(g_{1_i}g_{1_i}^\top|\widehat{\Theta}^{(j)}, x\right)\right\}^{-1};$$

3. Update $\Psi^{(j)}$ by

$$\Psi^{(j+1)} = \mathrm{diag}\left[-\Lambda_2^{(j)}(\Lambda_2^{(j)})^\top + \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu^{(j+1)}\right)\left(x_i - \mu^{(j+1)}\right)^\top - \frac{1}{n}\Lambda_1^{(j+1)}\left\{\sum_{i=1}^{n}\mathrm{E}\left(g_{1_i}g_{1_i}^\top|\widehat{\Theta}^{(j)}, x\right)\right\}(\Lambda_1^{(j+1)})^\top\right];$$

4. Find $\Lambda_2^{(j+1)}$ to maximize the actual constrained log-likelihood given $\mu^{(j+1)}$, $\Lambda_1^{(j+1)}$, and $\Psi^{(j+1)}$; that is,

$$\Lambda_2^{(j+1)} = \arg\max_{\Lambda_2}\sum_{i=1}^{n}\ln\{f(x_i)\}.$$

We remark that the density of the marginal distribution of x is given in Theorem 6, and the pdf of step 4 of the CM-steps is given by

$$f(x_i) = 2^{k_1}\phi_p\left\{x_i - \mu^{(j+1)}; 0, \left(\Lambda_1^{(j+1)}, \Lambda_2\right)\left(\Lambda_1^{(j+1)}, \Lambda_2\right)^\top + \Psi^{(j+1)}\right\}\Phi_{k_1}\left(\xi_i^{+(j+1)}; \Omega^{+(j+1)}\right), \quad x_i \in \mathbb{R}^p,$$

where

$$\xi_i^{+(j+1)} = (\Lambda_1^{(j+1)})^\top\left\{\left(\Lambda_1^{(j+1)}, \Lambda_2\right)\left(\Lambda_1^{(j+1)}, \Lambda_2\right)^\top + \Psi^{(j+1)}\right\}^{-1}\left(x_i - \mu^{(j+1)}\right)$$

and

$$\Omega^{+(j+1)} = I_{k_1} - (\Lambda_1^{(j+1)})^\top\left\{\left(\Lambda_1^{(j+1)}, \Lambda_2\right)\left(\Lambda_1^{(j+1)}, \Lambda_2\right)^\top + \Psi^{(j+1)}\right\}^{-1}\Lambda_1^{(j+1)}.$$

To choose the starting values of the parameters, we perform an ordinary factor analysis as previously mentioned.

The generalized skew-normal factor model can be extended to a generalized skew-t factor model using a similar approach as described in this subsection.

3. Statistical aspects

For both the skewed factor models and the normal factor model, the factor loadings, Λ, are determined only up to an orthogonal random sign matrix, P, if we relax the possible change of signs in the factor loadings. This is a motivation to adopt the skewed models defined in (3), (4) and (6). By doing so, we can handle skewed and/or heavy tailed data. Furthermore, the number of parameters is the same as the normal factor model and the t factor model since the shape parameters are determined by Theorem 1 and (7). In terms of the number of parameters, our skew-normal factor model is more parsimonious than that of Montanari and Viroli [25] since they have one additional shape parameter. For easier interpretation of the factors, factor rotation is common in classical factor analysis. For the same reason, we may apply orthogonal or oblique factor rotations.

In factor analysis, the interest is usually centered on the parameters in the factor model. However, the estimated values of the common factors, called factor scores, may also be required. These quantities are used for diagnostic purposes, as well as inputs to the subsequent analysis. Using the weighted least squares method [20], we have

$$fs = (\Lambda^\top \Psi^{-1} \Lambda)^{-1} \Lambda^\top \Psi^{-1} (x - \mu). \tag{8}$$

Replacing the parameters with their maximum likelihood estimates obtained by the ECME algorithms, the factor scores become

$$\widehat{fs}_i = (\widehat{\Lambda}^\top \widehat{\Psi}^{-1} \widehat{\Lambda})^{-1} \widehat{\Lambda}^\top \widehat{\Psi}^{-1} (x_i - \widehat{\mu})$$

for $i = 1, \ldots, n$. Note that the result of the weighted least squares method approach is the same for the skew-normal, skew-t, and generalized skew-normal models, since no distributional assumption is used. However, the estimated values of the parameters are different for each model.

If any factor rotation is applied, then we have new factor scores such that

$$\widehat{fs}_i^* = T \widehat{fs}_i,$$

where T is orthogonal or non-singular matrices corresponding to orthogonal or oblique factor rotations, respectively, for $i = 1, \ldots, n$. Obviously, the factor loading matrices will be changed to $\Lambda T^\top$ or ΛT^{-1} for orthogonal or oblique rotations, respectively.

