Computational Statistics and Data Analysis 71 (2014) 183–195

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Learning from incomplete data via parameterized $t$ mixture models through eigenvalue decomposition

Tsung-I Lin *

*Institute of Statistics, National Chung Hsing University, Taichung, Taiwan*
*Department of Public Health, China Medical University, Taichung, Taiwan*

ARTICLE INFO

*Article history:*
Received 9 January 2012
Received in revised form 17 February 2013
Accepted 17 February 2013
Available online 6 March 2013

*Keywords:*
Eigenvalue decomposition
EM-type algorithms
F–G algorithm
Integrated completed likelihood
Model-based clustering
Multivariate $t$ mixture models

ABSTRACT

A framework of using $t$ mixture models with fourteen eigen-decomposed covariance structures for the unsupervised learning of heterogeneous multivariate data with possible missing values is designed and implemented. Computationally flexible EM-type algorithms are developed for parameter estimation of these models under a missing at random (MAR) mechanism. For ease of computation and theoretical developments, two auxiliary indicator matrices are incorporated into the estimating procedure for exactly extracting the location of observed and missing components of each observation. Computational aspects related to the specification of starting values, convergence assessment and model choice are also discussed. The practical usefulness of the proposed methodology is illustrated with real data examples and a simulation study with varying proportions of missing values.


## 1. Introduction

Finite mixture models (FMMs) have attracted considerable attention and been widely used in many disciplines such as supervised and unsupervised clustering, pattern recognition, data mining, computer vision, signal and image processing, machine learning in bioinformatics, and so on. Practical applications may be found in monographs by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), Bishop (2006), and Frühwirth-Schnatter (2006). The Gaussian mixture (GMIX) model (e.g. Redner and Walker, 1984) has been found to be the most popular model-based tool because of its wide applicability and desirable properties. When handling those data with relatively a larger dimension $p$ than the number of observations $n$, however, GMIX may produce unreliable results due to singular or near-singular estimates of the component-covariance matrices.

Traditionally, in various applications, individuals among a population may often be divided into several nonoverlapping groups. Cluster analysis (or clustering) is a task of identifying a natural grouping of observations that are cohesive and separate from the other groups. The GMIX-based clustering method in which the component covariance matrix is parameterized via a variant of eigenvalue decomposition (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Bensmail and Celeux, 1996; Bensmail et al., 1997) is among the most popular model-based clustering techniques due to its versatility of use. Many freely available statistical packages designed for mixture analysis such as `mclust` (Fraley and Raftery, 2003) and `mixmod` (Biernacki et al., 2006) have often been used for model-based clustering. Recently, model-based classification techniques built on latent GMIX models were investigated by McNicholas (2010).

Mixtures of $t$ distributions as originally proposed by McLachlan and Peel (1998), known as $t$ mixture (TMIX) models, have been considered a standard choice in place of GMIX because of their robustness against atypical observations. Peel

* Correspondence to: Institute of Statistics, National Chung Hsing University, Taichung, Taiwan. Tel.: +886 4 22850420; fax: +886 4 22873028.
*E-mail address:* tilin@nchu.edu.tw.

0167-9473/$ – see front matter 
http://dx.doi.org/10.1016/j.csda.2013.02.020

and McLachlan (2000) adopted the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993) for parameter estimation and showed the robustness of the model in clustering. Shoham (2002) presented a robust clustering technique based on two variants of expectation maximization (EM) algorithms (Dempster et al., 1977). Recent developments in Bayesian missing data problems include Lin et al. (2004, 2009). More recently, Andrews and McNicholas (2012) have presented a new family of TMIX models with component covariance matrices parameterized by an eigen-decomposed structure and showed its effectiveness in clustering, classification and discriminant analysis.

Missing data occur frequently due to diverse reasons, especially encountered in areas such as censuses and surveys (Rubin, 1987). Simply deleting the cases with missing values from the analysis may yield substantial biases. Learning mixture models from incomplete data were pioneered by Ghahramani and Jordan (1994), who applied the EM algorithm to conduct maximum likelihood (ML) estimation of GMIX model with arbitrary patterns of missingness. Wang et al. (2004) presented an ordinary EM algorithm to cope with ML estimation of TMIX models in the presence of missing data.

