

# More Nonparametric Bayesian Models for Biostatistics

Peter Müller and Fernando Quintana

January 19, 2009

## Abstract

In this companion chapter to Dunson (2009) we discuss and extend some of the models and inference approaches introduced there. We elaborate on the discussion of random partition priors implied by the Dirichlet process (DP). We review some additional variations of dependent DP (DDP) models and we review in more detail the PT prior used briefly in Dunson (2009). Finally, we review variation of DP models for data formats beyond continuous responses.

## 1 Introduction

Dunson (2009) introduced many interesting applications of nonparametric priors for inference in biomedical problems. The focus of the discussion was on Dirichlet process (DP) priors and variations. While the DP prior defines a probability model for a (discrete) random probability distribution  $G$ , the primary objective of inference in many recent applications is not inference on  $G$ . Instead many applications of the DP prior exploit the random partition of the Polya Urn scheme that is implied by the configuration of ties among the random draws from a discrete measure with DP prior. When the emphasis is on inference for the clustering, it is helpful to recognize the DP as a special case of more general clustering models. In particular we will review the product partition (PPM) and species sampling models (SSM). We discuss these models in section 2. A definition and discussion of the SSM as a random probability measure also appears in Lijoi and Prünster (2009, Section 3.4). Another useful characterization of the DP is as a special case of the Polya tree (PT) prior. A particularly attractive feature of PT priors is the possibility to model absolutely continuous distributions. In section 3 we will define PT priors and discuss specific algorithms to implement inference. For more discussion of the PT prior see also Lijoi and Prünster (2009, Section 4.2). In Section 4 we discuss more variations of the dependent DP (DDP) models. In Section 5 we review some examples of DP priors for biostatistical applications that involve non-continuous data. Finally, in Section 6 we discuss implementation details. We show some example R code using the popular R package *DPpackage* to implement nonparametric Bayesian inference.

## 2 Random Partitions

Let  $[n] = \{1, \dots, n\}$  denote a set of experimental units. A partition is a family of subsets  $\rho_n = \{S_1, \dots, S_k\}$  with  $\bigcup S_j = [n]$  and  $S_j \cap S_\ell = \emptyset$  for all  $j \neq \ell$ . The partitioning subsets  $S_j$  define clusters of experimental units. Often it is convenient to describe a partition by cluster membership indicators  $s_i = j$  if  $i \in S_j$ . We use notation  $n_{nj} = |S_j|$  and  $\mathbf{n}_n = (n_1, \dots, n_k)$  to denote the cluster sizes and  $k_n = |\rho_n|$  to denote the number of clusters. When the sample size  $n$  is understood from the context we drop the subindex  $n$  in  $\rho$ ,  $n_j$ ,  $\mathbf{n}$  and  $k$ .

Several probability models  $p(\rho_n)$  are introduced in the literature. For an extensive recent review of probability models and Bayesian inference for clustering, see, for example, Quintana (2006). Popular models include the product partition models (PPM), species sampling models (SSM), model based clustering (MBC) and Voronoi tessellations. The PPM (Hartigan; 1990; Barry and Hartigan; 1993) requires a cohesion function  $c(S_j) \geq 0$  (see an example below). A PPM for a partition  $\rho$  and data  $y$  is defined as

$$p(\rho) \propto \prod c(S_i) \quad \text{and} \quad p(y \mid \rho) = \prod_{j=1}^k p_j(y_{S_j}). \quad (1)$$

Here,  $p_j$  is any sampling model for the observations in the  $j$ -th cluster. Model (1) is conjugate. The posterior  $p(\theta \mid y)$  is again in the same product form.

Most recent applications of such models in biomedical applications use the special case of DP priors (Ferguson; 1973; Antoniak; 1974). The DP implicitly defines a probability model on  $p(\rho_n)$  by defining a discrete random probability measure  $G$ . An i.i.d. sample  $x_i \sim G$ ,  $i = 1, \dots, n$ , includes with positive probability ties among the  $x_i$ . Let  $x_j^*$ ,  $j = 1, \dots, k \leq n$  denote the unique values of  $x_i$  and define clusters  $S_j = \{i : x_i = x_j^*\}$ . By defining the probability of ties the DP prior has implicitly defined  $p(\rho_n)$ . The probability model is known as the Polya urn and is a special case of the PPM with cohesions  $c(A) = M \times (|A| - 1)!$  (Quintana and Iglesias; 2003; Dahl; 2003).

