

Bayesian Semiparametric Structural Equation Models with Latent Variables

Mingan Yang and David B. Dunson¹

Biostatistics Branch, MD A3-03 National Institute of Environmental Health Sciences

Research Triangle Park, NC 27709, USA

¹*dunson1@niehs.nih.gov*

Structural equation models (SEMs) with latent variables are widely useful for sparse covariance structure modeling and for inferring relationships among latent variables. Bayesian SEMs are appealing in allowing for the incorporation of prior information and in providing exact posterior distributions of unknowns, including the latent variables. In this article, we propose a broad class of semiparametric Bayesian SEMs, which allow mixed categorical and continuous manifest variables while also allowing the latent variables to have unknown distributions. In order to include typical identifiability restrictions on the latent variable distributions, we rely on centered Dirichlet process (CDP) and CDP mixture (CDPM) models. The CDP will induce a latent class model with an unknown number of classes, while the CDPM will induce a latent trait model with unknown densities for the latent traits. A simple and efficient Markov chain Monte Carlo algorithm is developed for posterior computation, and the methods are illustrated using simulated examples, and several applications.

Key Words: Dirichlet process; Factor analysis; Latent class; Latent trait; Mixture model; Nonparametric Bayes; Parameter expansion.

1. Introduction

In the social sciences and increasingly in other application areas, it is routine to collect multivariate data, with the individual measurements having a variety of scales (continuous,

count, categorical). Often, these measurements are collected specifically with the goal of studying relationships among latent variables, such as life event-induced anxiety, that can only be measured indirectly through multiple manifest variables. In such settings, structural equation models (SEMs) provide a valuable tool for obtaining insight into the relationships between different latent variables and between latent and observed variables (Bollen, 1989). In addition, SEMs provide a flexible class of multivariate models for describing covariance structures in multivariate data.

In recent years, there has been increased interest in Bayesian SEMs due in part to advances in posterior computation that now allow Bayesian approaches to be implemented routinely in complex settings involving multilevel structures, missing data, censoring and other challenges. Using Markov chain Monte Carlo (MCMC) algorithms, one can obtain samples from the exact posterior distribution of all the unknowns, including the latent variables. These samples can be used to estimate exact posterior distributions, which provide a full probability characterization of uncertainty without needing to appeal to large sample assumptions. Asymptotic arguments may be difficult to justify for SEMs even in large samples, since the data can contain minimal information about certain parameters due to weak identifiability. In such settings, it is particularly important to include outside information and theory into the analysis, which can be accomplished within a Bayesian approach through an informative prior. For a recent review of Bayesian SEMs, refer to Palomo, Dunson and Bollen (2007).

Because Bayesian SEMs require a full likelihood specification, a concern is robustness to parametric assumptions, such as normality of the latent variables and measurement errors. Lee and Xia (2006) recently proposed robust maximum likelihood methods for nonlinear SEMs incorporating symmetric heavy-tailed distributions. Fahrmeir and Raach (2007) proposed an alternative semiparametric Bayesian approach, which characterizes the latent variables in a latent factor regression model using an additive model. This approach as-

sumes normally distributed latent variables, while allowing the mean of the latent variable distribution to vary flexibly with continuous predictors and spatial location. An alternative strategy to define flexible latent variable models is to use mixtures of parametric models. Jedidi, Jagpal and DeSarbo (1997) used a finite mixture of SEMs to allow heterogeneity across subgroups. Zhu and Lee (2001) considered Bayesian inference on finite mixtures of LISREL models. Fokoue and Titterington (2003) and Fokoue (2005) proposed mixtures of factor analysers, based on a finite mixture of normal factor models. Lubke and Muthén (2005) used factor mixture models to assess population heterogeneity. McLachlan, Bean and Jones (2007) proposed a robust extension of mixtures of factor analysers to allow heavy-tailed distributions within each component.

Instead of characterizing heterogeneity through finite mixtures of SEMs, our focus is on using Bayesian nonparametric methods to allow the latent variable distributions within an SEM to be unknown. There is a rich literature on the use of Dirichlet process (DP) priors (Ferguson, 1973, 1974) and DP mixtures (DPMs) to allow unknown latent variable distributions in hierarchical models. For example, Bush and MacEachern (1996), Kleinman and Ibrahim (1998), and Brown and Ibrahim (2003) used DPMs for random effects distributions. Ansari and Iyengar (2006) recently used DP components to define a semiparametric dynamic choice model. Dunson (2006) used dynamic mixtures of DPs to allow a latent variable distribution to change nonparametrically across groups. Burr and Doss (2005) used a conditional DP for the random effects distribution within a meta analysis application.

Unfortunately, direct application of Dirichlet processes and other priors for unknown distributions is problematic in general SEMs due to the need to incorporate constraints on the latent variable distributions for identifiability and interpretability. For example, it is standard practice to restrict the residual distributions on the latent variable level to have mean zero and variance one. Although this is straightforward when the latent variable distributions are normally distributed, it is quite challenging to incorporate mean and variance

constraints on unknown distributions. In fact, although there is a rich literature on incorporation of median and quantile constraints, until recently there were no approaches available for placing mean zero and variance one constraints on unknown latent variable distributions within a hierarchical model. To address this gap, two conceptually related centering approaches were independently developed by Dunson, Yang and Baird (2007) and Li, Müller and Lin (2007).

The focus of this article is on applying the centered DP (CDP) and CDP mixture (CDPM) models proposed by Dunson et al. (2007) to develop a general class of semiparametric Bayes SEMs that have unknown latent variable distributions. When a CDP is used for a latent variable distribution, one obtains a latent class model with infinitely many classes represented in the population, finitely many of which are represented in the current sample. When a CDPM is used, one instead treats the latent variables as continuous with an unknown smooth density, which can be skewed and multimodal. By centering the prior for this unknown density on the standard normal density, one utilizes Bayesian shrinkage to stabilize estimation of latent variable densities in small samples and weak identifiability situations. In practice, it is straightforward to define SEMs having both discrete and continuous latent variables. Posterior computation relies on a straightforward data augmentation parameter-expanded Gibbs sampling algorithm, which tends to be highly efficient.

Section 2 describes the class of SEMs to be considered. Section 3 describes the CDP and CDPM, discussing properties in the setting of SEMs. Section 4 describes the algorithm for posterior computation. Section 5 contains simulation examples, providing a proof of concept that non-normal latent variable densities can be estimated reliably in many cases. Section 6 contains several applications, and Section 7 discusses the results.

2. Semiparametric Bayes SEMs

SEMs provide a broad framework for modeling of multivariate data having mixed categorical

and continuous measurement scales. For subject i ($i = 1, \dots, n$), the observed data consist of $\mathbf{z}_i = (\mathbf{y}'_i, \mathbf{x}'_i)'$, where $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})'$ is a $p \times 1$ vector of outcome measurements and $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})'$ is a $q \times 1$ vector of predictor measurements. Following common practice, we assume that $y_{ij} = g_j(y_{ij}^*; \boldsymbol{\tau}_{y,j})$, for $j = 1, \dots, p$, and $x_{ij} = h_j(x_{ij}^*; \boldsymbol{\tau}_{x,j})$, for $j = 1, \dots, q$, where y_{ij} is linked to an underlying continuous variable y_{ij}^* through a link function g_j having parameters $\boldsymbol{\tau}_{y,j}$, and x_{ij} is linked to an underlying continuous variable x_{ij}^* through a link function h_j having parameters $\boldsymbol{\tau}_{x,j}$. Typically, for continuous measurements, identity links will be used, so that $y_{ij} = y_{ij}^*$ and $x_{ij} = x_{ij}^*$. In contrast, for categorical measurements, threshold links will be used mapping from \mathfrak{R} to $\{1, \dots, C\}$, with C the number of categories.

After relating the observations to underlying continuous variables, we specify the SEM in two components: (1) the measurement model, which relates the underlying variable to latent variables; and (2) the latent variable or structural model, which describes relationships among the latent variables. For the measurement model, we use a typical normal, linear form as in Bollen (1989):

$$\begin{aligned} \mathbf{y}_i^* &= \boldsymbol{\mu}_y + \boldsymbol{\Lambda}_y \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_{y,i}, & \boldsymbol{\epsilon}_{y,i} &\sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_y) \\ \mathbf{x}_i^* &= \boldsymbol{\mu}_x + \boldsymbol{\Lambda}_x \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{x,i}, & \boldsymbol{\epsilon}_{x,i} &\sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_x), \end{aligned} \quad (1)$$

where $\boldsymbol{\mu}_y, \boldsymbol{\mu}_x$ are intercepts, $\boldsymbol{\Lambda}_y$ is a $p \times r$ factor loadings matrix, $\boldsymbol{\eta}_i$ is a $r \times 1$ vector of latent response variables, $\boldsymbol{\epsilon}_{y,i}$ is a $p \times 1$ vector of idiosyncratic measurement errors, $\boldsymbol{\Sigma}_y$ is a $p \times p$ diagonal covariance matrix, $\boldsymbol{\Lambda}_x$ is a $q \times s$ factor loadings matrix, $\boldsymbol{\xi}_i$ is an $s \times 1$ vector of latent predictors, $\boldsymbol{\epsilon}_{x,i}$ is a $q \times 1$ vector of measurement errors, and $\boldsymbol{\Sigma}_x$ is a $q \times q$ diagonal covariance matrix.

To complete a specification of the SEM, we then choose a linear structural relations (LISREL) model:

$$\begin{aligned} \boldsymbol{\eta}_i &= \mathbf{B} \boldsymbol{\eta}_i + \boldsymbol{\Gamma} \boldsymbol{\xi}_i + \boldsymbol{\delta}_i, & \boldsymbol{\delta}_i &\sim F \\ \boldsymbol{\xi}_i &\sim G, \end{aligned} \quad (2)$$

where \mathbf{B} is an $r \times r$ matrix with zeros along the diagonal describing relationships among the different latent response variables, $\mathbf{\Gamma}$ is a $r \times s$ matrix describing relationships between $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$, and $\boldsymbol{\delta}_i$ is a residual. Our contribution relative to previous work on SEMs is to treat F and G as unknown discrete or continuous distributions using nonparametric Bayes methods. In particular, this is accomplished by letting $F \sim \mathcal{P}$ and $G \sim \mathcal{Q}$, where \mathcal{P} and \mathcal{Q} are priors with support on the space of distributions on \mathfrak{R}^r and \mathfrak{R}^s , respectively, subject to appropriate mean and/or variance constraints.

