Published in final edited form as:

J Am Stat Assoc. 2011 September 1; 106(495): 807–817. doi:10.1198/jasa.2011.ap10058.

Nonparametric Bayes Stochastically Ordered Latent Class Models

Hongxia Yang[Post Doctoral Researcher],

Mathematical Sciences Department, Watson Research Center, IBM, Yorktown Heights, NY 10598 (yangho@us.ibm.com)

Sean O'Brien[Assistant Professor], and

Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710 (obrie027@mc.duke.edu)

David B. Dunson[Professor]

Department of Statistical Science, Duke University, Durham, NC 27708 (dunson@stat.duke.edu)

Abstract

Latent class models (LCMs) are used increasingly for addressing a broad variety of problems, including sparse modeling of multivariate and longitudinal data, model-based clustering, and flexible inferences on predictor effects. Typical frequentist LCMs require estimation of a single finite number of classes, which does not increase with the sample size, and have a well-known sensitivity to parametric assumptions on the distributions within a class. Bayesian nonparametric methods have been developed to allow an infinite number of classes in the general population, with the number represented in a sample increasing with sample size. In this article, we propose a new nonparametric Bayes model that allows predictors to flexibly impact the allocation to latent classes, while limiting sensitivity to parametric assumptions by allowing class-specific distributions to be unknown subject to a stochastic ordering constraint. An efficient MCMC algorithm is developed for posterior computation. The methods are validated using simulation studies and applied to the problem of ranking medical procedures in terms of the distribution of patient morbidity.

Keywords

Factor analysis; Latent variables; Mixture model; Model-based clustering; Nested Dirichlet process; Order restriction; Random probability measure; Stick breaking

1. INTRODUCTION

Latent class models (LCMs) are routinely used for analysis and interpretation of multivariate data. LCMs comprise an extremely rich class of discrete mixture models, which allow units to be allocated to latent subpopulations or clusters, with the allocation probabilities potentially dependent on predictors. Suppose one collects response data $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})' \in \mathcal{R}^p$ and predictors $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})'$ for subjects $i = 1, \dots, n$. Then, a simple Gaussian LCM model could be specified as

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{k=1}^K \pi_k(\mathbf{x}_i) N_p(\mathbf{y}_i; \mu_k, \Sigma_k), \quad (1)$$

where $\pi_k(\mathbf{x}_i)$ is the probability of allocation to latent class k given predictors $\mathbf{x}_i$, the response data for subjects in class k are normally distributed with mean μ_k and covariance Σ_k , and K is

the number of latent classes. In routine applications of such models, $\pi_k(\mathbf{x}_i)$ is typically specified as a logistic regression model and the EM algorithm is used for maximum likelihood estimation.

There are a number of well-known issues that arise in considering model (1) and related LCMs. First, there is the so-called label ambiguity problem, which results because there is nothing distinguishing class k from k' a priori. The estimates produced by the EM algorithm correspond to a local mode, with an identical likelihood obtained for any permutation of the labels $\{1, \dots, K\}$ on the K clusters. Label ambiguity is even more of a problem in Bayesian analyses of LCMs relying on Markov chain Monte Carlo (MCMC) for posterior computation, as label switching makes it difficult to obtain meaningful posterior summaries of the cluster-specific parameters from the MCMC output, though postprocessing can potentially be used (Stephens 2000; Jasra, Holmes, and Stephens 2005). Although constraints on the component-specific parameters, such as ordered means, are widely used to avoid label ambiguity, it is typically not clear what constraints are appropriate in multivariate models such as (1) and partial ambiguity may remain even with constraints. A second well-known issue is uncertainty in the choice of K . Although standard analyses rely on selection criteria, such as the BIC, the theoretical justification for use of the BIC in mixture models such as LCMs is unclear. In addition, conditioning on a selected value in a two-stage procedure clearly ignores uncertainty in the selection process. A third issue with LCMs is sensitivity to parametric assumptions, with a very different number of clusters and allocation to clusters potentially obtained if one replaces the normality assumption in (1) with a multivariate t distribution or other choice.

Our motivation is drawn from an application to ranking of medical procedures in terms of the distribution of patient morbidity following the procedure. In particular, we would like to obtain clusters (latent classes) of procedures having a similar morbidity distribution, while also estimating an ordering in severity of the procedures. Ideally, we would like to avoid some of the problems arising in typical LCMs through stochastic ordering restrictions that are natural in many applications, with nonparametric Bayes methods used to allow infinitely many classes and avoid parametric assumptions on the class-specific distributions. We will focus on the setting in which subjects are nested within prespecified groups, with $i = 1, \dots, n$ indexing the groups and $j = 1, \dots, n_i$ the subjects in the i th group. In the motivating application, groups correspond to different medical procedures.

For illustration, initially consider the case in which y_{ij} is a single outcome for subject j in group i , there are no predictors, and we let $y_{ij} \sim F_i$, with F_i the distribution specific to group i . Then, taking a nonparametric Bayes approach, we require a prior for the collection of distributions $\{F_i\}_{i=1}^n$. Two possibilities that have been proposed in the literature include hierarchical Dirichlet process (HDP) (Teh et al. 2006) and nested Dirichlet process (nDP) mixtures (Rodriguex, Dunson, and Gelfand 2008). The HDP specification automatically allocates patients to clusters, with dependence incorporated in the cluster weights across the groups. The nDP is more relevant in clustering groups, with each cluster having a different distribution of subject-level outcomes. Specifically, the nDP mixture model would let $F_i(y) = F_i(y)$ with prior probability $1/(1 + \alpha)$, with α a precision parameter. The densities specific to each cluster are then modeled using separate DP mixture models.

This approach partly addresses our interests in allowing clustering of procedures based on the distribution of patient outcomes, while allowing the number of clusters (latent classes) to be unknown. However, there is no allowance for predictors that provide information about the cluster allocation and there is no natural way to obtain a ranking of the procedures. Potentially, one may rank the procedures based on the mean of F_i , but it is not clear that the mean is the best summary to rank on, as the proportion of subjects having extreme or life-

threatening adverse events may be more clinically relevant. With this motivation, we propose a nonparametric Bayes stochastically ordered LCM (SO-LCM) that is inspired by the nDP but has a fundamentally different structure.

Section 2 proposes the basic structure of the SO-LCM, with considerations of properties and extensions to more complex hierarchical models motivated in particular by the ranking of medical procedures application. Section 3 outlines an MCMC algorithm for posterior computation. Section 4 contains a simulation study assessing operating characteristics under a default prior. Section 5 applies the method to the medical procedures data, showing advantages relative to parametric methods, and Section 6 contains a discussion.

2. STOCHASTICALLY ORDERED LATENT CLASS PRIORS

2.1 Basic Formulation and Properties

Consider a collection of unknown distributions $P = \{P_1, \dots, P_n\}$, with $P \sim \mathcal{P}$ where $\mathcal{P}$ is a prior. In particular, the prior $\mathcal{P}$ is induced by letting,

$$P_i \sim \sum_{k=1}^{\infty} \pi_k(\mathbf{x}_i) \delta_{P_k^*}, \quad P_k^* = \sum_{l=1}^{\infty} v_l \delta_{\theta_{kl}}, \quad (2)$$

where $\pi_k(\mathbf{x}_i) = \Pr(P_i = P_k^* | \mathbf{x}_i)$ is the conditional probability of allocating distribution i to cluster k given predictors $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})'$, and each of the cluster-specific distributions is assumed to be discrete. In particular, the distribution P_k^* specific to cluster k has probability weights $\{v_l\}$ on atoms $\{\theta_{kl}\}$. This discreteness assumption will be relaxed later by using P_i as a mixture distribution within a continuous kernel. Note that instead of setting P_i equal to a convex combination of the P_k^* 's, related to the approaches of Dunson, Pillai, and Park (2007) and Müller, Quintana, and Rosner (2004), P_i is set exactly equal to P_k^* with probability $\pi_k(\mathbf{x}_i)$.