A typical recent use of the DP random partition model in biostatistics applications appears in Dahl and Newton (2007) who use a clustering model to improve power for multiple hypothesis tests by pooling tests within clusters. Dahl (2006) describes a similar model with focus on inference for the clustering only. Tadesse et al. (2005) combine inference on clustering, again based on the DP prior, with variable selection to identify a subset of genes whose sampling model is defined by the clustering.

Another class of random partition models are the species sampling models (SSM) (Pitman; 1996). An SSM defines a probability model  $p(\rho)$  that depends on  $\rho$  only indirectly through the cardinality of the partitioning subsets,  $p(\rho) = p(|S_1|, \dots, |S_k|)$ . The SSM can be alternatively characterized by a sequence of predictive probability functions (PPF) that describe how individuals are sequentially assigned to either already formed clusters or to start new ones. Let  $n_j = |S_j|$  and  $\mathbf{n}_n = (n_1, \dots, n_k)$ . The PPFs are the probabilities  $p_{nj}(\mathbf{n}_n) = Pr(s_{n+1} = j \mid \rho_n)$ ,  $j = 1, \dots, k_n + 1$ . Compare Theorem 11 in Lijoi and Prünster (2009, Section 3.4). The opposite is not true. Not every sequence of PPF's characterizes an SSM. Pitman (1996) states the conditions. Let  $\mathbf{n}^{j+}$  denote  $\mathbf{n}$  with  $n_j$  incremented by one.

Essentially the PPF's have to arise as  $p_{nj}(\mathbf{n}_n) = p(\mathbf{n}_{n+1}^{j+})/p(\mathbf{n}_n)$ , where  $p(\mathbf{n}_n)$  is a probability measure on  $\bigcup_n \{\mathbf{n}_n\}$  that is symmetric in its arguments. An important implication is that  $p(\mathbf{n}_n)$  has to arise as the marginal of  $p(\mathbf{n}_{n+1})$ , i.e.,  $p(\mathbf{n}_n) = \sum_{j=1}^{k_n+1} p(\mathbf{n}_{n+1}^{j+})$ . The probability model  $p(\mathbf{n}_n)$  is known as the exchangeable partition probability function. See Lijoi and Prünster (2009, Section 3.2) for more discussion. Again, the random clustering implied by the DP model is a special case of an SSM. I.e., the random partition model implied by the DP is a special case of both, PPM and SSM.

Model based clustering (Banfield and Raftery; 1993; Dasgupta and Raftery; 1998) implicitly defines a probability model on clustering by assuming a mixture model

$$p(y_i | \eta, k) = \sum_{j=1}^k \tau_j f_j(y_i | \theta_j),$$

where  $\eta = (\theta_1, \dots, \theta_k, \tau_1, \dots, \tau_k)$  are the parameters of a size  $k$  mixture model. Together with a prior  $p(k)$  on  $k$ , the mixture implicitly defines a probability model on clustering. Consider the equivalent hierarchical model

$$p(y_i | s_i = j, k, \eta) = f_j(y_i | \theta_j) \quad \text{and} \quad Pr(s_i = j | k, \eta) = \tau_{kj}. \quad (2)$$

The implied posterior distribution on  $(s_1, \dots, s_n)$  and  $k$  implicitly defines a probability model on  $\rho_n$ . Richardson and Green (1997) develop posterior simulation strategies for mixture of normal models. Green and Richardson (1999) discuss the relationship to DP mixture models.

Heard et al. (2006) model gene expression profiles with a hierarchical clustering model. The prior model on the random partition is an example of model based clustering with a mixture of regression models, a uniform prior on the number of clusters and a Dirichlet prior for cluster membership probabilities.