3. Centered Dirichlet Process Mixtures for Latent Variables

3.1 Dirichlet processes and Dirichlet process mixtures

In this section, we review some basic properties of the Dirichlet process (DP). First, suppose \mathcal{Q} corresponds to a $DP(\alpha G_0)$ prior, which is a Dirichlet process with precision parameter α and base distribution G_0 . For example, G_0 may correspond to a normal distribution. Then, any realization G from \mathcal{Q} will be discrete, implying that individuals will be grouped into clusters. This clustering will occur through the DP prediction rule, or Polya urn scheme, which was originally described by Blackwell and MacQueen (1973). Assuming that $\boldsymbol{\xi}_i \sim G$, with $G \sim DP(\alpha G_0)$, then the Polya urn scheme implies that $\boldsymbol{\xi}_1 \sim G_0$, $(\boldsymbol{\xi}_i | \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{i-1}) \sim (\frac{\alpha}{\alpha+i-1})G_0 + \frac{1}{\alpha+i-1} \sum_{h=1}^{i-1} \delta_{\boldsymbol{\xi}_h}$ where $\delta_{\boldsymbol{\xi}}$ is a degenerate distribution with all its mass on $\boldsymbol{\xi}$. Hence, the first subject has their latent variable value drawn from G_0 and as subjects are added they are either grouped with one of the existing subjects or allocated to a new cluster, with probability decreasing as the number of subjects increases. This allows there to be infinitely many latent classes represented in the population, with $k \leq n$ classes represented in the current sample. Allowing the number of latent classes to grow slowly with the sample size is more realistic, in most applications, than assuming a fixed, finite number of classes.

Another property of the DP, which we will utilize in describing a modification to incorporate identifiability constraints, is the stick-breaking representation of Sethuraman (1994).

In particular, $G \sim DP(\alpha G_0)$ implies that

$$G = \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_{\theta_h}, \quad V_h \stackrel{iid}{\sim} \text{beta}(1, \alpha), \quad \theta_h \stackrel{iid}{\sim} G_0 \quad (3)$$

where $V_h, h = 1, \dots, \infty$ is an infinite sequence of stick-breaking probabilities, and $\theta_h, h = 1, \dots, \infty$ an infinite sequence of random atoms. If $\xi_i \sim G$, then this formulation implies that $\xi_i = \theta_h$ with probability $\pi_h = V_h \prod_{l < h} (1 - V_l)$. The cluster probabilities are unknown, with the prior for the number of latent classes represented in a sample of n subjects controlled by α . If α is small, then the tendency is to group all the individuals together. By placing a hyperprior distribution on α , one can allow the data to inform more strongly about the estimated number of latent classes. It is clear from (3) that the latent classes having the highest weights, π_h , will be occupied quickly as subjects are added, but that there will be infinitely-many classes having extremely small probabilities. This allows for very rare traits (e.g., corresponding to a rare health condition or psychological disorder) that may be quite unlikely to observe in a moderate sized sample, but may be occasionally present in large samples.

Until this point, we have focused on using DPs for latent classes distributions. However, one can easily modify the formulation to accommodate continuous latent traits having unique values for each subject, with the density of these traits being completely unknown. To allow this, we simply add another level to the hierarchy and let $\xi_i \sim N(\mu_i, \sigma^2)$, with $\mu_i \sim G$ and $G \sim DP(\alpha G_0)$. Hence, instead of using a DP prior for the distribution of ξ_i directly, we characterize the distribution of ξ_i as a DP mixture (DPM) of normals, as in Escobar and West (1995). Any smooth density on the real line can be accurately approximated using such an approach. However, in the setting of SEMs, this flexibility raises questions about identifiability of the latent variable distributions.

3.2 Identifiability Issues

Because we do not observe the latent variables, ξ_i and η_i , directly for any of the subjects

under study, it is clear that some constraints are needed on the latent variable distributions and/or the parameters in (1) and (2). In the parametric case, such constraints were discussed in Bollen (1989), and there are two common strategies used. The first relies on fixing the diagonal elements of $\mathbf{\Lambda}_y$ and $\mathbf{\Lambda}_x$ as equal to one, in order to identify the scale of the latent variable distributions, while including sufficient numbers of structural zeros in the factor loadings matrices. In the Bayesian setting, such an approach is often unappealing in requiring one to have prior knowledge that a specific measurement is particularly relevant in defining a latent trait.

For example, suppose that $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})'$ consists of three measures of sperm concentration for individual i based on different technologies, and we consider the following factor model:

$$\begin{aligned} y_{ij} &= \mu_{y,j} + \lambda_{y,j}\eta_i + \epsilon_{y,ij}, & \epsilon_{y,ij} &\sim N(0, \sigma_{y,j}^2) \\ \eta_i &\sim F. \end{aligned} \tag{4}$$

Then, if F has an unknown mean and variance, we can still ensure identifiability by letting $\mu_{y,1} = 0$ and $\lambda_{y,1} = 1$. However, this implies that the first measure of sperm concentration is treated differently than the other two measures, which is artificial unless we have prior knowledge that the first measure is in some sense a gold standard. An alternative, which would treat the different sperm concentration measures as exchangeable, is to restrict F to have mean zero and variance one, while letting $\lambda_{y,j} \geq 0$ for $j = 1, 2, 3$ to remove sign ambiguity.

In the parametric case, restricting the mean and variance of F is straightforward, as we can simply let F correspond to the standard normal distribution. However, if we let F be unknown through using a DP or DPM prior, such constraints are not incorporated. One strategy is to choose the base distribution to be constrained. For example, one could let $F \sim DP(\alpha F_0)$, with F_0 chosen to correspond to the standard normal distribution. In this

case, the prior expectation of the mean of F is zero and the prior expectation of the variance of F is one. However, the posterior expectations of the mean and variance of F can deviate substantially from these prior expectations, leading to substantially biased inferences.

3.3 Centered Dirichlet Process

To address this problem, we propose to use a centering idea initially introduced by Dunson et al. (2007). In particular, if $\boldsymbol{\xi}_i \sim G$, with $G \sim \mathcal{Q}$, we say that \mathcal{Q} corresponds to a centered Dirichlet process (CDP) prior if

$$\begin{aligned} G &= \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_{\boldsymbol{\theta}_h}, \\ \boldsymbol{\theta}_h &= \boldsymbol{\Sigma}_G^{*-1/2} (\boldsymbol{\theta}_h^* - \boldsymbol{\mu}_G^*), \quad h = 1, \dots, \infty, \\ V_h &\sim \text{beta}(1, \alpha), \quad h = 1, \dots, \infty, \\ \boldsymbol{\theta}_h^* &\stackrel{iid}{\sim} G_0, \quad h = 1, \dots, \infty, \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_G^*$ and $\boldsymbol{\Sigma}_G^*$ correspond to the unconstrained mean and covariance under a $\text{DP}(\alpha G_0)$ prior,

$$\begin{aligned} \boldsymbol{\mu}_G^* &= \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \boldsymbol{\theta}_h, \\ \boldsymbol{\Sigma}_G^* &= \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) (\boldsymbol{\theta}_h - \boldsymbol{\mu}_G^*) (\boldsymbol{\theta}_h - \boldsymbol{\mu}_G^*)'. \end{aligned} \tag{6}$$

As shorthand notation for prior (5), we use $G \sim \text{CDP}(\alpha G_0)$. We will also consider a finite approximation, $\text{CDP}_N(\alpha G_0)$, which corresponds to letting $V_N = 1$, implying that terms $N + 1, \dots, \infty$ can be discarded in (5). Here, N can be viewed as an upper bound on the number of latent classes. Due to the standardization of the atoms, $\{\boldsymbol{\theta}_h\}$, both the $\text{CDP}(\alpha G_0)$ and $\text{CDP}_N(\alpha G_0)$ priors are constrained so that G has mean $\mathbf{0}$ and identity covariance, \mathbf{I} , a posteriori.

Note that one can induce a $\text{CDP}_N(\alpha G_0)$ prior on G by letting

$$\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_G^{*-1/2} (\boldsymbol{\xi}_i^* - \boldsymbol{\mu}_G^*), \quad i = 1, \dots, n,$$

$$\begin{aligned}
\xi_i^* &\sim G^*, \quad i = 1, \dots, n, \\
G^* &= \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_{\theta_h^*}.
\end{aligned} \tag{7}$$

This formulation, which treats the uncentered DP as a parameter-expanded version of the centered DP, will form the basis of our computational algorithm.

3.4 Centered Dirichlet Process Mixtures

The CDP is appropriate for modeling of an unknown discrete distribution having mean zero and identity covariance. Hence, the CDP can be used as a prior for a latent class model. For continuous latent traits, we instead propose a CDP mixture (CDPM), which has the following form

$$\begin{aligned}
\xi_i &= (\Sigma_{G^*} + \mathbf{I})^{-1/2} (\xi_i^* - \mu_{G^*}^*), \quad i = 1, \dots, n, \\
\xi_i^* &\sim N_r(\mu_i^*, \mathbf{I}), \quad i = 1, \dots, n, \\
\mu_i^* &\sim G^*,
\end{aligned} \tag{8}$$

where each of the terms is as defined in Section 3.3, but we now incorporate an additional level to draw ξ_i^* from a normal distribution centered on μ_i^* with identity covariance, with μ_i^* drawn from G^* instead of ξ_i^* being drawn from G^* directly. Marginalizing out the latent variables $\{\xi_i^*\}$, we obtain:

$$\begin{aligned}
\xi_i &\sim N_r\left(\mu_i, (\Sigma_G^* + \mathbf{I})^{-1}\right), \quad i = 1, \dots, n, \\
\mu_i &\sim G, \quad i = 1, \dots, n, \\
G &= \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_{\theta_h}, \\
\theta_h &= (\Sigma_G^* + \mathbf{I})^{-1/2} (\theta_h^* - \mu_{G^*}^*), \quad h = 1, \dots, \infty.
\end{aligned} \tag{9}$$

Thus, the induced mixture of normals distribution for ξ_i has mean and covariance,

$$\mathbb{E}(\xi_i | G) = \mathbf{0} \quad \text{and} \quad \mathbb{V}(\xi_i | G) = \Sigma_G^* (\Sigma_G^* + \mathbf{I})^{-1} + (\Sigma_G^* + \mathbf{I})^{-1} = \mathbf{I},$$

so that we obtain a highly flexible prior for the latent variable distribution that is restricted to have zero mean and identity covariance for identifiability. Following Muliere and Tardella (1998) and Ishwaran and James (2001), we can obtain an accurate approximation to the CDP and CDPM models by truncating the stick-breaking representation by letting $V_N = 1$. Here, the truncation level, N , is interpretable as an upper bound on the number of mixture components, or latent classes, needed to characterize the data. As it is well known that one can obtain an accurate approximation to any smooth density using a mix of a modest number of normal components (e.g., 5-7), an upper bound of 20 or so should be sufficient from a practical perspective. This produces an accurate approximation to the CDP and CDPM when α is small (e.g., $\alpha \approx 1$), which is often favored in applications as giving a sparse approximation to the data while allowing considerable flexibility.

