Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables

Xin-Yuan Song*,†, Ye-Mao Xia and Sik-Yum Lee

Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

SUMMARY

Recently, structural equation models (SEMs) have been applied for analyzing interrelationships among observed and latent variables in biological and medical research. Latent variables in these models are typically assumed to have a normal distribution. This article considers a Bayesian semiparametric SEM with covariates, and mixed continuous and unordered categorical variables, in which the explanatory latent variables in the structural equation are modeled via an appropriate truncated Dirichlet process with a stick-breaking procedure. Results obtained from a simulation study and an analysis of a real medical data set are presented to illustrate the methodology. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: blocked Gibbs sampler; model selection; modified truncated Dirichlet process; phenotype and genotype latent variables; structural equation models

1. INTRODUCTION

In biological and medical sciences, it is common to encounter latent variables or constructs that cannot be directly measured by a single observed variable, but instead are assessed through a number of observed variables. Structural equation models (SEMs) [1] are a flexible class of models for modeling of multivariate correlated data to assess the interrelationships among observed and latent variables. In general, SEMs are formulated by two major components. The first component is the measurement equation, which is basically a confirmatory factor analysis model for relating the latent variables with their associative observed variables. Its main purpose is to group the correlated observed variables to ‘measure’ their corresponding latent variables (factors) with the measurement errors being taken into account. The relationship among latent variables is then examined by the second component, which is a regression-type structural equation that regresses outcome

*Correspondence to: Xin-Yuan Song, Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.
†E-mail: xysong@sta.cuhk.edu.hk

Contract/grant sponsor: Research Grants Council of the Hong Kong Special Administration Region; contract/grant numbers: 404507, 450607

Received 19 August 2008
Copyright © 2009 John Wiley & Sons, Ltd.
Accepted 30 March 2009
(dependent) latent variables of interest on explanatory (independent) latent variables. Motivations for developing this methodology include: (i) Latent variables usually give more comprehensive interpretation than observed variables. For instance, it is more interesting to study the effect of the latent variable ‘blood pressure’ (which is formed by both systolic and diastolic blood pressures) on heart disease, rather than just the effect of systolic blood pressure on heart disease. (ii) As the number of latent variables is much less than the number of originally observed variables, using structural equation with latent variables has advantages over the ordinary regression model with originally observed variables. Through the user friendly software ([2] and others), SEMs have been extensively applied to behavioral, educational, and social–psychological sciences in the past years. Recently, this method has been applied to biomedical research in studying the characteristics of various diseases through models with latent variables. See below for an illustrative real example about diabetic kidney disease, and [3] for more detailed discussion on how SEMs are relevant to biomedical research. Here, we develop a Bayesian semiparametric approach for analyzing SEMs with covariates, and mixed continuous and unordered categorical variables.

The reason for considering mixed continuous and unordered categorical variable is based on various epidemiological and family-based cohort analyses, which demonstrated that diseases with complex traits such as diabetes have both genetic and environmental determinants [4, 5]. For example, continuous phenotype variables, such as glucose and lipid as well as unordered categorical genotype variables such as tumor necrosis factor beta, plasminogen activator inhibitor-1 SNP1, plasminogen activator inhibitor-1 SNP2, and the beta 1, 2, 3 adrenoreceptors, have been implicated in the early onset of the diabetic kidney disease [6–8]. In general, given the complexity of the pathways in relation to various diseases, it is desirable to develop an efficient SEM to investigate the effects of explanatory phenotype and genotype latent variables to the outcome latent variable (say diabetic kidney disease in the above example). As observed, phenotype variables are usually continuous, while observed genotype variables are unordered categorical (say with genotype carriers AA, Aa, and aa); it is necessary to develop an SEM with mixed continuous and unordered categorical variables [9].

The motivation of developing a Bayesian semiparametric approach is based on the major concern on the crucial normal assumption of the latent variables in the development of statistical methods for SEMs. Even for standard SEMs with continuous variables, violation of the above-mentioned normal assumption (it is impossible to check) would lead to bad effects on statistical inference [10, 11]. In the literature, development of robust methods for non-normal continuous data using the multivariate t-distribution has received much attention [12, 13]. However, the existing robust methods in [12, 13] cannot be applied to unordered categorical data. Moreover, as they are developed on the basis of the multivariate t-distribution, they are not effective in analyzing continuous data from skewed or mixture distributions. In this article, we propose a Bayesian semiparametric approach for analyzing SEMs with mixed continuous and unordered categorical variables with the explanatory latent variables modeled by an appropriate truncated Dirichlet process (DP). Bayesian semiparametric approaches based on the DP [14] have been widely applied to various statistical models and complex data, see [15–18]. In particular, Lee et al. [18] utilized an approximate truncation DP in developing Bayesian semiparametric SEMs with continuous variables. Although this approach works well for continuous variables, it is not suitable to model latent variables formed by unordered categorical variables, see Sections 2 and 3. Hence, we propose here a modified version of the truncated DP with appropriate priors for the explanatory latent variables in the proposed SEM.

In this paper, we also address the important issue about model selection. As far as we know, in the context of the Bayesian semiparametric SEMs, no model selection statistic has been developed.
Bayes factor [19] is a well-known Bayesian model selection statistic, and has been applied to many parametric SEMs, see [20] and the references therein. However, this statistic may not be a good choice for the current semiparametric model due to some computational difficulties. First, because of the complications induced by the semiparametric formulation of the model and the unordered categorical variables, the computational burden of the Bayes factor is extremely heavy. Moreover, path sampling [21], which is effective in computing Bayes factor for parametric SEMs [20], cannot give satisfactory result for semiparametric SEMs. Hence, we propose an attractive model selection statistic, say $L_{-\infty}$-measure, for the current semiparametric SEMs. As this statistic can be obtained from the available outputs in the estimation, its computational burden is very light.

We illustrate our methodology through an application relating to a study of diabetic kidney disease based on a data set obtained from type 2 diabetes patients involved in an applied genomic program conducted by the Institute of Diabetes, The Chinese University of Hong Kong. In this application, several observed phenotype and genotype variables were selected. Two observed continuous phenotype variables: logarithm urinary albumin creatinine ratio (ACR) and logarithm plasma creatinine (PCr) were used to form the outcome latent variable about diabetic kidney disease. On the basis of some preliminary data analysis and motivated by some medical findings, we selected some phenotype and genotype observed variables to form the explanatory latent variables that are expected to have substantial effects on diabetic kidney disease. In particular, Ravid et al. [5] identified an explanatory latent variable related to lipid control via continuous phenotype observed variables: non-high-density lipoprotein cholesterol (non-HDL), lower-density lipoprotein cholesterol (LDL), and logarithm plasma triglyceride (TG). Similarly, based on Wang et al. [22], continuous phenotype observed variables: fasting plasma glucose (FPG) and glycated hemoglobin (HbA1c) were selected to form an explanatory latent variable related to glycemic control. Furthermore, based on [7, 8], we are also interested in the following groups of correlated genotype variables that are unordered categorical with genotype carries AA, Aa, and aa: \{tumor necrosis factor beta (LTA2), plasminogen activator inhibitor1 SNP1 (PAI11), plasminogen activator inhibitor1 SNP2 (PAI12)\}, and \{beta2 adrenergic receptor SNP1 (ADR$\beta_{21}$), beta2 adrenergic receptor SNP2 (ADR$\beta_{22}$), beta3 adrenergic receptor (ADR$\beta_{3}$)\}. The phenotype and genotype observed variables are grouped into their corresponding latent variables through the measurement equation. Then, the outcome latent variable of diabetic kidney disease is regressed on the explanatory latent variables, such as lipid control, glycemic control, etc., through the structural equation. As the distributions of some observed variables are highly non-normal (even after the log transformation), we expect that the latent variables are not normal and the usual distributional assumption would be violated. Hence, the proposed Bayesian semiparametric SEM with continuous and unordered categorical variables is required in analyzing this data set.

The contributions of this article include (i) the formulation of a novel Bayesian semiparametric SEM with mixed continuous and unordered categorical variables, by an appropriately modified truncated Dirichlet process (MTDP) in modeling the latent variables, (ii) a novel implementation of the blocked Gibbs sampler with derivation of some new conditional distributions in drawing observations from the posterior distribution for estimation and other statistical inferences of the proposed model, (iii) a more computationally efficient model selection statistic for Bayesian semiparametric SEMs with mixed variables, and (iv) a novel application of the developed methodologies to the kidney disease study.

This article is organized as follows. The Bayesian semiparametric SEM with covariates, and mixed continuous and unordered categorical variables is presented in Section 2. Methods for estimation and model selection are discussed in Section 3. Results of a simulation study are
reported in Section 4 to illustrate the performance of the proposed approach. Section 5 applies the methodologies to a real data set about diabetic kidney disease. A discussion is given in Section 6, and some technical details are presented in the Appendix.

