

Nonparametric Bayes Models for High-Dimensional and Sparse Data

by

Hongxia Yang

Department of Statistical Science
Duke University

Date: _____

Approved:

David Dunson, Supervisor

David Banks

Fan Li

Sean O'Brien

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2010

ABSTRACT
(Statistical Science)

Nonparametric Bayes Models for High-Dimensional and
Sparse Data

by

Hongxia Yang

Department of Statistical Science
Duke University

Date: _____

Approved:

David Dunson, Supervisor

David Banks

Fan Li

Sean O'Brien

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2010

Copyright © 2010 by Hongxia Yang
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Current research has evolved at a dramatic rate in the past decade, with improvements in technology leading to a fundamental shift in the way in which data are collected and analyzed. It has become routine to collect large amounts of information and it has become necessary to consider new statistical paradigms that perform well in characterizing complex data from a broad variety of problems. We develop novel nonparametric Bayes models for high-dimensional and sparse data in this dissertation. Bayesian nonparametric methods are useful for modeling data without having to define the complexity of the entire model *a priori*, but rather allowing for this complexity determined by the data.

The flexibility of Bayesian nonparametric priors arises from the prior's definition over an infinite dimensional parameter space. Therefore, there are theoretically an infinite number of latent components and an infinite number of latent factors. Nevertheless, draws from each respective prior will produce only a small number of components or factors that appear in a given data set. As mentioned, the number of these components and factors, and their corresponding parameter values, are left for the data to decide. This dissertation is divided into four parts, which motivate novel Bayesian nonparametric methods and clearly illustrate their utilities:

- **Chapter 1:** In Chapter 1, we review the Dirichlet process (DP) in detail. There are many other ways of nonparametric modeling, but with the availability of efficient computation and complete set up of theories, the DP is most popular and has been developed and studied extensively. We will also review the most recent development

of the DP in this chapter.

- **Chapter 2:** We propose the multiple Bayesian elastic net (abbreviated as MBEN), a new regularization and variable selection method. High dimensional and highly correlated data are commonplace. In such situations, maximum likelihood procedures typically fail—their estimates are unstable, and have large variance. To address this problem, a number of shrinkage methods have been proposed, including ridge regression, the lasso and the elastic net; these methods encourage coefficients to be near zero (in fact, the lasso and the elastic net perform variable selection by forcing some regression coefficients to equal zero). In this paper we describe a semiparametric approach that allows shrinkage to multiple locations, where the location and scale parameters are assigned Dirichlet process hyperpriors. The MBEN prior encourages variables to cluster, so that strongly correlated predictors tend to be in or out of the model together. We apply the MBEN prior to a multi-task learning (MTL) problem, using text data from the Wikipedia. An efficient MCMC algorithm and an automated Monte Carlo EM algorithm enable fast computation in high dimensions. The methods are applied to Wikipedia data using shared words to predict article links.
- **Chapter 3:** 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.

- **Chapter 4:** In studies involving multi-level data structures, problems of data sparsity are often encountered and it becomes necessary to borrow information to improve inferences and predictions. This article is motivated by studies collecting data on different outcomes following congenital heart surgery. If there were sufficient numbers of patients receiving each type of procedure, one could potentially fit procedure-specific multivariate random effects model to relate the outcomes of surgery to patient predictors while allowing variability among hospitals. However, as there are approximately 150 procedures with many procedures conducted on few patients, it is important to borrow information. Allowing variability among hospitals, procedures and outcome types in the regression coefficients relating patient factors to outcomes, we obtain a three-way tensor of regression coefficient vectors. To borrow information in estimating these coefficients, we propose a Bayesian multi-way tensor co-clustering model. In particular, the model works by reducing the dimension of the table through separately clustering hospitals, procedures and outcome types. This soft probabilistic clustering proceeds via nonparametric Bayesian latent class models, which favor clustering of dimensions that have similar values for feature vectors. Efficient MCMC and fast approximation approaches are proposed for posterior computation. The methods are illustrated using simulated data, and applied to heart surgery outcome data from a Duke study.

To My Family

Contents

Abstract	iv
List of Tables	xi
List of Figures	xii
Acknowledgements	xiv
1 Introduction	1
1.1 The Dirichlet Process	1
1.1.1 The definition of Dirichlet Process	1
1.1.2 The Dirichlet Process Properties	3
1.2 Dirichlet Process Mixtures	5
1.3 Bayesian Nonparametric Regression using Dirichlet Process Mixtures	7
1.3.1 Dependent Dirichlet Process and its Extensions	7
1.3.2 Bayesian Nonparametric Inference on Stochastic Ordering	9
1.3.3 Hierarchical Dirichlet Process and the Nested Dirichlet Process	10
1.4 Inference of Dirichlet Process Mixture Modeling	12
1.4.1 Blocked Gibbs Sampler	12
1.4.2 Exact Block Gibbs Sampler	13
1.4.3 Pólya Urn Gibbs Sampling	14
1.5 Other Nonparametric Bayesian Modeling Processes	15
1.5.1 Chinese Restaurant Process	15

1.5.2	Indian Buffet Process	17
1.5.3	Pólya Tree Process	18
2	Multiple Bayesian Elastic Net	21
2.1	Introduction	21
2.2	Multiple Bayesian elastic net Prior	26
2.3	Posterior Computation	29
2.3.1	Weakly Informative Prior Specification	29
2.3.2	Gibbs Sampling Algorithm	30
2.4	Sparse Point Estimation via Automated Monte Carlo EM	33
2.5	Simulation Study	37
2.6	Wikipedia Project Application	41
2.7	Discussion	45
3	Nonparametric Bayes Stochastically Ordered Latent Class Models	53
3.1	Introduction	53
3.2	Stochastically Ordered Latent Class Priors	56
3.2.1	Basic Formulation and Properties	56
3.2.2	Applications to Ranking Medical Procedures	59
3.3	Posterior Computation	61
3.4	Simulation Study	64
3.4.1	Without predictors	65
3.4.2	With predictors	68
3.5	Medical Procedure Application	70
3.6	Discussion	77
4	Bayesian tensor co-clustering for flexible multilevel regression modeling	81
4.1	Introduction	81

4.2	Model Specification and Properties	84
4.3	Posterior Computation	90
4.4	Simulation Examples	93
4.5	Application to Congenital Heart Surgery Outcomes Data	95
4.6	Discussion	99
A	Additional Materials for Chapter 3	107
B	Additional Materials for Chapter 4	110
	Bibliography	115
	Biography	121

List of Tables

2.1	Table 2.1: Mean MSPE for the testing simulated examples based on 50 replications	47
2.2	Table 2.2: Single-task Learning Performance on the Central Page Normal Distribution	47
2.3	Table 2.3: Single-task Learning Performance on the Central Page Bayesian Inference	47
2.4	Table 2.4: Multi-task Learning Performance on the Central Page Normal Distribution	47
2.5	Table 2.5: Multi-task Learning Performance on the Central Page Bayesian Inference	47
3.1	Table 3.1: True distribution used in simulation study 4.1	79
3.2	Table 3.2: Posterior Probability for clustering (Epidemiology Application)	79
3.3	Table 3.3: Clustering comparison between Aristotle Level and SO-LCM .	80
4.1	Table 4.1: True distribution used in simulation study 4.1	100
4.2	Table 4.2: Simulation Study 1 (Generate Data from Full Model)	100
4.3	Table 4.3: Simulation Study 2 (Generate Data without Interaction)	100
4.4	Table 4.4: Simulation Study 3 (Generate Data with one factor)	100

List of Figures

1.1	Samples form a DP process with a standard normal distribution as base measure with different precision parameters.	6
1.2	Comparison Between CRP and IBP.	19
2.1	MBEN prior distribution with $\alpha = 1$, $c_1 = 0$ and $d_1 = 10$	48
2.2	Mean MSE for all coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)	48
2.3	Mean MSE for none null coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)	49
2.4	Mean MSE for null coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)	50
2.5	Bias for all coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)	51
2.6	Coefficient estimates from the lasso, elastic net, ridge regression, MBEN (Gibbs) and MBEN (MCEM) in simulation 4.	52
3.1	True distributions used in simulation study 4.1	65
3.2	Frequentist model-based clustering results implemented via the EM algorithm using the Mclust function in R in simulation study 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.	67

3.3	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 $P_i = P_{i'}$	68
3.4	True (solid lines) and estimated (dashed lines) densities from SO-LCM and DPM for distribution T_1	69
3.5	Posterior probability for ranking and clustering in study of section 4.2	71
3.6	Sorted Procedures Posterior Means for latent scores of each procedure with 95% Credible Intervals	74
3.7	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 $P_i = P_{i'}$	75
4.1	A: Multiple, overlapping co-clusters. B: Multiple, non-overlapping co-clusters. C: Single, non-overlapping co-clusters.	101
4.2	Prior Realizations of β_{hp}	101
4.3	Hospital-Procedure specific sample size	102
4.4	Posterior Marginal Likelihood Comparison	103
4.5	Posterior Mean of β_{hp}	104
4.6	Variation explained by hospitals and procedures for β_{hp}	105
4.7	R_h Estimation from Standard Method and the MCCI model	106

Acknowledgements

It is a pleasure to acknowledge with gratitude many people who made this thesis possible.

First of all, I would like to express my deep and sincere gratitude to my Ph.D. advisor David B. Dunson, for his enthusiasm, his inspiration, his patience and his great efforts in supervising me. Throughout my Ph.D study, he provided encouragement, excellent teaching and flows of research guidance. More importantly, his enthusiasm for and commitment to research work set a strong model for me to follow.

I would like to give my special thanks to Dr. David L. Banks and Dr. Mike West for their continuous and inspiring encouragement for me to fulfill my PhD study, for their support for me to visit Europe, Canada and many US cities for conferences and workshops. I would also thank Dr. Alan E. Gelfand for enrolling me to such a great department!

I would like to give thanks to other members of my committee, Dr. Fan Li and Dr. Sean O'Brien. I would like especially to thank Dr. Sean O'Brien to introduce and supervise me over the congenital heart surgery project, which inspires me to fulfill the thesis. I would also greatly appreciate Dr. Fan Li and Dr. Sayan Mukherjee for their great support for my prelim! I am also grateful to Dr. Merlise A. Clyde, Dr. Edwin S. Iversen and Dr. Scott C. Schmidler to give me great help when I started to do my research. I also appreciate Dr. Dalene K. Stangl for her great supervision for me to teach. I also benefitted a lot from a year-long program at SAMSI. I would like to extend my thanks to Dr. Jim O. Berger for organizing such an inspiring working group.

The department of Statistical sciences at Duke university has been very supportive. I

feel fortunate to spend three and a half years in such a stimulating and friendly environment. I've had many colleagues within the department whom I wish to thank. Especially thanks to Zhi Ouyang, Huiyan Sang, Chunlin Ji, Kai Mao, Simon Lunagomez, Avishek Chakraborty and Jarad B. Niemi for their great and patient help for my PhD study. I would like to thank all friends from my year for accompanying me through the FYE, the prelim, job search and defense.

I wish to thank my family for their supports. My special gratitude is due to my my parents and my parents in law. Most of all, I wish to thank my husband, Yi Li, who is now a graduate student at the Math Department, for his endless support and love. To them I dedicate this thesis.

Introduction

1.1 The Dirichlet Process

1.1.1 The definition of Dirichlet Process

The Dirichlet Process (DP) was developed by Ferguson (1973) as a prior probability model for random distributions G . DP models have enjoyed considerable popularity due to the ready availability of posterior simulation techniques, the analytic tractability of almost surely discrete realized probability functions, as well as the theoretical elegance of the model formulations. A $\text{DP}(\alpha G_0)$ prior for G consists of two parts: a parametric base distribution G_0 and a positive scalar parameter α , which can be interpreted as a precision parameter; larger values of α result in realizations G that are closer to G_0 . We write $G \sim \text{DP}(\alpha G_0)$ to indicate that a DP prior is used for the random distribution G . A more general version of the DP prior involves hyperpriors for α and/or parameters of G_0 .

The most commonly used DP definition is its constructive definition by Sethuraman (1994), which characterizes DP realizations as countable mixtures point masses (e.g., random discrete distributions). Specifically, a random distribution G generated from $\text{DP}(\alpha G_0)$

is of the form

$$G(\cdot) = \sum_{k=1}^K w_k \delta_{\theta_k}(\cdot), \quad (1.1)$$

where $\delta_{\theta}(\cdot)$ denotes a point mass at θ . The locations of the point masses, θ_k , are i.i.d. realizations from G_0 . The weights, w_k , are generated from a stick-breaking process: with $\{v_k : k = 1, 2, \dots\}$ drawing from a $\text{beta}(1, \alpha)$ distribution. In particular, $w_1 = v_1$ and $w_l = v_l \prod_{k < l} (1 - v_k)$. Note that the “stick-breaking” terminology arises, because starting with a unit probability stick, v_1 is the proportion of the stick broken off and assigned to θ_1 , v_2 is the proportion of the remaining $1 - v_1$ length stick allocated to θ_2 , and so on. The sequences $\{\theta_l, l = 1, 2, \dots\}$ and $\{v_k, k = 1, 2, \dots\}$ are independent.

For values of α close to zero, $v_1 \approx 1$ and essentially all the probability weight will be assigned to a single atom. For small values of α , such as the common used value $\alpha = 1$, most of the probability is assigned to the first few atoms. While for large values of α , each of the atoms is assigned vanishingly small weight, so that G resembles G_0 . Because the probability weights assigned to the atoms decrease stochastically as the index k grows, we are able to accurately represent realizations from G with only the first several atoms.

The DP can also be characterized by its predictive rule (Blackwell and MacQueen, 1973), which relates the DP to a Pólya urn process and is also the basis for the usual computational tools used to fit models based on the DP. If $(\theta_1, \dots, \theta_n)$ is an iid sample from G and $G \sim \text{DP}(\alpha G_0)$, we can integrate out the unknown G and obtain the following conditional predictive distribution of a new observation

$$\theta_{n+1} | \theta_1, \dots, \theta_n \sim \frac{\alpha}{\alpha + n} G + \sum_{l=1}^n \frac{1}{\alpha + n} \delta_{\theta_l}. \quad (1.2)$$

Exchangeability of the draws ensures that the full conditional distribution of any θ_l has this same form.

Finally, the DP can be alternatively obtained as the asymptotic limit of certain finite mixture models (Ishwaran and Zarepour, 2002). In particular, we consider the following finite dimensional Dirichlet-Multinomial prior

$$\begin{aligned}
 G^K(\cdot) &= \sum_{k=1}^K w_k \delta_{\theta_k}(\cdot), \\
 \mathbf{w} &\sim \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right), \\
 \theta_k &\sim G_0.
 \end{aligned}$$

The above definition differs from the truncated stick-breaking representation of the DP in the way that the weights have been defined.

1.1.2 The Dirichlet Process Properties

The clustering property of the DP has been widely exploited in recent years, since it has some appealing practical properties relative to alternative clustering procedures. In particular, it avoids assuming that all individuals can be clustered into a fixed number of groups, K . Instead, as is again clear from the stick-breaking form in (1.1), the DP assumes that there are infinitely many clusters represented in the overall population, with an unknown number of them observed in a finite sample of n subjects. When an $(n + 1)$ th subject is added, from model (1.2), with the positive probability $\alpha/(\alpha + n)$, the subject is assigned to a new cluster not yet represented in the sample.

In clustering there must be some implicit or explicit penalty for model complexity to avoid assigning everyone in the sample to their own cluster to obtain a higher likelihood. Hence, it is important not to view the DP model as a magic clustering approach, which avoids assuming a fixed number of clusters and specification of an arbitrary penalty. Instead, one must carefully consider how the penalty for model complexity or overfitting arises in the DP implementation, while also assessing the role of the hyperparameters α and parameters characterizing G_0 .

Antoniak (1974) studies the properties of draws from a distribution that follow a DP. In particular, he proves that given G_0 is nonatomic, the probability of $K(1 \leq K \leq n)$ distinct values on a sample $\theta_1, \dots, \theta_n$ of size n is

$$\mathbb{P}(K) = c_n(K)n!\alpha^K \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad (1.3)$$

where $c_n(K)$ is a constant that can be obtained by recurrence formulas for Stirling numbers. The expected number of distinct values can be calculated as

$$E(K|\alpha, n) = \sum_{i=1}^n \frac{\alpha}{\alpha + i - 1} \approx \alpha \log \left(\frac{\alpha + n}{\alpha} \right).$$

So the prior expectation for the number of clusters is proportional to $\alpha \log n$ and the number of clusters tends to increase slowly with the sample size at a rate determined by α .

From the stick-breaking construction we can easily see that draws from a DP are discrete distributions almost surely. It also provides a simple framework to calculate moments of the process. Note that for any measure set $A \in \mathfrak{B}$, $G(A)$ is a random quantity and

$$\mathbb{E}(G(A)) = \sum_{l=1}^{\infty} \mathbb{E}(w_k) E(\delta_{\theta_k}(A)) = G_0(A) \sum_{l=1}^{\infty} E(w_k) = G_0(A).$$

Similarly,

$$\mathbb{V}(P(A)) = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}.$$

We can better understand the role of G_0 and α in Figure 1.1, which shows the approximate simulations of a DP with G_0 a standard normal distribution and different values of the precision parameters α . During the simulations, we truncate the stick-breaking process when the leftover mass was smaller than 10^{-6} . As shown from the plots, for small values of α , the sampled distributions are widely around the baseline measure and most of the probability is allocated to the first few atoms. While as α getting larger, each of the atoms

is assigned vanishingly-small weight, so that G resembles G_0 . The resulting distributions look smoother and they tend to be closer and closer to the base measure G_0 . These results justify the interpretation of G_0 and α as the location and precision/roughness parameters respectively.

DP process also has the appealing conjugacy property. If $\theta_1, \dots, \theta_n \sim G$ with $G \sim \text{DP}(\alpha G_0)$, then

$$G|\theta_1, \dots, \theta_n \sim \text{DP}\left(\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}\right).$$

With squared error loss penalty, the optimal estimator for G is

$$\hat{G}(\cdot) = \frac{\alpha}{\alpha + n} G_0(\cdot) + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i}(\cdot),$$

which will converge to the empirical distribution as $n \rightarrow \infty$.

1.2 Dirichlet Process Mixtures

Since the DP models put probability one on the space of discrete measures, they cannot be used to model continuous data. Hence, the primary use of the DP is in nonparametric mixture modeling, by employing the DP as priors on the random mixing distribution over the parameters of a continuous kernel k ,

$$\begin{aligned} y &\sim f(\cdot), \\ f(\cdot) &= \int k(\cdot|\theta)G(d\theta), \\ G &\sim \text{DP}(\alpha G_0), \end{aligned}$$

resulting a DP mixture (DPM) model (Lo, 1984; Escobar, 1994; Escobar and West, 1998). The DPM induces a prior on f indirectly through a prior on the mixing distribution G . Popular choices are the DPM of Gaussian distributions, where $\theta = (\mu, \Sigma)$, which we call

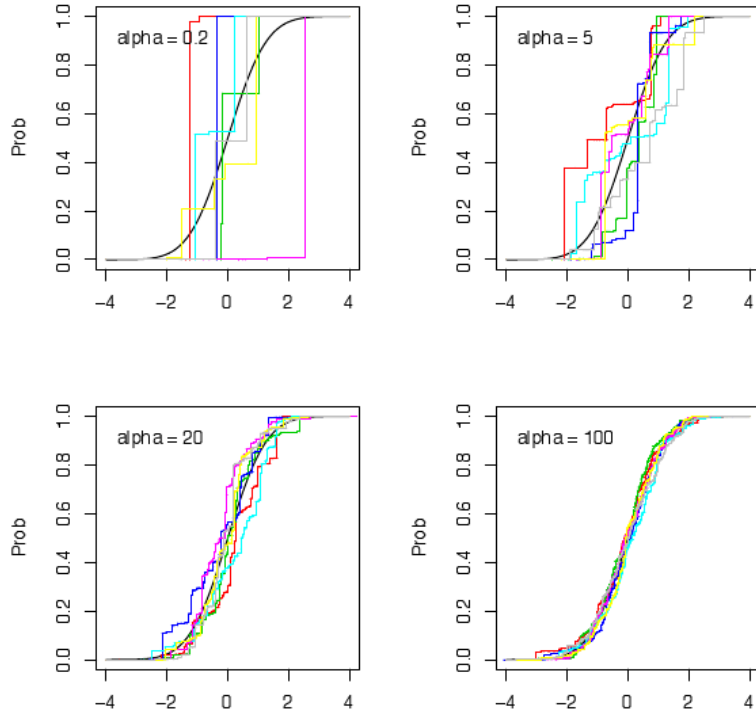


FIGURE 1.1: Samples from a DP process with a standard normal distribution as base measure with different precision parameters.

location-scale mixture of normals, or $\theta = \mu$, which we call location mixture of normals. $k(\cdot|\theta) = N_p(\cdot|\mu, \Sigma)$ is the p -variate normal kernel with mean μ and covariance matrix Σ .

Density estimates from location-scale DP mixtures can be interpreted as Bayesian kernel density estimates with adaptive bandwidth selection. It provides a direct link with well-known frequentist techniques and demonstrates the versatility of the model. Due to the discrete nature of the DP prior, the DPM model divides the observations into inde-

pendent groups, each one of them assumed to follow a distribution implied by the kernel k . Therefore, DPM models can be used for clustering as well as for density estimation. In this setting, the model automatically allows for an unknown number of clusters with model (1.3), which provides the implicit prior distribution.

1.3 Bayesian Nonparametric Regression using Dirichlet Process Mixtures

1.3.1 *Dependent Dirichlet Process and its Extensions*

More and more articles focus on flexible modeling of the conditional density of a response variable Y given multiple predictors $\mathbf{X} = (X_1, \dots, X_p)'$. $f(Y|\mathbf{X})$ are unknown and potentially change in shape as \mathbf{X} varies. The dependent Dirichlet process (DDP), which can be potentially used for modeling $f(Y|\mathbf{X})$ has been proposed by MacEachern (1999, 2000). Given an index set D , let $\{\theta(t) : t \in D\}$ and $\{v(t) : t \in D\}$ be stochastic processes over D such that $v(t) \sim \text{Beta}(1, \alpha(t))$ for $\forall t \in D$, then

$$G_t(\cdot) = \sum_{l=1}^{\infty} w_l(t) \delta_{\theta_l(t)}(\cdot),$$

where $\{\theta_l(t)\}_{l=1}^{\infty}$ and $\{v_l(t)\}_{l=1}^{\infty}$ are mutually independent collections of independent realizations of the stochastic processes $\{\theta(t) : t \in D\}$ and $\{v(t) : t \in D\}$ with $w_l(t) = v_l(t) \prod_{s < l} (1 - v_s(t))$. The defined $\mathbb{G}_D = \{G_t : t \in D\}$ is said to follow DDP.

Chung and Dunson (2009) extend the DDP in several aspects. They not only estimate the conditional response distribution addressing the distributional changes across the predictor space but also identify important predictors for the response distribution change both with local regions and globally. They first introduce the probit stick-breaking process (PSBP) as a new choice of prior for an uncountable collection of predictor-dependent random probability measures. The PSBP has distinct advantages over previous formulations in terms of computational tractability, which is particularly important in variable selection settings as marginal likelihoods need to be calculated. For modeling conditional

distributions, they propose a PSBP mixture (PSBPM) of normal linear regressions, resulting in an infinite mixture with mixing weights varying with predictors. More specifically, consider an uncountable collection of predictor-dependent random probability measures, $\mathbb{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, where \mathcal{X} is the sample space for the predictors $\mathbf{x} = (x_1, \dots, x_p)'$. The PSBP formulation for $G_{\mathbf{x}}$ is as:

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\theta_h}, \quad \forall \mathbf{x} \in \mathcal{X},$$

where $\pi_h(\mathbf{x})$ is a probability weight on the h th component. Let $\theta_h \sim G_0$ and introduce the following countable sequences of mutually independent random components:

$$\begin{aligned} \alpha_h &\sim \text{N}(\mu, 1), \\ \boldsymbol{\psi}_h &= \{\psi_{hj}\}_{j=1}^p \sim H_1, \\ \boldsymbol{\Gamma}_h &= \{\Gamma_{hj}\}_{j=1}^p \sim H_2, \end{aligned}$$

where H_1 and H_2 are distributions over a measurable Polish spaces $(\mathcal{L}_{\psi}, \mathcal{B}(\mathcal{L}_{\psi}))$ and $(\mathcal{L}_{\Gamma}, \mathcal{B}(\mathcal{L}_{\Gamma}))$ respectively. The probability weights $\pi_h(\mathbf{x})$ are then formed as

$$\begin{aligned} \pi_h(\mathbf{x}) &= \Phi(\eta_h(\mathbf{x})) \prod_{l < h} \{1 - \Phi(\eta_l(\mathbf{x}))\}, \\ \text{with } \eta_h(\mathbf{x}) &= \alpha_h - \sum_{j=1}^p \psi_{hj} |x_j - \Gamma_{hj}|, \quad \forall \mathbf{x} \in \mathcal{X}, \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal $\text{N}(0, 1)$. In order to address the curse of dimensionality in estimation and interest in testing and variable selection, they incorporate a variable selection structure through G_0 and H_1 as the following distributions for $\boldsymbol{\theta}_h$ and $\boldsymbol{\psi}_h$

$$\begin{aligned} \boldsymbol{\theta}_h &= (\boldsymbol{\beta}_h, \tau_h) \sim \text{N}_{p\gamma_h+1}(\boldsymbol{\beta}_{\gamma_h, h}; 0, \Sigma_{\gamma_h, h}) \times \delta_0(\boldsymbol{\beta}_{\bar{\gamma}_h, h}) \times \text{Gamma}(\tau_h; a_{\tau}, b_{\tau}), \\ \boldsymbol{\psi}_h &= \{\psi_{hj}\}_{j=1}^p \sim \prod_{j=1}^p \left\{ 1(\gamma_{hj} = 0) \delta_0(\psi_{hj}) + 1(\gamma_{hj} = 1) \text{N}_+(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}) \right\}, \end{aligned}$$

where N_+ denotes a truncated normal distribution bounded below by zero, $\beta_{\gamma_{h,j},h}$ is the vector of regression coefficients corresponding to $\gamma_{h,j} = 1$ including intercept, $\beta_{\bar{\gamma}_{h,j},h}$ is the coefficient vector with $\gamma_{h,j} = 0$, and $p_{\gamma_h} = \sum_{j=1}^p \gamma_{h,j}$. Note that $\gamma_{h,j}$ controls local inclusion of the j th predictor, with $\gamma_{h,j} = 0$ implying that $\psi_{h,j} = 0$ and $\beta_{h,j} = 0$. A value of $\beta_{h,j} = 0$ leads to the j th predictor assigned a coefficient of zero in the h th linear regression, while a value of $\psi_{h,j} = 0$ leads to excluding the j th predictor from the h th predictor-dependent stick-breaking weight in the expression for $\pi_h(\mathbf{x})$. Clearly, if $\gamma_{h,j} = 0$ for $h = 1, \dots, \infty$, then the j th predictor will be globally excluded from the model.

1.3.2 Bayesian Nonparametric Inference on Stochastic Ordering

Nonparametric Bayesian modeling can also be used to make inference on K group-specific distributions. For example, in toxicology studies, the groups may correspond to different doses of a potentially adverse chemical exposure. In such settings, it is of interest to assess whether the response distribution changes across groups, while also estimating the group-specific distributions. Also, it is common to have prior knowledge that the magnitude of a response for a particular experimental unit would not decrease if that unit had been exposed to a higher dose, which implies stochastic ordering in the response distribution. ? consider Bayesian inference on collections of unknown distributions subject to a partial stochastic ordering and propose classes of restricted dependent Dirichlet process (rDDP) priors. These rDDP priors have full support in the space of stochastically ordered distributions, and can be used for collections of unknown mixture distributions to obtain a flexible class of rDDP mixture models. Their proposed method is also the first paper to incorporate the stochastic ordering constraints in the DDP literature. Another contribution of their article is the development of a framework for testing of equalities in distributions between groups against stochastically ordered alternatives.

Let $(G_1, \dots, G_n) \in \mathcal{G}^n$ with G^n the set of $n \times 1$ collections of probability measures on

$(\mathcal{X}, \mathcal{B})$. In addition, define the following convex subset of \mathcal{G}^n :

$$C_E = \{(G_1, \dots, G_n) \in \mathcal{G}^n : G_i \leq G_j, \forall (i, j) \in E\},$$

where $E \subset (1, \dots, n)^2$ is a partial ordering. Here, $G_i \leq G_j$ if $G_i(x, \infty) \leq G_j(x, \infty)$ for all x , so that G_j is stochastically larger than G_i . The collections of probability measures belonging to C_E satisfy the partial ordering defined by E . As shown by Hoff (003a,b), C_E is a weakly closed convex set with extreme points $\{(\delta_{S_1}, \dots, \delta_{S_n}) : s \in \mathcal{S}_E\}$, where $\mathcal{S}_E = \{(s_1, \dots, s_n) \in \mathcal{X}^n : s_i \leq s_j, \forall (i, j) \in E\}$. The rDDP are formulated as following

$$G_k(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\Theta_{hk}}(\cdot),$$

$$\pi_h = v_h \prod_{l < h} (1 - v_l), k = 1, \dots, n,$$

where $v_h \stackrel{\text{i.i.d.}}{\sim} \text{beta}(1, \alpha)$, $h = 1, \dots, \infty$, are stick-breaking weights, and

$$\Theta_h = (\Theta_{h1}, \dots, \Theta_{hn})' \stackrel{\text{i.i.d.}}{\sim} G_0$$

are random atoms, with G_0 a Borel probability measure on \mathcal{S}_E .