There are two main distinct features of prior (2) relative to the nested Dirichlet process. First, we allow covariates to impact the allocation to clusters. In the motivating application to ranking of medical procedures, this is an important modification, as we have preliminary rankings of the different procedures by physicians. These rankings can serve as a predictor informing the allocation to clusters. Hence, instead of simply relying on the preliminary physician rankings or the outcomes data in isolation, we allow for a combination or fusion of these data in ranking the procedures. Second, as we are interested in ranking the procedures, we impose a stochastic ordering restriction on the cluster-specific distributions with $P_k^* \leq P_{k'}^*$ for all $k < k'$, where $P_k^* \leq P_{k'}^*$ denotes that P_k^* is stochastically no larger than $P_{k'}^*$ so that $P_k^*(a, \infty) \leq P_{k'}^*(a, \infty)$ for all a . This restriction implies that clusters with a higher index correspond to stochastically higher distributions.

Dunson and Peddada (2008) proposed a restricted dependent Dirichlet process (rDDP) prior for stochastically ordered distributions. Here, we apply the rDDP prior to the cluster-specific distributions $P^* = \{P_k^*\}_{k=1}^{\infty}$. We could have instead used an alternative stochastically ordered prior, such as the approaches proposed by Karabatsos and Walker (2007). Jara and Hanson (2011) proposed a class of mixtures of dependent tail-free processes, with particular applications to the mixture of Polya trees model, allowing each beta random variable in the tree to be covariate-dependent via a regression model. Potentially one can combine this approach with that of Karabatsos and Walker (2007) to obtain a new class of nonparametric Bayes stochastic ordering models. We used the rDDP mixture prior instead to avoid the partitioning effect of the Pólya tree prior. Such an effect can be removed using mixtures of

Pólya trees, though the computation can be more intensive for such models and the results still tend to be quite spiky looking densities. The estimates produced in DP mixtures of Gaussian kernels in our experience tend to match our prior beliefs for the latent variable density more closely.

The stochastic ordering prior from the rDDP is accomplished by first letting $v_l = v_l \prod_{s < l} (1 - v_s)$ with $v_l \sim \text{beta}(1, \alpha_2)$ independently for $l = 1, \dots, \infty$. Then, we let $\theta_l = \{\theta_{kl}\}_{k=1}^{\infty} \sim P_0$ independently for $l = 1, \dots, \infty$, with P_0 chosen so that $P_0(\theta_{1l} \leq \theta_{2l} \leq \dots) = 1$. The cluster k distribution, P_k^* , is marginally distributed according to a Dirichlet process prior with precision α_2 and base distribution P_{0k} , with P_{0k} the k th marginal distribution of P_0 . This implies that $\theta_{kl} \sim P_{0k}$ marginally, where θ_{kl} is the k th element of the multivariate vector θ_l . In addition, $\Pr(P_k^* \leq P_{k'}^*) = 1$ for all $k < k'$ a priori (and hence a posteriori). Dependence in the elements of P^* is incorporated through the use of fixed weights $\{v_l\}_{l=1}^{\infty}$ for all k and dependent atoms. This dependence structure allows flexible borrowing of information across the cluster-specific distributions.

As a specific choice of P_0 , let $\theta_{1l} = \gamma_{1l}^* \sim N(m_0, s_0^2)$ and $\gamma_{kl}^* = \theta_{kl} - \theta_{k-1,l}$, for $k = 2, \dots, \infty$, with

$$\gamma_{kl}^* \sim w_0 \delta_0 + (1 - w_0) N^+(0, \kappa^{-1}), \quad k=2, \dots, \infty, \quad (3)$$

where $w_0 = \Pr(\gamma_{kl}^* = 0)$ and N^+ denotes the normal distribution truncated to have positive support. By including positive mass at zero, the prior allows a subset of the atoms in P_k and $P_{k'}$ to be identical. This is appealing in allowing commonalities between the distributions specific to different latent classes. Also including a positive probability of zero values allows collapsing on an effectively lower-dimensional model through zeroing out the coefficients. This allows us to start with a very richly parameterized model and adaptively drop out parameters that are not needed. To allow the data to inform about the appropriate value for the point mass probability w_0 , we choose a hyperprior $w_0 \sim \text{beta}(a_{w_0}, b_{w_0})$, with $a_{w_0} = b_{w_0} = 1$ used routinely as a default.

To complete a specification of the SO-LCM, we require a prior for the predictor-dependent probabilities. For simplicity, we use the logistic regression-type model

$$\pi_k(\mathbf{x}) = \frac{\psi_k \exp\{\mathbf{x}'\beta_k\}}{\sum_{l=1}^K \psi_l \exp\{\mathbf{x}'\beta_l\}}, \quad (4)$$

$$\psi_k \sim \text{Gamma}(\alpha_1/K, 1), \quad \beta_k \sim H,$$

where $\psi_k \geq 0$ is a baseline weight for mixture component k , β_k are regression parameters controlling the impact of the predictors on the probabilities of allocation to each cluster (latent class), and H is a prior on the regression coefficients. For example, H can be chosen to be Gaussian or, to allow shrinkage towards zero for unimportant coefficients, we can choose a heavy-tailed Cauchy prior or a variable selection mixture prior with a mass at zero.

Unlike in typical generalized logistic regression models, we avoid placing identifiability constraints on the parameters, such as setting the coefficients equal to zero in a reference class. Unlike in frequentist models fitted by maximum likelihood, the choice of the reference class can impact the results, and it is important to maintain exchangeability of the cluster indices in model (4). Otherwise, there may be some bias introduced in which we favor stochastically smaller or larger distributions a priori. In Bayesian modeling, it is not necessary to satisfy frequentist identifiability criteria, and indeed it is often quite useful to consider over-parameterized models as long as inferences are based on identifiable quantities.

To further motivate models (2)–(4), it is useful to consider properties in the baseline case in which $\mathbf{x} = \mathbf{0}$, so that we obtain $\pi_k = \psi_k / \sum_{l=1}^K \psi_l$. In this case, the particular gamma prior that was chosen for the cluster-specific weight parameters leads to $(\pi_1, \dots, \pi_K) \sim \text{Dir}(\alpha_1/K, \dots, \alpha_1/K)$. This is the same distribution on the cluster-specific probabilities that was proposed by Ishwaran and Zarepour (2002) in developing a finite approximation to the Dirichlet process. It is straightforward to show (proof in Appendix) that the prior probability of clustering two groups in this baseline case is

$$\Pr(P_i = P_{i'}) = E \left(\sum_{k=1}^K \pi_k \pi_k \right) = \frac{1 + \alpha_1/K}{1 + \alpha_1}, \quad (5)$$

which simplifies to $1/(1 + \alpha_1)$ in the limit as $K \rightarrow \infty$. In addition, the prior probability that group i is stochastically less than group i' can be derived as

$$\Pr(P_i < P_{i'}) = \frac{1}{2} \{1 - \Pr(P_i = P_{i'})\} = \frac{\alpha_1}{2(1 + \alpha_1)} \left(1 - \frac{1}{K}\right), \quad (6)$$

which reduces to $\frac{\alpha_1}{2(1 + \alpha_1)}$ in the limit as $K \rightarrow \infty$.