### 3 Polya Trees

Lavine (1992, 1994) proposed Polya trees (PT) as a useful nonparametric Bayesian prior for random probability measures. In contrast to the DP, an appropriate choice of the PT parameters allows the analyst to specify absolutely continuous distributions. In the following discussion we briefly review the definition of the PT model, and give explicit implementation details for posterior inference under the PT. See also Lijoi and Prünster (2009, Section 4.2) for more discussion of the definition. The definition starts with a nested sequence  $\Pi = \{\pi_m, m = 1, 2, \dots\}$  of partitions of the sample space  $\Omega$ . Without loss of generality, we assume that the partitions are binary. We start with a partition  $\pi_1 = \{B_0, B_1\}$  of the sample space,  $\Omega = B_0 \cup B_1$ , and continue with nested partitions defined by  $B_0 = B_{00} \cup B_{01}$ ,  $B_1 = B_{10} \cup B_{11}$ , etc. Thus the partition at level  $m$  is  $\pi_m = \{B_\epsilon, \epsilon = \epsilon_1 \dots \epsilon_m\}$ , where  $\epsilon$  are all binary sequences of length  $m$ . A PT prior for a random probability measure  $G$  is defined by beta-distributed random branching probabilities. Let  $Y_{\epsilon_0} \equiv G(B_{\epsilon_0} | B_\epsilon)$ , and let  $\mathcal{A} \equiv \{\alpha_\epsilon\}$  denote a sequence of nonnegative numbers, one for each partitioning subset. If  $Y_{\epsilon_0} \sim \text{Beta}(\alpha_{\epsilon_0}, \alpha_{\epsilon_1})$  then we say that  $G$  has a PT prior,  $G \sim \text{PT}(\Pi, \mathcal{A})$ .

The parameters  $\alpha_\epsilon$  are usually chosen as  $\alpha_\epsilon = cm^r$  for level  $m$  subsets. For  $r = 2$  the random probability measure  $G$  is a.s. absolutely continuous. With  $r = -1/2$  the PT reduces to the DP as a special case. The partitioning subsets  $B_\epsilon$  can be chosen to achieve a desired prior mean  $G^*$ . Let  $q_{mk} = G^{*-1}(k/2^m)$ ,  $k = 0, \dots, 2^m$ , denote the inverse c.d.f. under  $G^*$  evaluated at dyadic fractions. If  $\alpha_{\epsilon_0} = \alpha_{\epsilon_1}$ , for example  $\alpha_\epsilon = cm^r$ , and the dyadic quantile sets  $[q_{mk}, q_{m,k+1})$  are used as the partitioning subsets  $B_\epsilon$  then  $E(G) = G^*$ . Alternatively the prior mean can be fixed to  $G^*$  by choosing  $\alpha_{\epsilon_0}/(\alpha_{\epsilon_0} + \alpha_{\epsilon_1}) = G^*(B_{\epsilon_0} | B_\epsilon)$  for any choice of the nested partitions  $\Pi$ .

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

$$\alpha'_\epsilon = \begin{cases} \alpha_\epsilon & \text{if } x_1 \notin B_\epsilon \\ \alpha_\epsilon + 1 & \text{if } x_1 \in B_\epsilon \end{cases} \quad (3)$$

The general case with a sample of size  $n > 1$  follows by induction.

The result (3) can be used to implement exact posterior predictive simulation, i.e., simulation from  $p(x_{n+1} | x_1, \dots, x_n)$ . In words, we “drop” a ball down (well, really up) the Polya tree. Starting with  $(B_0, B_1)$  at the root we generate the random probabilities  $(Y_{\epsilon_0}, Y_{\epsilon_1} = 1 - Y_{\epsilon_0})$  for picking the two nested partitions  $B_{\epsilon_0}$  and  $B_{\epsilon_1}$  at the next level. Recall that  $Y_{\epsilon_0} = P(x \in B_{\epsilon_0} | x \in B_\epsilon)$  and  $Y_{\epsilon_0} \sim \text{Beta}(\alpha'_{\epsilon_0}, \alpha'_{\epsilon_1})$ . Going down the tree we run into some good luck. At some level  $m$  we will drop the ball into a subset  $B_\epsilon$ ,  $\epsilon = \epsilon_1 \epsilon_2 \dots \epsilon_m$ , that does not contain any data point. From level  $m$  onwards, dropping the ball proceeds as if we had no data observed. Thus we can generate the posterior predictive draw from the base measure  $G^*$ , restricted to  $B_\epsilon$ . The following algorithm summarizes posterior predictive simulation of  $x_{n+1} \sim p(x_{n+1} | x_1 \dots x_n)$ :