2. MODEL DESCRIPTION

2.1. SEMs with mixed continuous and unordered categorical variable

Consider an \( p \times 1 \) observed random vector \( \mathbf{y}_i = (\mathbf{x}_i^T, \mathbf{u}_i^T)^T = (x_{i1}, \ldots, x_{ir}, u_{ir+1}, \ldots, u_{ip})^T \), in which \( \mathbf{x}_i = (x_{i1}, \ldots, x_{ir})^T \) contains continuous variables, and \( \mathbf{u}_i = (u_{ir+1}, \ldots, u_{ip})^T \) contains unordered categorical variables. The observed continuous variables in \( \mathbf{x}_i \) are related with a vector of correlated latent variables \( \mathbf{\omega}^c_i \) and a vector of covariates \( \mathbf{c}_i \) through the following confirmatory factor analysis model:

\[
x_{ij} = \mathbf{a}_j^T \mathbf{c}_i + \lambda_j \mathbf{\omega}^c_i + \epsilon_{ij}, \quad j = 1, \ldots, r
\]

For \( j = r+1, \ldots, p \), we assume that \( u_{ij} \) is an observed multinomial variable that takes one and only one of the possible values \( \{0, \ldots, K_j - 1\} \). For example, in a genetic application, \( u_{ij} \) denotes the \( K = 3 \) possible genotype carriers, say \{AA, Aa, aa\}. Similar to the literature [9, 23], \( u_{ij} \) is modeled through the \((K_j - 1) \times 1\) latent vector \( \mathbf{v}_{ij} = (v_{ij1}, \ldots, v_{ij(K_j-1)})^T \) as follows:

\[
u_{ij} = \begin{cases} 0 & \text{if } \max(\mathbf{v}_{ij}) < 0 \\ k & \text{if } \max(\mathbf{v}_{ij}) = v_{ijk} > 0, \quad k = 1, \ldots, K_j - 1 \end{cases}
\]

where \( \max(\mathbf{v}_{ij}) \) is the largest element in \( \mathbf{v}_{ij} \). To achieve our goal in relating the observed unordered categorical variables to latent variables, and with covariates \( \mathbf{c}_i \) and measurement errors being taken into account, we propose the following model for \( \mathbf{v}_{ij} \):

\[
\mathbf{v}_{ij} = \mathbf{1}_{K_j-1} \mathbf{a}_j^T \mathbf{c}_i + \mathbf{1}_{K_j-1} \lambda_j \mathbf{\omega}^c_i + \mathbf{\epsilon}_{ij}, \quad j = r+1, \ldots, p
\]

where \( \mathbf{1}_{K_j-1}(\times 1) \) is a vector of one’s, \( \lambda_j \) is a vector of parameters, \( \mathbf{\omega}^c_i \) is a vector of the latent variables that are formed by multinomial observed variables, and \( \mathbf{\epsilon}_{ij} \) is a vector of residual errors, which is distributed as \( \mathbf{N}(\mathbf{0}, \mathbf{\Psi}_j) \) and is independent of \( \mathbf{\omega}^c_i \). In this article, inspired by the suggestion of Song et al. [9] in identifying the unordered categorical variables, \( \mathbf{\Psi}_j \) is fixed at \( \mathbf{I}_{K_j-1} \), an identity matrix of order \( K_j - 1 \).

Let

\[
\mathbf{\Phi}_{ik} = \begin{cases} \{\mathbf{v}_{ij} : \max(\mathbf{v}_{ij}) < 0\}, & k = 0 \\ \{\mathbf{v}_{ij} : \max(\mathbf{v}_{ij}) = v_{ijk} > 0\}, & k = 1, \ldots, K_j - 1 \end{cases}
\]

Then, for \( i = 1, \ldots, n \), and \( j = r+1, \ldots, p \),

\[
P_r(\mathbf{u}_{ij} = k | \mathbf{a}_j, \lambda_j, \mathbf{\omega}_j^+) = P_r(\mathbf{v}_{ij} \in \mathbf{\Phi}_{ik} | \mathbf{a}_j, \lambda_j, \mathbf{\omega}_j^+),
\]

\[
= \int_{\mathbf{\Phi}_{ik}} \mathbf{\phi}_{K_j-1}(\mathbf{v}_{ij} | \mathbf{1}_{K_j-1}(\mathbf{a}_j^T \mathbf{c}_i + \lambda_j \mathbf{\omega}^c_i), \mathbf{1}_{K_j-1}) \, d\mathbf{v}_{ij}
\]

\[
= \begin{cases} \{1 - [1 - \Phi(\mathbf{a}_j^T \mathbf{c}_i + \lambda_j \mathbf{\omega}^c_i)]^{K_j-1}\} / (K_j - 1) & \text{if } k \neq 0 \\ \{1 - \Phi(\mathbf{a}_j^T \mathbf{c}_i + \lambda_j \mathbf{\omega}^c_i)^{K_j-1}\} & \text{if } k = 0 \end{cases}
\]

where $\phi_{K_j-1}(\cdot)$ is the density function of the $(K_j-1)$-variate normal distribution, and $\Phi(\cdot)$ is the univariate standard normal distribution function. A similar approach of modeling discrete outcomes via unobserved normal continuous variables has also been used in latent variable models for analyzing binary and ordered categorical variables [20].

Let $y_i^* = (x_i^T, v_{i,r+1}^T, \ldots, v_{i,p}^T)^T$ be the vector of observed and unobserved continuous variables corresponding to $y_i$. It follows from (1) and (3) that the measurement equation of the proposed SEM is

$$
y_i^* = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ir} \\ v_{i,r+1} \\ \vdots \\ v_{ip} \end{bmatrix} = \begin{bmatrix} a_1^T \\ \vdots \\ a_r^T \\ 1_{K_r+1-1}a_{r+1}^T \\ \vdots \\ 1_{K_p-1}a_p^T \end{bmatrix} \begin{bmatrix} \lambda_1^T \\ \vdots \\ \lambda_r^T \\ 0 \\ 1_{K_r+1-1}\lambda_{r+1}^T \\ \vdots \\ 1_{K_p-1}\lambda_p^T \end{bmatrix} \begin{bmatrix} \epsilon_i^T \\ \vdots \\ \epsilon_i^{r+1} \\ \epsilon_{i,r+1} \\ \vdots \\ \epsilon_i^p \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ir} \\ \epsilon_{i,r+1} \\ \vdots \\ \epsilon_{ip} \end{bmatrix} \tag{6}
$$

where $\epsilon_{i1}, \ldots, \epsilon_{ir}, \epsilon_{i,r+1}, \ldots, \epsilon_{ip}$ are the residual errors that are mutually independent and independent of $\omega_i^*$ and $\omega_i^+$. This equation is used to assess the interrelationships among the latent variables in $\omega_i^*$ and $\omega_i^+$ and the observed variables in $x_i$ and $u_i$ (indirectly through $v_{ij}$, $j = r+1, \ldots, p$). Equation (6) can be rewritten as

$$
y_i^* = CA\epsilon_i + CA\omega_i + \epsilon_i \tag{7}
$$

where $A = (a_1, \ldots, a_p)^T$, $\omega_i = (\omega_i^*, \omega_i^+)^T$ is an $m \times 1$ vector of latent variables, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{ir}, \epsilon_{i,r+1}, \ldots, \epsilon_{ip})^T$.

$$
C = \begin{bmatrix} I_r & 0 & \cdots & 0 \\ 0 & 1_{K_r+1-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{K_p-1} \end{bmatrix}
$$

and

$$
A^T = \begin{bmatrix} \lambda_1 & \cdots & \lambda_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \lambda_{r+1} & \cdots & \lambda_p \end{bmatrix}
$$

Let $\omega_i = (\eta_i^T, \xi_i^T)^T$ be a partition of $\omega_i$ with a vector $\eta_i$ ($m_1 \times 1$) of outcome latent variables and a vector $\xi_i$ ($m_2 \times 1$) of explanatory latent variables. Note that $\eta_i$ or $\xi_i$ may involve latent variables in $\omega_i^*$ and/or $\omega_i^+$. The proposed SEM is defined by (7) together with the following structural equation:

$$
\eta_i = Bf_i + \Pi \eta_i + \Gamma \xi_i + \xi_i \tag{8}
$$

where $B$, $\Pi$, and $\Gamma$ are $m_1 \times q_2$, $m_1 \times m_1$, and $m_1 \times m_2$ matrices of unknown parameters; $f_i$ ($q_2 \times 1$) is a vector of covariates and $\zeta_i$ is a vector of errors that is independent of $\xi_i$ and $\epsilon_i$. This regression-type structural equation is used to assess the effects of (i) some possible covariates, (ii) variables in $\eta_i$ on other variables in $\eta_i$, and (iii) variables in $\xi_i$ on $\eta_i$. Following Kleinman and Ibrahim [15] in
Bayesian semiparametric analysis of random effect models, we assume that $\tilde{z}_i$ are distributed as $\mathcal{N}(0, \Psi_{\tilde{z}})$ and $\tilde{z}_i$ are distributed as $\mathcal{N}(0, \Psi_{\tilde{z}})$, respectively. Hence, the covariance matrix of $\tilde{z}_i$, $\Psi_{\tilde{z}}$, is a diagonal block matrix, with the diagonal blocks equal to $\Psi_{\tilde{z}}$ and $\Psi_{\tilde{z}} = \mathbf{I}_{K_j - 1}$, respectively. For convenience, we also assumed that $\Psi_{\tilde{z}}$ and $\Psi_{\tilde{z}}$ are diagonal; and $\Pi_0 = \mathbf{I}_{m_1} - \Pi$ is non-singular and its determinant is independent of the elements $i$.

2.2. Semiparametric hierarchical modeling via DP

In classical structural equation modeling, it is assumed that $\tilde{z}_i$ follows a multivariate normal distribution. This assumption may not be true for many latent constructs in practice. Even for analyzing continuous data, applications of the standard parametric SEM with violation to this assumption would lead to biased statistical results [10, 11]. For unordered categorical data, the problem could be more serious. Hence, it is desirable to consider a semiparametric approach to relax the basic normal assumption of $\tilde{z}_i$. Moreover, such an approach allows us to better model the true distribution of $\tilde{z}_i$.

Let $\theta_k$ and $\theta_\psi$ be the vectors of unknown parameters, respectively, involved in (7) and (8), and let $\theta = (\theta_{k}, \theta_{\psi})$. The Bayesian semiparametric model can be reformulated as follows: For $i = 1, \ldots, n$,

$$
(y_i^0 | n_i, \tilde{z}_i, \theta) \sim \mathcal{N}(\mathbf{C} \mathbf{\alpha}_i + \mathbf{C} \mathbf{\omega}_i, \Psi_{\psi})
$$

$$
(n_i | \tilde{z}_i, \theta) \sim \mathcal{N}(\mathbf{\Pi}_0^{-1}(\mathbf{B} \mathbf{\tilde{z}} + \mathbf{\Gamma} \tilde{z}_i), \mathbf{\Pi}_0^{-1}\Psi_{\psi}\mathbf{\Pi}_0^{-T})
$$

(9)

$$
\theta \sim p(\theta), \quad (\tilde{z}_i | P) \sim P \text{ and } P \sim \mathcal{P}(\cdot)
$$

where $p(\theta)$ is the prior distribution of $\theta$ and $\mathcal{P}(\cdot)$ is a random probability measure. An important issue is to specify the prior of $P$. Among various easy-to-construct random probability measures that approximate the general DP; Ishwaran and James [24] proposed the truncated DP with stick-breaking priors. They also showed that the associative blocked Gibbs sampler generally overcome some limitations of the Pólya urn scheme that arise from the effect of marginalizing over the general DP. Specifically, the truncated DP with stick-breaking priors is defined as

$$
[\tilde{z}_1, \ldots, \tilde{z}_n | P] \sim P, \quad P \sim \mathcal{P}(\cdot) = \sum_{k=1}^{G} \pi_k \delta_{Z_k}(\cdot)
$$

