The Matrix Stick-Breaking Process for Flexible Multi-Task Learning

Ya Xue YX10@EE.DUKE.EDU
Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708 USA

David Dunson DUNSON@STAT.DUKE.EDU
Biostatistics Branch, National Institute of Environmental Health Sciences, RTP, NC 27709 USA

Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708 USA

Abstract

In multi-task learning our goal is to design regression or classification models for each of the tasks and appropriately share information between tasks. A Dirichlet process (DP) prior can be used to encourage task clustering. However, the DP prior does not allow local clustering of tasks with respect to a subset of the feature vector without making independence assumptions. Motivated by this problem, we develop a new multi-task-learning prior, termed the matrix stick-breaking process (MSBP), which encourages cross-task sharing of data. However, the MSBP allows separate clustering and borrowing of information for the different feature components. This is important when tasks are more closely related for certain features than for others. Bayesian inference proceeds by a Gibbs sampling algorithm and the approach is illustrated using a simulated example and a multi-national application.

1. Introduction

Multi-task learning (MTL) is a problem of increasing interest within the machine-learning community (Caruana, 1997; Thrun & Pratt, 1998). Hierarchical Bayesian modeling techniques have proven to be particularly powerful for this problem, providing a framework wherein an appropriate level of cross-task sharing is learned based on the data itself. Considering sample i $(i = 1, \dots, n_m)$ from task m $(m = 1, \dots, M)$, suppose that data consist of a response variable, y_{mi} and a feature vector, $\mathbf{x}_{mi} = (x_{mi1}, \dots, x_{mid})'$; for regression problems y_{mi} is real and for classification problems it is integer. For concreteness, we focus on the regression problem in this paper.

A common strategy for analysis is to use a hierarchical model of the form:

$$\begin{aligned} y_{mi} &\sim f(\mathbf{x}_{mi}, \boldsymbol{\theta}_m, \boldsymbol{\phi}) \\ \boldsymbol{\theta}_m &\sim G, \end{aligned}$$

where $f(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is the conditional distribution of y given feature vector $\mathbf{x}$ and parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_q)'$ is a vector of global parameters, $\boldsymbol{\theta}_m = (\theta_{m1}, \dots, \theta_{md})'$ is a vector of task-specific parameters, and G is the distribution of $\boldsymbol{\theta}_m$ across tasks. Typically, G is assumed to correspond to a normal distribution.

In multi-task learning we wish to borrow information across tasks in estimating the task-specific parameters. A natural choice of prior for G that induces clustering of tasks having identical coefficients is the Ferguson Dirichlet process (DP) (Ferguson, 1973). Refer to Mukhopadhyay and Gelfand (1997), Yu et al. (2004) and Xue et al. (2007) for articles on using the DP as a prior to encourage clustering between multiple tasks.

To illustrate an important drawback of DP, suppose:

$$\boldsymbol{\theta}_m \overset{iid}{\sim} G, \quad G \sim DP(\alpha G_0), \tag{1}$$

where α is a precision parameter and G_0 is the base measure of the DP. Following the Pólya urn result of Blackwell and MacQueen (1973), expression (1) implies that tasks m and m' are clustered together with prior probability, $\Pr(\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m'}) = 1/(1+\alpha)$. Hence, this formulation clusters the coefficients for all features in the same manner, and therefore it does not afford the flexibility to allow feature-dependent task clustering. This flexibility may be desirable. Take as an

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

example the multi-national validation study of the rodent uterotrophic bioassay in Kanno et al. (2001), which is a system for identifying suspected agonists or antagonists of estrogen. The primary goal of the study was to assess heterogeneity across labs in the different predictor (feature component) effects, while clustering labs with similar coefficients.

As an alternative approach to (1), we could specify independent DP priors for the coefficients as follows:

$$\theta_{mj} \overset{iid}{\sim} G_j, \quad G_j \sim DP(\alpha_j G_{0j}).$$

This approach allows differential clustering of the coefficients for different feature components, so is more flexible. However, independence is assumed across the feature components. This is unappealing, because $\theta_{mj} = \theta_{m'j}$ provides information that tasks m and m' are similar, which should intuitively increase the probability that $\theta_{mj'} = \theta_{m'j'}$, for $j' \neq j$.

Motivated by borrowing of information across related nonparametric Bayesian models, Müller et al. (2004) proposed an approach for incorporating dependence in G_j and $G_{j'}$ by using a mixture specification in which $G_j \sim \pi_0 F_0 + (1-\pi_0)F_j$, with $0 \leq \pi \leq 1$ a mixture probability, $F_0 \sim DP(\alpha F_0^*)$ a global distribution, and $F_j \sim DP(\beta F_j^*)$ a local deviation. A hierarchical DP (HDP), which also shares atoms across dependent distributions, was proposed by Teh et al. (2006). An alternative strategy, which allows the atoms in the different distributions to be dependent but not identical, was proposed by De Iorio et al. (2004), relying on the dependent DP (DDP) of MacEachern (1999).

For multi-task learning, these methods are not appropriate, because the regression coefficients for the different features are on intrinsically different scales. Hence, it is not reasonable to borrow information by allowing identical coefficients for different features. Instead, we propose a novel approach for borrowing of information across feature components in the clustering process using a matrix stick-breaking process (MSBP). We follow recent authors (Griffin & Steel, 2006; Dunson et al., 2007a) by incorporating dependency through the stick-breaking weights. However, our approach is fundamentally different from these approaches, and has different applications.

