Computational Statistics and Data Analysis 71 (2014) 220–240

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust growth mixture models with non-ignorable missingness: Models, estimation, selection, and application☆

Zhenqiu (Laura) Lu[a,*], Zhiyong Zhang[b]

[a] University of Georgia, United States

[b] University of Notre Dame, United States

HIGHLIGHTS

- Four non-ignorable missingness models are proposed.
- Three robust models to deal with outliers are proposed.
- A full Bayesian method is implemented.
- Model selection criteria are proposed in a Bayesian context.
- Three simulation studies and one real data case study are conducted.

ARTICLE INFO

Article history:
Received 3 July 2012
Received in revised form 25 July 2013
Accepted 26 July 2013
Available online 7 August 2013

Keywords:
Growth mixture models
Non-ignorable missing data
Robust methods
Bayesian method
Model selecting criteria

ABSTRACT

Challenges in the analyses of growth mixture models include missing data, outliers, estimation, and model selection. Four non-ignorable missingness models to recover the information due to missing data, and three robust models to reduce the effect of non-normality are proposed. A full Bayesian method is implemented by means of data augmentation algorithm and Gibbs sampling procedure. Model selection criteria are also proposed in the Bayesian context. Simulation studies are then conducted to evaluate the performances of the models, the Bayesian estimation method, and selection criteria under different situations. The application of the models is demonstrated through the analysis of education data on children's mathematical ability development. The models can be widely applied to longitudinal analyses in medical, psychological, educational, and social research.

1. Introduction

Mixture models offer natural models for unobserved population heterogeneity. The importance of mixture models, their enormous developments, and their frequent applications are not only remarked by a number of recent books but also by a diversity of journal publications. For example, Computational Statistics & Data Analysis has published two special issues on mixture models (Bohning and Seidel, 2003; Bohning et al., 2007) and the current issue is a new one. Latent growth models are used to study individuals' latent growth trajectories by analyzing the variables of interest on the same individuals repeatedly through time (e.g., Bollen and Curran, 2006; McArdle and Bell, 1999; Meredith and Tisak, 1990). These models are very popular in biological, psychological, educational, and social sciences (e.g., Collins, 1991; Fitzmaurice et al., 2004; Singer and Willett, 2003). By combining latent growth models and finite mixture models (e.g., McLachlan and Peel, 2000),

☆ Supplementary tables summarized from simulation results can be accessed from the web site of http://nd.psychstat.org/research/csda2013.

* Correspondence to: 325 Aderhold Hall, University of Georgia, Athens, GA 30602, United States. Tel.: +1 706 542 4540; fax: +1 706 542 4240.
E-mail address: zlu@uga.edu (Z. Lu).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.07.036

growth mixture models (GMMs, see, e.g., Lubke and Muthén, 2005; Muthén, 2004; Muthén et al., 2011), therefore, provide researchers with a flexible set of models for growth data with latent population heterogeneity.

However, with the increase in complexity of model specification comes an increase in difficulties estimating GMMs. First, missing data are almost inevitable (e.g., Little and Rubin, 2002; Yuan and Lu, 2008), especially in longitudinal studies (e.g., Jelicic et al., 2009; Roth, 1994). Little and Rubin (2002) distinguished *ignorable* and *non-ignorable* missingness mechanisms. Non-ignorable missingness is a crucial and serious concern, because not attending to it may result in severely biased statistical estimates, standard errors, and associated confidence intervals (e.g., Little and Rubin, 2002; Schafer, 1997; Zhang and Wang, 2012). However, most of the literature on the problems of missing data focuses on ignorable missingness (e.g., Schafer and Graham, 2002). Second, data may have outliers (e.g., Hoaglin et al., 1983), particularly in social and behavioral sciences (e.g., Micceri, 1989). The consequences of applying a normal distribution assumption to such data include unreliable parameter estimates (e.g., Pan and Fang, 2002), unreliable standard errors and confidence intervals, and misleading statistical tests and inference (e.g., Yuan and Bentler, 1998). Third, for complex models such as GMMs with missing data and outliers, maximum likelihood methods might fail or provide biased estimates (e.g., Yuan and Zhang, 2012). Most of the previous estimations have relied on maximum likelihood methods for parameter estimation and have carried out inferences through conventional likelihood procedures (e.g., Song et al., 2014). Fourth, even with effective estimation methods, model selection in such complex situations becomes extremely difficult. Traditional criteria for model selection, including Akaike's Information Criterion (AIC, Akaike, 1974), Bayesian Information Criterion (BIC, Schwarz, 1978), consistent Akaike's Information Criterion (CAIC, Bozdogan, 1987), sample-size adjusted Bayesian Information Criterion (ssBIC, Sclove, 1987), and Deviance Information Criterion (DIC, Spiegelhalter et al., 2002), are not uniformly effective due to latent effects and missing data (e.g., Celeux et al., 2006).

