Computational Statistics and Data Analysis 93 (2016) 119–130

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Finite mixture of nonlinear mixed-effects joint models in the presence of missing and mismeasured covariate, with application to AIDS studies

Xiaosun Lu, Yangxin Huang *, Yiliang Zhu

Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612, USA

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 8 April 2014
Accepted 9 April 2014
Available online 18 April 2014

Keywords:
AIDS clinical trials
Bayesian inference
Finite mixture models
NLME models
Longitudinal data analysis

ABSTRACT

It is a common practice to analyze longitudinal data using nonlinear mixed-effects (NLME) models. However, the following issues may standout. (i) Individuals may be possibly from a heterogeneous population following more than one mean trajectories, while a homogeneous population assumption for model structure may be unrealistically obscuring important features of between- and within-subject variations; (ii) some covariates may be missing and/or measured with errors. There has been few studies concerning both population heterogeneity and covariates measured with errors and missing data features simultaneously in longitudinal data analysis. A finite mixture of NLME joint (FMNLMEJ) models is developed to address simultaneous impact of both features under Bayesian framework, which offers a route to estimate not only model parameters but also probabilities of class membership. An AIDS data set is analyzed to demonstrate the methodologies in comparison of the proposed FMNLMEJ model with a commonly used NLME model.

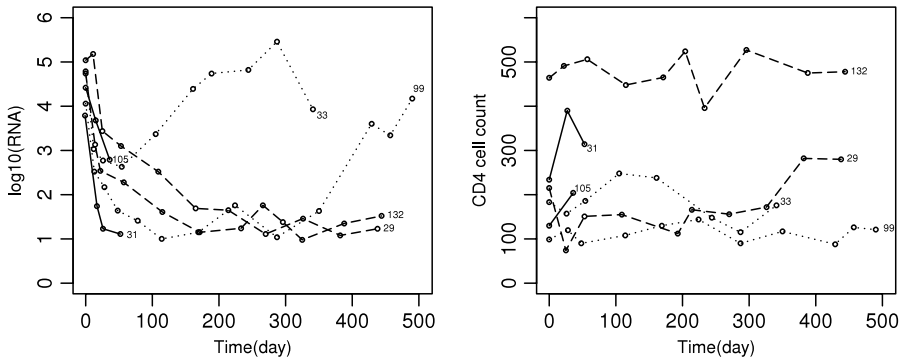
1. Introduction

HIV dynamic modeling based on HIV-1 RNA copies in plasma (viral load) after initiation of antiretroviral treatment (ART) has been one of the most important areas in AIDS research in the past two decades. One is interested in estimating parameters in the viral dynamic models to acquire more comprehensive understanding of the pathogenesis of HIV infection and to assess the effectiveness of ART (Perelson et al., 1997; Wu and Ding, 1999). Various statistical models and methods have been used in conjunction with HIV dynamic models for estimating those parameters including, but not limited to, linear and nonlinear regression (Perelson et al., 1997), nonlinear mixed-effects (NLME) modeling approach (Wu and Ding, 1999; Wu et al., 1998), nonparametric NLME modeling approach (Wu and Zhang, 2006; Liu and Wu, 2007), joint modeling approach via Monte Carlo EM algorithm (Liu and Wu, 2007; Wu, 2002), and Bayesian NLME modeling approach via Markov chain Monte Carlo (MCMC) procedure (Huang et al., 2006; Huang and Dagne, 2011; Huang et al., 2012). Those models and methods so far have provided us better understanding of the pathogenesis of HIV infection and the treatment effects of ART.

However, the majority of existing statistical models was based on a common homogeneous population assumption that all patients follow the same mean trajectory and the large inter-individual variation is accommodated by random-effects and/or time-varying covariates. This typical inter-individual variation feature is shown in Fig. 1(a), viral load trajectory profiles of six representative patients in an AIDS Clinical Trial (ACTG398) study (Hammer et al., 2002) (see Section 3 for

* Corresponding author. Tel.: +1 813 974 8209; fax: +1 813 974 4719.
E-mail address: yhuang@health.usf.edu (Y. Huang).

http://dx.doi.org/10.1016/j.csda.2014.04.003
0167-9473/



(a) Profiles of viral load in $\log_{10}$ scale. (b) Profiles of CD4 cell count.

