# Matrix-Variate Dirichlet Process Mixture Models

**Zhihua Zhang**
College of Comp. Sci. and Tech.
Zhejiang University
Zhejiang 310027, China
zhzhang@cs.zju.edu.cn

**Guang Dai**
College of Comp. Sci. and Tech.
Zhejiang University
Zhejiang 310027, China
daiguang116@gmail.com

**Michael I. Jordan**
Depts. of EECS and Statistics
University of California, Berkeley
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

## Abstract

We are concerned with a multivariate response regression problem where the interest is in considering correlations both across response variates and across response samples. In this paper we develop a new Bayesian nonparametric model for such a setting based on Dirichlet process priors. Building on an additive kernel model, we allow each sample to have its own regression matrix. Although this overcomplete representation could in principle suffer from severe overfitting problems, we are able to provide effective control over the model via a matrix-variate Dirichlet process prior on the regression matrices. Our model is able to share statistical strength among regression matrices due to the clustering property of the Dirichlet process. We make use of a Markov chain Monte Carlo algorithm for inference and prediction. Compared with other Bayesian kernel models, our model has advantages in both computational and statistical efficiency.

## 1 Introduction

In this paper we are concerned with a multivariate supervised learning problem based on a training data set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_1^n$. We in particular consider a regression problem where $\mathbf{x}_i \subset \mathbb{R}^p$ is an input vector and $\mathbf{y}_i \in \mathbb{R}^q$ is a $q$-dimensional continuous vector of responses.

In the univariate setting (i.e. $q = 1$), Gaussian processes (GPs) (Neal, 1999; Rasmussen and Williams, 2006) provide a flexible approach to regression. However, a typical treatment in the multivariate setting is to exploit several independent GPs, with one GP for each response variate. This treatment can not capture statistical relationships among the response variates. To cope with this problem, Boyle and Frean (2005) developed a so-called dependent GP model. The applications of the model are limited, however, because it is based on an $nq \times nq$ covariance matrix.

Other methods for capturing the dependency among multiple response variates include "cokriging" (Cressie, 1993), "curds and whey" (Breiman and Friedman, 1997) and semiparametric latent factor models (Teh et al., 2005). In this paper, we make use of Dirichlet processes to capture the relationship among the response variates. We also show that our approach provides leverage on problems where the data are not iid (independent and identically distributed).

Dirichlet processes (DPs) (Ferguson, 1973) or DP mixture models (Lo, 1984) are classical Bayesian nonparametric modeling tools. After Markov chain Monte Carlo (MCMC) algorithms were developed for DP mixture models in the 1990s (see, for example, (Bush and MacEachern, 1996; Escobar and West, 1995; MacEachern, 1998; Neal, 2000)), DP mixture models have seen a wide range of applications in the literature. A DP is a distribution on probability measures (i.e., it is a random measure) that yields clustering phenomena when one considers repeated draws from the random measure. This clustering property allows DPs to formalize the notion of "borrowing strength" across related studies (Antoniak, 1974; Ferguson, 1973).

In recent years, one of the most important developments in the DP literature is the dependent DP (DDP) of MacEachern (1999). The DDP is a general framework for describing dependency among a collection of random measures. This is achieved by treating the weights and the atoms in the stick-breaking representation of the DP (Sethuraman, 1994) as stochastic processes (De Iorio et al., 2004; Dunson et al., 2008; Griffin and Steel, 2006). However, this framework typi-

Appearing in Proceedings of the $13^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. 

cally leads to demanding computations because conventional ways to devise MCMC algorithms for DP mixture models based on the Pólya urn scheme (Blackwell and MacQueen, 1973) can no longer be used under the framework.

This paper is concerned with the formulation of DPs in dependent nonparametric models for multivariate-response regression problems. Our point of departure is an expansion of the regression function $f_j(\mathbf{x})$ in a series expansion using a combination of basis functions; that is,

$$f_j(\mathbf{x}) = b_{j0} + \sum_{l=1}^{n} b_{jl} K(\mathbf{x}_l, \mathbf{x}), \quad j = 1, \ldots, q$$

where the $b_{j0}$ are offset terms, the $b_{jl}$ are regression coefficients, and $K(\cdot, \cdot)$ is the kernel function.

Let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$ where $\mathbf{b}_j = (b_{j0}, b_{j1}, \ldots, b_{jn})'$ for $j = 1, \ldots, q$ be the regression matrix. In the usual setting, either the columns of $\mathbf{B}$ or the rows of $\mathbf{B}$ are assumed to be independent. This can yield a model that ignores the dependence between the response variates or between the response samples. To take an extreme alternative nonparametric approach, we might endow each sample with its own regression matrix. This could overfit, thus we envision a DP prior to provide a joint distribution on the regression matrices. The clustering property of DPs naturally allows the sharing of statistical strength between the samples and between the responses. Moreover, the clustering property is able to transfer statistical strength from existing regression matrices to new regression matrices, and thus yield out-of-sample prediction.

We refer to the resulting DP prior as a *matrix-variate DP* since it is developed for describing a set of random matrices. We employ the Pólya urn scheme for Bayesian inference. Our regression model is a conjugate model, and Bayesian inference for this model proceeds via a relatively straightforward merging of MCMC techniques (Bush and MacEachern, 1996; Escobar and West, 1995; MacEachern, 1998; Neal, 2000).

Our regression model not only captures the relationship among the response samples, but also the relationship among the response variates. The spatial DP model of Gelfand et al. (2005) is also able to model these two types of the relationships. Since the base measure in the spatial DP model is defined as a Gaussian process, this model typically requires repeatedly inverting $n \times n$ matrices, limiting their applications in large-scale datasets. However, our model can avoid this limitation.

It is worth noting that the kernel weighted mixture of DPs (Dunson et al., 2007) is to capture the relationship among the response samples, but it cannot be used to model the dependence among the response variates. Our model is also different from the method of Dunson et al. (2008) in which only one regression matrix for all samples is employed and a so-called matrix stick-breaking process is proposed to define a joint prior for the elements of this regression matrix.