(10)

where $\delta_{Z_k}(\cdot)$ denotes a discrete probability measure concentrated at $Z_k$, $\pi = \{\pi_k : k = 1, \ldots, G\}$ is a random vector chosen to be independent of $Z_k$, and

$$
\pi_1 = V_1, \quad \pi_k = (1 - V_1) \ldots (1 - V_{k-1}) V_k, \quad k = 2, \ldots, G - 1
$$

$$
\pi_G = (1 - V_1) \ldots (1 - V_{G-1})
$$

(11)

such that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{G} \pi_k = 1$. Following Sethuraman [25], one may take $V_k$ as independent $\operatorname{Beta}(1, x)$, where $x > 0$ is a hyper-parameter. Moreover, $Z_k$ are assumed to be i.i.d. $\mathcal{N}(\mathbf{\mu}_z, \Sigma_z)$. While the above truncated DP with stick-breaking priors works well for Bayesian semiparametric SEMs with continuous variables [18], it cannot be directly applied to our model with unordered categorical variable because the location of $P$ will affect the probability of the unordered categorical variables, see below.
Let \( \Lambda_\eta \) and \( \Lambda_v \) be the submatrices of \( \Lambda \) that, respectively, contain the \( m_1 \times 1 \) and \( m_2 \times 1 \) rows of \( \Lambda \) corresponding to \( \eta \) and \( \xi \). It follows from (7) and (8) that

\[
p(y_{i}^{*} | \theta, P) = \int p(y_{i}^{*} | \eta_{i}, \xi_{i}, \theta) p(\eta_{i} | \xi_{i}, \theta) P(d\xi_{i}) d\eta_{i}, \tag{12}\]

and its mean vector is given by

\[
\mu_{y_{i}^{*}} | \theta, P = \text{CAG} + \text{CAG}^{1}_{\theta} \text{F} + (\text{CAG}^{1}_{\theta} \text{G} + \text{CAG}^{\theta}) \int \xi_{i} P(d\xi_{i})
\]

Based on (4) and (5) in modeling the unordered categorical variable \( v_{ij} \) through \( v_{ij}^{+} \), we require \( \mu_{P} = \int \xi_{i} P(d\xi_{i}) = 0 \) with probability one. That is, the random probability measure \( P \) in relation to semiparametric modeling of \( \xi \) is not centered at zero, and we require to construct an appropriate \( P \) whose location is centered at zero. Note that truncated DP with stick-breaking priors defined by (10) and (11) is defined on an arbitrary \( P \). Hence, this approach is not appropriate because the location of the random probability measure is not centered at zero. To solve this problem, a modified approach is given below.

Based on the reasoning given in Appendix A, we propose the following MTDP for modeling \( \xi_{i} \)

\[
\xi_{il} | P_{l} \sim P_{l}, \quad l = 1, \ldots, m_2 \tag{13}
\]

\[
P_{l} \sim \frac{1}{2} \sum_{k=1}^{G} \pi_{kl} \delta_{Z_{kl}}(\cdot) + \frac{1}{2} \sum_{k=G+1}^{2G} \pi_{kl} \delta_{Z_{kl}}(\cdot)
\]

in which \( Z_{l} = (Z_{1l}, \ldots, Z_{Gl})^{T} \) and \( Z_{l}^{+} = (Z_{G+1,l}, \ldots, Z_{2Gl})^{T} \) are independent with the following distributions:

\[
Z_{kl} \sim \text{N}(\mu_{zl}, \sigma_{zl}) I\{Z_{kl} \leq 0\}, \quad k = 1, \ldots, G \tag{14}
\]

\[
Z_{kl} \sim \text{N}(\mu_{zl}, \sigma_{zl}) I\{Z_{kl} > 0\}, \quad k = G + 1, \ldots, 2G
\]

in which \( I\{\cdot\} \) denotes the indicator function. The random weights \( \pi^{-}_{l} = (\pi_{1l}, \ldots, \pi_{Gl})^{T} \) and \( \pi^{+}_{l} = (\pi_{G+1,l}, \ldots, \pi_{2Gl})^{T} \) are independent, and both are constructed by slight modification of the stick-breaking priors in (11). That is, for \( \pi^{-} \)

\[
\pi_{1l} = V_{1l}, \quad \pi_{kl} = (1 - V_{1l}) \ldots (1 - V_{k-1,l}) V_{kl}, \quad k = 2, \ldots, G - 1 \tag{15}
\]

\[
\pi_{Gl} = (1 - V_{1l}) \ldots (1 - V_{G-1,l})
\]

where \( V_{kl} \) are independently distributed as Beta(1, \( \alpha_{l} M_{l}^{-} \)) with \( \alpha_{l} > 0 \) and \( M_{l}^{-} = 1 - \Phi(\mu_{zl} / \sqrt{\sigma_{zl}}) \); for \( \pi^{+} \)

\[
\pi_{G+1,l} = U_{1l}, \quad \pi_{G+k,l} = (1 - U_{1l}) \ldots (1 - U_{k-1,l}) U_{kl}, \quad k = 2, \ldots, G - 1 \tag{16}
\]

\[
\pi_{2G,l} = (1 - U_{1l}) \ldots (1 - U_{G-1,l})
\]

where \( U_{kl} \) are independent Beta(1, \( \alpha_{l} M_{l}^{+} \)) random variables with \( M_{l}^{+} = \Phi(\mu_{zl} / \sqrt{\sigma_{zl}}) \).

Theoretical justification of using above MTDP and the stick-breaking priors in modeling \( \xi \) is given in Appendix A. The empirical performance is revealed by the simulation study in Section 4.
3. PARAMETERS ESTIMATION AND MODEL SELECTION

3.1. Estimation by the blocked Gibbs sampler

Let \( Z_l = (Z_{1l}, \ldots, Z_{2G,l})^T \), \( Z = (Z_1, \ldots, Z_{m_2}) \), \( \pi_l = (\pi_{1l}, \ldots, \pi_{2G,l})^T \), and \( \pi = (\pi_1, \ldots, \pi_{m_2}) \). The key character for the blocked Gibbs sampler is to recast the model completely by introducing the cluster variables. Let \( L = \{L^1, \ldots, L^{m_2}\} \) with \( L_l^i = (L_{1l}^i, \ldots, L_{l}^i)^T \), such that \( \xi_{il} = Z_{l_{il}} \), then (9) and (13) can be rewritten as

\[
[y_i^s | \eta_i, Z, L, 0] \sim N(CA\eta_i + CA\omega_i, \Psi_e) \\
[\eta_i | Z, L, 0] \sim N((\Pi_0^{-1}Bf_i + \Pi_0^{-1}T^T \xi^l, \Pi_0^{-1} \Psi_e \Pi_0^{-T})) \\
[L_l^i | \pi] \sim N(1/2 \sum k \pi_{kl} \delta_k (\cdot) + 1/2 \sum \pi_{kl} \delta_k (\cdot), l = 1, \ldots, m_2 \\
(\pi, Z) \sim p(\pi) p(Z) \quad \text{and} \quad 0 \sim p(0)
\]

where \( \omega_i = (\eta_i^T, \xi_i^T)^T \) with \( \xi_{il} = Z_{l_{il}} \), \( p(\pi) \) and \( p(Z) \) are prior densities of \( \pi \) and \( Z \), respectively. Let \( \Lambda_i = (A, A), \Lambda_i^e = (B, B, \Gamma), \Psi_{ex} = \text{diag}(\psi_{ex1}, \ldots, \psi_{exr}) \), and \( \Psi_e = \text{diag}(\psi_{e1}, \ldots, \psi_{em}) \). Let \( \lambda_{ik} \) be the \( k \)th row of \( \Lambda_i \), and \( \lambda_{ij} \) be the \( j \)th row of \( \Lambda_i^e \). For convenience, we assume that \( p(A_i, \Psi_e) = \prod p(\lambda_{ik} | \psi_{exk}) p(\psi_{exk}) \prod p(\lambda_{ij} | \psi_{eij}) p(\psi_{eij}) \). The following conjugate-type prior distributions for the components of \( \theta \) are used:

\[
p(\lambda_{ik}) \sim N(\lambda_{0k}, \mathbf{H}_{0k}), \quad k = r + 1, \ldots, p \\
p(\lambda_{ik} | \psi_{exk}) \sim N(\lambda_{0k}, \psi_{exk} \mathbf{H}_{0k}), \quad p(\psi_{exk}^{-1}) \sim \text{Gamma}(\alpha_{0k}, \beta_{0k}), \quad k = 1, \ldots, r \\
p(\lambda_{ij} | \psi_{eij}) \sim N(\lambda_{0j}, \psi_{eij} \mathbf{H}_{0j}), \quad p(\psi_{eij}^{-1}) \sim \text{Gamma}(\alpha_{0j}, \beta_{0j}), \quad j = 1, \ldots, m_1
\]

where \( \lambda_{0k}, \alpha_{0k}, \beta_{0k}, \lambda_{0j}, \alpha_{0j}, \beta_{0j} \), and the positive definite matrices \( \mathbf{H}_{0k} \) and \( \mathbf{H}_{0j} \) are the hyper-parameters. Let \( \beta \) be the parameter vector containing the unknown parameters in \( \mu_z = (\mu_{z1}, \ldots, \mu_{zm})^T, \Sigma_z = \text{diag}(\sigma_{z1}, \ldots, \sigma_{zm}) \), and \( \alpha = (\alpha_1, \ldots, \alpha_{m})^T \), which are involved in the non-parametric component. The following conjugate prior distributions for these parameters are used:

\[
\gamma_{zl} \sim \text{Gamma}(\gamma_{1l}, \gamma_{2l}), \quad \gamma_{zl} \sim \text{Gamma}(\gamma_{1l}, \gamma_{2l})
\]

where \( \sigma_M, \gamma_{1l}, \gamma_{2l}, \tau_1, \tau_2 \) are hyper-parameters.