4. Posterior Computation

4.1 *PX-Blocked Gibbs Sampler*

For posterior computation in the semiparametric SEMs proposed in Section 2, with the CDP or CDPM priors of Section 3, we propose a parameter-expansion (PX) blocked Gibbs sampler. PX algorithms have been increasingly widely used to accelerate EM (Liu et al., 1998) and Gibbs sampling convergence (Liu and Wu, 1999). The basic idea of PX data augmentation algorithms is that one can reduce autocorrelation in Gibbs sampling by introducing extra variables, which are not identified but are only incorporated for computational reasons to reduce posterior dependence in the draws of the parameters of interest. As shown by Ghosh and Dunson (2008) for parametric latent factor models, PX Gibbs samplers can have dramatically improved performance relative to typical Gibbs samplers, particularly when such algorithms perform poorly. Unfortunately, such poor performance is standard in latent factor and structural equation models, and autocorrelation in MCMC samples is often very high, making it necessary to collect hundreds of thousands or even millions of samples in

many cases to be assured of convergence and obtain estimates with negligible Monte Carlo error.

In addition to improving convergence and mixing rates, parameter expansion can be used to induce new classes of priors having appealing properties (Gelman, 2004). This approach was used by Ghosh and Dunson (2008), in the setting of factor analysis, to induce heavy-tailed default priors for the factor loadings, which provide a robust plug-in specification in the absence of informative prior knowledge. Here, we induce priors for the parameters in SEM-CDP or SEM-CDPM models through a PX specification. In particular, we treat an SEM with a typical DP or DPM prior on the latent variable distributions as a parameter-expanded version of the SEM-CDP or SEM-CDPM model. Here, the extra or redundant parameters correspond to the mean and covariance of the latent variable distributions. In the SEM-DP or SEM-DPM models, we follow standard practice in PX analyses and do not worry about lack of identifiability due to the redundancy in the parameterization.

We then run a blocked Gibbs sampler, as described in Ishwaran and James (2001), for the SEM-DP or SEM-DPM model. A naive analysis utilizing the results from this blocked Gibbs sampler will perform poorly, with high autocorrelation in the Gibbs samples and unreliable inferences due to the non-identifiability problem. However, prior to using the results for inferences, we apply a simple post-processing approach, which consists of transforming the draws from the “working” model parameterization (corresponding to the SEM-DP or SEM-DPM) to the “inferential” model parameterization (corresponding to the SEM-CDP or SEM-CDPM). Note that all of the latent variable distributions need not be constrained to have mean zero and variance identity in the SEM-CDP or SEM-CDPM analyses; instead we can incorporate only those constraints that would have been included in a parametric analysis with normally distributed latent traits.

4.2 Outline of Sampling Steps

The blocked Gibbs sampler is a standard approach for implementing posterior computation for DPMs and the adaptation to structural equation models is straightforward. Hence, we provide only a brief outline of the basic steps. Note the blocked Gibbs sampler relies on truncating the stick-breaking representation of the DP shown in (3) by letting $V_N = 1$, so that the $N + 1, \dots, \infty$ terms in the sum can be discarded. For sake of brevity, focus on the case in which

$$\begin{aligned}\boldsymbol{\delta}_i^* &\sim N_r(\boldsymbol{\mu}_{\delta,i}^*, \mathbf{I}), & \boldsymbol{\mu}_{\delta,i}^* &\stackrel{iid}{\sim} G_\delta^*, & G_\delta^* &\sim DP(\alpha G_{\delta,0}), \\ \boldsymbol{\xi}_i^* &\sim N_s(\boldsymbol{\mu}_{\xi,i}^*, \mathbf{I}), & \boldsymbol{\mu}_{\xi,i}^* &\stackrel{iid}{\sim} G_\xi^*, & G_\xi^* &\sim DP(\alpha G_{\xi,0}),\end{aligned}\tag{10}$$

where the $*$ superscript denotes that these are the unknowns under the working model, $G_{\delta,0}$ corresponds to $N_r(\mathbf{0}, \mathbf{I})$ and $G_{\xi,0}$ corresponds to $N_s(\mathbf{0}, \mathbf{I})$. This specification implies that the working model assigns F^* and G^* DPM of normal priors, with the covariance fixed to correspond to the identity matrix and with DP priors placed on the subject-specific location parameters. Hence, the latent variables are modeled as continuous variables with unknown distributions.

Using the truncation approximation, we let

$$G_\delta^* = \sum_{h=1}^N V_{\delta,h} \prod_{l < h} (1 - V_{\delta,l}) \delta_{\boldsymbol{\theta}_{\delta,h}^*} \quad \text{and} \quad G_\xi^* = \sum_{h=1}^N V_{\xi,h} \prod_{l < h} (1 - V_{\xi,l}) \delta_{\boldsymbol{\theta}_{\xi,h}^*},$$

where $V_{\delta,h} \stackrel{iid}{\sim} \text{beta}(1, \alpha_\delta)$, $V_{\xi,h} \stackrel{iid}{\sim} \text{beta}(1, \alpha_\xi)$, $\boldsymbol{\theta}_{\delta,h}^* \stackrel{iid}{\sim} N_r(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\theta}_{\xi,h}^* \stackrel{iid}{\sim} N_s(\mathbf{0}, \mathbf{I})$ independently, for $h = 1, \dots, N$, *a priori*, with $V_{\delta,N} = V_{\xi,N} = 1$. We introduce latent class index $S_{\delta,i} = h$ denoting that $\boldsymbol{\mu}_{\delta,i}^* = \boldsymbol{\theta}_{\delta,h}^*$ and $S_{\xi,i} = h$ denoting that $\boldsymbol{\mu}_{\xi,i}^* = \boldsymbol{\theta}_{\xi,h}^*$. Then, we alternate between the following steps after choosing initial values:

1. Allocate individuals to latent classes for $\boldsymbol{\delta}_i^*$ and $\boldsymbol{\xi}_i^*$ by sampling $S_{\delta,i}$ and $S_{\xi,i}$ from their multinomial full conditional posterior distribution. Note that individuals within a class will vary in the exact values of the latent traits, $\boldsymbol{\delta}_i^*$ and $\boldsymbol{\xi}_i^*$, but will have the same class-specific mean.

2. Update the stick-breaking random variables, $V_{\delta,h}$ and $V_{\xi,h}$ by sampling from their beta full conditional posterior distributions.
3. Update the DP precision parameters, α_δ and α_ξ by sampling from their gamma full conditional posterior distributions, assuming gamma priors.
4. Update the component-specific parameters, $\theta_{\delta,h}^*$, by sampling from the conjugate multivariate normal obtained in updating the $N_r(\mathbf{0}, \mathbf{I})$ prior with the normal likelihood for $\boldsymbol{\delta}_i$ for those subjects with $S_{\delta,i} = h$, $h = 1, \dots, N$.
5. Update the component-specific parameters, $\theta_{\xi,h}^*$ by updating the $N_s(\mathbf{0}, \mathbf{I})$ prior with the normal likelihood for $\boldsymbol{\xi}_i$ for those subjects with $S_{\xi,i} = h$, $h = 1, \dots, N$.
6. Update the free elements in $\mathbf{B}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\mu}_y^*, \boldsymbol{\mu}_x^*, \boldsymbol{\Lambda}_y^*, \boldsymbol{\Lambda}_x^*, \boldsymbol{\Sigma}_y^*, \boldsymbol{\Sigma}_x^*$ using the full conditional posterior distributions derived as for a typical parametric SEM.

This steps are repeated for a large number of iterations, with a burn-in discarded to allow convergence.

Upon convergence, this algorithm produces draws from the posterior under the SEM-DPM working model. As mentioned above, we need to then post-process the samples to obtain inferences under the SEM-CDPM inferential model. This is accomplished by transforming each of the samples. In describing this transformation, we focus on the simple case in which all the latent variable distributions are constrained to have zero mean and identity covariance, though modifications to constrain a subset of the latent variables and only constrain the mean or variance are straightforward.

1. Calculate $\boldsymbol{\mu}_{G_\delta}^*$, $\boldsymbol{\Sigma}_{G_\delta}^*$, $\boldsymbol{\mu}_{G_\xi}^*$ and $\boldsymbol{\Sigma}_{G_\xi}^*$ relying on expression (6), noting that the $N + 1, \dots, \infty$ terms are zero under the truncation approximation.
2. Let $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_{G_\xi}^{*-1/2}(\boldsymbol{\xi}_i^* - \boldsymbol{\mu}_{G_\xi}^*)$ and $\boldsymbol{\delta}_i = \boldsymbol{\Sigma}_{G_\delta}^{*-1/2}(\boldsymbol{\delta}_i^* - \boldsymbol{\mu}_{G_\delta}^*)$

3. Calculate the parameters for the inferential models

$$\begin{aligned}
B &= \Sigma_{G_\delta}^{*1/2} B^* \\
\Gamma &= \Sigma_{G_\delta}^{*-1/2} \Gamma^* \Sigma_{G_\xi}^{*1/2} \\
\Lambda_x &= \Lambda_x^* \Sigma_{G_\xi}^{*1/2} \\
\Lambda_y &= \Lambda_y^* (I - B)^{-1} \Sigma_{G_\delta}^{*1/2} \\
\mu_x &= \mu_x^* + \Lambda_x^* \mu_{G_\delta}^* \\
\mu_y &= \mu_y^* + \Lambda_y^* (I - B)^{-1} (\Gamma \mu_{G_\xi}^* + \mu_{G_\delta}^*)
\end{aligned}$$

5. Application 1: Uterine Fibroids and Bleeding

5.1 Background and Description

We initially consider an application to data from an NIEHS study of uterine fibroids (Baird et al. (2003)), a common reproductive tract tumor, which rarely becomes malignant, but leads to substantial morbidity. In cross-sectional analyses of data from this study, fibroid size was related to increased bleeding (Wegienka et al. 2003). The goal of the current study was to assess whether the current presence and size of uterine fibroids predict the future level of bleeding.