2. Matrix Stick-Breaking Process

We first provide a brief review of the stick-breaking formulation of the DP (Sethuraman, 1994), which is useful for inferring properties. Assuming $G \sim DP(\alpha G_0)$, we have

$$G = \sum_{h=1}^{\infty} \left\{ V_h \prod_{l<h} \overline{V}_l \right\} \delta_{\Theta_h}, \tag{2}$$

$$V_h \overset{iid}{\sim} \text{beta}(1, \alpha), \quad \Theta_h \overset{iid}{\sim} G_0,$$

where $\mathbf{V} = \{V_h, h = 1, \dots, \infty\}$ is an infinite sequence of stick-breaking random probabilities, $\overline{V}_h = 1 - V_h$, and $\mathbf{\Theta} = \{\Theta_h, h = 1, \dots, \infty\}$ is an infinite sequence of random atoms. Because the probability weights, $\pi_h = V_h \prod_{l<h} \overline{V}_l$, are stochastically decreasing in h, an accurate approximation to (2) can be obtained by truncating the infinite sum at N terms, with $V_N = 1$ (Ishwaran and James, 2001, and references therein).

We propose a *matrix stick-breaking process* (MSBP), which is motivated by the desire to borrow information across feature components and tasks simultaneously. We borrow information by incorporating dependence in the prior distributions for the coefficients $\{\theta_{mj}\}$, recalling that θ_{mj} is the coefficient for feature j in task m. We start by assuming

$$\begin{aligned} \theta_{mj} &\overset{ind}{\sim} G_{mj}, \quad m = 1, \dots, M, \quad j = 1, \dots, d, \\ \mathcal{G} &\sim \mathcal{P}, \end{aligned} \tag{3}$$

where $\mathcal{G} = \{G_{mj}, m = 1, \dots, M, j = 1, \dots, d\}$ is a matrix of random probability measures, and $\mathcal{P}$ is a probability measure on $(\Omega, \mathcal{F})$, with Ω the space of $M \times d$ matrices with the (m, j) element a probability measure on $(\mathcal{X}_j, \mathcal{B}_j)$. Here, $\mathcal{F}$ is a σ-algebra of subsets of Ω, $\theta_{mj} \in \mathcal{X}_j$ (typically, $\mathcal{X}_j = \Re$), and $\mathcal{B}_j$ is a Borel σ-algebra of subsets of $\mathcal{X}_j$.

Our focus is on the specification for $\mathcal{P}$. Assuming each element in $\mathcal{G}$ has a stick-breaking representation, we let:

$$G_{mj} = \sum_{h=1}^{N} \left\{ V_{mjh} \prod_{l<h} \overline{V}_{mjl} \right\} \delta_{\Theta_{jh}}, \quad \Theta_{jh} \overset{ind}{\sim} G_{0j}, \tag{4}$$

where $\mathbf{V} = \{V_{mjh}, m = 1, \dots, M, j = 1, \dots, d, h = 1, \dots, N\}$ is an array of random stick-breaking weights, and $\mathbf{\Theta} = \{\Theta_{jh}\}$ is a $d \times N$ matrix of random atoms. The rows $(j = 1, \dots, d)$ of $\mathbf{\Theta}$ correspond to the different feature components, while the columns $(h = 1, \dots, N)$ correspond to the different clusters. Here, $V_{mjN} = 1$, for all m, j, to ensure that the elements of $\boldsymbol{\pi}_{mj} = \{V_{mjh} \prod_{l<h} \overline{V}_{mjl}, h = 1, \dots, N\}$ sum to one, for each m, j, so that (4) is a valid probability measure.

Dependence within rows and columns of $\mathcal{G}$ will be incorporated through (i) dependent stick-breaking weights and (ii) a common parametric prior, $G_0 = \bigotimes_{j=1}^{d} G_{0j}$, across the different tasks. Focusing on the stick-breaking component, we let

$$\begin{aligned} &V_{mjh} = U_{mh} W_{jh}, \\ &U_{mh} \overset{iid}{\sim} \text{beta}(1, \alpha), \quad W_{jh} \overset{iid}{\sim} \text{beta}(1, \beta), \end{aligned} \tag{5}$$

