An Autoregressive Approach to Nonparametric Hierarchical Dependent Modeling

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

Dakan Wang
Department of Computer Science
Stanford University
Stanford, CA 94305
vondrak@stanford.edu

Edward Y. Chang
Google Research
Beijing 100084, China
edchang@google.com

Abstract

We propose a conditional autoregression framework for a collection of random probability measures. Under this framework, we devise a conditional autoregressive Dirichlet process (DP) that we call one-parameter dependent DP (ωDDP). The appealing properties of this specification are that it has two equivalent representations and its inference can be implemented in a conditional Pólya urn scheme. Moreover, these two representations bear a resemblance to the Pólya urn scheme and the stick-breaking representation in the conventional DP. We apply this ωDDP to Bayesian multivariate-response regression problems. An efficient Markov chain Monte Carlo algorithm is developed for Bayesian computation and prediction.

1 Introduction

Dirichlet processes (DPs) (Ferguson, 1973) or DP mixture models (Lo, 1984) are important nonparametric Bayesian modeling tools. After Markov chain Monte Carlo (MCMC) algorithms were developed for DP mixture models in the 1990s, DP mixture models have been used very successfully 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 (Ferguson, 1973; Antoniak, 1974).

In recent years, one of the most important developments in the DP literature is the notion of dependent DPs (DDPs), which provides a general framework to describe dependency among a collection of stochastic processes (MacEachern, 1999).

This paper is concerned with the concrete formulation of DDPs in dependent nonparametric modeling for a collection of related random probability distributions or stochastic functions. A principled approach to this direction is to treat the weights in the stick-breaking representation (Sethuraman, 1994) as stochastic functions (De Iorio et al., 2004; Griffin and Steel, 2006; Petrone et al., 2009). Such treatments are demanding computationally because conventional approaches for devising MCMC algorithms for DP mixture models based on the Pólya urn scheme (Blackwell and MacQueen, 1973) can no longer be used. A popular approach is to truncate the stick-breaking representation, but this forfeits some of the guarantees associated with MCMC algorithms (Ishwaran and James, 2001). Recently, Lin et al. (2010) proposed a new approach for constructing DDPs based on Poisson processes, and Zhang et al. (2010) developed a matrix-variate DP.

Other ways of achieving dependence among random measures include the hierarchical DP model (Teh et al., 2006), the use of linear combinations of realizations of independent DPs (Müller et al., 2004) and kernel-weighted mixture of DPs (Dunson et al., 2007). These are specialized approaches that can make use of generalized Pólya urn schemes for posterior inference and prediction.

In the spirit as the combination approach to combining random measures we propose an *autoregressive* model that yields conditional autoregressive DPs. We refer to our approach as the ωDDP. The ωDDP specification results in a conditional Pólya urn scheme, which can be used to devise efficient MCMC algorithms for posterior inference and prediction. Moreover, there exists an interesting resemblance between our ωDDP and the

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

conventional DP.

The second contribution of this paper is to exploit ωDDP in multivariate-response regression problems (Breiman and Friedman, 1997), giving rise to a nonparametric hierarchical model. 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,

$$f_j(\mathbf{x}) = u_j + \sum_{l=1}^{k} b_{jl} g_l(\mathbf{x}), \quad j = 1, \ldots, m$$

where u_j are offset terms, b_{jl} are regression coefficients, and $g_l(\mathbf{x})$ are basis functions whose type is usually prespecified. In the parametric setting, the regression vectors $\mathbf{b}_j = (b_{j1}, \ldots, b_{jk})^T$ are assumed to be fixed (but unknown) constants for all samples. This can yield an underfitted model if the order k does not match well to the complexity inherent in the samples. To take an extreme alternative nonparametric approach we might endow each sample with its own regression vector. This would overfit, thus we envision a two-stage form of coupling among these regression vectors.

In particular, we make use of DP priors to provide a joint distribution to the regression vectors for each response at the first stage, and then incorporate these DP priors thorough ωDDP at the second stage. The clustering property of DPs naturally allows the sharing of statistical strength between and within the two stages, but also allows no sharing. Moreover, the clustering property is able to transfer statistical strength from existing regression vectors to new regression vectors, and thus yield out-of-sample prediction.

We employ the conditional 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.

Our regression model not only captures the relationship among the output samples, but also the relationship among the output 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 to repeatedly invert $n \times n$ matrices where n is the number of training samples, limiting their applications in large-scale datasets. However, our model can avoid this limitation. Note that the kernel weighted mixture of DPs (Dunson et al., 2007) is also able to capture relationships among the output samples, but it cannot be used to model the dependence among the output variates.

2 Dirichlet Process Mixtures

In a Dirichlet Process Mixture (DPM) model, the samples $\mathbf{z}_i$ for $i = 1, \ldots, n$ are assumed to be drawn from a mixture component parameterized by $\boldsymbol{\theta}_i \in \Theta$. The $\boldsymbol{\theta}_i$s are in turn generated by the distribution G, which is assumed to follow a Dirichlet process prior. If G is drawn from the Dirichlet process $\mathrm{DP}(G_0, \alpha)$ with base measure G_0 and concentration parameter α over $(\Theta, \mathcal{B})$ then for any finite partition $(B_1, \ldots, B_k)$ of $\mathcal{B}$,

$$(G(B_1), \ldots, G(B_k)) \sim \mathrm{Dir}(\alpha G_0(B_1), \ldots, \alpha G_0(B_k)).$$

Here $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ denotes the Dirichlet distribution with positive parameters $\alpha_1, \ldots, \alpha_k$.