Similar to Montanari and Viroli [25], the suggested skew-normal, skew-t, and generalized skew-normal factor models are not a full exponential family and the validity of the regularity conditions needed for the general theory of likelihood-based statistical inference is not provided. We can therefore not use the usual asymptotic distribution theory to test the goodness-of-fit and to test the nested hypotheses on the number of factors. Montanari and Viroli [25] suggested using the average of the final communalities. This measure could be applied using the calculated final communalities, which are given in the formulas of $\text{var}(x_i)$ excluding ψ_i terms.

Instead of the average of the final communalities, the minimum Akaike Information Criterion estimate (MAICE) criterion [28] could be adopted. That is, we compute the value of the AIC for models with different number of factors, and then we select the model that yields the minimum AIC estimate. The definition of AIC is:

$$\text{AIC} = -2 \ln(\text{maximized likelihood}) + 2 \times (\text{number of independent parameters}).$$

Similarly one may use the Bayesian Information Criterion (BIC) defined as in [30]:

$$\text{BIC} = -2 \ln(\text{maximized likelihood}) + \ln n \times (\text{number of independent parameters}).$$

In either case, we select a model that reflects a proper compromise between the percent of variation in each original variable accounted for by the factors and the parsimony and interpretability of the model.

4. Monte Carlo simulations

We consider a simulation example to compare the performance of the proposed skew-normal, skew-t and generalized skew-normal factor models with the common Gaussian case. In the simulation study, we consider the following models and their associated first and second moments:

I. Normal factor (nf) model: $y = \mu + \Lambda f + \epsilon$,

$$\text{E}_1(y) = \mu, \qquad \text{var}_1(y) = \Lambda\Lambda^\top + \Psi, \qquad \text{cov}_1(y, f) = \Lambda.$$

II. Skew-normal factor (snf) model: $y = \mu + \Lambda |f| + \epsilon$,

$$\text{E}_2(y) = \mu + \sqrt{\frac{2}{\pi}} \Lambda 1_k, \qquad \text{var}_2(y) = \left(1 - \frac{2}{\pi}\right) \Lambda\Lambda^\top + \Psi, \qquad \text{cov}_2(y, |f|) = \left(1 - \frac{2}{\pi}\right) \Lambda.$$

III. Skew-t factor (stf) model: $y = \mu + \Lambda|f| + \epsilon$, where ϵ and f follow multivariate-t distributions with $df = \nu$. From Theorem 4,

$$\mathrm{E}_3(y) = \mu + 2c_\nu \Lambda 1_k, \qquad \mathrm{var}_3(y) = \Lambda(aI_k + bJ_k)\Lambda^\top + \frac{\nu}{\nu-2}\Psi, \qquad \mathrm{cov}_3(y, |f|) = \Lambda(aI_k + bJ_k).$$

IV. Generalized skew-normal factor (gsnf) model: $y = \mu + \Lambda_1|f_1| + \Lambda_2 f_2 + \epsilon$,

$$\mathrm{E}_4(y) = \mu + \sqrt{\frac{2}{\pi}}\Lambda_1 1_{k_1}, \qquad \mathrm{var}_4(y) = \Lambda\Lambda^\top - \frac{2}{\pi}\Lambda_1\Lambda_1^\top + \Psi, \qquad \mathrm{cov}_4\{y, (|f_1|, f_2)^\top\} = \left(\left(1 - \frac{2}{\pi}\right)\Lambda_1, \Lambda_2\right).$$

V. t factor (tf) model: $y = \mu + \Lambda f + \epsilon$, where ϵ and f follow multivariate-t distributions with $df = \nu$,

$$\mathrm{E}_5(y) = \mu, \qquad \mathrm{var}_5(y) = \frac{\nu}{\nu-2}(\Lambda\Lambda^\top + \Psi), \qquad \mathrm{cov}_5(y, f) = \frac{\nu}{\nu-2}\Lambda.$$

VI. Gamma factor (gf) model: $y = \mu + \Lambda f + \epsilon$, where the f_i's follow a Gamma distribution with shape parameter α and scale parameter β,

$$\mathrm{E}_6(y) = \mu + \alpha\beta\Lambda 1_k, \qquad \mathrm{var}_6(y) = \alpha\beta^2\Lambda\Lambda^\top + \Psi, \qquad \mathrm{cov}_6(y, f) = \alpha\beta^2\Lambda.$$

Note that in models I, II, IV and VI, $f \sim \mathcal{N}_k(0, I_k)$, and $\epsilon \sim \mathcal{N}_p(0, \Psi)$. Furthermore E_j, var_j and cov_j are the mean, variance and covariance structures associated with the jth model. For simulation purposes, we choose $\nu = 4, \alpha = 1, \beta = 0.75$,

$$\mu = (12, 24, 36, 48, 60)^\top, \qquad \Lambda = \begin{pmatrix} 0 & 0 & 0 & 4 & 5 \\ 2 & 4 & 6 & 0 & 0 \end{pmatrix}^\top \quad \text{and} \quad \Psi = \mathrm{diag}(1, 2, 3, 4, 5).$$