To accommodate missing data, let \( x_i = (x_{io}, x_{im}) \) and \( u_i = (u_{io}, u_{im}) \) where \( x_{io} \) and \( u_{io} \) are observed, while \( x_{im} \) and \( u_{im} \) are missing. For brevity, we assume that the missing data are missing at random (MAR). Let \( X_o = \{x_{io}, \ldots, x_{no}\}, X_m = \{x_{1m}, \ldots, x_{nm}\}, U_o = \{u_{io}, \ldots, u_{no}\}, U_m = \{u_{im}, \ldots, u_{nm}\} \), \( \Omega = \{o_1, \ldots, o_n\} \), and \( \Omega_l = \{\eta_1, \ldots, \eta_m\} \). Further, let \( V_o \) and \( V_m \) be the collection of \( v_{ij} \) that correspond to the \( U_o \) and \( U_m \), respectively. Data augmentation is used to cope with the posterior analysis in relation to the complicated \( p(0, \beta | X_o, U_o) \). Specifically, the observed data \( \{X_o, U_o\} \) are augmented with the missing quantities \( \{X_m, V_o, V_m, \Omega_1, \pi, Z, L\} \) in the posterior analysis. A sequence of random observations will be generated from the joint posterior distribution \( p(X_m, V_o, V_m, \Omega_1, \pi, Z, L | \beta, X_o, U_o) \) by the blocked Gibbs sampler [24] coupled with the Metropolis–Hastings (MH) algorithm [26, 27]. Some nice features of the blocked Gibbs sampler

Copyright © 2009 John Wiley & Sons, Ltd.
DOI: 10.1002/sim
have been shown [24], for example, it avoids marginalizing over the prior and hence allowing the prior to be directly involved in the Gibbs sampler. As the proposed semiparametric method for SEMs with continuous and unordered categorical variables as described in Section 2 does not exist in the literature, it is necessary to derive some new conditional distributions for developing of the block Gibbs sampler that involves innovative implementation of the MH algorithm. Some related details are given in Appendix B.

3.2. Model selection

Model selection is an important issue in SEMs. A common statistic for Bayesian model selection is the Bayes factor or the marginal likelihood [19]. Path sampling [21] has been applied in computing Bayes factor for many parametric SEMs [20]. However, as the computation of path sampling involves an approximation of a huge dimensional integral over $\Omega$ and other terms, and the distribution of $\xi_j$ in the proposed Bayesian semiparametric SEMs with mixed continuous and unordered categorical variables is more complex (see (13)) than that in parametric SEMs, path sampling cannot give accurate results for the current model. Moreover, the more sophisticated [28, 29] algorithms would also encounter serious difficulties and computational intensity. Inspired by the predictive approach for model selection and its applications to various data structures [30–33], we consider the $L_v$-measure that is based on the future values of a replicate experiment and the idea that good models should give good predictions close to what has been observed.

Let $\mathbf{y}_o^{rep} = (y_{1,o}^{rep}, \ldots, y_{n,o}^{rep})$ be a future response vector of a replicate experiment with the same sampling density as $p(\mathbf{y}_o | \mathbf{0})$ with $\mathbf{Y}_o = \{\mathbf{X}_o, \mathbf{U}_o\}$. In this article, we first consider the following:

$$L_1(\mathbf{Y}_o, \mathbf{B}, \delta) = E[\text{tr}(\mathbf{y}_o^{rep} - \mathbf{B})^T(\mathbf{y}_o^{rep} - \mathbf{B})] + \delta \text{tr}(\mathbf{y}_o - \mathbf{B})$$

(20)

where the expectation is taken with respect to the posterior predictive distribution of $p(\mathbf{y}_o^{rep} | \mathbf{Y}_o)$. Note that this statistic can be regarded as a multivariate version of the model assessment criterion considered by Ibrahim et al. [32] for analyzing generalized linear models and survival models. By selecting $\mathbf{B}$ as the minimizer of (20), it can be shown that for some $0 \leq v < 1$, we have the following multivariate version of $L_v$-measure:

$$L_v(\mathbf{Y}_o) = \sum_{i=1}^{n} \text{tr}[\text{Cov}(\mathbf{y}_{i,o}^{rep} | \mathbf{Y}_o)] + v \sum_{i=1}^{n} \text{tr}[\{E(\mathbf{y}_{i,o}^{rep} | \mathbf{Y}_o) - \mathbf{y}_{i,o}\} \{E(\mathbf{y}_{i,o}^{rep} | \mathbf{Y}_o) - \mathbf{y}_{i,o}\}^T]$$

(21)

where $v = \delta / (\delta + 1)$. This statistic, which we call the $L_v$-measure, is a sum of two components. The first component related to the variability of the prediction, while the second component measures how close its predictions to the observed data. Clearly, a small value of the $L_v$-measure indicates that the corresponding model gives prediction close to the observed values and the variability of the prediction is low as well. Hence, the model with the smallest $L_v$-measure is selected from a collection of competing models. In the context of a linear model, we note that $v = 0.5$ is a desirable and justifiable choice for model selection [32]. Thus, this will be used in our empirical illustrations.

In applying the $L_v$-measure for model assessment and model selection for the proposed semiparametric SEM, we have to evaluate $\text{Cov}(\mathbf{y}_{i,o}^{rep} | \mathbf{Y}_o)$ and $E(\mathbf{y}_{i,o}^{rep} | \mathbf{Y}_o)$ that involve intractable multiple integrals. These integrals can be approximated by the corresponding ergodic averages of Markov chain Monte Carlo (MCMC) samples, $\{\mathbf{X}_m(t), \mathbf{Y}_m(t), \mathbf{Y}_{o,m}^{(t)}, \mathbf{0}_{m}^{(t)}, \mathbf{x}_{m}^{(t)}, \mathbf{Z}_{m}^{(t)}, \mathbf{L}_{m}^{(t)}, \mathbf{p}_{m}^{(t)}: t = 1, \ldots, T\}$, which are generated by the blocked Gibbs sampler coupled with the MH algorithm in the estimation. Hence, the computation burden of $L_v(\mathbf{Y}_o)$ is very light.
4. A SIMULATION STUDY

We conducted a simulation study to roughly reveal the performances of the proposed Bayesian semi-parametric approach for estimation and model selection, and to compare them with the parametric approach. We consider a situation with six continuous variables, and three unordered categorical variables, which take one and only one values in \{0, 1, 2\}. The measurement equation of the true model is given by (see (7))

\[
x_{ij} = a_jc_i + \mathbf{k}_j^T \mathbf{w}_i^* + \epsilon_{ij}, \quad j = 1, \ldots, 6
\]

\[
u_{ijl} = a_jc_i + \mathbf{k}_j^T \mathbf{w}_i^* + \epsilon_{ij}, \quad l = 1, 2, \quad j = 7, 8, 9
\]

Covariates \(c_i\) are generated from the standard normal distribution. The specifications of \(A\) and \(\Lambda\) in (7) and their true parameter values are given by:

\[
A^T = \begin{bmatrix}
0.7 & 0.7 & 0.7 & 0.7 & 0.7 & 0.7 & 0.7
\end{bmatrix}
\]

\[
\Lambda^T = \begin{bmatrix}
1^* & 0.8 & 0.8 & 0^* & 0^* & 0^* & 0^*
\end{bmatrix}
\]

in which the elements with an asterisk are treated as fixed for identifying the model. For \(j = 1, \ldots, 6\), the true value of \(\psi_{eij}\) is 1.0. Let \(\mathbf{w}_i = (\eta_i, \xi_{i1}, \xi_{i2})^T\), then the structural equation of the true model is

\[
\eta_i = b_f i + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \zeta_i
\]

where the covariates \(f_i\) are sampled from a Bernoulli distribution that takes 1 with probability 0.7 and 0 with probability 0.3. The true values of the parameters involved in (23) are given by \(b = 1.0, (\gamma_1, \gamma_2) = (0.5, 0.7)\), and \(\psi_{e} = 1.0\).

Simulation study with various non-normal distributions of \(\xi_i\) has been conducted. As the results are similar, we only present details obtained under the following two true distributions for \(\xi_i\): (I) \(\xi_i\) are drawn from the following mixture of normal distributions, \(0.5N(\mu_1, \Phi_1) + 0.5N(\mu_2, \Phi_2)\), where \(\mu_1 = (-2.0, 1.5)^T\), \(\mu_2 = (2.0, -1.5)^T\), and \(\Phi_1 = \Phi_2 = I_2\); and (II) \(\xi_{ij}\) were independently drawn from \(\sum_{j=0}^{7} N(3((2/3)^j - 1) + 1.919, (2/3)^{2j})/8\) (see [33]) for \(j = 1, 2\). Hence, the true distributions of \(\xi_i\) under (I) and (II) are bimodal and skewed, respectively. Based on these settings, random samples \(\{y_{il}, i = 1, \ldots, n\}\) with \(n = 800\) were generated according to (2), (7), and (8).