As is well known, integrating over G results in a Pólya urn scheme for the $\boldsymbol{\theta}_i$ (Blackwell and MacQueen, 1973); that is,

$$\boldsymbol{\theta}_1 \sim G_0(\boldsymbol{\theta}_1),$$
$$\boldsymbol{\theta}_i | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1} \sim \frac{\alpha G_0(\boldsymbol{\theta}_i) + \sum_{l=1}^{i-1} \delta(\boldsymbol{\theta}_i | \boldsymbol{\theta}_l)}{\alpha + i - 1},$$

where $\delta(\boldsymbol{\theta}_i | \boldsymbol{\theta}_l)$ is a point mass at $\boldsymbol{\theta}_l$. It is easy to see that as $\alpha \to 0$, all the $\boldsymbol{\theta}_i$ are identical to $\boldsymbol{\theta}_1$, which in turn follows G_0. When $\alpha \to \infty$, the $\boldsymbol{\theta}_i$ becomes iid G_0. Since the $\boldsymbol{\theta}_i$ are exchangeable, the Pólya urn scheme can be written as

$$\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i} \sim \frac{\alpha G_0(\boldsymbol{\theta}_i) + \sum_{l \neq i} \delta(\boldsymbol{\theta}_i | \boldsymbol{\theta}_l)}{\alpha + n - 1}, \tag{1}$$

were $\boldsymbol{\theta}_{-i}$ represents $\{\boldsymbol{\theta}_l : l \neq i\}$.

3 Nonparametric Hierarchical Dependent Modeling

In this paper, all vectors are represented in the column form. We let $\mathbf{1}_m$ denote the $m \times 1$ vector of 1's, $\mathbf{I}_m$ denote the $m \times m$ identity matrix, and $\mathbf{0}$ denote the zero vector (or matrix) whose dimensionality is dependent upon the context.

In order to model relationships among multiple studies, we consider a nonparametric hierarchical model. Let $\mathbf{y}_{\cdot j} = (y_{1j}, \ldots, y_{n_j j})^T$ denote the response vector in study j. The model is

$$\left[y_{ij} | \mathbf{b}_{ij}\right] \overset{ind}{\sim} p(y_{ij} | \mathbf{b}_{ij}), \; j = 1, \ldots, m \text{ and } i = 1, \ldots, n_j; \tag{2}$$
$$\left[\mathbf{b}_{ij} | G_j\right] \overset{iid}{\sim} G_j, \; i = 1, \ldots, n_j \text{ for each } \; j.$$

In general, there are two extreme constructions for the G_j. For the first construction, G_j are treated as independent DPs given hyperparameters $\boldsymbol{\theta}$, so the model is equivalent to the m separate submodels. For the second one, the model is treated as a single conventional

DP, i.e., $G_1 = \cdots = G_m$. As discussed by Müller et al. (2004), the first case allows too little sharing of strength in many applications, while the second case enforces too much sharing.

3.1 Conditional Autoregressive DPs

Let $\mathcal{G} = \{G_j, j = 1, \ldots, m\}$ denote a collection of random probability measures on $(\Phi, \mathcal{B})$. We model the G_j in autoregressive form of

$$G_j = \omega_{jj} G_j^* + \sum_{l \neq j} \omega_{jl} G_l, \quad j = 1, \ldots, m, \tag{3}$$

where $0 \leq \omega_{jl} < 1$ and $\sum_{l=1}^m \omega_{jl} = 1$. From (3), we get the following conditional autoregressive model

$$E(G_j(B)|G_l(B), l \neq j) = \omega_{jj} G_0(B) + \sum_{l \neq j} \omega_{jl} G_l(B)$$

for any Borel set $B \in \mathcal{B}$. We thus say the G_j defined by (3) follow *conditional autoregressive* DPs. We denote $G_j \sim \mathrm{DDP}(\mathcal{G}^*, \boldsymbol{\omega}_j)$ where $\boldsymbol{\omega}_j = (\omega_{j1}, \ldots, \omega_{jm})^T$ and $\mathcal{G}^* = \{G_1^*, \ldots, G_m^*\}$.

Using the induction principle (details are given in Appendix), we can express (3) in the following form

$$\begin{aligned} G_1 &= \beta_{11} G_1^* \\ G_2 &= \beta_{21} G_1 + \beta_{22} G_2^* \\ \vdots &= \quad \vdots \\ G_m &= \beta_{m1} G_1 + \cdots + \beta_{m,m-1} G_{m-1} + \beta_{mm} G_m^*, \end{aligned} \tag{4}$$

where the β_{jl} satisfy $\beta_{jl} \geq 0$, $\beta_{ll} > 0$ and $\sum_{l=1}^j \beta_{jl} = 1$, and the G_j^* are independent from $\mathrm{DP}(G_0, \nu_j)$. Thus, models (4) and (3) are mutually equivalent. Model (4) shows that the conditional autoregressive DP can serve for countably infinite random probability distributions. That is, given a new study $m+1$ (out-of-sample study), we always have

$$G_{m+1} = \beta_{(m+1),(m+1)} G_{m+1}^* + \sum_{l=1}^m \beta_{(m+1),l} G_l.$$

On the other hand, let us denote $\boldsymbol{\Omega} = [\omega_{jl}]$ $(m \times m)$, $\mathrm{dg}(\boldsymbol{\Omega}) = \mathrm{diag}(\omega_{11}, \ldots, \omega_{mm})$, $\mathbf{M} = \boldsymbol{\Omega} - \mathrm{dg}(\boldsymbol{\Omega})$ and $\mathbf{A} = [a_{ij}] = (\mathbf{I}_m - \mathbf{M})^{-1} \mathrm{dg}(\boldsymbol{\Omega})$. Then, $\mathbf{A}$ is nonnegative and row stochastic, i.e. $\mathbf{A} \geq \mathbf{0}$ and $\mathbf{A}\mathbf{1}_m = \mathbf{1}_m$ (the proof is given in Appendix B). Thus, the G_j can be expressed as

$$G_j = \sum_{l=1}^m a_{jl} G_l^*, \quad G_l^* \overset{ind}{\sim} \mathrm{DP}(G_0, \nu_l), \tag{5}$$

for $j = 1, \ldots, m$ (also see Dunson et al., 2007, Theorem 2). Conversely, given a row stochastic matrix $\mathbf{A}$ for (5), we cannot always obtain a row stochastic matrix $\boldsymbol{\Omega}$ for (3). Thus, it is not always possible to derive (3) from (5). However, when the inverse $\mathbf{B} = [b_{jl}]$ of row stochastic $\mathbf{A}$ satisfies $b_{jj} > 0$ and $b_{jl} \leq 0$ for $j \neq l$; namely, $\mathbf{B}$ is an M-matrix (Saad, 2003), we can derive an $\boldsymbol{\Omega}$ and obtain (3) from (5). Summarizing, we are able to show (the proof involves straightforward algebraic manipulations):