Previous Bayesian nonparametric analyses of stochastically ordered distributions have made strict constraints except when \mathcal{X} is discrete. In addition to not allowing uncertainty in group differences, strict constraints can lead to a tendency to cover estimate group differences, particularly when the true difference is small, sample sizes are small to moderate, and the number of groups is moderate to large. By incorporating prior mass at the boundary (?), one obtains a shrinkage estimator of the density, which borrows information across groups.

1.3.3 Hierarchical Dirichlet Process and the Nested Dirichlet Process

The hierarchical Dirichlet process (HDP) (Teh et al., 2006) places a prior on a collection of exchangeable distributions $\{G_1, \dots, G_n\}$. Conditional on a probability measure G_0 , the

distributions in the collection are assumed to be iid samples from a regular DP centered around G_0 . In order to induce dependence, G_0 is in turn given another DP prior. In summary,

$$\begin{aligned} G_i | G_0 &\sim \text{DP}(\alpha G_0), \\ G_0 &\sim \text{DP}(\beta H). \end{aligned}$$

Since, by construction, G_0 is almost surely discrete, the distributions G_i share the same set of random atoms, but assign strictly different (although dependent) weights to each one of them. α and β control the variance around H and the dependence between distributions.

In current research, people would like to borrow information from multicenter studies. In such studies, subjects in different centers may have different outcome distributions. Rodríguez et al. (2008) propose the nested DP (nDP) prior, which can be placed on the collection of distributions of the different centers, with centers drawn from the same DP component automatically clustered together. A collection of distributions $\{G_1, \dots, G_n\}$ is said to follow a nDP prior if

$$\begin{aligned} G_j(\cdot) &\sim \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*(\cdot)}, \\ G_k^*(\cdot) &= \sum_{l=1}^{\infty} w_{lk} \delta_{\theta_{lk}^*(\cdot)}, \end{aligned}$$

where $\theta_{lk} \sim G_0$, $w_{lk} = u_{lk} \prod_{s < l} (1 - u_{sk})$, $\pi_k = v_k \prod_{s < k} (1 - v_s)$, $v_k \sim \text{beta}(1, \alpha)$ and $u_{lk} \sim \text{beta}(1, \beta)$. The nDP is a novel extension of the Dirichlet process for a family of *a priori* exchangeable distributions that allows us to simultaneously cluster groups and observations within groups. Moreover, the groups are clustered by their entire distribution rather than by particular features of it.

The HDP specification automatically allocates subjects 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.

1.4 Inference of Dirichlet Process Mixture Modeling

The discussion thus far has led to methods for generating $G \sim \text{DP}(\alpha G_0)$. In this section, we discuss different Markov chain Monte Carlo (MCMC) inferences for DPM.

1.4.1 Blocked Gibbs Sampler

Consider the blocked stick-breaking process prior representation for the DPM as following for $i = 1, \dots, n$

$$\begin{aligned} (y_i | \boldsymbol{\theta}, \mathbf{K}, \phi) &\stackrel{\text{i.i.d.}}{\sim} \pi(y_i | \theta_{K_i}, \phi), \\ (K_i | \mathbf{p}) &\stackrel{\text{i.i.d.}}{\sim} \sum_{k=1}^N p_k \delta_k(\cdot), \\ (\mathbf{p}, \boldsymbol{\theta}) &\sim \pi(\mathbf{p}) \times G_0^N(\boldsymbol{\theta}), \\ \phi &\sim \pi(\phi). \end{aligned}$$

where p_k follows the stick breaking process, with $p_k = w_k \prod_{l < k} (1 - w_l)$ and $w_k \sim \text{beta}(1, \alpha)$ and N is the upper bound to approximate the DP process.

Let $\{K_1^*, \dots, K_m^*\}$ denote the set of current m unique values of \mathbf{K} . Ishwaran and James (2001) propose the following blocked Gibbs sampling algorithm

(a) Conditional for $\boldsymbol{\theta}$: for each $k \in \mathbf{K} - \{K_1^*, \dots, K_m^*\}$.

$$f(\theta_{K_j^*} | \dots) \propto G_0(d\theta_{K_j^*}) \prod_{\{i: K_i = K_j^*\}} f(y_i | \theta_{K_j^*}, \phi), \text{ for } j = 1, \dots, m.$$

(b) Conditional for \mathbf{K} :

$$(K_i | \dots) \stackrel{\text{i.i.d.}}{\sim} \sum_{k=1}^N p_{k,i} \delta_k(\cdot), \quad i = 1, \dots, n,$$

where $(p_{1,i}, \dots, p_{N,i}) \propto (p_1 f(y_i | \theta_1, \phi), \dots, p_N f(y_i | \theta_N, \phi))$.

(c) Conditional for \mathbf{p} :

$$p_1 = w_1,$$

$$p_k = (1 - w_1)(1 - w_2) \cdots (1 - w_{k-1})w_k, \text{ for } k = 2, \dots, N - 1,$$

where $w_k \stackrel{\text{ind}}{\sim} \text{beta}(a_k + M_k, b_k + \sum_{l=k+1}^N M_l)$, for $k = 1, \dots, N - 1$ and M_k records the number of K_i values that equal k .

(d) Conditional for ϕ :

$$f(\phi | \dots) \propto \pi(d\phi) \prod_{i=1}^n f(y_i | \theta_{K_i}, \phi).$$

The blocked Gibbs sampler works by directly sampling values from the posterior of the random measure and can be viewed as a more general approach because it works without requiring an explicit prediction rule. The blocked Gibbs avoids some of the limitations seen with the Pólya urn approach and should be simpler for nonexperts to use.

1.4.2 Exact Block Gibbs Sampler

We can escape to prefix the upper bound N by using the following exact block Gibbs sampler (Yau et al., 2010):

(a) The joint prior distribution of the indicator K_i and a latent variable q_i can be written as

$$f(K_i, q_i | \mathbf{p}) = \sum_{l: v_l > q_i} \delta_l(\cdot) = \sum_{l=1}^{\infty} 1(q_i < p_l) \delta_l(\cdot).$$

- i. Sample $q_i \sim \text{Unif}(0, p_{K_i})$ for $i = 1, \dots, n$ with $p_l = w_l \prod_{s < l} (1 - w_s)$.
- ii. Sample the stick-breaking random variables

$$w_k \stackrel{\text{ind}}{\sim} \text{beta}(a_k + M_k, b_k + \sum_{l=k+1}^L M_l),$$

for $k = 1, \dots, L - 1$ and M_k records the number of K_i values that equal k . L is the minimum value satisfying $p_1 + p_2 + \dots + p_L > 1 - \min\{q_i\}$.

iii. Sample K_i for $i = 1, \dots, n$ form the multinomial conditional with

$$\Pr(K_i = l) \propto 1(q_i < p_l) f(y_i | \theta_{K_i}, \phi).$$

All the other sampling steps are left the same as the block Gibbs sampler described in the above.

1.4.3 Pólya Urn Gibbs Sampling

We iteratively draw values from the conditional distribution of $(\theta_i | \boldsymbol{\theta}_{-i}, \phi, y)$, for $i = 1, \dots, n$. In particular, each iterations of the Gibbs sampler draws the following samples:

(a) $(\theta_i | \boldsymbol{\theta}_{-i}, \phi, y)$ for each $i = 1, \dots, n$: The required conditional distributions are defined by

$$\mathcal{P}(\theta_i \in \cdot | \boldsymbol{\theta}_{-i}, \phi, y) = q_0^* \mathcal{P}(\theta_i \in \cdot | \phi, y_i) + \sum_{j=1}^m q_j^* \delta_{\theta_j^*}(\cdot),$$

where

$$q_0^* \propto (\alpha + m) \int f(y_i | \theta, \phi) G_0(d\theta),$$

$$q_j^* \propto (M_j - 1) f(y_i | \theta_j^*, \phi),$$

and these values are subject to the constraint that they sum to 1, that is, $\sum_{j=0}^m q_j^* = 1$.

Here we are dropping the dependence on i for notational simplicity and we write $\{\theta_1^*, \dots, \theta_m^*\}$ for the set of unique values in $\boldsymbol{\theta}_{-i}$, where each value occurs with frequency M_j for $j = 1, \dots, m$.

The other updating steps remain the same. Although here we focus on its application to stick-breaking priors (such as the DP), in principle, the Pólya urn Gibbs sampler can be applied to any random probability measure with a known prediction rule. Refer to Ishwaran and James (2001) for more detailed discussion.

1.5 Other Nonparametric Bayesian Modeling Processes

1.5.1 Chinese Restaurant Process

The most common choice of infinite-capacity prior is known as the Chinese Restaurant Process (CRP). A draw from this distribution can be generated by sequentially assigning observations to classes with probability

$$\Pr(K_n = c | K_{1:n-1}) \propto \begin{cases} m_c, & \text{if } c \leq C \text{ (i.e., } c \text{ is an old class),} \\ \alpha, & \text{otherwise (i.e., } c \text{ is a new class).} \end{cases}$$

where m_c is the number of observations currently assigned to class c and C is the number of classes for which $m_c > 0$. In this setting, the parameter α is referred to as the concentration parameter. Intuitively, a larger value of α will produce more clusters, and in the limit as $\alpha \rightarrow \infty$ all observations will be assigned to a unique class, whereas in the limit as $\alpha \rightarrow 0$ all observations will be assigned to the same class.

There are two important invariant properties related to the CRP. First, under the distribution the assignment vector $\mathbf{K} = (1, 2, 2)$ has the same posterior probability as $\mathbf{K} = (2, 1, 1)$. These vectors are equivalent up to a “label switch”. Second, the cluster assignments under the CRP are exchangeable, despite being drawn from a sequential process. This means that the joint distribution $P(\mathbf{K})$ is unchanged if the order of datapoints is shuffled. Moreover, this distribution is marginally invariant: removing a single datapoint leaves the distribution over the other datapoints unchanged.

The CRP derives its name from the following metaphor: imagine a Chinese restaurant with an infinite number of tables, where each table corresponds to a class and customers correspond to observations. The first customer enters and sits at the first table. The second customer enters and sits at the first table with probability $1/(1 + \alpha)$, and the second table with probability $\alpha/(1 + \alpha)$. This process continues until all customers have been seated, at which point the assignment of customers to tables defines a partition.

The customers of a CRP are exchangeable: under any permutation of their ordering, the probability of a particular configuration is the same. Exchangeability is a reasonable assumption in some clustering applications, but in many it is not. Consider data ordered in time, such as a time-stamped collection of news articles. In this setting, each article should tend to cluster with other articles that are nearby in time. Or, consider spatial data, such as pixels in an image or measurements at geographic locations. Here again, each data should tend to cluster with other data that are nearby in space. And in our Wikipedia study, we would like articles share more characteristics to “sit at the same table”. While the traditional CRP mixture provides a flexible prior over partitions of the data, it cannot accommodate such non-exchangeability.

Blei and Frazier (2010) proposes the dependent Chinese restaurant process (dCRP), a flexible class of distributions over partitions that allows for non-exchangeability. This class can be used to model dependencies between data in finite clustering models, including dependencies across time or space. The key to the dCRP is that it represents the partition with customer assignments, rather than table assignments. While the traditional CRP connects customers to tables, the dCRP connects customers to other customers. The partition of the data, i.e., the table assignment representation, arises from these customer connections. Let c_i denote the α_i assignment, the index of the “customer” with whom the α_i is sitting. Let d_{ij} denote the difference of the characteristics between subject i and subject j , and let f be a decay function. The dCRP independently draws the customer assignments conditioned on the distance measurements,

$$\Pr(c_i = j | \dots) \propto \begin{cases} f(d_{ij}), & \text{if } j \neq i, \\ a, & \text{if } i = j. \end{cases}$$

We should notice that the customer assignments do not depend on other customer assignments, only the distances between customers. Also notice that j ranges over the entire set of “customers”, and so any customer may sit with any other. Define $R(c_{1:n})$ to contain one

customer index from each table and $R(c_{-i})$ to contain one index from each cluster in the seating assignment c_{-i} with the outgoing link of customer i removed. Then the prior for c_i given c_{-i} is

$$p(c_i|c_{-i}, \mathbf{d}) \propto \begin{cases} a1(\alpha_i = \alpha_{c_i}), & \text{if } c_i = i, \\ f(d_{ij})1(\alpha_i = \alpha_{c_i}), & \text{if } c_i \neq i \text{ and } R(c_{-i}) \text{ does not change,} \\ f(d_{ij})1(\alpha_i = \alpha_{c_i})/p(\alpha_i|G_0), & \text{if } c_i \text{ joins two tables in } R(c_{-i}), \end{cases} \quad (1.6)$$

where \mathbf{d} denotes the set of all distance measurements between customers and G_0 is the base measure. More details about the derivation of model (1.6) can be seen from Blei and Frazier (2009).

1.5.2 Indian Buffet Process

Latent factor models typically assume that the observed data is generated by a noisy weighted combination of latent factors:

$$\mathbf{y}_n = \sum_{k=1}^K \phi_k z_{nk} + \epsilon,$$

where ϵ is a vector of Gaussian noise terms, z_{nk} is a binary “mask” variable that indicates whether factor k is “on” for observation n , and ϕ_k is a vector of weights expressing how strongly each factor influences the observation. In reality, the number of factors is unknown. Therefore, we would like to avoid specifying K and instead allow it to be unbounded. In this case, $\mathbf{Z} = (z_{nk})$ is a binary matrix with a finite number of rows (each corresponding to an observation) and an infinite number of columns (each corresponding to a latent factor).

Like the CRP, the infinite-capacity distribution over \mathbf{Z} has been furnished with a similarly colorful culinary metaphor: Indian Buffet Process (IBP). IBP can be derived as following: a customer (observation) enters a buffet with an infinite number of dishes (factors)

arranged in a line. The customer samples dish k in proportion to its popularity m_k (the number of prior customers who have sampled the dish). When the customer has sampled all the previously sampled dishes (i.e., those for which $m_k > 0$), it samples an additional $\text{Poisson}(\alpha)$ dishes that have never been sampled before. When all N customers have navigated the buffet, the resulting binary matrix \mathbf{Z} (encoding which customers sampled which dishes) is a draw from the IBP.

The IBP plays exactly the same role for latent factor models that the CRP plays for the mixture models: it functions as an infinite-capacity prior over the space of latent variables, allowing an unbounded number of latent factors. Whereas in the CRP, each observation is associated with only one latent class, in the IBP each observation is associated with a theoretically infinite number of latent factors. We can see the comparisons from Figure 1.2. Customers are represented by numbered squares; tables (in the CRP, top) and dishes (in the IBP, bottom) are represented by circles. Arrows show the assignment of a new customer. On the right of Figure 1.2 are the matrices produced by the CRP (top) and (IBP) (bottom) respectively. Rows correspond to observations. Columns correspond to classes in the CRP and factors in the IBP. The key difference is that in the CRP, each customer sits at a single table, whereas in the IBP, a customer can sample several dishes. That is, the CRP assigns each observation to a single class, whereas, the IBP assigns each observation to potentially multiple factors.

1.5.3 Pólya Tree Process

Another useful characterization of the DP is as a special case of the Pólya tree (PT) prior (Lavine, 1992, 1994). A particularly attractive feature of PT priors is the possibility to model absolutely continuous distributions. The definition of PT starts with a nested sequence $\Pi = \{\pi_m, m = 1, 2, \dots\}$ of partitions of the sample space Ω . Assuming that the partitions are binary, we start with a partition $\pi_1 = \{B_0, B_1\}$ of the sample space $\Omega = B_0 \cup B_1$ and continue with nested partitions defined by $B_0 = B_{00} \cup B_{01}$,

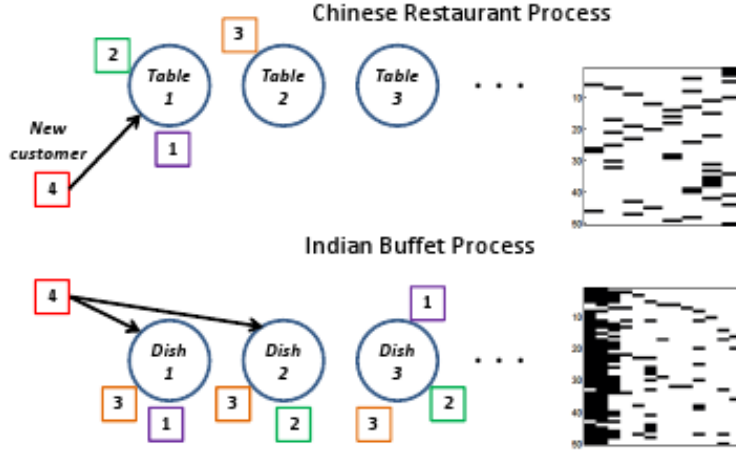


FIGURE 1.2: (Left) The generative process of the CRP (top) and IBP (bottom). (Right) Matrices produced by the CRP (top) and IBP (bottom)

$B_1 = B_{10} \cup B_{11}$, etc. Thus the partition level m is $\pi_m = \{B_\epsilon, \epsilon = \epsilon_1 \dots \epsilon_m\}$, where ϵ are all binary sequences of length m . A PT prior for a random probability measure G is defined by a beta-distributed random branching probabilities. Let $Y_{\epsilon_0} = G(B_{\epsilon_0}|B_\epsilon)$, and $\mathcal{A} \sim \text{beta}(\alpha_{\epsilon_0}, \alpha_{\epsilon_1})$, then we say that G has a PT prior with $G \sim \text{PT}(\Pi, \mathcal{A})$.

The parameters α_ϵ are usually chosen as $\alpha_\epsilon = cm^r$ for level m subsets. For $r = 2$ the random probability measure G is a.s. absolutely continuous. With $r = -1/2$ the PT reduces to the DP as a special case. The partitioning subsets B_ϵ can be chosen to achieve

a desired prior mean G_0 . Let $q_{mk} = G^{*-1}(k/2^m)$, $k = 0, \dots, 2^m$, denote the inverse c.d.f. under G_0 evaluated at dyadic fractions. If $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$, for example $\alpha_\epsilon = cm^r$, and the dyadic quantile sets $[q_{mk}, q_{m,k+1})$ are used as the partitioning subsets B_ϵ then $E(G) = G_0$. Alternatively the prior mean can be fixed by G_0 by choosing $\alpha_{\epsilon 0}/(\alpha_{\epsilon 0} + \alpha_{\epsilon 1}) = G_0(B_{\epsilon 0}|B_\epsilon)$ for any choice of the nested partitions Π .

The main attraction of PT models for nonparametric Bayesian inference is the simplicity of posterior updating. Assuming $x_i \stackrel{\text{i.i.d.}}{\sim} G$ for $i = 1, \dots, n$ and $G \sim \text{PT}(\Pi, \mathcal{A})$. Consider first $n = 1$, i.e., a single sample from the unknown distribution G . The posterior $p(G|x)$ is again a Pólya tree, $p(G|x) = \mathcal{P}(\Pi, \mathcal{A}')$ with the beta parameters in \mathcal{A}' defined as

$$\alpha'_\epsilon = \begin{cases} \alpha_\epsilon, & \text{if } x_1 \notin B_\epsilon, \\ \alpha_\epsilon + 1, & \text{if } x_1 \in B_\epsilon. \end{cases}$$

The general case with a sample of size $n > 1$ follows by induction. The above result can be used to implement exact posterior predictive simulation, i.e., simulation from $p(x_{n+1}|x_1, \dots, x_n)$.

Multiple Bayesian Elastic Net

2.1 Introduction

Highly correlated relevant features are frequently encountered in variable selection problems. One example in text mining concerns estimation of the probability that two Wikipedia articles are linked, based on their lexical content. Some sets of words are strongly associated and informative about the presence of a link; other sets of words may be associated but uninformative. One wants to select all words in a useful, strongly associated set simultaneously as a group, for better model interpretation and robustness. Similarly, irrelevant lexical sets should be excluded as a group, providing a sparse solution that tends to avoid overfitting.

For a bag-of-words model, the number of features is equal to the number of distinct words in the articles. For the Wikipedia problem, the sample size (the number of links and non-links to a specific article within a narrow topic area) is usually insufficient to allow accurate selection of important predictors; thus we must borrow strength across data from multiple articles. This paper focuses on the problem of flexibly borrowing strength across data sources (articles) in selecting predictors (words) from among a very large number of

candidate features.

We initially consider a linear regression model of the form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{X} is the $n \times p$ matrix of *standardized* data, each row of which corresponds to a sample, $\boldsymbol{\beta}$ is the $p \times 1$ regression coefficient vector, \mathbf{y} is the $n \times 1$ vector of *centered* regression responses, and $\boldsymbol{\epsilon}$ is an $n \times 1$ additive noise vector. The ordinary least squares (OLS) estimate can be obtained by minimizing the residual sum of squares, but it is well known that the OLS does poorly in both prediction and interpretation when data are sparse relative to the number of parameters. When $p > n$, the OLS will yield an estimator that is not unique since \mathbf{X} is not of full rank. Penalization techniques that promote shrinkage of $\boldsymbol{\beta}$ have been proposed to improve OLS. Ridge regression Hoerl and Kennard (1988) minimizes the residual sum of squares subject to a penalty on the L_2 -norm of the coefficients. For the problem of multicollinearity, ridge regression improves prediction accuracy, but it cannot perform variable selection.

An alternative technique called the lasso was proposed by Tibshirani (1996). The lasso is a penalized least squares method subject to an L_1 -penalty on the regression coefficients, leading to continuous shrinkage and automatic variable selection simultaneously. However, as summarized by Zou and Hastie (2005), the lasso has several important limitations. In the $p > n$ case, the convex optimization algorithm forces the lasso to select at most n variables before it saturates. Also, if there is a group of predictors that are highly correlated with each other, the lasso tends to select only one variable from the group, almost arbitrarily. Even for the usual $n > p$ situations, if there are strong correlations among predictors, ridge regression will outperform the lasso in terms of predictive performance Zou and Hastie (2005).

Zou and Hastie (2005) proposed the elastic net, a new regularization of the lasso, for use when there are unknown groups of multicollinear predictors. The elastic net estimator

can be expressed as

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p |\beta_j|^2,$$

where λ_1 and λ_2 are tuning parameters. The elastic net estimator can be interpreted as a stabilized version of the lasso. As pointed by Zou and Hastie (2005), the elastic net often outperforms the lasso. In addition, the elastic net encourages the grouping of covariates, so that strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors p is much bigger than the number of observations n ; potentially, the elastic net can select all p predictors.

The Bayesian elastic net model was first studied in Bornn et al. (2007). Their model assumed

$$\begin{aligned} \mathbf{y} &\sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I}), \quad \tau^2 \sim \text{IG}(c_0, d_0), \\ \beta_j &\sim \mathbf{N}(0, \tau^2(\alpha_j + \lambda)^{-1}), \\ \alpha_j &\sim \eta \left(\frac{\alpha_j}{\alpha_j + \lambda} \right)^{1/2} \text{IG}(\alpha_j; 1, \frac{\gamma}{2}), \end{aligned} \quad (2.1)$$

where η is a normalizing constant and IG denotes the Inverse-Gamma distribution. By integrating out $\boldsymbol{\alpha}$, the likelihood becomes

$$p(\mathbf{y}; \boldsymbol{\beta}, \tau^2, \gamma) \propto f(\tau^2, \gamma) \exp \left\{ -\frac{1}{2\tau^2} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\sqrt{\gamma\tau^2}|\boldsymbol{\beta}| + \lambda\|\boldsymbol{\beta}\|^2) \right\},$$

where $f(\tau^2, \gamma)$ is the joint prior for τ^2 and γ . The likelihood after integrating out $\boldsymbol{\alpha}$ is similar to the elastic net penalization problem and thus the resulting posterior estimates are a Bayesian generalization of the elastic net. The MAP estimate of $\boldsymbol{\beta}$ is no longer a linear function of \mathbf{y} ; it is given by

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\sqrt{\gamma\tau^2}|\boldsymbol{\beta}| + \lambda\|\boldsymbol{\beta}\|^2 \}. \quad (2.2)$$

It is clear that γ and λ control the shrinkage of coefficients, with larger values of γ and λ inducing more shrinkage. Chen et al. (2010) introduce different γ_j for each precision parameter α_j , with further Gamma priors being imposed on them. In this way, they achieve sparsity and grouped variable selection simultaneously.

There are many advantages to embedding shrinkage priors in a hierarchical Bayesian formulation. In addition to the usual ease-of-interpretation of hierarchical models, the Bayesian formulation produces valid measures of uncertainty (which can be problematic for the frequentist lasso). Also, posterior computation for the Bayesian lasso can be based on a geometrically ergodic Markov chain using a block Gibbs sampler Kyung et al. (2010).

We extend the previous work on Bayesian elastic nets in several important ways. First, we improve the flexibility of shrinkage through mixing elastic net priors, which automatically allows adaptive shrinkage, not only to zero but also to a sparse set of non-zero values. Multiple shrinkage was originally suggested by George (1986), with MacLehose and Dunson (2010) proposing a Dirichlet process (DP) Ferguson (1974) mixture of double exponential priors. Nott (2008) also investigates a generalization of ridge regression in the linear model using the DP prior. The approach in MacLehose and Dunson (2010) inherits some of the disadvantages of the lasso, such as a tendency to select only one variable from a highly correlated set of variables. In addition, when using Markov chain Monte Carlo (MCMC), it is infeasible to handle very large values of p and/or n . For applications with large, high-dimensional data, such as the Wikipedia project, computational speed is important—standard Bayesian methods of posterior computation using MCMC are too time consuming to implement. To achieve fast computation in large p , this paper uses Monte Carlo EM sparse point estimation Booth and Hobert (1999).

In addition, we extend the new MBEN model to the problem of logistic regression, in order to deal with classification problems that have sparse but correlated sets of covariates. A further extension to MTL, motivated by fitting the Wikipedia data, is also considered. The fact that the topics of some of the articles are related implies that what is learned

about classifiers from one article is transferable to another. By learning the classifiers in parallel through a unified representation, the transferability of expertise between articles is exploited to the benefit of all. This expertise transfer is particularly important since there is only a limited amount of data for learning each classifier. By exploring data from related articles, the inference for each article is strengthened. More specifically, the model in the Wikipedia network describes the link from article i to article m as

$$\text{logit Pr}(y_i^{(m)} = 1 | \mathbf{x}_i^{(m)}) = \mathbf{x}_i^{(m)'} \boldsymbol{\beta}^{(m)}, \quad \text{for } m = 1, \dots, M, \quad i = 1, \dots, n^{(m)}, \quad (2.3)$$

where $\mathbf{x}_i^{(m)}$ is the frequency of a common lexical word (e.g., “Gaussian”) being used both in article i and the central article m . The MBEN prior is shared across the M articles (so the sparseness properties are shared, but not the exact regression weights). In this setting the M articles cluster, and within each cluster, similar components of $\boldsymbol{\beta}^{(m)}$ are important. Liu et al. (2010) also use hierarchical shrinkage approaches for multi-task learning, but their approach is set up under a hierarchical parametric model in which information sharing across studies only affects the precision parameters controlling shrinkage to zero. Our work is also different from Xue et al. (2007), since their paper does not do multiple shrinkage but assumes that two tasks either have identical coefficients for all predictors or have completely different coefficients, which is too restrictive, especially when p is large.

In section 2, we introduce the proposed MBEN approach and also consider modifications needed for logistic regression and multi-task learning. In section 3, we describe an efficient Gibbs sampling algorithm for posterior computation. Section 4 outlines a Monte Carlo EM algorithm for sparse point estimation. Data simulation is performed in section 5 and section 6 applies the approach to the Wikipedia project. Section 7 contains a discussion.

2.2 Multiple Bayesian elastic net Prior

The prior distribution in (2.1) induces shrinkage toward the prior mean of zero; however, as pointed by MacLehose and Dunson (2010), in many situations shrinkage toward non-null values will be beneficial. Their approach is more flexible than the Bayesian lasso in that they allow multiple shrinkage instead of just shrinking towards zero. Thus the data help to choose the best hyperparameters while favoring sparsity through the use of a carefully-tailored hyperprior. Building on their approach and the model (2.1), we introduce the following MBEN prior,

$$\begin{aligned}
 \beta_j &\sim \mathbf{N}(\mu_j, \tau^2(\alpha_j + \lambda)^{-1}), \quad \lambda \sim \mathbf{G}(r_0, s_0) \\
 (\mu_j, \alpha_j) &\sim \eta \left\{ \pi \delta_0(\mu_j) \left(\frac{\alpha_j}{\alpha_j + \lambda} \right)^{1/2} \mathbf{IG}(\alpha_j; 1, \frac{\gamma}{2}) \mathbf{G}(\gamma; a_0, b_0) + (1 - \pi) D \right\} \\
 \pi &\sim \mathbf{beta}(1, \alpha_0), \quad D \sim DP(\alpha_0 D_0) \\
 D_0 &\equiv \mathbf{N}(\mu_j; c_1, d_1) \left(\frac{\alpha_j}{\alpha_j + \lambda} \right)^{1/2} \mathbf{IG}(\alpha_j; 1, \frac{\gamma}{2}) \mathbf{G}(\gamma; a_1, b_1)
 \end{aligned} \tag{2.4}$$

where $\delta_0(\mu_j)$ indicates that the random variable μ_j has a degenerate distribution with all its mass at 0, and $\mathbf{G}(a_1, b_1)$ denotes the Gamma distribution with mean a_1/b_1 and variance a_1/b_1^2 . Thus, with probability π a coefficient is shrunk toward zero as in standard elastic net estimation. With probability $1 - \pi$ the coefficient is shrunk toward a non-zero mean, μ_j .