Now, from each of the above models (I, II, III, IV, V and VI), we simulate $M = 100$ datasets with sample size $n = 200$. We use the factanal function in the R package stats [29] to obtain the MLEs with no rotation for the normal factor model (nf fit). The results are used as initial values to obtain the MLEs based on the proposed models (snf, stf and gsnf fits) separately for each dataset. For a given dataset, among the first four generating models (I, II, III and IV), we expect to observe a better fit for the model that data come from. Furthermore, the datasets obtained from the last two generating models (V and VI) can be used to monitor the robustness of the four model fits. In order to check these conjectures, we compare estimators of the factor models using the following five measurements:

(a) Sum of absolute deviation of mean (SADM):

$$\mathrm{SADM}_j = \frac{1}{M}\sum_{m=1}^{M} \|\widehat{\mathrm{E}_j(y_m)} - \mathrm{E}_T(y)\|_1,$$

where $\mathrm{E}_T(y)$ denotes the true mean vector. SADM_j measures how well we estimate the mean vector using model j.

(b) Sum of absolute deviation of variance (SADV):

$$\mathrm{SADV}_j = \frac{1}{M}\sum_{m=1}^{M} \|\widehat{\mathrm{var}_j(y_m)} - \mathrm{var}_T(y)\|_1,$$

where $\mathrm{var}_T(y)$ denotes the true covariance matrix for the observed values. SADV_j measures how well we estimate the variance structure using model j.

(c) Sum of absolute deviation of covariance (SADC):

$$\mathrm{SADC}_j = \frac{1}{M}\sum_{m=1}^{M} \|\widehat{\mathrm{cov}_j(y_m, f)} - \mathrm{cov}_T(y, f)\|_1,$$

where $\mathrm{cov}_T(y, f)$ denotes the true covariance between the observed and latent variables. SADC_j measures how well we estimate the covariance structure using model j.

(d) Sum of absolute deviation of factor scores (SADF):

$$\mathrm{SADF}_j = \frac{1}{nM}\sum_{m=1}^{M} \|\widehat{Fs_j(y_m)} - Fs(y_m)\|_1,$$

where $Fs(y_m)$ is an $n \times k$ matrix of true factor scores obtained from (8) for the m-th dataset and $\widehat{Fs_j(y_m)}$ is the associated estimate using model j.

(e) We also use the AIC and BIC as the final measurement for comparison.

Table 1 shows that, overall, the normal factor (nf) fit does a good job in estimating the mean (SADM) and variance structure (SADV) for all of the data generative cases (nf, snf, stf, gsnf, tf and gf), while the measurements for evaluating the covariance structure (SADC), factor scores (SADF) and AIC/BIC indicate clear advantages for the fits associated with the same data generative models. This can be further illustrated using Fig. 1, which confirms that the AIC associated with

Table 1
Comparing the fitted models (nf, snf, stf and gsnf) based on SADM, SADV, SADC, SADF and AIC/BIC for $M = 100$ simulated datasets, with sample size $n = 200$, generated from the associated models (I—nf, II—snf, III—stf, IV—gsnf, V—tf and VI—gf). Values in parentheses are empirical standard errors.

Generating model	Fitted model	SADM	SADV	SADC	SADF	AIC/BIC
I—nf	nf	0.318 (0.014)	1.697 (0.078)	0.490 (0.014)	7.592 (0.254)	2014.7/2080.7 (5.4)
	snf	0.537 (0.026)	6.128 (0.235)	3.362 (0.116)	25.50 (0.490)	2110.8/2176.7 (6.3)
	stf	0.553 (0.025)	6.808 (0.258)	3.512 (0.140)	26.15 (0.562)	2111.3/2180.6 (6.1)
	gsnf	0.437 (0.022)	4.299 (0.247)	4.278 (0.245)	29.84 (1.514)	2056.4/2122.4 (5.7)
II—snf	nf	0.285 (0.012)	1.631 (0.055)	1.858 (0.021)	15.08 (0.297)	2435.0/2501.0 (4.7)
	snf	0.279 (0.012)	1.600 (0.052)	0.338 (0.011)	3.131 (0.134)	2386.0/2452.0 (4.5)
	stf	0.290 (0.013)	1.735 (0.059)	0.431 (0.014)	3.181 (0.141)	2388.3/2457.6 (4.5)
	gsnf	0.280 (0.012)	1.648 (0.056)	1.026 (0.020)	9.294 (0.344)	2402.5/2468.4 (4.7)
III—stf	nf	0.911 (0.021)	6.792 (0.213)	3.881 (0.062)	23.04 (1.820)	1990.8/2056.8 (12.3)
	snf	1.190 (0.031)	8.364 (0.159)	5.850 (0.025)	22.31 (1.371)	1879.1/1945.1 (11.3)
	stf	0.312 (0.014)	3.375 (0.314)	1.777 (0.126)	14.22 (1.865)	1735.0/1804.3 (7.4)
	gsnf	1.140 (0.031)	7.774 (0.174)	5.350 (0.038)	26.14 (1.850)	1917.6/1983.6 (11.3)
IV—gsnf	nf	0.294 (0.013)	1.702 (0.062)	1.333 (0.018)	15.30 (0.856)	2308.4/2374.3 (4.4)
	snf	0.355 (0.016)	2.311 (0.093)	1.127 (0.061)	14.13 (0.633)	2317.5/2383.5 (4.7)
	stf	0.369 (0.016)	2.689 (0.107)	1.244 (0.086)	14.70 (0.734)	2319.4/2388.7 (4.8)
	gsnf	0.291 (0.012)	1.701 (0.062)	0.542 (0.016)	8.435 (0.726)	2277.2/2343.2 (4.4)
V—tf	nf	0.312 (0.014)	2.828 (0.175)	2.558 (0.099)	22.862 (1.173)	1990.8/2056.7 (12.1)
	snf	0.429 (0.024)	5.975 (0.416)	5.867 (0.145)	41.773 (0.953)	2110.0/2176.0 (12.4)
	stf	0.578 (0.027)	15.638 (0.985)	6.147 (0.261)	41.658 (1.116)	1999.2/2068.5 (8.6)
	gsnf	0.405 (0.021)	3.778 (0.248)	6.176 (0.256)	42.604 (1.781)	2022.7/2088.7 (11.7)
VI—gf	nf	0.286 (0.013)	2.354 (0.097)	1.362 (0.042)	11.582 (0.303)	2268.5/2334.5 (6.1)
	snf	0.388 (0.018)	2.306 (0.086)	0.966 (0.016)	5.224 (0.183)	2149.5/2215.4 (5.4)
	stf	0.302 (0.015)	2.711 (0.100)	0.882 (0.018)	4.086 (0.174)	2145.8/2215.0 (5.3)
	gsnf	0.360 (0.017)	2.397(0.091)	1.206 (0.025)	7.921 (0.263)	2196.1/2262.1 (5.6)