Theorem 1 *Assume that* $\mathbf{A} = [a_{jl}]$ *in (5) satisfies* $a_{jl} \geq 0$, $a_{jj} > 0$ *and* $\sum_{l=1}^m a_{jl} = 1$, *and it is nonsingular. If* $\mathbf{B} = \mathbf{A}^{-1}$ *is an* M*-matrix, then there exists an* $\boldsymbol{\Omega} = (\mathrm{dg}(\mathbf{B}))^{-1}[\mathbf{I}_m + \mathrm{dg}(\mathbf{B}) - \mathbf{B}]$ *such that (3) holds.*

3.2 One-parameter Dependent DP

In this paper we present a family of special matrices $\mathbf{A}$; that is,

$$\mathbf{A} = \frac{1}{\omega + m}\Big(\omega \mathbf{I}_m + \mathbf{1}_m \mathbf{1}_m^T\Big), \quad \text{for } \omega > 0.$$

Thus, (5) reduces to

$$G_j = \frac{\omega + 1}{\omega + m} G_j^* + \frac{1}{\omega + m} \sum_{l \neq j} G_l^*. \tag{6}$$

Noting that $\mathbf{A}^{-1} = \frac{1}{\omega}\big((\omega{+}m)\mathbf{I}_m - \mathbf{1}_m \mathbf{1}_m^T\big)$ which is an M-matrix, we obtain $\boldsymbol{\Omega}$ given by

$$\boldsymbol{\Omega} = \frac{1}{\omega{+}(m{-}1)}\big((\omega - 1)\mathbf{I}_m + \mathbf{1}_m \mathbf{1}_m^T\big).$$

Subsequently, we have

$$G_j = \frac{\omega}{\omega{+}m{-}1} G_j^* + \frac{1}{\omega{+}m{-}1} \sum_{l \neq j} G_l. \tag{7}$$

Again, using the induction principle, we can equivalently express (7) as

$$G_j = \frac{\omega}{\omega{+}j{-}1} G_j^* + \frac{1}{\omega{+}j{-}1} \sum_{l=1}^{j-1} G_l. \tag{8}$$

Since the resulting dependent DP is scaled only by a single parameter ω, we call it *one-parameter* DDP (ωDDP) and denote by $G_j \sim \mathrm{DDP}(\mathcal{G}^*, \omega)$. It is easily seen from either (7) or (6) that, as $\omega \to \infty$, the G_j $(= G_j^*)$ are mutually independent from $\mathrm{DP}(G_0, \nu_j)$, while as $\omega \to 0$, we have $G_1 = \cdots = G_m$ $(= \frac{1}{m}\sum_{l=1}^m G_l^*)$. It is also worth pointing out that our ωDDP bears a close resemblance to the conventional DP; in particular, (6) corresponds to the (truncated) stick-breaking representation and (7) (or (8)) corresponds to the Pólya urn scheme (or the Chinese restaurant process).

3.3 Conditional Pólya urn Scheme

Applying DDP$(\mathcal{G}^*, \omega)$ to model (2), we have

$$\big[\mathbf{b}_{ij}|G_j\big] \overset{iid}{\sim} G_j, \quad i = 1, \ldots, n_j \text{ for each } \; j; \qquad (9)$$
$$G_j \sim \mathrm{DDP}(\mathcal{G}^*, \omega), \quad G_j^* \overset{ind}{\sim} \mathrm{DP}(G_0, \nu_j).$$

See Figure 1-(a) for a graphical model representation. By introducing indicators r_{ij}, the above model can be equivalently expressed as

$$\big[\mathbf{b}_{ij}|r_{ij} = h\big]{\sim}G_h^*, \quad G_h^*{\sim}\mathrm{DP}(G_0, \nu_h), \; h = 1, \ldots, m;$$
$$\big[r_{ij}|\omega\big] \sim \mathrm{Multinomial}(\{1, \ldots, m\}, \boldsymbol{\gamma}_j),$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jm})^T$ with $\gamma_{jj} = (\omega{+}1)/(\omega{+}m)$ and $\gamma_{jl} = 1/(\omega{+}m)$ for $j \neq l$. As is well known (Blackwell and MacQueen, 1973), integrating over G_h^* results in a Pólya urn scheme for $\mathbf{b}_{ij}$; that is,

$$\big[\mathbf{b}_{ij}|\mathbf{b}_{-ij}, r_{ij} = h, \mathbf{r}_{-ij}\big] \sim \frac{\nu_h G_0 + \sum_{l \neq i} \delta(\mathbf{b}_{ij}|\mathbf{b}_{lh})}{\nu_h + n_h - 1}, \qquad (10)$$

where $\delta(\mathbf{x}|\mathbf{b})$ is a point mass at $\mathbf{b}$, $\mathbf{b}_{-ij}$ represents $\{\mathbf{b}_{lk} : l \neq i \text{ or } k \neq j\}$, and similarly for $\mathbf{r}_{-ij}$. Let $\Phi_h = \{\boldsymbol{\phi}_{kh}, k = 1, \ldots, c_h\}$ denote the set of distinct values among the $\{\mathbf{b}_{ij} : r_{ij} = h\}$, η_{kh} denote occurrences of $\boldsymbol{\phi}_{kh}$, and $\eta_h = \sum_k \eta_{kh}$, for $h = 1, \ldots, m$. The set of configuration indicators $S = \{s_{ij}\}$ is defined by $s_{ij} = (lh)$ if and only if $\mathbf{b}_{ij} = \boldsymbol{\phi}_{lh}$. Thus, (S, Φ) is an equivalent representation of the $\mathbf{b}_{ij}$, and hence (10) reduces to

$$\big[\mathbf{b}_{ij}|\mathbf{b}_{-ij}, r_{ij} = h, \mathbf{r}_{-ij}\big] \sim \frac{\nu_h G_0 + \sum_{l=1}^{c_h} \eta_{lh}^{-} \delta(\mathbf{b}_{ij}|\boldsymbol{\phi}_{lh})}{\nu_h + \eta_h^{-}}, \qquad (11)$$

where η_{kh}^{-} represents the number of clustering (kh), with $\mathbf{b}_{ij}$ removed, and similarly for η_h^{-}. By marginalizing over r_{ij}, the conditional prior of $\mathbf{b}_{ij}$ on $\mathbf{r}_{-ij}$ is given by

$$\big[\mathbf{b}_{ij}|\mathbf{b}_{-ij}, \mathbf{r}_{-ij}\big] \qquad (12)$$
$$\sim \sum_{h=1}^{m} \frac{\nu_h \gamma_{jh}}{\nu_h + \eta_h^{-}} G_0 + \sum_{h=1}^{m} \sum_{l=1}^{c_h} \frac{\gamma_{jh} \eta_{lh}^{-}}{\nu_h + \eta_h^{-}} \delta(\mathbf{b}_{ij}|\boldsymbol{\phi}_{lh}).$$