To simplify our presentation, we will employ the notation of Gupta and Nagar (2000) for matrix-variate distributions. That is, for an $s \times t$ random matrix $\mathbf{Z}$, $\mathbf{Z} \sim N_{s,t}(\mathbf{M}, \mathbf{A} \otimes \mathbf{B})$ means that $\mathbf{Z}$ follows a matrix-variate normal distribution with mean matrix $\mathbf{M}$ $(s \times t)$ and covariance matrix $\mathbf{A} \otimes \mathbf{B}$, where $\mathbf{A}$ $(s \times s)$ and $\mathbf{B}$ $(t \times t)$ are positive definite. For an $s \times s$ random matrix $\mathbf{C}$, $\mathbf{C} \sim W_s(r, \mathbf{D})$ means that $\mathbf{C}$ follows a *Wishart distribution* with $r$ degrees of freedom and an $s \times s$ positive definite parameter matrix $\mathbf{D}$.

The rest of this paper is organized as follows. Section 2 presents a Bayesian nonparametric regression model based on the matrix-variate DP mixture prior. An experimental analysis is presented in Section 3 and we summarize in Section 4.

## 2 Matrix-variate DP mixture priors for Multivariate Regression

We consider the following regression model

$$\mathbf{y}_i = \mathbf{B}_i' \mathbf{g}_i + \epsilon_i, \tag{1}$$

where $\mathbf{g}_i = (1, K(\mathbf{x}_1, \mathbf{x}_i), \ldots, K(\mathbf{x}_n, \mathbf{x}_i))'$ $((n+1) \times 1)$ for short and $\mathbf{B}_i$ $((n+1) \times q)$ is the regression matrix corresponding to $\mathbf{x}_i$. Unlike a conventional regression model, the current model allows each input sample $\mathbf{x}_i$ to have its own regression matrix $\mathbf{B}_i$.

### 2.1 Matrix-variate DP Priors

To capture relationships among the $\mathbf{B}_i$, we introduce a DP prior to model the joint distribution of the $\mathbf{B}_i$. In particular, we assume that $p(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ follows a DP mixture model:

$$\begin{aligned} [\mathbf{y}_i | \mathbf{B}_i, \mathbf{\Sigma}] &\overset{ind}{\sim} N_q(\mathbf{y}_i | \mathbf{B}_i' \mathbf{g}_i,\ \tau \mathbf{\Sigma}), \quad i = 1, \ldots, n; \\ [\mathbf{B}_i | G] &\overset{iid}{\sim} G, \quad i = 1, \ldots, n; \\ G &\sim \mathrm{DP}(\nu G_0). \end{aligned} \tag{2}$$

Here $\mathbf{\Sigma}$ is a $q \times q$ positive definite matrix, $\nu > 0$ is the *concentration parameter* of the DP prior and $G_0$ is the *base distribution*. In this paper, we define $G_0$ as

$$G_0(\cdot | \mathbf{\Sigma}, \mathbf{\Lambda}) = N_{n+1,q}(\mathbf{0},\ \mathbf{\Lambda} \otimes \mathbf{\Sigma}),$$

where $\mathbf{0}$ represents the zero vector (or matrix) whose dimensionality is dependent upon the context and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n+1})$ is a diagonal matrix with $\lambda_i > 0$

for $i = 1, \ldots, n+1$. Since the base measure follows a matrix-variate distribution, we refer to the resulting DP as a *matrix-variate DP*.

In addition, we assume that $\nu$, $\tau^{-1}$ and $\lambda_i^{-1}$ follow Gamma distributions: $Ga(\nu|a_\nu,\ b_\nu)$, $Ga(\tau^{-1}|\frac{a_\tau}{2},\ \frac{b_\tau}{2})$ and $Ga(\lambda_i^{-1}|\frac{a_i}{2},\ \frac{b_i}{2})$; and we assume that $\mathbf{\Sigma}^{-1}$ follows Wishart distribution: $W_q(\mathbf{\Sigma}^{-1}|\rho,\ \mathbf{R})$.

As was showed by Blackwell and MacQueen (1973), integrating over $G$ results in a Pólya urn scheme for the $\mathbf{B}_i$; that is,

$$\mathbf{B}_1 \sim G_0,$$
$$[\mathbf{B}_i|\mathbf{B}_1, \ldots, \mathbf{B}_{i-1}] \sim \frac{\nu G_0 + \sum_{l=1}^{i-1} \delta(\mathbf{B}_i|\mathbf{B}_l)}{\nu + i - 1},$$

where $\delta(\mathbf{B}_i|\mathbf{B}_l)$ is a point mass at $\mathbf{B}_l$. It is easily seen that as $\nu \to 0$, all the $\mathbf{B}_i$ are identical to $\mathbf{B}_1$, which follows $G_0$. When $\nu \to \infty$, the $\mathbf{B}_i$ become iid $G_0$. Since the $\mathbf{B}_i$ are exchangeable, the Pólya urn scheme can be written as

$$[\mathbf{B}_i|\mathbf{B}_{-i}] \sim \frac{\nu N_{n+1,q}\big(\mathbf{B}_i|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}\big) + \sum_{l\neq i} \delta(\mathbf{B}_i|\mathbf{B}_l)}{\nu + n - 1}, \tag{3}$$

where $\mathbf{B}_{-i}$ represents $\{\mathbf{B}_l : l \neq i\}$.