Fig. 1. (a) Profile of viral load (response) in $\log_{10}$ scale for six representative patients. Trajectory class 1: decrease rapidly and constantly in a short-term period (solid lines, ID: 31, 105); class 2: decrease at the beginning and then maintain stable at a low level (dashed lines, ID: 29, 132) and class 3: decrease at the beginning, but rebound later (dotted line, ID: 33, 99). (b) Profile of corresponding CD4 (covariate) cell count in an AIDS clinical trial study.

details of this study and data). Based on the profiles in Fig. 1(a), viral load trajectories can be roughly classified into three classes: (i) decrease rapidly and constantly in a short-term period (two solid lines, ID: 31, 105), (ii) decrease at the beginning and then stay stable at a low level (two dashed lines, ID: 29, 132), and (iii) decrease at the beginning, but rebound later (two dotted line, ID: 33, 99). Along with these observations, we can reasonably assume that patients are from a population which consists of three relatively homogeneous classes and, thus, it is motivated to consider a finite mixture of NLME models for such data set. Instead of considering individual variation around a single mean trajectory, a finite mixture model allows different classes of individuals to vary around several different mean trajectories. In this sense, finite mixture modeling may better capture inter-individual variation because it does not require that all individuals follow the same mean trajectory function over time. Specific finite mixture models such as growth mixture model (GMM) and latent curve model (LCM), a special type of GMM, are often employed in social longitudinal studies (Muthén and Shedden, 1999; Muthén et al., 2002; Pauler and Laird, 2000) to cluster individuals based on mixture model fitting. However, most finite mixture models are currently based on linear (polynomial) (Muthén and Shedden, 1999; Muthén et al., 2002) or piecewise linear (Pauler and Laird, 2000) mean functions, partially because inference process can be conveniently carried out since the likelihood function of a mixture model based on these linear mean functions has a closed form Muthén and Shedden (1999). When a mixture model is extended to incorporate nonlinear mean functions, inferential procedures can become complicated dramatically because a closed form of likelihood function is no longer available. We demonstrate a Bayesian inferential approach with application to an AIDS clinic data to estimate both model parameters and class membership probabilities based on a finite mixture model with nonlinear mean functions. It was noticed that, if the data exhibit heterogeneity, finite mixture models not only fit data "better" than the single-class model, but also provide an efficient modeling approach-based clustering and classification. Toward this end, probability of belonging to each class can be obtained at both the individual and population level. The estimated probabilities may help clinicians to refine the treatment strategy. For example, clinicians would be able to plan a better general treatment strategy if they know the proportion of patients who have viral load rebound and to provide a better individualized treatment if the probability of a patient experiencing viral load rebound is known.

Another challenge when we analyzing longitudinal data is possibly missing and mismeasured covariate which could lead to bias estimates (Wu, 2002). In the scenario of HIV dynamics, the covariates such as CD4 are often measured with substantial errors, and the covariates usually contain missing data possibly because viral loads and CD4 are not measured on the same scheme. Various covariate mixed-effect models were investigated in the literature (Wu, 2002; Huang et al., 2006; Huang and Dagne, 2011; Huang et al., 2012), but there has been relatively few studies concerning longitudinal data with both heterogeneous response and missing and mismeasured covariates. We provide a unified approach to simultaneously address these typical features often observed in longitudinal data.

The rest of this paper is organized as follows. Section 2 presents a finite mixture of NLME joint (FMNLMEJ) model in the presence of missing and mismeasured covariates and associated Bayesian inferential method in a general form. In Section 3, we describe an AIDS data set that motivated this research and discuss the specific FMNLMEJ model, formulated by three mean functions of different mixture components for viral load response. In Section 4, we apply the proposed methodologies to an AIDS data set described in Section 3 and report the analysis results. Finally, we conclude the paper with a discussion in Section 5.

2. Mixture of joint models and Bayesian inferential procedures

To set up the notation, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be the vector of observed outcomes on individual i, in which y_{ij}, $(i = 1, \ldots, n;\ j = 1, \ldots, n_i)$, is observed measurement at time t_{ij}; $\boldsymbol{x}_{ij}$ be a $(p \times 1)$ vector including time t_{ij} and possibly other covariates of the ith individual.