Computational Statistics and Data Analysis 56 (2012) 4190–4203

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A Bayesian approach for generalized random coefficient structural equation models for longitudinal data with adjacent time effects

Xin-Yuan Song[a,*], Nian-Sheng Tang[b], Sy-Miin Chow[c]

[a] Department of Statistics, Chinese University of Hong Kong, Hong Kong
[b] Department of Statistics, Yunan University, China
[c] Department of Psychology, University of North Carolina at Chapel Hill, USA

ARTICLE INFO

Article history:
Received 28 July 2011
Received in revised form 22 March 2012
Accepted 24 April 2012
Available online 7 May 2012

Keywords:
Adjacent time effects
Bayesian approach
Longitudinal data
Random regression coefficients
Structural equation models

ABSTRACT

This paper proposes a generalized random coefficient structural equation model for analyzing longitudinal data by incorporating the correlated structure due to adjacent time effects and by allowing structural parameters to vary across individuals. The coregionalization for modeling multivariate spatial data is adopted to formulate the correlated structure between adjacent time points. A Bayesian approach coupled with the Gibbs sampler and the Metropolis–Hastings algorithm is developed to obtain the Bayesian estimates of unknown parameters and latent variables simultaneously. A simulation study and a real example related to an emotion study are presented to illustrate the newly developed methodology.

1. Introduction

Latent variables that cannot be directly measured by a single variable are common in practical research. Structural equation modeling is a very popular multivariate method that has been widely used to assess regression-type relationships among latent variables. In general, structural equation models (SEMs) combine the idea of factor analysis and regression through two of its components. The first component is a confirmatory factor analysis model for measuring the latent variables via several response variables with the measurement errors being taken into account. The second component is a regression-type structural equation for assessing the effects of explanatory latent variables on the outcome latent variables of interest. Through the usage of some software (LISREL, EQS), SEMs have been widely applied to behavioral, social, and biomedical sciences (Bollen, 1989; Jöreskog and Sörbom, 1996; Bentler and Wu, 2002; Sanchez et al., 2005; Lee, 2007).

One of our objectives is to establish a novel SEM for analyzing multivariate response variables measured at a large number of time points. Let $\boldsymbol{y}_{ij}$ be a $p \times 1$ random vector that represents response variables of the ith individual ($i = 1, \ldots, I$) measured at time j ($j = 1, \ldots, T_i$). In many longitudinal datasets, for instance those related to finance and medical research, characteristics of an individual (e.g. a stock or a patient) at time j are highly correlated with those at times $j-l$ or $j+l$ when l is small (e.g. $l = 1$ or 2), but slightly correlated with those at times $j - l$ or $j + l$ when l is large. Hence, in model building, there is a need to incorporate a component that considers this kind of correlation structures induced by adjacent time effects. For large p and T_i, the dimension of this correlation matrix can be very high. Inspired by the spirit of SEM to express the $p \times 1$ vector of response variables through a smaller number of latent variables, we model the correlation structure due to adjacent time effects by introducing a linear model of coregionalization to the latent variables in structural equation.

* Corresponding author. Tel.: +852 26097935; fax: +852 26035188.
E-mail address: xysong@sta.cuhk.edu.hk (X.-Y. Song).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.04.016

Conventional analysis of the latent variable assumes a common model for all individuals and ignore the influence of possible unobserved heterogeneity. This may result in biased parameter estimates and misleading inference (see Ansari et al., 2002, and references therein). To account for unobserved heterogeneity, theoretically one can formulate a SEM with an unknown parameter vector at each individual for assessing the changes of parameters with respect to individual. As the whole model will consist of many SEMs that may differ across individuals, the total number of parameters can become very large when sample size increases. This problem is more serious when the correlation structure due to time effects is incorporated in the model. Hence, in our model building, a common measurement equation is formulated for all individuals, that is the number of factors, the loading matrix, and other associated parameters in the measurement equation are invariant for all individuals. In formulating the crucial structural equation, we consider random coefficients in order to capture how regression coefficients vary across different individuals. Here, random coefficients are used to reveal the structural heterogeneity among individuals. These random coefficients can also be used to explain that the importance of latent and observed explanatory variables considered in structural equation models vary across individuals. Finally, nonignorable missing data are accommodated in the proposed SEM to deal with the frequent occurrence of missing data in longitudinal studies.

Although some latent variable models or specific SEMs have been developed to analyze longitudinal data, their objectives and/or formulations are quite different from the proposed model. For instance, longitudinal models in the statistical literature, such as the linear mixed model (Verbeke and Molenbergh, 2000), the generalized linear mixed model (Diggle et al., 2002), and the dynamic latent variable model (Dunson, 2003), involve neither a structural equation with explanatory latent variables, nor the correlation structure due to adjacent time effects. Song et al. (2008) recently proposed a two-level longitudinal SEM for assessing various model characteristics that dynamically change over time. Again, their model did not consider the correlation structure due to adjacent time effects, and the regression coefficients in their structural equation were not random.

The paper is organized as follows. We describe a generalized random coefficient SEM for longitudinal data with adjacent time effects in Section 2. In Section 3, a Bayesian approach compiled with Markov Chain Monte Carlo (MCMC) methods is developed. To demonstrate the newly developed methodology, results obtained from a simulation study and an analysis of a real dataset are presented in Section 4. A discussion is given in Section 5. Technical details are provided in Appendices A and B.