### Algorithm 1

1. *Initialize:*  $\epsilon = \emptyset$  (nil).
2. *Iteration:* Loop over  $m = 1, 2, \dots$ :
  - (a) *Posterior PT Parameters:* Find  $n_{\epsilon_0} = \sum I(x_i \in B_{\epsilon_0})$  and  $n_{\epsilon_1} = \sum I(x_i \in B_{\epsilon_1})$ , the number of data points in the two partitioning subsets for  $B_\epsilon = B_{\epsilon_0} \cup B_{\epsilon_1}$ . Let  $\alpha'_{\epsilon_0} = \alpha_{\epsilon_0} + n_{\epsilon_0}$ , and same for  $\alpha'_{\epsilon_1}$ .
  - (b) *Generate Random Branching Probability:* Generate  $Y_{\epsilon_0} \sim \text{Beta}(\alpha'_{\epsilon_0}, \alpha'_{\epsilon_1})$ , and set  $\epsilon_m \sim \text{Bernoulli}(1 - Y_{\epsilon_0})$ .
3. *Stop of the Recursion:* Stop the iteration over  $m$  for the smallest  $m^*$  such that  $n_\epsilon = 0$  at  $m = m^*$ .
4. *Generate  $x_{n+1}$ :* Draw  $x_{n+1} \sim G^*(x_{n+1} | B_\epsilon)$ .

We can use the same algorithm to generate a prior predictive, i.e., marginal samples  $x_i \sim G$ , i.i.d., with  $G \sim \text{PT}(\Pi, \mathcal{A})$ .

1. Generate  $x_1 \sim G^*$
2. Iterate over  $i = 2, \dots, n$ :  
Use Algorithm 1 to generate  $x_i \sim p(x_i | x_1, \dots, x_{i-1})$ .

A minor variation of the algorithm computes the posterior mean  $E(G | x_1, \dots, x_n)$ . Let  $x = x_1, \dots, x_n$  denote the data. Consider some maximum level  $M$ , say  $M = 10$ . For all levels  $m = 1, \dots, M$  compute  $\bar{Y}_{\epsilon_1 \epsilon_2 \dots \epsilon_{m-1} 0} = E(Y_{\epsilon_1 \epsilon_2 \dots \epsilon_{m-1} 0} | x) = \alpha'_{\epsilon_1 \epsilon_2 \dots 0} / (\alpha'_{\epsilon_1 \epsilon_2 \dots 0} + \alpha'_{\epsilon_1 \epsilon_2 \dots 1})$ . Record  $\bar{Y}_{\epsilon_1 \epsilon_2 \dots \epsilon_{m-1} 1} = 1 - \bar{Y}_{\epsilon_1 \epsilon_2 \dots \epsilon_{m-1} 0}$  as the complement to one. Computing  $\bar{Y}_\epsilon$  is most elegantly implemented as a recursion. Let  $\epsilon = \epsilon_1 \dots \epsilon_M$  denote a dyadic number,  $\epsilon \in [0, 1]$ . We find

$$E(G(\epsilon) | x_1, \dots, x_n) \approx \prod_{m=1}^M \bar{Y}_{\epsilon_1 \epsilon_2 \dots \epsilon_m} \equiv \bar{G}(\epsilon).$$

Here  $G(\cdot)$  is the p.d.f. for the random measure  $G$ .

To generate random draws  $G \sim p(G | x_1, \dots, x_n)$  proceed as above replacing  $\bar{Y}_\epsilon$  by  $Y_{\epsilon 0} \sim \text{Beta}(\alpha'_{\epsilon 0}, \alpha'_{\epsilon 1})$ . Let  $G(\epsilon) = \prod_{m=1}^M Y_{\epsilon_1 \epsilon_2 \dots \epsilon_m}$ . Plotting  $G$  against  $\epsilon$  shows a random posterior draw of  $G$ . Plotting multiple draws  $G_j$ ,  $j = 1, 2, \dots, J$  in the same figure illustrates posterior uncertainty on the random measure.

An important simplification for applications is achieved by restricting the PT prior to finitely many levels, e.g.  $m \leq 10$ . Actual inference differs little, but the required computational effort is greatly reduced. Posterior predictive draws and posterior estimated densities can still be carried out exactly, as outlined above. Nonparametric Bayesian inference under the PT prior and the finite PT prior for some important models is implemented in the public domain R package *DPpackage*. See Section 6 for an example.