Because the DP prior implies that D is almost surely discrete, the prior will automatically group the p coefficient-specific hyperparameters, $\{\mu_j, \alpha_j\}_{j=1}^p$, into L clusters, $\{\mu_l^*, \alpha_l^*\}$, with $L \leq p$. One of these clusters will most likely correspond to $\mu_j = 0$, while the other clusters will not have zero means. The prior on the number of clusters is controlled by α_0 , with smaller α_0 inducing fewer clusters. When the data are informative about the number of clusters and the cluster-specific hyperparameters, the procedure adaptively shrinks coefficients toward the non-zero locations suggested by the available data.

Our proposed MBEN prior can be seen more clearly through the equivalent stick breaking form:

$$\begin{aligned}
\beta_j &\sim \mathbf{N}(\mu_{k_j}^*, \tau^2(\alpha_{k_j}^* + \lambda)^{-1}), \quad \lambda \sim \mathbf{G}(r_0, s_0), \quad \tau^2 \sim \mathbf{IG}(c_0, d_0) \\
k_j &\sim \sum_{t=1}^{\infty} \pi_t \delta_t, \\
(\mu_t^*, \alpha_t^*) &\sim \begin{cases} \eta_0 \delta_0(\mu_t^*) (\frac{\alpha_t^*}{\alpha_t^* + \lambda})^{1/2} \mathbf{IG}(\alpha_t^*; 1, \frac{\gamma_0}{2}), & \text{for } t = 1, \\ \eta_1 \mathbf{N}(\mu_t^*; c_1, d_1) (\frac{\alpha_t^*}{\alpha_t^* + \lambda})^{1/2} \mathbf{IG}(\alpha_t^*; 1, \frac{\gamma_1}{2}), & \text{for } t > 1, \end{cases} \\
\gamma_l &\sim \mathbf{G}(a_l, b_l), \quad \text{for } l = 0, 1
\end{aligned} \tag{2.5}$$

where the random variable π_t is constructed as $\pi_t = V_t \prod_{h < t} (1 - V_h)$, $V_t \sim \text{beta}(1, \alpha_0)$, and η_0 and η_1 are normalizing constants. The prior on the number of clusters is controlled by α_0 , with smaller values favoring fewer clusters. However, the data are strongly informative about the number of clusters and the cluster-specific hyperparameters. Coefficients are shrunk adaptively toward non-zero locations suggested by the data. An infinite number of (μ_t^*, α_t^*) are drawn from the prior distribution, with the j th coefficient falling into the k_j th of these components. Coefficients with prior parameters drawn from the first component have a standard elastic net prior. Thus for $t > 1$, π_t is the prior probability of falling into the t th component and it is constructed through a stick-breaking process. Figure 2.1 shows a random draw from the prior distribution for one predictor-specific coefficient.

Each mixture component has a distinct prior mean and variance. Although there are infinitely-many components available, the intrinsic Bayes penalty for model complexity will tend to favor allocation of the coefficients to very few components relative to p . One can reinforce allocation to few components through choosing a small value of α_0 . By choosing a relatively large value of d_1 , one can favor a wide variety of prior means, which also tends to upweight allocation to a few components that are widely distributed. Even with such priors, the data are strongly informative about the number of occupied components and the distribution of the prior means.

For the Wikipedia project, responses are binary, $\mathbf{z} = (z_1, \dots, z_n)'$ with $z_i \in \{0, 1\}$, where $z_i = 0$ means there is no link from the current article to article i , and $z_i = 1$ otherwise. We augment the data using the O'Brien and Dunson (2004) algorithm by assuming the outcome $z_i = 1$ when the latent variable $y_i > 0$. With $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i / \phi^{1/2}$, $\epsilon_i \sim \mathbf{N}(0, \tilde{\tau}^2)$ and $\phi \sim \mathbf{G}(\nu/2, \nu/2)$. The scale mixture of normals for ϵ_i with $\tilde{\tau}^2 = \pi^2(\nu - 2)/(3\nu)$ and $\nu = 7.3$ is a near exact representation of the logistic distribution. It is easy to apply the MBEN prior to the logistic regression model as

$$\begin{aligned}
z_i &= 1(y_i > 0), \\
y_i &= \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i, \\
\epsilon_i &= \tilde{\epsilon}_i / \phi^{1/2}, \quad \tilde{\epsilon}_i \sim \mathbf{N}(0, \tilde{\tau}^2), \quad \phi \sim \mathbf{G}(\nu/2, \nu/2), \\
\beta_j &\sim \text{MBEN}(\mu_{k_j}^*, \alpha_{k_j}^*, \tau^2, \lambda, \gamma).
\end{aligned} \tag{2.6}$$

The proposed MBEN model can be readily extended to a multi-task setting, in which multiple regression or classification tasks are performed jointly. For example, the regression coefficients can differ from task to task, but the sparsity pattern of the regression coefficients may be shared, thereby imposing the belief that irrelevant features are the same or similar across the tasks. Suppose there are M tasks, each represented as in model (2.6),

$$\mathbf{y}^{(m)} = \mathbf{X}^{(m)} \boldsymbol{\beta}^{(m)} + \boldsymbol{\epsilon}^{(m)}, \quad m = 1, \dots, M,$$

where $\mathbf{X}^{(m)}$ is the $n^{(m)} \times p$ design matrix for task m , $n^{(m)}$ is the number of samples for task m , $\boldsymbol{\beta}^{(m)}$ is the $p \times 1$ dimensional regression coefficients for task m and $\boldsymbol{\epsilon}^{(m)}$ is the residual for task m . The Multiple Bayesian elastic net prior is shared across the M tasks (the sparseness properties are shared, but not the exact regression coefficients). For classification problems, $\mathbf{y}^{(m)}$ is not directly observed, instead only the label information

$\mathbf{z}^{(m)}$ is known with $z_i^{(m)} \in \{1, 0\}$. The multi-task model can be expressed as

$$\begin{aligned}
z_i^{(m)} &= \mathbb{1}(y_i^{(m)} > 0), \\
y_i^{(m)} &= \mathbf{x}_i^{(m)'} \boldsymbol{\beta}^{(m)} + \epsilon_i^{(m)}, \\
\epsilon_i^{(m)} &= \tilde{\epsilon}_i^{(m)} / \phi^{(m)1/2}, \quad \tilde{\epsilon}_i^{(m)} \sim \mathbf{N}(0, \tilde{\tau}^{(m)2}), \quad \phi^{(m)} \sim \mathbf{G}(\nu/2, \nu/2), \\
\beta_j^{(m)} &\sim \text{MBEN}(\mu_{k_j}^*, \alpha_{k_j}^*, \tau^{(m)2}, \lambda, \gamma).
\end{aligned} \tag{2.7}$$

The above form is for multi-task classification; for regression one can simply remove the first line of (2.7). The task-dependent $\boldsymbol{\beta}^{(m)}$ share the same MBEN prior, implying that the multiple tasks share similar non-zero coefficients and similar weights on each component. In this setting, the M tasks cluster, and within each cluster similar components of $\boldsymbol{\beta}^{(m)}$ are important.

2.3 Posterior Computation

2.3.1 Weakly Informative Prior Specification

Prior specification is of great importance in Bayesian modeling. As we are interested in high-dimensional settings and automated methods that can be applied quickly to new data sets, subjective prior elicitation based on expert knowledge is not feasible. Hence, we follow Gelman et al. (2008) in recommending default hyperparameter values, with predictors standardized to eliminate sensitivity to measurement units.

We first specify a_0 and b_0 , the hyperparameters in the component having mean fixed at zero. Coefficients assigned to this component should have values close to zero. We choose a_0 and b_0 such that $\int_{-\epsilon}^{\epsilon} \text{MBEN}(\beta_j; 0, \tau^2, \alpha_j, \lambda, \gamma_0) d\beta_j = z$, where z is the prior probability that the coefficient drawn from the cluster with mean zero has a null effect and ϵ is a small positive constant. Setting $z = 0.90$ and $\epsilon = 0.1$ implies $a_0 = b_0 = 10^{-2}$. For the remaining components centered away from zero, if we choose $a_1 = b_1 = 1$, the priors will have prior credible intervals of unit width. We recommend choosing smaller

values for these hyperparameters which are large enough to encourage shrinkage but not so large as to overwhelm the data and arbitrarily force a large number of components to be generated. The DP precision parameter α_0 is set to be 1 as a common default choice in DP models.

The parameters c_1 and d_1 , the prior mean and variance for the location parameter of the non-zero components, are chosen so that μ_t^* will have a relatively wide range of support. We also want to avoid setting d_1 to be too large, leading to proposing unlikely prior locations and over-favoring of allocation to the zero component. Therefore, we choose $c_1 = 0$ and $d_1 = 5$. For priors of λ , we set the shape parameter $r_0 = 1/100$ and the rate parameter $s_0 = 1/100$ so that the prior mean for λ concentrates on 1 and the prior variance is 100. As for the original elastic net or the Bayesian elastic net, one may use cross validation to set λ and the range of λ is dictated by the desired level of model sparseness. Our paper gives a fully Bayesian approach and has the advantage of adjusting λ automatically during the updating steps.

2.3.2 Gibbs Sampling Algorithm

We propose a Gibbs sampling algorithm which is a hybrid of the slice sampling scheme of Damien and Wakefield (1999) and the exact block Gibbs sampler of Yau et al. (2010). The exact block Gibbs sampler combines characteristics of the retrospective sampler of Papaspiliopoulos and Roberts (2008) and the slice sampler of Walker (2007), modifying the block Gibbs sampler of Ishwaran and James (2001) to avoid truncation approximations. We shall focus on the model with $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I})$ and $\tau^2 \sim \text{IG}(c_0, d_0)$.

1. Update $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ by sampling from the conditional posterior

$$(\boldsymbol{\beta} | \dots) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \boldsymbol{\Omega})^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(\mathbf{X}'\mathbf{y} + \boldsymbol{\Omega}\boldsymbol{\mu}^*)$ and $\boldsymbol{\Omega}$ is a $p \times p$ diagonal matrix with j th element $\alpha_{k_j}^* + \lambda$.

2. Update parameter λ through

$$\begin{aligned}
p(\lambda|\dots) &\propto \prod_{j=1}^p (\alpha_{k_j}^* + \lambda)^{\frac{1}{2}} \exp\left\{-\frac{(\alpha_{k_j}^* + \lambda)(\beta_j - \mu_{k_j}^*)^2}{2\tau^2}\right\} \prod_{t=1}^L \left(\frac{\alpha_t^*}{\alpha_t^* + \lambda}\right)^{\frac{1}{2}} \\
&\quad \lambda^{r_0-1} \exp(-s_0\lambda) \\
&\propto \prod_{t=1}^L (\alpha_t^* + \lambda)^{\frac{m_t-1}{2}} \lambda^{r_0-1} \exp\left[-\lambda\left\{\frac{\sum_{j=1}^p (\beta_j - \mu_{k_j}^*)^2}{2\tau^2} + s_0\right\}\right].
\end{aligned}$$

With the slice sampling scheme by Damien and Wakefield (1999), we first generate

$$\begin{aligned}
w_t &\sim \text{Unif}\left[0, (\alpha_t^* + \lambda)^{\frac{m_t-1}{2}}\right], \quad \text{for } t = 1, \dots, L, \\
w_0 &\sim \text{Unif}\left[0, \lambda^{r_0-1}\right].
\end{aligned}$$

We then get the range for λ as $\lambda \in \left[\max\left(\max_t(w_t^{\frac{2}{m_t-1}} - \alpha_t^*), 0\right), (w_0)^{\frac{1}{r_0-1}}\right] \doteq [l_0, u_0]$. Hence the posterior distribution for λ is a truncated exponential distribution with parameter $\frac{\sum_{j=1}^p (\beta_j - \mu_{k_j}^*)^2}{2\tau^2} + s_0$ within the range $[l_0, u_0]$.

3. Update τ^2 from its full conditional

$$\tau^2 \sim \text{IG}\left(\frac{n+p}{2} + c_0, \frac{1}{2}\left\{\sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \sum_{j=1}^p (\alpha_{k_j}^* + \lambda)(\beta_j - \mu_{k_j}^*)^2\right\} + d_0\right).$$

4. The joint prior distribution of k_j and a latent variable u_j can be written as

$$f(k_j, u_j|\pi) = \sum_{t:\pi_t > u_j} \delta_j(\cdot) = \sum_{t=1}^{\infty} \mathbf{1}(u_j < \pi_t) \delta_t(\cdot).$$

Implement Exact Block Gibbs sampler steps:

- i. Sample $u_j \sim \text{uniform}(0, \pi_{k_j})$, for $j = 1, \dots, p$ with $\pi_t = V_t \prod_{s < t} (1 - V_s)$.

ii. Sample the stick-breaking random variables

$$V_t \sim \text{beta}\left(1 + m_t, \alpha_0 + \sum_{s=t+1}^L m_s\right),$$

for $t = 1, \dots, L$, with L the minimum value satisfying $\pi_1 + \dots + \pi_L > 1 - \min\{u_j\}$.

iii. Sample μ_t^* for $t = 1, \dots, L$ by

i For $t = 1$, since the prior for μ_1^* has unit probability mass at 0, the posterior distribution still has probability mass at 0.

ii For $2 \leq t \leq L$,

$$\mu_t^* \sim \text{N}\left(\frac{\tau^{-2}(\alpha_t^* + \lambda) \sum_{j:k_j=t} \beta_j + d_1^{-1} c_1}{\tau^{-2} m_t (\alpha_t^* + \lambda) + d_1^{-1}}, \left\{ \frac{m_t (\alpha_t^* + \lambda)}{\tau^2} + \frac{1}{d_1} \right\}^{-1}\right),$$

where m_t is the number of regression coefficients assigned to component t .

iv. Sample α_t^* for $t = 1, \dots, L$ by

$$p(\alpha_t^* | \dots) \propto (\alpha_t^*)^{-3/2} \exp\left\{-\frac{\alpha_t^* \sum_{k_j=t} (\beta_j - \mu_t^*)^2}{2\tau^2}\right\} \exp\left(-\frac{\gamma/2}{\alpha_t^*}\right) (\alpha_t^* + \lambda)^{(m_t-1)/2}.$$

We still apply the slice sampling scheme of Damien and Wakefield (1999) to get the above posterior distribution of α_t^* at the $(i+1)$ th iteration,

$$\begin{aligned} \max \left\{ 0, (\alpha_t^{*(i)} + \lambda) \exp\left(-\frac{2e_3}{m_t - 1} - \lambda\right) \right\} &\leq \alpha_t^{*(i+1)} \\ &\leq \min \left\{ \alpha_t^{*(i)} \exp\left(\frac{2}{3}e_1\right), \frac{\gamma}{2e_2 + (\alpha_t^{*(i)})^{-1}\gamma} \right\}. \end{aligned}$$

where e_1, e_2 and e_3 are independent exponential random variates with mean 1.

Denote this range as $\alpha_t^{*(i+1)} \in [l_t, u_t]$ and

$$\begin{aligned} p(\alpha_t^{*(i+1)} | \dots) &\sim \text{Exp}\left(\frac{\sum_{k_j=t} (\beta_j - \mu_t^*)^2}{2\tau^2}\right) \mathbf{1}(l_t \leq \alpha_t^{*(i+1)} \leq u_t) \\ &\doteq \text{Exp}(\theta_t) \mathbf{1}(l_t \leq \alpha_t^{*(i+1)} \leq u_t) \end{aligned}$$

where $\text{Exp}(\theta_t)$ is the exponential distribution with parameter θ_t . If no observation is assigned to a specific cluster, then μ_t^* and α_t^* are drawn directly from the prior distribution.

v. Sample k_j for $j = 1, \dots, p$ from the multinomial conditional with

$$\Pr(k_j = t | \dots) \propto \mathbf{1}(u_j < \pi_t) \text{N}(\beta_j | \mu_t^*, \alpha_t^*), \quad \text{for } t = 1, \dots, L.$$

5. Sample the mixing parameter γ with the conjugate prior through

$$\gamma_0 \sim \mathbf{G}\left(a_0 + 1, 1/(2\alpha_1^*) + b_0\right), \quad \gamma_1 \sim \mathbf{G}\left(a_1 + L - 1, \sum_{t=2}^L 1/(2\alpha_t^*) + b_1\right).$$

The above updating algorithm can be easily modified to fit the logistic regression model and the MTL model.

2.4 Sparse Point Estimation via Automated Monte Carlo EM

In the Wikipedia project, some articles have tens of thousands of predictors (stemmed words) but only a few dozen links to other articles, so inference is difficult. Although the MCMC algorithm proposed in the previous section for fully Bayes posterior computation is quite efficient, problems can arise when scaling computation to very large p , particularly if it is important to produce results quickly. From the MCMC implementation, one can obtain posterior probabilities of allocation to the zero component for each of the predictors and the tendency will be to shrink the coefficient for unimportant predictors close to zero. However, none of the estimated coefficients will be exactly zero. From the standpoint

of producing a simpler model that may be more interpretable and which provides dimensionality reduction (thus eliminating the need to store all the predictors), it is appealing to obtain coefficient estimates that are exactly zero.

In order to address the dual goal of substantially shortened computational time and sparse estimation, this section proposes an expectation maximization (EM) algorithm for maximum a posteriori (MAP) estimation under a finite approximation to the model proposed in Section 2. In particular, in place of the Dirichlet process, we use the finite approximation to the Dirichlet process proposed by Ishwaran and Zarepour (2002). This approximation is useful since the EM algorithm has difficulty with the stick-breaking representation because of the non-exchangeability of the components. In particular, the EM algorithm converges to a local mode, leading to inferior inference.

More specifically, in expression (2.6), we let

$$k_j \sim \sum_{t=1}^L \pi_t \delta_t, \quad (\pi_1, \dots, \pi_L)' \sim \text{Dirichlet}(\alpha_0/L, \dots, \alpha_0/L).$$

The hierarchical decomposition of the MBEN prior allows use of the EM algorithm to implement the elastic net criterion in (2.2). This can be done simply, by regarding $\mathbf{k} = (k_1, \dots, k_p)'$, $\boldsymbol{\mu}^* = (\mu_{k_1}^*, \dots, \mu_{k_p}^*)'$ and $\boldsymbol{\alpha}^* = (\alpha_{k_1}^*, \dots, \alpha_{k_p}^*)$ as *hidden/missing data*. Set $\boldsymbol{\phi} = (\lambda, \boldsymbol{\mu}^*, \boldsymbol{\alpha}^*, \gamma)$ and $\boldsymbol{\psi} = (\boldsymbol{\beta}, \tau^2)$. Each iteration consists of an E-step and an M-step. The $(r + 1)$ th E-step entails the calculation of

$$Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(r)}) = E[\log\{f(\mathbf{y}, \boldsymbol{\phi}, \mathbf{k}; \boldsymbol{\psi})\} | \mathbf{y}; \boldsymbol{\psi}^{(r)}], \quad (2.8)$$

where $f(\mathbf{y}, \boldsymbol{\phi}, \mathbf{k}; \boldsymbol{\psi})$ represents the joint density of the *complete data* and $\boldsymbol{\psi}^{(r)}$ denotes the value of $\boldsymbol{\psi}$ from the r th iteration. The new value $\boldsymbol{\psi}^{(r+1)}$ is obtained by maximizing $Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(r)})$, which serves as the M-step.

We cannot get an analytical evaluation of equation (2.8) under our structure with the MBEN prior. A modified Monte Carlo EM algorithm by Wei and Tanner (1990) avoids

this difficulty by replacing the expectation in the E-step with a Monte Carlo approximation. Booth and Hobert (1999) improves the Monte Carlo EM algorithm by suggesting a rule for automatically increasing the Monte Carlo sample size after iterations in which the true EM step is swamped by Monte Carlo error. More specifically, a Monte Carlo approximation of $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(r)})$ is given by

$$Q_m(\boldsymbol{\psi}|\boldsymbol{\psi}^{(r)}) = \frac{1}{m} \sum_{l=1}^m \log \{f(\mathbf{y}, \boldsymbol{\phi}_{r,l}, \mathbf{k}_{r,l}; \boldsymbol{\psi})\}. \quad (2.9)$$

The Monte Carlo EM algorithm involves the use of Q_m in place of Q . An appropriate value for m is chosen after each iteration, and the algorithm is stopped when changes in the parameter estimates are small (after taking account of Monte Carlo error).

Define

$$Q^{(1)}(\boldsymbol{\psi}|\boldsymbol{\psi}') = \frac{\partial}{\partial \boldsymbol{\psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}'), \quad Q^{(2)}(\boldsymbol{\psi}|\boldsymbol{\psi}') = \frac{\partial^2}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} Q(\boldsymbol{\psi}|\boldsymbol{\psi}').$$

Then

$$\begin{aligned} \mathbf{0} &= Q_m^{(1)}(\boldsymbol{\psi}^{(r+1)}|\boldsymbol{\psi}^{(r)}) \approx Q_m^{(1)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)}) + (\boldsymbol{\psi}^{(r+1)} - \boldsymbol{\psi}^{*(r+1)})' Q_m^{(2)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)}), \\ &\Rightarrow \boldsymbol{\psi}^{*(r+1)} = \boldsymbol{\psi}^{(r+1)} + Q_m^{(2)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)})^{-1} Q_m^{(1)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)})' \end{aligned} \quad (2.10)$$

where $\boldsymbol{\psi}^{*(r+1)}$ satisfies $Q^{(1)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)}) = \mathbf{0}$. Conditionally on $\boldsymbol{\psi}^{(r)}$, $\boldsymbol{\psi}^{(r+1)}$ is approximately normal with mean $\boldsymbol{\psi}^{*(r+1)}$ and variance

$$\text{var}(\boldsymbol{\psi}^{(r+1)}|\boldsymbol{\psi}^{(r)}) \approx Q_m^{(2)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)})^{-1} \text{var}\{Q_m^{(1)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)})\} Q_m^{(2)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)})^{-1} \quad (2.11)$$

A sandwich estimate of $\text{var}(\boldsymbol{\psi}^{(r+1)}|\boldsymbol{\psi}^{(r)})$ is obtained by substituting $\boldsymbol{\psi}^{(r+1)}$ in place of $\boldsymbol{\psi}^{*(r+1)}$ on the right hand side of the above equation with the estimate

$$\begin{aligned} \hat{\text{var}}\{Q_m^{(1)}(\boldsymbol{\psi}^{*(r+1)}|\boldsymbol{\psi}^{(r)})\} &= \frac{1}{m^2} \sum_{l=1}^m \left(\frac{\partial}{\partial \boldsymbol{\psi}} \log \{f(\mathbf{y}, \boldsymbol{\phi}_{r,l}, \mathbf{k}_{r,l}; \boldsymbol{\psi}^{(r+1)})\} \right) \\ &\times \left(\frac{\partial}{\partial \boldsymbol{\psi}} \log \{f(\mathbf{y}, \boldsymbol{\phi}_{r,l}, \mathbf{k}_{r,l}; \boldsymbol{\psi}^{(r+1)})\} \right)'. \end{aligned}$$

After the $(r+1)$ th iteration, we construct an approximate $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\psi}^{*(r+1)}$ by using the normal approximation shown in the above. If the previous value $\boldsymbol{\psi}^{(r)}$ lies in that region, then the EM step was swamped by Monte Carlo error, and m should be increased by $m = m + m/k$, $k \in \{2, 3, 4\}$, as suggested by Booth and Hobert (1999).

The algorithm is stopped when

$$\max_i \left(\frac{|\psi_i^{(r+1)} - \psi_i^{(r)}|}{|\psi_i^{(r)}| + \delta_1} \right) < \delta_2, \quad (2.12)$$

where $\delta_1 = \delta_2 = 0.001$ is the common choice. Since the expectation in the E-step cannot be calculated analytically in our context, we run the risk of stopping prematurely, where the algorithm stops when $\boldsymbol{\psi}^{(r+1)}$ is very close to $\boldsymbol{\psi}^{(r)}$ only because a large Monte Carlo error is associated with $\boldsymbol{\psi}^{(r+1)}$. To reduce this risk, we use a second convergence criterion in conjunction with criterion (2.12) suggested by Booth and Hobert (1999). Their criterion is based on the change in the parameter estimates relative to their standard errors. More specifically, the algorithm will stop if

$$\max_i \left\{ \frac{|\psi_i^{(r+1)} - \psi_i^{(r)}|}{\sqrt{\text{var}(\hat{\psi}_i) + \delta'_1}} \right\} < \delta'_2, \quad (2.13)$$

where δ'_1 and δ'_2 should be suitably small and need not be the same as δ_1 and δ_2 in (2.12). The variance of $\hat{\psi}_i$ can be estimated by using the inverse of an estimate of the observed Fisher information evaluated at the current parameter estimate. Louis (1982) showed that the observed information matrix can be written as the sum of $-Q^{(2)}(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}})$ and $-\text{var} \left[\frac{\partial}{\partial \boldsymbol{\psi}} \log \{f(\mathbf{y}, \boldsymbol{\phi}; \boldsymbol{\psi})\} | \mathbf{y}, \hat{\boldsymbol{\psi}} \right]$, which is the complete information minus the missing information. More details about stopping rules and updating schemes for m are given in Booth and Hobert (1999).

Generalizing the above analysis, we have the following automated Monte Carlo EM algorithm steps:

1. Set starting values $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_0, \tau_0^2)$ and $r = 0$.
2. Gibbs sampling. In the r th iteration, we repeatedly sample k_1, \dots, k_p and $\phi_{k_1}, \dots, \phi_{k_p}$ for m times, as follows:
 - i. Update $p(\pi_1, \dots, \pi_L | \dots) \sim \text{Dirichlet}(\alpha_0/L + m_1, \dots, \alpha_0/L + m_L)$, where m_l is the number that $k_j = l$ for $j = 1, \dots, p$.
 - ii. Update k_l from a multinomial distribution with probabilities

$$\Pr(k_j = l) \propto \pi_l \mathbf{N}(\beta_j; \mu_{k_j}^*, \tau^2(\alpha_{k_j}^* + \lambda)^{-1}) f(\mu_{k_j}^*, \alpha_{k_j}^*),$$

where $f(\mu_{k_j}^*, \alpha_{k_j}^*)$ is the prior for $(\mu_{k_j}^*, \alpha_{k_j}^*)$. Label switching moves by Papaspiliopoulos and Roberts (2008) are needed here to assist the algorithm to jump across modes.

- iii. Sample $\lambda, \gamma, (\alpha_1^*, \dots, \alpha_L^*)$ and $(\mu_1^*, \dots, \mu_L^*)$ through the same steps as in the above MCMC algorithm, where $\{\mu_l^*, \alpha_l^*\}$ are cluster specific parameters.
3. (E-Step). Calculate equation (2.9) by Monte Carlo integration.
4. (M-Step). Obtain $\boldsymbol{\psi}^{*(r+1)}$ from (2.10) and the $100(1 - \alpha)\%$ confidence ellipsoid for $\boldsymbol{\psi}^{*(r+1)}$ with (2.11) to decide whether to change m .

Repeat Steps 2 through 4 until the algorithm reaches the stopping rule.

2.5 Simulation Study

The intuitive appeal of allowing shrinkage of parameter estimates to multiple locations is clear, but the potential gain must be confirmed by simulation. We examine the performance of our two approaches in comparison with the lasso Tibshirani (1996), the LARS-EN Zou

and Hastie (2005) and ridge regression Hoerl and Kennard (1988). Results for the latter three methods are obtained from **R** package **elasticnet**, **lars** (both with ten fold cross validation) and **MASS**. Our first example is similar to the one from the original elastic net paper Zou and Hastie (2005). Specifically, we simulate data from the true model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathbf{N}_n(0, \sigma^2 I_n),$$

where \mathbf{X} is the predictor matrix with dimension $n \times p$, where $n = 200$, $p = 40$ and $\sigma^2 = 1$.

We simulate 50 data sets. The predictors \mathbf{X} are generated by

$$\begin{aligned} \mathbf{x}_i &= \mathbf{Z}_1 + \epsilon_i^x, & \mathbf{Z}_1 &\sim \mathbf{N}(0, I_n), & i &= 1, \dots, 5, \\ \mathbf{x}_i &= \mathbf{Z}_2 + \epsilon_i^x, & \mathbf{Z}_2 &\sim \mathbf{N}(0, I_n), & i &= 6, \dots, 10, \\ \mathbf{x}_i &= \mathbf{Z}_3 + \epsilon_i^x, & \mathbf{Z}_3 &\sim \mathbf{N}(0, I_n), & i &= 11, \dots, 15, \\ \mathbf{x}_i &\stackrel{\text{i.i.d.}}{\sim} \mathbf{N}(0, I_n), & & & i &= 16, \dots, 40, \end{aligned}$$

where the ϵ_i^x are independent and identically distributed $\mathbf{N}(0, 0.01I_n)$, for $i = 1, \dots, 15$. The true regression coefficient is set to $\boldsymbol{\beta} = (3, \dots, 3, 0, \dots, 0)'$ with the first 15 elements equaling 3 and the rest 0. Thus, we have three equally important groups, and within each group there are five highly correlated members. The last 25 are pure noise features. An ideal method would select out the 15 true features while setting the coefficients for the 25 noise features equal to zero. A second set of simulations is performed for 50 additional data sets, each with $p > n$: here $n = 150$ and $p = 200$. The first 15 elements of $\boldsymbol{\beta}$ are set to 3 and the rest to 0. A third simulation is performed for another 50 data sets under the extreme situation where $p \gg n$: here $n = 200$ and $p = 2000$. Finally, we compare the performances of all of the methods when the true effects are not sparse, with $n = 200$ and $p = 200$ and the first 90 coefficients being 3 and the rest being 0.