For a new $\mathbf{b}_{0j}$, it follows from (12) that the conditional predictive prior for $\mathbf{b}_{0j}$ as

$$\big[\mathbf{b}_{0j}|\{\mathbf{b}_{ij}\}, \mathbf{r}_{-ij}\big] \qquad (13)$$
$$\sim \sum_{h=1}^{m} \frac{\nu_h \gamma_{jh}}{\nu_h + \eta_h} G_0 + \sum_{h=1}^{m} \sum_{l=1}^{c_h} \frac{\gamma_{jh} \eta_{lh}}{\nu_h + \eta_h} \delta(\mathbf{b}_{ij}|\boldsymbol{\phi}_{lh}).$$

Note that there are possibly some $h \in \{1, \ldots, m\}$ such that $\eta_h = 0$.

4 Nonparametric Models for Multivariate-Response Regression

In this section we apply our ωDDP to multivariate regression problems. We are now given a set of training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ where $\mathbf{x}_i$ is a $d{\times}1$ input vector and $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T \in \mathbb{R}^m$ is an m-dimensional response vector.

We consider the following regression model

$$y_{ij} = u_j + \mathbf{g}_i^T \mathbf{b}_{ij} + \epsilon_j,$$

where the ϵ_j are independent normal errors with mean 0 and variance σ^2, $\mathbf{g}_i = \mathbf{g}(\mathbf{x}_i) = (g_1(\mathbf{x}_i), \ldots, g_k(\mathbf{x}_i))^T$ are basis functions and $\mathbf{b}_{ij}$ are $k{\times}1$ regression vectors. Unlike the conventional regression model, the current model allows each sample $\mathbf{x}_i$ to have its own $\mathbf{b}_{ij}$. In this paper, we specifically define $g_j(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_j)$, where $K(\mathbf{x}, \mathbf{x}_j)$ is a reproducing kernel.For simplicity, we let $y_{ij} \leftarrow y_{ij} - \frac{1}{n}\sum_{i=1}^n y_{ij}$ and set $u_j = 0$ for $j = 1, \ldots, m$.

To capture the relationships among the y_{ij}, we model them as (9) where we set $\nu_1 = \cdots = \nu_m = \nu$ and define G_0 as

$$G_0(\cdot|\tau, \boldsymbol{\Sigma}) = N_n(\mathbf{0}, \; \tau^{-1}\mathbf{K}^{-1}),$$

where $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$ is the $n{\times}n$ kernel matrix. We further assume that σ^{-2}, τ, ν follow Gamma distributions $\Gamma(\sigma^{-2}|\frac{a_\sigma}{2}, \; \frac{b_\sigma}{2})$, $\Gamma(\tau|\frac{a_\tau}{2}, \; \frac{b_\tau}{2})$ and $\Gamma(\nu|\frac{a_\nu}{2}, \; \frac{b_\nu}{2})$, respectively. In addition, ω and the other hyperparameters of the priors for ν, τ and σ^2 are fixed.

4.1 Inference

Posterior inference is achieved by generating realizations of the parameters from the conditional joint density $[\mathbf{b}, \sigma^2, \tau, \nu|\mathbf{Y}]$. Using the same notations as those in the previous section, we present a Gibbs sampler, which consists of the following steps:

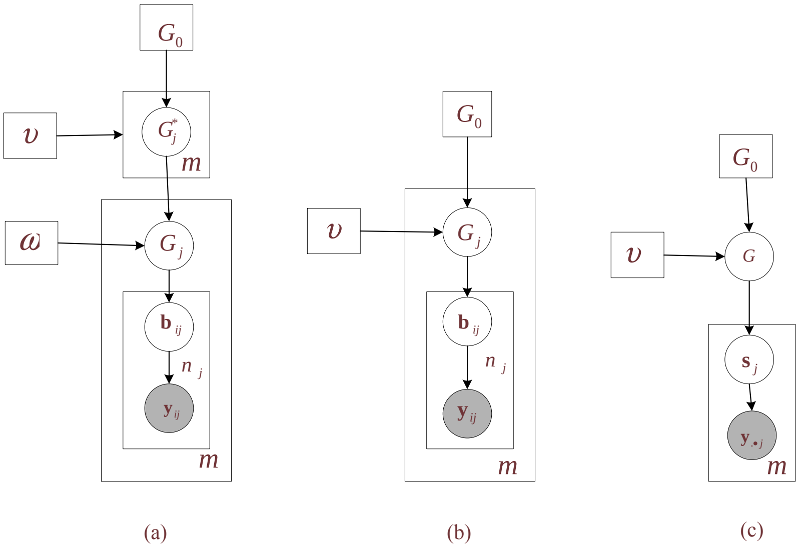
(a) Update $(\mathbf{b}_{ij}, r_{ij}, s_{ij})$ from $[(\mathbf{b}_{ij}, r_{ij}, s_{ij})|(\mathbf{b}_{-ij}, \mathbf{r}_{-ij}, \mathbf{s}_{-ij}), \nu, \tau, \sigma^2, \mathbf{Y}]$ for $j = 1, \ldots, m$ and $i = 1, \ldots, n$;

(b) Update $\boldsymbol{\phi}_{kh}$ from $[\boldsymbol{\phi}_{kh}|\mathbf{r}, \tau, \nu, \sigma^2, \mathbf{Y}]$ for $h = 1, \ldots, m$ and $k = 1, \ldots, c_h$;

(c) Update σ^{-2}, τ and ν from $[\sigma^{-2}|\mathbf{Y}, \mathbf{b}, a_\sigma, b_\sigma]$, $[\tau|\{\Phi_h\}_{h=1}^m, a_\tau, b_\tau]$ and $[\nu|\{\Phi_h\}_{h=1}^m, a_\nu, b_\nu]$.