In the analysis, \(G = 50\) is taken in the MTDP (see (13)). Prior inputs for the hyper-parameters in the prior distributions relating to the parameters involved in the non-parametric component (see (19)) are: \(\sigma_M = 100, \xi_1 = \xi_2 = 0.01\), and \(\tau_1 = \tau_2 = 2.0\); while prior inputs in the prior distributions corresponding to the parameters involved in the parametric component (see (18)) are: \(\alpha_{0k} = \alpha_{0j} = 9, \beta_{0k} = \beta_{0j} = 8, \mathbf{H}_{0k}\) and \(\mathbf{H}_{0j}\) are diagonal matrices with the diagonal elements 1.0, and elements in \(\{k_{0k}, k_{0j}\}\) are equal to the true values. A few test runs revealed that the blocked Gibbs sampler converged in less than 2000 iterations. To be conservative, we collect 2000 observations after 3000 ‘burn-ins’ in computing the absolute bias (BIAS), and the root mean squares (RMS) of the estimates and the true values in 100 replications. Results obtained by using the MTDP with the stick-breaking priors (see (13)–(16)), and a parametric approach based on the
Table I. Performance of the semiparametric and parametric Bayesian estimates.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\xi_1 \sim$ MTDP (I)</th>
<th>$\xi_1 \sim$ MTDP (II)</th>
<th>$\xi_1 \sim$ Normal (I)</th>
<th>$\xi_1 \sim$ Normal (II)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIAS</td>
<td>RMS</td>
<td>BIAS</td>
<td>RMS</td>
</tr>
<tr>
<td>$a_1 = 0.7$</td>
<td>0.008</td>
<td>0.056</td>
<td>-0.008</td>
<td>0.058</td>
</tr>
<tr>
<td>$a_2 = 0.7$</td>
<td>0.001</td>
<td>0.044</td>
<td>0.003</td>
<td>0.046</td>
</tr>
<tr>
<td>$a_3 = 0.7$</td>
<td>0.006</td>
<td>0.048</td>
<td>-0.005</td>
<td>0.049</td>
</tr>
<tr>
<td>$a_4 = 0.7$</td>
<td>-0.008</td>
<td>0.085</td>
<td>0.001</td>
<td>0.042</td>
</tr>
<tr>
<td>$a_5 = 0.7$</td>
<td>-0.004</td>
<td>0.075</td>
<td>-0.002</td>
<td>0.038</td>
</tr>
<tr>
<td>$a_6 = 0.7$</td>
<td>-0.007</td>
<td>0.076</td>
<td>-0.003</td>
<td>0.041</td>
</tr>
<tr>
<td>$a_7 = 0.7$</td>
<td>-0.019</td>
<td>0.105</td>
<td>0.009</td>
<td>0.073</td>
</tr>
<tr>
<td>$a_8 = 0.7$</td>
<td>0.005</td>
<td>0.090</td>
<td>0.014</td>
<td>0.068</td>
</tr>
<tr>
<td>$a_9 = 0.7$</td>
<td>0.011</td>
<td>0.093</td>
<td>0.003</td>
<td>0.060</td>
</tr>
<tr>
<td>$\xi_{12} = 0.8$</td>
<td>0.002</td>
<td>0.031</td>
<td>0.004</td>
<td>0.039</td>
</tr>
<tr>
<td>$\xi_{13} = 0.8$</td>
<td>0.003</td>
<td>0.035</td>
<td>0.007</td>
<td>0.036</td>
</tr>
<tr>
<td>$\xi_{14} = 0.8$</td>
<td>0.004</td>
<td>0.019</td>
<td>0.009</td>
<td>0.056</td>
</tr>
<tr>
<td>$\xi_{16} = 0.8$</td>
<td>0.002</td>
<td>0.021</td>
<td>0.012</td>
<td>0.056</td>
</tr>
<tr>
<td>$\xi_{18} = 0.8$</td>
<td>0.029</td>
<td>0.102</td>
<td>0.057</td>
<td>0.139</td>
</tr>
<tr>
<td>$\xi_{31} = 0.8$</td>
<td>0.025</td>
<td>0.093</td>
<td>0.030</td>
<td>0.131</td>
</tr>
<tr>
<td>$\psi_{c1} = 1.0$</td>
<td>-0.003</td>
<td>0.069</td>
<td>0.000</td>
<td>0.072</td>
</tr>
<tr>
<td>$\psi_{c2} = 1.0$</td>
<td>0.002</td>
<td>0.064</td>
<td>-0.020</td>
<td>0.069</td>
</tr>
<tr>
<td>$\psi_{c3} = 1.0$</td>
<td>-0.009</td>
<td>0.063</td>
<td>0.003</td>
<td>0.066</td>
</tr>
<tr>
<td>$\psi_{c4} = 1.0$</td>
<td>-0.009</td>
<td>0.077</td>
<td>0.021</td>
<td>0.085</td>
</tr>
<tr>
<td>$\psi_{c5} = 1.0$</td>
<td>0.008</td>
<td>0.065</td>
<td>0.002</td>
<td>0.061</td>
</tr>
<tr>
<td>$\psi_{c6} = 1.0$</td>
<td>-0.004</td>
<td>0.065</td>
<td>0.004</td>
<td>0.066</td>
</tr>
<tr>
<td>$\alpha_1 = 1.0$</td>
<td>0.001</td>
<td>0.054</td>
<td>-0.000</td>
<td>0.072</td>
</tr>
<tr>
<td>$\gamma_1 = 0.5$</td>
<td>0.070</td>
<td>0.090</td>
<td>0.015</td>
<td>0.059</td>
</tr>
<tr>
<td>$\gamma_2 = 0.7$</td>
<td>0.051</td>
<td>0.119</td>
<td>0.044</td>
<td>0.141</td>
</tr>
<tr>
<td>$\psi_0 = 1.0$</td>
<td>-0.068</td>
<td>0.128</td>
<td>0.001</td>
<td>0.090</td>
</tr>
</tbody>
</table>

The above model with measurement equation (22) and structural equation (23).

$M_1$: The above model with measurement equation (22) and structural equation (23).

$M_2$: Same measurement equation in (22), and $\eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \xi_i$.

$M_3$: $x_{ij} = \lambda_j^T \omega^T_i + \epsilon_{ij}, \ j = 1, \ldots, 6$. $v_{ijl} = \lambda_j^T \omega_{ij}^T + \epsilon_{ij}, \ l = 1, 2, j = 7, 8, 9$, with the structural equation $\eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \xi_i$. 

The above model with measurement equation (22) and structural equation (23).
Figure 1. True densities (solid lines) and estimated densities via MTDP in 100 replications.

Table II. Mean and standard deviation (SD) of $L_{0.5}$-measures in 100 replications.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_1$</td>
<td>10930</td>
<td>229</td>
</tr>
<tr>
<td>$M_2$</td>
<td>11570</td>
<td>311</td>
</tr>
<tr>
<td>$M_3$</td>
<td>11500</td>
<td>341</td>
</tr>
<tr>
<td>(II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_1$</td>
<td>9470</td>
<td>183</td>
</tr>
<tr>
<td>$M_2$</td>
<td>9490</td>
<td>197</td>
</tr>
<tr>
<td>$M_3$</td>
<td>9810</td>
<td>175</td>
</tr>
</tbody>
</table>

For each of the 100 replications, observations obtained through the blocked Gibbs sampler are used to compute $L_{0.5}$-measures (see (21)) that correspond to $M_1$, $M_2$, and $M_3$. The $L_{0.5}$-measure selects the true model if its value corresponding to $M_1$ is the smallest. Under (I), with a bimodal distribution of $\xi_i$, it correctly selected $M_1$ in 84 replications, and incorrectly selected $M_2$ and $M_3$ in the other 7 and 9 replications, respectively. Under (II) with a skewed distribution of $\xi_i$, it correctly selected $M_1$ in all replications. To give more information, the mean and standard deviations of the $L_{0.5}$-measures obtained from the 100 replications are presented in Table II. We see that under [I] and [II] the mean of the $L_{0.5}$ measures under the correct model is smallest.

We have conducted additional simulation studies with smaller sample sizes, and obtained the expected result that larger sample sizes give better performances. However, we have to admit
that our simulation is limited in scope, because (i) it just considered a single SEM with one
set of population values and (ii) it only focused on two non-normal distributions of \( \xi_i \). In order
to draw more definite conclusions, we need to conduct a more comprehensive simulation study.
As the main emphasis of this paper is not on simulation, such a comprehensive simulation study
is not considered here. Owing to the complexity of the model and the large number of parameters
involved, for most simulation studies in the literature of SEMs, their Monte Carlo sizes are 100
replications. Of course, increasing the Monte Carlo size can reduce the Monte Carlo error. However,
it requires more computing effort and gives no effect in improving the generalizability for drawing
definite conclusions.

5. A REAL EXAMPLE: DIABETIC NEPHROPATHY STUDY

We illustrate the proposed methods through analysis of a real data set about the study of diabetic
kidney disease related to type 2 diabetes patients as mentioned in the ‘Introduction’. Basically,
we are interested in incorporating multiple genetic and phenotypic pathways to study diabetic
kidney disease. Data are obtained from high-risk diabetes patients who underwent a comprehensive
assessment of complications based on the European Diabetes protocol. The outcome variable of
diabetic kidney disease was assessed as a latent variable reflected by two observed continuous
phenotype variables: \( \log(\text{ACR}), \log(\text{PCr}) \). Based on the justification as given in the ‘Introduction’,
we considered the following groups of continuous phenotype observed variables: \{non-HDL, LDL, 
\log(TG)\} and \{\text{FGP, HbA1c}\}, and the following genotype observed variables that are unordered
categorical: \{LTA2, PAI11, PAI12, ADR \}. The sample size of the data set is 353. There are some missing data that were treated as MAR. To roughly unify the scale,
the continuous variables were standardized using the fully observed data points. The genotype
variables LTA2, PAI11, PAI12, ADR\_21, ADR\_22, and ADR\_3 are coded with \{AA, Aa, aa\}; their
frequencies are, respectively, equal to \{78, 177, 95\}, \{61, 178, 112\}, \{72, 173, 106\}, \{127, 164, 48\},
\{266, 79, 0\}, and \{307, 43, 0\}.

Let \( y = (\text{ACR}, \text{PCr}, \text{HbA1c}, \text{FGP}, \text{non-HDL}, \text{LDL}, \text{TG}, \text{LTA2}, \text{PAI11}, \text{PAI12}, \text{ADR}\_21, \text{ADR}\_22, \text{ADR}\_3)^T \)
be the vector of the observed variables. Based on the objective of this example
and the medical motivation given in the ‘Introduction’, it is natural to group (i) \{ACR, PCr\} to
an outcome latent variable that can be interpreted as ‘diabetic kidney disease, \( \eta \)’; (ii) \{HbA1c,
\text{FGP} and \{non-HDL, LDL, TG\} to two explanatory phenotype latent variables that can be,
respectively, interpreted as ‘glycemic control, \( \xi_1 \)’ and ‘lipid control \( \xi_2 \)’; (iii) \{LTA2, PAI11,
PAI12\} and \{ADR\_21, ADR\_22, ADR\_3\} to two explanatory genotype latent variables that can be,
respectively, interpreted as ‘plasminogen activator inhibitor, \( \xi_3 \)’ and ‘\( \beta \)-adrenergic receptor,
\( \xi_4 \)’. Hence, the following loading matrix \( \Lambda \) in the measurement equation (see (7)) with \( \omega_i = 
(\eta_i, \xi_{i1}, \xi_{i1}, \xi_{i3}, \xi_{i4})^T \) is considered:

\[
\Lambda^T = \begin{bmatrix}
1^* & \lambda_{21} & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 1^* & \lambda_{42} & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^*
\end{bmatrix}
\]
Although other structure of \( A \) could be used, we consider the above non-overlapped structure for clear interpretations of the latent variables.