In each simulation, for the lasso and elastic net, 50 observations are used to select tuning parameters, 50 observations are used to fit the model and test errors are calculated from the remaining observations. We use the first 100 observations to fit the models and

the remaining test samples to calculate the test errors in our two proposed methods and ridge regression.

The MBEN (MBEN updated by MCMC algorithm) and MBEN (MCEM) (MBEN updated by Monte Carlo EM algorithm) tend to identify the correct coefficient clustering. For instance, prior location and scale parameters are often grouped into one cluster for the first 15 coefficients and a second cluster for the remaining coefficients. Each of the β coefficients is shrunk toward a cluster-specific prior mean, e.g., the first 15 or 90 coefficients are shrunk towards a prior mean that is close to 3 and the remaining coefficients are shrunk towards the null component.

We first compare the mean squared errors (MSE) of the estimated coefficients from the lasso, the elastic net, the ridge regression, the MBEN and the MBEN (MCEM) in the first three simulations. The MBEN and the MBEN (MCEM) have the smallest MSE for all the coefficients (MSEALL) over the first three simulated datasets (Figure 2.2). The elastic net performs robustly with only slight differences from the MBEN methods. The lasso performs the worst in terms of the MSEALL, with the MSEALL decreasing as the dimension increases. In terms of the mean MSE for the non-null coefficients (MSENULL, Figure 2.3), the lasso performs the worst in the first two simulations, while ridge regression performs worst in the third simulation. It is surprising that the elastic net performs worse than ridge regression in estimating the non-null coefficients in the second simulation. In the third simulation, the lasso, the elastic net and the MBEN and the MBEN (MCEM) perform almost identically. For the MSE of the null coefficients (MSENULL, Figure 2.4), since the estimates from the ridge regression cannot be exactly zero, ridge regression performs poorly in all the three simulations. In terms of the bias (Figure 2.5), the elastic net performs the worst in the first simulation, and ridge regression performs worst in the second and the third simulations. The MBEN and the MBEN (MCEM) obtain the smallest bias throughout the first three simulations. The superior performances of the MBEN and the MBEN (MCEM) result from including prior locations that are distinct from 0. How-

ever this could also lead to slightly poorer performance in estimating coefficients with null effects (Figure 2.4), since they may be shrunk toward a non-null component; however, it is difficult to see any deterioration as the dimension increases. In the fourth simulation, the lasso and the elastic net perform the worst in terms of the MSEALL, the MSENNULL and the bias. The reason can be seen clearly from Figure 2.6: both the lasso and the elastic net over-shrink the estimates when the true coefficients are not sparse.

Performances from all the methods are also summarized in Table 2.1 in terms of the mean-squared prediction errors (MSPE). The numbers in parentheses are the corresponding standard errors (of the means) estimated by the bootstrap with $B = 500$ resamplings on the 50 mean-squared errors. We note that ridge regression performs only slightly worse than the lasso and the elastic net in simulation 1 but is very poor in simulations 2 through 4. This makes sense since the ridge regression cannot obtain the exact zero estimates and the MSE from the null effects aggregates as the dimension increases. The lasso and the elastic net perform very robustly through the first three simulations and stand just behind the MBEN and MBEN (MCEM), but both are much worse in the fourth simulation. The MBEN and MBEN (MCEM) are more accurate than all the other methods across all four simulation studies. The MBEN (MCEM) performs almost the same as the MBEN but is, on average, 2 to 4 times faster, especially when the dimension becomes large. In the first three simulations, the MBEN (MCEM) converged in 45, 59, 90 and 65 iterations. The values of m increased from the initial value of 100 to 2550, 4750, 6775 and 5320 in the final iteration, respectively.

In the simulation studies, we also notice that the cross-validation is not robust and sometimes can cause unreasonable estimates when using the lasso and the elastic net. The MBEN and the MBEN (MCEM) are fully Bayesian methods and choose the tuning parameters automatically in each iteration, leading to more robust inference. Also, our proposed MBEN and MBEN (MCEM) are not much more time consuming than the lasso and the elastic net. In particular, cross-validation with the elastic net can be slow, and

possibly becomes stuck when the dimension is relatively large (e.g., $p = 2000$). The **R** function **enet** sometimes stops because the variances of some of the columns of the design matrix resulting from the cross-validation are zero. It happens when the design matrix is sparse, which is the situation for the Wikipedia.

2.6 Wikipedia Project Application

The Wikipedia is a free-content encyclopedia project based on an openly-editable model. Wikipedia’s articles provide links to guide the users to related articles with additional information. These articles are written by thousands of volunteer editors that collaborate to build a consensus on additions and changes. Since its creation in 2001, the Wikipedia has grown rapidly into one of the most popular reference sites on the internet.

Articles in the Wikipedia are organized into categories, and a single article may belong to multiple categories. Most categories have a number of subcategories, and the numbers of subcategories and the articles within them change over time. The numbers cited in this paper are accurate as of June 25, 2010.

There are 56 subcategories for “Statistics” and we focus on two of the subcategories. The first subcategory is “Continuous Distributions”. The second subcategory is “Bayesian Statistics”. Our analysis considers the corpus formed by the union of the articles in “Continuous Distributions” and “Bayesian Statistics”. We view each article in the corpus as a bag-of-words. From those bags we extracted the vocabulary, that is, the set of W words that appear, in total, in the entire corpus (we used a Perl script stemmer to identify similar words; e.g., “average”, “averages”, and “averaging” are treated as the same word). We count how many times word w appears in article n , and denote this by x_{nw} .

Besides the counts, we computed the importance of word w in article n , denoted by x'_{nw} according to the formula

$$x'_{nw} = \frac{x_{nw}}{\sum_w x_{nw}} \log \left(\frac{N}{1 + \sum_n 1(x_{nw} > 0)} \right),$$

where $1(x_{nw} > 0)$ indicates whether a word w appears in article n , and N is the total number of articles in the corpus (in this case, $N = 160$; 95 from “Continuous Distributions” and 65 from “Bayesian Statistics”, and no article is common to both subcategories). The “normal distribution” is included in the 65 links from “Bayesian Statistics” and the “Bayesian inference” is also included in the 95 links from “Continuous Distributions”. We exclude these two links and keep the remaining 158 links for analysis. This importance measure is known as the TF/IDF (term frequency/inverse document frequency) transformation and the importance is proportional to the number of times the word appears in a document, and inversely proportional to the number of times it appears in the corpus Spärck Jones (1972).

The immediate goal is to build a statistical model to predict which articles receive links from the selected central articles, using lexical information from all of the articles in these two subcategories. That goal is part of a larger project, to build a dynamic network model for growth and change in the Wikipedia. This represents a typical example of a general learning scenario called single-task learning (STL) when considering each central article independently and multi-task learning (MTL) when considering all articles simultaneously.

STL refers to the approach of making inference one task at a time, using only the data that directly correspond to the problem; it assumes that tasks are drawn independently from a pool in which each task requires unrelated sets of covariates. In contrast, MTL regards the tasks as potentially having association structure, in that similar sets of covariates are relevant across many or all tasks. This perspective allows transference of expertise among tasks, benefitting inference in all of the tasks and allowing joint solutions within a unified representation. By exploiting data from related tasks in MTL, the data for each task is partially combined and the resulting algorithm is improved.

In applying STL, we selected “normal distribution” and “Bayesian inference” as the central articles and consider all the articles in subcategory “continuous distribution” and

“Bayesian Statistics” as the corpus. We chose these two articles because they are lengthy, have large numbers of links to other articles, and contain a mix of distinct and common vocabulary. Also, these two central articles have similarities that suggest what is learned from one task could be transferable to another, enabling later comparison with MTL.

We first did analyses for these two central pages separately. For this application, to eliminate uninformative words (e.g., “of”, “the”, and so fourth) we use only those words whose TF/IDF score is greater than a threshold. This threshold was chosen after some exploration of a number of documents; our value is 0.0006. This leads to a set of 1052 and 1053 words, sorted in decreasing order according to the TF/IDF score. Using all the other articles in the category Statistics, we create a matrix of size 1052×158 and 1053×158 , where the rows are the selected words from “normal distribution” and “Bayesian inference”, with each column representing the counts of those words between the “normal distribution” (or “Bayesian inference”) and the other 158 articles in the two subcategories. The response is a 0-1 vector, indicating the presence of a link between the “normal distribution” (or “Bayesian inference”) and article j .

The number of times that a word appears in a given article is not the only variable included in the analysis. We can determine the divergence between the importance of a selected word in “normal distribution” (or “Bayesian inference”) and the importance of the same word in any other article. This can be done using the cross-entropy measure $H(p, q)$. The cross-entropy measure for two discrete distributions p and q is defined as:

$$H(p, q) = - \sum_z p(z) \log q(z). \quad (2.14)$$

In this case $p(z)$ is the distribution of the word counts in “normal distribution” or “Bayesian inference” and $q(z)$ is the distribution of the same words in article j , $j = 1, \dots, 158$. We expect that if a word w is important for “normal distribution” (or “Bayesian inference”) and also important for article j , then it is more likely that there is a link between them. Small values of $H(p, q)$ indicate less divergence between the articles.

The results for the MBEN that follow are based on 50,000 iterations obtained after a burn-in period of 5,000 iterations. The results for the MBEN (MCEM) are obtained after satisfying conditions (2.12) and (2.13) when setting δ_1 , δ_2 , δ'_1 , and δ'_2 to 0.001. No evidence of lack of convergence was found from the visual inspection of trace plots or from the Gelman and Rubin (1992) convergence test .

We use multiple random training-tests splits to choose 79 articles as the training set and the remaining 79 articles as the testing set. The predictive performances are summarized across different splits. We randomly select the training set and the testing set to escape the alphabetical order of the articles, otherwise perhaps Asymptotics, Asymptotic Normality, Asymptotic Limits, etc., are all in one or the other of the training/testing set. This random selection can also help test the robustness of all the methods. The design matrix \mathbf{X} is composed of the frequency of the bag of words and the entropies, with the first column being all 1's to account for the mean effect. For the STL, the validation and testing results are given in Table 2.2-3 for the lasso, the elastic net and the proposed MBEN and MBEN (MCEM). From the tables, we notice that estimation of the lasso and the elastic net yield inferior performances compared to that of the MBEN and the MBEN (MCEM). It is probably because of the lack of robustness to choice of the tuning parameters for the lasso and the elastic net, while our proposed methods update the tuning parameters in each iteration and obtain rather robust estimations. For the MTL, since the lasso and the elastic net do not allow different predictors across the tasks, we can only show the performances resulted from the MBEN and the MBEN (MCEM).

In STL, for “normal distribution”, the chosen words of the MBEN are “**random**”, “**estimator**”, “scores”, “sum”, “characteristic”, “intensity”, “n-1”, “errors”, “variable”, “~”, “bell”, “ σ^6 ” and **cross entropy**. For “Bayesian inference”, the significant words are “inference”, “bowl”, “guilty”, “posterior”, “theorem”, “test”, “probabilities”, “positive”, “court”, “odds”, “observed”, “let”, “testimony”, “0.5” and **cross entropy**. We notice that the words chosen for the “Bayesian inference” are related to legal issues. Words

in bold face are also chosen in the MBEN (MCEM).

In MTL with the MBEN, there are two other words are selected for “normal distribution”, which are “mean” and “ σ^4 ”; “theorem” is recognized as another significant word for “Bayesian inference”. For estimation with the MBEN (MCEM), only the word “mean” is selected for the normal distribution and no new word is selected for the “Bayesian Inference”. Training errors and testing errors are decreased as can be seen from Table 4 and Table 5.

2.7 Discussion

We have extended the elastic net model developed by Zou and Hastie (2005) to a semi-parametric Bayesian version, with tuning parameters marginalized out using a Bayesian approach instead of through cross-validation. An efficient MCMC algorithm and an automated Monte Carlo EM algorithm enable fast computation in high dimensions. The proposed multiple Bayes elastic net (MBEN) prior can be easily inserted into general Bayesian models, including generalized linear models, nonparametric regression models having many kernels or basis functions, and even hierarchical regression models for functional data and multi-task learning. A logistic link regression model has been developed for independent tasks and a multi-task MBEN model has been developed for simultaneously fitting multiple related models.

The performance of the new model has been assessed in several simulation examples. In sparse settings, we consider a low dimensional example with $p < n$, a high dimensional example with $p > n$ and an extremely high dimensional example with $p \gg n$. In non-sparse settings, we consider a relatively high dimension example with 90 out of 200 coefficients being non-zero. Performance is compared to the lasso, the elastic net and the ridge regression, and we find that the new model is superior according to a variety of criteria. The MSE for coefficient estimates is strongly decreased for coefficients having

non-null effects but is only slightly increased for coefficients with null effects. The MSPE decreases significantly compared to ridge regression, the lasso and the elastic net. Our proposed MBEN and MBEN (MCEM) perform robustly under different settings, while the lasso and the elastic net behave poorly when the true coefficients are not sparse and ridge regression performs poorly when the settings are sparse. The MBEN and the MBEN (MCEM) also show advantages in choosing robust tuning parameters.

The proposed methodology has been motivated by the challenging Wikipedia project, in which bags-of-words from articles are employed to try to infer links among articles. We have presented both single-task and multi-task results on classification performance and have also inferred words that are relevant for this classification task. The results from this application are interpretable and successfully distinguish links and non-links. As wanted, the MBEN selects multiple correlated words. The MBEN priors may be easily extended to other models and should be of use to researchers in many other settings.

Table 2.1:

Method	Results for the Data Simulation			
	Simulation 1	Simulation 2	Simulation 3	Simulation 4
lasso	3.10 (1.68)	2.42 (0.82)	15.91 (4.12)	23.93 (6.77)
elastic net	1.58 (0.48)	2.08 (0.88)	14.07 (4.04)	65.43 (57.48)
Ridge Regression	5.42 (2.59)	39.57 (13.19)	447.15 (65.27)	125.21 (87.69)
MBEN(Gibbs)	0.87 (0.10)	0.97 (0.13)	8.93 (1.21)	12.39 (2.49)
MBEN (MCEM)	0.97 (0.12)	1.00 (0.24)	9.34 (1.45)	16.78 (2.98)

Table 2.2:

Method	Training Error	Testing Error	# Words Chosen
lasso	10/79	16/79	10
elastic net	10/79	13/79	11
MBEN(Gibbs)	6/79	10/79	13
MBEN (MCEM)	7/79	11/79	11

Table 2.3:

Method	Training Error	Testing Error	# Words Chosen
lasso	13/79	17/79	9
elastic net	7/79	16/79	13
MBEN(Gibbs)	6/79	10/79	15
MBEN (MCEM)	6/79	12/79	14

Table 2.4:

Method	Training Error	Testing Error	# Words Chosen
MBEN(Gibbs)	5/79	6/79	15
MBEN (MCEM)	5/79	7/79	14

Table 2.5:

Method	Training Error	Testing Error	# Words Chosen
MBEN(Gibbs)	4/79	8/79	16
MBEN (MCEM)	6/79	10/79	14

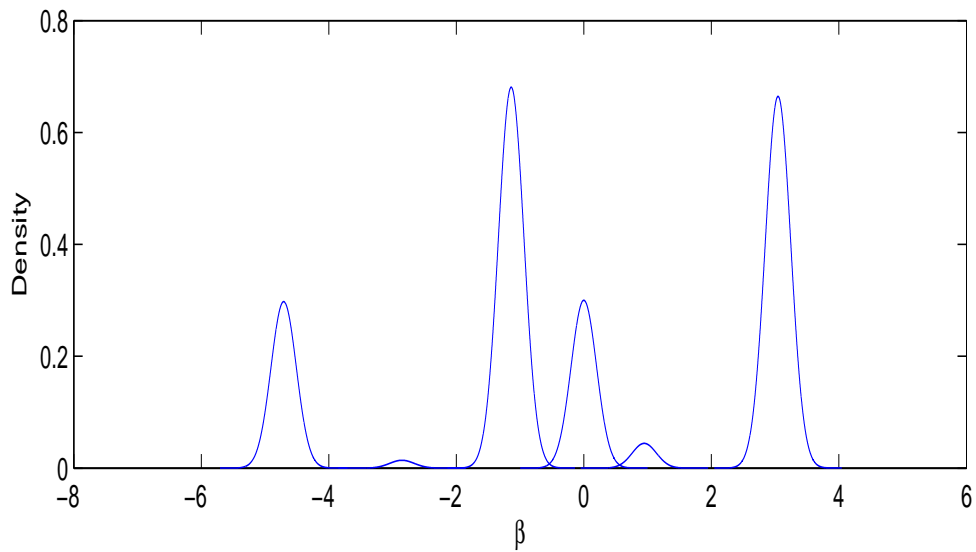


FIGURE 2.1: MBEN prior distribution with $\alpha = 1$, $c_1 = 0$ and $d_1 = 10$.

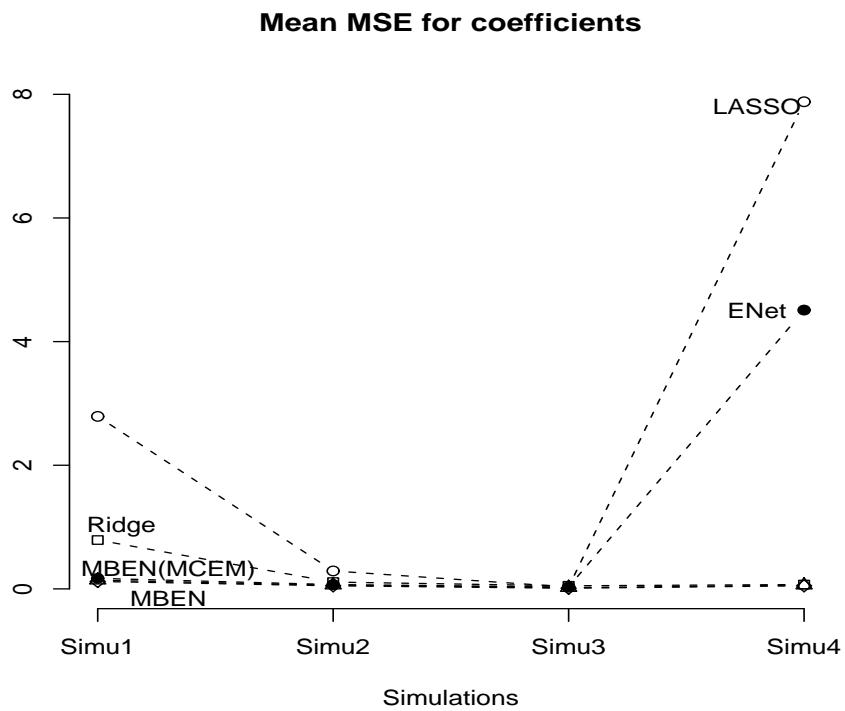


FIGURE 2.2: Mean MSE for all coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)

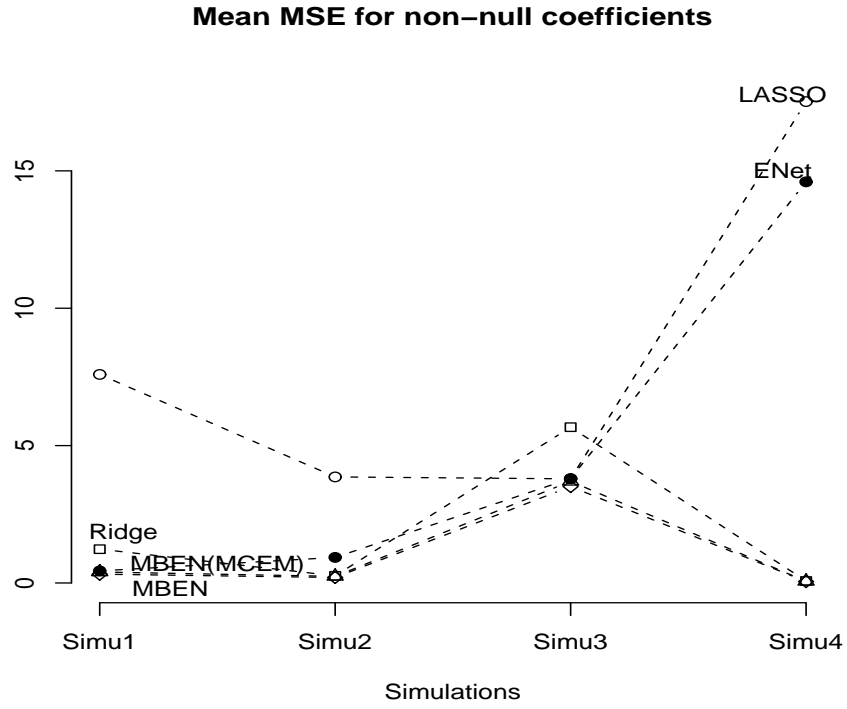


FIGURE 2.3: Mean MSE for none null coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)

Mean MSE for the null coefficients

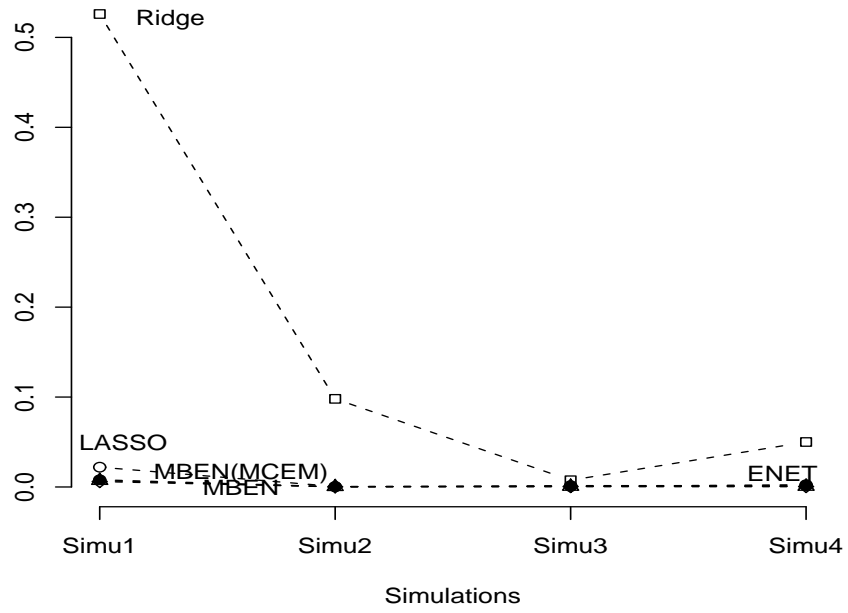


FIGURE 2.4: Mean MSE for null coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)

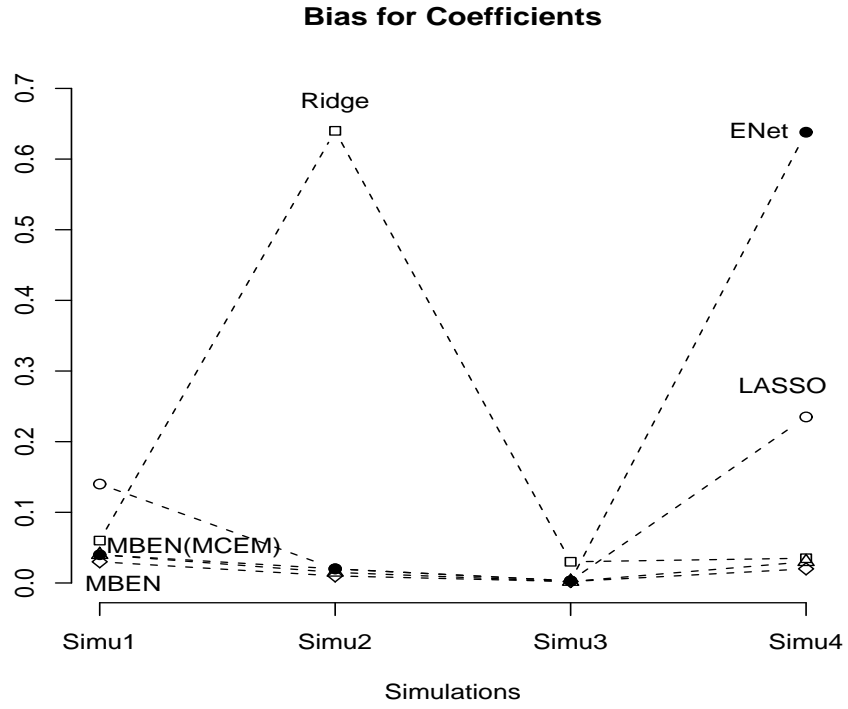


FIGURE 2.5: Bias for all coefficient estimates from the lasso (circle), elastic net (filled circle), ridge regression (square), MBEN (Gibbs) (diamond) and MBEN (MCEM) (triangle)

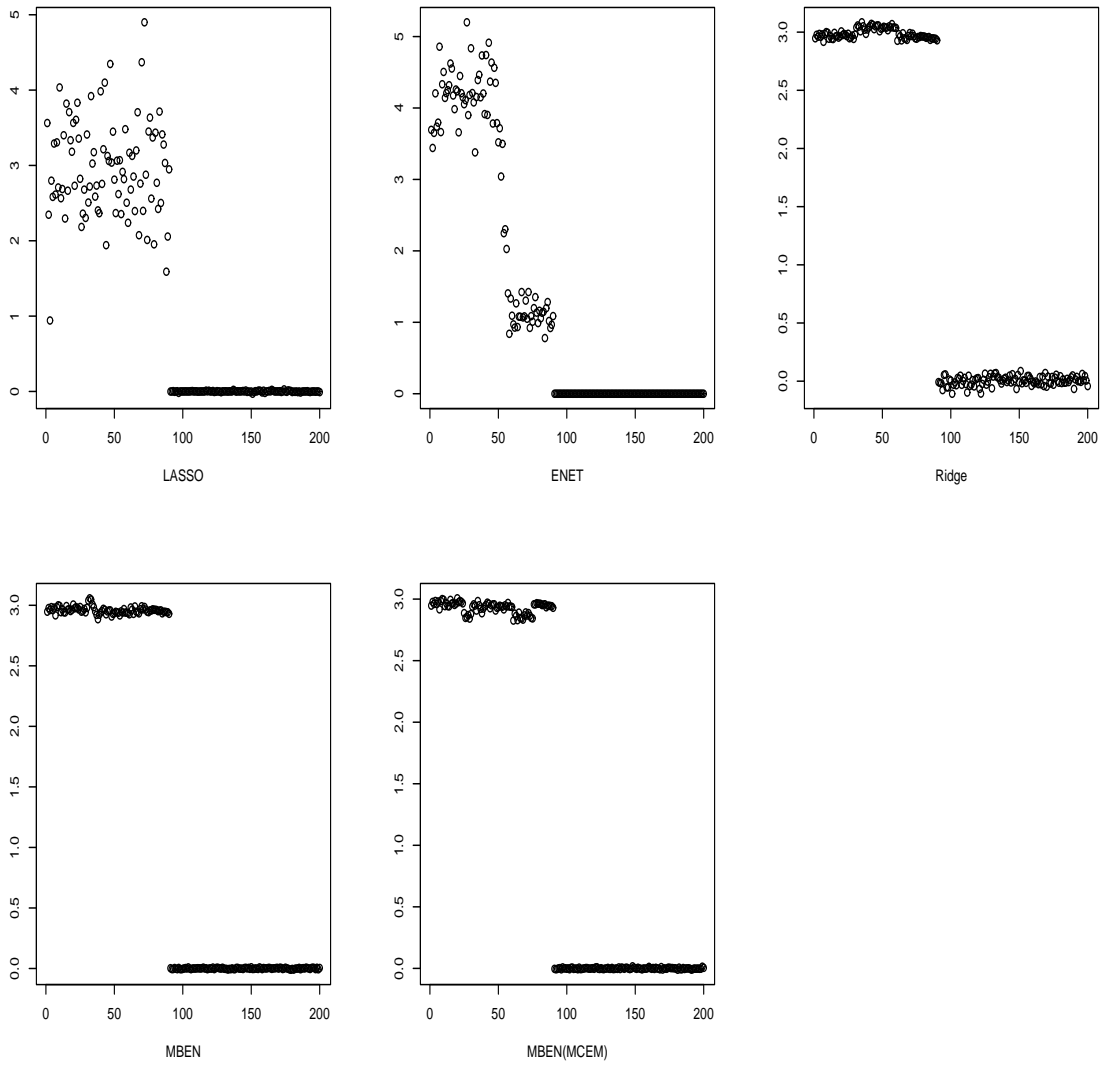


FIGURE 2.6: Coefficient estimates from the lasso, elastic net, ridge regression, MBEN (Gibbs) and MBEN (MCEM) in simulation 4.

