Journal of Multivariate Analysis 143 (2016) 398–413

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Extending mixtures of factor models using the restricted multivariate skew-normal distribution

Tsung-I Lin [a,b,*], Geoffrey J. McLachlan [c], Sharon X. Lee [c]

[a] Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan
[b] Department of Public Health, China Medical University, Taichung 404, Taiwan
[c] Department of Mathematics, University of Queensland, St Lucia, 4072, Australia

ARTICLE INFO

Article history:
Received 27 January 2014
Available online 19 October 2015

AMS subject classifications:
62H25
62H30
65C60

Keywords:
Clustering
Data reduction
ECM algorithm
Factor analyzer
rMSN distribution
Skewness

ABSTRACT

The mixture of factor analyzers (MFA) model provides a powerful tool for analyzing high-dimensional data as it can reduce the number of free parameters through its factor-analytic representation of the component covariance matrices. This paper extends the MFA model to incorporate a restricted version of the multivariate skew-normal distribution for the latent component factors, called mixtures of skew-normal factor analyzers (MSNFA). The proposed MSNFA model allows us to relax the need of the normality assumption for the latent factors in order to accommodate skewness in the observed data. The MSNFA model thus provides an approach to model-based density estimation and clustering of high-dimensional data exhibiting asymmetric characteristics. A computationally feasible Expectation Conditional Maximization (ECM) algorithm is developed for computing the maximum likelihood estimates of model parameters. The potential of the proposed methodology is exemplified using both real and simulated data.

1. Introduction

Factor analysis (FA) is a popular technique for explaining the covariance relationships among many variables through a fewer number of unobservable random quantities known as *latent factors*. Finite mixture models (FMMs) have been widely used as flexible means to model heterogeneous data, in particular, for density estimation and clustering. There are a number of monographs on mixture models; see, for example, [14,19,26,38,46,50,57,68] and the references contained therein. Mixtures of factor analyzers (MFAs), introduced by Ghahramani and Hinton [28], provide a global non-linear approach to dimension reduction via the adoption of component distributions having a factor-analytic representation for the component-covariance matrices; see also [51]. McLachlan et al. [48,52] exploited the MFA model for clustering microarray gene-expression profiles. For data with clusters having longer than the normal tails, McLachlan et al. [47] adopted the family of multivariate t-distributions for the component factors and errors to establish a robust extension of MFA. More recently, Baek et al. [9] proposed mixtures of common factor analyzers (MCFA) in which the factors are taken to have a common distribution before their transformation to be white noise. A robust version of MCFA using t-component distributions, called mixtures of common factor t analyzers (MCtFA), was subsequently provided by Baek et al. [8]. Wang [72,73] extended the MCFA and MCtFA approaches to accommodate high-dimensional data with possibly missing values. Bayesian treatments of the MFA model have been investigated by Ghahramani and Beal [27] via a variational approximation and Utsugi and Kumagai [70] using the Gibbs sampler and a deterministic algorithm.

* Corresponding author at: Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan.
E-mail address: tilin@nchu.edu.tw (T.-I. Lin).

http://dx.doi.org/10.1016/j.jmva.2015.09.025
0047-259X/

For computational convenience and mathematical tractability, component errors and latent factors in the traditional MFA model are routinely assumed to follow multivariate normal distributions. However, in many applied problems, the data to be analyzed may contain a group or groups of observations whose distributions are moderately or severely skewed. Just like other normal-based mixture models, a slight deviation from normality may seriously affect the estimates of mixture parameters and/or lead to spurious groups, subsequently misleading inference from the data. Wall et al. [71] conducted several simulation studies to explore the influence of non-normal latent factors in the estimation of parameters.

In recent years, there has been growing interest in studying mixtures of skew-normal distributions [37,40], both in the univariate and multivariate cases, as a more general tool for handling heterogeneous data involving asymmetric behavior across sub-populations. Pyne et al. [65] proposed mixtures of multivariate skew-normal and t-distributions based on a restricted variant of the skew-elliptical family of distributions of Sahu et al. [66], which we shall refer to as the restricted multivariate skew-normal (rMSN) distribution. The use of "restricted" was adopted by Lee and McLachlan [33] since it is obtained by imposing the restriction that the p latent skewing variables are all equal in the form of the class of skew elliptical distributions proposed by Sahu et al. [66]. The latter class without this restriction was referred to as "unrestricted". The rMSN distribution is equivalent to the skew normal distribution proposed by Azzalini and Dalla Valle [7]. Lee and McLachlan [34] gave a systematic overview of various existing multivariate skew distributions and clarified their conditioning-type and convolution-type representations. Also, Lee and McLachlan [35] have provided the `EMMIXuskew` package, which implements a closed-form expectation–maximization (EM) algorithm for computing the maximum likelihood (ML) estimates of the parameters for mixtures of unrestricted skew-normal and skew-t distributions.