The Gibbs sampler exploits the simple structure of the conditional posterior for each $\mathbf{b}_{ij}$. In terms of the conditional Pólya urn scheme in (12), the conditional distribution is given by

$$[\mathbf{b}_{ij}|\mathbf{b}_{-ij}, r_{-ij}, \nu, \boldsymbol{\gamma}, \tau, \sigma^2, \mathbf{Y}]$$
$$\propto q_0 N_n(\mathbf{b}_{ij}|\sigma^{-2}\mathbf{Q}_i \mathbf{g}_i y_{ij}, \; \mathbf{Q}_i) + \sum_{h=1}^{m} \sum_{l=1}^{c_h} q_{lh} \frac{\gamma_{jh} \eta_{lh}^{-}}{\nu + \eta_h^{-}} \delta(\mathbf{b}_{ij}|\boldsymbol{\phi}_{lh}),$$



(a)

(b)

(c)

Figure 1: Graphical Representations: (a) ωDDP mixture model, (b) m independent DP mixture models, and (c) DP mixture model.

where

$$q_{lh} = N(y_{ij}|\mathbf{g}_i^T \boldsymbol{\phi}_{lh},\ \sigma^2),$$

$$q_0 = N(y_{ij}|0,\ (\tau^{-1}\mathbf{g}_i^T \mathbf{K}^{-1}\mathbf{g}_i + \sigma^2)) \sum_{h=1}^{m} \frac{\nu \gamma_{jh}}{\nu + \eta_h^-},$$

and

$$\begin{aligned}\mathbf{Q}_i &= (\tau\mathbf{K} + \sigma^{-2}\mathbf{g}_i\mathbf{g}_i^T)^{-1} \\ &= \tau^{-1}\mathbf{K}^{-1} - \tau^{-1}\mathbf{K}^{-1}\mathbf{g}_i(\tau\sigma^2 + \mathbf{g}_i^T\mathbf{K}^{-1}\mathbf{g}_i)^{-1}\mathbf{g}_i^T\mathbf{K}^{-1}.\end{aligned}$$

Thus, given $\mathbf{b}_{-ij}$, with probability proportional to $q_{lh}\frac{\gamma_{jh}\eta_{lh}^-}{\nu+\eta_h^-}$, we draw $\mathbf{b}_{ij}$ from distribution $\delta(\cdot|\boldsymbol{\phi}_{lh})$, or with probability proportional to q_0, we draw $\mathbf{b}_{ij}$ from $N_n(\cdot|\sigma^{-2}\mathbf{Q}_i\mathbf{g}_i y_{ij},\ \mathbf{Q}_i)$.

To speed mixing of the Markov chain, Bush and MacEachern (1996) suggested resampling the $\boldsymbol{\phi}_{kh}$ after every step. For each $h = 1, \ldots, m$ and $k = 1, \ldots, c_h$, we have

$$[\boldsymbol{\phi}_{kh}|\mathbf{Y}, S, \tau, \sigma^2] \propto N_n(\boldsymbol{\phi}_{kh}|\mathbf{0},\ \tau^{-1}\mathbf{K}^{-1}) \times \prod_{(ij):\ s_{ij}=(kh)} N(y_{ij}|\mathbf{g}_i^T\boldsymbol{\phi}_{kh},\ \sigma^2),$$

from which it follows that the conditional density of $\boldsymbol{\phi}_{kh}$ is given by

$$[\boldsymbol{\phi}_{kh}|\mathbf{Y}, S, \mathbf{K}, \tau, \sigma^2] \sim N_n\Big(\boldsymbol{\phi}_{kh}|\sigma^{-2}\boldsymbol{\Psi}_{kh} \sum_{(ij):\ s_{ij}=(kh)} y_{ij}\mathbf{g}_i,\ \boldsymbol{\Psi}_{kh}\Big)$$

with $\boldsymbol{\Psi}_{kh} = (\tau\mathbf{K} + \sigma^{-2}\sum_{(ij):\ s_{ij}=(kh)} \mathbf{g}_i\mathbf{g}_i^T)^{-1}$ for each $h = 1, \ldots, m$ and $k = 1, \ldots, c_h$.

Given the prior of σ^{-2}, we then obtain the update of σ^{-2} as

$$[\sigma^{-2}|\mathbf{y}, \mathbf{b}, a_\sigma, b_\sigma] \sim \Gamma\Big(\sigma^{-2}\Big|\frac{a_\sigma + nm}{2}, \frac{b_\sigma + \sum_{j=1}^m \sum_{i=1}^n (y_{ij} - \mathbf{g}_i^T\mathbf{b}_{ij})^2}{2}\Big).$$

Since τ is only dependent on the $\boldsymbol{\phi}_{kh}$, we use the Gibbs sampler to update them from their own conditional distributions as

$$[\tau|\boldsymbol{\phi}, a_\tau, b_\tau] \sim \Gamma\Big(\tau\Big|\frac{a_\tau + n\sum_{h=1}^m c_h}{2}, \frac{b_\tau + \sum_{h=1}^m \sum_{k=1}^{c_h} \boldsymbol{\phi}_{kh}^T\mathbf{K}\boldsymbol{\phi}_{kh}}{2}\Big).$$

As for the update of ν, it is immediately obtained from MacEachern (1998).

The main computational burden of the algorithm comes from the calculation of $\boldsymbol{\Psi}_{kh}$. However, we can use the Sherman-Morrison-Woodbury formula to calculate $\boldsymbol{\Psi}_{kh}$. This formula allows us to invert an $\eta_{kh} \times \eta_{kh}$ matrix instead of an $n \times n$ matrix. Thus, when reproducing kernels as basis functions are used for a large-scale dataset, the algorithm is still efficient.