Nonparametric Bayes Stochastically Ordered Latent Class Models

3.1 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 sub-populations or clusters, with the allocation probabilities potentially dependent on predictors. Suppose one collects response data $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})' \in \mathfrak{R}^p$ 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

$$f(\mathbf{y}_i | \mathbf{x}_i) = \sum_{k=1}^K \pi_k(\mathbf{x}_i) \mathbf{N}_p(\mathbf{y}_i; \mu_k, \Sigma_k), \quad (3.1)$$

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 (4.5) 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 post-processing can potentially be used Stephens (2000); Jasra et al. (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 (4.5) 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 (4.5) 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 pre-specified

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 et al. (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.

3.2 Stochastically Ordered Latent Class Priors

3.2.1 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,

$$P_i \sim \sum_{k=1}^{\infty} \pi_k(\mathbf{x}_i) \delta_{P_k^*}, \quad P_k^* = \sum_{l=1}^{\infty} v_l \delta_{\theta_{kl}}, \quad (3.2)$$

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.

There are two main distinct features of prior (3.2) 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). We used the rDDP mixture prior instead to avoid the partitioning effect of the Polya tree prior. Such an effect can be removed using mixtures of Polya 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 $v_l = \nu_l \prod_{s<l} (1 - \nu_s)$ with $\nu_l \sim \text{beta}(1, \alpha_2)$ 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 $\theta_{1l} = \gamma_{1l}^* \sim \text{N}(m_0, s_0^2)$ and $\gamma_{kl}^* = \theta_{kl} - \theta_{k-1,l}$, for $k = 2, \dots, \infty$, with

$$\gamma_{kl}^* \sim w_0 \delta_0 + (1 - w_0) \mathbf{N}^+(0, \kappa^{-1}), \quad k = 2, \dots, \infty, \quad (3.3)$$

where $w_0 = \Pr(\gamma_{kl}^* = 0)$ and \mathbf{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

$$\pi_k(\mathbf{x}) = \frac{\psi_k \exp\{\mathbf{x}'\boldsymbol{\beta}_k\}}{\sum_{l=1}^K \psi_l \exp\{\mathbf{x}'\boldsymbol{\beta}_l\}}, \quad \psi_k \sim \text{Gamma}(\alpha_1/K, 1), \quad \boldsymbol{\beta}_k \sim H, \quad (3.4)$$

where $\psi_k \geq 0$ is a baseline weight for mixture component k , $\boldsymbol{\beta}_k$ 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 (3.4). 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 model (3.2) - (3.4), 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 A) that the prior probability of clustering two groups in this baseline case is,

$$\Pr(P_i = P_{i'}) = \mathbb{E}\left(\sum_{k=1}^K \pi_k \pi_k\right) = \frac{1 + \alpha_1/K}{1 + \alpha_1}, \quad (3.5)$$

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,

$$\Pr(P_i < P_{i'}) = \frac{1}{2} \left\{ 1 - \Pr(P_i = P_{i'}) \right\} = \frac{\alpha_1}{2(1 + \alpha_1)} \left(1 - \frac{1}{K} \right), \quad (3.6)$$

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 6 (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, (3.5) and (3.6) 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.

3.2.2 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,

$$\begin{aligned}
y_{ijt}^* &= h_t(y_{ijt}), \quad h_t(y) = y, t = 1, \dots, p_1, \quad h_t(y) = 1(y > 0), t = p_1 + 1, \dots, p \\
\mathbf{y}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\Lambda}\eta_{ij} + \boldsymbol{\epsilon}_{ij}, \quad \epsilon_{ijt} \sim \mathbf{N}(0, \sigma_{it}^2), \\
\eta_{ij} &\sim f_i, \quad f_i(\eta) = \int \mathcal{K}(\eta; \theta) dP_i(\theta), \\
\mathcal{K}(\eta; \theta) &= \int \mathbf{N}(\eta; \theta, \varphi) dQ(\varphi), \tag{3.7}
\end{aligned}$$

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 $\mathcal{K}(\cdot; \theta)$ 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 non-negative so that $\lambda_t \geq 0$ for $t = 1, \dots, p$. For the scale mixture component, we let $Q \sim DP(\alpha_0 Q_0)$ 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 $\mathcal{K}(\cdot; \theta)$ is monotonically stochastically increasing in θ , we maintained 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 (3.7).

To complete a Bayesian specification of the SO-LCM model in (3.7), we choose priors as follows. The intercept vector is assigned a normal prior, $\mu_t \sim \text{N}(\mu_0, \sigma_0^2)$ for $t = 1, \dots, p$, and the factor loadings are assigned robust truncated Cauchy priors by letting $\lambda_t \sim \text{N}^+(0, \tau)$ 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.

3.3 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. (2010)

for posterior computation and update polychotomous weights through Holmes and Held (2006). We will focus on the simple model $y_{ij} \sim f_i$, $f_i(\eta) = \int \mathcal{K}(\eta; \theta) dP_i(\theta)$, $\mathcal{K}(\eta; \theta) = \int \mathbf{N}(\eta; \theta, \varphi) dQ(\varphi)$, $\{P_i\} \sim \text{SO-LCM}$, as in expression (3.2) and (3.4), 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

$$\zeta_i \sim \mathcal{M}(1; \pi_1(\mathbf{x}_i), \dots, \pi_K(\mathbf{x}_i)), \quad \pi_k(\mathbf{x}_i) = \frac{\psi_k \exp(\mathbf{x}_i' \boldsymbol{\beta}_k)}{\sum_{l=1}^K \psi_l \exp(\mathbf{x}_i' \boldsymbol{\beta}_l)}, \quad (3.8)$$

where ζ_i is the procedure cluster indicator and $\mathcal{M}(1; \cdot)$ denotes the single sample multinomial distribution. Defining $\boldsymbol{\beta}_{[k]} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}_{k+1}, \dots, \boldsymbol{\beta}_K)$, we have

$$\begin{aligned} L(\boldsymbol{\beta}_k | \boldsymbol{\zeta}, \boldsymbol{\beta}_{[k]}) &\propto \prod_{i=1}^n \prod_{k=1}^K [\pi_k(\mathbf{x}_i)]^{1(\zeta_i=j)} \\ &\propto \prod_{i=1}^n [\chi_{ij}]^{1(\zeta_i=j)} [1 - \chi_{ij}]^{1(\zeta_i \neq j)} \end{aligned} \quad (3.9)$$

where

$$\begin{aligned} \chi_{ij} &= \frac{\exp\{\mathbf{x}_i' \boldsymbol{\beta}_j + \log(\psi_j) - C_{ij}\}}{1 + \exp\{\mathbf{x}_i' \boldsymbol{\beta}_j + \log(\psi_j) - C_{ij}\}} = \frac{\exp(\hat{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}_j - C_{ij})}{1 + \exp(\hat{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}_j - C_{ij})}, \\ C_{ij} &= \log \left[\sum_{k \neq j} \exp\{\mathbf{x}_i' \boldsymbol{\beta}_k + \log(\psi_k)\} \right] = \log \left[\sum_{k \neq j} \exp(\hat{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}_k) \right] \end{aligned} \quad (3.10)$$

where $\hat{\mathbf{x}}_i = (\mathbf{x}_i', 1)'$ and $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{\beta}_j', \log \psi_j)'$. The prior for $\log(\psi_k)$ from (3.4) can be approximated as $\log(\psi_k) \stackrel{\mathcal{D}}{\approx} \mathbf{N}(m, v)$. We use this approximation to obtain an efficient Metropolis independence chain proposal. The conditional likelihood

$L(\hat{\beta}_j|\zeta, \hat{\beta}_{[j]})$ 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

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

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 re-run.

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

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

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

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

4. Sample γ_l^* from

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

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 (3.3).

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 $\text{Gamma}(a_2, b_2)$ directly from

$$(\alpha_2 | \dots) \sim \text{Gamma} \left(a_2 + K(L-1), b_2 - K \sum_{l=1}^{L-1} \log(1 - v_l) \right)$$

Note that this algorithm can be generalized easily to accommodate model (3.7), so the details are omitted.

3.4 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 (3.7) is studied and both simulations mimic the structure of the medical procedure data.

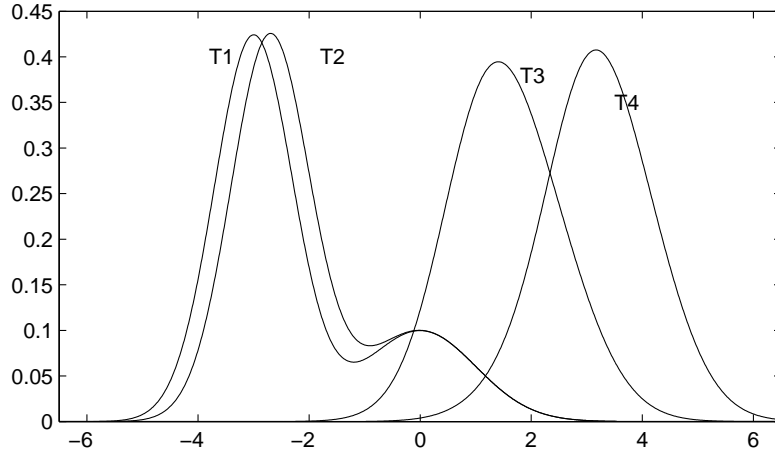


FIGURE 3.1: True distributions used in simulation study 4.1

3.4.1 Without predictors

Data y_{ij} are generated according to (3.7), 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

$$\boldsymbol{\mu} = (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4)',$$

$\boldsymbol{\Lambda} = \mathbf{1}'_{p \times 1}$ 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 fifteen procedures being generated from mixture distribution T_1 , the second fifteen procedures generated from T_2 , the third fifteen procedures generated from T_3 and the last fifteen 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 3.1 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 et al. (2005) to the data $\{\bar{\mathbf{y}}_1, \dots, \bar{\mathbf{y}}_n\}$ using the **R** functions available in the package described in Fraley et al. (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 3.2 plots the BIC for the simulated data vs 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 (3.3), 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

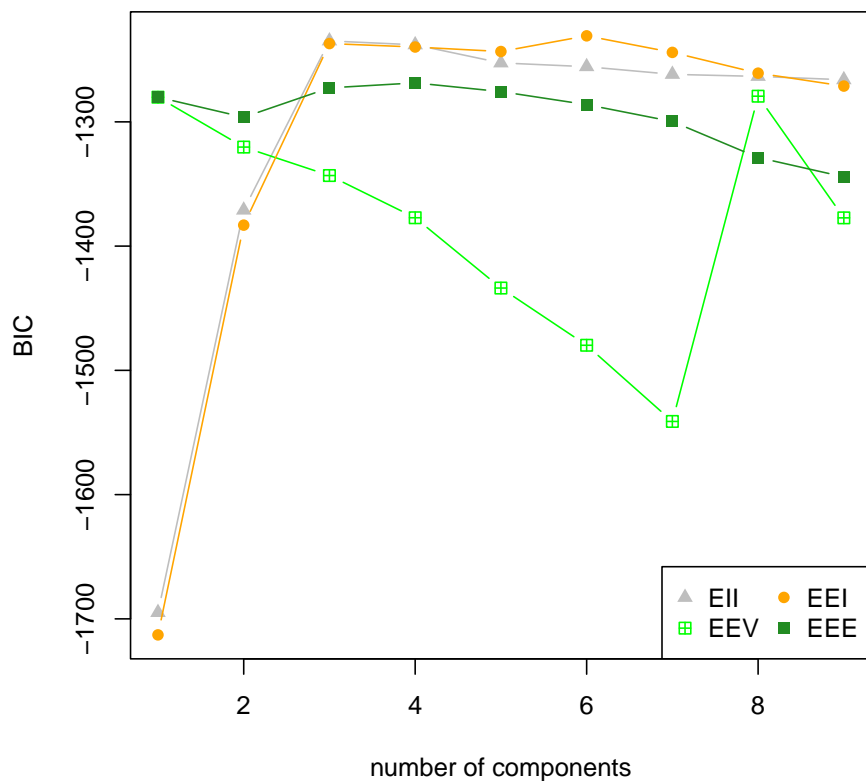


FIGURE 3.2: Frequentist model-based clustering results implemented via the EM algorithm using the `Mclust` function in **R** in simulation study 4.1, with the different symbols representing different model assumptions. EII: spherical, equal volume; EEI: spherical, equal volume and shape; EEV: spherical, equal volume but varying orientation; EEE: ellipsoidal, equal volume and shape.

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 $P_i < P_{i'}$ and Figure 3(b) is the clustering plot with the (i, j) th entry identifying the probability for $P_i = P_{i'}$. Figure 3 illustrates that there is not enough information in the data to differentiate the first thirty 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

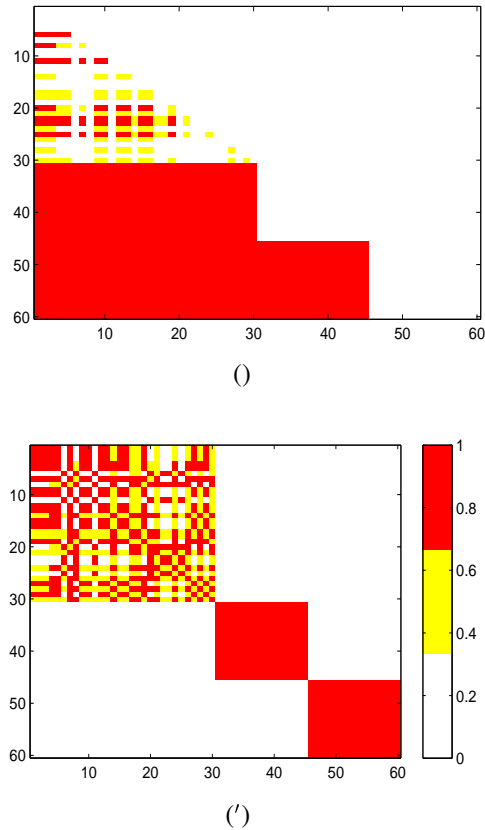


FIGURE 3.3: Posterior probability for ranking and clustering in study of section 4.1 with entry (i, j) in (a) being the lower triangular matrix identifying the probability for $P_i < P_{i'}$ and in (b) the probability for $P_i = P_{i'}$.

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.

3.4.2 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-

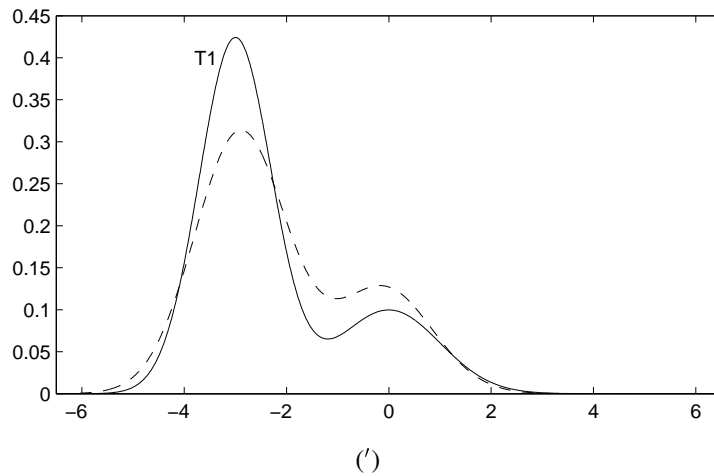
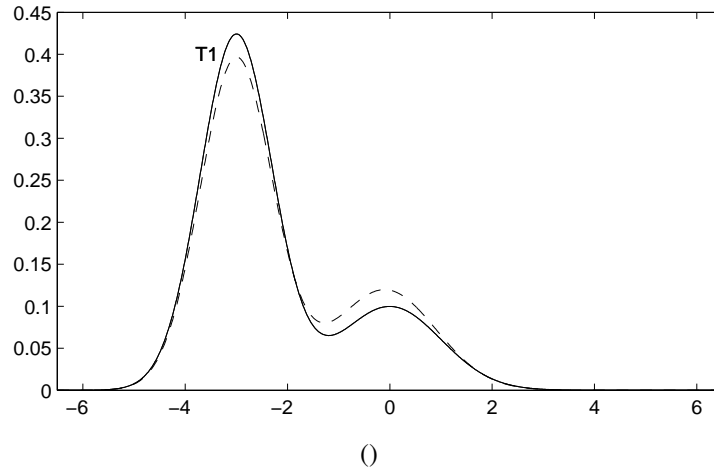


FIGURE 3.4: True (solid lines) and estimated (dashed lines) densities from SO-LCM and DPM for distribution T_1

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 fifteen procedures drawn uniformly from $(-1.5, -1)$, the second fifteen procedures drawn uniformly from $(-0.6, -0.4)$, the third fifteen procedures drawn uniformly from $(0.4, 0.6)$ and the last fifteen 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 (3.4) 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 fifteen 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 (3.4) and we additionally choose a $N(0, 10 I)$ 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 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 4.1 such that the first thirty procedures are ranked consistently with the true order and are clustered correctly.

3.5 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 proce-

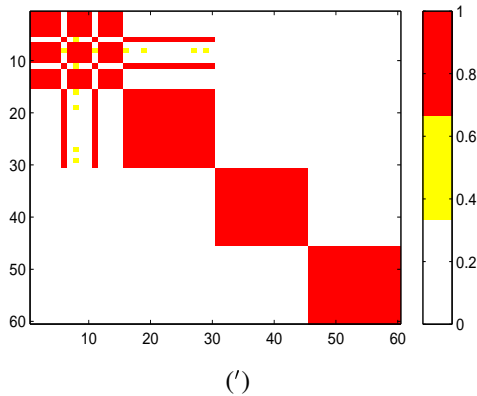
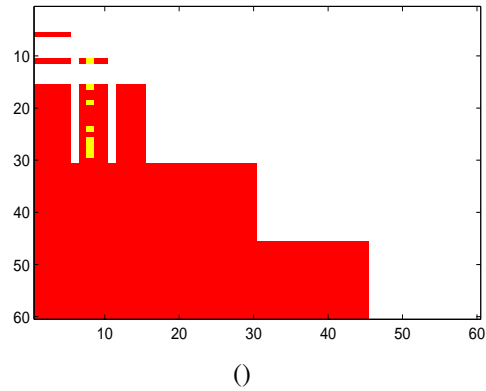


FIGURE 3.5: Posterior probability for ranking and clustering in study of section 4.2

dures 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 6 categories based on their estimated risk of in-hospital mortality. Similarly, the Aristotle method groups over 160 types of procedures into 4 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 post-procedural 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. Post-operative morbidity was regarded as a patient-level unobserved latent variable. Indicators of morbidity included a single continuous variable, post-operative length of stay (PLOS), modeled as $y_1 = \log(1 + \text{PLOS})$; and 6 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 7 endpoints is:

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

Let f_i denote the density of η_{ij} among patients undergoing the i th type of procedure, i.e. $\eta_{ij} \sim f_i$. 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 $\psi_k \sim \text{Gamma}(\alpha_1/K, 1)$, $\alpha_1 \sim \text{Gamma}(1, 6/\ln 145)$, and $\beta_k \stackrel{\text{ind}}{\sim} \text{N}(0, 10)$, $\alpha_0 \sim \text{Gamma}(1, 1)$ and $\alpha_2 \sim \text{Gamma}(1, 1)$. The prior for α_1 is chosen based on expert elicitation to favor 6 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 $\mu_t \stackrel{\text{ind}}{\sim} \text{N}(0, 10)$ and $\lambda_t \stackrel{\text{ind}}{\sim} \text{N}^+(0, \tau)$, $t = 1, 2, \dots, p$, 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 each figure, procedures are sorted in ascending order based on the posterior mean $E[S_i]$. In Figure 3.6, the posterior means of each procedure ($1/n_i \sum_{j=1}^{n_i} \theta_{ij}$) are plotted on the vertical axis along with 95% credible intervals. The relatively wide

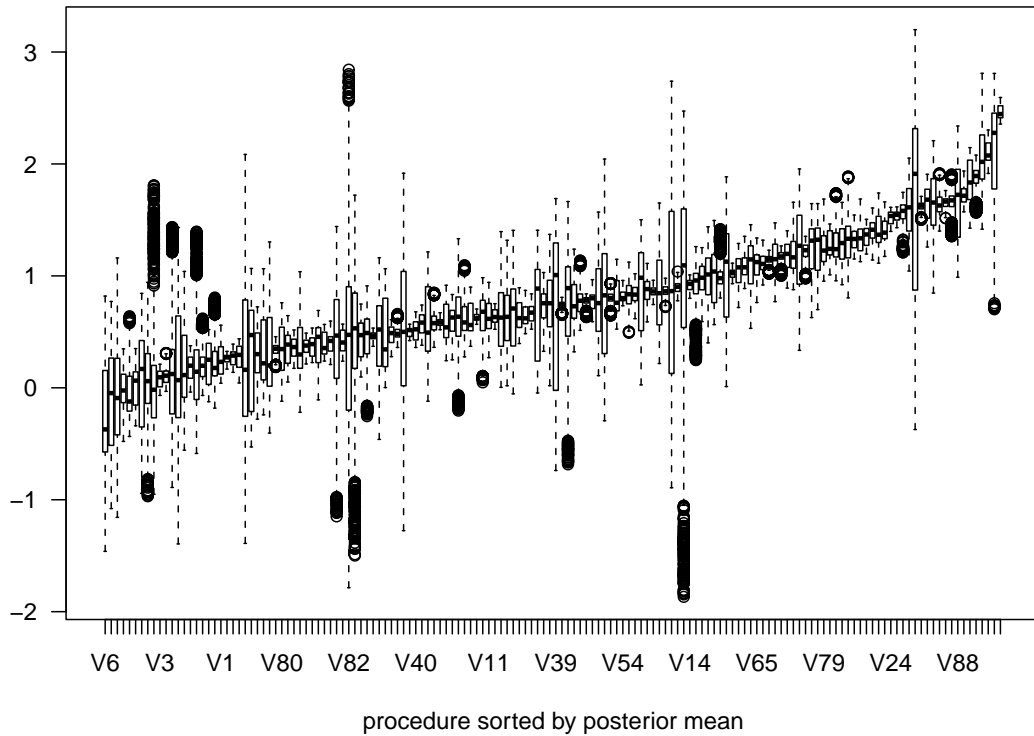


FIGURE 3.6: Sorted Procedures Posterior Means for latent scores of each procedure with 95% Credible Intervals

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 3.7(a) and 3.7(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 8 clusters, with 32% probability assigned to the posterior mode of 5.

We also propose a way to obtain an optimal point estimate of the ranked clustering as

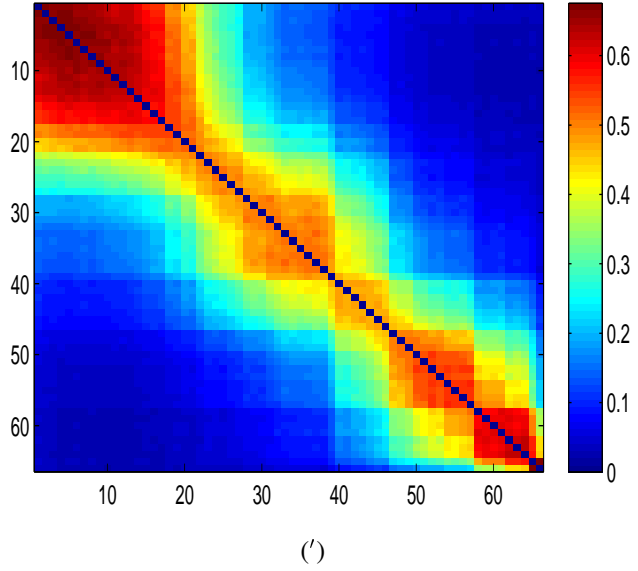
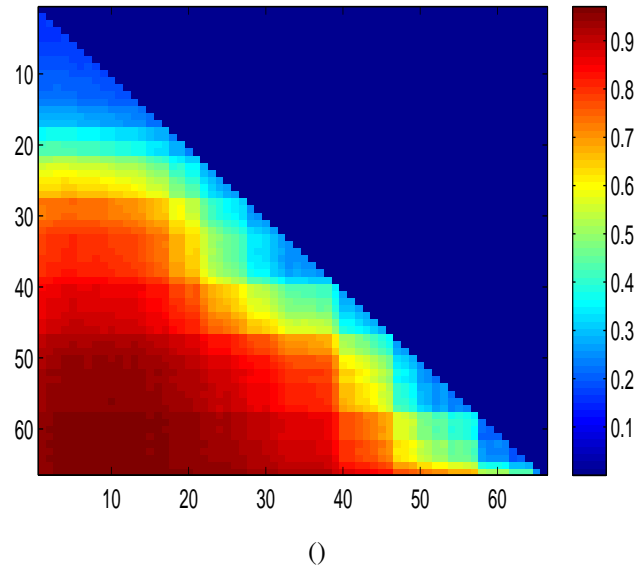


FIGURE 3.7: Ranking and Clustering for selected 66 procedures with more than 200 patients, with entry (i, j) in (a) being the lower triangular matrix identifying the probability for $P_i < P_{i'}$ and in (b) the probability for $P_i = P_{i'}$.

follows. For a vector-valued parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_I)$, define the corresponding ranks $\mathbf{R} = (R_1, \dots, R_I)$ in equation (3.11), and let $\tilde{\mathbf{R}} = (\tilde{R}_1, \dots, \tilde{R}_I)$ denote a possible point estimate of \mathbf{R} .

$$R_i(\boldsymbol{\theta}) = \text{rank}(\theta_i) = \sum_{j=1}^I 1(\theta_i \geq \theta_j), \quad (3.11)$$

with the smallest θ having rank 1 and the largest having rank I . Denote that the true rank for $\boldsymbol{\theta}$ is \mathbf{R} , the estimated rank is $\tilde{\mathbf{R}}$, R_i is the rank variable for object i and \tilde{R}_i is the estimated rank (we drop the dependency on $\boldsymbol{\theta}$ for notational convenience). To find the optimal ranked clustering, we define the following loss function $L(\mathbf{R}, \tilde{\mathbf{R}})$ as

$$\begin{aligned} L(\mathbf{R}, \tilde{\mathbf{R}}) &= \sum_{(i,j) \in \mathcal{M}} \{2 \times 1(\tilde{R}_i < \tilde{R}_j, R_i > R_j) + 2 \times 1(\tilde{R}_i > \tilde{R}_j, R_i < R_j) \\ &\quad + 1(\tilde{R}_i = \tilde{R}_j, R_i \neq R_j) + 1(\tilde{R}_i \neq \tilde{R}_j, R_i = R_j)\}. \end{aligned} \quad (3.12)$$

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

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

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 smallest risk as the estimate.

To compare the performance of our method with that of the Aristotle level (with 4 levels) based on the Bayes risk, we let $K = 4$ so that we will only obtain 4 clusters. The

Bayes risks for the ranked clustering obtained from the Aristotle score is 4900.5. Our optimal Bayesian ranked clustering achieves 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 were 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 were assigned to the lowest risk category according to our method. The correlation between these two ranked clusterings is 0.44. The correlation between the ranked clustering of the SO-LCM and the $\log(1 + \text{PLOS})$ is 0.81 and the correlation between the ranked clustering of the Aristotle level and the $\log(1 + \text{PLOS})$ is 0.58.

3.6 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 (3.3) independently for each ele-

Table 3.1:

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 3.2:

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

ment 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.

Table 3.3:

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

Bayesian tensor co-clustering for flexible multilevel regression modeling

4.1 Introduction

Many applications collect matrix data having large numbers of rows and columns. For example, the rows may correspond to movie viewers and the columns to movies, with the elements of the matrix being movie rankings. In the well known Netflix problem, the focus is on filling in the missing elements of the enormous matrix based on data for a sparse number of cells. In other cases, such as our motivating application, the elements of the matrix are not observed directly but correspond to vectors of parameters in a model. In such settings, data are sparse and hence some strong dimensionality reduction or borrowing of information across the cells is needed. In our application, we have the additional challenge of having more than two factors, and hence need to borrow information across cells in a multiway array (tensor).

In the matrix case, a simple and popular approach for dimensionality reduction is co-clustering, which refers to simultaneous grouping of rows and columns. Each co-cluster is a submatrix of the full data matrix, and co-clustering algorithms are characterized by

how the rows and columns are assigned in clusters. As described by Meeds and Roweis (2007) in Figure 4.1, cells can either belong to multiple clusters (as shown in A and B) or to a single cluster (as shown in C), with clustering overlapping (as in A) or not (as in B and C). For robustness and to avoid inducing a very large number of clusters, certain cells are not assigned to any cluster and are instead allocated to a background noise component. Relative to black box methods for dimensionality reduction of large matrices or tensors, such as singular value decompositions, co-clustering approaches are appealing due to their transparent interpretability. For example, grouping movie viewers and movies into clusters is intuitive and the clustering produced may be of interest in itself.

Shafiei and Milios (2006) consider co-clustering models for simultaneously clustering documents and terms. Their latent Dirichlet co-clustering model depicts each document as a random mixture of document topics, where each topic is a distribution over segments of the document. Each of these segments in the document can be modeled as a mixture of word topics where each topic is a distribution over words. They use the Dirichlet distribution to model the mixing proportion of document-topics and word-topics in a document. They did not consider interactions and available features which are informative about cluster allocation. In their modeling, they also assume that the number of word and document topics are known and fixed, which is too restrictive.

