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.