### 2.2 Posterior Inference

The discreteness of the random distribution $G$ plays a central role in Bayesian inference and computation, because with positive probability, some of the $\mathbf{B}_i$ are identical. This is the well known clustering property of the DP. Assume that there are $c$ distinct values among the $\mathbf{B}_i$ as $\mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_c\}$, and that there are $n_k$ occurrences of $\mathbf{Q}_k$ such that $\sum_{k=1}^c n_k = n$. The vector of configuration indicators $\mathbf{w} = (w_1, \ldots, w_n)$ is defined by $w_i = k$ if and only if $\mathbf{B}_i = \mathbf{Q}_k$ for $i = 1, \ldots, n$. Thus $(\mathcal{Q}, \mathbf{w})$ is an equivalent representation of the $\mathbf{B}_i$, and hence (3) reduces to

$$[\mathbf{B}_i|\mathbf{B}_{-i}] \tag{4}$$
$$\sim \frac{\nu N_{n+1,q}\big(\mathbf{B}_i|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}\big) + \sum_{k=1}^c n_{k(-i)}\delta(\mathbf{B}_i|\mathbf{Q}_k)}{\nu + n - 1},$$

where $n_{k(-i)}$ refers to the cardinality of cluster $k$ with $\mathbf{B}_i$ removed, and

$$\mathbf{Q}_k \overset{iid}{\sim} N_{n+1,q}\big(\mathbf{Q}_k|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}\big),\ k = 1, \ldots, c.$$

Hence, we can express the joint distribution of $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]'$ $(n\times q)$ as

$$[\mathbf{Y}|\mathbf{w}, \mathcal{Q}, \tau] \sim \prod_{k=1}^c \prod_{i:\ w_i=k} N_q(\mathbf{y}_i|\mathbf{Q}_k'\mathbf{g}_i,\ \tau\mathbf{\Sigma}).$$

Integrating out the $\mathbf{Q}_k$ yields the marginal distribution of $\mathbf{Y}$ as

$$[\mathbf{Y}|\mathbf{w}, \tau, \mathbf{\Lambda}, \mathbf{\Sigma}] \sim \prod_{k=1}^c N_{n_k,q}\big(\mathbf{Y}_k|\mathbf{0},\ (\tau\mathbf{I}_{n_k} + \mathbf{G}_k\mathbf{\Lambda}\mathbf{G}_k')\otimes\mathbf{\Sigma}\big), \tag{5}$$

where $\mathbf{Y}_k$ and $\mathbf{G}_k$ are respectively $n_k\times q$ and $n_k\times(n+1)$ matrices consisting of those $\mathbf{y}_i$ and $\mathbf{g}_i$ with $w_i = k$. For each $k = 1, \ldots, c$, we have

$$[\mathbf{Q}_k|\mathbf{Y}, \mathbf{w}, \mathbf{\Lambda}, \mathbf{\Sigma}, \tau] \sim N_{n+1,q}\big(\mathbf{Q}_k|\mathbf{\Theta}_k\mathbf{G}_k'\mathbf{Y}_k,\ \tau\mathbf{\Theta}_k\otimes\mathbf{\Sigma}\big) \tag{6}$$

where $\mathbf{\Theta}_k = (\tau\mathbf{\Lambda}^{-1} + \mathbf{G}_k'\mathbf{G}_k)^{-1}$.

Posterior inference is achieved by generating realizations of the parameters from the conditional joint density $[\mathbf{B}, \tau, \mathbf{\Lambda}|\mathbf{Y}]$. We use Gibbs sampler, which consists of the following steps (see the Appendix for a detailed presentation):

(a) Update $(\mathbf{B}_i, w_i)$ from $[(\mathbf{B}_i, w_i)|(\mathbf{B}_{-i}, \mathbf{w}_{-i}), \mathbf{\Lambda}, \tau, \mathbf{\Sigma}, \mathbf{Y}]$ for $i = 1, \ldots, n$;

(b) Update $\mathbf{Q}_k$ from $[\mathbf{Q}_k|\mathbf{w}, \mathbf{\Lambda}, \nu, \tau, \mathbf{Y}]$ for $k = 1, \ldots, c$;

(c) Update $\tau^{-1}$, $\mathbf{\Sigma}^{-1}$ and $\lambda_i^{-1}$ $(i = 1, \ldots, n+1)$ from $[\tau^{-1}|\mathbf{Y}, \mathbf{B}, \mathbf{\Sigma}, a_\tau, b_\tau]$, $[\mathbf{\Sigma}^{-1}|\mathbf{Y}, \mathbf{B}, \tau, \mathbf{R}, \rho]$ and $[\lambda_i^{-1}|\{\boldsymbol{\beta}_i^{(k)}\}_{k=1}^c, \mathbf{\Sigma}, a_i, b_i]$ where $\boldsymbol{\beta}_i^{(k)}$ is the $i$th row of $\mathbf{Q}_k$;

(d) Update $\nu$ from $p(\nu|a_\nu, b_\nu, c)$.

Our method groups the regression matrices $\mathbf{B}_i$ into $c$ clusters by using the matrix-variate DP prior. The main computational burden of our method comes from the calculation of $\mathbf{\Theta}_k$, but fortunately we can use the Sherman-Morrison-Woodbury formula (Golub and Loan, 1996) to calculate $\mathbf{\Theta}_k$ efficiently. In particular, we have

$$\begin{aligned}\mathbf{\Theta}_k &= (\tau\mathbf{\Lambda}^{-1} + \mathbf{G}_k'\mathbf{G}_k)^{-1} \\ &= \tau^{-1}\mathbf{\Lambda} - \tau^{-1}\mathbf{\Lambda}\mathbf{G}_k'(\tau\mathbf{I}_{n_k} + \mathbf{G}_k\mathbf{\Lambda}\mathbf{G}_k')^{-1}\mathbf{G}_k\mathbf{\Lambda}.\end{aligned}$$

Thus, the formula allows us to invert an $n_k\times n_k$ matrix instead of an $n\times n$ matrix. Since $n_k$ is typically far smaller than $n$, the algorithm is still efficient for a large-scale dataset.

### 2.3 Prediction

Given a new input vector $\mathbf{x}_0$, we wish to predict the corresponding response $\mathbf{y}_0$. Let $\mathbf{B}_0$ be the associated regression matrix. Prediction in our model is based on the cluster structure of the $\mathbf{B}_i$.