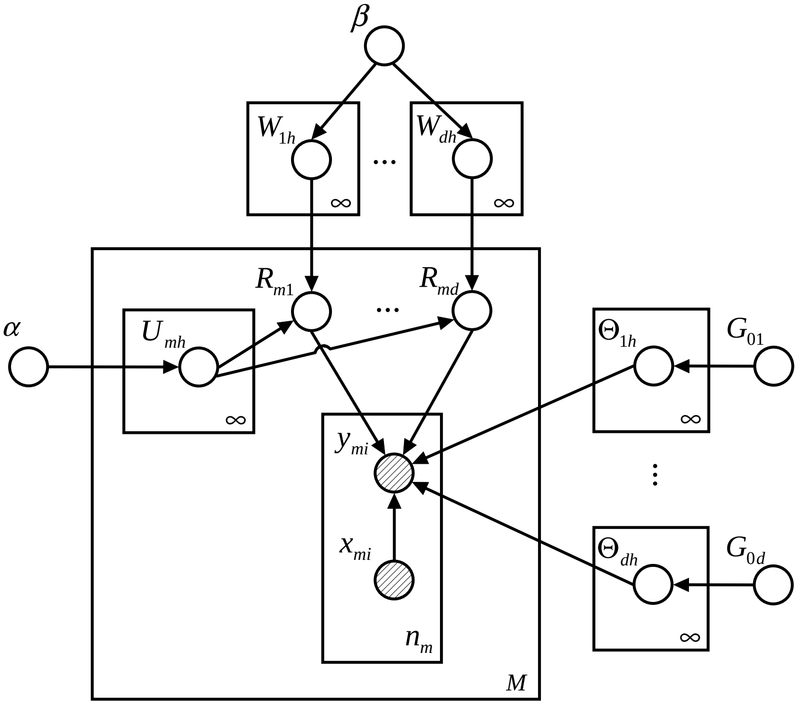


Figure 1. Graphical representation of the MTL model with prior $\text{MSBP}(\alpha, \beta, G_0)$.

so that the probability, V_{mjh}, is decomposed into the product of U_{mh}, which measures the tendency to allocate task m to cluster h, and W_{jh}, which measures the tendency to allocate coefficients for feature component j to cluster h. We use $\text{MSBP}_N(\alpha, \beta, G_0)$ to denote the choice of $\mathcal{P}$ specified in (4) and (5), with $\text{MSBP}(\alpha, \beta, G_0)$ denoting the limiting case as $N \to \infty$.

Figure 1 shows the graphical representation of the multi-task learning model with the MSBP prior. We introduce a latent indicator $\mathbf{R}$. Let $R_{mj} = h$ denote that $\theta_{mj} = \Theta_{jh}$, so that the coefficient j for task m is allocated to the hth cluster.

To provide an intuitive explanation for the form imposed in (5), first consider the sticks W_{jh}, $h = 1, \ldots, \infty$. If W_{jh} is large for a particular index h^*, then the corresponding parameter Θ_{jh^*} is likely to be shared among multiple tasks. Specifically, all tasks for which U_{mh^*} is large are likely to share the parameter Θ_{jh^*}. Hence, if there is sharing of parameter Θ_{jh^*} among some tasks, it is likely there will be sharing of this same parameter among other tasks, particularly for a large number of tasks. We also note that this sharing among tasks is encouraged by large U_{mh^*}. Importantly, since U_{mh^*} may be large for multiple different tasks m, this implies that if parameter sharing occurs for one predictor among these multiple tasks, then it is also likely that there will be sharing for other predictors (those predictors with indices j with large W_{jh^*}). We therefore recognize the following key properties of the MSBP: (i) if a given parameter for predictor j, Θ_{jh^*}, is shared among some of the tasks, it is more likely to be shared among other tasks (those tasks with large U_{mh^*}), and (ii) if sharing occurs between multiple predictors for a subset of tasks, then it is more encouraged that sharing will occur between other predictors within these tasks. These two properties motivate the form of the MSBP, and are demonstrated explicitly through subsequent analysis below.

To further motivate this choice, it is useful to consider some special cases. First, note that in the limit as $\alpha \to 0$, $U_{mh} \to 1$ a.s. and $V_{mjh} = W_{jh}$. Then, we have

$$G_{mj} = \sum_{h=1}^{N} \left\{ W_{jh} \prod_{l<h} \overline{W}_{jl} \right\} \delta_{\Theta_{jh}} = G_j,$$

where $\Theta_{jh} \overset{ind}{\sim} G_{0j}$ and $W_{jh} \overset{iid}{\sim} \text{beta}(1, \beta)$. Note that this is the truncation approximation to a Dirichlet process prior, as described by Ishwaran and James (2001), and in the limit as $N \to \infty$, $G_j \sim DP(\beta G_{0j})$. This special case corresponds to choosing independent DP priors for the distribution of the regression coefficients for the different feature components. Hence, there is no borrowing of information across the feature components, only within a particular feature component across the tasks. In the further special case in which

$\alpha \to 0$ and $\beta \to \infty$, we instead have $\theta_{mj} \sim G_{0j}$, which corresponds to a parametric hierarchical model. In addition, when $\alpha \to 0$ and $\beta \to 0$, we instead have that $\theta_{mj} = \theta_j$, so that the coefficients for the different tasks are identical and the data are pooled.

Borrowing of information across feature components occurs for $\alpha > 0$, with the random variables $\mathbf{U}_m = \{U_{mh}, h = 1, \ldots, N\}$ controlling the tendency of coefficients for task m to be allocated to particular clusters with high probability. The dependence structure and other properties are described in detail in the next subsection focusing on the case in which $N \to \infty$. Refer to Dunson et al. (2007b) for proofs of the theorems.

2.1. Basic Properties

Letting $\pi_{mjh} = V_{mjh} \prod_{l<h} \overline{V}_{mjl}$, for $m = 1, \ldots, M, j = 1, \ldots, d, h = 1, \ldots, \infty$, G_{mj} is a well defined probability measure if and only if the random weights $\boldsymbol{\pi}_{mj} = \{\pi_{mjh}, h = 1, \ldots, \infty\}$ sum to one almost surely.