The uterine fibroid study was conducted by NIEHS in 1996 in collaboration with George Washington University Medical Center. Members aged 35-49 of an urban prepaid health plan in Washington D.C. were selected for the study, out of 1430 participants, 1245 were premenopausal. In the study, information on menstrual, medical and reproductive history as well as any previous fibroid diagnoses and treatment were collected by phone interview. Detailed information on fibroid location and size were collected by ultrasound examination during a clinic visit or from recent medical records if available. After 3-5 years, we attempted to re-contact the premenopausal women, 981 of whom were interviewed and asked about symptoms. If women had had a myomectomy, hysterectomy, or menopause prior to followup,

they were asked about symptoms prior to those events. Generally, African-American women have higher risk of uterine fibroids than other ethnic groups (Baird et. al., 2003). Our interest is in assessing how fibroid size at baseline and African American ethnicity relate to bleeding at the follow-up.

Size of the fibroid is categorized as 0, 1, 2 or 3, corresponding to none, small ($< 2\text{cm}$), medium (between 2 and 4cm) or large ($> 4\text{cm}$). The following data are available on the intensity of bleeding at follow-up:

- Count data:
 - Y_1 : number of days during menses of real blood flow.
 - Y_2 : number of days of spotting.
 - Y_3 : number of days each month in using more than 8 pads or tampons.
- Binary data:
 - Y_4 : Is there intermenstrual spotting?
- Ordinal data (1-5 scale):
 - Y_5 : How often do you have menstrual periods?
 1. Did not have any period.
 2. Too irregular to say.
 3. Less frequently than once a month (>34 days).
 4. About once a month (27-34 days).
 5. More frequently than once a month (<27 days).
 - Y_6 : How often do you have gushing-type bleeding?
 1. Just once.

- 2. During occasional periods.
 - 3. Most periods.
 - 4. Every period.
- Y_7 : How much did the menstrual bleeding limit social activities?
- 1. Not at all.
 - 2. A little.
 - 3. Some.
 - 4. A lot.

Summary statistics for the bleeding symptom data are provided in Table 1. For flexibility in modeling and because most women had values close to 0, we treat the count data as ordinal data for our analysis.

Letting η_i denote the latent bleeding intensity score for woman i , we used model

$$\eta_i = \mathbf{x}_i' \boldsymbol{\beta} + \delta_i, \quad \delta_i \sim G \quad (11)$$

to relate fibroid size and African American ethnicity to bleeding intensity. The vector \mathbf{x}_i is coded without an intercept and with indicators for (x_{i1}) small, (x_{i2}) medium and (x_{i3}) large fibroids as well as (x_{i4}) African American ethnicity. To relate the bleeding score η_i , to the ordered categorical symptom data, we used a continuation ratio measurement model:

$$P(y_{ij} = c | y_{ij} \geq c, \tau, \lambda, \eta) = \Phi(\tau_{jc} - \lambda_j \eta_i), \quad c = 1, \dots, C_j, \quad (12)$$

where C_j is the number of categories for symptom type j , $\Phi(\cdot)$ is the CDF of the standard normal function, $\lambda_1, \dots, \lambda_7$ are the loading factors for symptoms $Y_1 - Y_7$.

5.2 Simulation Experiment

We assessed the performance of the approach through a simulation example designed to mimic the fibroid data described in section 5.1. In this application, we are interested in

inference under the latent factor regression model (11) with the same sample size and \mathbf{x}_i values from the real data. For the simulation, we assume that the true parameter values are $\beta = (1, 1, 1, 1)'$, $\Lambda = (\lambda_1, \dots, \lambda_7)' = (1, 1, 1, 1, 1, 1, 1)'$ and the latent variable density η_i is the following mixture of four normals:

$$0.15N(-1.92, 0.24) + 0.05N(-0.95, 0.24) + 0.15N(0.024, 0.24) + 0.65N(0.51, 0.24)$$

which has mean 0 and variance 1.

The values of $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and n are taken directly from the observed data. One of our goals was to assess whether the data contain sufficient information to reliably estimate the latent variable density.

We analyzed the simulation data using a CDPM prior for G , applying the algorithm of section 4. The DP precision parameter, α , was treated as unknown using a gamma(1, 1) hyperprior, while G_0 was assumed to correspond to $N(0, 1)$. Conditionally-conjugate priors were chosen for the remaining parameters: the intercept τ_{jc} was placed a prior $N(0, 4.0)$, the loading factor λ_j with $N(0, 10.0)$ for $j = 1, \dots, 7$, the vector β with independent $N(0, 5.0I)$.

A blocked Gibbs sampler was implemented in each case, with the chain run for 100,000 iterations. After a 20,000 iteration burn-in, we take every 20th sample, resulting in a total of 4000 samples. To assess convergence, we ran several independent chains with widely dispersed starting values; for sensitivity to prior specification, we also tried with varied variances: priors with variance/2, priors with variance $\times 2$, priors with variance $\times 5$. With all these trials, we do not see much differences between the results.

Table 2 presents posterior summaries of the model parameters for cases of CDP and CDPM, while Figure 1 plots the estimated and true latent variable distributions of CDPM. From these results, we can see that our approach can produce good results. The estimated latent variable density is very close to the true density, suggesting that the data are informative.

The centered Dirichlet process mixture (CDPM) model results are much more accurate than the results for the DPM model, as expected due to the non-identifiability problem. In general, the closer the latent variable distribution is to the base G_0 , the better the performance of the DPM model. However, the performance of the DPM degrades in the presence of deviation from G_0 , while the CDPM results are robust to the shape of the latent variable density.

5.3 Analysis of Real Data

We implemented the analysis as in the simulation example, and again found the results robust to the prior specification. Posterior summaries of the parameters are provided in Table 3. These results suggest a significant increase in bleeding intensity with increasing fibroid size and for African American women compared with other races. For small fibroids compared with no fibroids, the expected change in the latent bleeding intensity score is 0.05 and the 95% credible interval (CI) includes 0. Note that the latent variable regression coefficients have a clear interpretation due to the incorporation of the variance=1 constraint. In particular, the coefficients for the indicators represent the number of standard deviations the mean bleeding intensity score shifts between the categories. Hence, a shift of 0.05 is clearly not a clinically significant change. However, the estimated shift of $\hat{\beta}_1 + \hat{\beta}_2 = 0.05 + 0.45 = 0.50$ between no fibroids and size category 2 is significant. The estimated shift between no fibroids and size category 3 is $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 1.26$. Hence, fibroid size explains a sizable proportion of the variability in the latent bleeding score.

Interestingly, African American ethnicity is also a significant predictor of bleeding intensity, even adjusting for fibroid size. Although it is known that African Americans have a higher fibroid prevalence, so that it would not be surprising to see more fibroid related bleeding, the occurrence of higher bleeding rates adjusting for fibroid sizes is interesting. It may be that future development of fibroids between the screening examination and the

measurement of bleeding symptoms at the follow-up time may explain this difference.

The estimated latent bleeding intensity residual density is plotted in Figure 2. Interestingly, the density is quite similar to a normal density, though we have demonstrated power to detect non-normality in the simulation example.

It is important to assess which symptoms provide the most information about the latent bleeding intensity score for a woman and hence are most sensitive to fibroid size. With this goal in mind, we plot the predicted mean symptom score in different fibroid size categories for African American women in Figure 3. The plot for white women and other ethnicities shows a very similar pattern. For symptoms 2, 4 and 5, there are essentially no differences across the fibroid size categories and the factor loading parameters are low, suggesting that the bleeding intensity score has low correlation with these symptoms. Symptoms 2 and 4 relate to spotting, while symptom 5 relates to frequency of menstrual periods. In contrast, for symptom 1, there is a moderate shift across fibroid size categories, while for symptoms 3, 6 and 7, the shift is large, with non-overlapping 95% predictive intervals. These findings are quite plausible biologically, as symptoms 3 and 6 relate to frequency of severe bleeding, while symptom 7 measures bleeding that is sufficient to limit activities.

6. Application 2: Education and Fertility Data

6.1 Background and Description

As a second example, we analyze data from the widely used Switzerland Socio-Economic study (Walle, 1980). The data set contains information on fertility, education and other variables between 1870 and 1930. This data set is interesting in that it collected data around the time of the demographic transition. This transition is characterized by a shift from high fertility and high mortality to low fertility and low mortality. Our focus is on studying how the relationship between education and fertility changes near the time of demographic transition. In this data set, both education and fertility are measured with multiple items,

so it is most natural to consider a structural equation model with latent variables. However, it is appealing to avoid assuming normally distributed latent variables.

We analyzed county-level data for four different years, including 1870 ($j = 1$), 1888 ($j = 2$), 1910 ($j = 3$) and 1930 ($j = 4$). Let $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})'$ denote the measures of education level in county i at year j , and let $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2}, y_{ij3}, y_{ij4})'$ denote the measures of fertility. Education level was measured by scores of military recruits, with x_{ij1} =proportion with highest grade, x_{ij2} =proportion with lowest grade, and x_{ij3} =proportion with education beyond primary school. Fertility was measured by four indexes, including y_{ij1} = total fertility (IF), y_{ij2} = marital fertility (IG), y_{ij3} = unmarried fertility (IH), and y_{ij4} = ratio of marital fertility to total fertility. There was a change of grading system in 1888.