To illustrate the \( L_v \)-measure in model selection, we consider the following competing models with different fixed covariates:

\[
M_1: y_{ik} = a_k c_{i1} + a_k c_{i2} + \lambda_k^T \omega_i + \epsilon_{ik}, \quad \eta_i = b_1 z_1 + b_2 z_2 + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i3} + \gamma_4 \xi_{i4} + \xi_i \\
M_2: y_{ik} = a_k c_{i1} + \lambda_k^T \omega_i + \epsilon_{ik}, \quad \eta_i = b_1 z_1 + b_2 z_2 + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i3} + \gamma_4 \xi_{i4} + \xi_i \\
M_3: y_{ik} = a_k c_{i2} + \lambda_k^T \omega_i + \epsilon_{ik}, \quad \eta_i = b_1 z_1 + b_2 z_2 + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i3} + \gamma_4 \xi_{i4} + \xi_i \\
M_4: y_{ik} = a_k c_{i1} + a_k c_{i2} + \lambda_k^T \omega_i + \epsilon_{ik}, \quad \eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i3} + \gamma_4 \xi_{i4} + \xi_i
\]

where \( c_{i1} \) and \( c_{i2} \) are continuous fixed covariates related to age-onset and duration of disease, and \( z_{i1} \) and \( z_{i2} \) are binary variables related to smoking and alcohol.

The proposed Bayesian semiparametric approach with \( G = 50 \) in the MTDP (see (13)) was applied to obtain the parameter estimates. The following prior inputs for the hyper-parameters in the conjugate prior distributions (see (18) and (19)) were used: \( \alpha_{0k} = 15, \beta_{0k} = (\alpha_{0k} - 1) \psi_{e_{ik}}, \lambda_{0k} = \lambda_{e_{ik}}, h_{0k} = I_7, \phi_{0} = 15, \phi_0 = (\phi_0 - 1) \psi_{e_{ik}}, \lambda_{z_{00}} = \lambda_{z}, h_{z_{00}} = I_6, \tau_1 = \tau_2 = 2.0, \gamma_1 = \gamma_2 = 0.01, \sigma_M = 10000, \)

where \( \hat{\theta} \) denotes the Bayesian estimates of \( \theta \) obtained from analysis a ‘control-group’ sample of diabetes patients. Results were obtained from 10 000 observations sampled by the blocked Gibbs sampler after 10 000 ‘burn-in’ iterations.

The \( L_{0.5} \)-measure corresponding to \( M_1, M_2, M_3, \) and \( M_4 \) are equal to 59 270, 58 822, 59 206, and 59 245, respectively. Hence, \( M_2 \) is selected. To reveal the convergence of the blocked Gibbs sampler, the plot of the ‘estimated potential scale reduction (EPSR)’ values against the iteration numbers in the analysis of \( M_2 \) is displayed in Figure 2, which indicates that the algorithm converges after several thousands iterations. The estimated densities based on 10 000 simulated observations of \( \xi_1, \xi_2, \xi_3, \) and \( \xi_4 \) are used to provide some idea about the distributions of the latent variables. These plots are presented in Figure 3, and they indicate that the distributions of \( \xi_1, \xi_2, \xi_3, \) and \( \xi_4 \) are not normal. A path diagram to illustrate the relationships among observed and latent variables is presented in Figure 4. Parameter estimates, together with the 5 and 95 per cent quantiles are

![Figure 2. EPSR values against the number of iteration in the analysis of kidney disease data under \( M_2 \).](image-url)
Figure 3. Estimated densities of $\xi_1$, $\xi_2$, $\xi_3$, and $\xi_4$.

reported in Table III. We note that most of the 5 and 95 per cent quantile ranges are reasonably short. The estimates of the coefficients $\hat{a}_{kj}$ corresponding to the fixed covariate age-onset and the factor loading estimates $\hat{\lambda}_{kj}$ in the measurement equation can be interpreted according to a standard confirmatory factor analysis model; the estimated structural equation is

$$\eta_i = -0.023z_1 + 0.031z_2 - 0.122\xi_1 + 0.553\xi_2 + 0.726\xi_3 + 0.563\xi_4$$

This structural equation can be interpreted according to a standard regression model. It seems that the effects of ‘smoking’ and ‘alcohol’ are minor, but the other phenotype and genotype latent variables have substantial effects on diabetic kidney disease. To save space, other straightforward interpretations are not discussed.

To assess the sensitivity of the results to the prior inputs of the hyper-parameters in prior distributions corresponding to the parametric component, the data set was reanalyzed with some perturbations of the above ad hoc prior inputs, for example, $H_{0k} = 10 \times I$ and $H_{0j} = 10 \times I$. We obtained close Bayesian estimates and very similar estimated densities of the latent variables as given above.

To reveal the differences between a semiparametric approach and a parametric approach, the data set has been reanalyzed by using a parametric approach under the assumption that $\xi_i$ is distributed as $N(0, \Phi)$. The prior distribution of $\Phi$ is taken as $IW(\hat{\Phi}, 20)$, where $IW(\cdot, \cdot)$ is an inverted Wishart distribution and $\hat{\Phi}$ is the Bayesian estimate of $\Phi$ obtained from the analysis of a ‘control-group’ sample of diabetes patients. The prior distributions of other parameters are the same as before. For comparison, the obtained estimates are also presented in Table III. We observe that some of the parameter estimates, in particular, those associated with unordered categorical variables, such as $\hat{\lambda}_{94}, \hat{\lambda}_{12.5}, \hat{\gamma}_3$, and $\hat{\gamma}_4$, are quite different from the estimates obtained from the semiparametric approach. Given the fact that the distributions of $\xi_1, \ldots, \xi_4$ are not normal.
Figure 4. The path diagram of the selected model for analyzing the real data set.

(see Figure 3) and the results are obtained from the simulation study reported in Section 4, we expect that the semiparametric solution is better.

The computer program for analyzing this data set is written in C language; the code is available in web site: http://www.sta.cuhk.edu.hk/sylee/BSSEM/code.zip.

6. CONCLUDING REMARKS

In this article, we introduce a Bayesian semiparametric SEM for analyzing inter-relationships of observed and latent variables with mixed continuous and unordered categorical data. To relax the crucial assumption on the normal distribution of the latent variables, we propose a modified version of the truncated DP (MTDP) and stick-breaking priors to model the explanatory latent variables. An efficient blocked Gibbs sampler is implemented to obtain the estimates of the unknown parameters and the latent variables. The statistic $L_v$-measure is established for model selection. A key advantage of this statistic is that it uses the available outputs in the estimation, and hence its computational burden is very light. It is clear from a theoretical and methodological point of
view that the Bayesian semiparametric approach is better than the traditional parametric approach in the sense that the normality assumption of the explanatory latent variables is relaxed. Moreover, results obtained from the simulation study and the analysis of a real medical data set also indicate that Bayesian semiparametric approach has substantial advantages. In medical studies, as many phenotype and genotype latent variables that are formed by non-normal observed continuous and unordered categorical variables are not normal; the proposed methodologies are useful in handling these kind of problems.

There are some limitations of the current approach that suggest certain future research topics. First, as individual element $\xi_i|y$ in $\xi_i$ was modeled by a specific $P_i$ via univariate random variables $Z_{kl}$ (see (13) and (14)), the covariance matrix of $\xi_i$ is restricted to be diagonal. One more general
approach to relax this restriction is to model the whole $\xi_i$. However, it is not straightforward to construct an appropriate $P$ with a multivariate random vector $Z_k$ such that $P$ is centered at zero. Moreover, the development of the blocked Gibbs vector $Z_k$ such that $P$ is centered at zero. Second, the developed methodologies can neither be applied to hierarchic data nor to longitudinal data. Multilevel SEMs [34, 35] and longitudinal SEMs [36] are required to handle such complicated data. As the amount of latent variables in those more complex SEMs is rather large, we expect to encounter some theoretical and computational difficulties in the development of the Bayesian semiparametric models and the algorithm for achieving the solution. Third, because the current approach can only handle missing data that are MAR, generalization of treating non-ignorable missing data in the observed variables and/or covariates is necessary.

APPENDIX A

To construct an appropriate random probability measure, we consider a finite non-null measure $v$ on the space of real numbers $R$ and the set containing all the subsets of $(R, \mathcal{B})$, where $R$ is a real-valued space and $\mathcal{B}$ is a Borel $\sigma$-field on $R$. Further, we denote $\mathcal{D}_v$ be the Dirichlet process (DP) over all the distributions defined on $(R, \mathcal{B})$ with the parameter $v$. Let $v_+(B) = v(B \cap (0, \infty)) + 0.5v(0)$, and $v_-(B) = v(B \cap (-\infty, 0)) + 0.5v[0]$, for $B \in \mathcal{B}$. Now, we consider

$$P = \frac{1}{2}P_+ + \frac{1}{2}P_- \quad (A1)$$

where $P_+$ and $P_-$ are taken independently from $\mathcal{D}_{v_+}$ and $\mathcal{D}_{v_-}$, respectively. Thus, the median of $P$ is equal to 0. This defines a random probability measure $\mathcal{D}_v^\pi$ on the set of all distribution functions $P$ on $R$ with median equal to 0. Similar random probability measure has been used in Bayesian semiparametric estimation of the median in a distribution [37] and median regression modeling [38]. It can be shown by similar reasoning as in Sethuraman [25] that $\mathcal{D}_v$ generates $P_+ = \sum_{k=1}^{\infty} \pi_k \delta_{z_k}$ with $\pi_k = \prod_{j=1}^{k-1} (1 - v_j^+) v_k^+$, where $v_k^+ \sim \text{Beta}(1, v_+(R))$ and $z_k \sim v_+/v_+(R)$. Similarly, $\mathcal{D}_v$ generates $P_- = \sum_{k=1}^{\infty} p_k \delta_{w_k}$ with $p_k = \prod_{j=1}^{k-1} (1 - v_j^-) v_k^-$, where $v_k^- \sim \text{Beta}(1, v_-(R))$, and $w_k \sim v_-/v_-(R)$. Note that $(z_k)_{k=1}^{\infty}$, $(w_k)_{k=1}^{\infty}$, $(\pi_k)_{k=1}^{\infty}$, and $(p_k)_{k=1}^{\infty}$ are independent. For $v(\cdot) = \alpha H_0(\cdot)$ with some distribution $H_0(\cdot)$, we have $v_k^- \sim \text{Beta}(1, \alpha M^-)$ with $M^- = H_0((\infty, 0))$ and $w_k \sim H_0^-$, where $H_0^-$ is a truncated distribution function at $(-\infty, 0)$. We also have similar results for $v_k^+$ and $\pi_k$. Consequently, $\mathcal{D}_v^\pi$ generates

$$P = \frac{1}{2}P_+ + \frac{1}{2}P_- = \frac{1}{2} \sum_{k=1}^{\infty} \pi_k \delta_{z_k} + \frac{1}{2} \sum_{k=1}^{\infty} p_k \delta_{w_k} \quad (A2)$$

This gives (13), the MTDP for modeling $\xi_i$.