Lemma 1. For $\mathcal{G} = \{G_{mj}, m = 1, \ldots, M, j = 1, \ldots, d\}$, with the elements defined in (4) and (5) for $N \to \infty$, we have $\sum_{h=1}^{\infty} \pi_{mjh} = 1$ a.s. for all m, j.

Theorem 1 provides the prior mean and variance of the random measure G_{mj}.

Theorem 1. Letting G_{mj} denote the random measure defined in (4) and (5), for $N \to \infty$, and $A \in \mathcal{B}_j$, we have

$$\begin{aligned} \mathrm{E}\big\{G_{mj}(A)\big\} &= G_{0j}(A), \\ \mathrm{V}\big\{G_{mj}(A)\big\} &= \tfrac{2}{(\alpha+2)(\beta+2)-2} G_{0j}(A)\{1 - G_{0j}(A)\}. \end{aligned}$$

Note that the prior for G_{mj} is centered on G_{0j}, which corresponds to a probability measure obeying a parametric law. For example, a convenient choice is $G_{0j}(A) = \int_A (2\pi\psi_j^{-1})^{-1/2} \exp\big\{ - \psi_j/2(z - \mu_j)^2\big\} dz$. In this case, the prior is centered on a normal hierarchical model having $\theta_{mj} \sim \mathrm{N}(\mu_j, \psi_j^{-1})$. Confidence in this normal model is controlled by the precision parameters α and β, with $\mathrm{V}\{G_{mj}(A)\} \to 0$ in the limit as either α or $\beta \to \infty$.

Theorem 2 characterizes the correlation between the random measures, G_{mj} and $G_{m'j}$, corresponding to the priors on the jth coefficient for tasks m and m'.

Theorem 2. Letting $\mathcal{G}$ denote the array of random measures defined in (4) and (5), for $N \to \infty$, and $A \in \mathcal{B}_j$, we have

$$\rho = \mathrm{corr}\big\{G_{mj}(A), G_{m'j}(A)\big\} = \frac{\alpha + \beta + \alpha\beta/2 + 1}{2\alpha + \beta + \alpha\beta + 1}.$$

The expression in Theorem 2 is particularly useful in being free from the set A, so that it can be used as a general summary of correlation in the random measures. Focusing on limiting cases, we obtain (i) $\lim_{\alpha\to 0} \rho = 1$, (ii) $\lim_{\beta\to 0} \rho = (1+\alpha)/(1+2\alpha)$, (iii) $\lim_{\alpha\to\infty} \rho = 1/2$, and (iv) $\lim_{\beta\to\infty} \rho = \frac{1+\alpha/2}{1+\alpha}$. In general, $0 \le \rho \le 1$, with the correlation coefficient increasing as α decreases.

2.2. Truncation Approximations

The $N \to \infty$ formulation of the MSBP is appealing in avoiding the need to choose a bound N on the number of components. However, in practice, computation for the infinite-dimensional specification is infeasible and it is useful to consider finite N approximations. In this subsection, we assess the approximation error using an approach related to Ishwaran and James (2001).

Theorem 3. Let $\boldsymbol{\pi}_{mj} = \{\pi_{mjh}, h = 1, \ldots, \infty\}$ denote the random weights within the measure G_{mj}, where $\mathcal{G} \sim \mathrm{MSBP}(\alpha, \beta, G_0)$. For any $N, r \ge 1$, let

$$\Gamma_{mj}(N,r) = \left(\sum_{h=N}^{\infty} \pi_{mjh}\right)^r, \Upsilon_{mj}(N,r) = \sum_{h=N}^{\infty} (\pi_{mjh})^r.$$

Then

$$\begin{aligned} E\big\{\Gamma_{mj}(N,r)\big\} &= \left\{ \sum_{k=0}^{r} C_r^k (-1)^{r-k} \mu_{r-k}(\alpha)\mu_{r-k}(\beta) \right\}^{N-1}, \\ E\big\{\Upsilon_{mj}(N,r)\big\} &= \frac{\mu_r(\alpha)\mu_r(\beta)E\big\{\Gamma_{mj}(N,r)\big\}}{1 - \sum_{k=0}^{r} C_r^k(-1)^{r-k}\mu_{r-k}(\alpha)\mu_{r-k}(\beta)}, \end{aligned}$$

where $C_r^k = \frac{r!}{(r-k)!k!}$ and $\mu_r(\lambda) = \prod_{l=1}^{r} \frac{l}{l+\lambda}$ is the rth non-central moment of beta$(1, \lambda)$, with $\mu_0(a, b) \equiv 1$.

Note that the expressions for $\mathrm{E}\big\{\Gamma_{mj}(N,r)\big\}$ and $E\big\{\Upsilon_{mj}(N,r)\big\}$ are free of m, j, so the subscripts can be excluded in discussing these expectations. An accurate truncation approximation can be produced when $\mathrm{E}\big\{\Gamma(N,r)\big\} \approx 0$ and $\mathrm{E}\big\{\Upsilon(N,r)\big\} \approx 0$ for all r. In general, these quantities decay to 0 exponentially fast with increasing N, with the rate of decay increasing as α and β decrease. For example, $\mathrm{E}\big\{\Gamma(N,1)\big\} = \left\{1 - \left(\frac{1}{1+\alpha}\right)\left(\frac{1}{1+\beta}\right)\right\}^{N-1}$. Values of α and β less than one are typically recommended in applications, as discussed in the following sections. A reasonable strategy for choosing N is to plug in an upper bound on α and β, and then choose N so that $\mathrm{E}\big\{\Gamma(N,1)\big\} = \epsilon$, for some arbitrarily small positive constant ϵ.