Let $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \xi_{i3}, \xi_{i4})'$ denote the latent education level of county i at the four times, and let $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \eta_{i3}, \eta_{i4})'$ denote the latent fertility. We define the following measurement models:

$$\begin{aligned} \mathbf{x}_{i1} &= \mu_{x1} + \Lambda_x^1 \boldsymbol{\xi}_{i1} + \epsilon_{i1}^x, & \epsilon_{i1}^x &\sim N(0, \Sigma_x^1) \\ \mathbf{x}_{ik} &= \mu_{x2} + \Lambda_x \boldsymbol{\xi}_{ik} + \epsilon_{ik}^x, & \epsilon_{ik}^x &\sim N(0, \Sigma_x) \\ \mathbf{y}_{ik} &= \mu_y + \Lambda_y \boldsymbol{\eta}_{ik} + \epsilon_{ik}^y, & \epsilon_{ik}^y &\sim N(0, \Sigma_y), \quad k = 2, 3, 4 \\ \mu_{x1} &= (\mu_x^1, \mu_x^2, \mu_x^3)', & \mu_{x2} &= (\mu_x^4, \mu_x^5, \mu_x^6)' \end{aligned}$$

Here, we use μ_{x1} , μ_{x2} to denote the different means for 1870 and later years respectively due to change of grading system .

We assume the latent variable model has the following semiparametric form:

$$\begin{aligned} \eta_i &= \Gamma \boldsymbol{\xi}_i + \delta_i, & \delta_i &\sim N(\mu_{\delta_i}, \Sigma_{\delta_i}) \\ (\mu_{\delta_i}, \Sigma_{\delta_i}) &\sim F, & F &\sim DP(\alpha_1 F_0), \quad F_0 = N(\mu_{\delta_i} | \mu_0, \Sigma_{\delta_i}) IW(\Sigma_{\delta_i} | \Phi_0, m) \\ \boldsymbol{\xi}_i &\sim N(\mu_i^\xi, \Sigma_i^\xi), & (\mu_i^\xi, \Sigma_i^\xi) &\sim G, \quad G \sim DP(\alpha G_0) \\ G_0 &= N(\mu_0^\xi, \Sigma_0^\xi) IW(\Sigma_0^\xi | \Phi_1, q), & \alpha_1 &\sim \text{Gamma}(g, h), \quad \alpha \sim \text{Gamma}(g_1, h_1) \end{aligned}$$

We use a more general location-scale mixture of normals for these data instead of the location mixture proposed earlier. The location-scale mixture has the advantage of allowing multivariate densities to be characterized with fewer components. To solve the identifiability issue, we constrain the first element of ξ_i and δ_i to have mean 0 and variance 1. The post-process transformation of the parameters for the inferential model from the working model is similar to those in Section 4.2. We ran a simulation study to assess the ability for this model and sample size, and obtained good results: Table 4 summarizes the results of the estimated parameters and true values. Figures 4 and 5 provide the density estimation of the response latent variables and residuals respectively.

6.2 Analysis and Results

A blocked Gibbs sampler was implemented with the chain run for 20,000 iterations after a 20,000 iteration burn-in. Similar procedures for convergence assessment are taken as the last one, we do not see much difference among the results.

We standardize the data before analysis. The results are provided in Table 5. The correlation coefficient between the latent response variables and latent predictors shows a bigger negative value across time: the coefficient of the first year 1870 has a estimated mean value of -0.47 though the 95%CI (-1.54, 0.28) covers 0; the other three coefficients get more negative values -0.86, -0.76 and -1.39 for the year of 1888, 1910 and 1930 respectively and the corresponding 95% CIs do not cover 0, which indicates a stronger negative relationship between education and fertility across time. Thus we can see a clear picture of such demographic transition: in the year of 1870, we do not see a significantly negative relationship between education and fertility; in the middle of decline of birth, we see a fairly strong negative relation between education and fertility; in 1910, we got a similar negative relation as that of 1888, but a more pronounced one for the year of 1930. This may indicate a time-lag for the year 1910 and 1930.

Figure 6 provides the density estimation of both the residuals δ and the education latent variables ξ across time: obvious, we can see the mean of the education latent variables get bigger across time; the residuals get smaller across time, but since there is an increasing negative relation coefficient between education latent variables and fertility latent variables, thus we should expect a more pronounced decreasing fertility latent variables adjusted for education latent variables across time. Figure 7 provides the density estimate of residuals δ across time and also the normal density with the same mean and variance for the corresponding year. We can see the residual plots are definitely non-normal for the year of 1888, 1910 and 1930. Similarly, Figure 8 provides the density estimate of the education latent variables, again the plots indicate the year of 1910 and 1930 are non-normal. Thus in combination of the above residual and the education latent variables density plots, we see our nonparametric analysis more potent and reasonable. Figure 9 gives a direct relation between the fertility latent variables and the education latent variables, we can see the first plot (the year 1870), there is an influential point on the left corner, without which, the negative relation between education and fertility is much weakened. That explains why the correlation coefficient has a negative value estimate but the 95%CI includes 0.

Prior sensitivity and convergence are assessed as previous sections, all the results indicate the results and convergence are good.

7. Discussion

The structural equation models (SEMs) attract more attention due to increased interest in knowing complex multivariate relationship in biological, medical and other fields. We motivate our method from the epidemiology study and a Socio-Economic study of fertility and education. Generally, researchers of such fields are concerned with results due to strict parametric assumption or even asymptotic properties with small sample size.

We propose a flexible method which can be applied in both discrete or continuous vari-

ables with constrain on mean and variance Generally, identifiability issue arise for such models even with parametric ones; also indirect information about the latent variables provided by data makes it difficult for verification. Our method avoids the above strict parametric assumption and provide an easy post-process procedure, which can be easily applied in many fields.

References

- Ansari, A. and Iyengar, R. (2006). Semiparametric Thurstonian models for recurrent choice: A Bayesian analysis. *Psychometrika*, 71, 631-657.
- Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D. and Schectman, J.M. (2003), "High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence," *American Journal of Obstetrics and Gynecology*, 188, 100-107.
- Blackwell, D. and MacQueen, J.B. (1973). Ferguson distribution via Polya Urn schmes. *Annals of Statistics*, 2, 353-355.
- Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.
- Brown, E.R. and Ibrahim, J.G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59, 221-228.
- Burr, D. and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100, 242-251.
- Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83, 275-285.
- Dunson, D.B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7, 551-568.

- Dunson, D.B., Yang, M. and Baird, D. (2007). Semiparametric Bayes hierarchical models with mean and variance constraints. *Discussion Paper*, Department of Statistical Science, Duke University.
- Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.
- Fahrmeir, L. and Raach, A. (2007). A Bayesian semiparametric latent variable model for mixed responses. *Psychometrika*, in press.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 2, 209-230.
- Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, 2, 615-629.
- Fokoue, E. (2005). Mixtures of factor analyzers: an extension with covariates. *Journal of Multivariate Analysis*, 95, 370-384.
- Fokoue, E. and Titterington, D.M. (2003). Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation. *Machine Learning*, 50, 73-94.
- Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99, 537-545.
- Ghosh, J. and Dunson, D.B. (2008). Default priors and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, under invited revision.
- Ishwaran, H. and Takahara, G. (2002). Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association*, 97, 1154-1166.

- Jedidi, K., Jagpal, H.S. and DeSarbo, W.S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16, 39-59.
- Kleinman, K.P. and Ibrahim, J.G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54, 921-938.
- Lee, S.Y. and Xia, Y.M. (2006). Maximum likelihood methods in treating outliers and symmetrically heavy-tailed distributions for nonlinear structural equation models with missing data. *Psychometrika*, 71, 565-585.
- Li, Y., Müller, P., and Lin, X. "Bias-Corrected Inference in Semiparametric Bayesian Mixed Models" *Technical Report*.
- Liu, C.H., Rubin, D.B. and Wu, Y.N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85, 755-770.
- Liu, J.S. and Wu, Y.N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94, 1264-1274.
- Lubke, G.H. and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10, 21-39.
- McLachlan, G.J., Bean, R.W. and Jones, L.B.T. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics & Data Analysis*, 51, 5327-5338.
- Müller, P. and Rosner, G.L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, 92, 1279-1292.

- Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J. (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine*, 26, 2088-2112.
- Palomo, J., Dunson, D.B., and Bollen, K.A. (2007). Bayesian structural equation modeling. *Handbook of Latent Variable and Related Methods*, ed. S-Y. Lee, Elsevier.
- Zhu, H.T. and Lee, S.Y. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, 66, 133-152.
- Walle F. (1980). Education and the Demographic Transition in Switzerland. *Population and Development Review*, 3, 463-472.

Table 1: Empirical means within different fibroid size and ethnicity categories for the seven bleeding symptoms

Symptoms	Whites and other				African American			
	Fibroid size				Fibroid size			
	0	1	2	3	0	1	2	3
Y_1	4.05	3.80	4.92	4.81	3.94	3.88	4.63	5.88
Y_2	1.97	2.29	2.63	2.15	1.81	1.59	1.76	2.06
Y_3	0.51	0.25	0.98	1.41	0.75	1.07	1.73	2.45
Y_4	0.92	0.87	0.87	0.98	0.94	0.97	0.91	0.90
Y_5	2.86	2.65	2.79	2.75	2.80	2.84	2.91	2.91
Y_6	1.57	1.41	2.00	2.00	1.79	2.15	2.39	2.80
Y_7	1.27	1.27	1.40	1.75	1.29	1.54	1.75	1.93
n	1453	496	601	383	826	550	1079	759

Table 2: Parameter estimation of DPM & CDPM for simulation

Parameter	True value	DPM		CDPM	
		Estimate	95 % CI	Estimate	95 % CI
β_1	1.00	1.66	(1.27,2.09)	0.88	(0.68,1.07)
β_2	1.00	1.61	(1.26,2.00)	0.85	(0.68,1.01)
β_3	1.00	1.76	(1.38,2.24)	0.93	(0.74,1.14)
β_4	1.00	1.73	(1.44,2.11)	0.92	(0.78,1.05)
λ_1	1.00	0.54	(0.44,0.63)	1.02	(0.89,1.15)
λ_2	1.00	0.56	(0.46,0.65)	1.06	(0.92,1.20)
λ_3	1.00	0.56	(0.45,0.67)	1.06	(0.91,1.21)
λ_4	1.00	0.58	(0.48,0.69)	1.10	(0.94,1.29)
λ_5	1.00	0.62	(0.48,0.79)	1.18	(0.97,1.42)
λ_6	1.00	0.61	(0.49,0.72)	1.14	(0.98,1.32)
λ_7	1.00	0.58	(0.47,0.68)	1.09	(0.95,1.24)