the same data generative model is the smallest for all 4×100 simulated datasets. In terms of robustness, when the latent variable, f, is generated from a symmetric model (V), the normal factor (nf) fit outperforms the fits that consider skewness, and vice versa when the common factor, f, is coming from skewed family of distributions, i.e. Gamma distribution (VI), the skewed factor models fit better.

Fig. 2 compares the performance of four different fits based on SADM, SADV, SADC and SADF for the case in which data are generated from the skew-normal factor model. The results for the rest of the generating models (nf, stf, gsnf, tf and gf) are omitted here for brevity and similarity.

5. Open/closed book exam scores data

To illustrate the performance of the proposed methods on real data, we use the classic open/closed book dataset by Mardia et al. [24]. A total of $n = 88$ students were tested for their ability in five content areas: mechanics (MC), vectors (VC), algebra (LO), analysis (NO), and statistics (SO). The first two tests were closed book (C), and the last three were open book (O). The test scores are represented as Y, an 88×5 data matrix. This dataset has been studied widely using two-factor models (i.e., [24,16,33]). Using the Shapiro–Wilk test for multivariate normality, the p-value is 0.003. The p-values for marginal normalities are 0.057, 0.928, 0.264, 0 and 0.016, respectively. Hence, there is indication of non-normality.

Next we present the MLEs based on the proposed two-factor models (skew-normal, skew-t and generalized skew-normal) and the regular Gaussian model (with no rotation). Given that $\{\mathsf{MC}, \mathsf{VC}, \mathsf{LO}, \mathsf{NO}, \mathsf{SO}\}^\top$ is the order of the content areas, the results are as follows.

Normal factor model: $y = \mu + \Lambda f + \epsilon$, assuming $f \sim \mathcal{N}_2(0, I_2)$ and $\epsilon \sim \mathcal{N}_5(0, \Psi)$. The MLEs are:

$$\widehat{\mu} = \begin{bmatrix} 38.95 \\ 50.59 \\ 50.60 \\ 46.68 \\ 42.31 \end{bmatrix}, \quad \widehat{\Lambda} = \begin{bmatrix} 10.99 & 6.52 \\ 9.14 & 4.10 \\ 9.56 & -0.53 \\ 11.57 & -2.98 \\ 12.55 & -3.45 \end{bmatrix}, \quad \widehat{\Psi} = \begin{bmatrix} 142.46 \\ 72.43 \\ 21.29 \\ 77.53 \\ 128.34 \end{bmatrix},$$

and AIC $= 3430.2$.

Skew-normal factor model: $y = \mu + \Lambda|f| + \epsilon$, assuming $|f| \sim \mathcal{TN}_2(0, I_2, 0)$ and $\epsilon \sim \mathcal{N}_5(0, \Psi)$. The MLEs are:

$$\widehat{\mu} = \begin{bmatrix} 43.77 \\ 54.32 \\ 52.58 \\ 46.71 \\ 26.25 \end{bmatrix}, \quad \widehat{\Lambda} = \begin{bmatrix} 9.61 & -15.65 \\ 8.21 & -12.90 \\ 10.63 & -13.15 \\ 13.82 & -13.93 \\ 26.57 & -6.66 \end{bmatrix}, \quad \widehat{\Psi} = \begin{bmatrix} 189.04 \\ 92.50 \\ 16.49 \\ 90.33 \\ 39.84 \end{bmatrix},$$

and AIC $= 3426.7$.