4.2 Prediction

Given a new input vector $\mathbf{x}_0$, we predict the corresponding response $\mathbf{y}_0 = (y_{01}, \ldots, y_{0m})^T$. Let the $\mathbf{b}_{0j}$ be the associated regression vectors. Prediction is based on the cluster structure of the $\mathbf{b}_{ij}$. A non-Bayesian approach is to choose the c_h with the highest posterior probability among those drawn from the MCMC algorithm. Let $\hat{\mathbf{b}}_{kh}$, $k = 1, \ldots, c_h$ be the M-CMC approximations of the $\mathbf{b}_{kh}$ associated with c_h, for $h = 1, \ldots, m$. Consequently, y_{0h} is predicted as

$$\hat{y}_{0h} = \frac{1}{c_h} \sum_{k=1}^{c_h} \mathbf{g}_0^T \hat{\mathbf{b}}_{kh}$$

where $\mathbf{g}_0 = (g_1(\mathbf{x}_0), \ldots, g_n(\mathbf{x}_0))^T$. This approach requires large storage to record the results for the computation of posterior probabilities, thus it is not feasible in practice.

In this paper we are interested in Bayesian nonparametric prediction. In particular, we utilize the conditional predictive prior for $\mathbf{b}_{0j}$ in (13). Let $\{\mathbf{b}^{(t)}, (\sigma^2)^{(t)}, \tau^{(t)}, \nu^{(t)}\}$, $t = 1, \ldots, T$, be the MCMC realizations of the parameters after the burn-in period. We consider a Bayesian averaging approach (Raftery et al., 1997). The approach is to draw $\mathbf{b}_{0j}^{(t)}$ from (13) with the parameter realizations. We thus have $\hat{\mathbf{b}}_{0j} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{b}_{0j}^{(t)}$, and hence $\hat{y}_{0j} = \mathbf{g}_0^T \hat{\mathbf{b}}_{0j}$.

5 Example Studies

We conduct some numerical experiments to test the performance of our proposed Bayesian regression method and compare it with the model shown in Figure 1-(b). Specifically, $\mathbf{y}_{\cdot j} = (y_{1j}, \ldots, y_{nj})^T$, $j = 1, \ldots, m$, are modeled as m mutually independent DP (iDP) mixture models. namely, for $j = 1, \ldots, m$,

$$\begin{array}{rcl} y_{ij}|\mathbf{b}_{ij}, \sigma^2 & \overset{ind}{\sim} & N_n(\mathbf{y}_{\cdot j}|\mathbf{g}_i^T \mathbf{b}_{ij},\ \sigma^2), \quad i = 1, \ldots, n; \\ \mathbf{b}_{ij}|G_j & \overset{iid}{\sim} & G_j, \quad i = 1, \ldots, n; \\ G_j|\nu, G_0 & \overset{iid}{\sim} & \mathrm{DP}(G_0, \nu); \\ G_0(\cdot|\mathbf{K}, \tau) & = & N_n(\cdot|\mathbf{0},\ \tau^{-1}\mathbf{K}^{-1}). \end{array}$$

The above iDPs can be easily implemented by setting ω to a very large value in our ωDDP model. In addition to the above iDP model, we further use another baseline which assumes that the m independent models follow the settings in the most recent literature (Shahbaba and Neal, 2009; Hannah et al., 2010). Namely, for $j = 1, \ldots, m$, the specification is

$$\begin{array}{rcl} (\mathbf{x}_i, y_{ij})|\mathbf{b}_{ij}, \sigma^2, \mu_i, \Sigma_i & \overset{ind}{\sim} & N(y_{ij}|\mathbf{x}_i^T \mathbf{b}_{ij},\ \sigma^2) \times \\ & & N_n(\mathbf{x}_i|\mu_i, \Sigma_i), i = 1, \ldots, n; \\ (\mathbf{b}_{ij}, \mu_i, \Sigma_i)|G_j & \overset{iid}{\sim} & G_j, \quad i = 1, \ldots, n; \\ G_j|\nu, G_0 & \overset{iid}{\sim} & \mathrm{DP}(G_0, \nu). \end{array}$$

We here refer to this model as dpReg. It has been illustrated by Shahbaba and Neal (2009) that the dpReg model can handle nonlinear data fairly well.

Table 1: Summary of the four used datasets: d–the dimension of $\mathbf{x}$, m–the dimension of $\mathbf{y}$, k–the number of instances; n–the number of training data

Dataset	d	m	k	n
Chemometrics	22	6	58	35
biscuit	700	4	70	39
forest fire	7	6	517	150
robot arm	12	6	600	300

In the third counterpart which is also used for comparison, $p(\mathbf{y}_{\cdot 1}, \ldots, \mathbf{y}_{\cdot m})$ follows a DP mixture model and the base measure G_0 is defined as a Gaussian process. In particular, we establish the following model:

$$\begin{array}{rcl} \mathbf{y}_{\cdot j}|\mathbf{s}_j, \sigma^2 & \overset{ind}{\sim} & N_n(\mathbf{y}_{\cdot j}|\mathbf{K}\mathbf{s}_j,\ \sigma^2\mathbf{I}_n), \quad j = 1, \ldots, m; \\ \mathbf{s}_j|G & \overset{iid}{\sim} & G, \quad j = 1, \ldots, q; \\ G|\nu, G_0 & \sim & \mathrm{DP}(G_0, \nu); \\ G_0(\cdot|\mathbf{K}, \tau) & = & N_n(\cdot|\mathbf{0},\ \tau^{-1}\mathbf{K}^{-1}). \end{array}$$

This model is equivalent to the spatial Dirichlet process (sDP) mixture model of Gelfand et al. (2005), which is also a special formulation of dependent DPs. We use the MCMC algorithms built on the Pólya urn scheme for these models.