Table 3: Parameter estimation of DPM & CDPM for real data analysis

Parameter	DPM		CDPM	
	Estimate	95 % CI	Estimate	95 % CI
β_1	0.07	(-0.21, 0.35)	0.05	(-0.18,0.28)
β_2	0.53	(0.29, 0.91)	0.45	(0.25,0.66)
β_3	0.91	(0.60, 1.45)	0.76	(0.51,1.01)
β_4	0.54	(0.34,0.87)	0.46	(0.28,0.63)
λ_1	0.51	(0.27,0.71)	0.60	(0.43, 0.83)
λ_2	0.02	(0.00,0.06)	0.02	(0.00,0.08)
λ_3	1.17	(0.73,1.55)	1.37	(1.04, 1.86)
λ_4	0.02	(0.00,0.07)	0.02	(0.00,0.05)
λ_5	0.12	(0.05,0.19)	0.14	(0.06, 0.18)
λ_6	0.96	(0.61,1.23)	1.13	(0.88, 1.50)
λ_7	0.80	(0.51,1.01)	0.93	(0.73, 1.23)

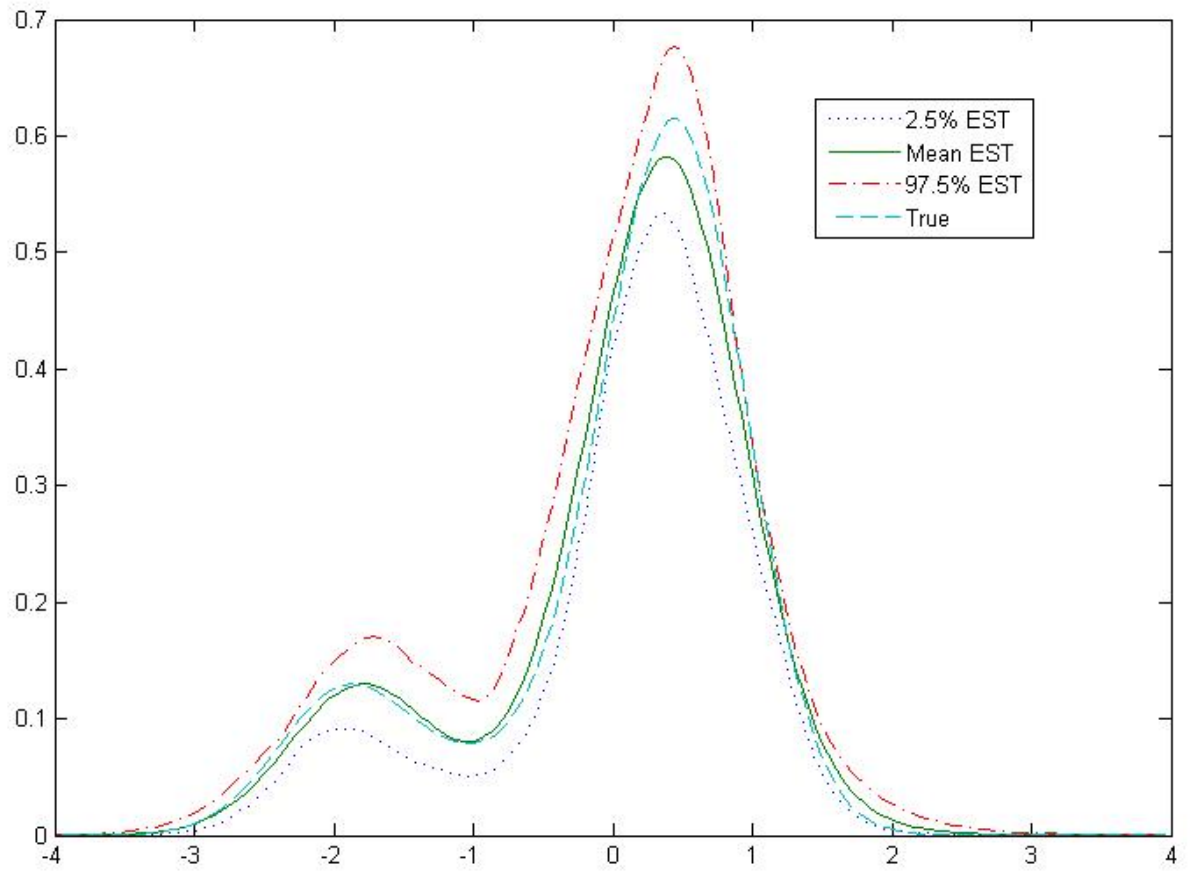


Figure 1: True and estimated latent variable densities in simulation example.

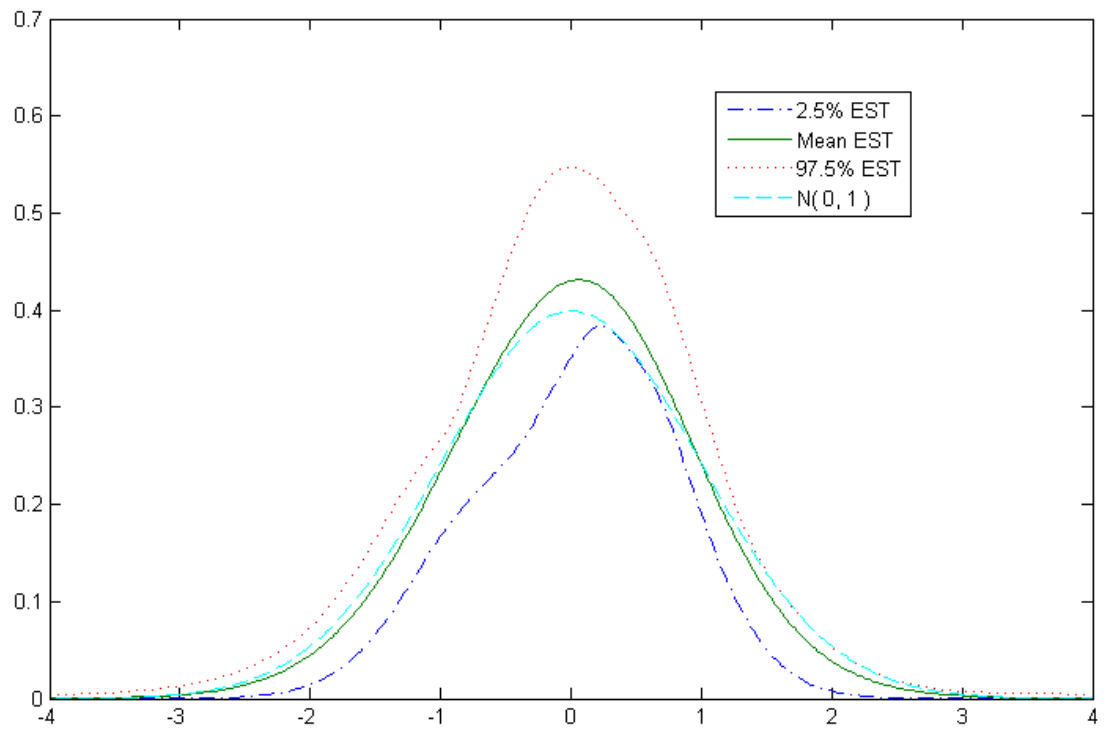


Figure 2: Estimated density of the latent bleeding intensity score in the fibroid data application

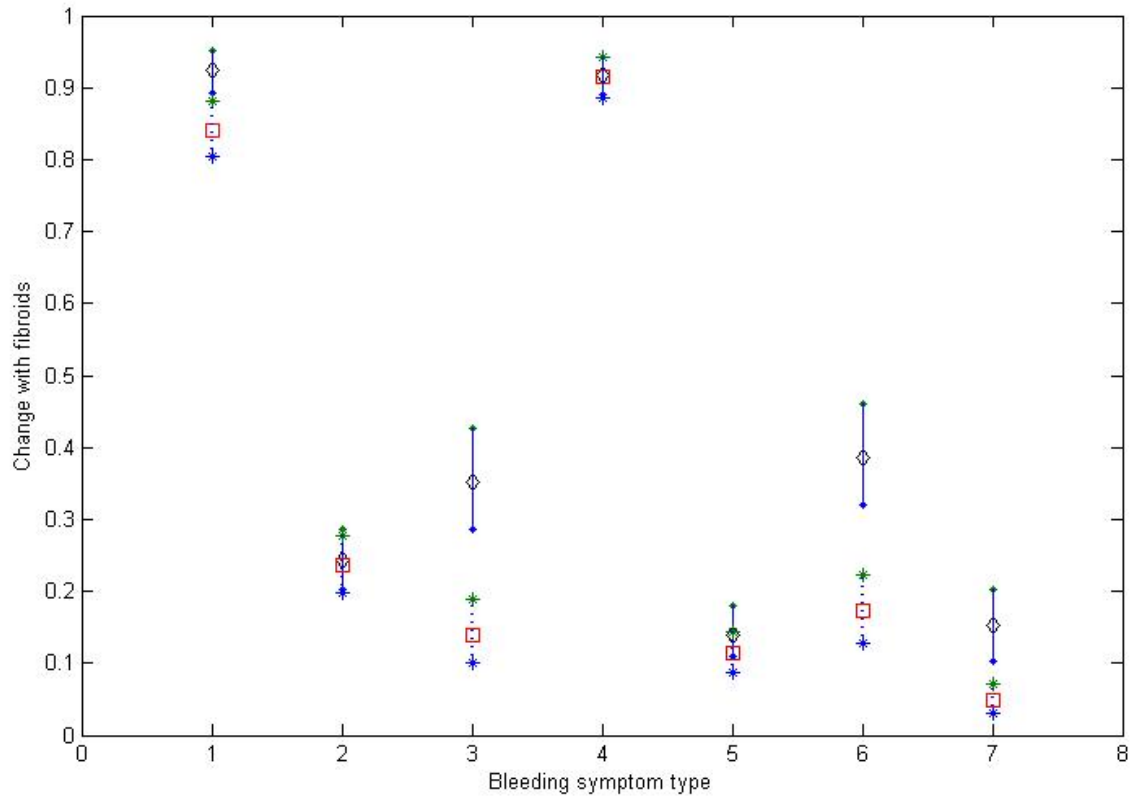


Figure 3: Comparison of bleeding symptoms for African American women with large fibroid size (solid line, diamond: estimated mean) vs. no fibroids (dashed line, square:estimated mean)