APPENDIX B

The full conditional distribution $p(X_m|V_o, V_m, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o)$ is

$$p(X_m|V_o, V_m, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o) = \prod_{i=1}^{n} p(x_{i,m}|\omega_i, 0)$$
and

\[ p(x_{i,m} | \omega_i, 0) \overset{D}{=} N(a_{i,m}c_I + \Lambda_{i,m}\omega_i, \Psi_{\omega i}) \]

where \( a_{i,m} \) is an \( r_i \times 1 \) subvector of \( A \) with rows corresponding to \( x_{i,m} \), \( \Lambda_{i,m} \) is an \( r_i \times q \) submatrix of \( A \) with rows correspond to \( x_{i,m} \), and \( \Psi_{\omega i} \) is the submatrix of \( \Psi_{\omega x} \) corresponding to \( x_{i,m} \).

The full conditional distribution \( p(V_o | X_m, V_m, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o) \) is given by

\[
p(V_o | X_m, V_m, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o) = \prod_{i=1}^{n} \prod_{j=r+1}^{p} p(v_{ij} | u_{ij}, \omega_i, 0)
\]

and

\[
p(v_{ij} | u_{ij}, \omega_i, 0) \propto N(1_{K_j-1}a_j^Tc_I + 1_{K_j-1}\lambda_j^T\omega_i, 1)I\{(−∞, 0]\}
\]

It can be shown that if \( u_{ij} = k = 0 \),

\[
p(v_{ij} | u_{ij}, \omega_i, 0) = \prod_{i=1}^{K_j-1} p(v_{ij} | u_{ij}, \omega_i, 0) \propto \prod_{i=1}^{K_j-1} N(a_j^Tc_I + \lambda_j^T\omega_i, 1)I\{(−∞, \max_{l \neq k} v_{ijl} \lor 0]\}
\]

and if \( u_{ij} = k \neq 0 \),

\[
p(v_{ij} | u_{ij}, v_{ijl} \neq l, \omega_i, 0) \propto \begin{cases} N(a_j^Tc_I + \lambda_j^T\omega_i, 1)I\{(−∞, \max_{l \neq k} v_{ijl} \lor 0]\} & \text{if } l \neq k \\ N(a_j^Tc_I + \lambda_j^T\omega_i, 1)I\{(\max_{l \neq k} v_{ijl} \lor 0, +∞]\} & \text{if } l = k \end{cases}
\]

where \( a \lor b \) means the maximum of \( a \) and \( b \).

The full conditional distribution \( p(V_m | X_m, V_o, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o) \) is given by

\[
p(V_m | X_m, V_o, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o) = \prod_{i=1}^{n} \prod_{j=r+1}^{p} p(v_{ij,m} | \omega_i, 0)
\]

and

\[
p(v_{ij,m} | \omega_i, 0) \overset{D}{=} N(a_j^Tc_I + \lambda_j^T\omega_i, 1)
\]

where \( v_{ij,m} \) is the \( k \)th component of \( v_{ij,m} \).

The full conditional distribution \( p(\Omega_1 | X_m, V_o, V_m, 0, \pi, Z, L, \beta, X_o, U_o) \) is given by

\[
p(\Omega_1 | X_m, V_o, V_m, 0, \pi, Z, L, \beta, X_o, U_o) = \prod_{i=1}^{n} p(\eta_i | \xi_i, \gamma_i, 0)
\]

and

\[
[\eta_i | \xi_i, \gamma_i, 0] \sim N_{m_1}(\mu_{\eta_i}, \Sigma_{\eta}), \quad i = 1, \ldots, n
\]

where \( \Sigma_{\eta} = (\Lambda_{\eta}^T\Psi^{-1}_M\Lambda_{\eta} + \Pi_0^T\Psi^{-1}_x\Pi_0)^{-1} \) and \( \mu_{\eta_i} = \Sigma_{\eta}(\Lambda_{\eta}^T\Psi^{-1}_M(\vec{y}_i - \Lambda c_i - \Lambda_{\xi_i}^{\vec{y}_i}) + \Pi_0^T\Psi^{-1}_x(\vec{f}_i + \Gamma_{\xi_i}^{\vec{y}_i})) \).

The full conditional distribution \( p(0 | X_m, V_o, V_m, \Omega_1, 0, \pi, Z, L, \beta, X_o, U_o) \) is further partitioned into the following parts. Let \( w_i = (c_i^T, \omega_i^T)^T, g_i = (f_i^T, \omega_i^T)^T, G = (g_1, \ldots, g_n)^T, W = (w_1, \ldots, w_n)^T, \) ...
X = (x_1, ..., x_n)^T, and Y* = (y_1*, ..., y_n*)^T. We denote x(k) the kth column of X, and ω_{1(j)} the jth column of Ω_1. It can be shown that under the prior (18), the full conditional distribution \( p(\theta | \beta, L, Z, \pi, \Omega_1, V_m, V_o, X_m, U_o, X_o) \) is given as follows:

\[
p(A_v, \Psi_v | \Omega, Y^*) \sim \left\{ \prod_{k=1}^{r} N(m_{ekk}, \psi_{ekk}, \Sigma_{ekk}) \right\} \cdot \text{Gamma}(n/2 + \omega_{0k}, \beta_{ek}) \times \prod_{k=r+1}^{p} N(m_{ek}, \Sigma_{ek})
\]

\[
p(A_v, \Psi_v | \Omega, Y^*) \sim \prod_{j=1}^{m_1} N(m_{\zeta_{j}}, \psi_{\zeta_{j}}, \Sigma_{\zeta_{j}}) \cdot \text{Gamma}(n/2 + \omega_{0j}, \beta_{\zeta_{j}})
\]

in which

\[
m_{ekk} = \Sigma_{ekk}[H_{e0k}^{-1} \lambda_{e0k} + W^T x(k)], \quad \Sigma_{ekk} = [H^{-1} + W^T W]^{-1}
\]

\[
m_{ek} = \Sigma_{ek k}[H_{e0k}^{-1} \lambda_{e0k} + (K_k - 1)W^T \tilde{v}(k)], \quad \Sigma_{ek} = [H^{-1} + (K_k - 1)W^T W]^{-1}
\]

\[
\beta_{ek} = \beta_{e0k} + 2^{-1}[\lambda_{e0k}^H H_{e0k}^{-1} \lambda_{e0k} - m_{ekk}^{-1} m_{ekk} + x(k)x(k)^T]
\]

\[
m_{\zeta_{j}} = \Sigma_{\zeta_{j}}[H_{\zeta_{0j}}^{-1} \lambda_{\zeta_{0j}} + G^T \omega_{1(j)}], \quad \Sigma_{\zeta_{j}} = [G^T G + H_{\zeta_{0j}}^{-1}]^{-1}
\]

\[
\beta_{\zeta_{j}} = \beta_{\zeta_{0j}} + 2^{-1}[\lambda_{\zeta_{0j}}^H H_{\zeta_{0j}}^{-1} \lambda_{\zeta_{0j}} + m_{\zeta_{j}}^{-1} m_{\zeta_{j}} + \omega_{1(j)}^T \omega_{1(j)}]
\]

where \( \tilde{v}(k) = (\tilde{v}_1, ..., \tilde{v}_{nk})^T \).

To simulate \( (\pi, Z) \) from \( p(\pi, Z|X_m, V_o, V_m, \Omega_1, 0, L, \beta, X_o, U_o) \), we draw \( (\pi_l, Z_l) \) from \( p(\pi_l, Z_l|X_m, V_o, V_m, \Omega_l, 0, L, \beta, X_o, U_o) \), for \( l = 1, ..., m_2 \). To simulate \( (\pi_l, Z_l) \) from \( p(\pi_l, Z_l|X_m, V_o, V_m, \Omega_1, 0, L, \beta, X_o, U_o) \), let \( Z_{-l} = (Z_1, ..., Z_{l-1}, Z_{l+1}, ..., Z_{m_2}) \), \( \pi_{-l} = (\pi_1, ..., \pi_{l-1}, \pi_{l+1}, ..., \pi_{m_2}) \), and \( L_{-l} = (L^1, ..., L^{l-1}, L^{l+1}, ..., L^{m_2}) \). Let \( \{L^{l*_j}, j = 1, ..., n_l\} \) be the unique set of \( L^{l_j} \) values in \( L^l \) and \( n_j = \# \{i: L^l_i = L^{l*_j}\} \), the number of \( L^{l_j} \) equal to \( L^{l*_j} \), then

\[
p(\pi_l, Z_l|X_m, V_o, V_m, \Omega_1, 0, L, \beta, X_o, U_o)
\]

\[
\propto \prod_{j=1}^{n_l} \left\{ p(Z_{L^{l*_j}_j}|\mu_{z_l}, \sigma_{z_l}) \times \prod_{i:L^l_i = L^{l*_j}} p(y^{n_l}_i|\eta_i, 0, Z^{l*_j}, L_{-l}, Z_{-l}) \right\} \times p(Z_{-(-L^{l*_j})}, \mu_{z_l}, \sigma_{z_l}) p(L|\pi_l) p(\pi_l)
\]

where \( Z_{-(-L^{l*_j})} \) corresponds to those values in \( Z_l \) excluding \( Z_{L^{l*_j}_j} = \{Z_{L^{l*_j}_1}, ..., Z_{L^{l*_j}_{m_2}}\} \). Now, let \( \tilde{y}_l = (x_l^T, \tilde{v}_{l(r+1)}, ..., \tilde{v}_{lp})^T \) with \( \tilde{v}_{lj} = \sum_{k=1}^{K_{r+1}} v_{ijk} / (K_{j} - 1) \), \( j = r+1, ..., p \), and

\[
\Psi_{-M}^{-1} = \begin{bmatrix}
\Psi_{ex}^{-1} & 0 & \cdots & 0 \\
0 & (K_{r+1} - 1)I_{K_{r+1} - 1} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (K_{p} - 1)I_{K_{p} - 1}
\end{bmatrix}
\]
It can be shown that for $j = 1, \ldots, n_l$

$$Z_{L_j^l, l} \sim \begin{cases} 
N(\mu_{wjl}, \sigma_{wjl}) I\{(\neg \infty, 0]\} & \text{if } L_j^l \leq G \\
N(\mu_{wjl}, \sigma_{wjl}) I\{(0, \infty]\} & \text{if } L_j^l > G 
\end{cases}$$

where $\sigma_{wjl} = [\sigma_{\zeta l}^{-1} + n_j l(M_{\zeta l} \lambda_{\zeta l} + \gamma_{\zeta l}^{-1}(\gamma_l))]^{-1}$, and

$$\mu_{wjl} = \sigma_{wjl} \left\{ \gamma_{\zeta l} M^{-1} \sum_{l': l' = L_j^l} (\tilde{y}_{l} - A \delta_{l} - A \eta_b - \sum_{l' = 1, l' \neq l}^m \lambda_{l' l} \tilde{\xi}_{l'}) + \sigma_{\zeta l}^{-1} \mu_{l} \right\}$$

where $\lambda_{\zeta l}$ and $\gamma_l$ are the $l$th columns of $\Lambda_{\zeta}$ and $\Gamma$, respectively.

