Journal of Multivariate Analysis 149 (2016) 54–64

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Maximum likelihood inference for the multivariate t mixture model

Wan-Lun Wang [a,*], Tsung-I Lin [b,c]

[a] Department of Statistics, Graduate Institute of Statistics and Actuarial Science, Feng Chia University, Taichung 40724, Taiwan
[b] Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan
[c] Department of Public Health, China Medical University, Taichung 404, Taiwan

ARTICLE INFO

Article history:
Received 23 November 2015
Available online 6 April 2016

AMS subject classifications:
62F05
62F12
62F40

Keywords:
Degrees of freedom
Information matrix
Multivariate t distribution
Robust sandwich estimator
Score vector

ABSTRACT

Multivariate t mixture (TMIX) models have emerged as a powerful tool for robust modeling and clustering of heterogeneous continuous multivariate data with observations containing longer than normal tails or atypical observations. In this paper, we explicitly derive the score vector and Hessian matrix of TMIX models to approximate the information matrix under the general and three special cases. As a result, the standard errors of maximum likelihood (ML) estimators are calculated using the outer-score, Hessian matrix, and sandwich-type methods. We have also established some asymptotic properties under certain regularity conditions. The utility of the new theory is illustrated with the analysis of real and simulated data sets.

1. Introduction

Finite mixture models (FMM) have become one of the most widely used statistical tools for modeling heterogeneous multivariate data arisen from a wide range of research disciplines, such as statistical pattern recognition, econometrics, bioinformatics, and biomedical sciences, to name just a few. For mathematical and computational convenience, the multivariate Gaussian (normal) distribution is the most commonly assumed for mixture components. However, the normality assumption is not always realistically applicable to any data sources. The estimates of component means, variances and covariances, as well as the identification of clustering can be dramatically affected by observations that exhibit atypically longer-than-normal tails in multivariate Gaussian mixture (GMIX) models being fitted. To circumvent such obstacles, Peel and McLachlan [18] proposed multivariate t mixture (TMIX) models and provided the expectation conditional maximization (ECM) algorithm [16] for computing maximum likelihood (ML) estimates of parameters. Conceptually, the TMIX model which, as the name suggests, imposes multivariate t distributions [13] for each component, has long been recognized as a robust approach to handling population heterogeneity and heavy tails in multivariate data.

We say a p-dimensional random vector $\boldsymbol{X}$ follows a multivariate t distribution with location vector $\boldsymbol{\mu}$, positive definite scale-covariance matrix $\boldsymbol{\Sigma}$ and degrees of freedom (DOF) ν, denoted by $\boldsymbol{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, if it has the probability density

* Corresponding author.
E-mail address: wlunwang@fcu.edu.tw (W.L. Wang).

http://dx.doi.org/10.1016/j.jmva.2016.03.009
0047-259X/

function (pdf)

$$t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\boldsymbol{\Sigma}|^{-1/2}}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{p/2}}\left(1+\frac{\Delta}{\nu}\right)^{-(\nu+p)/2}, \tag{1}$$

where $\Delta = (\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. The DOF ν, which controls the thickness of the tails, is used to adjust robustness of inference. Assume that $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ forms a p-dimensional random sample of size n arising from a population with g subclasses $\mathcal{C}_1, \ldots, \mathcal{C}_g$. Each $\boldsymbol{y}_j$ follows a g-component TMIX model, denoted by

$$f(\boldsymbol{y}_j; \boldsymbol{\theta}) = \sum_{i=1}^{g}\pi_i t_p(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{2}$$

where π_i $(i = 1, \ldots, g)$ are mixing proportions satisfying $\sum_{i=1}^{g}\pi_i = 1$. The TMIX model defined in (2) includes the GMIX model as a limiting case when $\nu_i \to \infty$ for all i. To prevent overfitting, it might be useful in practice to impose some constraints on component parameters. Andrews et al. [2] introduced a 'tClass' family of four TMIX models defined by constraining, or not, $\boldsymbol{\Sigma}$ and ν_i to be equal across components. They found in simulations that an appropriate tClass model can be effectively determined by the penalized likelihood criteria such as Bayesian information criterion (BIC [19]) or integrated completed likelihood (ICL [5]). Moreover, their application to real data indicated that the constrained tClass models with equal covariance matrices and/or equal DOF are more likely to be chosen. In a recent related work, Andrews and McNicholas [1] demonstrated in real-data examples that TMIX models with constrained DOF may yield better clustering performance than the unconstrained ones.