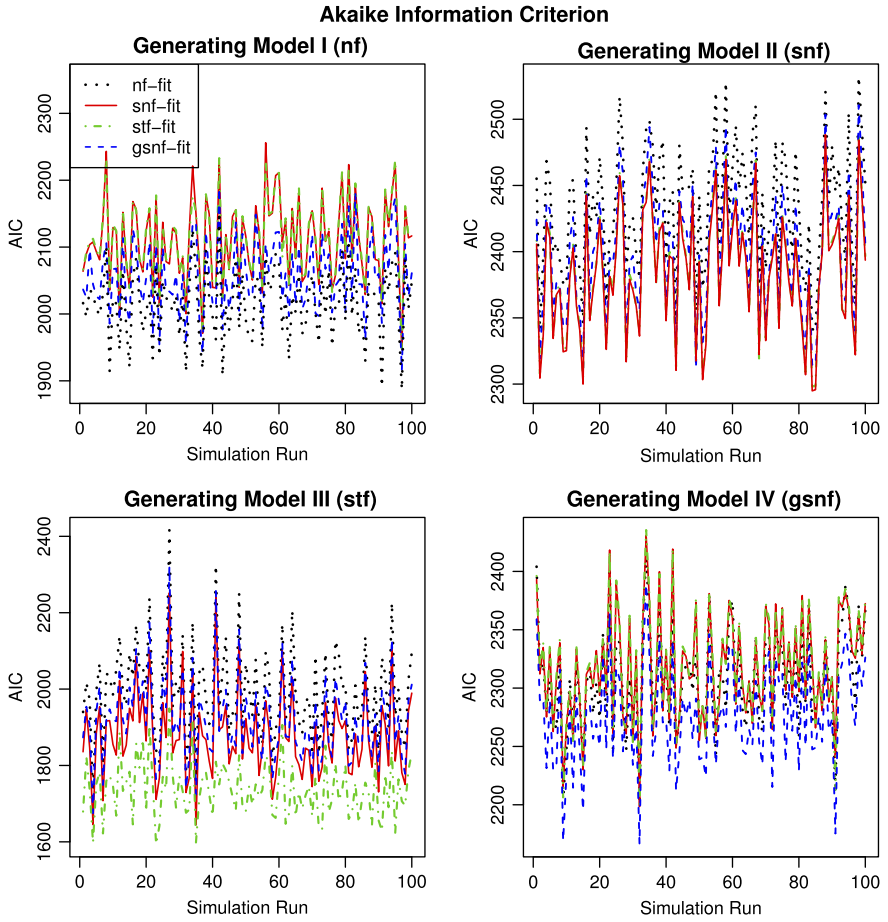

Fig. 1. Comparing the fitted models (nf, snf, stf and gsnf) based on AIC for 400 simulated datasets, with sample size $n = 200$, generated from models I—nf, II—snf, III—stf and IV—gsnf.

Skew-t factor model: $y = \mu + \Lambda|f| + \epsilon$, assuming $f \sim \mathrm{t}_2(0, I_2, \nu)$ and $\epsilon \sim \mathrm{t}_5(0, \Psi, \nu)$. The MLEs are:

$$\widehat{\mu} = \begin{bmatrix} 45.65 \\ 56.11 \\ 53.68 \\ 50.20 \\ 28.87 \end{bmatrix}, \quad \widehat{\Lambda} = \begin{bmatrix} 8.61 & -15.96 \\ 7.27 & -13.15 \\ 9.33 & -12.58 \\ 11.67 & -14.96 \\ 24.89 & -8.56 \end{bmatrix}, \quad \widehat{\Psi} = \begin{bmatrix} 160.84 \\ 79.50 \\ 18.89 \\ 75.14 \\ 31.76 \end{bmatrix},$$

and $\widehat{\nu} = 17$, AIC $= 3424.8$.

Generalized skew-normal factor model: $y = \mu + \Lambda_1|f_1| + \Lambda_2 f_2 + \epsilon$, assuming $|f_1| \sim \mathcal{TN}(0, 1, 0)$, $f_2 \sim \mathcal{N}(0, 1)$ and $\epsilon \sim \mathcal{N}_5(0, \Psi)$. The MLEs are:

$$\widehat{\mu} = \begin{bmatrix} 36.40 \\ 48.46 \\ 45.49 \\ 39.48 \\ 23.80 \end{bmatrix}, \quad \widehat{\Lambda} = \begin{bmatrix} 3.20 & 11.12 \\ 2.67 & 9.35 \\ 6.41 & 8.56 \\ 9.02 & 9.89 \\ 23.21 & 7.84 \end{bmatrix}, \quad \widehat{\Psi} = \begin{bmatrix} 174.79 \\ 80.80 \\ 23.41 \\ 90.47 \\ 36.30 \end{bmatrix},$$

and AIC $= 3428.3$.