Hence, α_1 is a key hyperparameter controlling the prior on clustering and ordering of the groups. For greater flexibility, we recommend letting $\alpha_1 \sim \text{Gamma}(a_1, b_1)$. In many applications, it is appealing to favor a slow rate of introduction of new clusters with sample size. As in the DP, clusters are introduced at a rate proportion to $\alpha_1 \log n$ when K is sufficiently large. In order to favor few clusters relative to the number of groups n , one can choose the hyperparameters a_1, b_1 so that the prior is concentrated at values close to zero. In the application to ranking of medical procedures in terms of their severity, our physician collaborators have a strong preference for parsimony and expect a model with six (or fewer) clusters to fit the data adequately. This knowledge is used to elicit the a_1, b_1 hyperparameters. In the case in which covariates are included, (5) and (6) can potentially be extended, and it will be the case that prior clustering and ordering probabilities depend on the relative values of the predictors for the two groups. However, it is not straightforward to obtain simple analytic forms.

2.2 Applications to Ranking Medical Procedures

In the motivating application to ranking medical procedures based on the distribution of patient morbidity following each procedure, response data consist of a vector

$\mathbf{y}_{ij}^* = (y_{ij1}^*, \dots, y_{ijp}^*)'$ of p measures of morbidity on the j th patient having procedure i , for $i = 1, \dots, n$ and $j = 1, \dots, n_i$. The first p_1 elements of $\mathbf{y}_{ij}^*$ are continuous and the next p_2 elements are binary with $p_1 + p_2 = p$. Higher values of each of the measurements imply higher morbidity, and we relate the measurements to a latent morbidity score for each patient within each procedure through the following factor model,

$$\begin{aligned} y_{ijt}^* &= h_t(y_{ijt}), \quad h_t(y) = y, \quad t = 1, \dots, p_1, \\ h_t(y) &= 1(y > 0), \quad t = p_1 + 1, \dots, p, \\ \mathbf{y}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad \boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{ij}^2), \\ \boldsymbol{\eta}_{ij} &\sim f_i, \quad f_i(\boldsymbol{\eta}) = \int \mathcal{K}(\boldsymbol{\eta}; \boldsymbol{\theta}) dP_i(\boldsymbol{\theta}), \\ \mathcal{K}(\boldsymbol{\eta}; \boldsymbol{\theta}) &= \int N(\boldsymbol{\eta}; \boldsymbol{\theta}, \boldsymbol{\varphi}) dQ(\boldsymbol{\varphi}), \end{aligned} \quad (7)$$

where y_{ijt} is a continuous variable underlying y_{ijt}^* , with $y_{ijt}^* = y_{ijt}$ for continuous responses and $y_{ijt}^* = 1(y_{ijt} > 0)$ for binary responses, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ is a $p \times 1$ intercept vector, $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_p)'$ is a $p \times 1$ vector of factor loadings, η_{ij} is a latent morbidity score for the j th patient having procedure i and $\mathcal{K}; \theta$ is an unknown unimodal kernel that is symmetric about θ . The procedure-specific latent variable density functions f_i are modeled as a flexible location mixture of scale mixture of Gaussian kernels. By using an unknown kernel, we favor fewer and more biologically interpretable clusters. Letting $\sigma_{it}^{-2} = c_i d_t$ for continuous responses, we obtain an additive log-linear model for the residual precision, with c_i a procedure-specific multiple and d_t a response type specific multiple, while fixing $\sigma_{it}^{-2} = 1$ for binary responses. This allows the residual variance to change for the different procedures, while also allowing a shift specific to each measure of morbidity. The constraint on the residual variances for the continuous variables underlying the binary responses is a standard identifiability condition. Because higher values of y_{ijt}^* imply higher morbidity, we constrain the factor loadings to be nonnegative so that $\lambda_t \geq 0$ for $t = 1, \dots, p$. For the scale mixture component, we let $Q \sim \text{DP}(\alpha_0 Q_0)$ where $Q_0 = \text{Inv-Gamma}(c_0, d_0)$ is the base measure. Q can also be denoted as $Q(\cdot) = \sum_{t=1}^{\infty} u_t \delta_{\varphi_t^*}$ with $\varphi_t^* \sim \text{Inv-Gamma}(c_0, d_0)$.

We avoid using P_i directly as the distribution of the latent factor scores within procedure i , since that would assume that the factor scores follow a discrete distribution. It seems more biologically realistic to allow a continuum of patient morbidity, while allowing patients with similar but not identical morbidity to be clustered. This is accomplished by the proposed model in that patients allocated to the same mixture component will be clustered. As mentioned above, we are more interested in clustering and ranking of the medical procedures instead of the patients. Because $\mathcal{K}; \theta$ is monotonically stochastically increasing in θ , we maintain the stochastic ordering restriction in the P_i 's. Note that two procedures i and i' having $P_i = P_{i'}$, which is allowed by the proposed prior, will also have $f_i = f_{i'}$ and hence have the same morbidity density. In addition, $f_i < f_{i'}$ (the distribution of patient morbidity under procedure i is stochastically less than that under procedure i') if and only if $P_i < P_{i'}$. Hence, the clustering and ranking properties of the prior for $\{P_i\}$ proposed above extend directly to the continuous latent factor model in (7).

To complete a Bayesian specification of the SO-LCM model in (7), we choose priors as follows. The intercept vector is assigned a normal prior, $\mu_t \sim N(\mu_0, \sigma_0^2)$ for $t = 1, \dots, p$, and the factor loadings are assigned robust truncated Cauchy priors by letting $\lambda_t \sim N^+(0, \tau)$ for $t = 1, \dots, p$ with $\tau \sim \text{Inv-Gamma}(1/2, 1/2)$. We use a common precision τ to induce dependent shrinkage across the loadings. The multiplicative terms in the variance model, $\{c_i\}$ and $\{d_t\}$, are assigned gamma priors. Elicitation of the different hyperparameters in these priors is considered later.

3. POSTERIOR COMPUTATION

Due to the structure of the model described in Section 2.1, it becomes straightforward to adapt previously proposed algorithms for posterior computation in DPMs and logistic regression models. We will focus on the exact block Gibbs sampler (Yau et al. 2011) for posterior computation and update polychotomous weights through Holmes and Held (2006). We will focus on the simple model $y_{ij} \sim f_i$, $f_i(\eta) = \int \mathcal{K}(\eta; \theta) dP_i(\theta)$, $\mathcal{K}(\eta; \theta) = \int N(\eta; \theta, \varphi) dQ(\varphi)$, $\{P_i\} \sim \text{SO-LCM}$, as in expression (2) to (4), and $Q \sim \text{DP}(\alpha_0 Q_0)$. Denote the procedure location cluster index by ζ_i , the patient location cluster index by ξ_{ij} and the precision cluster index by ε_{ij} . In the sequel, let $\zeta_i = k$, $\xi_{ij} = l$, and $\varepsilon_{ij} = t$ iff $P_i = P_k^*$, $\theta_{ij} = \theta_{kl}^*$, and $\varphi_{ij} = \varphi_t^*$. The sampling steps are as follows:

1. Sample $\pi_k(\mathbf{x}_i)$ through the following steps. The polychotomous generalization of the logistic regression model is defined via

$$\zeta_i \sim \mathcal{M}(1; \pi_1(\mathbf{x}_i), \dots, \pi_K(\mathbf{x}_i)),$$

$$\pi_k(\mathbf{x}_i) = \frac{\psi_k \exp(\mathbf{x}_i' \beta_k)}{\sum_{l=1}^K \psi_l \exp(\mathbf{x}_i' \beta_l)}, \quad (8)$$

where ζ_i is the procedure cluster indicator and $\mathcal{M}(1; \cdot)$ denotes the single sample multinomial distribution. Defining $\beta_{[k]} = (\beta_1, \dots, \beta_{k-1}, \beta_{k+1}, \dots, \beta_K)$, we have

$$L(\beta_k | \zeta, \beta_{[k]}) \propto \prod_{i=1}^n \prod_{k=1}^K [\pi_k(\mathbf{x}_i)]^{1(\zeta_i=j)} \propto \prod_{i=1}^n [\chi_{ij}]^{1(\zeta_i=j)} [1 - \chi_{ij}]^{1(\zeta_i \neq j)}, \quad (9)$$

where

$$\chi_{ij} = \frac{\exp(\mathbf{x}_i' \beta_j + \log(\psi_j) - C_{ij})}{1 + \exp(\mathbf{x}_i' \beta_j + \log(\psi_j) - C_{ij})} = \frac{\exp(\hat{\mathbf{x}}_i' \hat{\beta}_j - C_{ij})}{1 + \exp(\hat{\mathbf{x}}_i' \hat{\beta}_j - C_{ij})},$$

$$C_{ij} = \log \left[\sum_{k \neq j} \exp\{\mathbf{x}_i' \beta_k + \log(\psi_k)\} \right] = \log \left[\sum_{k \neq j} \exp(\hat{\mathbf{x}}_i' \hat{\beta}_k) \right], \quad (10)$$

where $\hat{\mathbf{x}}_i = (\mathbf{x}_i', 1)'$ and $\hat{\beta}_j = (\beta_j', \log \psi_j)'$. The prior for $\log(\psi_k)$ from (4) can be approximated as $\log(\psi_k) \stackrel{\mathcal{D}}{\approx} N(m, v)$. We use this approximation to obtain an efficient Metropolis independence chain proposal. The conditional likelihood $L(\beta_j | \zeta, \beta_{[j]})$ has the form of a logistic regression on class indicator $1(\zeta_i = j)$, which allows us to use the algorithm of Holmes and Held (2006). Details are in the Appendix.

2. Sample the procedure cluster indicators ζ_i , for $i = 1, \dots, n$, from a multinomial distribution with probabilities

$$\Pr(\zeta_i = k | \dots) \propto \pi_k(\mathbf{x}_i) \prod_{j=1}^{n_i} \sum_{l=1}^L v_l N(y_{ij}; \theta_{kl}, \varphi_{ij})$$

and construct $m_k = \sum_{i=1}^n 1(\zeta_i = k)$. For K , we first choose a reasonable upper bound and then monitor the maximum index of the occupied components. If all the MCMC samples have maximum indices several units below the upper bound, then the upper bound is sufficiently high, while otherwise the upper bound can be increased, with the analysis rerun.

3. The joint prior distribution of the group indicator ξ_{ij} and a latent variable q_{ij} can be written as

$$f(\xi_{ij}, q_{ij} | v) = \sum_{l: v_l > q_{ij}} \delta_l(\cdot) = \sum_{l=1}^{\infty} 1(q_{ij} < v_l) \delta_l(\cdot).$$

Implement the Exact Block Gibbs sampler steps:

- i. Sample $q_{ij} \sim \text{Unif}(0, v_{\xi_{ij}})$ for $j = 1, \dots, n_i$ with $v_l = v_l \prod_{s < l} (1 - v_s)$.
- ii. Sample the stick-breaking random variables

$$v_l \sim \text{beta} \left(1 + \sum_{k=1}^K n_{kl}, \alpha_2 + \sum_{s=l+1}^L \sum_{k=1}^K n_{ks} \right), l=1, \dots, L,$$

where n_{kl} is the number of observations assigned to atom l of distribution k with $n_{kl} = \sum_{i=1}^n \sum_{j=1}^{n_i} 1(\zeta_i = k, \xi_{ij} = l)$ and L the minimum value satisfying $v_1 + \dots + v_L > 1 - \min\{q_{ij}\}$.

- iii. Sample ξ_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, n_i$ from the multinomial conditional with

$$\Pr(\xi_{ij} = l) \propto 1(q_{ij} < v_l) N(y_{ij}; \theta_{\zeta_{il}}^*, \varphi_{ij}).$$

4. Sample γ_l^* from

$$p(\gamma_l^* | \dots) \propto \left\{ \prod_{\{(i,j): \xi_{ij}=l\}} N(y_{ij}; \theta_{\zeta_{il}}^*, \varphi_{ij}) \right\} P_0(\gamma_l^*),$$

where $\theta_{kl} = w_k' \gamma_l^*$, $w_k = (1_k', \mathbf{0}_{K-k}')'$, $\gamma_l^* = (\gamma_{1l}^*, \dots, \gamma_{Kl}^*)$ as defined in (3).

5. φ_{ij} is updated through the exact blocked Gibbs sampler similar to the above steps.
6. Use random walk Metropolis–Hastings method to update concentration parameter α_1 .
7. Sample concentration parameter α_2 with conjugate prior $\text{Gamma}(\alpha_2, b_2)$ directly from

$$(\alpha_2 | \dots) \sim \text{Gamma} \left(\alpha_2 + K(L-1), b_2 - K \sum_{l=1}^{L-1} \log(1 - v_l) \right).$$

α_0 is updated similarly.

Note that this algorithm can be generalized easily to accommodate model (7), so the details are omitted.

4. SIMULATION STUDY

We separate this section into two parts. Predictors are not considered in the first simulation but will be considered in the second simulation. Model (7) is studied and both simulations mimic the structure of the medical procedure data.

4.1 Without Predictors

Data $\mathbf{y}_{ij}$ are generated according to (7), with one continuous response ($p_1 = 1$) and six binary responses ($p_2 = 6$) for each of 100 patients ($j = 1, \dots, 100$) in each of 60 procedures ($i = 1, \dots, 60$). Parameters for the data-generating model are $\boldsymbol{\mu} = (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4)'$, $\boldsymbol{\Lambda} = \mathbf{1}_p \times \mathbf{1}_1$, and $\boldsymbol{\Sigma} = \text{diag}(0.5, 1, 1, 1, 1, 1, 1)$. The latent morbidity η_{ij} is generated from one of four mixtures of Gaussian components outlined in Table 1, with the first 15

procedures being generated from mixture distribution T_1 , the second 15 procedures generated from T_2 , the third 15 procedures generated from T_3 and the last 15 procedures generated from T_4 , where $T_1 < T_2 < T_3 < T_4$ such that the generated latent morbidity distributions are stochastically ordered. As shown in Figure 1 and Table 1, distributions share components with each other and the ordering of the distributions is subtle.

To obtain an initial clustering of the medical procedures using standard methods, we first averaged the severity data for the different patients having each procedure to obtain

$\bar{\mathbf{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$ as a $p = 7$ dimensional summary of severity for procedure i . We then applied model-based clustering (Fraley and Raftery 2002; Fraley, Raftery, and Wehrens 2005) to the data $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ using the R functions available in the package described in Fraley, Raftery, and Wehrens (2005). These approaches rely on fitting of finite mixture models with the EM algorithm, with the model fit for a variety of choices of the number of mixture components, which also corresponds to the number of clusters. The BIC is used to select the optimal number of clusters. Figure 2 plots the BIC for the simulated data versus the number of clusters for four different options on the cluster shapes. In this case, the best model according to BIC is EEI (equal size and shape) with six clusters. Note that this approach does not utilize the patient-specific data and instead clusters based on the mean severity measure across patients, while the proposed approach should have advantages in clustering procedures based on the entire distribution across patients.