Table 4: Parameter estimation of CDPM for simulation of Switzerland Socio-Economic data

Parameter	True value	CDPM	
		Estimate	95 % CI
$\sigma_{y,1}^2$	0.50	0.52	(0.49, 0.56)
$\sigma_{y,2}^2$	0.50	0.50	(0.46, 0.54)
$\sigma_{y,3}^2$	0.50	0.50	(0.46,0.53)
$\sigma_{y,4}^2$	0.50	0.51	(0.47, 0.54)
$\sigma_{x,1}^2$	0.50	0.52	(0.49,0.55)
$\sigma_{x,2}^2$	0.50	0.48	(0.47,0.53)
$\sigma_{x,3}^2$	0.50	0.53	(0.48,0.54)
$\lambda_{y,1}^*$	1.00	0.96	(0.91, 1.01)
$\lambda_{y,2}^*$	1.00	0.97	(0.91, 1.02)
$\lambda_{y,3}^*$	1.00	0.96	(0.91, 1.02)
$\lambda_{y,4}^*$	1.00	0.97	(0.91, 1.02)
γ_1^*	1.00	1.02	(0.92, 1.11)
γ_2^*	0.50	0.52	(0.47, 0.56)
γ_3^*	0.80	0.82	(0.75, 0.91)
γ_4^*	1.00	1.01	(0.93, 1.10)
$\lambda_{x,1}^*$	0.50	0.50	(0.48, 0.53)
$\lambda_{x,2}^*$	1.00	1.01	(0.95, 1.06)
$\lambda_{x,3}^*$	1.50	1.51	(1.43, 1.59)

Table 5: Parameter estimation of CDPM for Switzerland Socio-Economic data

Parameter	Estimate	CDPM
		95 % CI
$\sigma_{y,1}^2$	0.15	(0.09, 0.22)
$\sigma_{y,2}^2$	0.12	(0.07, 0.22)
$\sigma_{y,3}^2$	0.85	(0.58, 1.30)
$\sigma_{y,4}^2$	1.10	(0.73, 1.74)
$\sigma_{x,1}^2$	0.27	(0.17, 0.44)
$\sigma_{x,2}^2$	0.15	(0.07, 0.33)
$\sigma_{x,3}^2$	0.58	(0.36, 1.02)
$\lambda_{y,1}^*$	0.41	(0.27, 0.60)
$\lambda_{y,2}^*$	0.44	(0.27, 0.64)
$\lambda_{y,3}^*$	0.19	(0.07, 0.35)
$\lambda_{y,4}^*$	0.05	(0.00, 0.19)
γ_1^*	-0.47	(-1.54, 0.28)
γ_2^*	-0.86	(-1.88, -0.11)
γ_3^*	-0.75	(-1.52, -0.23)
γ_4^*	-1.39	(-2.62, -0.62)
$\lambda_{x,1}^*$	0.30	(0.17, 0.48)
$\lambda_{x,2}^*$	0.58	(0.39, 0.83)
$\lambda_{x,3}^*$	0.06	(0.00, 0.20)

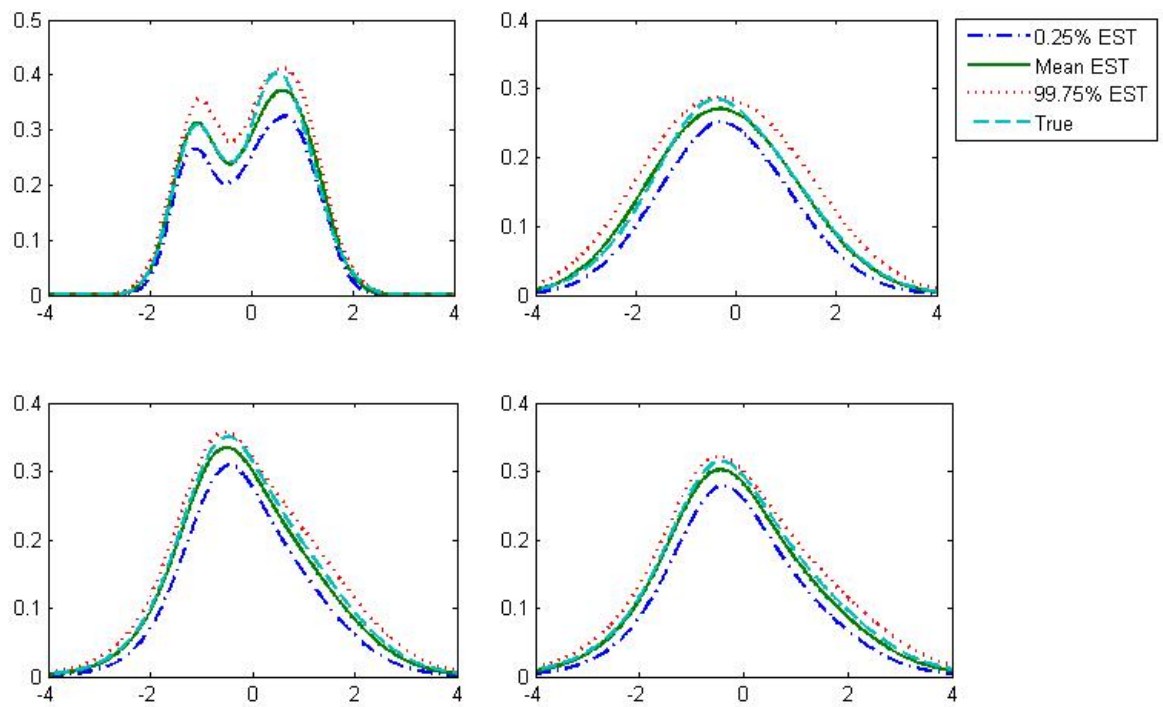


Figure 4: Density estimate of education latent variables

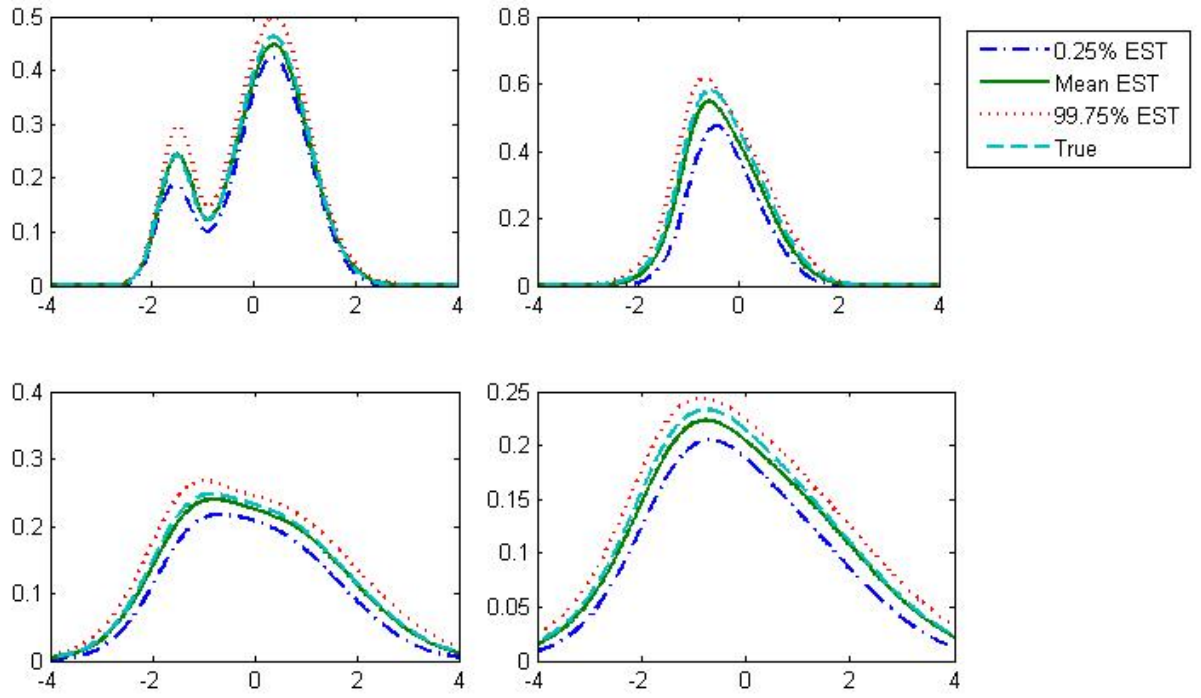


Figure 5: Density estimate of residuals

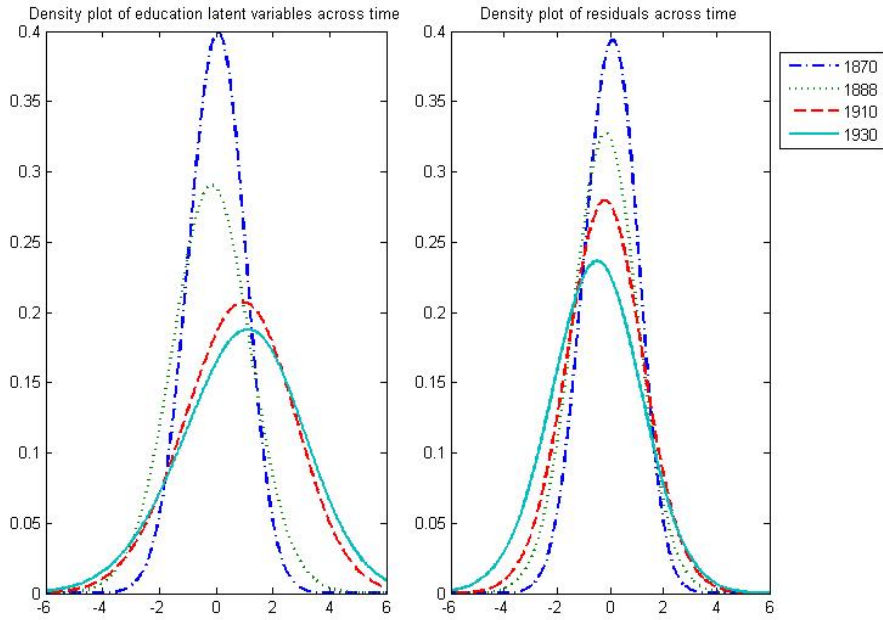


Figure 6: Density estimate of residuals and education latent variables across time