Few studies have discussed how to address these common problems in longitudinal research in the framework of GMMs. Lu et al. (2011) discussed GMMs with non-ignorable missing data using Bayesian methods. However, they (1) considered only one type of non-ignorable missingness, (2) assumed data are normally distributed without any outlier, and (3) did not propose any model selection criterion.

This article extends the study of Lu et al. (2011) and addresses these challenges in GMMs: missing data, outliers, estimation, and model selection. Regarding missing data, we propose new types of non-ignorable missingness in GMMs and investigate their influences on model estimation under different situations. Regarding outliers, we use robust models (e.g., Lange et al., 1989) to minimize the effects of contaminated data. Because convenient robust methods often lead to other problems such as under-estimation of standard errors (e.g., Poon and Poon, 2002), we adopt t-distributions to deal with heavy-tailed data (Lin et al., 2004; Zhang et al., 2013). Regarding estimation methods, as Bayesian methods provide many advantages of estimating complex models (e.g., Dunson, 2000), we propose a full Bayesian approach, which is flexible enough to estimate a variety of models with different missing data mechanisms, contaminated data, and mixture structure. Regarding model selection, we propose several selection criteria in the Bayesian context. The performances of these criteria are investigated under different situations.

In the next section of this article, Section 2, we propose GMMs with different types of missing data and outliers. In Section 3, we present Bayesian estimation methods. In Section 4, we propose Bayesian model selection criteria. In Section 5, we conduct three simulation studies on Bayesian GMMs under different conditions. In Section 6, we demonstrate the application of the GMMs and the Bayesian method by analyzing real education data on children's mathematical ability development. In Section 7, we draw conclusions. The Appendices present the technical details of our analyses.

2. Models

The density function of a growth mixture model is

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i), \tag{1}$$

where π_k is the invariant class probability (or weight) for class k, ($k = 1, \ldots, K$), satisfying $0 \le \pi_k \le 1$ and $\sum_{k=1}^{K} \pi_k = 1$ (e.g., McLachlan and Peel, 2000), and $f_k(\mathbf{y}_i)$ is the density for the kth class, in which $\mathbf{y}_i$ is a $T \times 1$ vector of outcomes for participant i ($i = 1, \ldots, N$) following a latent growth model

$$\begin{cases} \mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{e}_i, \\ \boldsymbol{\eta}_i = \boldsymbol{\beta} + \boldsymbol{\xi}_i, \end{cases} \tag{2}$$

where $\boldsymbol{\eta}_i$ is a $q \times 1$ vector of latent effects, $\boldsymbol{\Lambda}$ is a $T \times q$ matrix of factor loadings for $\boldsymbol{\eta}_i$, $\mathbf{e}_i$ is a $T \times 1$ vector of residual or measurement errors, $\boldsymbol{\beta}$ is a $q \times 1$ vector of fix-effects, and $\boldsymbol{\xi}_i$ captures the variation of $\boldsymbol{\eta}_i$.

In the Extended Growth Mixture Models (EGMMs, Muthén and Shedden, 1999), π_k is not invariant any more for all individuals in class k. It is allowed to vary individually depending on covariates, so it is expressed as $\pi_{ik}(\mathbf{x}_i)$. In this study, a probit link function is used

$$\begin{cases} \pi_{i1}(\mathbf{x}_i) = \Phi(X_i' \boldsymbol{\varphi}_1), \\ \pi_{ik}(\mathbf{x}_i) = \Phi(X_i' \boldsymbol{\varphi}_k) - \Phi(X_i' \boldsymbol{\varphi}_{k-1}), \quad (k = 2, 3, \ldots, K-1) \\ \pi_{iK}(\mathbf{x}_i) = 1 - \Phi(X_i' \boldsymbol{\varphi}_{K-1}), \end{cases} \tag{3}$$