For the SO-LCM estimation, parameters α_0 , α_1 , and α_2 are fixed to be 1 and a normal inverse-gamma prior distribution, $\text{NIG}(0, 0.1, 2, 3)$ is chosen for the baseline measure m_0 and s_0^2 described in (3), implying that $E(m_0 | s_0^2) = 0$, $V(m_0 | s_0^2) = 10s_0^2$, $E(s_0^2) = 1$, and $V(s_0^2) = 3$. Additionally, we assign priors $w_0 \sim \text{beta}(1, 1)$, $\kappa \sim \text{Gamma}(1/2, 1/2)$, representing a robust and flexible default prior for the base measure P_0 . Without predictors, the prior for the cluster-specific allocation probabilities turns out to be $(\pi_1, \dots, \pi_K)' \sim \text{Dir}(\alpha_1/K, \dots, \alpha_1/K)$. Posterior samples under this SO-LCM prior are obtained through the algorithm described in Section 3 with prespecified truncation bounds $K = 20$. This truncation tends to be accurate for $\alpha_1 \leq 1$, where such values of α_1 favor a small number of mixture components. In this particular application, mixture components close to the upper bound are not occupied in any of the MCMC samples after the burn-in period; 20,000 iterations are found to be enough for parameters to converge. All results are based on 20,000 samples obtained after a burn-in period of 20,000 iterations.

For each pair of distributions P_i and $P_{i'}$, ($i < i'$), the probability $\Pr(P_i = P_{i'})$ was estimated as the proportion of posterior samples for which P_i and $P_{i'}$ are assigned to the same cluster; and $\Pr(P_i < P_{i'})$ is calculated as the proportion of posterior samples for which P_i is assigned to a cluster with stochastically less morbidity than $P_{i'}$. Results are shown in Figure 3, where Figure 3(a) is the ranking plot with the (i, j) th entry of the lower triangular matrix identifying the probability for $P_i < P_{j'}$ and Figure 3(b) is the clustering plot with the (i, j) th entry identifying the probability for $P_i = P_{j'}$. Figure 3 illustrates that there is not enough information in the data to differentiate the first 30 procedures, which is not surprising given the very subtle differences in T_1 and T_2 shown in Figure 1. However, the true rankings and clusterings in the medical procedures are otherwise accurately reflected in the results. The estimated density of T_1 is shown in Figure 4(a). For comparison, this density is also estimated under a DPM prior with the same base measure and precision parameter $\alpha_2 = 1$ [in Figure 4(b)]. The estimate obtained using the SO-LCM prior distribution appears to capture both the small and large modes more accurately than the DPM alternative.

4.2 With Predictors

Potentially, the incorporation of predictors may improve the ability to detect subtle differences in the distributions of patient morbidity between procedures. To assess this, we repeated the simulation study of Section 4.1 but modified the model to allow predictor-dependent mixture weights. Mimicking the real data, we assumed there was a single predictor corresponding to an initial physician severity score obtained from their clinical experience and not from examination of the current data. In particular, similar to range of the potentially useful auxiliary covariate: Aristotle Basic Complexity (ABC) level, we let predictors for the first 15 procedures drawn uniformly from $(-1.5, -1)$, the second 15 procedures drawn uniformly from $(-0.6, -0.4)$, the third 15 procedures drawn uniformly from $(0.4, 0.6)$ and the last 15 procedures drawn uniformly from $(1, 1.5)$. Data are then generated from the assumed model exactly as described in Section 4.1 but assuming a logistic regression model (4) for the weights with $\psi = (0.5, 1, 1.65, 1.45)'$ and $\beta = (-1.5, -0.5, 1, 1.2)'$. Procedures with the first 15 predictors are then assigned to the first cluster and so forth.

In the analysis, priors are specified as described in Section 4.1 and model (4), and we additionally choose a $N(0, 10I)$ prior for β to complete the specification. The MCMC algorithm was run for 20,000 iterations following a 20,000 iteration burn-in. Apparent convergence was rapid and mixing was adequate. The truncation level of $K = 20$ was still sufficiently high.

Figure 5(a) depicts the ranking performance and Figure 5(b) depicts the clustering plot. Both ranking and clustering performances are improved compared to study in Section 4.1 such that the first 30 procedures are ranked consistently with the true order and are clustered correctly.

5. MEDICAL PROCEDURE APPLICATION

The push for accountability in medicine has led to a proliferation of “report cards” evaluating health care providers in various therapeutic areas such as adult and pediatric cardiac surgery, treatment of heart attacks, and management of chronic conditions. In adult cardiac surgery, report cards typically focus on a single commonly performed procedure, coronary artery bypass grafting (CABG), and a single endpoint, short-term all-cause mortality. Regression models are used to adjust for differences in each provider’s case mix that may impact short-term mortality. Although regression models are straightforward to apply in adult cardiac surgery, the development of such models for pediatric and congenital heart surgery is relatively challenging. In congenital heart surgery, the total number of cases is much smaller than CABG—there are literally hundreds of different types of surgical procedures—and no single type of procedure accounts for the majority of cases. To address the challenge of multiple rare procedures, researchers have proposed methods to allow procedures with similar mortality and morbidity risk to be grouped together for analysis. Two widely used methods are the Risk Adjustment for Congenital Heart Surgery (RAHCS-1) methodology (Jenkins 2004) and the Aristotle Basic Complexity Levels (Lacour-Gayet et al. 2004). RACHS-1 groups more than 100 types of congenital heart surgery procedures into six categories based on their estimated risk of in-hospital mortality. Similarly, the Aristotle method groups over 160 types of procedures into four categories (levels) based on their potential for mortality, morbidity, and technical difficulty. For both RACHS-1 and Aristotle, procedure categories were determined by panels of subject matter experts without using a formal statistical framework. In this section, our goal is to show that the SO-LCM methodology provides a useful statistical framework for grouping procedures into categories of risk and for choosing the number of categories. More formally, we sought

to identify clusters of congenital procedures with similar distributions of postprocedural morbidity.

Data for this analysis were obtained from the Society of Thoracic Surgeons (STS) Congenital Heart Surgery database. The study population consisted of $N = 79,635$ patients who underwent one of 145 types of congenital cardiovascular procedures at an STS-participating center during the years 2002–2008. Postoperative morbidity was regarded as a patient-level unobserved latent variable. Indicators of morbidity included a single continuous variable, postoperative length of stay (PLOS), modeled as $y_1 = \log(1 + \text{PLOS})$; and six binary (yes/no) variables: y_2 = renal failure, y_3 = stroke, y_4 = heart block, y_5 = requirement for extracorporeal membrane oxygenation or ventricular assist device, y_6 = phrenic nerve injury, and y_7 = in-hospital mortality. Responses from different patients were assumed to be independent. Multiple responses from the same patient were conditionally independent given the latent morbidity variable. The joint model for all seven endpoints is

$$\begin{aligned} y_{ij1} | \eta_{ij} &\sim N[\mu_1 + \lambda_1 \eta_{ij}, \sigma_{i1}^2] \text{ (PLOS)}, \\ y_{ij2} | \eta_{ij} &\sim \text{Bernoulli}[\Phi(\mu_2 + \lambda_2 \eta_{ij})] \text{ (stroke)}, \\ y_{ij3} | \eta_{ij} &\sim \text{Bernoulli}[\Phi(\mu_3 + \lambda_3 \eta_{ij})] \text{ (renal failure)}, \\ y_{ij4} | \eta_{ij} &\sim \text{Bernoulli}[\Phi(\mu_4 + \lambda_4 \eta_{ij})] \text{ (heart block)}, \\ y_{ij5} | \eta_{ij} &\sim \text{Bernoulli}[\Phi(\mu_5 + \lambda_5 \eta_{ij})] \text{ (ECMO/VAD)}, \\ y_{ij6} | \eta_{ij} &\sim \text{Bernoulli}[\Phi(\mu_6 + \lambda_6 \eta_{ij})] \text{ (phrenic nerve injury)}, \\ y_{ij7} | \eta_{ij} &\sim \text{Bernoulli}[\Phi(\mu_7 + \lambda_7 \eta_{ij})] \text{ (in-hospital mortality)}. \end{aligned}$$

