Computational Statistics and Data Analysis 71 (2014) 43–51

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixture models for clustering multilevel growth trajectories

S.K. Ng [a,*], G.J. McLachlan [b]

[a] School of Medicine, Griffith Health Institute, Griffith University, Australia

[b] Department of Mathematics, University of Queensland, Australia

ARTICLE INFO

Article history:
Received 30 June 2012
Received in revised form 28 November 2012
Accepted 14 December 2012
Available online 21 December 2012

Keywords:
Mixture models
Random effects
Multilevel growth trajectories
EM algorithm

ABSTRACT

Mixture model-based methods assuming independence may not be valid for clustering growth trajectories arising from multilevel studies because longitudinal data collected from the same unit are often correlated. A mixture of mixed effects models is considered to capture the correlation using multilevel and multivariate random effects. Furthermore, the mixing proportions are allowed to depend on covariates. The additional information is thus incorporated into the mixture model to adjust for individual probabilities of membership of the components. The proposed method is illustrated using simulated and real multilevel growth trajectory data sets from various scientific fields.

1. Introduction

Mixture model-based clustering methods implemented via the EM algorithm are being commonly used in a wide range of applications in the cluster analysis of multivariate data (McLachlan et al., 2004; McLachlan and Peel, 2000; Ng et al., 2012). With this approach to clustering, a common assumption is to take all the observations on the entities to be independent of one another. We let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n where $\boldsymbol{Y}_j$ is a p-dimensional random vector with probability density function being modeled as a mixture of g multivariate normal component densities $\phi(\boldsymbol{y}_j; \boldsymbol{\psi}_h)$, where $\boldsymbol{\psi}_h$ is the vector of unknown parameters in the hth component density ($h = 1, \ldots, g$). The independence assumption implies that the likelihood function can be expressed as

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} \sum_{h=1}^{g} \pi_h(\boldsymbol{x}_j; \boldsymbol{\alpha}) \phi(\boldsymbol{y}_j; \boldsymbol{\psi}_h), \tag{1}$$

where $\boldsymbol{\Psi}$ is the vector containing all the unknown parameters in the mixture model and the mixing proportions $\pi_h(\boldsymbol{x}_j; \boldsymbol{\alpha})$ depend on a vector of covariates $\boldsymbol{x}_j$ associated with the response $\boldsymbol{y}_j$, and where $\boldsymbol{\alpha}$ contains the unknown parameters in the mixing proportions. For many applied problems in the context of social, medical, and health sciences, the data collected could exhibit a hierarchical or multilevel structure (Ng and McLachlan, 2007; Ng et al., 2004). Data collected from the same unit (such as hospital) are correlated and the independence assumption for cluster analysis is no longer valid. Ignoring the interdependence between hierarchical or multilevel data can result in overlooking the importance of certain unit-specific effects and lead to spurious or misleading clustering results (Goldstein, 2010; Ng and McLachlan, 2007).

In clustering growth trajectories, growth mixture models (Muthén, 2004; Muthén and Asparouhov, 2009; Vermunt, 2007) and mixture latent growth models (Vermunt, 2003, 2007) have been adopted to identify different classes of trajectory

* Correspondence to: School of Medicine, Griffith University, Meadowbrook, QLD 4131, Australia. Tel.: +61 7 33821525; fax: +61 7 33821338.
E-mail addresses: s.ng@griffith.edu.au (S.K. Ng), g.mclachlan@uq.edu.au (G.J. McLachlan).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.12.007

patterns and predictors of membership in these classes. The growth mixture models extend the conventional growth model (Raudenbush and Bryk, 2002) and the latent class growth analysis (LCGA) approach (Nagin and Land, 1993) to allow the presence of different clusters of trajectory patterns and heterogeneity in individual trajectories that vary around the mean trajectory within a cluster. The growth mixture models are thus flexible to model individual growth trajectories from unobserved subpopulations (latent trajectory clusters) with individual variation in growth parameters that are captured by random effects.