Using the AIC, the skew-t model achieves the smallest value, while it estimates an extra parameter for degrees of freedom ($\widehat{\nu} = 17$). The other three models have an equal number of parameters, and based on the AIC, the skew-normal fit is ranked the second best, before the generalized skew-normal and normal factor models.

6. Discussion

Based on selection mechanisms, we developed three skewed factor models: the skew-normal, skew-t, and generalized skew-normal factor models. To estimate relevant parameters, we developed ECME algorithms. Their efficiency was

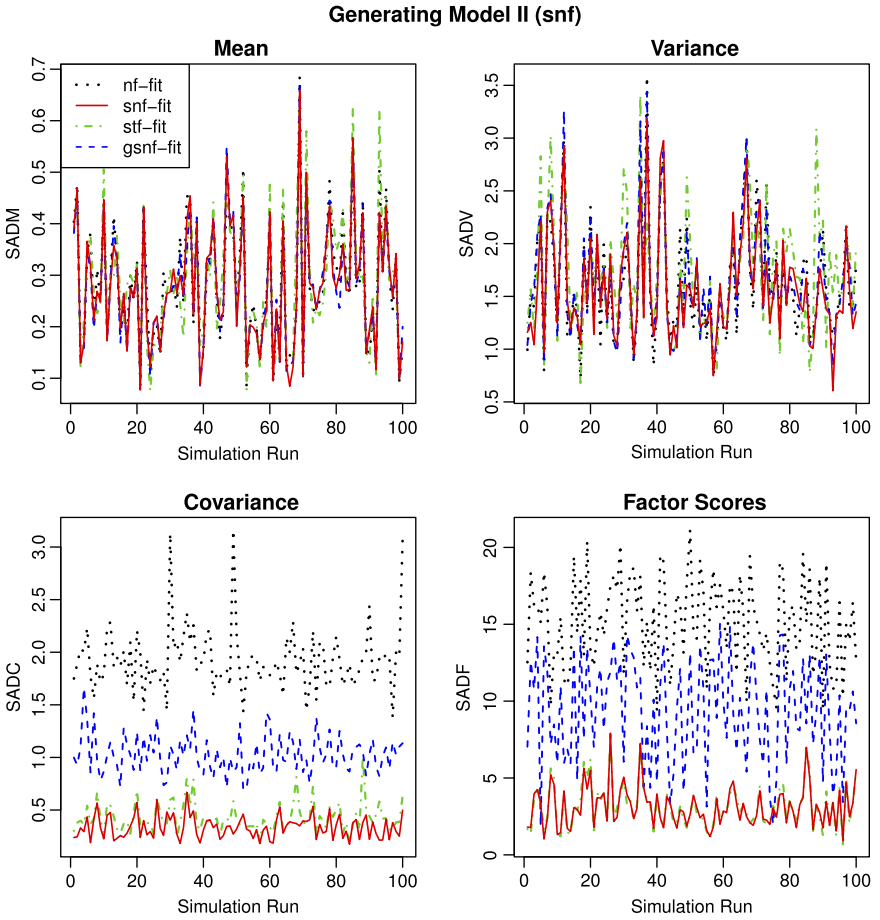

Fig. 2. Comparing the fitted models (nf, snf, stf and gsnf) based on SADM, SADV, SADC and SADF for $M = 100$ simulated datasets, with sample size $n = 200$, generated from model II (snf).

demonstrated via a simulation study. We developed one connection between the skewed factor models and normal factor models in terms of the factor loadings that differ only up to a possible change of signs of each row of the factor loadings. This is the reason why we developed the skewed factor models. In this way, we can handle skewed and/or heavy tailed data. Finally, a real data example illustrated the need for skewed factor models.

When analyzing a real dataset, we suggest to test for multivariate normality before applying any skewed factor models or a traditional normal factor model. If multivariate normality is rejected, then one can fit some skewed factor models and use criterions such as AIC or BIC to choose the best model.

Acknowledgments

The authors thank the Editor, Associate Editor and three referees for valuable suggestions that improved the paper. Kim's research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2015R1D1A1A01059161). Arellano-Valle's research was supported by Grant FONDECYT (Chile) 1120121. Genton's research was supported by King Abdullah University of Science and Technology (KAUST).