Let f_i denote the density of η_{ij} among patients undergoing the i th type of procedure, that is, $\eta_{ij} \sim f_i$. Our goal is to estimate the densities $f_1, f_2, \dots, f_{145}$ nonparametrically under the assumption that they have an unknown stochastic ordering. This assumption facilitates ranking of the procedures and is less restrictive than alternative parametric models which assume a Gaussian distribution and procedure-specific location parameters. An important consideration for the analysis is that the procedure-specific sample sizes are small and highly variable (median = 50; range 10 to 2000). To account for these low sample sizes, we propose a method of analysis that permits borrowing of information across procedures and incorporates external prior information. A potentially useful auxiliary covariate is each procedure's Aristotle Basic Complexity (ABC) level. As noted above, the ABC level represents the average subjective ranking by an international panel of congenital heart surgeons. Large ABC values imply that the procedure is considered to be a difficult operation with high potential for mortality and morbidity.

The procedure-specific morbidity distributions are estimated nonparametrically under the SO-LCM model as in Section 4.2. Hyperparameters are $\psi_k \sim \text{Gamma}(\alpha_1/K, 1)$, $\beta_k \stackrel{\text{ind}}{\sim} N(0, 10)$, $\alpha_0 \sim \text{Gamma}(1, 1)$, $\alpha_1 \sim \text{Gamma}(1, 6/\ln 145)$, and $\alpha_2 \sim \text{Gamma}(1, 1)$. The prior for α_1 is chosen based on expert elicitation to favor six or fewer clusters in the procedures. The prespecified truncation bound $K = 20$ is found to be enough since mixture components close to the upper bound are not occupied after the algorithm converges. The baseline distribution P_0 is constructed as in (3) with $w_0 \sim \text{beta}(1, 1)$, $\kappa \sim \text{Gamma}(1/2, 1/2)$, and

$(m_0, s_0^2) \sim \text{NIG}(0, 0.1, 2, 3)$. Priors for the outcomes model are

$\mu_t \stackrel{\text{ind}}{\sim} N(0, 10)$ and $\lambda_t \stackrel{\text{ind}}{\sim} N^+(0, \tau)$, $t = 1, 2, \dots, p$, where $\tau \sim \text{Inv-Gamma}(1/2, 1/2)$. Estimates are calculated with 50,000 MCMC iterations after a burn-in period of 20,000 iterations. We took a long burn-in and collection interval to be conservative, but similar results are obtained using shorter chains.

Figures 6 and 7 summarize posterior inferences for procedures ($n = 66$) having at least 200 occurrences in the database. For the i th procedure, let $S_i = \sum_{h=1}^{66} I(f_i \leq f_h)$ denote the number of procedures having a morbidity distribution that is stochastically no smaller than f_i . In Figure 6, the posterior means of each procedure (θ_i) are plotted on the vertical axis along with 95% credible intervals and procedures are sorted in ascending order based on the posterior mean $E[S_i]$. The relatively wide probability intervals indicate that there is uncertainty regarding the true rank ordering of the procedure-specific morbidity distributions. Nonetheless, several procedures have narrow intervals and are clearly distinguished as either low (e.g., atrial septal defect repair) or high (e.g., Norwood operation) latent morbidity. Ranking and clustering performances are depicted in Figure 7(a) and (b).

The 145 procedures can be grouped into four, five, or six homogeneous clusters according to posterior clustering probabilities shown in Table 2. The data suggest high (99%) posterior probability of fewer than eight clusters, with 32% probability assigned to the posterior mode of 5.

Quite often, an optimal point estimate of the ranked clustering is required and we propose the following method which obtains much smaller Bayes risk than the existed popular Aristotle score for ranking congenital heart surgery. For a vector-valued parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, denote the corresponding true ranks as $\mathbf{R} = (R_1, \dots, R_n)$ and $\tilde{\mathbf{R}} = (\tilde{R}_1, \dots, \tilde{R}_n)$ a possible point estimate of $\mathbf{R}$ (we drop the dependency on $\boldsymbol{\theta}$ for notational convenience). To find the optimal ranked clustering, we define the following loss function $L(\mathbf{R}, \tilde{\mathbf{R}})$:

$$L(\mathbf{R}, \tilde{\mathbf{R}}) = \sum_{(i,j) \in \mathcal{M}} \{2 \times 1(\tilde{R}_i < \tilde{R}_j, R_i > R_j) + 2 \times 1(\tilde{R}_i > \tilde{R}_j, R_i < R_j) + 1(\tilde{R}_i = \tilde{R}_j, R_i \neq R_j) + 1(\tilde{R}_i \neq \tilde{R}_j, R_i = R_j)\}.$$

Here $\mathcal{M} = \{(i, j) : i < j; i, j \in \{1, \dots, n\}\}$. We penalize estimates of ties when the true ranks are strictly ordered half as much as estimates in the wrong direction. The posterior expected loss is

$$\begin{aligned} E(L(\mathbf{R}, \tilde{\mathbf{R}}|\mathbf{y})) &= \sum_{(i,j) \in \mathcal{M}} \{2 \times 1(\tilde{R}_i < \tilde{R}_j) \Pr\{R_i > R_j|\mathbf{y}\} + 2 \times 1(\tilde{R}_i > \tilde{R}_j) \Pr\{R_i < R_j|\mathbf{y}\} + 1(\tilde{R}_i = \tilde{R}_j) \Pr\{R_i \neq R_j|\mathbf{y}\} \\ &\quad + 1(\tilde{R}_i \neq \tilde{R}_j) \Pr\{R_i = R_j|\mathbf{y}\}, \end{aligned}$$

where $\Pr\{R_i > R_j|\mathbf{y}\}$, $\Pr\{R_i < R_j|\mathbf{y}\}$, $\Pr\{R_i = R_j|\mathbf{y}\}$, and $\Pr\{R_i \neq R_j|\mathbf{y}\}$ are estimated through the MCMC outputs. As a pragmatic approach to avoid additional computation, we estimate the Bayes risk for each MCMC ranked clustering sample, and choose the sample with the smallest risk as the optimal estimate.

To compare the performance of our method with that of the Aristotle level (with four levels) based on the Bayes risk, we let $K = 4$ so that we will only obtain four clusters. The Bayes risks for the ranked clustering obtained from the Aristotle score is 4900.5. Our optimal Bayesian ranked clustering achieves much smaller risk: 749.8.

We compare groupings based on Aristotle to our final point estimate in Table 3. Several procedures that were predicted to be relatively low risk by the Aristotle score are actually moderate risk or high risk according to our proposed methodology. Among 23 procedures that were Aristotle Category 1 (lowest risk), only 14 of these procedures are assigned to the

lowest risk category according to our method. The correlation between these two ranked clusterings (from the SO-LCM and the Aristotle score) is 0.44. The correlation between the ranked clustering of the SO-LCM and the $\log(1 + \text{PLOS})$, where $\log(1 + \text{PLOS})$ is considered as a critical impact factor for surgery procedure ranking, is 0.81 and the correlation between the ranked clustering of the Aristotle level and the $\log(1 + \text{PLOS})$ is only 0.58.