In this paper we are interested in Bayesian prediction. Using the structure of the DP prior, we have

$$[\mathbf{B}_0|\mathcal{Q},\mathbf{w},\nu,\mathbf{\Lambda}] \tag{7}$$
$$\sim \frac{\nu}{\nu+n}N_{n+1,q}(\mathbf{B}_0|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}) + \frac{1}{\nu+n}\sum_{k=1}^{c} n_k\delta(\mathbf{B}_0|\mathbf{Q}_k).$$

Let $\{\mathcal{Q}^{(t)},\mathbf{\Sigma}^{(t)},\tau^{(t)},\mathbf{\Lambda}^{(t)},\nu^{(t)}\}$, $t=1,\ldots,T$ be the MCMC realizations of the parameters after the burn-in. We present two Bayesian approaches. The first approach is to draw $\mathbf{B}_0^{(t)}$ from (7) with the parameter realizations. We thus have $\hat{\mathbf{B}}_0 = \frac{1}{T}\sum_{t=1}^{T}\mathbf{B}_0^{(t)}$, and hence $\hat{\mathbf{y}}_0 = \hat{\mathbf{B}}_0'\mathbf{g}_0$.

The second approach is based on the posterior distribution of $y_0$, which is given by

$$p(\mathbf{y}_0|\mathbf{x}_0,\mathbf{Y}) = \int p(\mathbf{y}_0|\mathbf{x}_0,\mathbf{B}_0,\mathbf{Y})p(\mathbf{B}_0|\mathcal{Q},\mathbf{w})p(\mathcal{Q}|\mathbf{Y},\mathbf{w})d\mathcal{Q}d\mathbf{B}_0.$$

Integrating over $\mathbf{B}_0$, we have

$$\begin{aligned}\hat{p}(\mathbf{y}_0|\mathbf{x}_0,\mathbf{Y}) &= \frac{1}{(\nu+n)T}\sum_{t=1}^{T}\Bigg\{\sum_{k=1}^{c} n_k p\big(y_0|\mathbf{x}_0,\mathbf{Q}_k^{(t)},\tau^{(t)}\mathbf{\Sigma}^{(t)}\big) \\ &\quad +\nu\int p(\mathbf{y}_0|\mathbf{x}_0,\mathbf{B}_0)N_{n+1,q}\big(\mathbf{B}_0|\mathbf{0},\mathbf{\Lambda}^{(t)}\otimes\mathbf{\Sigma}^{(t)}\big)d\mathbf{B}_0\Bigg\} \\ &= \frac{1}{(\nu+n)T}\sum_{t=1}^{T}\Bigg\{\sum_{k=1}^{c} n_k N_q\big(\mathbf{y}_0|(\mathbf{Q}_k^{(t)})'\mathbf{g}_0,\tau^{(t)}\mathbf{\Sigma}^{(t)}\big) \\ &\quad +\nu N_q\big(\mathbf{y}_0|\mathbf{0},\ (\tau^{(t)}+\mathbf{g}_0'\mathbf{\Lambda}^{(t)}\mathbf{g}_0)\mathbf{\Sigma}^{(t)}\big)\Bigg\}.\end{aligned}$$

Thus, we obtain the following prediction of $\mathbf{y}_0$:

$$\hat{\mathbf{y}}_0 = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_0^{(t)}, \tag{8}$$

where $\mathbf{y}_0^{(t)}$ are the MCMC realizations of $\mathbf{y}_0$ from $\hat{p}(\mathbf{y}_0|\mathbf{x}_0,\mathbf{Y})$ after the burn-in.

## 3 Experimental Analysis

In this section we conducted numerical experiments to analyze the performance of our proposed Bayesian regression based on the matrix-variate Dirichlet process (MDP) mixture model.

Our analysis was implemented on the four datasets: `chemometrics`, `Boston housing`, `forest fires`, `automobile`, and `robot arm`. The `chemometrics` data taken from Skagerberg et al. (1992) were used in Breiman and Friedman (1997) to analyze their regression methods, and it contains 56 samples. The `robot arm` dataset was used by Teh et al. (2005) for modeling the domain of multi-joint robot arm dynamics. The `Boston housing`, `forest fires`, and `automobile` datasets were taken from UCI, and they are available from `http://archive.ics.uci.edu/ml/datasets.html`. The Table 1 summarized these benchmark datasets.

In our experiments, for all the four datasets the input samples were standardized to have zero mean and unity variance. Moreover, each dataset was randomly partitioned into two disjoint subsets as the training and test datasets, according to percentages listed in the last column of Table 1. In addition, each sample in all the datasets was divided into ($q=$)6 responses and $p$ input variables. Note that the `Boston housing` data are typically used for univariate response regression problems. Here we formulated this dataset as a 6-response regression problem for our purpose.

We compared our method with the Gaussian process-based regression (GPR) (Rasmussen and Williams, 2006), the support vector regression (SVR) (Schölkopf and Smola, 2002), the independent DP (iDP) mixture-based regression, and the spatial DP (sDP) mixture-based regression (Gelfand et al., 2005). It should be mentioned that: (1) the iDP-based regression is to model the columns of $\mathbf{Y}=[y_{ij}]$ ($n\times q$) as $q$ mutually independent DP mixture models, so it does not consider correlations between the response variates; (2) the sDP mixture model is a specification of nonparametric dependent modeling that can effectively take advantage of the correlations among the response variates.

For the sake of simplicity, we used the RBF Gaussian kernel function with a single scale parameter, i.e., $K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/2\theta^2)$. Although in principle we can estimate the scale parameter $\theta^2$ using MCMC algorithms (see, e.g., (Gelfand et al., 2005; Zhang and Jordan, 2006)), we simply specified $\theta^2$ as the mean of Euclidean distances among the training dataset. In general, such a choice was empirically verified to be effective. For those methods based on the MCMC, we ran each MCMC algorithm for $5,000$ sweeps, discarding the first $1,000$ as the burn-in, and retaining every fifth realization of the parameters after the burn-in for inference and prediction. The hyperparameters were set as follows: $a_\nu=10$, $b_\nu=1$, $a_\tau=10$, $b_\tau=1$, $a_i=40$, $b_i=1$, $\rho=q+1$, and $\mathbf{R}=\mathbf{I}_q+\frac{1}{q}\mathbf{1}_q\mathbf{1}_q'$, where $\mathbf{I}_q$ is the $q\times q$ identity matrix and $\mathbf{1}_q$ is the $q\times 1$ vector of ones.