There have been a few different proposals of mixtures of skew factor models in the literature, see, for instance, mixtures of shifted asymmetric Laplace factor analyzers of Franczak et al. [24], mixtures of generalized hyperbolic factor analyzers of Tortora et al. [69], and mixtures of skew-t factor analyzers (MSTFA) of Murray et al. [61]. An unrestricted version of MSTFA was considered by Murray et al. [62]. Notice that the form of the skew-t distribution used in Murray et al. [61] arises as a special case of the generalized hyperbolic distribution [10], called the generalized hyperbolic skew-t (GHST) distribution. More recently, Murray et al. [63] have put forward a skew version of the MCFA model in which the common factors follow the GHST distribution. The model is henceforth referred to as mixtures of common skew-t factor analyzers (MCSTFA). We should emphasize that the GHST distribution differs from the restricted skew-t distribution in a number of ways, such as different behavior in its tails, for example in the univariate case, with one polynomial and the other exponential [1]. Also, it does not become a skew normal distribution as a limiting case [36].

In this paper, we propose mixtures of skew-normal factor analyzers (MSNFA) where the latent component factors are assumed to follow the family of rMSN distributions in an attempt to model the data adequately in the presence of skewed sub-populations. The proposed model, which is a generalization of the MFA model, can be viewed as a novel approach to achieving dimensionality reduction and representing appropriately non-normal data. ML estimates of the parameters in the model can be computed via the closed-form EM implementations [16,58], and the estimated factor scores can be obtained as by-products within the estimation procedure. The asymptotic covariance matrix of the estimated mixture parameters is obtained by inverting an approximation to the observed information matrix [30].

The rest of the paper is organized as follows. In Section 2, we establish notation and provide a preliminary account of the rMSN distribution. In Section 3, we briefly present the formulation of the skew-normal factor analysis (SNFA) model and study its related properties. Section 4 extends the work to the MSNFA model and presents an EM-type algorithm for obtaining the ML estimates of model parameters. Section 5 describes some practical issues, including the specification of starting values, the stopping rule, model selection and two indices for performance evaluation. The proposed methodology is illustrated through both real and simulated data in Section 6. Some concluding remarks are given in Section 7.

2. The restricted multivariate skew-normal distribution

We begin with a brief review of the rMSN distribution and a study of some essential properties. A unification of families of MSN distributions and several variants and extensions can be found in [2,4]. To establish notation, let $\phi_p(\cdot\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the probability density function (pdf) corresponding to $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, a p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$, and $\Phi(\cdot)$ the cumulative distribution function (cdf) of the standard normal distribution. Further, let $TN(\mu,\sigma^2;(a,b))$ denote the truncated normal distribution for $N(\mu,\sigma^2)$ lying within a truncated interval (a,b).

Following Lee and McLachlan [33], a $p\times 1$ random vector $\boldsymbol{X}$ is said to follow a rMSN distribution with location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$ and skewness vector $\boldsymbol{\lambda}$, denoted by $\boldsymbol{X}\sim rSN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})$, if it can be represented as

$$\boldsymbol{X}=\boldsymbol{\lambda}|U_1|+\boldsymbol{U}_2,\qquad U_1\perp\boldsymbol{U}_2, \tag{1}$$

where $U_1\sim N(0,1)$, $\boldsymbol{U}_2\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the symbol '$\perp$' indicates independence. Letting $W=|U_1|$, a two-level hierarchical representation of (1) is

$$\begin{aligned}\boldsymbol{X}\mid(W=w)&\sim N_p(\boldsymbol{\mu}+\boldsymbol{\lambda}w,\boldsymbol{\Sigma}),\\ W&\sim TN(0,1;(0,\infty))\,.\end{aligned} \tag{2}$$

For computing the moments of W, we use the following proposition.