6. DISCUSSION

We have formulated a novel extension of the nested Dirichlet process (nDP) to the latent class modeling with a partial stochastic ordering that allows us to simultaneously rank and cluster procedures. The procedures are clustered by their entire distribution rather than by particular features of it. We avoid some of the problems arising in typical LCMs through stochastic ordering restrictions and we also relax parametric assumptions on the class-specific distributions through nonparametric Bayes methods. Similar to the nDP, the SO-LCM also allows us to cluster subjects within procedures. The SO-LCM is also straightforward to be imbedded for stochastically ordered mixture distributions within a large hierarchical model.

Although inspired by the pioneering work of the nDP, this article makes several important contributions. First, the stochastically ordered priors that allow covariates to impact the allocation to clusters are developed to apply to nonparametrically estimate densities for multiple procedures subject to a stochastic ordering constraint. In addition, we can also test the hypothesis of equalities between procedures against stochastically ordered alternatives. After examining some of the theoretical properties of the model, we describe a computationally efficient implementation and demonstrate the flexibility of the model through both a simulation study and an application where the SO-LCM is used within a hierarchical model. Heat maps are also offered to summarize the ranking and clustering structures generated by the model.

It is straightforward to make several generalizations of the SO-LCM. One natural generalization is to include hyperparameters in the prior on the regression coefficients of the predictor dependent probabilities H and the baseline measure P_0 . For H , we can choose a heavy-tailed Cauchy prior or a variable selection mixture prior with a mass at zero to shrink unimportant coefficients towards zero. We note that, conditional on P_0 , the distinct atoms $\{P_k^*\}_{k=1}^\infty$ are assumed to be independent. Therefore, including hyperparameters in P_0 allows us to parametrically borrow information across the distinct distributions.

Another natural generalization of the SO-LCM is to replace the $\text{beta}(1, \alpha_2)$ stick-breaking densities with more general forms $\text{beta}(a_k, b_k)$ as considered in Ishwaran and James (2001), with the SO-LCM corresponding to the special case $a_k = 1, b_k = \alpha_2$. Richer classes of priors that encompass the SO-LCM as a particular case will be obtained, though in some regression contexts it does not always outperform the DP model with $\text{beta}(1, \alpha_2)$ in terms of the log-predictive marginal likelihood.

We can also generalize the procedure to incorporate multivariate latent factors whose distributions are stochastically ordered. This generalization is inspired by the valuable suggestion from the editors. In having univariate stochastic ordering on the latent variable level, we actually induce multivariate stochastic ordering for the responses (albeit in a somewhat restrictive manner). To directly place the stochastic ordering constraint on multivariate distributions, we can adopt multivariate monotone functions. In particular, in place of the scalar θ_{kl} we could have a vector $\boldsymbol{\theta}_{kl}$ with P_0 chosen (e.g., multivariate truncated normal) so that the different elements are appropriately ordered to satisfy the constraint. In

the simple ordering case, we could just let (3) independently for each element of the θ_{kl} vector instead of just for the θ_{kl} scalar. We could even have different orders for different variables and could have some variables with no ordering.

APPENDIX

Clustering Probability

Under (2), the probability that $P_i = P_{i'}$ so that groups i and i' are allocated to the same cluster is

$$\begin{aligned}\mathbb{P}(P_i = P_{i'}) &= \mathbb{E} \left(\sum_{k=1}^K \pi_k \pi_k \right) \\ &= \sum_{k=1}^K \mathbb{E}(\pi_k^2) \\ &= \sum_{k=1}^K \text{Var}(\pi_k) \\ &\quad + \mathbb{E}(\pi_k)^2 = \sum_{k=1}^K \frac{\alpha_1/K(\alpha_1 - \alpha_1/K)}{\alpha_1^2(\alpha_1 + 1)} \\ &\quad + \left(\frac{\alpha_1/K}{\alpha_1} \right)^2 \\ &= K \left(\frac{\alpha_1/K(\alpha_1 - \alpha_1/K)}{\alpha_1^2(\alpha_1 + 1)} + \left(\frac{\alpha_1/K}{\alpha_1} \right)^2 \right) \\ &= \frac{\alpha_1 - \alpha_1/K}{\alpha_1(\alpha_1 + 1)} + \frac{1}{K}.\end{aligned}$$

As K goes to infinity,

$$\lim_{K \rightarrow \infty} \frac{\alpha_1 - \alpha_1/K}{\alpha_1(\alpha_1 + 1)} + \frac{1}{K} = \frac{1}{\alpha_1 + 1},$$

MCMC Supplement

We would introduce a set of variables d_{ij} , $i = 1, \dots, p$, $j = 1, \dots, K$, and define $\hat{y}_{ij} = 1$ if the i th observation belongs to class j , $j \in \{1, \dots, K\}$, and $\hat{y}_{ij} = 0$ otherwise. Notice that the equivalent representation of equation (8) is

$$\begin{aligned}\hat{y}_{ij} &= \begin{cases} 1, & s_{ij} \geq 0, \\ 0, & \text{else,} \end{cases} \\ s_{ij} &= \mathbf{x}_{ij}' \hat{\beta}_j - C_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, d_{ij}), \\ d_{ij} &= (2e_{ij})^2, \quad e_{ij} \sim KS, \quad \hat{\beta}_j \sim \pi(\hat{\beta}_j), \\ \lambda_j &\sim \text{Gamma}\left(\frac{\alpha_1}{K}, 1\right).\end{aligned}\tag{A.1}$$

Parameters of equation (A.1) are updated through following steps:

1. Sampling $\hat{\beta}_j$ in the case of a normal prior on β_j , $\pi(\beta_j) = N(b_0, v_0)$, the full conditional distribution of β_j given $\mathbf{s}_j$ and $\mathbf{d}_j$ is still normal,

$$\begin{aligned}\hat{\beta}_j | \mathbf{s}_j, \mathbf{d}_j &\sim \text{MN}(B_j, V_j), \\ B_j &= V_j(v_0^{-1}b_0 + \hat{\mathbf{x}}'W_j(\mathbf{s}_j + C_j)), \\ V_j &= (v_0^{-1} + \hat{\mathbf{x}}'W_j\hat{\mathbf{x}})^{-1}, \quad W_j = \text{diag}(d_{1j}^{-1}, \dots, d_{nj}^{-1}),\end{aligned}$$

where $\hat{\mathbf{x}} = (\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_p)'$, $\mathbf{s}_j = (s_{1j}, \dots, s_{nj})'$, $C_j = (C_{1j}, \dots, C_{nj})'$, and $\mathbf{d}_j = (d_{1j}, \dots, d_{nj})'$.