The regression performance is measured in terms of two quantities: the root-mean-square error (RMSE)

and the mean absolute percentage error (MAPE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_j(\mathbf{x}_i) - \tilde{y}_j(\mathbf{x}_i))^2}$$

and

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_j(\mathbf{x}_i) - \tilde{y}_j(\mathbf{x}_i)}{y_j(\mathbf{x}_i)}\right|,$$

where $m$ is the number of the test data, $y_j(\mathbf{x}_i)$ is the $j$th response corresponding to input vector $\mathbf{x}_i$, and $\tilde{y}_j(\mathbf{x}_i)$ is the $j$th prediction computed from a regression method. It is well known that RMSE is a measure of the absolute deviation of the estimated quantity from the actual quantity, giving more weight to large errors, whereas MAPE is usually used to compare the accuracy of the model on different series.

Tables 2 and 3 show the estimated prediction errors (RMSE and MAPE) for the four datasets. Note the average error listed in the last column of Tables 2 and 3 is calculated according to the definitions of RMSE and MAPE on all responses. As we can see, the regression methods based on both the MDP and sDP mixture models tend to out-perform the other three methods, presumably because the MDP and sDP mixture models can take advantage of the correlations between the responses.

The computation of the sDP mixture model is demanding because its MCMC algorithm involves the computation of $n\times n$ matrices at each sweep. In addition, this algorithm needs to calculate the densities of $n$-variate normal distributions. In the experiments, we found that the ratios (say, $r$) between some of these values become very large. This results in a slowly mixing Markov chain. To alleviate this problem, we applied a simple truncation trick; namely, $r$ is set to 0.001 if $r < 0.001$ and set to 1000 if $r > 1000$. As discussed in Section 2.2, the MDP and iDP mixture models are efficient computationally. Moreover, their MCMC algorithms only involve calculating the densities of $q$-variate or univariate normal distributions (see Section 2.2). Thus, they work very well without the need for the truncation trick.

Finally, we report the posterior distribution of the number $c$ of clusters for our MDP mixture model in Figure 1. It is worth noting that the sDP model captures the correlation between $q$ GPs by using the clustering property of DPs. In our regression problems, $q$ (= 6) takes a small value, thus the $q$ GPs were very often clustered into one cluster during the MCMC sweeps. It seems somewhat strong for sDP to borrow the strength across the response variates. However, our model captures the correlation between the response variates in terms of a full covariance matrix $\mathbf{R}$, which was defined as an equicorrelation matrix.

Table 1: Summary of the benchmark datasets: $p-$the dimension of the input vector; $q-$the dimension of the output vector; $k-$the size of the dataset; $n-$the number of the training data.

| Dataset | $p$ | $q$ | $k$ | $n/k$ |
|---|---|---|---|---|
| Chemometrics | 22 | 6 | 56 | 60% |
| Boston housing | 8 | 6 | 506 | 60% |
| Forest fires | 7 | 6 | 517 | 30% |
| Automobile | 20 | 6 | 205 | 50% |
| Robot arm | 12 | 6 | 1500 | 60% |

## 4 Conclusion

We have derived a new Bayesian nonparametric kernel regression method based on the matrix-variate Dirichlet process mixture prior and introduced an MCMC algorithm for inference and prediction. Possible extensions of the approach would involve using other exponential family models (Ibrahim and Kleinman, 1998; Xue et al., 2007). For example, we can develop extensions to multi-class classification problems by mixing the matrix-variate DP with a generalized additive model.

## Appendix: The MCMC Algorithm

We can use Gibbs sampling to draw $[\mathbf{B}, \tau, \mathbf{\Sigma}, \mathbf{\Lambda}, \nu|\mathbf{Y}]$. The required full conditionals are

(a) $[(\mathbf{B}_i, w_i)|(\mathbf{B}_{-i}, \mathbf{w}_{-i}), \nu, \mathbf{\Lambda}, \mathbf{\Sigma}, \mathbf{Y}]$ for $i = 1, \ldots, n$;

(b) $[\mathbf{Q}_k|\mathbf{w}, \mathbf{\Lambda}, \tau, \mathbf{\Sigma}, \nu, \mathbf{Y}]$ for $k = 1, \ldots, c$;

(c) $[\tau^{-1}|\mathbf{Y}, \mathbf{B}, \mathbf{\Sigma}, a_\tau, b_\tau]$;

(d) $[\mathbf{\Sigma}^{-1}|\mathbf{Y}, \mathbf{B}, \tau, \mathbf{R}, \rho]$;

(e) $[\lambda_i^{-1}|\{\boldsymbol{\beta}_i^{(k)}\}_{k=1}^c, \mathbf{\Sigma}, a_i, b_i]$ for $i = 1, \ldots, n{+}1$;

(f) $[\nu|a_\nu, b_\nu, c]$.

The Gibbs sampler exploits the simple structure of the conditional posterior for each $\mathbf{B}_i$. That is, for $i = 1, \ldots, n$, the conditional distribution is given by

$$[\mathbf{B}_i|\mathbf{B}_{-i},\ \mathbf{Y}, \mathbf{\Lambda}, \mathbf{\Sigma}, \tau] \tag{9}$$
$$\propto q_0 N(\mathbf{y}_i|\mathbf{B}_i'\mathbf{g}_i,\ \tau\mathbf{\Sigma}) N_{n+1}(\mathbf{B}_i|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}) + \sum_{j\neq i} q_j \delta(\mathbf{B}_i|\mathbf{B}_j),$$

where $q_j = N_q(\mathbf{y}_i|\mathbf{B}_j'\mathbf{g}_i,\ \tau\mathbf{\Sigma})$ and

$$q_0 = \nu \int N_q(\mathbf{y}_i|\mathbf{B}_i'\mathbf{g}_i,\ \tau\mathbf{\Sigma}) N_{n+1,q}(\mathbf{B}_i|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}) d\mathbf{B}_i$$
$$= \nu N_q(\mathbf{y}_i|\mathbf{0},\ (\mathbf{g}_i'\mathbf{\Lambda}\mathbf{g}_i + \tau)\mathbf{\Sigma}).$$