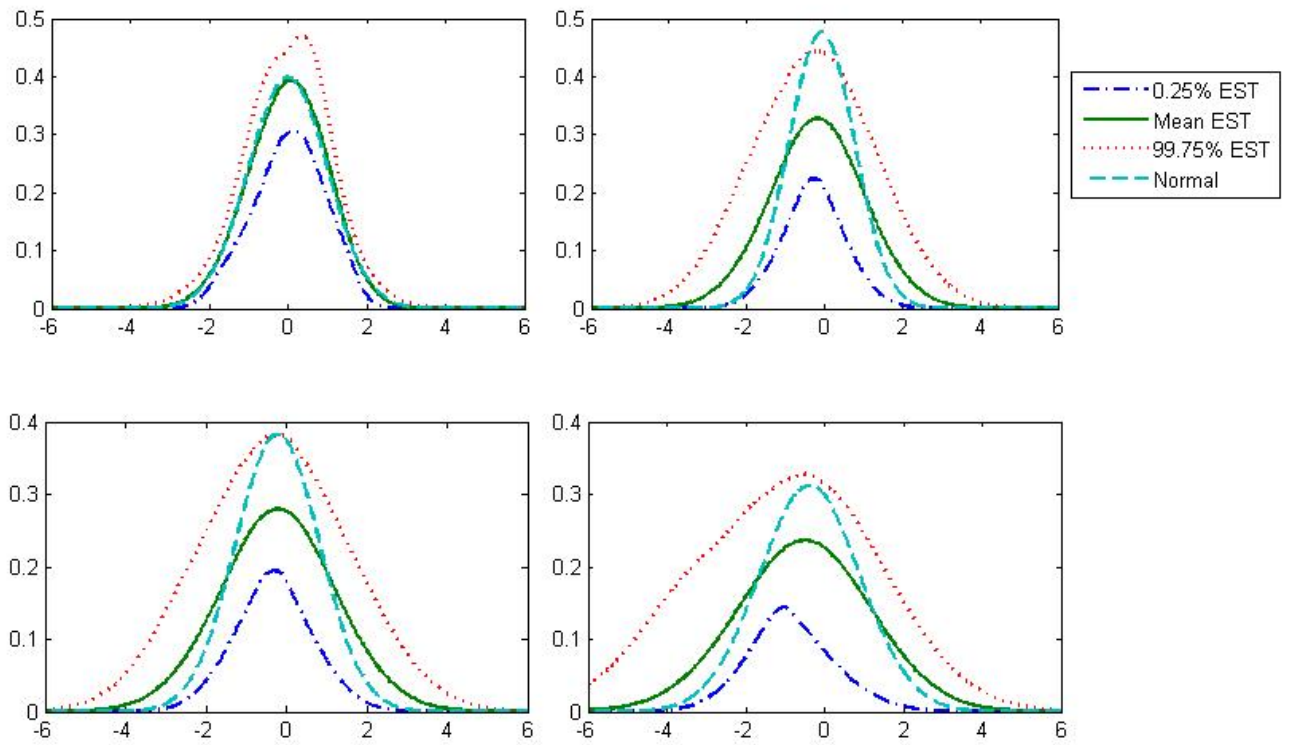


Figure 7: Density estimate of residuals across time

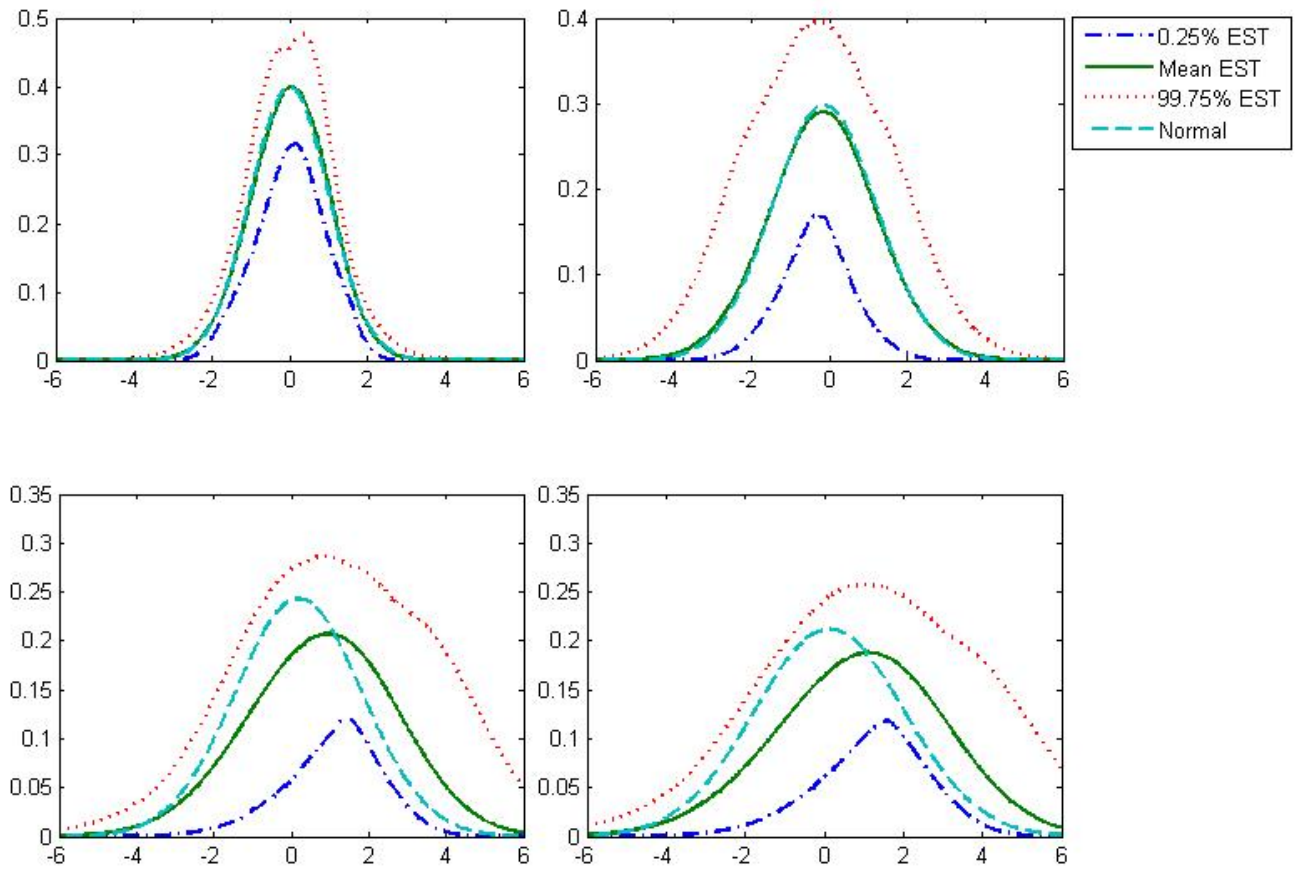


Figure 8: Density estimate of education latent variables

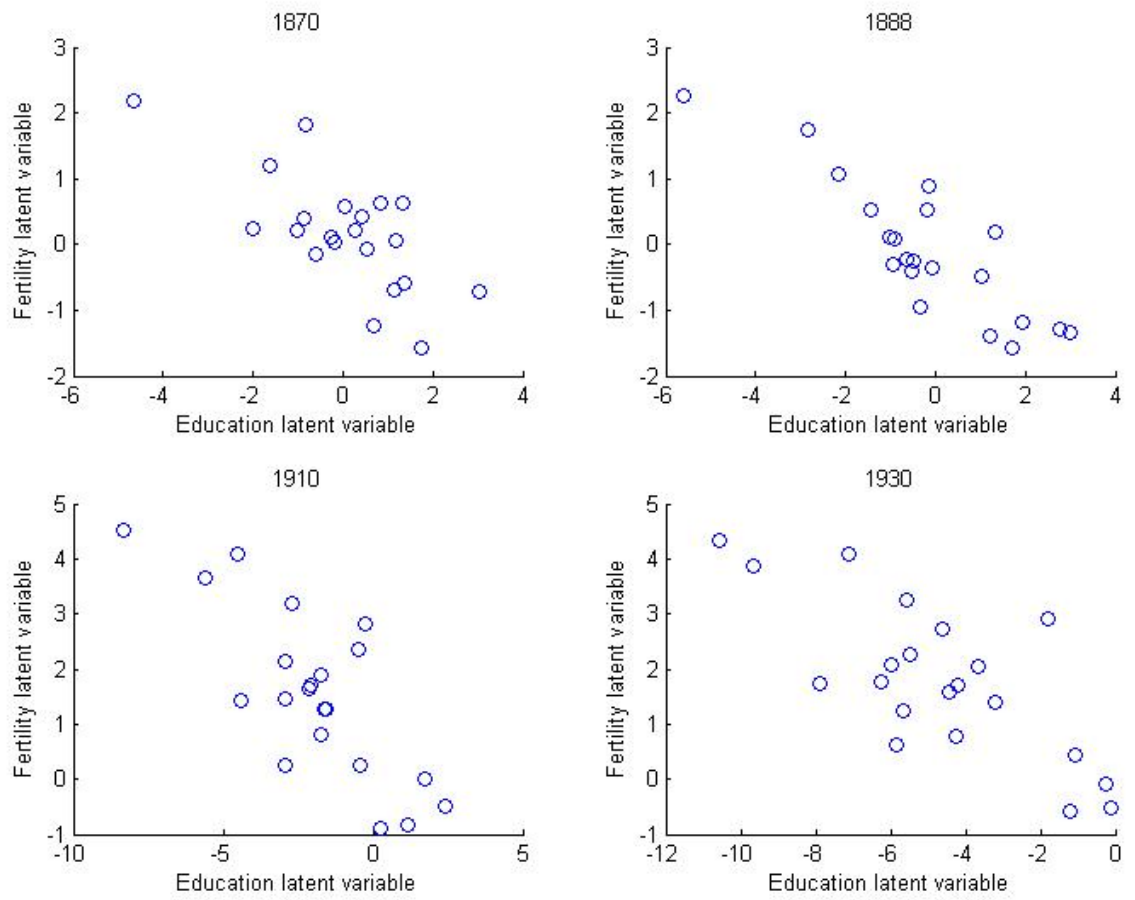


Figure 9: Plot of fertility latent variables Vs. education latent variable across time