Applications of PT models in biomedical problems are less common than the widely used DP. The limited use of PT priors for nonparametric Bayesian data analysis is due in part to the awkward sensitivity of posterior inference to the choice of the partitioning subsets. Consider a model for density estimation,  $x_i \sim G$ , i.i.d.,  $i = 1, \dots, n$ , with a PT prior for the unknown distribution,  $G \sim \text{PT}(\mathcal{A}, \mathcal{P})$ . The posterior estimate  $\bar{G} = E(G | x)$  shows characteristic discontinuities at the boundaries of the partitioning subsets. An exception is the special case of the DP, i.e., when  $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$ . In that case the probability model  $p(G)$  remains invariant under any change of the partitioning sequence. Several authors have proposed variations of the PT prior model to mitigate this undesirable feature of posterior inference. Hanson and Johnson (2002) and Hanson (2006) defined mixture of Polya trees, with the mixture being with respect to the centering measure  $G^*$ . Paddock et al. (2003) introduced an explicit perturbation of the partition boundaries. The posterior dependence on the partition boundaries is less of an issue when the focus of the inference is not on the density itself. For example, Branscum et al. (2008) develop inference for ROC curves with a PT prior for the distribution of the recorded measurements for true positives and negatives. Hanson and Yang (2007) use PT priors for survival data.

## 4 More DDP Models

Many applications in Biostatistics involve hierarchical models across different subpopulations and naturally lead to models that include several random probability measures. Appropriate nonparametric probability models require modeling of dependent families of random distributions. One of the most commonly used models to achieve this aim are variations of dependent DP (DDP) models. Since the first proposal of DDP models in MacEachern (1999) many authors have developed variations and implementations for specific problems. Some are discussed in (Dunson; 2009, sections 5 and 6). In this section we review more related models that find use in Biostatistics.

### 4.1 The ANOVA DDP

DDP models define probability models for families of random probability measures  $\{G_x, x \in X\}$  in such a way that marginally each  $G_x$  follows a DP prior,  $G_x \sim \text{DP}(c, G_x^*)$ . Let  $G_x = \sum_h w_{xh} \delta_{\mu_{xh}}$ . By the definition of the DP prior the point masses  $\mu_{xh}$ ,  $h = 1, 2, \dots$ , are i.i.d. draws from a base measure, and the weights are generated by the stick breaking process based on independent beta random variables. See, for example, Dunson (2009), equation (2) for a definition of the DP prior. The DDP introduces the desired dependence of the random measures  $G_x$  across  $x$  by defining dependence of the locations  $\mu_{xh}$  and/or the weights  $w_{xh}$  across  $x$ . The independence across  $h$  remains untouched. Alternative variations of the DDP model differ in the definition of the dependence structure. Consider the case when  $X$  is categorical, for example, when  $x$  indexes different studies, different subpopulations and/or centers. Perhaps the easiest definition of dependence across  $\{\mu_{xh}, x \in X\}$  for categorical factors  $x$  is the ANOVA model. Without loss of generality assume  $x = (u, v)$  is bivariate with  $u \in \{0, \dots, n_u\}$  and  $v \in \{0, \dots, n_v\}$  referring to two factors, for example,  $u$  might be treatment history and  $v$  might be an indicator for a relevant molecular marker. We can define a DDP model by assuming  $\mu_{xh} = M_h + A_{uh} + B_{vh}$  where  $(M_h, A_{uh}, B_{vh}, u = 0, \dots, n_u, v = 0, \dots, n_v)$  are overall mean and main effects in an ANOVA model. For identifiability we fix  $A_{0h} = B_{0h} = 0$ . The weights  $w_{hx}$  are assumed constant across  $x$ ,  $w_{hx} \equiv w_h$ . This defines the ANOVA DDP of De Iorio et al. (2004).

An interesting application of the ANOVA DDP to modeling survival data appears in De Iorio et al. (2008). The model implements non-parametric regression for event time data without relying on restrictive assumptions like proportional or additive hazards. In particular, inference allows survival functions to cross.