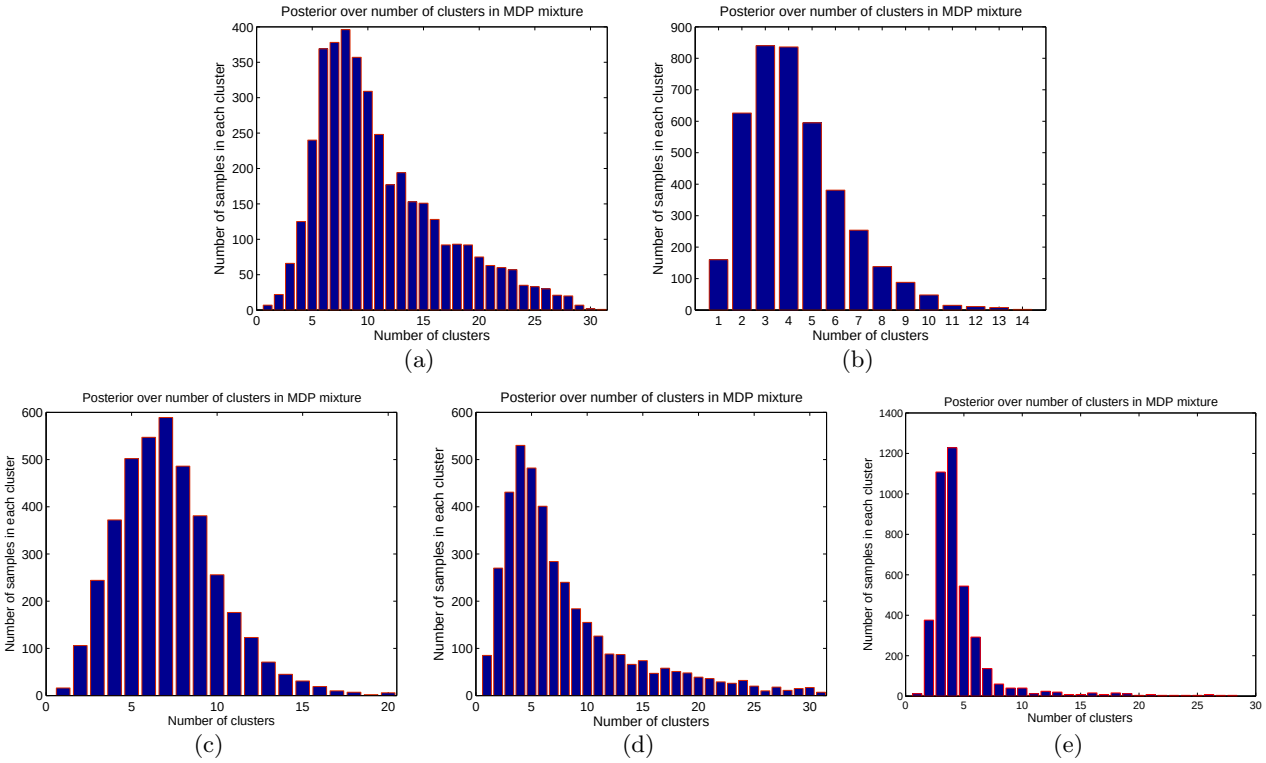


Figure 1: Posterior distribution of the number $c$ of clusters, visited in the $4,000$ sweeps after the burn-in, for the different datasets: (a) `chemometrics`; (b) `Boston housing`; (c) `forest fires`; (d) `automobile`; (e) `robot arm`.

Table 2: Evaluation results on the root-mean-square error (RMSE) for all datasets.

| Dataset | Method | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | Ave. |
|---|---|---|---|---|---|---|---|---|
| Chemometrics | GP | 0.2087 | 0.1252 | 0.6296 | 0.3581 | 0.5099 | 0.3892 | 0.4073 |
| | SVR | **0.1970** | **0.1218** | 0.5755 | 0.3568 | 0.5000 | 0.3827 | 0.3891 |
| | iDPs | 0.4146 | 0.3900 | **0.2722** | 0.4294 | 0.3154 | 0.4484 | 0.3836 |
| | sDP | 0.2668 | 0.2719 | 0.3631 | 0.2730 | **0.2685** | 0.3113 | 0.2924 |
| | MDP | 0.2015 | 0.2553 | 0.2727 | **0.2396** | 0.2818 | **0.2945** | **0.2576** |
| Boston housing | GP | 2.2907 | 2.5563 | 2.5267 | 4.3983 | 1.7337 | 3.1622 | 2.9016 |
| | SVR | 0.9505 | 0.9725 | 1.0298 | 2.1701 | 1.0999 | 1.2499 | 1.3160 |
| | iDPs | 0.8448 | 0.9062 | 0.6672 | 0.9945 | 0.6493 | 0.8097 | 0.8212 |
| | sDP | **0.8304** | **0.5386** | 0.5698 | 1.0216 | **0.3963** | 0.5978 | 0.6591 |
| | MDP | 0.8332 | 0.5544 | **0.5440** | **0.9900** | 0.4026 | **0.5787** | **0.6505** |
| Forest fires | GP | 3.8510 | 3.8755 | 0.9889 | 2.5171 | 1.0531 | 3.4328 | 2.8884 |
| | SVR | 2.6902 | 2.6721 | 0.9924 | 2.2670 | 1.0162 | 1.6733 | 2.0139 |
| | iDPs | 1.2047 | 1.2268 | 3.5418 | 1.1959 | 3.8747 | 3.0723 | 2.6263 |
| | sDP | 1.1166 | 1.1087 | 1.6035 | 1.0266 | 1.4227 | 1.6172 | 1.3379 |
| | MDP | **1.0529** | **1.0451** | **0.6632** | **1.0013** | **0.5007** | **0.7364** | **0.8599** |
| Automobile | GP | **0.2355** | 0.4754 | 0.8225 | 0.6811 | 0.6087 | 1.0763 | 0.7013 |
| | SVR | 0.2515 | **0.4128** | 0.7915 | 1.0152 | 0.9969 | 0.6204 | 0.7382 |
| | iDPs | 0.6414 | 0.7566 | 0.7481 | 0.3740 | 0.6085 | 0.8037 | 0.6727 |
| | sDP | 0.2487 | 0.4813 | 0.7052 | 0.3637 | **0.4284** | 0.4593 | 0.4686 |
| | MDP | 0.2542 | 0.4357 | **0.5877** | **0.3592** | 0.4336 | **0.3914** | **0.4223** |
| Robot arm | GP | 0.4117 | 0.5109 | 0.3717 | 0.4958 | 0.5189 | 0.4191 | 0.4590 |
| | SVR | 0.4200 | 0.4933 | 0.3643 | 0.4668 | 0.4986 | 0.4053 | 0.4441 |
| | iDPs | 0.4322 | 0.4786 | 0.4041 | 0.4369 | 0.4555 | 0.3637 | 0.4301 |
| | sDP | 0.4249 | 0.4640 | 0.3729 | **0.4296** | **0.4400** | 0.3765 | 0.4193 |
| | MDP | **0.4083** | **0.4413** | **0.3631** | 0.4429 | 0.4432 | **0.3701** | **0.4129** |