2.3. Clustering Properties

As discussed before, a primary motivation for the MSBP over the DP is that the event, $\theta_{mj} = \theta_{m'j}$, provides information that tasks m and m' are similar, which should lead to an increased prior probability of $\theta_{mj'} = \theta_{m'j'}$, for any $j' \neq j$. This property is apparent from Theorem 4, which also provides closed form expressions for marginal and conditional prior clustering probabilities in terms of the precision parameters, α and β.

Theorem 4. Under the $\mathcal{G} \sim \text{MSBP}(\alpha, \beta, G_0)$ prior, the probability that tasks m and m' have identical coefficients for feature j is

$$\Pr(\theta_{mj} = \theta_{m'j}) = \frac{1}{(\alpha+1)(\beta+2)-1},$$

while the corresponding conditional probability given $\theta_{mj'} = \theta_{m'j'}$ is

$$\begin{aligned} &\Pr(\theta_{mj} = \theta_{m'j} \,|\, \theta_{mj'} = \theta_{m'j'}) \\ &= \frac{2\beta_1(\alpha_2^2\beta_2 - 4)}{2\alpha_2\beta_2(\alpha_1\beta_1+\alpha)(\alpha_1\beta_1+\beta) - (\alpha_2\beta_2-2)(\alpha_2\beta_2-4)}, \end{aligned}$$

where $\alpha_1 = \alpha+1$, $\alpha_2 = \alpha+2$, $\beta_1 = \beta+1$ and $\beta_2 = \beta+2$.

In addition, we have $\Pr(\theta_{mj} = \theta_{m'j}) < \Pr(\theta_{mj} = \theta_{m'j} \,|\, \theta_{mj'} = \theta_{m'j'})$.

From the simple expression for $\Pr(\theta_{mj} = \theta_{m'j})$, it is clear that the clustering probability ranges between 0 and 1 depending on the values of α and β, converging to 1 in the limit as $\alpha, \beta \to 0$ and to 0 as either α or $\beta \to \infty$. As expected,

$$\lim_{\alpha \to 0} \Pr(\theta_{mj} = \theta_{m'j} \,|\, \theta_{mj'} = \theta_{m'j'}) = \frac{1}{\beta+1},$$

which corresponds to the clustering probability in the special case of $\theta_{mj} \overset{iid}{\sim} G_j$, for $m = 1, \ldots, M$, with $G_j \sim DP(\beta G_{0j})$, independently for $j = 1, \ldots, d$. If $\alpha \to \infty$ or $\beta \to \infty$, $\Pr(\theta_{mj} = \theta_{m'j} | \theta_{mj'} = \theta_{m'j'}) = \Pr(\theta_{mj} = \theta_{m'j}) = 0$, and none of the tasks are clustered together, so that borrowing of information relies entirely on the base parametric model.

3. Posterior Computation

For posterior computation, we propose a modification of the blocked Gibbs sampling algorithm of Ishwaran and James (2001), relying on an $\text{MSBP}_N(\alpha, \beta, G_0)$ prior and using a data augmentation scheme to facilitate efficient updating. As defined in Section 2, $R_{mj} = h$ indicates that the coefficient j for task m is allocated to the hth cluster. Here we introduce two additional latent indicator vectors. Under expressions (3) - (5), R_{mj} can be expressed as $R_{mj} = \min\{l : S_{mjl} = T_{mjl} = 1\}$, with $S_{mjl} \overset{iid}{\sim} \text{Bernoulli}(U_{mh})$ and $T_{mjl} \overset{iid}{\sim} \text{Bernoulli}(W_{jh})$, for $l = 1, \ldots, N-1$, and $S_{mjN} = T_{mjN} = 1$. After augmenting the data in this manner, it is straightforward to update each of the unknowns based on their full conditional posterior distributions. Due to space limitation we give only the conditional posterior distributions for S, T, U and W.

The conditional distribution for (S_{mjh}, T_{mjh}), for $h = 1, \ldots, R_{mj}$, sets $S_{mjh} = T_{mjh} = 1$, for $h = R_{mj}$, and otherwise samples with probabilities: $\kappa_{st} = \Pr(S_{mjh} = s, T_{mjh} = t)$, where

$$\begin{aligned} \kappa_{00} &= \frac{(1-U_{mh})(1-W_{jh})}{1-U_{mh}W_{jh}}, \\ \kappa_{10} &= \frac{U_{mh}(1-W_{jh})}{1-U_{mh}W_{jh}}, \quad \kappa_{01} = \frac{(1-U_{mh})W_{jh}}{1-U_{mh}W_{jh}}. \end{aligned}$$

The conditional distributions for U_{mh}, $m = 1, \ldots, M, h = 1, \ldots, N-1$, and $W_{jh}, j = 1, \ldots, d, h = 1, \ldots, N-1$, have the simple forms:

$$\begin{aligned} &(U_{mh} | S, \alpha) \\ &\sim \text{beta}\Big(1 + \sum_{j: R_{mj} \geq h} S_{mjh}, \alpha + \sum_{j: R_{mj} \geq h} (1 - S_{mjh})\Big), \\ &(W_{jh} | T, \beta) \\ &\sim \text{beta}\Big(1 + \sum_{m: R_{mj} \geq h} T_{mjh}, \beta + \sum_{m: R_{mj} \geq h} (1 - T_{mjh})\Big). \end{aligned}$$

We also note that the regression model which has been the principal focus of this paper can easily be adapted to the classification problem using the method of Albert and Chib (1993).