Figures 1 and 2 show the estimated survival functions for data from a study of infant and childhood mortality in Colombia (Somoza; 1980). A questionnaire was administered to a sample of women between the ages of 15 and 49 eliciting their maternity history, educational level, age, union status and information on the sex, date of birth and survival status (at the date of interview) of all their children and, if applicable, age at death. We consider data on a sub-sample of 1437 children (corresponding to the oldest child for each mother). The response of interest is the survival time (in years) of a child at the time of the maternal interview. The covariates of interest are: gender (male/female), birth cohort (1=1941-59;

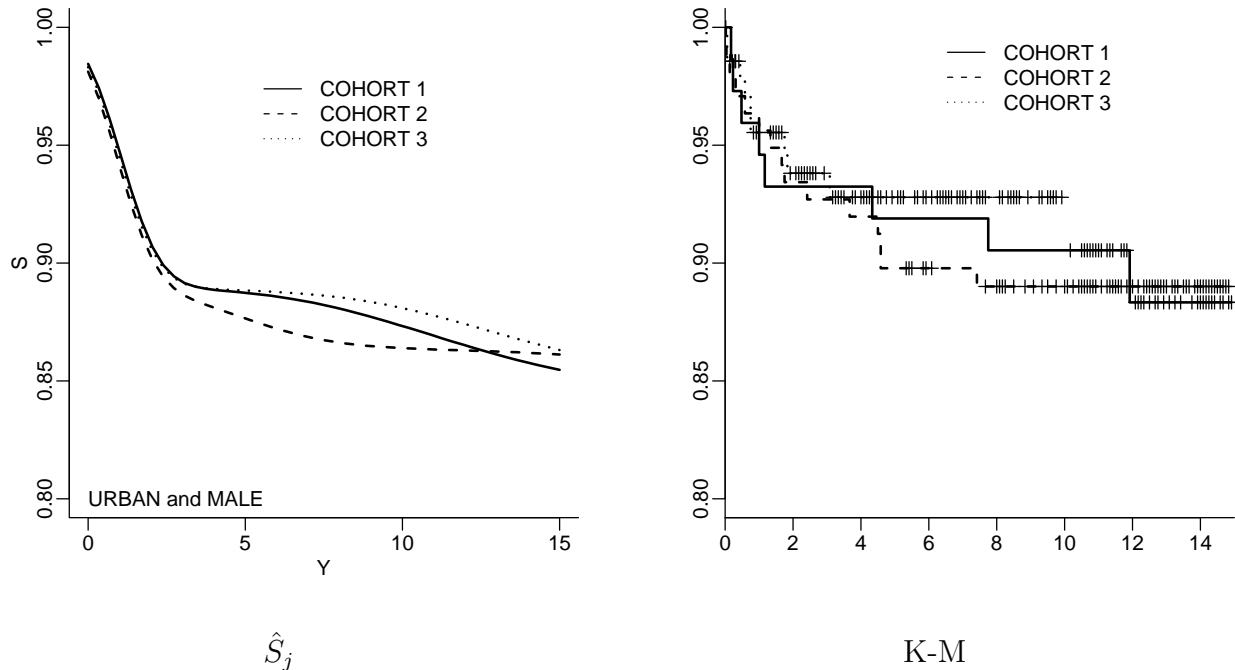


Figure 1: Colombian children data. Posterior survivor functions for urban male children from the three birth cohorts, under the DDP ANOVA model (left panel) and raw estimates (KM) from the data (right panel). The solid line corresponds to children in the first birth cohort, the dashed line represents a child in the second birth cohort and the dotted line refers to children in the first birth cohort.

2=1960-67; 3=1968-76) and a binary variable indicating whether a child was born in a rural area (yes/no). Around 87% of the observations in the dataset were censored. The original research was conducted to investigate how patterns of childhood mortality have changed over time. Also of interest are urban/rural and gender differences.

Inference under the DDP ANOVA model is implemented in the R packages *DPpackage* and *ddpanova*. See Section 6 for an example and additional details.