2. Following advice of Holmes and Held (2006), we update $\{\mathbf{s}_j, \mathbf{d}_j\}$ jointly given $\hat{\beta}_j$,

$$\pi(\mathbf{s}_j, \mathbf{d}_j | \hat{\beta}_j, \hat{\mathbf{y}}_j) = \pi(\mathbf{s}_j | \hat{\beta}_j, \hat{\mathbf{y}}_j) \pi(\mathbf{d}_j | \mathbf{s}_j, \hat{\beta}_j)$$

followed by an update to $\hat{\beta}_j | \mathbf{s}_j, \mathbf{d}_j$. The marginal densities for the s_{ij} 's are independent truncated logistic distributions,

$$s_{ij} | \hat{\beta}_j, \hat{\mathbf{y}}_j \propto \begin{cases} \text{Logistic}(\hat{\mathbf{x}}_i' \hat{\beta}_j + \log(\lambda_j) - C_{ij}, 1) I(s_{ij} > 0), & \hat{\mathbf{y}}_j = 1, \\ \text{Logistic}(\hat{\mathbf{x}}_i' \hat{\beta}_j + \log(\lambda_j) - C_{ij}, 1) I(s_{ij} \leq 0), & \text{else,} \end{cases}$$

where $\text{Logistic}(a, b)$ denotes the density function of the logistic distribution with mean a and scale parameter b .

3. Sampling $\mathbf{d}_j$ through rejection sampling. As advised by Holmes and Held (2006), we use Generalized Inverse Gaussian distribution

$g(d_{ij}) = \text{GIG}(0.5, 1, r^2) = \frac{r}{\text{Inv-Gamma}(1, r)}$, where $r^2 = (s_{ij} - \hat{\mathbf{x}}_i' \hat{\beta}_j + C_{ij})^2$, as rejection sampling density. Following a draw from $g(\cdot)$ the sample is accepted with probability $\alpha(\cdot)$,

$$\alpha(d_{ij}) = \frac{L(d_{ij})\pi(d_{ij})}{Mg(d_{ij})},$$

where $M \geq \sup_{d_{ij}} \frac{L(d_{ij})\pi(d_{ij})}{g(d_{ij})}$, $L(d_{ij})$ denotes the likelihood,

$L(d_{ij}) \propto d_{ij}^{-1/2} \exp(-0.5r^2/d_{ij})$, and $\pi(d_{ij})$ is the prior,

$$\pi(d_{ij}) = \frac{1}{4} d_{ij}^{-1/2} \text{KS}\left(\frac{1}{2} d_{ij}^{1/2}\right),$$

where $\text{KS}(\cdot)$ denotes the Kolmogorov–Smirnov density.

REFERENCES

- Dunson D, Peddada S. Bayesian Nonparametric Inference on Stochastic Ordering. *Biometrika*. 2008; 95:859–874.
- Dunson D, Pillai N, Park J. Bayesian Density Regression. *Journal of the Royal Statistical Society, Ser. B*. 2007; 69:163–183.

- Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*. 2002; 97:611–631.
- Fraley C, Raftery AE, Wehrens R. mclust: Model-Based Cluster Analysis. R package version 2.1-11. 2005
- Holmes C, Held L. Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*. 2006; 1:145–168.
- Ishwaran H, James L. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*. 2001; 96:161–173.
- Ishwaran H, Zarepour M. Exact and Approximate Sum-Representations for the Dirichlet Process. *Canadian Journal of Statistics*. 2002; 30:269–283.
- Jara A, Hanson T. A Class of Mixtures of Dependent Tail-Free Processes. *Biometrika*. 2011 to appear.
- Jasra A, Holmes CC, Stephens DA. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modelling. *Statistical Science*. 2005; 20:50–67.
- Jenkins K. Risk Adjustment for Congenital Heart Surgery: The RACHS-1 Method. *Seminars in Thoracic and Cardiovascular Surgery: Pediatric Cardiac Surgery Annual*. 2004; 7:180–184. [PubMed: 15283367]
- Karabatsos G, Walker S. Bayesian Nonparametric Inference of Stochastically Ordered Distributions, With Polya Trees and Bernstein Polynomials. *Statistics and Probability Letters*. 2007; 77:907–913.
- Lacour-Gayet F, Clarke D, Jacobs J, Comas J, Daebritz S, Daenen W, Gaynor W, Hamilton L, Jacobs M, Maruszewski B, Pozzi M, Spray T, Stellin G, Tchervenkov C, Mavroudis C. The Aristotle Committee. The Aristotle Score: A Complexity-Adjusted Method to Evaluate Surgical Results. *European Journal of Cardio-Thoracic Surgery*. 2004; 25:911–924. [PubMed: 15144988]
- Müller P, Quintana F, Rosner G. A Method for Combining Inference Across Related Nonparametric Bayesian Models. *Journal of the Royal Statistical Society, Ser. B*. 2004; 66:735–749.
- Rodriguez A, Dunson D, Gelfand A. The Nested Dirichlet Process. *Journal of the American Statistical Association*. 2008; 103:1131–1154.
- Stephens M. Dealing With Label Switching in Mixture Models. *Journal of the Royal Statistical Society, Ser. B*. 2000; 62:795–809.
- Teh Y, Jordan M, Beal M, Blei D. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*. 2006; 101:1566–1581.
- Yau C, Papaspiliopoulos O, Roberts G, Holmes C. Nonparametric Hidden Markov Models With Application to the Analysis of Copy-Number-Variation in Mammalian Genomes. *Journal of the Royal Statistical Society, Ser. B*. 2011; 73:37–57.

Figure 1.
True distributions used in simulation study in Section 4.1.

Figure 2.

Frequentist model-based clustering results implemented via the EM algorithm using the `mclust` function in R in simulation study in Section 4.1, with the different symbols representing different model assumptions. EII: spherical, equal volume; EEI: spherical, equal volume, and shape; EEV: spherical, equal volume but varying orientation; EEE: ellipsoidal, equal volume, and shape. Refer to Fraley, Raftery, and Wehrens (2005) for more details. The online version of this figure is in color.

Figure 3. Posterior probability for ranking and clustering in study of Section 4.1 with entry (i, j) in (a) being the lower triangular matrix identifying the probability for $P_i < P_{i'}$ and in (b) the probability for $P_i = P_{i'}$.

Figure 4. True (solid lines) and estimated (dashed lines) densities from SO-LCM and DPM for distribution T_1 .

Figure 5. Posterior probability for ranking and clustering in study of Section 4.2 with entry (i, j) in (a) being the lower triangular matrix identifying the probability for $P_i < P_{i'}$ and in (b) the probability for $P_i = P_{i'}$.

Figure 6.
Sorted procedures posterior means for latent scores of each procedure with 95% credible intervals.

Figure 7. Ranking and clustering for selected 66 procedures with more than 200 patients, with entry (i, j) in (a) being the lower triangular matrix identifying the probability for $P_i < P_{i'}$ and in (b) the probability for $P_i = P_{i'}$.

Table 1

Parameters for the true distributions used in simulation study in Section 4.1

Dist	Comp1		Comp2		Comp3		Comp4	
	w	(μ, σ^2)	w	(μ, σ^2)	w	(μ, σ^2)	w	(μ, σ^2)
T_1	0.75	$(-3, 0.5)$	0.25	$(0, 1)$				
T_2	0.75	$(-2.5, 0.5)$	0.25	$(0, 1)$				
T_3	0.2	$(1, 0.5)$	0.5	$(1.5, 1)$	0.3	$(2, 1)$		
T_4	0.2	$(2, 1)$	0.3	$(2.5, 0.5)$	0.4	$(3, 1)$	0.1	$(3.5, 0.5)$

Table 2

Posterior probability for clustering (epidemiology application)

Number of clusters	Posterior probability
1	0.01
2	0.02
3	0.12
4	0.21
5	0.32
6	0.25
7	0.06
8	0.01

Table 3

Clustering comparison between Aristotle level and SO-LCM

SO-LCM	Aristotle			
	Level 1	Level 2	Level 3	Level 4
Level 1	15	18	15	2
Level 2	4	17	28	11
Level 3	4	6	6	11
Level 4	0	0	1	7