4. Simulation Example

We first considered a simple simulation example to illustrate the approach. We assumed $y_{mi} \sim N(\mathbf{x}'_{mi}\boldsymbol{\theta}_m, \phi^{-1})$, with $d = 5, n_m = 15, M = 8, \phi = 16$, and

$$\begin{array}{lcrrrrr} \boldsymbol{\theta}_1 & = & (\;1 & 1 & 1 & 1 & 1)' \\ \boldsymbol{\theta}_2 & = & (\;1 & 1 & 1 & 1 & -1)' \\ \boldsymbol{\theta}_3 & = & (\;1 & 1 & -1 & -1 & 2)' \\ \boldsymbol{\theta}_4 & = & (\;1 & 1 & -1 & -1 & -2)' \\ \boldsymbol{\theta}_5 & = & (-1 & -1 & 2 & 2 & 3)' \\ \boldsymbol{\theta}_6 & = & (-1 & -1 & 2 & 2 & -3)' \\ \boldsymbol{\theta}_7 & = & (-1 & -1 & -2 & -2 & 4)' \\ \boldsymbol{\theta}_8 & = & (-1 & -1 & -2 & -2 & -4)' \end{array} \tag{6}$$

The $MSBP_N(\alpha, \beta, G_0)$ model is implemented for this synthetic data set, with $N = 20$. The base distribution is specified as $G_{0j} \sim N(\mu_j, \psi_j^{-1})$, where $\mu_j \sim N(0, 1)$ and $\psi_j \sim Ga(1, 1)$ for all j. We place a Gamma prior $Ga(0.1, 0.1)$ on ϕ. As noted in Section 2.3, the precision parameters, α and β, control the prior distribution on the number of clusters. In order for the data to inform more strongly about the clustering, we chose $Ga(1, 1)$ hyperpriors on α and β. The hyperparameter values were chosen to favor few clusters.

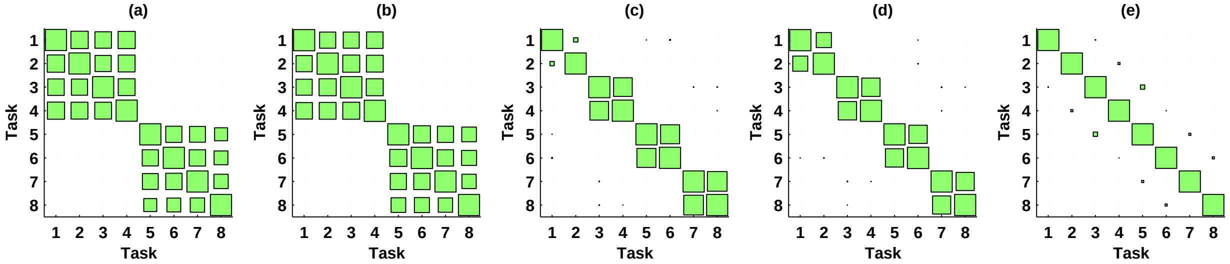


Figure 2. Pairwise posterior probabilities of two tasks being assigned to the same cluster for the simulation example analyzed using the MSBP model.

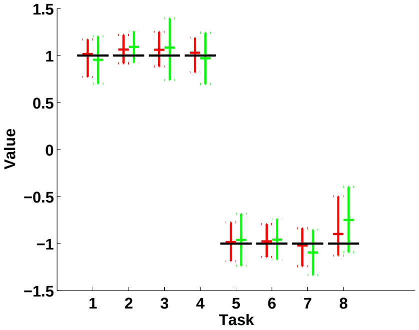


Figure 3. Posterior means and 95% credible intervals for the study-specific coefficients from the simulation example. The black solid bar indicates the true value, the red shows the estimates for the MSBP model, and the green shows the estimates based on the normal hierarchical base model.

The Gibbs sampling algorithm is used to obtain samples of the posteriors under the MSBP. The results shown below are based on $100,000$ samples obtained by thinning the MCMC chain by a factor of 20, after a burn-in period of $5,000$ iterations. Rapid convergence has been observed in the diagnostic tests as described in Geweke (1992) and Raftery and Lewis (1992). In addition, mixing was good.

Figure 2 plots the posterior probability of two tasks being assigned to the same cluster separately for each of the five feature components. The size of the shaded box is proportional to the posterior probability of pairwise clustering. It is apparent that the true clustering structure is well represented, with tasks having the same coefficient for a feature component clustered together with high probability. Figure 3 shows the task-specific posterior means and 95% credible intervals for one of the five coefficients in red, along with the true values (black) and results for a normal hierarchical analysis (green). The normal analysis corresponded to letting $\alpha \to 0$ and $\beta \to \infty$, as described in Section 2.

It is clear that the posterior densities are concentrated around the true values. In addition, the 95% credible intervals from the MSBP analysis are narrower in each case, with the difference considerable when uncertainty in the parameters is high for the base parametric model.