Denote the unique parameters in each component by $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i^\top, \boldsymbol{f}_i^\top, \nu_i)^\top$, and the entire parameter by $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_g\}$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{g-1})^\top$ and $\boldsymbol{f}_i$ contains distinct elements in $\boldsymbol{\Sigma}_i$. Accordingly, the log-likelihood function of $\boldsymbol{\theta}$ for a set of observations $\boldsymbol{y}$ is

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{j=1}^{n}\ln f(\boldsymbol{y}_j; \boldsymbol{\theta}). \tag{3}$$

Owing to the complexity of the likelihood function which involves logarithms of a sum, the task of obtaining the score vector and Hessian matrix of (3) is tedious and challenging.

Recently, Boldea and Magnus [6] derived the score vector and Hessian matrix for the GMIX model in explicit expressions and utilized the formulae to estimate the information matrix as well as the standard errors of parameters. For fitting the TMIX models, so far the users must resort to the bootstrap technique [4,9] to calculate the standard errors. Unfortunately, such a resampling procedure can be very time-consuming or even infeasible.

In this paper, we carry out analytical derivations of the score vector and Hessian matrix for the four members of tClass models to estimate the information matrix. We offer the closed-form expressions of Hessian matrices under a general case and three special cases, including (i) equal scale-covariance matrix, (ii) equal DOF, and (iii) equal scale-covariance matrix and equal DOF. Having obtained these results, the variance–covariance matrix of ML estimators can be approximated by using either the outer product of score vector, the Hessian matrix or the robust sandwich-type estimation procedure. In addition, we investigate the asymptotic properties of ML estimators, which are useful for estimating the precision of parameters, constructing confidence intervals, and undertaking the hypothesis testing.

The rest of this paper is structured as follows. In Section 2, we establish the notation. Section 3 formulates the main theoretic result in Theorem 1 and presents three important special cases in Theorems 2–4. In Section 4, we discuss the estimation of variance–covariance matrix of ML estimators and study their asymptotic properties. An application to uranium exploration data set is illustrated in Section 5. Section 6 conducts simulations to examine the finite-sample behavior of the proposed methods. The performance is also compared with the bootstrap-based procedures. Section 7 ends this paper with a short discussion. Proofs of the theoretical results are sketched in the supplemental material (see Appendix A).

2. Notation

We begin by defining the notation to be used throughout the paper. Let $\mathrm{vec}(\boldsymbol{M})$ be the operator that vectorizes a matrix by stacking its columns, and $\mathrm{vec}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_g)$ be the operator that vectorizes a set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_g\}$ with possibly distinct dimensions by stacking them in turn as a pooled column vector. Since the component scale-covariance $\boldsymbol{\Sigma}_i$ in (2) is symmetric, we use $\boldsymbol{f}_i = \mathrm{vech}(\boldsymbol{\Sigma}_i)$ to denote the $p(p+1)/2 \times 1$ vector that contains unique sub-diagonal elements in $\mathrm{vec}(\boldsymbol{\Sigma}_i)$. In addition, we introduce a $p^2 \times p(p+1)/2$ duplication matrix $\boldsymbol{D}$ such that

$$\boldsymbol{D}\mathrm{vech}(\boldsymbol{\Sigma}_i) = \mathrm{vec}(\boldsymbol{\Sigma}_i),$$

which uniquely transforms the half-vectorization of a matrix to its vectorization. Recall from (2) that the mixing proportions π_i have to be all positive and sum to one. Let $\otimes$ be the Kronecker product, which maps two arbitrarily dimensioned matrices into a larger matrix with a specific block structure. Because the differential of (1) involves the first-order and second-order