Appendix

Proof of Theorem 1. The proof of part (i) is direct from (3) and the proof of part (ii) follows from the definition of the k-dimensional half-normal distribution. Part (iii) is proved here. Since, by Bayes' formula, the conditional density of $|f|\,|x$ becomes

$$\frac{\phi_p(x;\mu+\Lambda|f|,\Psi)\phi_k(|f|;0,I_k)}{\int_{\mathbb{R}^k_+}\phi_p(x;\mu+\Lambda|f|,\Psi)\phi_k(|f|;0,I_k)d|f|},$$

where the numerator can be expressed as the joint density form of $(x, |f|)$ using standard properties of the multivariate normal distribution, we can find the conditional distribution of $|f|\,|x$ that yields the result. □

Proof of Theorem 2. First find the joint distribution of y and f using the joint distribution of f and ϵ:

$$\begin{pmatrix} f \\ y \end{pmatrix} \sim \mathcal{N}_{k+p}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} I_k & \Lambda^\top \\ \Lambda & \Lambda\Lambda^\top + \Psi \end{pmatrix}\right).$$

Then find the conditional distribution of $y|f > 0$ using the standard properties of the multivariate normal distribution. Since $x \overset{d}{=} [y|f > 0]$, the result follows. □

Proof of Theorem 3. Given η, first find the joint distribution of f and y using the joint distribution of f and ϵ:

$$\begin{pmatrix} f \\ y \end{pmatrix}\Bigg| \eta \sim \mathcal{N}_{k+p}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, W(\eta)\begin{pmatrix} I_k & \Lambda^\top \\ \Lambda & \Lambda\Lambda^\top + \Psi \end{pmatrix}\right).$$

Then find the conditional distribution of $y|f > 0, \eta$. Since $W(\eta) = 1/\eta$ and $\eta \sim \mathcal{G}(\nu/2, \nu/2)$, after cumbersome algebra, we have the result. □

Proof of Theorem 4. Given η, the proof is then similar to that of Theorem 1. □

Proof of Lemma 2. Let $Y|\eta \sim \mathcal{N}_k(\mu, \Sigma/\eta)$. Then, $Y \sim \mathrm{t}_k(\mu, (\beta/\alpha)\Sigma, 2\alpha)$ by the property of scale mixtures of normal distributions. For any $a \in \mathbb{R}^k$,

$$\mathrm{E}\left\{\Phi_k\left(a - \mu; \frac{1}{\eta}\Sigma\right)\right\} = \mathrm{E}\{\Pr(Y \le a|\eta)\} = \Pr(Y \le a).$$

Hence, the result follows. □

Proof of Theorem 5. The conditional density of $X|\eta$ is

$$f(x|\eta) = 2^k \phi_p\{x; \mu, W(\eta)(\Lambda\Lambda^\top + \Psi)\}\Phi_k\{\xi; W(\eta)\Omega\}, \quad x \in \mathbb{R}^p.$$

Let $h(\eta|x)$ and $h_0(\eta|x)$ be the conditional densities of $\eta|X = x$ and $\eta|X_0 = x$, respectively. Since $h(\eta|x) = 2^k f_0(x|\eta)\Phi_k\{\xi; W(\eta)\Omega\}h(\eta)/f(x)$, where $f_0(x|\eta) = \phi_p\{x; \mu, W(\eta)(\Lambda\Lambda^\top + \Psi)\}$ is the conditional density of $X_0|\eta$, we have

$$\begin{aligned}
\mathrm{E}\{S(\eta)|X = x\} &= \int_0^\infty S(\eta)h(\eta|x)d\eta \\
&= \int_0^\infty S(\eta)\frac{f(x|\eta)h(\eta)}{f(x)}d\eta \\
&= \int_0^\infty S(\eta)\frac{2^k f_0(x|\eta)\Phi_k\{\xi; W(\eta)\Omega\}h(\eta)}{f(x)}d\eta \\
&= \frac{2^k f_0(x)}{f(x)}\int_0^\infty S(\eta)\Phi_k\{\xi; W(\eta)\Omega\}\frac{f_0(x|\eta)h(\eta)}{f_0(x)}d\eta \\
&= \frac{2^k f_0(x)}{f(x)}\int_0^\infty S(\eta)\Phi_k\{\xi; W(\eta)\Omega\}h_0(\eta|x)d\eta.
\end{aligned}$$

The result thus follows. □

Proof of Corollary 1. When $W(\eta) = 1/\eta$ and $\eta \sim \mathcal{G}(\nu/2, \nu/2), X_0 \sim \mathrm{t}_p(x - \mu; \Lambda\Lambda^\top + \Psi, \nu)$, the marginal distribution of X is given at Theorem 3, and the conditional distribution of $\eta|X_0 = x$ is $\mathcal{G}((p + \nu)/2, \{\nu + q(x - \mu)\}/2)$, where $q(x - \mu) = (x - \mu)^\top(\Lambda\Lambda^\top + \Psi)^{-1}(x - \mu)$. Hence, taking $S(\eta) = \eta$ for (i), $S(\eta) = \Phi_k^{-1}(\xi; \eta^{-1}\Omega)F_l(-\xi_l)$ for (ii), $S(\eta) = \eta^{-1}\Phi_k^{-1}(\xi; \eta^{-1}\Omega)F_{lq}(-\xi_l, -\xi_q)$ for (iii) and using Lemma 2 and Theorem 5, we have the respective results. □