Agarwal and Merugu (2007) propose a regression model based on co-clustering. Let $\mathbf{Y} = (y_{ij}) \in \mathcal{R}^{m \times n}$ denote the response matrix and let $\mathbf{X} = (\mathbf{x}_{ij}) \in \mathcal{R}^{m \times n \times s}$ denote the tensor corresponding to s prespecified covariates with $\mathbf{x}_{ij} \in \mathcal{R}^s$. Given $k \times l$ blocks (I, J) with prior probabilities π_{IJ} , the marginal distribution of response given covariates is

$$p(y_{ij} | \mathbf{x}_{ij}) = \sum_{I,J} \pi_{IJ} f_{\psi}(y_{ij}; \boldsymbol{\beta}' \mathbf{x}_{ij} + \delta_{I,J}), \quad i = 1, \dots, m, j = 1, \dots, n,$$

where f_{ψ} is an exponential family distribution with cumulant $\psi(\cdot)$, $\boldsymbol{\beta} \in \mathcal{R}^s$ denotes the regression coefficients associated with the prespecified covariates, π_{IJ} denotes the prior probability of class (I, J) and $\delta_{I,J}$ denotes the interaction effects associated with this class.

They showed that accounting for interaction $\delta_{I,J}$ directly in the predictive model along with information in the covariates often leads to better predictions. The number of clusters in their model is also prefixed.

In our motivating application to data on different outcomes following congenital heart surgery, it is appealing to cluster hospitals and procedures to achieve dimensionality reduction, while producing clusters that are informative and interpretable to the physicians. However, there are some key disadvantages to current co-clustering approaches that are important to address. The first is that the number of co-clusters can be large. If we cluster hospitals separately from procedures, the number of cells in the resulting hospital \times procedure co-cluster matrix will be large enough that we would still need to borrow information strongly across co-clusters. This problem becomes even more of an issue in generalizing beyond matrix co-clustering to tensor co-clustering by also including outcome type. An additional issue is that it is not realistic to treat rows as exchangeable or columns as exchangeable, as features are often available that are informative about cluster allocation. For example, we may have information on the type of hospital or procedure.

To address these issues, this article proposes an ANOVA co-clustering model with interactions. Each cell of the tensor is assigned to a cluster corresponding to each dimension. Including a main effect for each dimension in an additive model for the cell-specific parameters, we then cluster the main effects. This greatly reduces the dimensionality relative to the usual co-clustering approach through the use of an additive model. To avoid ignoring interactions, we also allocate cells to interaction clusters and add an interaction term. By including zero clusters in the main effect and interaction components, we allow collapsing on simplified models when appropriate. To allow features to inform the allocation to clusters and avoid exchangeability assumptions, we rely on the probit stick-breaking process (PSBP) of Chung and Dunson (2009). Using a full probability model based on a Bayesian approach, a simple MCMC algorithm can be implemented for posterior computation. As clustering is soft and probabilistic, we obtain unique parameter

estimates for each cell. Even if clustering is not of interest in itself, the proposed framework provides a useful strategy for dimensionality reduction and borrowing information for high-dimensional categorical covariates with interactions.

Section 2 proposes the basic structure of the Bayesian matrix co-clustering with interactions with considerations of properties. Section 3 outlines an efficient Gibbs sampler algorithm for posterior computation. Simulation studies are conducted in section 4 and we describe the application to the heart surgery outcomes data in section 5. Section 6 contains a discussion.

4.2 Model Specification and Properties

Focusing on the heart surgery application for concreteness, let y_{hpi} be a binary indicator of an outcome for the i th ($i = 1, \dots, n_{hp}$) patient receiving procedure p ($p = 1, \dots, P$) in hospital h ($h = 1, \dots, H$). We consider the following generalized linear model

$$\Pr(y_{hpi} = 1 \mid \mathbf{x}_{hpi}) = g(\mathbf{x}'_{hpi}\boldsymbol{\beta}_{hp}), \quad (4.1)$$

where $\mathbf{x}_{hpi} = (x_{hpi1}, \dots, x_{hpiq})'$ is a vector of patient level predictors that may depend on procedure type, $\boldsymbol{\beta}_{hp} = (\beta_{hp1}, \dots, \beta_{hpq})'$ are regression coefficients specific to procedure p and hospital h , and $g(\cdot)$ is a monotone link function mapping from $\mathfrak{R} \rightarrow [0, 1]$. For example, we will focus on probit and logistic link functions. For many values of (h, p) , there will be few patients and hence it is important to borrow information across hospitals and related procedures in estimating these regression coefficient vectors. One natural approach for borrowing information is to use a hierarchical model that expresses $\boldsymbol{\beta}_{hp}$ as a sum of main effects for hospital and procedure while allowing interactions,

$$\boldsymbol{\beta}_{hp} = \boldsymbol{\beta}_0 + \boldsymbol{\rho}_h + \boldsymbol{\kappa}_p + \boldsymbol{\psi}_{hp}, \quad (4.2)$$

with $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0q})'$ an overall mean vector, $\boldsymbol{\rho}_h = (\rho_{h1}, \dots, \rho_{hq})'$ main effects for hospital h , $\boldsymbol{\kappa}_p = (\kappa_{p1}, \dots, \kappa_{pq})'$ main effects for procedure p , and $\boldsymbol{\psi}_{hp} = (\psi_{hp1}, \dots, \psi_{hpq})'$

interactions between hospital h and procedure p . Such interactions are important, as certain hospitals may have good patient outcomes for certain procedures but not others.

To complete a specification of the hierarchical model, it is necessary to choose random effects distributions for each of the factors in (4.2),

$$\boldsymbol{\rho}_h \sim P_{\boldsymbol{\rho}}, \quad \boldsymbol{\kappa}_p \sim P_{\boldsymbol{\kappa}}, \quad \psi_{hp} \sim P_{\psi}, \quad (4.3)$$

with Gaussian distributions centered at zero with unknown covariance providing a standard choice for $P_{\boldsymbol{\rho}}, P_{\boldsymbol{\kappa}}, P_{\psi}$. However, there are some clear disadvantages of Gaussians in this context. Firstly, the results may be sensitive to the assumed shape of the random effects distributions, with the Gaussian distribution not allowing groups of hospitals or procedures having similar performance or outliers. Secondly, in applications involving many hospitals and procedures, it is appealing to consider an approach that leads to dimensionality reduction while also leading to insight into whether two hospitals or procedures are similar. Finally, there is typically interest in variable selection in which certain predictors may have zero coefficients and may have constant coefficients across hospitals or procedures.

To address these interests, we propose to follow a Bayesian approach and to specify a multivariate normal distribution for β_0 while using independent zero-inflated Dirichlet process priors for $P_{\boldsymbol{\rho}}, P_{\boldsymbol{\kappa}}$ and P_{ψ} . These priors are related to those described by MacLachlan et al. (2007) and Dunson et al. (2008a),

$$\boldsymbol{\rho}_h \sim P_{\boldsymbol{\rho}} = \sum_{l=1}^{\infty} \pi_{1l} \delta_{\boldsymbol{\rho}_l^*}, \quad (4.4)$$

where $\pi_{1l} = \pi_{1l}^* \prod_{s < l} (1 - \pi_{1s}^*)$, $\pi_{1l}^* \sim \text{beta}(1, \alpha_{\rho})$ and $\boldsymbol{\rho}_l^*$ is a $q \times 1$ vector of coefficients with elements drawn *a priori* from a distribution consisting of a point mass at zero with probability $\tilde{\pi}_{\rho}$ given a beta hyperprior $\text{beta}(a_{\rho}, b_{\rho})$ and a continuous density $G_{0\rho}$:

$$\boldsymbol{\rho}_l^* \sim G_{\rho}, \quad G_{\rho} = \tilde{\pi}_{\rho} \delta_0 + (1 - \tilde{\pi}_{\rho}) G_{0\rho}, \quad (4.5)$$

where the elements of $G_{0\rho}$ follow the Laplace distribution and can be expressed as a scale mixture of normals (with an exponential mixing density),

$$\begin{aligned} G_{0\rho} &\equiv \text{MN}(\boldsymbol{\rho}_l^*; \boldsymbol{\mu}_\rho, \Gamma_\rho), \quad \mu_{\rho,s} \sim \text{N}(0, \gamma_{\rho,s}), \\ \gamma_{\rho,s} &\stackrel{\text{ind}}{\sim} \text{Exp}(\gamma_{\rho,s}; 2/\tau_\rho), \quad \tau_\rho \sim \text{Gamma}(\tau_\rho; r_\rho, \delta_\rho), \end{aligned} \quad (4.6)$$

where $\Gamma_\rho = \text{diag}(\gamma_{\rho,1}, \dots, \gamma_{\rho,q})$. $P_{\boldsymbol{\kappa}}$ and $P_{\boldsymbol{\psi}}$ are defined similarly to $P_{\boldsymbol{\rho}}$:

$$\begin{aligned} P_{\boldsymbol{\kappa}} &= \sum_{j=1}^{\infty} \pi_{2j} \delta_{\boldsymbol{\kappa}_j^*}, \quad \pi_{2j} = \pi_{2j}^* \prod_{s<j} (1 - \pi_{2s}^*), \quad \pi_{2j}^* \sim \text{beta}(1, \alpha_\kappa), \\ P_{\boldsymbol{\psi}} &= \sum_{k=1}^{\infty} \pi_{3k} \delta_{\boldsymbol{\psi}_k^*}, \quad \pi_{3k} = \pi_{3k}^* \prod_{s<k} (1 - \pi_{3s}^*), \quad \pi_{3k}^* \sim \text{beta}(1, \alpha_\psi), \end{aligned}$$

where $\boldsymbol{\kappa}_j^*$, $\boldsymbol{\psi}_k^*$ are defined similarly to $\boldsymbol{\rho}_l^*$ with a mixture of a degenerate distribution at zero and a non-atomic multivariate Laplace distribution. We refer to the model in (4.2) - (4.6) as the matrix co-clustering with interactions prior (abbreviated as MCCI).

Following Dunson et al. (2008a), we can reexpress (4.4) as follows:

$$\begin{aligned} P_{\boldsymbol{\rho}} &= \tilde{\pi}_\rho \delta_0 + (1 - \tilde{\pi}_\rho) \sum_{l=1}^{\infty} \pi_{1l} \delta_{\tilde{\boldsymbol{\rho}}_l} \\ &= \tilde{\pi}_\rho \delta_0 + (1 - \tilde{\pi}_\rho) G_{\rho}^*, \end{aligned} \quad (4.7)$$

where $\tilde{\boldsymbol{\rho}}_l \stackrel{\text{iid}}{\sim} G_{0\rho}^*$, with $G_{0\rho}^*$ denoting the non-atomic Laplace prior, and $G_{\rho}^* \sim DP(\alpha_\rho G_{0\rho}^*)$. Hence the random distribution $P_{\boldsymbol{\rho}}$ can be formulated as a mixture of a degenerate distribution with all its mass at zero and a DP with non-atomic base measure. From model (4.7), we can show that Conditional on $\tilde{\boldsymbol{\rho}}_l$, the prior distribution for the number of procedures having zero coefficients is binomial.

Coefficients from the same hospital belong to the same cluster if $\boldsymbol{\kappa}_p = \boldsymbol{\kappa}_{p'}$ and $\boldsymbol{\psi}_{hp} = \boldsymbol{\psi}_{hp'}$, whereas coefficients from different hospital belong the same cluster if $\boldsymbol{\rho}_h = \boldsymbol{\rho}_{h'}$,

$\kappa_p = \kappa_{p'}$ and $\psi_{hp} = \psi_{h'p'}$. In particular, given that the coefficients are not null, the probabilities of belonging to the same cluster is

$$\begin{aligned}
\Pr(\beta_{hp} = \beta_{h'p'} | \beta_{hp} \neq \beta_0, \beta_{h'p'} \neq \beta_0) &= \tilde{\pi}_\kappa^2 \tilde{\pi}_\psi^2 (1 - \tilde{\pi}_\rho) + \tilde{\pi}_\kappa^2 \frac{(1 - \tilde{\pi}_\psi)^2}{1 + \alpha_\psi} + \tilde{\pi}_\psi^2 \\
&\quad \times \frac{(1 - \tilde{\pi}_\kappa)^2}{1 + \alpha_\kappa} + \frac{(1 - \tilde{\pi}_\kappa)^2 (1 - \tilde{\pi}_\psi)^2}{(1 + \alpha_\kappa)(1 + \alpha_\psi)}, \\
\Pr(\beta_{hp} = \beta_{h'p'} | \beta_{hp} \neq \beta_0, \beta_{h'p'} \neq \beta_0) &= \tilde{\pi}_\rho^2 \frac{(1 - \tilde{\pi}_\kappa)^2 (1 - \tilde{\pi}_\psi)^2}{1 + \alpha_\kappa} + \tilde{\pi}_\kappa^2 \frac{(1 - \tilde{\pi}_\rho)^2}{1 + \alpha_\rho} \\
&\quad \times \frac{(1 - \tilde{\pi}_\psi)^2}{1 + \alpha_\psi} + \tilde{\pi}_\psi^2 \frac{(1 - \tilde{\pi}_\rho)^2 (1 - \tilde{\pi}_\kappa)^2}{1 + \alpha_\rho} + \tilde{\pi}_\rho^2 \tilde{\pi}_\kappa^2 \frac{(1 - \tilde{\pi}_\psi)^2}{1 + \alpha_\psi} + \tilde{\pi}_\rho^2 \tilde{\pi}_\psi^2 \frac{(1 - \tilde{\pi}_\kappa)^2}{1 + \alpha_\kappa} \\
&\quad + \tilde{\pi}_\psi^2 \tilde{\pi}_\kappa^2 \frac{(1 - \tilde{\pi}_\rho)^2}{1 + \alpha_\rho} + \frac{(1 - \tilde{\pi}_\rho)^2 (1 - \tilde{\pi}_\kappa)^2 (1 - \tilde{\pi}_\psi)^2}{1 + \alpha_\rho} \frac{(1 - \tilde{\pi}_\kappa)^2}{1 + \alpha_\kappa} \frac{(1 - \tilde{\pi}_\psi)^2}{1 + \alpha_\psi}.
\end{aligned} \tag{4.8}$$

Similar results compared to model (4.8) will hold for coefficients from same procedures but different hospitals. After some algebra, we can show that coefficients from the same hospital (or procedure) are clustered together with higher probability than from different hospitals (or procedures). Denote $\mathbf{a} = (a_\rho, a_\kappa, a_\psi)$ and $\mathbf{b} = (b_\rho, b_\kappa, b_\psi)$ and integrate out $\tilde{\pi}_\rho, \tilde{\pi}_\kappa$ and $\tilde{\pi}_\psi$, the probability that β_{hp} is null is

$$\Pr(\beta_{hp} = \beta_0 | \mathbf{a}, \mathbf{b}) = \frac{a_\rho}{a_\rho + b_\rho} \frac{a_\kappa}{a_\kappa + b_\kappa} \frac{a_\psi}{a_\psi + b_\psi}.$$

Similarly, the probability that only the main effect of hospital is null is

$$\Pr(\rho_h = \mathbf{0}, \kappa_p \neq \mathbf{0}, \psi_{hp} \neq \mathbf{0} | \mathbf{a}, \mathbf{b}) = \frac{a_\rho}{a_\rho + b_\rho} \frac{b_\kappa}{a_\kappa + b_\kappa} \frac{b_\psi}{a_\psi + b_\psi}.$$

Similar results for main effect of procedure and the interactions can be obtained.

The prior assigns high probability in sparse locations of this parameter space favoring strong shrinkage towards zero for most of the elements. We give an intuition for the

motivation through plots (in Figure 4.2) showing multiple realizations for β_{hp} with $q = 1$, $\tilde{\pi}_\rho = \tilde{\pi}_\kappa = \tilde{\pi}_\psi = 1/2$.

To allow predictors to provide information about the cluster allocation, we can employ PSBP instead of Dirichlet process in model (4.4), where the probability of allocation to cluster h would depend on hospital-specific predictors u_h . More specifically,

$$\rho_h \sim P\rho = \sum_{l=1}^{\infty} \pi_l(u_h) \delta_{\rho_l^*}. \quad (4.9)$$

P_κ are defined similarly with coefficients v_p , while P_ψ remain the same as defined in model (4.4). We name this prior as the predictor dependent matrix co-clustering with interactions prior (abbreviated as PMCCI). The predictor-dependent weights are constructed as $\pi_l(u_h) = V_l(u_h) \prod_{k < l} \{1 - V_k(u_h)\}$, where $0 < V_l(u_h) < 1$ constitute predictor-dependent probabilities. We set $V_l(u)$ predictor-dependent probit functions:

$$V_l(u) = \int_{-\infty}^{g_l(u)} \mathbf{N}(x; 0, 1) dx, \quad g_l(u) = \zeta_{l0} + \sum_{h=1}^H \zeta_{lh} \mathcal{K}(u, u_h; \varrho_l),$$

where $\mathcal{K}(u, u_h; \varrho_l)$ is a kernel characterized by parameter ϱ_l with $\varrho_l \sim \text{Gamma}(a_\varrho, b_\varrho)$ and $\{\zeta_{lh}\}_{h=0,H}$ are a sparse set of real numbers. To impose sparseness on $\{\zeta_{lh}\}_{h=0,H}$, we choose priors $\zeta_{lh} \sim \mathbf{N}(0, \alpha_{lh}^{-1})$ and $\alpha_{lh} \sim \text{Gamma}(a_0, b_0)$, with (a_0, b_0) setting most α_{lh} large to let most $\{\zeta_{lh}\}_{h=0,H}$ near zero. A radial basis function (RBF) can be used for kernel \mathcal{K} , e.g., $\mathcal{K}(u, u_h; \varrho_l) = \exp\{-\|u_h - u\|_2 / \varrho_l\}$.

Let $\Theta_{\mathbf{B}}$ denote the collection of tensors with finite components:

$$\Theta_{\mathbf{B}} = \{\mathbf{B} = (\beta_{hpl}), h = 1, \dots, H, p = 1, \dots, P, \max_{1 \leq l \leq q} |\beta_{hpl}| < \infty\}.$$

Theorem 4.2.1. *Define $\|\mathbf{B} - \mathbf{B}^0\|_\infty = \max_{h,p,l} |\beta_{hpl} - \beta_{hpl}^0|$. Let $\mathcal{N}_{(\infty, \epsilon)}(\mathbf{B}^0) = \{\mathbf{B} : \|\mathbf{B} - \mathbf{B}^0\|_\infty < \epsilon\}$ denote an L_∞ neighborhood around an arbitrary $\mathbf{B}^0 \in \Theta_{\mathbf{B}}$. Letting $\mathbf{B} \sim \Pi$ denote the prior for \mathbf{B} in model (4.2) - (4.6), for any $\epsilon > 0$, and $\mathbf{B}^0 \in \Theta_{\mathbf{B}}$, the probability $\Pi\{\mathcal{N}_{(\infty, \epsilon)}(\mathbf{B}^0)\} > 0$.*

Proof: Details are in the Appendix.

Theorem 2.1 shows that our proposed MCCI prior has large support, which places positive probability in arbitrarily small neighborhoods around any $\mathbf{B}^0 \in \Theta_{\mathbf{B}}$. Large support for PMCCI prior can be similarly obtained.

Theorem 4.2.2. *Let $G^{\mathbf{T}}$ be a random distribution drawn from the PMCCI prior with $\mathbf{T} = (T_1, T_2, T_3)$ components, baseline measure G_{ρ} , G_{κ} and G_{ψ} as described in (4.5). Let G^{∞} denote the case $\mathbf{T} = (\infty, \infty, \infty)$. For $\mathbf{y} = (y_{hpi})$, with $h = 1, \dots, H$, $p = 1, \dots, P$ and $i = 1, \dots, n_{hp}$, let*

$$p^{\mathbf{T}}(\mathbf{y}) = \int \left\{ \prod_{h,p} \int \prod_i k(y_{hpi} | \beta_{hp}) P(d\beta_{hp}) \right\} G^{\mathbf{T}}(dP).$$

$p^{\infty}(\mathbf{y})$ is defined similarly. Then

$$\begin{aligned} \|p^{\mathbf{T}}(\mathbf{y}) - p^{\infty}(\mathbf{y})\| &\leq 4 \left[1 - \left\{ 1 - \left(\frac{1}{2}\right)^{T_1} \right\}^{\sum_h n_{hp}} \left\{ 1 - \left(\frac{1}{2}\right)^{T_2} \right\}^{\sum_p n_{hp}} \left\{ 1 - \left(\sum_{h,p} n_{hp}\right) \right. \right. \\ &\quad \left. \left. \times \exp\left(-\frac{T_3 - 1}{\alpha_{\psi}}\right) \right\} \right]. \end{aligned}$$

Proof: More Details are in the Appendix.

If $G^{\mathbf{T}}$ is drawn from the MCCI prior, similar to the above steps, we have

$$\begin{aligned} \|p^{\mathbf{T}}(\mathbf{y}) - p^{\infty}(\mathbf{y})\| &\leq 4 \left[1 - \left\{ 1 - \left(\sum_h n_{hp}\right) \exp\left(-\frac{T_1 - 1}{\alpha_{\rho}}\right) \right\} \left\{ 1 - \left(\sum_p n_{hp}\right) \exp\left(-\frac{T_2 - 1}{\alpha_{\kappa}}\right) \right\} \right. \\ &\quad \left. \left\{ 1 - \left(\sum_{h,p} n_{hp}\right) \exp\left(-\frac{T_3 - 1}{\alpha_{\psi}}\right) \right\} \right]. \end{aligned}$$

Theorem 2.2 is especially important for computational purposes. It ensures that samples obtained from the posterior distribution of the truncated process can be used to generate arbitrarily accurate inferences on measurable functionals of the infinite process.

4.3 Posterior Computation

For the posterior computation, we propose a data augmentation Gibbs sampling algorithm. We consider both the MCCI and the PMCCI priors. With the prior MCCI, we may follow the Pólya urn scheme proposed by Dunson et al. (2008a). With the prior PMCCI, updating schemes are generalized from the PSBP by Rodriguez and Dunson (2009). First, let $y_{hpi} = 1(y_{hpi}^* > 0)$, where $y_{hpi}^* = \mathbf{x}'_{hpi}\boldsymbol{\beta}_{hp} + \epsilon_{hpi}/\phi_{hpi}^{1/2}$, with $\phi_{hpi} \sim \text{Ga}(\nu/2, \nu/2)$ and $\epsilon_{hpi} \sim \text{N}(0, \sigma^2)$. Following O'Brien and Dunson (2004), we may set $\sigma^2 = \pi(\nu - 2)/3\nu$ with $\nu = 7.2$ to obtain an almost exact approximation to the logistic density. The sampling steps are as follows,

1. Denote $\tilde{y}_{hpi}^{(0)} = y_{hpi}^* - \mathbf{x}'_{hpi}(\boldsymbol{\beta}_{hp} - \boldsymbol{\beta}_0)$ and sample $\boldsymbol{\beta}_0$ by

$$p(\boldsymbol{\beta}_0 | \dots) \sim \text{MN}\left(\left(\Sigma_0^{-1} + \sum_{hpi} \frac{1}{\sigma_{hpi}^2} \mathbf{x}_{hpi} \mathbf{x}'_{hpi}\right)^{-1} \sum_{hpi} \frac{\tilde{y}_{hpi}^{(0)}}{\sigma_{hpi}^2} \mathbf{x}_{hpi}, \left(\Sigma_0^{-1} + \sum_{hpi} \frac{1}{\sigma_{hpi}^2} \mathbf{x}_{hpi} \mathbf{x}'_{hpi}\right)^{-1}\right),$$

where $\sigma_{hpi}^2 = \sigma^2/\phi_{hpi}$ and the prior for $\boldsymbol{\beta}_0$ is $\text{MN}(\mathbf{0}, \Sigma_0)$.

2. i. Let the (h) superscript denote a quantity obtained excluding element h . With the prior MCCI, the conditional prior distribution of $\boldsymbol{\rho}_h$ given $\boldsymbol{\rho}^{(h)}$ is

$$\left(\frac{\alpha_\rho(1 - \tilde{\pi}_\rho)}{\alpha_\rho + H - m_{\rho_0^*} - 1}\right) \text{MN}(\boldsymbol{\mu}_\rho, \Gamma_\rho) + \tilde{\pi}_\rho \delta_0 + \sum_{l=2}^{k_{\rho^*}} \left(\frac{m_{\rho_l^*}(1 - \tilde{\pi}_\rho)}{\alpha_\rho + H - m_{\rho_0^*} - 1}\right) \delta_{\rho_l^*}, \quad (4.10)$$

where $m_{\rho_l^*}$ is the number of elements of $\boldsymbol{\rho}^{(h)}$ equal to $\boldsymbol{\rho}^{*(h)}$ and k_{ρ^*} is the unique number of values of $\boldsymbol{\rho}^*$. As shorthand, let $\mathbf{u}_\rho = (u_{\rho,0}, u_{\rho,1}, \dots, u_{\rho,k_{\rho^*}})$ denote the probability weights on the mixture components in expression (4.10). The full

conditional posterior distribution of $\boldsymbol{\rho}_h$ is

$$p(\boldsymbol{\rho}_h | \dots) \sim w_{\rho,h0} \text{MN}(E_{\rho,h}, V_{\rho,h}) + \sum_{l=1}^{k_{\rho^*}} w_{\rho,hl} \delta_{\rho_l^*}.$$

Denote $\tilde{y}_{hpi}^{(h)} = y_{hpi}^* - \mathbf{x}'_{hpi}(\boldsymbol{\beta}_{hp} - \boldsymbol{\rho}_h)$ and the conditional posterior mean and covariance matrix in the multivariate normal component are

$$V_{\rho,h} = (\Gamma_{\rho}^{-1} + \sum_{pi} 1/\sigma_{hpi}^2 \mathbf{x}_{hpi} \mathbf{x}'_{hpi})^{-1}, \quad E_{\rho,h} = V_{\rho,h}(\Gamma_{\rho}^{-1} \boldsymbol{\mu}_{\rho} + \sum_{pi} \tilde{y}_{hpi}^{(h)}/\sigma_{hpi}^2 \mathbf{x}_{hpi}).$$

and the updated mixture weights are defined as:

$$w_{\rho,h0} = c_u \frac{u_{\rho,0} \text{MN}(0; \boldsymbol{\mu}_{\rho}, \Gamma_{\rho}) \prod_{pi} \text{N}(\tilde{y}_{hpi}^{(h)}; 0, \sigma_{hpi}^2)}{\text{MN}(0; E_{\rho,h}, V_{\rho,h})},$$

$$w_{\rho,hl} = c_u u_{\rho,l} \prod_{pi} \text{N}(\tilde{y}_{hpi}^{(h)}; \mathbf{x}'_{hpi} \rho_l^*, \sigma_{hpi}^2).$$

where c_u is the normalizing constant. Let $\xi_h = l$ if $\rho_h = \rho_l^*$ for $l = 1, \dots, k_{\rho^*}$, the full conditional posterior for ξ_h is

$$p(\xi_h | \dots) \sim \text{Multinomial}(0, 1, \dots, k_{\rho^*}; w_{\rho,h0}, w_{\rho,h1}, \dots, w_{\rho,hk_{\rho^*}}).$$

In step (1), we sample from these multinomial distributions. When $\xi_h = 0$, a new value for $\boldsymbol{\rho}_h$ is drawn from $\text{MN}(E_{\rho,h}, V_{\rho,h})$. In step (2), the unique values ρ_l^* are updated by

$$p(\rho_l^* | \dots) \sim \text{MN}(E_{\rho^*,h}, V_{\rho^*,h}),$$

$$\text{with } V_{\rho^*,h} = \left(\sum_{hpi} \frac{1}{\sigma_{hpi}^2} \mathbf{x}_{hpi} \mathbf{x}'_{hpi} 1(\xi_h = l) + \Gamma_{\rho}^{-1} \right)^{-1},$$

$$E_{\rho^*,h} = V_{\rho^*,h} \left(\sum_{hpi} \frac{1}{\sigma_{hpi}^2} \mathbf{x}_{hpi} \tilde{y}_{hpi}^{(h)} 1(\xi_h = l) + \Gamma_{\rho}^{-1} \boldsymbol{\mu}_{\rho} \right).$$

ii. For the PMCCI prior, the full conditional distribution for the indicators is multinomial with probability given by

$$\Pr(\xi_h = l | \dots) \propto u_{\rho,l} w_{\rho,hl} \prod_{p,i} \mathbf{N}(y_{hpi}^* | \boldsymbol{\rho}_l^*).$$

In order to sample the values of the latent process $g_l(u_h)$ and the corresponding weights $w_{\rho,hl}$, we introduce a collection of conditionally independent latent variables $z_{\rho,hl} \sim \mathbf{N}(g_l(u_h), 1)$. If we define $\xi_h = l$ if and only if $z_{\rho,hl} \geq 0$ and $z_{\rho,hr} < 0$ for $r < l$. We have

$$\begin{aligned} \Pr(\xi_h = l) &= \Pr(z_{\rho,hl} \geq 0, z_{\rho,hr} < 0 \text{ for } r < l) \\ &= \Phi(g_l(u_h)) \prod_{r < l} (1 - \Phi(g_r(u_h))) = w_{\rho,hl}. \end{aligned}$$

This data augmentation scheme simplifies computation as it allows us to implement another Gibbs sampling scheme. We can impute the augmented variables by sampling from its full conditional distribution conditional on the other values of the latent process and the indicator variables,

$$z_{\rho,hl} \propto \begin{cases} \mathbf{N}(g_l(u_h), 1)^-, & \xi_h > l \\ \mathbf{N}(g_l(u_h), 1)^+, & \xi_h = l. \end{cases}$$

$g_l(u_h)$ can be updated through the conjugate full conditional posterior distribution.