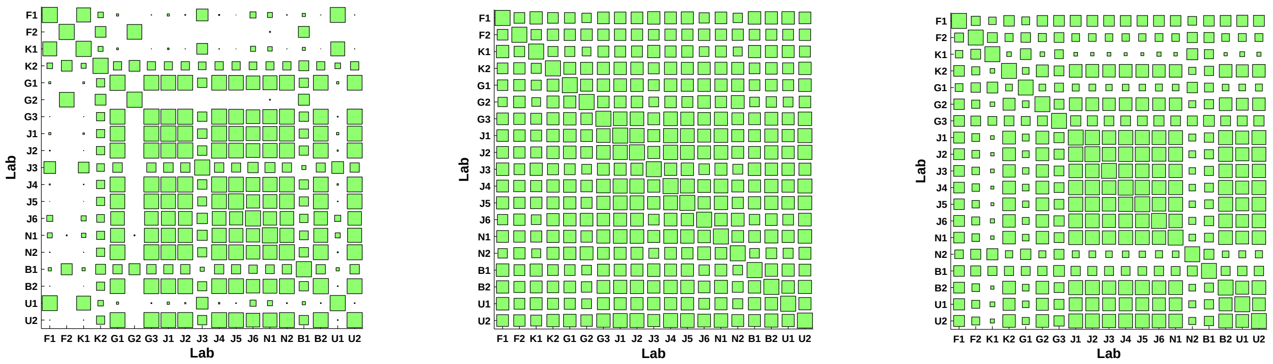
5. Multi-national Bioassay Application

We illustrate the methodology using an uterotrophic bioassay study (Kanno et al., 2001), for which data were collected for 2681 female rats from studies conducted in 19 laboratories from 8 nations. There were four protocols (A/B/C/C'), with two relying on an immature female rat model and two using an adult ovariectomized rat model. Under each protocol, there were 11 treatment groups, with 6 animals per group and the groups including an untreated control, a vehicle control, and seven dose levels of EE, with the final two groups having both EE and ZM exposure. Refer to Kanno et al. (2001) for more details.

The outcomes of the bioassay were wet and blotted rat uterus weights. To reduce measurement error, we focus here on the blotted uterus weights. The bioassay data can be modeled using $y_{mi} \sim N(\mathbf{x}'_{mi}\boldsymbol{\theta}_m, \phi^{-1})$, where y_{mi} is the log-transformed blotted uterus weight for rat i in lab m and $\mathbf{x}_{mi} = (x_{mi1}, \ldots, x_{mi6})'$, with $x_{mi1} = 1$, $x_{mi2}, x_{mi3}, x_{mi4}$ 0/1 indicators of protocol B, C, and C', respectively, x_{mi5} dose of EE, and x_{mi6} dose of ZM.

The primary focus of the study is on assessing heterogeneity among the labs in the effects of the different feature components, with a particular emphasis on assessing variability in the slopes, θ_{m5} and θ_{m6}. With this goal in mind, we repeated the analysis conducted in Section 4 for the simulation example, using the same priors and computational implementation.

For sake of brevity, we focus our discussion on the results for the intercept (θ_{m1}) and two dose effects $(\theta_{m5}, \theta_{m6})$. Figure 4 presents pairwise posterior prob-



(a) Intercept

(b) EE coefficient

(c) ZM coefficient

Figure 4. Pairwise posterior probabilities of two laboratories being assigned to the same cluster within the rat uterotrophic bioassay application based on the MSBP analysis. Results shown are for the intercept and slopes for EE and ZM.

abilities of labs being clustered together for these coefficients. Note that two labs being assigned to the same cluster implies an identical value for the regression coefficient, though this *soft* probabilistic clustering leads to posterior mean estimates that are different for the different labs. The labs are denoted as follows: F1 and F2 for two labs in France; K1 and K2 for two labs in Korea; G1, G2 and G3 for three labs in Germany; J1, . . ., J6 for six labs in Japan; N1 and N2 for two labs in Netherlands; B1 and B2 for two labs in the UK; and U1 and U2 for two labs in the USA.