## 4.2 Classification with DDP Models

A minor extension of ANOVA DDP models allows their use for nonparametric classification. Let  $x \in X \equiv \{1, \dots, k\}$  index  $k$  subpopulations and assume a semi-parametric hierarchical model for outcomes  $y$  across the  $k$  subpopulations. Let  $y_i$  denote the outcome for the  $i$ -th experimental unit and let  $x_i \in X$  denote the known class label, i.e., the index of the subpopulation containing the  $i$ -th patient,  $i = 1, \dots, n$ . For example, De la Cruz-Mesía et al. (2007) consider data for pregnant women. For each woman  $y_i = (y_{i1}, \dots, y_{in_i})$  are repeated measurements of a hormone ( $\beta$ -HCG) which shows dramatic changes during pregnancy. Pregnancies are classified as normal  $x_i = 0$  vs. not normal  $x_i = 1$  (spontaneous abortions or other types of adverse pregnancy outcomes). We consider the classification problem of

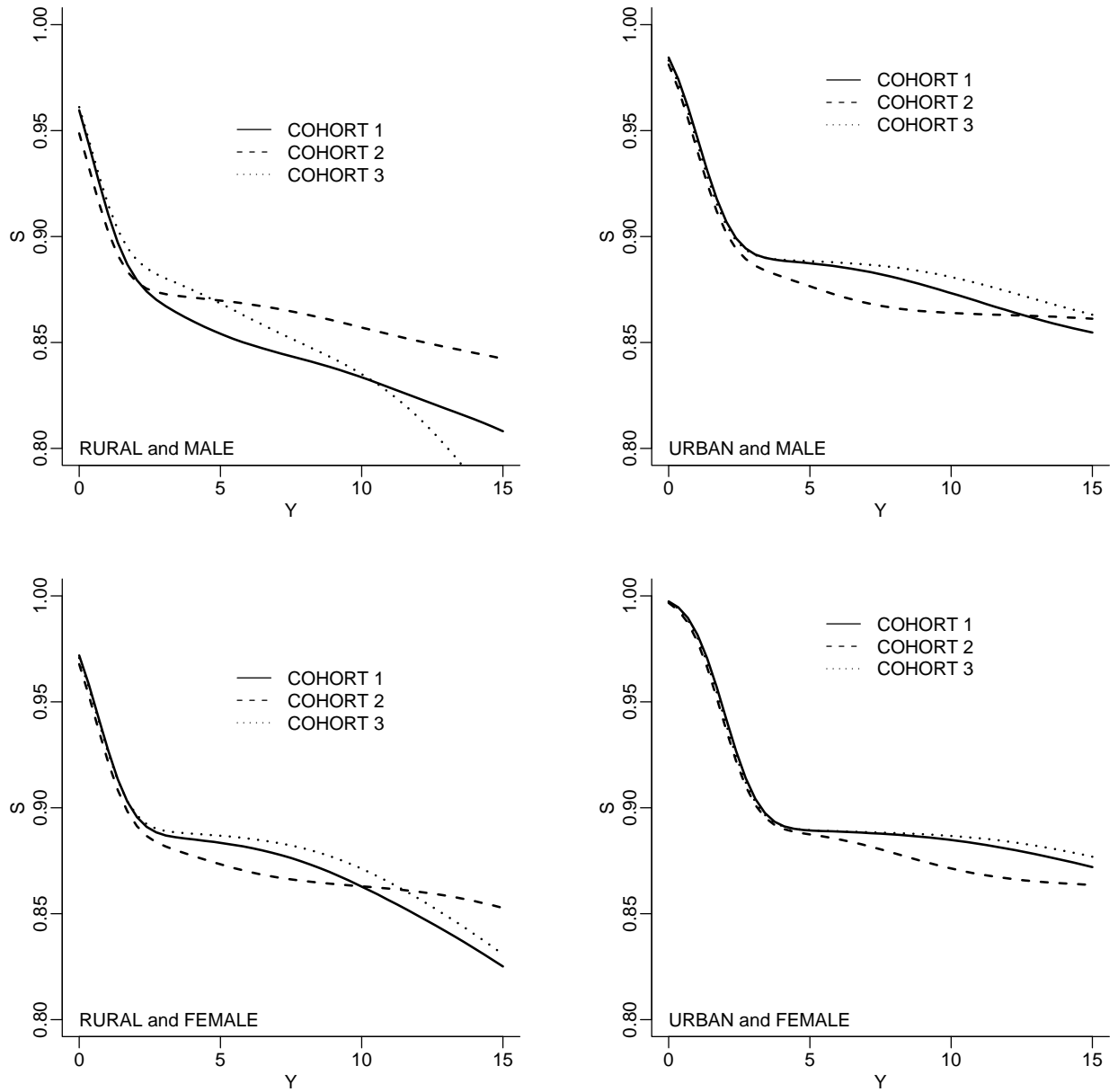


Figure 2: Colombian children data. Posterior survivor functions for children from the three birth cohorts, arranged by rural versus urban and male versus female. The solid line corresponds to children in the first birth cohort, the dashed line represents a child in the second birth cohort and the dotted line refers to children in the first birth cohort.