The use of random-effects modeling in a mixture framework has been considered in another topic concerning the analysis of gene expression data with repeated measurements (Celeux et al., 2005; Grün et al., 2012; Ng et al., 2006), where the major aim is to reveal groups of genes with similar profiles that may be related to the same underlying biological process or molecular pathway (McLachlan et al., 2004). Besides the difference in aims, another distinct feature of this type of applications is that no covariate risk factor is usually available for the clustering of genes. This implies that $\pi_h(\boldsymbol{x}_j; \boldsymbol{\alpha}) = \pi_h$ in (1), which is the prior probability that the jth gene belongs to the hth component given observed gene expression profile $\boldsymbol{y}_j$. Extended from a multivariate Gaussian mixture model (McLachlan and Peel, 2000), Celeux et al. (2005) considered a mixture of linear mixed-effects models (LMMs) with a single random effects term to account for the correlation between repeated measurements at time t for each gene. Specifically, the unconditional distribution of all $\boldsymbol{y}_j$ that arise from the hth component is given by

$$\boldsymbol{y}^h \sim N_{n_h \times r}(\boldsymbol{W}\boldsymbol{\beta}_h, \theta_h \boldsymbol{U}\boldsymbol{U}^T + \sigma_h^2 \boldsymbol{I}), \tag{2}$$

where $\boldsymbol{\beta}_h$ is the fixed effect vector for the hth component, θ_h is the random effect variance, σ_h^2 is the residual variance, $\boldsymbol{W}$ and $\boldsymbol{U}$ are design matrices for the corresponding fixed and gene-specific random effects (Celeux et al., 2005). In (2), n_h and r are, respectively, the number of genes belonging to the hth cluster and the number of repeated measurements for each tissue sample, $\boldsymbol{I}$ is an identify matrix, and the superscript T represents vector transpose. Ng et al. (2006), on the other hand, considered a mixture of LMMs with two random effects terms to separately account for the correlation between repeated measurements and within tissue samples, respectively. As the tissue-specific random effects induce dependency among the expression values of genes from the same component and from the same tissue, their model can handle correlated gene-expression profiles without the requirement of independence assumption for the genes as with other methods. More recently, Grün et al. (2012) considered a mixture of linear additive models (LAMs) for the clustering of time-course gene expression data, where random effects on individual genes are incorporated in the component densities and estimated using regularized likelihood approaches. In contrast to the mixture of LMMs, the mixture of LAMs assume that the repeated observations for the same gene are independent given the component membership (Grün et al., 2012).

In this paper, we extend earlier work of Ng et al. (2006) aforementioned by incorporating a bi-level multivariate random effects structure within the mixture models framework and allowing the mixing proportions to depend on covariate risk factors. We wish to focus on the applications of this new mixture of mixed effects models for clustering multilevel growth trajectories that are obtained from hierarchical units such that their trajectories are correlated within a unit. The random-effects modeling approach proposed in this paper is thus different from those considered for the analysis of gene expression data (Celeux et al., 2005; Grün et al., 2012; Ng et al., 2006), where only individual-level random effects are adopted.

In contrast to the existing growth mixture model approaches, our method does not require the independence assumption for individual trajectories, which will not hold in practice for data with hierarchical or multilevel structure. Moreover, the proposed model allows bi-level multivariate random effects for capturing the variation among higher level study units and the individual level variation, respectively. Furthermore, the second extension facilitates the provision of a better clustering result where there exists additional information on an individual's risk factors that have an impact on membership of subpopulations. The extensions thus create a wider applicability of mixture model-based approaches for clustering hierarchically structured trajectory data. Simulated multilevel data and a real example will be given to illustrate the proposed method.

2. Mixtures of random effects models

With a bi-level hierarchical data structure, it is assumed that there are M higher level units, and within each unit there are n_i study subjects ($i = 1, \ldots, M$). Thus, the total number of participants is $n = \sum_{i=1}^{M} n_i$. The objectives are to identify the subpopulation structure within the participants and the risk factors that have impact on the trajectory patterns of an outcome measure. We denote $\boldsymbol{y}_{ij}$ the observed p-dimensional trajectory for the jth individual in the ith unit and $\boldsymbol{x}_{ij}$ a vector of risk factors associated with $\boldsymbol{Y}_{ij}$.

In this paper, we formulate a LMM (McCulloch and Searle, 2001) for the mixture components in which covariance information can be incorporated into the clustering process. Specifically, it is assumed that the effects imposed by the higher level units are random and shared among participants collected from the same unit. In addition, random effects are introduced to capture individual variation in trajectories that vary around the mean trajectory within a unit at various time periods s, where $s \leq p$. Let $\boldsymbol{b}_{hi} = (b_{hi1}, \ldots, b_{hip})^T$ and $\boldsymbol{c}_{hij} = (c_{hij1}, \ldots, c_{hijs})^T$ the unobserved unit-specific and individual-specific random effects, respectively. This bi-level random-effects modeling is in contrast to that considered in Ng et al. (2006) with individual-specific random effects only. With reference to the mixture framework of multivariate normal component