In this paper, we consider the learning of TMIX models with 14 parsimonious eigen-decomposed covariance structures wherever missing data occur. In what follows, the missingness of data is assumed to be missing at random (MAR) with an ignorable mechanism (cf. Rubin, 1976; Schafer, 1997; Little and Rubin, 2002). In this setup, the missingness is unrelated to the missing values, and likelihood inference can ignore the missing data mechanism. Note that the proposed strategy is also valid if mechanism is missing completely at random (MCAR), which is a special case of MAR. In a view of computational aspects, we establish workable EM-type algorithms for ML estimation of model parameters as well as imputation of each missing value. To reduce complications during the estimation procedure, we introduce two permutation matrices for indexing the observed and missing components of each individual item. Further, we offer a conditional predictor to retrieve the missing components and a classifier to allocate partially observed vectors.

The outline of the paper is as follows. In Section 2, for the sake of completeness, we give a brief sketch of TMIX models and the background related to implementing eigenvalue decomposition and diagonalization for component covariance matrices simultaneously. Section 3 presents the development of EM-type algorithms for obtaining ML estimates of model parameters and retrieving a plausible imputed value for each missing cell. Some practical issues including the specification of starting values, the stopping rule and the model selection criterion are addressed in Section 4. In Section 5, we provide results for the simulated data, and in Section 6, we illustrate the usefulness of the proposed method with two real-world data sets. Concluding remarks are made in Section 7.

## 2. Preliminaries

We begin by introducing the TMIX model and briefly describing some related properties. Next, we summarize some features of the 14 possible parameterizations for component covariance matrices. Besides, we review the F–G algorithm (Flury and Gautschi, 1986), which is a complex iterative procedure for solving ML solutions of model parameters with common orientations across different components.

### 2.1. The TMIX model

Consider $n$ independent $p$-dimensional feature vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ which come independently from a nonhomogeneous population with $g$ subgroups. Suppose that each observation $\boldsymbol{y}_j$ is from a $g$-component mixture of multivariate $t$ distributions with density

$$f(\boldsymbol{y}_j \mid \boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i t_p(\boldsymbol{y}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \quad \sum_{i=1}^{g} w_i = 1, \tag{1}$$

where $\boldsymbol{\Theta} = (\boldsymbol{w}, \boldsymbol{\Psi}, \boldsymbol{\nu})$ represents all unknown parameters and $t_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denotes a $p$-variate $t$ density function with location vector $\boldsymbol{\mu}$, positive definite scaling covariance matrix $\boldsymbol{\Sigma}$ and degrees of freedom (df) $\nu \in \mathbb{R}^+ = (0, \infty)$. Here the vector $\boldsymbol{w} = (w_1, \ldots, w_g)$ consists of the mixing proportions, $\boldsymbol{\Psi} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g)$, and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_g)$. Specifically,

$$t_p(\boldsymbol{y}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i) = \frac{\Gamma\left(\frac{\nu_i + p}{2}\right) |\boldsymbol{\Sigma}_i|^{-1/2}}{\Gamma\left(\frac{\nu_i}{2}\right) (\pi \nu_i)^{p/2}} \left(1 + \frac{\Delta_{ij}}{\nu_i}\right)^{-(\nu_i + p)/2}, \tag{2}$$

where $\Delta_{ij} = (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i)$ denotes the Mahalanobis distance between $\boldsymbol{y}_j$ and $\boldsymbol{\mu}_i$ with respect to $\boldsymbol{\Sigma}_i$. The log-likelihood for data consisting of $n$ observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ arisen from a $g$-component TMIX model is given by

$$\ell(\boldsymbol{\Theta} \mid \boldsymbol{y}) = \sum_{j=1}^{n} \log \left( \sum_{i=1}^{g} w_i t_p(\boldsymbol{y}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i) \right).$$

The ML estimates of $\boldsymbol{\Theta}$ can be solved by

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\Theta} \mid \boldsymbol{y}),$$