Table 3: Evaluation results (%) on the mean absolute percentage error (MAPE) for all datasets.

| Dataset | Method | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | Ave. |
|---|---|---|---|---|---|---|---|---|
| Chemomtrics | GP | 0.1569 | 0.0941 | 0.4685 | 0.2720 | 0.3830 | 0.2915 | 0.2777 |
| | SVR | **0.1470** | **0.0915** | 0.4392 | 0.2697 | 0.3767 | 0.2872 | 0.2686 |
| | iDPs | 0.3079 | 0.2525 | **0.2065** | 0.2573 | 0.2142 | 0.3082 | 0.2578 |
| | sDP | 0.1991 | 0.1928 | 0.2727 | 0.2076 | **0.2114** | 0.2499 | 0.2223 |
| | MDP | 0.1605 | 0.2017 | 0.2188 | **0.1845** | 0.2298 | **0.2436** | **0.2065** |
| Boston housing | GP | 1.0661 | 1.4774 | 1.4978 | 2.6362 | 1.0269 | 1.5832 | 1.5479 |
| | SVR | 0.3170 | 0.4944 | 0.5117 | 0.9246 | 0.5732 | 0.7304 | 0.5919 |
| | iDPs | 0.3569 | 0.6095 | 0.5169 | 0.5498 | 0.4728 | 0.5433 | 0.5082 |
| | sDP | 0.2838 | **0.3344** | 0.3877 | 0.5826 | 0.2996 | **0.3858** | 0.3790 |
| | MDP | **0.2456** | 0.3389 | **0.3796** | **0.5497** | **0.2906** | 0.3994 | **0.3673** |
| Forest fires | GP | 2.9336 | 2.9121 | 0.6710 | 1.8789 | 0.5971 | 2.4587 | 1.9086 |
| | SVR | 1.7718 | 1.7424 | **0.4662** | 1.4312 | 0.5025 | 0.9875 | 1.1503 |
| | iDPs | 1.0000 | 0.9587 | 2.8157 | 1.0172 | 3.0546 | 2.4526 | 1.8831 |
| | sDP | 0.9425 | 0.8676 | 1.4386 | 0.8776 | 1.1048 | 1.4033 | 1.1057 |
| | MDP | **0.9210** | **0.7948** | 0.5087 | **0.8798** | **0.2969** | **0.6201** | **0.6702** |
| Automobile | GP | 0.1204 | 0.2396 | 0.5144 | 0.4521 | 0.3832 | 0.6592 | 0.3948 |
| | SVR | **0.1128** | **0.2132** | 0.5811 | 0.6895 | 0.7213 | 0.3943 | 0.4521 |
| | iDPs | 0.4949 | 0.6072 | 0.6300 | 0.2760 | 0.5323 | 0.3350 | 0.4792 |
| | sDP | 0.1409 | 0.2948 | 0.5473 | 0.2649 | **0.3056** | 0.3012 | 0.3091 |
| | MDP | 0.1487 | 0.2778 | **0.4256** | **0.2583** | 0.3161 | **0.2506** | **0.2795** |
| Robot arm | GP | 0.2186 | 0.2441 | 0.2036 | 0.2120 | 0.2527 | 0.2074 | 0.2231 |
| | SVR | **0.2133** | 0.2362 | **0.2001** | 0.2037 | 0.2436 | 0.1983 | 0.2159 |
| | iDPs | 0.2208 | 0.2593 | 0.2318 | 0.2192 | 0.2553 | 0.1989 | 0.2309 |
| | sDP | 0.2152 | 0.2494 | 0.2021 | **0.2001** | 0.2426 | **0.1829** | **0.2154** |
| | MDP | 0.2195 | **0.2326** | 0.2076 | 0.2120 | **0.2367** | 0.1946 | 0.2172 |

According to (4), (9) thus reduces to

$$[\mathbf{B}_i|\mathbf{B}_{-i},\mathbf{y}_i,\mathbf{\Lambda},\mathbf{\Sigma},\tau] \propto q_0 N_{n+1,q}(\mathbf{B}_i|\mathbf{A}_i\mathbf{g}_i\mathbf{y}_i',\ \tau\mathbf{A}_i\otimes\mathbf{\Sigma}) + \sum_{k=1}^{c} n_{k(-i)}q_k\delta(\mathbf{B}_i|\mathbf{Q}_k),$$

where $\mathbf{A}_i=(\tau\mathbf{\Lambda}^{-1}+\mathbf{g}_i\mathbf{g}_i')^{-1}$. Thus, given $\mathbf{B}_{-i}$, with probability proportional to $n_{k(-i)}q_k$, we draw $\mathbf{B}_i$ from distribution $\delta(\cdot|\mathbf{Q}_k)$, or with probability proportional to $q_0$, we draw $\mathbf{B}_i$ from $N_{n+1,n}(\cdot|\mathbf{A}_i\mathbf{g}_i\mathbf{y}_i',\ \tau\mathbf{A}_i\otimes\mathbf{\Sigma})$. Here we again use the Sherman-Morrison-Woodbury formula to calculate $\mathbf{A}_i$. That is,

$$\mathbf{A}_i=(\tau\mathbf{\Lambda}^{-1}+\mathbf{g}_i\mathbf{g}_i')^{-1}=\tau^{-1}\mathbf{\Lambda}-\tau^{-1}\mathbf{\Lambda}\mathbf{g}_i(\tau+\mathbf{g}_i'\mathbf{\Lambda}\mathbf{g}_i)^{-1}\mathbf{g}_i'\mathbf{\Lambda},$$

which involves reciprocal computations.