3. The full conditional posterior distribution for $\tilde{\pi}_\rho$ is

$$p(\tilde{\pi}_\rho | \dots) \sim \text{Beta}(a_\rho + \sum_h 1(\xi_h = 1), b_\rho + H - \sum_h 1(\xi_h = 1)).$$

4. ϕ_{hpi} is updated by

$$\phi_{hpi} \sim \text{Ga}\left(\frac{v+1}{2}, \frac{1}{2\sigma^2}(y_{hpi}^* - \mathbf{x}'_{hpi} \boldsymbol{\beta}_{hp})^2 + \frac{v}{2}\right).$$

4.4 Simulation Examples

Data are generated according to model (4.1), (4.2) and (4.3). We let $H = P = 4$, $q = 3$ and 20 patients at hospital h receive procedure p . More specifically, $\boldsymbol{\rho}$, $\boldsymbol{\kappa}$ and $\boldsymbol{\psi}$ are generated through multivariate mixture normal T_1 , T_2 and T_3 listed in Table 1 respectively. $\boldsymbol{\beta}_0$ is generated from the multivariate normal distribution with mean $(0, 0.1, 0.15)$ and covariance matrix $\text{diag}(0.5, 0.1, 0.1)$. Residuals are generated independently from $N(0, 0.5)$. For modeling with the PMCCI prior, we let the hospital-specific predictors as $-2, -1, 1, 2$ for the 4 hospitals respectively and procedure-specific predictors as $-4, -3, 3, 4$ for the 4 procedures respectively. Hospital and procedure coefficients are then generated from the mixture multivariate normal distribution with weights (w as in Table 1) calculated from (4.11) based on the predictors. The components of the multivariate normal distributions are the same as listed in Table 1. Hyperparameters are set as following. A larger r_ρ and/or a smaller δ_ρ lead more coefficients close to zero. The prior has fatter tails and larger variance as δ_ρ increases. We set r_ρ to introduce more shrinkage and let $\delta_\rho \sim \text{Gamma}(1, 1)$ to make the priors more flexible. Similarly, $r_\kappa = r_\psi = 1$ and $\delta_\kappa \sim \text{Gamma}(1, 1)$, $\delta_\psi \sim \text{Gamma}(1, 1)$. Following the common practice, we set $\alpha_\rho, \alpha_\kappa, \alpha_\psi \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(1, 1)$. We also let $\tilde{\pi}_\rho, \tilde{\pi}_\kappa$ and $\tilde{\pi}_\psi \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(a, b)$ to express prior ignorance. Results are based on 5,000 samples after a burn-in period of 2,500 iterations.

We run three data simulation examples. For the first simulation, we generate data from the full model (4.2). In the second simulation, data are generated without interaction $\boldsymbol{\psi}_{hp}$. In the last simulation, we generate data with only one factor $\boldsymbol{\rho}_h$ and the overall mean $\boldsymbol{\beta}_0$. Park and Hastie (2008) also proposed to use a variant of logistic regression with L_2 regularization (abbreviated as PLR) to fit the logistic modeling. We compare the performance of our proposed methods (with consideration of predictors interactions, PMCCI; with consideration of only interactions, MCCI; and without consideration of interactions, abbreviated as MCC) with their penalized logistic regression method. **R** package *stepPlr*

is applicable for the Park and Hastie (2008)'s method.

The current standard modeling for such problems is as following

$$\begin{aligned}
\text{logit Pr}(y_{hpi} = 1) &= \mu + \alpha_h + \beta_p + \mathbf{x}'_{hpi}\boldsymbol{\gamma}, \\
\mu &\sim \text{N}(0, \psi_0), & \psi_0 &\sim \text{IG}(a_0, b_0), \\
\alpha_h &\sim \text{N}(0, \psi_1), & \psi_1 &\sim \text{IG}(a_1, b_1), \\
\beta_p &\sim \text{N}(0, \psi_2), & \psi_2 &\sim \text{IG}(a_2, b_2), \\
\boldsymbol{\gamma} &\sim \text{MN}(\mathbf{0}, \boldsymbol{\Sigma}), & \boldsymbol{\Sigma}_{ii} &\stackrel{\text{i.i.d.}}{\sim} \text{IG}(a_3, b_3),
\end{aligned} \tag{4.12}$$

where y_{hpi} is the binary mortality response from patient i in hospital h receiving procedure p , \mathbf{x}_{hpi} is the patient related predictors, μ is the overall mean effect, α_h is the random effect for hospital h , β_p is the random effect for procedure p and $\boldsymbol{\gamma}$ is the regression coefficients for the patient related predictors. We choose hyperparameters $a_0 = b_0 = a_1 = b_1 = a_2 = b_2 = a_3 = b_3 = 1/2$ to obtain robust Cauchy priors for μ , α_h , β_p and γ_i .

Each simulation is repeated for 50 times. Comparisons are made based on the mean squared error (MSE) for the coefficients β_{hp} and the predicted mis-classification rates of the responses. Since β_{hp} can not be obtained from the standard modeling, we only compare the predicted mis-classification rates for the standard modeling with other methods. Performances from all the methods are summarized in Table 2-4. The numbers in parentheses are the corresponding standard errors (of the means) estimated by the bootstrap with $B = 500$ resamplings on the 50 mean-squared errors. Our proposed PMCCI, MCCI and MCC greatly outperform the PLR and the standard modeling in all simulations, especially for the MSE of the estimated coefficients for PLR and classification rates for the standard methods. PLR cannot do well when the sample size is small ($n = 20$), while borrowing information across hospitals and procedures is quite important and can greatly improve the performance of the PMCCI, MCCI and MCC. Interestingly, we notice that even generating data without interaction, the MCCI outperforms the MCC. The PMCCI performs

the best through the three simulation studies.

4.5 Application to Congenital Heart Surgery Outcomes Data

The growing call for accountability in medicine has led to increased demand for public reporting of health care providers' performance. Many profiling methods based on "risk-adjusted" outcomes have been developed, either within the framework of frequentist statistics (e.g., Landrum et al. (2000); Austin et al. (2003); Shahian et al. (2005); Timbie and Normand (2008); Landon et al. (2006)) or Bayesian inference (e.g., Austin et al. (2003); Lin et al. (2006, 2009); Racz and Sedransk (2010)). Extensive critiques (Shahian et al., 2001) and comparisons (Austin et al., 2001) have been conducted. Despite the publication of methodological recommendations and standards (Shahian and Normand, 2008), many aspects remain controversial.

Due to its substantial public health burden, social and financial costs, cardiovascular disease has figured prominently in public reporting efforts in recent years (Krumholz et al., 2006). At least 9 US states now publish hospital- or surgeon-specific risk-adjusted mortality outcomes for coronary artery bypass grafting (CABG) surgery. Reporting efforts for cardiac surgery have recently expanded beyond CABG to include other types of operations, such as aortic and mitral heart valve surgery. In addition to adult cardiac surgery, at least one state (New York State Department of Health, 2007) has begun reporting hospital-specific risk-adjusted mortality rates for surgical treatment of congenital heart defects.

When comparing mortality and other outcomes across providers, regression modeling is used to adjust for non-random allocation of patients to providers. Although such modeling is conceptually straightforward, developing appropriate models for congenital heart surgery patients is challenging, in part because the patient population is remarkably heterogeneous. Unlike cardiac surgery for acquired heart disease, there are literally hundreds of different types of congenital heart surgery operations, and they are performed

on patients ranging from neonates to the elderly. In addition, there are a large number of suspected and established risk factors which may confound the observed differences in outcomes when comparing centers. In theory, one could account for all of these factors in a regression model. In practice, however, we are limited by small numbers of cases in each age- and procedure- subgroup. Previous investigators have adjusted for the type of procedure by grouping procedures into a small number of categories, and modeling the categories using a series of category indicator variables. Although separate models for each procedure would be desirable from the standpoint of achieving correct model specification, the small number of cases available makes procedure-specific regression models highly challenging.

Therefore, it is of great significance for us to develop a new set of advanced profiling methods that can scale well in high-dimensional, multi-level, and sparse data. Our proposed co-clustering modeling strategy is highly flexible and adaptable and incorporates a sparseness-favoring structure, which combats the curse of dimensionality. In this section, we illustrate these properties by using our method to estimate and compare risk-adjusted outcomes for 16 hospitals performing 13 types of congenital heart surgery procedures.

Data for this analysis were obtained from the Society of Thoracic Surgeons (STS) Congenital Heart Surgery database. To create the study population, we first identified $N=79,635$ patients from 76 hospitals who underwent one of 145 types of congenital cardiovascular procedures at an STS-participating center during the years 2002-2008. We excluded hospitals with fewer than 1500 operation records and procedures with fewer than 1500 records, leaving a final population of 16,762 records for 13 types of procedures at 16 hospitals. As can be seen from Figure 4.3, the procedure-specific sample sizes are small and highly variable (median = 46; range 0 to 354).

The endpoint of interest for this analysis was post-operative length of stay (PLOS) in days, modeled as $y_{hpi} = \log(1 + \text{PLOS}_{hpi})$. Length of stay is an important determinant of resource usage (and hence cost) and is commonly used as an indirect indicator of

a patient’s overall health status following surgery. Patient-level predictors were age (in years) on the date of surgery; presence of any non-cardiac genetic abnormality (yes/no); and presence of any of the three risk factors (acidosis, shock or preop ventilatory support). The distribution of log-PLOS was modeled as $y_{hpi} \sim N(x'_{hpi}\beta_{hp}, \sigma_p^2)$, where σ_p^2 denotes the variance of log-PLOS for the p th procedure type. The variables y_{hpi} were assumed to be independent conditional on covariates x_{hpi} and parameters (β_{hp}, σ_p^2) .

Procedure-and-hospital-specific regression functions from model (4.1) and (4.2) were estimated nonparametrically under the MCCI prior in section 2. Flexible hyperpriors were set similarly to section 4. In each model, the prior for σ_p^2 was $\sigma_p^2 \sim \text{IG}(1/2, 1/2)$. Estimates were 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 were obtained using shorter chains. Analyses were compared with the standard model (4.12) to show the improved performance of our proposed modeling method.

We first compared the fit of the standard model (4.12) with our proposed model by using the log posterior marginal likelihood because it is insensitive to the choice of the prior distribution. Figure 4.4 shows the log posterior marginal likelihood across $16 \times 13 = 208$ specific combinations of hospitals and procedures for the standard method (right panel) and our proposed model (left panel). The improved fit of our model is apparent. We also compared the mean squared error (MSE) of y_{hpi} and \hat{y}_{hpi} , where the latter is the prediction value for y_{hpi} given x_{hpi} . The MSE for the standard method is 0.81, while our proposed MCCI has the MSE for 0.27.

The posterior means of β_{hp} are plotted in Figure 4.5. To quantify variation in regression coefficients β_{hp} across procedures and hospitals, let $\beta_{..} = (HP)^{-1} \sum_{h=1}^H \sum_{p=1}^P \beta_{hp}$ denote the overall mean, let $\beta_{h.} = P^{-1} \sum_{p=1}^P \beta_{hp}$ denote the mean for procedure h , and let $\beta_{.p} = H^{-1} \sum_{h=1}^H \beta_{hp}$ denote the mean for procedure p . Finally, let $\beta_{hp}^{(q)}$ denote the q th element of β_{hp} with analogous definitions for $\beta_{..}^{(q)}$, $\beta_{h.}^{(q)}$, and $\beta_{.p}^{(q)}$. The proportion of variation

in $\beta_{hp}^{(q)}$ that is explained by hospitals and procedures is, respectively,

$$v_{\text{hos}}^{(q)} = \frac{P \sum_h (\beta_h^{(q)} - \beta_{..}^{(q)})^2}{\sum_h \sum_p (\beta_{hp}^{(q)} - \beta_{..}^{(q)})^2}, \quad \text{and} \quad v_{\text{proc}}^{(q)} = \frac{H \sum_p (\beta_p^{(q)} - \beta_{..}^{(q)})^2}{\sum_h \sum_p (\beta_{hp}^{(q)} - \beta_{..}^{(q)})^2}.$$

Note that the proportion of variation in $\beta_{hp}^{(q)}$ that is explained by a least squares additive model of the form $\alpha_h + \beta_p$ is equal to $v_h^{(q)} + v_p^{(q)}$. Posterior means of $v_h^{(q)}$ and $v_p^{(q)}$ are summarized in Figure 4.6. As we might expect, regression coefficients appear to vary more substantially by procedure than by hospital.

The primary goal of our analysis was to compare the distribution of PLOS across providers in a manner that adjusts for non-random patient allocation. Conceptually, we would like to compare the average expected PLOS of the patients treated by a particular provider to the average PLOS that would be expected if the same patients had been randomly assigned to another provider. To fix ideas, let $Y = \text{PLOS}$ and let

$$g_h(x; p) = E[Y|X = x, \text{procedure} = p, \text{hospital} = h]$$

denote the true unknown regression relationship between patient-level covariates x and outcome Y for patients undergoing procedure p at hospital h . The regression function g_h encapsulates the h -th provider's 'quality' for performing procedure p on patients with covariates x . For a particular procedure p and covariate profile x , the performance of provider h relative to the other $H - 1$ providers may be summarized by the ratio

$$R_h(x; p) = \frac{g_h(x; p)}{\frac{1}{H-1} \sum_{j \neq h}^H g_j(x; p)}.$$

Although the performance of a provider is likely to differ as a function of x and p , it is convenient to have a global measure of performance which averages over x and p . Such a measure may be defined as

$$R_h = \frac{\int g_h(x; p) dF_h(x, p)}{\int \frac{1}{H-1} \sum_{j \neq h}^H g_j(x; p) dF_h(x, p)}.$$

where F_h is the actual observed distribution of covariates and procedures in the h -th provider's patient population. Comparisons of R_h based on the standard method and our proposed method are summarized in boxplots in Figure 4.7. As we can see, estimates from the MCCI model are more precise in the sense of having shorter confidence intervals. Though against our intuition, the much more flexible Bayesian nonparametric models typically produce narrower credible intervals than parametric hierarchical models such as Gaussian models. Kyung et al. (2009) prove that the variance on the fixed effects is always smaller in a linear random effects model with a DP prior being placed on the random effects distribution instead of a normal prior. Dunson et al. (2008b) also observe much decreases in credible interval width in more complex Bayesian nonparametric hierarchical models.

4.6 Discussion

In this paper, we have presented a general model of tensor co-clustering with interactions. Our proposed Bayesian multi-way tensor co-clustering (BMTCC) model allows borrowing information across the tensors. In particular, the model works by reducing the dimension of the tensor through separately clustering different dimensions. The BMTCC model inherits the strengths and robustness of Bayesian modeling, is designed to work with sparse tensors, and can use any exponential family distribution as the generative model, thereby making it suitable for a wide range of tensor related analysis.

Unlike existing co-clustering algorithms, BMTCC includes the interactions among different dimensions, which is quite important in practical analysis. We also allow features to inform the allocation to clusters and avoid exchangeability assumptions.

Finally, the main effects and interactions obtained from the BMTCC can be effectively used for visualization, subsequent predictive modeling and decision making.

Table 4.1:

Dist	Comp1			Comp2		
	w	μ	Σ	w	μ	Σ
T_1	0.5	(0,1,1)	diag(0.5,1,2)	0.5	(0,-2,-3)	diag(0.5,1,1)
T_2	0.5	(0,0,0.4)	diag(0.5,0.1, 0.2)	0.5	(0,1.5,2)	diag(0.5,0.5,1)
T_3	0.3	(0,1,1)	diag (0.5,1,1)	0.7	(0,-1,-1)	diag(0.5,0.5,1)

Table 4.2:

Method	MSE	Mis-Classification Rate
PMCCI	0.73 (0.08)	0.15 (0.02)
MCCI	0.84 (0.10)	0.16 (0.03)
MCC	1.06 (0.10)	0.16 (0.03)
PLR	14.67 (8.4)	0.18 (0.04)
SM	**	0.35 (0.12)

Table 4.3:

Method	MSE	Mis-Classification Rate
PMCCI	0.73 (0.10)	0.10 (0.01)
MCCI	0.84 (0.11)	0.10 (0.01)
MCC	1.06 (0.12)	0.12 (0.02)
PLR	21.9 (9.14)	0.12 (0.02)
SM	**	0.34 (0.15)

Table 4.4:

Method	MSE	Mis-Classification Rate
PMCCI	0.70 (0.10)	0.13 (0.02)
MCCI	0.62 (0.12)	0.14 (0.02)
MCC	0.64(0.13)	0.15 (0.02)
PLR	17.62 (12.26)	0.16 (0.03)
Standard	**	0.33 (0.11)

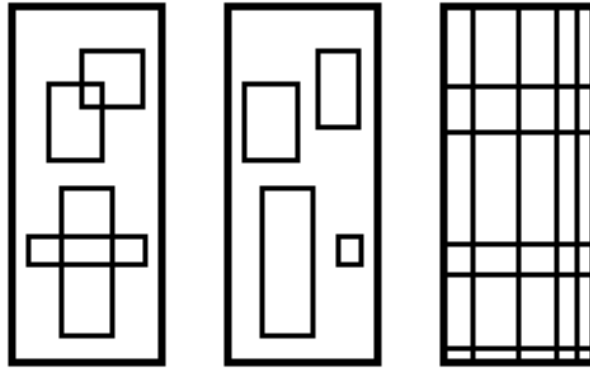


FIGURE 4.1: **A:** Multiple, overlapping co-clusters. **B:** Multiple, non-overlapping co-clusters. **C:** Single, non-overlapping co-clusters.

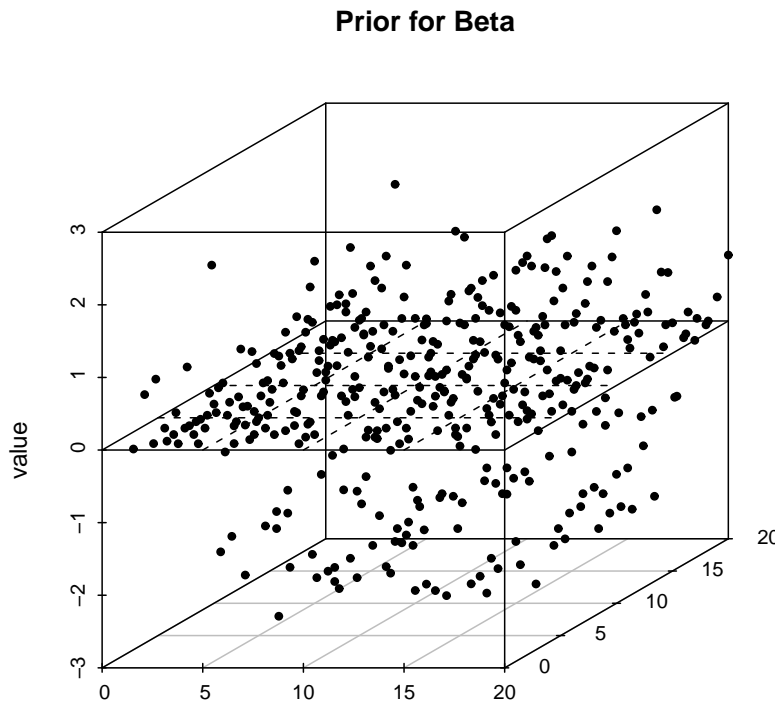


FIGURE 4.2: Prior Realizations of β_{hp}

Hospital-procedure specific sample size

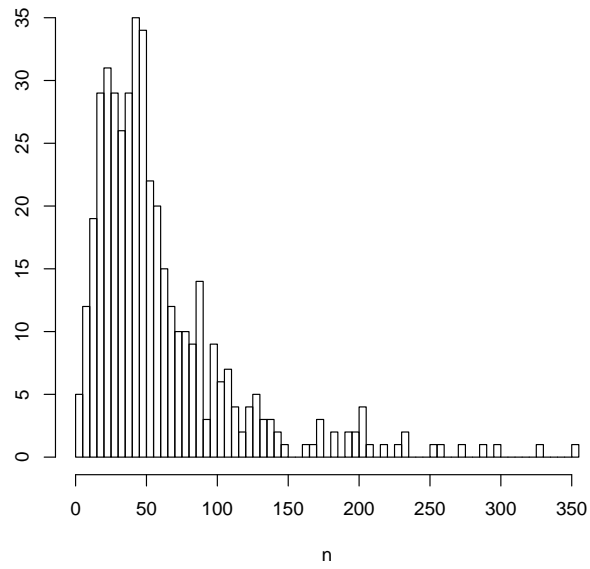
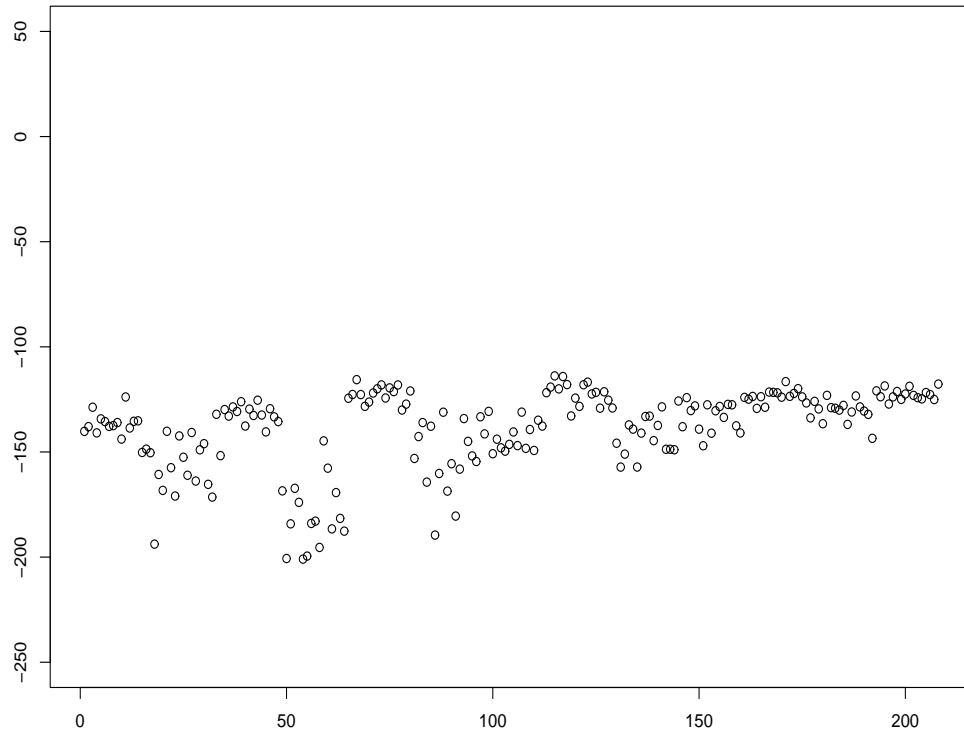