2. Model specification

Let y_{ijk} be the kth ($k = 1, \ldots, p$) component of $\boldsymbol{y}_{ij}$ observed at time j ($j = 1, \ldots, T_i$) of the ith ($i = 1, \ldots, I$) individual. We assume that given $\boldsymbol{\omega}_{ij}$, y_{ijk} are conditionally independent and come from the exponential family with a canonical parameter ϑ_{ijk} and a mean that is a function of the vector of latent variables $\boldsymbol{\omega}_{ij}$. That is, y_{ijk} have probability density function:

$$p(y_{ijk}|\boldsymbol{\omega}_{ij}) = \exp\left[\{y_{ijk}\vartheta_{ijk} - b(\vartheta_{ijk})\}/\psi_k + c(y_{ijk}, \psi_k)\right]$$

with $E(y_{ijk}|\boldsymbol{\omega}_{ij}) = \dot{b}(\vartheta_{ijk})$ and $\mathrm{Var}(y_{ijk}|\boldsymbol{\omega}_{ij}) = \psi_k\ddot{b}(\vartheta_{ijk})$, where $b(\cdot)$ and $c(\cdot)$ are specific known differentiable functions. As the exponential family distribution includes a lot of distributions, such as binomial, Poisson, normal, and gamma, as its special cases, we allow different kinds of manifest variables in the analysis. Following Lee and Tang (2006), we consider the following measurement equation to identify the latent variables in $\boldsymbol{\omega}_{ij}$ via the manifest variables (indicators) in $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijp})^T$:

$$g(\vartheta_{ijk}) = \mu_k + \boldsymbol{\Lambda}_k^T\boldsymbol{\omega}_{ij}, \tag{1}$$

where μ_k is an intercept, and $\boldsymbol{\Lambda}_k$ is a $q \times 1$ vector of loading factors. Let $\boldsymbol{\omega}_{ij} = (\boldsymbol{\eta}_{ij}^T, \boldsymbol{\xi}_{ij}^T)^T$ be a partition of $\boldsymbol{\omega}_{ij}$ into outcome latent variables in $\boldsymbol{\eta}_{ij}$ ($q_1 \times 1$) and explanatory latent variables in $\boldsymbol{\xi}_{ij}$ ($q_2 \times 1$), $q_1 + q_2 = q$. The following structural equation is used to model the relationship between $\boldsymbol{\eta}_{ij}$ and $\boldsymbol{\xi}_{ij}$:

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\Pi}_i\boldsymbol{\eta}_{ij} + \boldsymbol{B}_i\boldsymbol{x}_{ij} + \boldsymbol{\Gamma}_i\boldsymbol{G}(\boldsymbol{\xi}_{ij}) + \boldsymbol{\delta}_{ij}, \tag{2}$$

where $\boldsymbol{x}_{ij}$ is an $s \times 1$ vector of covariates; $\boldsymbol{G}(\boldsymbol{\xi}_{ij}) = (g_1(\boldsymbol{\xi}_{ij}), \ldots, g_t(\boldsymbol{\xi}_{ij}))^T$ is a $t \times 1$ nonzero vector-valued function with differential functions $g_1, \ldots, g_t$ and $t \geq q_2$; $\boldsymbol{\Pi}_i$ ($q_1 \times q_1$), $\boldsymbol{B}_i$ ($q_1 \times s$), and $\boldsymbol{\Gamma}_i$ ($q_1 \times t$) are structural coefficient matrices denoting the effect of $\boldsymbol{\eta}_{ij}$, $\boldsymbol{x}_{ij}$, and $\boldsymbol{\xi}_{ij}$ on $\boldsymbol{\eta}_{ij}$, respectively; $\boldsymbol{\delta}_{ij}$ ($q_1 \times 1$) is a vector of residuals, and it is assumed that $\boldsymbol{\xi}_{ij}$ and $\boldsymbol{\delta}_{ij}$ are independent. The covariates $\boldsymbol{x}_{ij}$ can be explanatory variables or other variables that are significant to explain $\boldsymbol{\eta}_{ij}$.

Unlike conventional SEMs, here we allow structural coefficients $\boldsymbol{\pi}_i = (\boldsymbol{\Pi}_i, \boldsymbol{B}_i, \boldsymbol{\Gamma}_i)$ to vary across individuals. To account for structural heterogeneity, we model structural parameters in $\boldsymbol{\pi}_i$ via the following equation:

$$\boldsymbol{\pi}_i = \boldsymbol{z}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{v}_i, \tag{3}$$

where $\boldsymbol{z}_i$ is a $q_1 \times \kappa$ matrix of individual-level covariates that are useful in explaining $\boldsymbol{\pi}_i$, $\boldsymbol{\beta}$ is a $\kappa \times (q_1 + s + t)$ regression coefficients, $\boldsymbol{U}_i$ is a $q_1 \times 1$ indicator vector of 0's and 1's, $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{i,q_1+s+t})$ is a $1 \times (q_1 + s + t)$ random vector that is independent of $\boldsymbol{\delta}_{ij}$ and $\boldsymbol{\xi}_{ij}$. Here we assume that $\boldsymbol{v}_i$ is distributed as $N(\boldsymbol{0}, \boldsymbol{\Upsilon})$, where $\boldsymbol{\Upsilon}$ is a $(q_1 + s + t) \times (q_1 + s + t)$ covariance matrix that represents the covariation of the structural parameters resulting from unobserved individual-level variables. The linear mixed effect model (3) includes both fixed effect $\boldsymbol{\beta}$ and random effect $\boldsymbol{v}_i$, in which the fixed effect $\boldsymbol{\beta}$