To speed mixing of the Markov chain, Bush and MacEachern (1996) suggested resampling the $\mathbf{Q}_k$ after every step. For each $k=1,\ldots,c$, we have

$$[\mathbf{Q}_k|\mathbf{Y},\mathbf{w},\tau,\mathbf{\Lambda},\mathbf{\Sigma}] \propto N_{n+1,q}(\mathbf{Q}_k|\mathbf{0},\ \mathbf{\Lambda}\otimes\mathbf{\Sigma}) \prod_{i:\ w_i=k} N_q\big(\mathbf{y}_i|\mathbf{Q}_k'\mathbf{g}_i,\ \tau\mathbf{\Sigma}\big),$$

from which it follows that the conditional density of $\mathbf{Q}_k$ is given by (6).

Given the prior of $\tau^{-1}$, we then obtain the update of $\tau^{-1}$ as

$$[\tau^{-1}|\mathbf{Y},\mathbf{B},\mathbf{\Sigma},a_\tau,b_\tau] \sim Ga\Big(\tau^{-1}\Big|\frac{a_\tau+nq}{2},\frac{b_\tau+\sum_{i=1}^n(\mathbf{y}_i-\mathbf{B}_i'\mathbf{g}_i)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{B}_i'\mathbf{g}_i)}{2}\Big).$$

The update of $\mathbf{\Sigma}$ is given by

$$[\mathbf{\Sigma}^{-1}|\mathbf{Y},\mathbf{B},\tau,\rho,\mathbf{R}] \sim W_q\Big(\mathbf{\Sigma}^{-1}\Big|\rho+n,\mathbf{R}+\tau^{-1}\sum_{i=1}^n(\mathbf{y}_i-\mathbf{B}_i'\mathbf{g}_i)(\mathbf{y}_i-\mathbf{B}_i'\mathbf{g}_i)'\Big).$$

Since the $\lambda_i$ for $i=1,2,\ldots,n+1$ are only dependent on the $\mathbf{Q}_k$, we use the Gibbs sampler to update them from their own conditional distributions as

$$[\lambda_i^{-1}|\mathbf{Q},a_i,b_i] \sim\ Ga\Big(\eta\Big|\frac{a_i+qc}{2},\frac{b_i+\sum_{k=1}^c(\boldsymbol{\beta}_i^{(k)})'\mathbf{\Sigma}^{-1}\boldsymbol{\beta}_i^{(k)}}{2}\Big),$$

where $(\boldsymbol{\beta}_i^{(k)})'$ is the $i$th row of $\mathbf{Q}_k$.

As for the estimate of $\nu$, we follow the data augmentation technique proposed by Escobar and West (1995). That is, given the currently sampled values of $c$ and $\nu$, ones sample an random variable $\omega$ from Beta distribution $Be(\nu+1,n)$; ones then sample a new $\nu$ from the following mixture as

$$[\nu|\omega,c] \sim \pi_0 Ga(a_\nu+c,b_\nu-\log(\omega)) + (1-\pi_0)Ga(a_\nu+c-1,b_\nu-\log(\omega))$$

with $\pi_0=\frac{\nu+c-1}{a_\nu+c-1+n(b_\nu-\log(\omega))}$.

**Acknowledgements**

Zhihua Zhang acknowledges support from Doctoral Program of Specialized Research Fund of Chinese Universities and from Chinese Universities Scientific Fund.

## References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*, 1152–1174.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*, 353–355.

Boyle, P. and M. Frean (2005). Dependent Gaussian processes. In *Advances in Neural Information Processing Systems 17*.

Breiman, L. and J. Friedman (1997). Predicting multivariate responses in multiple linear regression (with discussion). *Journal of the Royal Statistical Society, B 59*(1), 3–54.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika 83*, 275–285.

Cressie, N. (1993). *Statistics for Spatial Data* (Revised ed.). New York: Wiley.

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

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society Series B 69*(2), 163–183.

Dunson, D. B., Y. Xue, and L. Carin (2008). The matrix stick-breaking process. *Journal of the American Statistical Association 103*(481), 317–327.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

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

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association 100*, 1021–1035.

Golub, G. H. and C. F. V. Loan (1996). *Matrix Computations* (Third ed.). Baltimore: The Johns Hopkins University Press.

Griffin, J. E. and M. F. J. Steel (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association 101*(473), 179–194.

Gupta, A. and D. Nagar (2000). *Matrix Variate Distributions*. Chapman & Hall/CRC.

Ibrahim, J. G. and K. P. Kleinman (1998). Semiparametric Bayesian methods for random effects models. In D. Dey, P. Müller, and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 89–114. New York: Springer-Verlag.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics 12*(1), 351–357.

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In D. Dey, P. Müller, and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 23–43. New York: Springer-Verlag.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *The Section on Bayesian Statistical Science*, pp. 50–55. American Statistical Association.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics*, Volume 6, pp. 475–501. Oxford University Press.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*, 249–265.

Rasmussen, C. E. and C. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. The MIT Press.

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

Skagerberg, B., J. MacGregor, and C. Kiparissides (1992). Multivariate data analysis applied to low-density polythylene reactors. *Chemometrics and intelligent laboratory systems 14*, 341–356.

Teh, Y. W., M. Seeger, and M. I. Jordan (2005). Semiparametric latent factor models. In *Proceedings of the Eighth Conference on Artificial Intelligence and Statistics (AISTATS)*.

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

Zhang, Z. and M. I. Jordan (2006). Bayesian multicategory support vector machines. In *the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*.