The components of $(Z_{(-L_j^l), l}|m_z, \Sigma_z)$ are i.i.d. with the following distribution:

$$Z_{L_j^l, l} \sim \begin{cases} 
N(\mu_{zjl}, \sigma_{zjl}) I\{(\neg \infty, 0]\}, \ L_j^l \leq G \text{ and } L_j^l \neq L_j^l, \ j = 1, \ldots, n_l \\
N(\mu_{zjl}, \sigma_{zjl}) I\{(0, \infty]\}, \ L_j^l > G + 1 \text{ and } L_j^l \neq L_j^l, \ j = 1, \ldots, n_l 
\end{cases}$$

Let $n_{kl}^- = \#(L_j^l = k, \ k \leq G)$, $n_{kl}^+ = \#(L_j^l = G + k, \ k \leq G)$, $a_{kl}^- = 1 + n_k^-$, $b_{kl}^- = M_l \chi_l + \sum_{j = k + 1}^G n_j^-$, $a_{kl}^+ = 1 + n_k^+$, and $b_{kl}^+ = M_l \chi_l + \sum_{j = k + 1}^G n_j^+$. The conditional distribution for $p(\pi|L, \alpha)$ can be obtained as follows:

$$\pi_{kl}^* = V_{kl}^*, \ \pi_{kl} = (1 - V_{kl}^*) (1 - V_{k-1,l}^*) V_{kl}^*, \ k = 2, \ldots, G - 1$$

$$\pi_{G+1,l}^* = U_{kl}^*, \ \pi_{G+k,l} = (1 - U_{kl}^*) (1 - U_{k-1,l}^*) U_{kl}^*, \ k = 2, \ldots, G - 1$$

where $V_{kl}^* \sim \text{Beta}(a_{kl}^+, b_{kl}^-)$, and $U_{kl}^* \sim \text{Beta}(a_{kl}^-, b_{kl}^+)$. To simulate $L$ from $p(L|X_m, V_o, V_m, \Omega_1, 0, \pi, Z, X_o, U_o)$, we draw $L^l$ from $p(L^l|X_m, V_o, V_m, \Omega_1, 0, \pi, Z, L_{l-1}, \beta, X_o, U_o), l = 1, \ldots, m_2$. It can be shown that the conditional distribution of $L^l$ is given by

$$[L^l|X_m, V_o, V_m, \Omega_1, 0, \pi, Z, L_{l-1}, \beta, X_o, U_o] \sim \sum_{k = 1}^{2G} \pi_{kl}^* \delta_k(\cdot)$$

where $\pi_{kl}^* \sim \frac{1}{2} \pi_k p(y_{l}^*|\eta_b, 0, Z_{kl}, Z_{l-1}) p(\eta_b|0, Z_{kl}, Z_{l-1}), \ k = 1, \ldots, 2G$, satisfying $\sum_{k = 1}^{2G} \pi_{kl}^* = 1.0$.

Note that the conditional distribution $p(\beta|X_m, V_o, V_m, \Omega_1, 0, \pi, Z, L, X_o, U_o)$ only depends on the $\pi$ and $Z$. To simulate $\beta$, we draw $p(\beta|\mu_z, \Sigma_z, \pi, Z)$ and $p(\mu_z, \Sigma_z|\alpha, \pi, Z)$. Moreover, $p(\alpha|\mu_z, \Sigma_z, \pi, Z) = \prod_{l = 1}^{m_2} p(\alpha_l|\mu_z, \Sigma_z, \pi) = \prod_{l = 1}^{m_2} p(\alpha_l|\mu_z, \Sigma_z, \pi)$, and

$$(z_l|\mu_z, \Sigma_z, \pi) \sim \text{Gamma}(2G + a_1 - 1, a_2 - M^{-1} \sum_{k = 1}^{G-1} \log(1 - V_{kl}^*) - M^+ \sum_{k = 1}^{G-1} \log(1 - U_{kl}^*)$$

Copyright © 2009 John Wiley & Sons, Ltd.


DOI: 10.1002/sim
It follows from (17) that \( p(\mathbf{\mu}, \mathbf{\Sigma}|\mathbf{z}, \mathbf{Z}, \pi) \) can be obtained as a product of the following \( p(\mu_{zl}, \sigma_{zl}|\mathbf{z}, \mathbf{Z}, \pi): \)

\[
p(\mu_{zl}, \sigma_{zl}|\mathbf{z}, \mathbf{Z}, \pi) \propto \left( \prod_{i=1}^{2G} p(Z_{il}|\mu_{zl}, \sigma_{zl}) p(\mu_{zl}) p(\sigma_{zl}) \right) \left( \prod_{i=1}^{G} p(V_{il}|\mu_{zl}, \sigma_{zl}) p(U_{il}|\mu_{zl}, \sigma_{zl}) \right) \tag{B1}
\]

The MH algorithm can be used to draw observations from this target distribution. Under \( \mu_{zl}=0 \) with \( M^{-}=M^{+}=0.5 \), (A3) does not involve \( \pi \), and the conditional distribution of \( \sigma_{zl} \) is given by

\[
[\sigma_{zl}^{-1}|\mathbf{Z}] \sim \text{Gamma} \left( \zeta_1 + G, \ z_2 + \frac{1}{2} \sum_{i=1}^{2G} Z_{il}^2 \right)
\]

Finally, in the derivation of the full conditional distributions relating to \( \mathbf{\mu}, \mathbf{\Sigma}, \mathbf{z} \), it can be shown that

\[
\log p(\mu_{zl}|\sigma_{zl}, \mathbf{z}, \pi) \propto (G-1)[\log M_{zl}^{-} + \log M_{zl}^{+}] + \zeta_1 [M_{zl}^{-} \log \pi_{Gl} + M_{zl}^{+} \log p_{Gl}]
\]

\[
\times \log p(\mu_{zl}|\sigma_{zl}) + \sum_{i=1}^{2G} \log p(Z_{il}|\mu_{zl}, \sigma_{zl})
\]

\[
\log p(\sigma_{zl}^{-1}|\mu_{zl}, \mathbf{z}, \pi) \propto (G-1)[\log M_{zl}^{-} + \log M_{zl}^{+}] + \zeta_1 [M_{zl}^{-} \log \pi_{Gl} + M_{zl}^{+} \log p_{Gl}]
\]

\[
\times \log p(\sigma_{zl}^{-1}|\zeta_1, \zeta_2) + \sum_{i=1}^{2G} \log p(Z_{il}|\mu_{zl}, \sigma_{zl})
\]

\[
- \left. \frac{d^2 \log p(\mu_{zl}|\sigma_{zl}, \mathbf{Z}, \pi)}{d \mu_{zl}^2} \right|_{\mu_{zl}=0} = \sigma_{zl}^{-1}[2G + 4(G-1)/\pi] + \sigma_{zl}^{-1}
\]

In the implementation of the algorithm, we draw a candidate \( \mu_{zl}^{*} \) from the proposal distribution \( \mathcal{N}(\cdot|\mu_{zl}, r \sigma_{zl}) \) at the current values \( \mu_{zl} \), where \( r \) is a constant for adjusting the acceptance rate, accept \( \mu_{zl}^{*} \) with the probability

\[
\min \left\{ 1, \ \frac{p(\mu_{zl}^{*}|\sigma_{zl}, \mathbf{Z}, \pi)}{p(\mu_{zl}|\sigma_{zl}, \mathbf{Z}, \pi)} \right\}
\]

Independent sampler are used to sample \( \sigma_{zl}^{-1} \). Specifically, draw \( \sigma_{zl}^{-1*} \) from \( q(\cdot|\mu_{zl}, \mathbf{Z}) \) with

\[
q(\cdot|\mu_{zl}, \mathbf{Z}) = \text{Gamma} \left( G + \zeta_1, \ z_2 + \frac{1}{2} \sum_{i=1}^{2G} (Z_{il} - \mu_{zl})^2 \right)
\]

accept \( \sigma_{zl}^{-1*} \) with the probability

\[
\min \left\{ 1, \ \frac{p(\sigma_{zl}^{-1*}|\mu_{zl}, \mathbf{Z}, \pi) q(\sigma_{zl}^{-1}|\mu_{zl}, \mathbf{Z})}{p(\sigma_{zl}^{-1}|\mu_{zl}, \mathbf{Z}, \pi) q(\sigma_{zl}^{-1*}|\mu_{zl}, \mathbf{Z})} \right\}
\]
ACKNOWLEDGEMENTS

The research described herein was fully supported by grants (404507, 450607) from the Research Grants Council of the Hong Kong Special Administration Region. The authors are thankful to three reviewers for their constructive comments for improving the paper, and to Prof. J. C. N. Chan and her colleagues, Department of Medicine and Therapeutics, The Chinese University of Hong Kong, for allowing the use of their diabetic nephropathy data.

REFERENCES


34. Lee SY, Song XY. Maximum likelihood analysis of a general latent variable model with hierarchy mixed data. 
   *Biometrics* 2004; **60**:624–636.