We adopted four datasets to compare our algorithm with iDPs and dpReg: the Chemometrics data, the biscuit data, the forest fire data, and the robot-arm data. The Chemometrics data was introduced in (Skagerberg et al., 1992). According to the suggestion in (Breiman and Friedman, 1997), we instead use the logarithms of the responses values for our experimental analysis. The biscuit dataset was respectively used by Breiman and Friedman (1997) and Brown et al. (2001) to analyze their regression methods. The forest fire data is available in the UCI machine learning repository, and the robot arm data was used by Teh et al. (2005). The information about each dataset is summarized in Table 1

For each dataset, we estimate the parameters from the training dataset and evaluate the performance on the test dataset. For regression response j, we adopt the root mean squared error to be the performance metric:

$$E_j = \sqrt{\frac{1}{n_2} \sum_{i=1}^{n_2} (y_j(\mathbf{x}_i) - \widehat{y}_j(\mathbf{x}_i))^2}, j = 1, \ldots, m.$$

Here n_2 is the number of instances in the test dataset, $y_j(\mathbf{x}_i)$ is the true regression response and $\widehat{y}_j(\mathbf{x}_i)$ is the predicted response from a regression model.

We run each MCMC algorithm for $10,000$ sweeps, discarding the first $5,000$ sweeps as the burn-in, and average the estimated parameters in each iteration after the burn-in for prediction. For all experiments, the input data is standardized to have zero mean and unity variance, and K is chosen to be the RBF Gaussian Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{\delta^2})$. The scale parameter δ is set to be the mean of the Euclidean distances over all training instance pairs. The hyperparameters for the scale factors are set as follows: $\omega = 10$, $a_\tau = 10$, $b_\tau = 1$, $a_\sigma = 10$, $b_\sigma = 1$, $a_\nu = 20$ and $b_\nu = 1$.

Table 2: Predictive Squared Errors For the Biscuit Data.

Method	y_1	y_2	y_3	y_4	Average
sDP	**0.0533**	0.4960	0.3466	0.0547	0.2377
iDPs	0.2123	0.5261	0.2693	**0.0366**	0.2611
ωDDP	0.0644	**0.3965**	**0.2004**	0.0544	**0.1789**

Tables 2, 3, 4 and 5 show the estimated prediction errors for all datasets. In the four tables, the best entry in each column is highlighted. We can see that ωDDP is the winner in most cases and consistently beat other algorithms with respect to the average mean squared error. Thus, these results empirically demonstrate that our dependent DP model is effective in the real-world applications. Notice that for the biscuit data, we do not report the results with dpReg. This is because the input dimension is 700, and dpReg is devised only for input data with moderate dimensions.

We also do not report the results with sDP for the forest fire data and the robot arm data due to that the computation of the sDP mixture model is demanding. First, the MCMC algorithm for sDP involves the computation of $n \times n$ matrices at each sweep. Second, this algorithm needs to calculate the densities of n-variate normal distributions $N_n\big(\,\cdot\,|\mathbf{0},\ \tau^{-1}\mathbf{K}_n + \sigma^2\mathbf{I}_n\big)$ and $N_n(\cdot|\mathbf{K}\mathbf{s}_j,\ \sigma^2\mathbf{I}_n)$. In the experiments, we find that the ratios (say, r) between some of these values get very large. This results in a slowly mixing Markov chain. To alleviate this problem, we apply 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 4.1, the ωDDP and iDP mixture models are efficient computationally. Moreover, their MCMC algorithms only involve calculating the densities of univariate normal distributions (see Section 4.1). Thus, they work very well without the need for the above trick.

Although the parameter ω in ωDPP can be adaptively learnt, we we simply specify $\omega = 10$ in our experiments. Basically, the choice of ω depends on the dimension of output vector (i.e. m). The higher the dimension, the smaller ω should be. However, ω should not be too big; otherwise it just degenerates to independent DPs. In the experiments, we find that as long as we do not set ω as some extreme values (too big or too small), the performances make little difference.

6 Discussion

One further extension to the ωDDP model is to consider the following setting

$$\begin{aligned}
\big[y_{ij}|\mathbf{b}_{ij}\big] &\overset{ind}{\sim} p(y_{ij}|\mathbf{b}_{ij}),\\
\big[\mathbf{b}_{ij}|G_{ij}\big] &\overset{iid}{\sim} G_{ij},\\
G_{ij} &= \sum_{k=1}^{m}\sum_{l=1}^{n_k} a_{lk}^{(ij)} G_{lk}^*,\\
G_{lk}^* &\overset{iid}{\sim} \mathrm{DP}(G_0, v_{lk}).
\end{aligned}$$

Here $a_{lk}^{(ij)} \geq 0$ and $\sum_{k=1}^{m}\sum_{l=1}^{n_k} a_{lk}^{(ij)} = 1$. This setting can capture dependency between the studies as well as the instances. When $m = 1$, the setting tries to explore dependency between the instances. In this case, Dunson et al. (2007) proposed a kernel weighted mixture of DPs as

$$G_{\mathbf{x}} = \sum_{l=1}^{n} b_l(\mathbf{x}) G_l^*,\ G_l^* \overset{iid}{\sim} \mathrm{DP}(G_0, \nu),\ \text{for } l = 1, \ldots, n,$$

where $b_l(\mathbf{x})$ is a kernel-based weight. This formulation is flexible to model local dependence between the instances. However, the weight matrix $\mathbf{A}$ in the kernel weighted mixture of DPs is defined through a kernel function and there does not exist a guarantee to have a corresponding matrix $\mathbf{\Omega}$ from such $\mathbf{A}$, Thus, (3) would no longer hold in their method. This implies that a corresponding autoregressive form for $G_{\mathbf{x}}$ cannot be obtained.

Note that the kernel weighted mixture of DPs was originally proposed to capture relationships among the output samples (Dunson et al., 2007). However, it cannot directly model the dependence among the output variates because it uses the $b_l(\mathbf{x})$ for weights. The goal of our ωDPP model is to describe the dependence among the output variates. In this case, it is clear that the weight matrix $\mathbf{A}$ of our model cannot be defined via a kernel function $b_l(\mathbf{x})$ as in the kernel weighted mixture of DPs.

7 Conclusion

We have proposed an autoregressive approach to nonparametric hierarchical dependent modeling problems. In particular, we have devised an ωDDP and exploited its application to Bayesian multivariate regression. The novel aspect is that we can develop an MCMC algorithm, based on the conditional Pólya urn scheme, for Bayesian computation and prediction.

Table 3: Predictive Squared Errors For the (Log-transformed) Chemometrics Data.