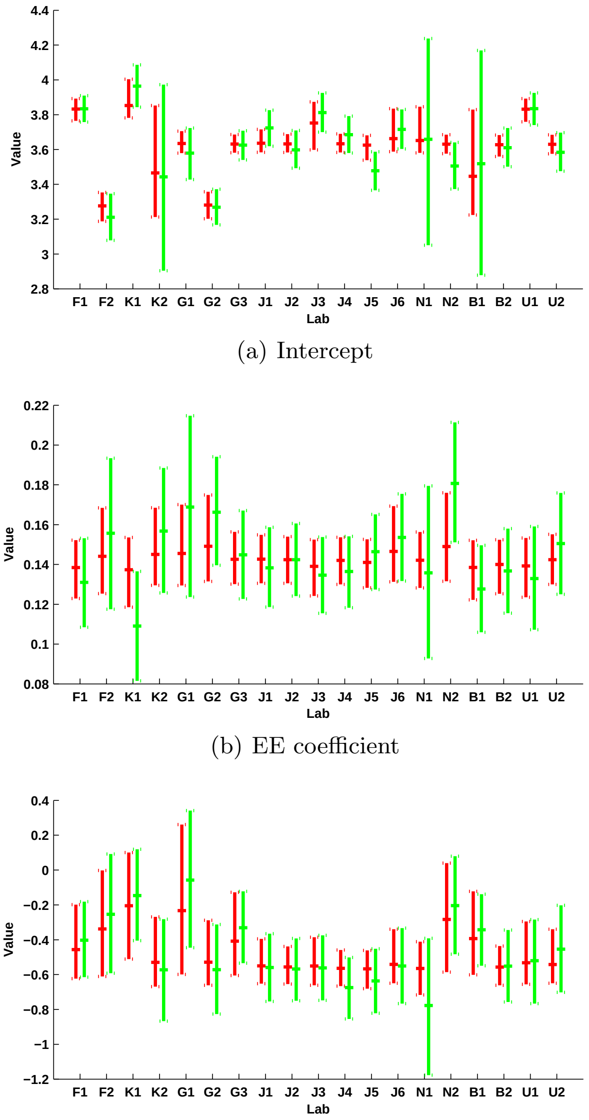
From Fig. 4, it is apparent that there is substantial evidence of heterogeneity among the labs in the intercept, as the posterior probabilities of certain labs being clustered together is small. For example, the clustering probability is close to zero for labs U1 and G2. However, there is a large group of labs, which all have a moderate to high probability of being clustered together. These results are consistent with exploratory plots of the data and with our prior knowledge that rodent body weights can vary across labs. Such variability is a well known problem in carcinogenicity studies, as body weight is an important feature of tumor response. However, in the current study, we are more interested in assessing heterogeneity in the estimated dose response across labs.

It is clear from Fig. 4 that the EE slope is consistent across labs, as each of the labs has a moderate to high posterior probability of being clustered together with any of the other labs. These results are reassuring that different labs should obtain consistent results in future uterotrophic bioassay studies seeking to identify chemicals having estrogen agonist effects. The results for the estrogen antagonist, ZM, are somewhat less consistent. Although most pairs of labs having a high posterior probability of being clustered together for the ZM coefficient, there are a couple of labs that have slight divergent results.

Figure 4 only provides pairwise probabilities of being clustered together. If two labs have a low probability of being clustered together that does not necessarily imply that the coefficients for those labs have a biologically significant difference. To assess the magnitude of the difference, we plot the lab-specific coefficients and 95% credible intervals in Fig. 5. The red lines are the results for the MSBP, while the green lines provide the results under the base parametric normal model. As in the simulation study, the normal model results tend to have wider credible intervals. From this figure, we can see that there are considerable differences in the intercepts, with labs F2 and G2 having control animals with low uterus weighs, and labs F1, K1 and U1 having unusually high uterus weights. However, the variability across labs in the EM and ZM coefficients is not biologically significant, taking into account the level of uncertainty in the lab-specific coefficients.

6. Conclusions

Motivated by the problem of flexible borrowing of information across feature components and tasks, this paper has proposed a new class of priors for a matrix of random probability measures. The proposed matrix stick-breaking process is a natural generalization of stick-breaking representations of the Dirichlet process. The MSBP should be broadly useful for borrowing information across related semiparametric models. The proposed computational implementation is efficient in cases we have considered and is no more difficult to implement than Gibbs samplers used for DP mixture



(a) Intercept

(b) EE coefficient

(c) ZM coefficient

Figure 5. Posterior means and 95% credible intervals for the lab-specific coefficients in the rat uterotrophic bioassay application. The red lines show the results for the MSBP analysis, while the green lines show the results based on the base normal hierarchical model.

models (Ishwaran & James, 2001).

References

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

De Iorio, M., Müller, P., Rosner, G., & MacEachern, S. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, *99*, 205–215.

Dunson, D., Pillai, N., & Park, J.-H. (2007a). Bayesian density regression. *Journal of the Royal Statistical Society B*, in press.

Dunson, D., Xue, Y., & Carin, L. (2007b). *The matrix stick-breaking process: Flexible Bayes meta analysis* (Technical Report 2007-03). Institute of Statistics and Decision Sciences, Duke University.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209–230.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, *4*, 169–193.

Griffin, J., & Steel, M. (2006). Order-based dependent Dirichlet process. *Journal of the American Statistical Association*, in press.

Ishwaran, H., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*, 161–173.

Kanno, J., Onyon, L., Haseman, J., Fenner-Crisp, P., Ashby, J., & Owens, W. (2001). The OECD program to validate the rat uterotrophic bioassay to screen compounds for *in vivo* estrogenic responses: Phase 1. *Environmental Health Perspectives*, *109*, 785–794.

MacEachern, S. (1999). Dependent nonparametric process. *ASA Proceeding of the Section on Bayesian Statistical Science*. Alexandria, VA.

Mukhopadhyay, S., & Gelfand, A. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, *92*, 633–639.

Müller, P., Quintana, F., & Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society Series B*, *66*, 735–749.

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler? *Bayesian Statistics*, *4*, 763–773.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566–1581.

Thrun, S., & Pratt, L. (Eds.). (1998). *Learning to learn*. Boston, MA: Kluwer Academic Publishers.

Xue, Y., Liao, X., Carin, L., & Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, *8*, 35–63.

Yu, K., Tresp, V., & Yu, S. (2004). A nonparametric hierarchical Bayesian framework for information filtering. *Proc. the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.