FIGURE 4.3: Hospital-Procedure specific sample size



Standard Method

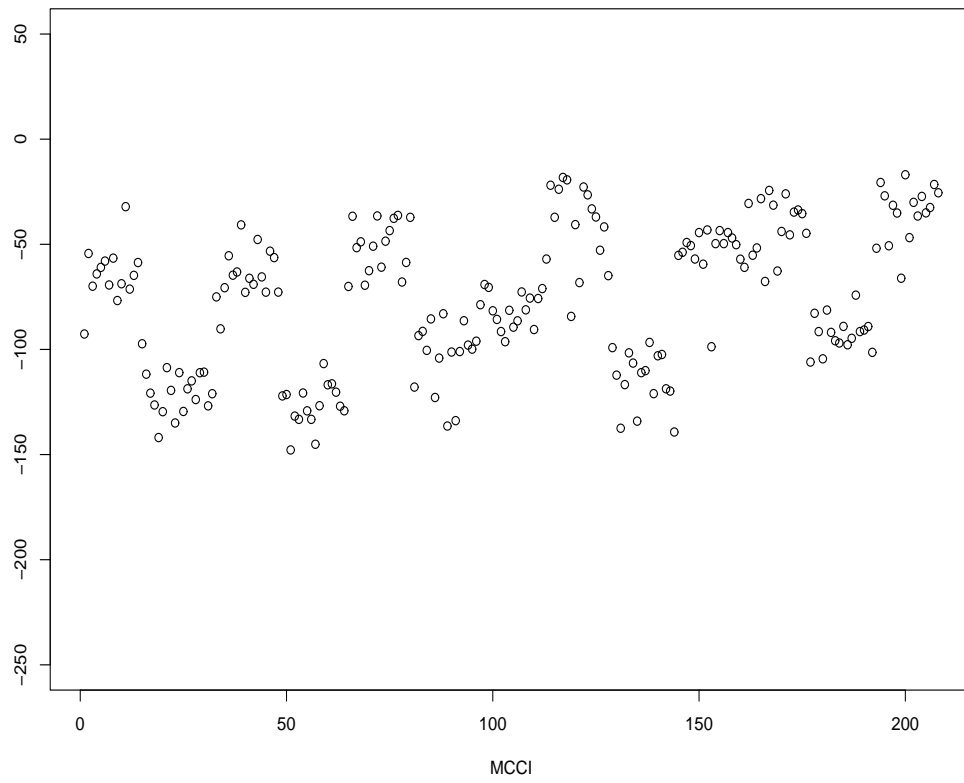


FIGURE 4.4: Posterior Marginal Likelihood Comparison

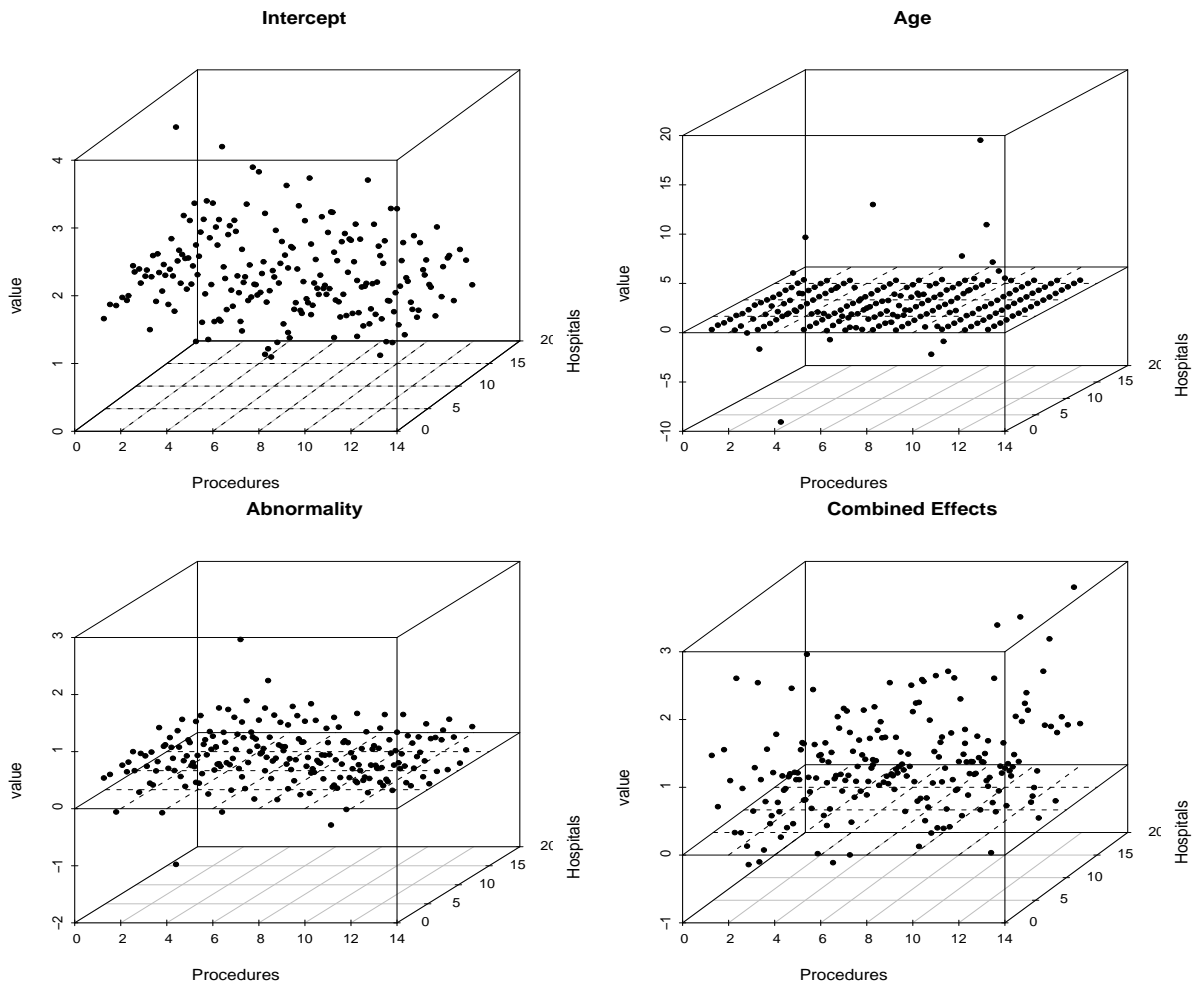


FIGURE 4.5: Posterior Mean of β_{hp}

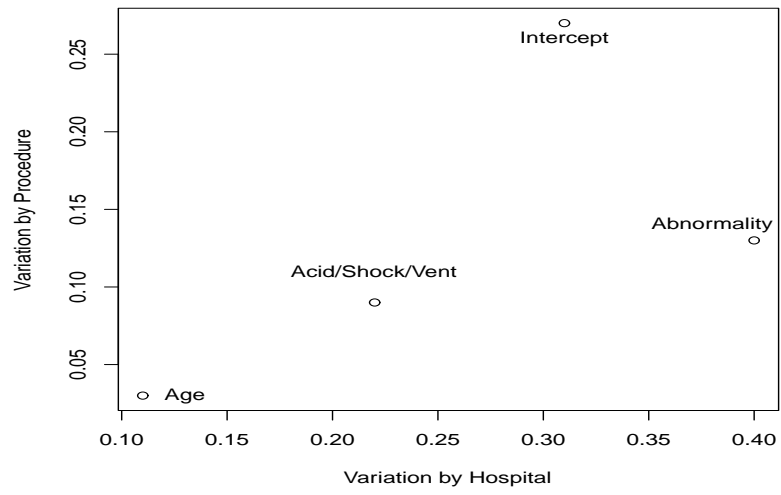


FIGURE 4.6: Variation explained by hospitals and procedures for β_{hp}

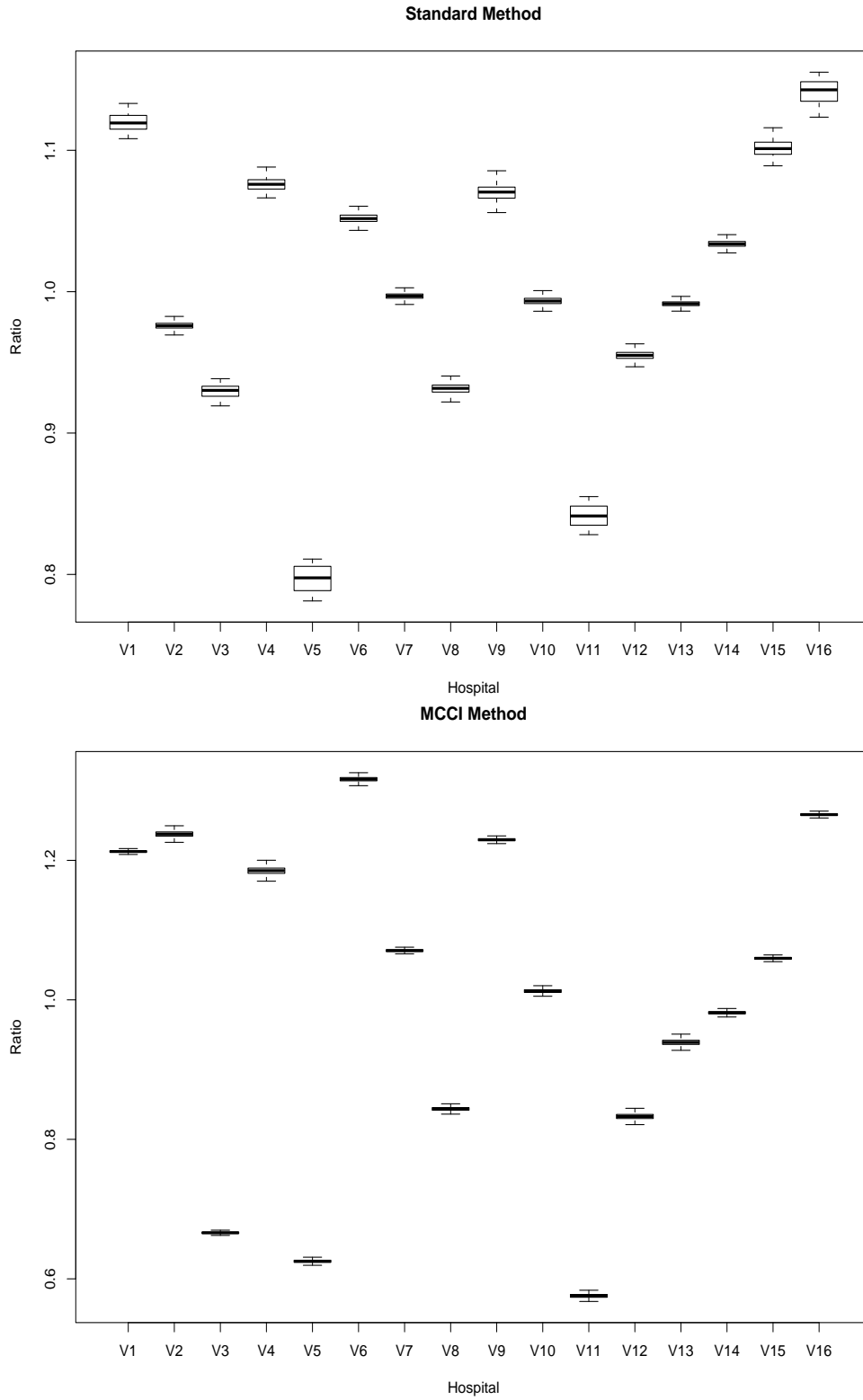


FIGURE 4.7: R_h Estimation from Standard Method and the MCCI model

Appendix A

Additional Materials for Chapter 3

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

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

As K goes to infinity,

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

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 (3.8) is,

$$\hat{y}_{ij} = \begin{cases} 1, & s_{ij} \geq 0, \\ 0, & \text{else.} \end{cases}$$

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

Parameters of equation (A.2) are updated through following steps,

1. Sampling $\hat{\boldsymbol{\beta}}_j$ in the case of a normal prior on $\hat{\boldsymbol{\beta}}_j$, $\pi(\hat{\boldsymbol{\beta}}_j) = \text{N}(b_0, v_0)$, the full conditional distribution of $\hat{\boldsymbol{\beta}}_j$ given $s_{.j}$ and $d_{.j}$ is still normal,

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

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

2. Following advice of Holmes and Held (2006), we update $\{s_{.j}, d_{.j}\}$ jointly given $\hat{\boldsymbol{\beta}}_j$,

$$\pi(s_{.j}, d_{.j} | \hat{\boldsymbol{\beta}}_j, \hat{\mathbf{y}}_{.j}) = \pi(s_{.j} | \hat{\boldsymbol{\beta}}_j, \hat{\mathbf{y}}_{.j}) \pi(d_{.j} | s_{.j}, \hat{\boldsymbol{\beta}}_j)$$

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

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

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

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

$$g(d_{lj}) = \text{GIG}(0.5, 1, r^2) = \frac{r}{\text{InvGamma}(1, r)}$$

, where $r = (s_{ij} - \hat{x}'_i \hat{\beta}_j)^2$, as rejection sampling density. Following a draw from $g(\cdot)$ the sample is accepted with probability $\alpha(\cdot)$,

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

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

$$L(d_{lj}) \propto d_{lj}^{-1/2} \exp(-0.5r^2/d_{lj})$$

, and $\pi(d_{lj})$ is the prior,

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

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

Appendix B

Additional Materials for Chapter 4

Proof for Theorem 4.2.1

Proof. The probability allocated to $\mathcal{N}_{(\infty, \epsilon)}(\mathbf{B}^0)$ can also be defined as

$$\int 1(\|\mathbf{B} - \mathbf{B}^0\|_\infty < \epsilon) d\boldsymbol{\beta}_0 \prod_{h=1}^H d\boldsymbol{\rho}_h \prod_{p=1}^P d\boldsymbol{\kappa}_p \prod_{h,p} d\boldsymbol{\psi}_{hp} \quad (\text{B.1})$$

Also

$$\|\mathbf{B} - \mathbf{B}^0\|_\infty = \max_{h,p} \|\boldsymbol{\beta}_{hp} - \boldsymbol{\beta}_{hp}^0\|_\infty \leq \max_{h,p} \left(\|\boldsymbol{\beta}_0\|_\infty + \|\boldsymbol{\rho}_h\|_\infty + \|\boldsymbol{\kappa}_p\|_\infty + \|\boldsymbol{\psi}_{hp} - \boldsymbol{\beta}_{hp}^0\|_\infty \right).$$

It is straightforward to show for $h = 1, \dots, H$ and $p = 1, \dots, P$ and $\forall \epsilon > 0$

$$\begin{aligned} \|\boldsymbol{\beta}_0\|_\infty &< \epsilon, & \|\boldsymbol{\rho}_h\|_\infty &< \epsilon, \\ \|\boldsymbol{\kappa}_p\|_\infty &< \epsilon, & \|\boldsymbol{\psi}_{hp} - \boldsymbol{\beta}_{hp}^0\|_\infty &< \epsilon \end{aligned}$$

implies that $\|\mathbf{B} - \mathbf{B}^0\|_\infty < \epsilon$. Hence to show (B.1) is strictly positive, it suffices to show

$$\begin{aligned} \Pr(\|\boldsymbol{\beta}_0\|_\infty < \epsilon, \|\boldsymbol{\rho}_h\|_\infty < \epsilon, \|\boldsymbol{\kappa}_p\|_\infty < \epsilon, \|\boldsymbol{\psi}_{hp} - \boldsymbol{\beta}_{hp}^0\|_\infty < \epsilon, \\ h = 1, \dots, H, p = 1, \dots, P) > 0 \quad (\text{B.2}) \end{aligned}$$

We also know that

(1) β_0 follows a multivariate normal distribution and

$$\rho_h \sim \tilde{\pi}_\rho \delta_0 + (1 - \tilde{\pi}_\rho) G_\rho^*, \quad \kappa_p \sim \tilde{\pi}_\kappa \delta_0 + (1 - \tilde{\pi}_\kappa) G_\kappa^*, \quad \psi_{hp} \sim \tilde{\pi}_\psi \delta_0 + (1 - \tilde{\pi}_\psi) G_\psi^*,$$

where G_ρ^* , G_κ^* and G_ψ^* are $\text{DP}(\alpha_\rho G_{0\rho}^*)$, $\text{DP}(\alpha_\kappa G_{0\kappa}^*)$ and $\text{DP}(\alpha_\psi G_{0\psi}^*)$ distributions which have full supports on the parameter space \mathcal{R}^q .

(2) For $\mathbf{B} = (\beta_{hp})$, with positive probability, the component of each cell (h, p) comes from a different cluster of the base measure and are independent of each other.

(B.2) follows directly from conditions (1) and (2). □

Proof for Theorem 4.2.2

Proof. By integrating over P , we can write p^T and p^∞ in terms of the distribution for $\mathbf{y} = (y_{hpi})$ under G^T and G^∞ respectively and call these two sampling distributions $\pi_{\mathbf{T}}(d\mathbf{B})$ and $\pi_\infty(d\mathbf{B})$.

$$\begin{aligned}
\|p^{\mathbf{T}}(\mathbf{y}) - p^\infty(\mathbf{y})\| &= \int |p^{\mathbf{T}}(\mathbf{y}) - p^\infty(\mathbf{y})| d\mathbf{y} \\
&= \int \int \left| \prod_{h,p,i} k(y_{hpi} | \beta_{hp}) (\pi_{\mathbf{T}}(d\mathbf{B}) - \pi_\infty(d\mathbf{B})) \right| d\mathbf{y} \\
&\leq \int \int \prod_{h,p,i} k(y_{hpi} | \beta_{hp}) d\mathbf{y} |\pi_{\mathbf{T}}(d\mathbf{B}) - \pi_\infty(d\mathbf{B})| \\
&= 2D(\pi_{\mathbf{T}}, \pi_\infty),
\end{aligned}$$

where $D(\mathcal{P}_1, \mathcal{P}_2) = \sup_A |\mathcal{P}_1(A) - \mathcal{P}_2(A)|$ is the total variation distance between two probability measures \mathcal{P}_1 and \mathcal{P}_2 . We can write $\rho_h = \rho_{\xi_h}^*$, $\kappa_p = \kappa_{\chi_p}^*$ and $\psi_{hp} = \psi_{\varepsilon_{hp}}^*$. The sampled values \mathbf{B} under $\pi_{\mathbf{T}}$ and π_∞ are identical when ξ_h , χ_p and ε_{hp} are sampled from values smaller than $\mathbf{T} = (T_1, T_2, T_3)$ th term for $h = 1, \dots, H$, $p = 1, \dots, P$ and $i = 1, \dots, n_{hp}$. Thus

$$\begin{aligned}
D(\pi_{\mathbf{T}}, \pi_\infty) &\leq 2 \left\{ 1 - \pi_{\mathbf{T}}(\xi_h < T_1, \chi_p < T_2, \varepsilon_{hp} < T_3, \text{ for } h = 1, \dots, H, p = 1, \dots, P, \right. \\
&\quad \left. i = 1, \dots, n_{hp}) \right\} \\
&= 2 \left[1 - \pi_{T_1} \{ \xi_h < T_1 \} \pi_{T_2} \{ \chi_p < T_2 \} \pi_{T_3} \{ \varepsilon_{hp} < T_3 \} \right] \\
&= 2 \left[1 - E \left\{ \prod_{h,i} \left(\sum_{l=1}^{T_1-1} \pi_{1l}(u_h) \right) \right\} E \left\{ \prod_{p,i} \left(\sum_{j=1}^{T_2-1} \pi_{2j}(v_p) \right) \right\} E \left\{ \left(\sum_{k=1}^{T_3-1} \pi_{3k} \right)^{\sum n_{hp}} \right\} \right]
\end{aligned}$$

We first notice that with Jensen's inequality

$$E\{\log(1 - V_l(u_h))\} \leq \log\{1 - E(V_l(u_h))\} = \log\left\{1 - E(\Phi(g_l(u_h)))\right\} < 0.$$

Therefore, $\sum_{l=1}^{\infty} E\{\log(1 - V_l(u_h))\} = -\infty$ and by theorem 2 in Ishwaran and James (2001), $\sum_{l=1}^{\infty} \pi_{1l}(u_h) = 1$ almost surely. With this conclusion and the Jensen's inequality, we have

$$\begin{aligned} E\left\{\prod_{h,i} \left(\sum_{l=1}^{T_1-1} \pi_{1l}(u_h)\right)\right\} &\geq \prod_{h,i} \left\{\sum_{l=1}^{T_1-1} E(\pi_{1l}(u_h))\right\} \\ &= \prod_{h,i} \left[1 - E\left(V_{T_1}(u_h) \prod_{l=1}^{T_1-1} \{V_l(u_h)\}\right)\right] \end{aligned}$$

We know that

$$\begin{aligned} E(V_l(u_h)) &= E(\Phi(g_l(u_h))) = E\left[\frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{g_l(u_h)}{\sqrt{2}}\right)\right)\right] \\ &= \frac{1}{2}\left[1 + E\left(\frac{\sqrt{2}}{\sqrt{\pi}}\left(g_l(u_h) - \frac{g_l(u_h)^3}{3} + \frac{g_l(u_h)^5}{10} - \frac{g_l(u_h)^7}{42} + \dots\right)\right)\right], \end{aligned}$$

where erf is the error function and with the taylor expansion one can get the above last equation. It is straightforward to verify that $E(g_l(u_h)^k) = 0$ with k being odd numbers. For example with $\zeta_{lh} \sim N(0, \alpha_{lh}^{-1})$ for $h = 1, \dots, H$:

$$\begin{aligned} E(g_l(u_h)^k) &= E\left(\zeta_{l0}^k + a_1 \zeta_{l0}^{k-1} \sum_{j=1}^H \zeta_{lj} \mathcal{K}(u_h, u_j; \varrho_l) + a_2 \zeta_{l0}^{k-2} \left(\sum_{j=1}^H \zeta_{lj} \mathcal{K}(u_h, u_j; \varrho_l)\right)^2\right. \\ &\quad \left.+ \dots + a_k \left(\sum_{j=1}^H \zeta_{lj} \mathcal{K}(u_h, u_j; \varrho_l)\right)^k\right). \end{aligned}$$

Since ζ_{lh} , $\mathcal{K}(u_h, u_j; \varrho_l)$ are independent of each other and $E(\zeta_{lh}^k) = 0$ when k are odd numbers, it can be shown that $E(g_l(u_h)^k) = 0$.

It turns out that

$$E\left\{\prod_{h,i} \left(\sum_{l=1}^{T_1-1} \pi_{1l}(u_h)\right)\right\} \geq \left[1 - \left(\frac{1}{2}\right)^{T_1}\right]^{\sum_h n_{hp}}.$$

and

$$E\left\{\prod_{p,i} \left(\sum_{j=1}^{T_2-1} \pi_{2j}(v_p)\right)\right\} \geq [1 - (\frac{1}{2})^{T_2}]^{\sum_p n_{hp}}.$$

Following Ishwaran and James (2001), we have

$$\begin{aligned} E\left\{\left(\sum_{k=1}^{T_3-1} \pi_{3k}(u_h)\right)^{\sum n_{hp}}\right\} &\approx \left(1 - \exp\left(-\frac{T_3-1}{\alpha_\psi}\right)\right)^{\sum n_{hp}} \\ &\approx 1 - \left(\sum_{h,p} n_{hp}\right) \exp\left(-\frac{T_3-1}{\alpha_\psi}\right). \end{aligned}$$

Generalize the above results, we have

$$D(\pi_{\mathbf{T}}, \pi_{\infty}) \leq 2 \left[1 - \left\{1 - \left(\frac{1}{2}\right)^{T_1}\right\}^{\sum_h n_{hp}} \left\{1 - \left(\frac{1}{2}\right)^{T_2}\right\}^{\sum_p n_{hp}} \left\{1 - \left(\sum_{h,p} n_{hp}\right) \exp\left(-\frac{T_3-1}{\alpha_\psi}\right)\right\} \right].$$

□

Bibliography

- Agarwal, D. and Merugu, S. (2007), “Predictive Discrete Latent Factor Models for Large Scale Dyadic Data,” *KDD-2007*.
- Antoniak, C. (1974), “Mixtures of Dirichlet processes with application to Bayesian non-parametric problems,” *Annals of Statistics*, 2, 1152–1174.
- Austin, P., Naylor, C., and Tu, J. (2001), “A Comparison of a Bayesian versus a frequentist method for profiling hospital performance,” *J Eval Clin Pract*, 7, 35–45.
- Austin, P., Alter, D., and Tu, J. (2003), “The Use of Fixed- and Random-Effects Models for Classifying Hospitals as Mortality Outliers: A Monte Carlo Assessment,” *Medical Decision Making*, 23, 526–539.
- Blackwell, D. and MacQueen, J. (1973), “Ferguson distribution via Pólya urn schemes,” *Annals of Statistics*, 1, 353–355.
- Blei, D. and Frazier, P. (2009), “Distance dependent Chinese restaurant processes,” *arXiv*.
- Blei, D. and Frazier, P. (2010), “Distance dependent Chinese restaurant processes,” *Proceedings for the 27th International Conference on Machine Learning, Haifa, Israel, 2010*.
- Booth, J. and Hobert, J. (1999), “Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm,” *Journal of Royal Statistical Society, Series B*, 61, 265–285.
- Bornn, L., Doucet, A., and Gottardo, R. (2007), “The Bayesian Elastic Net,” *CMS-MITACS Joint Conference*, pp. –.
- Chen, M., Carlson, D., Zaas, A., Woods, C., Ginsburg, G., Hero, A., Lucas, J., and Carin, L. (2010), “The Bayesian Elastic Net: Calssifying Multi-Task Gene-Expression Data,” *IEEE Trans. Biomedical Engineering (To appear)*, pp. –.
- Chung, Y. and Dunson, D. (2009), “Nonparametric Bayes Conditional Distribution Modeling with Variable Selection,” *Journal of the American Statistical Association*, 104, 1646–1660.

- Damien, P. and Wakefield, J. (1999), “Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables,” *Journal of Royal Statistical Society, Series B*, 61, 331–344.
- Dunson, B., Herring, A., and Engel, S. (2008a), “Bayesian selection and clustering of polymorphisms in functionally related genes,” *Journal of the American Statistical Association*, 103, 534–546.
- Dunson, D. and Peddada, S. (2008), “Bayesian nonparametric inference on stochastic ordering,” *Biometrika*, 95, 859–874.
- Dunson, D., Xue, Y., and Carin, L. (2008b), “The matrix stick-breaking process: flexible Bayes meta analysis,” *Journal of the American Statistical Association*, 103, 317C327.
- Escobar, M. (1994), “Estimating normal means with a Dirichlet process prior,” *Journal of the American Statistical Association*, 89, 268–277.
- Escobar, M. and West, M. (1998), “Computing nonparametric hierarchical models,” *Practical nonparametric and semi-parametric Bayesian statistics*, pp. 1–22.
- Ferguson, T. (1973), “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, 1, 615–629.
- Ferguson, T. (1974), “Prior distributions on spaces of probability measures,” *Annals of Statistics*, 2, 209–230.
- Fraley, C. and Raftery, A. E. (2002), “Model-based clustering, discriminant analysis, and density estimation,” *Journal of the American Statistical Association*, 97, 611–631.
- Fraley, C., Raftery, A. E., and Wehrens, R. (2005), “mclust: Model-based Cluster Analysis,” *R package version 2.1-11*.
- Gelman, A. and Rubin, D. B. (1992), “Inference from iterative simulation using multiple sequences,” *Statistical Science*, 7, 457–511.
- Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008), “A default prior distribution for logistic and other regression models,” *The Annals of Applied Statistics*, 2, 1360–1383.
- George, E. (1986), “Minimax multiple shrinkage estimation,” *The Annals of Statistics*, 14, 188–205.
- Hoerl, A. and Kennard, R. (1988), “Ridge regression,” *Encyclopedia of Statistical Science*, 8, 129–136.
- Hoff, P. (2003a), “Bayesian methods for partial stochastic ordering,” *Biometrika*, 90, 303–317.

- Hoff, P. (2003b), “Nonparametric estimation of convex models via mixtures,” *Annals of Statistics*, 31, 174–200.
- Holmes, C. and Held, L. (2006), “Bayesian auxiliary variable models for binary and multinomial regression,” *Bayesian Analysis*, 1, 145–168.
- Ishwaran, H. and James, L. (2001), “Gibbs Sampling Methods for Stick-Breaking Priors,” *Journal of the American Statistical Association*, 96, 161–173.
- Ishwaran, H. and Zarepour, M. (2002), “Dirichlet prior sieves in finite normal mixtures,” *Statistica Sinica*, 12, 941–963.
- Jasra, A., Holmes, C. C., and Stephens, D. A. (2005), “Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling,” *Statistical Science*, 20, 50–67.
- Jenkins, K. (2004), “Risk adjustment for congenital heart surgery: the RACHS-1 method,” *Seminars in thoracic and cardiovascular surgery. Pediatric cardiac surgery annual*, 7, 180–184.
- Karabatsos, G. and Walker, S. (2007), “Bayesian nonparametric inference of stochastically ordered distributions, with Polya trees and Bernstein polynomials,” *Statistics and Probability Letters*, 77, 907–913.
- Krumholz, H., Brindis, R., Brush, J., Cohen, D., and etc (2006), “Standards for Statistical Models Used for Public Reporting of Health Outcomes,” *Circulation*, 113, 456–462.
- Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2009), “Characterizing the Variance Improvement in Linear Dirichlet Random Effects Models,” *Statistics and Probability Letters*, 79, 2343–2350.
- Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), “Penalized Regression, Standard Errors, and Bayesian Lassos,” *Technical report, Center for Applied Statistics, Washington University*, pp. –.
- 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., Tcherwenkov, C., Mavroudis, C., and Committee, T. A. (2004), “The Aristotle score: a complexity-adjusted method to evaluate surgical results,” *European Journal of Cardio-Thoracic Surgery*, 25, 911–924.
- Landon, B., Normand, S., Lessler, A., O’Mallery, A., Schmaltz, S., Loeb, J., and McNeil, B. (2006), “Quality of Care for the Treatment of Acute Medical Conditions in US Hospitals,” *Arch Intern Med*, 166, 2511–2517.

- Landrum, M., Bronskill, S., and Normand, S. (2000), “Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers,” *Health Services And Outcomes Research Methodology*, 1, 23–47.
- Lavine, M. (1992), “Some aspects of Pólya tree distributions for statistical modeling,” *The Annals of Statistics*, 20, 1222–1235.
- Lavine, M. (1994), “More aspects of Pólya tree distributions for statistical modeling,” *The Annals of Statistics*, 22, 1161–1176.
- Lin, R., Louis, T., Peddock, S., and Ridgeway, G. (2006), “Loss Function Based Ranking in Two-Stage, Hierarchical Models,” *Bayesian Analysis*, 4, 915–946.
- Lin, R., Louis, T., Peddock, S., and Ridgeway, G. (2009), “Ranking USRDS provider specific SMRs from 1998-2001,” *Health Serv Outcomes Res Methodol*, 9, 22–38.
- Liu, F., Dunson, D., and Zou, F. (2010), “High-Dimensional Variable Selection in Meta Analysis for Censored Data,” *Biometrics (To appear)*, pp. –.
- Lo, A. (1984), “On a class of Bayesian nonparametric estimates: I. density estimates,” *Annals of Statistics*, 12, 351–357.
- Louis, T. (1982), “Finding the Observed Information Matrix when Using the EM Algorithm,” *Journal of Royal Statistical Society, Series B*, 44, 226–233.
- MacEachern, S. (1999), “Dependent nonparametric processes,” *In ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.
- MacEachern, S. (2000), “Dependent Dirichlet processes,” *Tech. rep., Ohio State University, Department of Statistics*.
- MacLehose, R. and Dunson, D. (2010), “Bayesian semi parametric multiple shrinkage,” *Biometrics*, 66, 455–462.
- MacLehose, R., Dunson, D., Herring, A., and Hoppin, J. (2007), “Bayesian methods for highly correlated exposure data,” *Epidemiology*, 18, 199–207.
- Meeds, E. and Roweis, S. (2007), “Nonparametric Bayesian Biclustering,” *Technical Report, Department of Computer Science, University of Toronto*.
- Nott, D. (2008), “Predictive performance of Dirichlet process shrinkage methods in linear regression,” *Computational Statistics and Data Analysis*, 52, 3658–3669.
- O’Brien, S. and Dunson, D. (2004), “Bayesian multivariate logistic regression,” *Biometrics*, 60, 739–746.
- Papaspiliopoulos, O. and Roberts, G. (2008), “Retrospective MCMC for Dirichlet process hierarchical models,” *Biometrika*, 95, 169–186.

- Park, M. and Hastie, T. (2008), “Penalized logistic regression for detecting gene interactions,” *Biostatistics*, 9, 30–50.
- Racz, M. and Sedransk, J. (2010), “Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes,” *Journal of the American Statistical Association*, 105, 49–58.
- Rodriguex, A., Dunson, D., , and Gelfand, A. (2008), “The Nested Dirichlet Process,” *Journal of the American Statistical Association*, 103, 1131–1154.
- Rodriguez, A. and Dunson, D. (2009), “Nonparametric Bayesian models through probit stickbreaking processes,” *Technical Report UCSC-SOE-09-12, University of California Santa Cruz*.
- Rodríguez, A., Dunson, D. B., and Gelfand, A. (2008), “The Nested Dirichlet Process,” *Journal of the American Statistical Association*, 103, 1131–1154.
- Sethuraman, J. (1994), “A constructive definition of dirichlet priors,” *Statistica Sinica*, 4, 639–650.
- Shafiei, M. and Milios, E. (2006), “Latent Dirichelt Co-Clustering,” *Proceedings of the Sixth International Conference on Data Mining*.
- Shahian, D. and Normand, S. (2008), “Comparison of “Risk-Adjusted” Hospital Outcomes.” *Health Serv Outcomes Res Methodol*, 117, 1955–1963.
- Shahian, D., Normand, S., Torchiana, D., Lewis, S., Pastore, J., Kuntz, R., and Dreyer, P. (2001), “Cardiac Surgery Report Cards: Comprehensive Review and Statistical Critique,” *Ann Thorac Surg*, 72, 2155–2168.
- Shahian, D., Torchiana, D., Shemin, R., Rawn, J., and Norman, S. (2005), “Massachusetts Cardiac Surgery Report Card: Implications of Statistical Methods.” *Ann Thorac Surg*, 80, 2106–2113.
- Spärck Jones, K. (1972), “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, 28, 11–21.
- Stephens, M. (2000), “Dealing with label switching in mixture models,” *Journal of the Royal Statistical Society, Ser. B*, 62, 795–809.
- Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), “Sharing clusters among related groups: Hierarchical Dirichlet processes,” *Journal of the Amercian Statistical Association*, 101, 1566–1581.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of Royal Statistical Society, Series B*, 58, 267–288.

- Timbie, J. and Normand, S. (2008), “A Comparison of methods for combining quality and efficiency performance measures: Profiling the value of hospital care following acute myocardial infarction,” *Statistics in Medicine*, 27, 1351–1370.
- Walker, S. (2007), “Sampling the Dirichlet mixture model with slices,” *Communications in Statistics: Simulation and Computation*, 36, 45–54.
- Wei, G. and Tanner, M. (1990), “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms,” *Journal of the American Statistical Association*, 85, 699–704.
- Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007), “Multi-Task Learning for Classification with Dirichlet Process Priors,” *Journal of Machine Learning Research*, 8, 35–63.
- Yau, C., Papaspiliopoulos, O., Roberts, G., and Holmes, C. (2010), “Nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes,” *Journal of Royal Statistical Society: Series B (to appear)*.
- Zou, H. and Hastie, T. (2005), “Regularization and variable selection via the elastic net,” *Journal of Royal Statistical Society*, 67, 301–320.

Biography

Hongxia Yang is a Ph.D. student of Statistical Science at Duke University in Durham, NC. She received the B.S. degree in statistics (2007) from Nankai University in Tianjin. Her research interests are in Bayesian statistics, with particular applications to nonparametric Bayesian statistics applied to biostatistics and bioinformatics. Her work in Bayesian statistics involves parametric and nonparametric Bayesian methodology, latent variable methods, and machine learning.