Proof of Theorem 6. First find the joint distribution of y and $f = (f_1^\top f_2^\top)^\top$ using the joint distribution of f and ϵ:

$$\begin{pmatrix} f_1 \\ f_2 \\ y \end{pmatrix} \sim \mathcal{N}_{k+p}\left(\begin{pmatrix} 0 \\ 0 \\ \mu \end{pmatrix}, \begin{pmatrix} I_{k_1} & 0 & \Lambda_1^\top \\ 0 & I_{k_2} & \Lambda_2^\top \\ \Lambda_1 & \Lambda_2 & \Lambda\Lambda^\top + \Psi \end{pmatrix}\right).$$

And then find the conditional distribution of $y|f_1 > 0$ similarly to the proof of Theorem 2. □

Proof of Theorem 7. The proofs of parts (i) and (ii) are similar to those of Theorem 1. Part (iii) is proved by changing f to f_1 and using a similar approach as that in Theorem 1. □

References

[1] D.F. Andrews, C.L. Mallows, Scale mixtures of normal distribution, J. R. Stat. Soc. Ser. B 36 (1974) 99–102.
[2] R.B. Arellano-Valle, A. Azzalini, On the unification of families of skew-normal distributions, Scand. J. Statist. 33 (2006) 561–574.
[3] R.B. Arellano-Valle, H. Bolfarine, On some characterizations of the t distribution, Statist. Probab. Lett. 25 (1995) 179–185.
[4] R.B. Arellano-Valle, M.D. Branco, M.G. Genton, A unified view on skewed distributions arising from selections, Canad. J. Statist. 34 (2006) 581–601.
[5] R.B. Arellano-Valle, M.G. Genton, Multivariate unified skew-elliptical distributions, Chil. J. Stat. 2 (2010) 17–34.
[6] A. Azzalini, A class of distributions which includes the normal ones, Scand. J. Statist. 12 (1985) 171–178.
[7] A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew normal distribution, J. R. Stat. Soc. Ser. B 61 (1999) 579–602.
[8] A. Azzalini, A. Capitanio, Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, J. R. Stat. Soc. Ser. B 65 (2003) 367–389.
[9] A. Azzalini, A. Capitanio, The Skew-Normal and Related Families, in: IMS Monograph, Cambridge University Press, 2014.
[10] A. Azzalini, A. Dalla Valle, The multivariate skew-normal distribution, Biometrika 83 (1996) 715–726.
[11] A. Azzalini, M.G. Genton, Robust likelihood methods based on the skew-t and related distributions, Internat. Statist. Rev. 76 (2008) 106–129.
[12] L. Bagnato, M. Minozzo, A latent variable approach to modelling multivariate geostatistical skew-normal data, in: Studies in Theoretical and Applied Statistics, Springer, 2014.
[13] M.D. Branco, D.K. Dey, A general class of multivariate skew elliptical distributions, J. Multivariate Anal. 79 (2001) 99–113.
[14] M.D. Branco, D.K. Dey, Regression model under skew elliptical error distribution, J. Math. Sci. 1 (2002) 151–168.
[15] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via EM algorithm (with discussion), J. R. Stat. Soc. Ser. B 39 (1977) 1–38.
[16] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, in: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1994.
[17] M.G. Genton, Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality, Chapman & Hall/CRC, Boca Raton, FL, 2004, Edited Volume.
[18] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., Johns Hopkins, Baltimore, MD, 1996.
[19] A.K. Gupta, Multivariate skew t-distribution, Statistics 37 (2003) 359–363.
[20] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, sixth ed., Pearson Prentice Hall, 2007.
[21] S. Kotz, S. Nadarajah, Multivariate t Distributions and their Applications, Cambridge University Press, 2004.
[22] P. Leppard, G.M. Tallis, Algorithm AS 249: Evaluation of the mean and covariance of the truncated multinormal distribution, Appl. Stat. 38 (1989) 543–553.
[23] C. Liu, D.B. Rubin, The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence, Biometrika 81 (1994) 633–648.
[24] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, INC., 1979.
[25] A. Montanari, C. Viroli, A skew-normal factor model for the analysis of student satisfaction towards university courses, J. Appl. Stat. 37 (2010) 473–487.
[26] A. Mooijaart, Factor analysis for non-normal variables, Psychometrika 50 (1985) 323–342.
[27] G. Pison, P.J. Rousseeuw, P. Filzmoser, C. Croux, Robust factor analysis, J. Multivariate Anal. 84 (2003) 145–172.
[28] S.J. Press, Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, second ed., Dover Publications, Inc., Florida, 2005.
[29] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2014.
[30] G.E. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[31] C.E. Spearman, General intelligence objectively determined and measured, American Journal of Psychology 15 (1904) 201–292.
[32] G.M. Tallis, The moment generating function of the truncated multi-normal distribution, J. R. Stat. Soc. Ser. B 23 (1961) 223–229.
[33] M. Watanabe, K. Yamaguchi, The EM Algorithm and Related Statistical Models, in: Statistics: A Series of Textbooks and Monographs, CRC Press, 2003.
[34] Y.F. Yung, Finite mixtures in confirmatory factor-analysis model, Psychometrika 62 (1997) 297–330.