Method	y_1	y_2	y_3	y_4	y_5	y_6	Average
sDP	0.0093	0.0463	**0.0016**	0.0005	0.0003	0.0004	0.0098
dpReg	**0.0082**	0.0662	0.0040	0.0010	0.0006	0.0006	0.0135
iDPs	0.0136	0.0753	0.0019	0.0005	0.0003	0.0004	0.0153
ωDDP	0.0085	**0.0456**	**0.0016**	**0.0003**	**0.0002**	**0.0002**	**0.0095**

Table 4: Predictive Squared Errors For the Forest Fire Data.

Method	y_1	y_2	y_3	y_4	y_5	y_6	Average
dpReg	0.6909	0.5869	**0.8829**	1.0459	**0.1115**	1.3874	0.7842
iDPs	0.6530	0.5831	0.9034	1.0012	0.1145	**1.3384**	0.7656
ωDDP	**0.6170**	**0.5531**	0.9022	**0.9977**	0.1271	1.3461	**0.7572**

Table 5: Predictive Squared Errors For the Robot Arm Data.

Method	y_1	y_2	y_3	y_4	y_5	y_6	Average
dpReg	0.3984	0.6136	0.4451	**0.3650**	0.5737	**0.3227**	0.4531
iDPs	0.7487	0.5938	0.6597	0.5003	0.6030	0.5609	0.6111
ωDDP	**0.3937**	**0.5812**	**0.4316**	0.3676	**0.5615**	0.3231	**0.4431**

Our approach can readily be applied to Bayesian multicategory classification problems. We only need to add an extra stage into the hierarchical model for regression, constructing a new hierarchical model for classification. The extra stage is used to relate class labels with a set of auxiliary variables, which in turn play the same role as the responses in the regression model. To implement an MCMC algorithm for the classification model, it is only required to insert a Gibbs sampling that updates the auxiliary variables into the MCMC algorithm for the regression model. In fact, this implies the use of data augmentation methodology in the current classification model.

Acknowledgements

Zhihua Zhang acknowledges support from the Natural Science Foundations of China (No. 61070239) and the Google visiting faculty program.

A The Derivation of Obtaining (4) from (3)

The proof is done via induction. First, when $m = 1$, it is clear to obtain (4) from (3). Suppose the result holds under $m{-}1$. We now consider the case m. For $j = 1, \ldots, m{-}1$, multiplying $G_m = \omega_{mm} G_m^* + \sum_{l \neq m} \omega_{ml} G_l$ by ω_{jm} and then adding with $G_j = \omega_{jj} G_j^* + \sum_{l \neq j} \omega_{jl} G_l$ yield

$$G_j = \beta_{jj} \frac{\omega_{jj} G_j^* {+} \omega_{jm}\omega_{mm} G_m^*}{\omega_{jj} {+} \omega_{jm}\omega_{mm}} + \sum_{j \neq l, m} \beta_{jl} G_l$$

where $\beta_{jj} = \frac{\omega_{jj}+\omega_{jm}\omega_{mm}}{1-\omega_{jm}\omega_{mj}}$ and

$$\beta_{jl} = \frac{\omega_{jl} + \omega_{jm}\omega_{ml}}{1{-}\omega_{jm}\omega_{mj}} \quad \text{for } l \neq j, m.$$

It is easily verified that $\sum_{l=1}^{m-1} \beta_{jl} = 1$. Thus, regarding $\frac{\omega_{jj} G_j^* + \omega_{jm}\omega_{mm} G_m^*}{\omega_{jj}+\omega_{jm}\omega_{mm}}$ as a new G_j^* and using the induction assumption, we complete the proof.

B The Proof of $\mathbf{A} \geq 0$ and $\mathbf{A}\mathbf{1}_m = \mathbf{1}_m$

The proof can be immediately obtained from Theorem 2 in Dunson et al. (2007). Here we present a simpler proof. First, we have

$$\begin{aligned}
\mathbf{A} &= [\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}) {-} \mathbf{\Omega}]^{-1} \mathrm{dg}(\mathbf{\Omega}) \\
&= [\mathbf{I}_m - (\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}))^{-1}\mathbf{\Omega}]^{-1} (\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}))^{-1} \mathrm{dg}(\mathbf{\Omega}) \\
&= \left\{ \sum_{t=0}^{\infty} ((\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}))^{-1}\mathbf{\Omega})^t \right\} (\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}))^{-1} \mathrm{dg}(\mathbf{\Omega}) \\
&\geq 0
\end{aligned}$$

due to that $\mathbf{\Omega} \geq 0$, $\mathbf{\Omega}\mathbf{1}_m = \mathbf{1}_m$ and $(\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}))^{-1} = \mathrm{diag}(1/(1{+}\omega_{11}), \cdots, 1/(1{+}\omega_{mm})) \geq 0$. Second, it is direct to obtain $\mathbf{A}\mathbf{1}_m = \mathbf{1}_m$ from the fact that $\mathrm{dg}(\mathbf{\Omega})\mathbf{1}_m = [\mathbf{I}_m {+} \mathrm{dg}(\mathbf{\Omega}) {-} \mathbf{\Omega}]\mathbf{1}_m$.

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.

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.

Brown, P. J., T. Fearn, and M. Vannucci (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association 96*, 398–408.

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

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.

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.

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

Hannah, L. A., D. M. Blei, and W. B. powell (2010). Dirichlet process mixtures of generalized linear models. In *The Thirteenth International Conference on AI and Statistics*.

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

Lin, D., E. Grimson, and J. Fisher (2010). Construction of dependent Dirichlet processes based on Posson processes. In *Advances in Neural Information Processing Systems 23*.

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.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society Series B 66*(3), 735–749.

Petrone, S., M. Guindani, and A. E. Gelfand (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society, B 71*(4), 755–782.

Raftery, A. E., D. Madigan, and D. Hoeting (1997). Bayesian model averaging for linear regression. *Journal of the American Statistical Association 92*, 179–191.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (Second ed.). Philadelphia: SIAM.

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

Shahbaba, B. and R. Neal (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research 10*(2), 1829–1850.

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. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

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)*.

Zhang, Z., G. Dai, and M. I. Jordan (2010). Matrix-variate Dirichlet process mixture models. In *The Thirteenth International Conference on AI and S-tatistics (AISTATS)*.