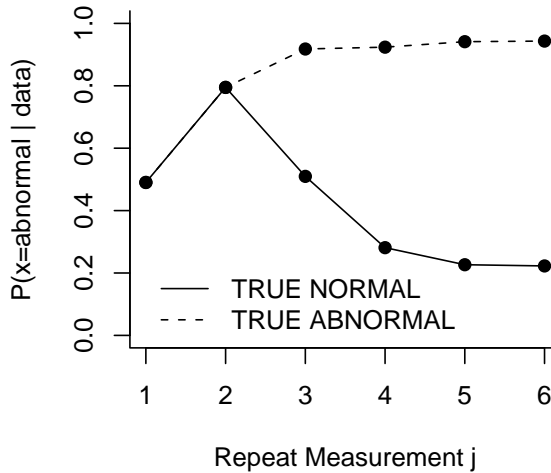


Figure 3: Pregnancy data: sequentially updated classification probabilities  $p(x_{n+1} = 1 \mid y_{n+1,1}, \dots, y_{n+1,j}, data)$  for a future case. The probabilities are plotted against  $j$ . The solid (dashed) line is for a future case with normal (abnormal) pregnancy.

predicting (unknown)  $x_{n+1}$  for a future patient,  $i = n + 1$  conditional observed responses  $y_{n+1}$ , i.e.,

$$p(x_{n+1} \mid y_{n+1}, y_1, x_1, \dots, y_n, x_n).$$

Let  $p(y_i \mid x_i = x) = \int p(y_i \mid \theta_i) dG_x(\theta_i)$  be a semiparametric sampling model for outcomes in group  $x$ . See for example De la Cruz-Mesía et al. (2007) for details of the model for the pregnancy data. Marginally, for each  $x \in X$  we use a DP mixture model, i.e., we assume a DP prior for  $G_x$ . The submodels for all  $x$  are combined into one encompassing hierarchical model by linking the marginal DP priors through an ANOVA DDP across  $x$ . Finally the model is completed by assuming a marginal distribution  $p(x_i)$  for  $x_i$ , for example  $x_i \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$ . The ANOVA DDP model defines  $p(y_1, \dots, y_{n+1} \mid x_1, \dots, x_{n+1})$ . Together with the marginal model  $p(x_1, \dots, x_{n+1}) = \prod p(x_i)$  we can use Bayes theorem to derive the desired  $p(x_{n+1} \mid y_{n+1}, x_1, y_1, \dots, x_n, y_n)$ . One of the attractions of this principled model-based approach is the possibility for coherent sequential updating. Assume  $y_i = (y_{i1}, \dots, y_{in_i})$  are repeated measurement data as in the pregnancy example. The classification probability  $p(x_{n+1} \mid y_{n+1,1}, \dots, y_{n+1,j}, y_1, x_1, \dots, y_n, x_n)$  can be sequentially updated and reported as increasingly more data becomes available for  $j = 1, \dots, n_i$ . Figure 3 shows sequentially updated classification probabilities  $p(x_{n+1} = 1 \mid y_{n+1,1}, \dots, y_{n+1,j}, y_1, x_1, \dots, y_n, x_n)$  for the pregnancy example. Probabilities are plotted against  $j$  for two hypothetical future patients. The first patient (solid line) is a woman with a truly normal pregnancy. The second case

(dashed line) is a truly abnormal pregnancy. See De la Cruz-Mesía et al. (2007) for details of the simulation. Note how the two classification probabilities start to diverge from the third repeat measurement onwards.

## 5 Other Data Formats

Many discussions of DP models, including most of the discussion in Dunson (2009) focus on continuous outcomes. But many biomedical data analysis problems involve other data formats. We briefly review two applications of nonparametric Bayesian inference to categorical and binary data.

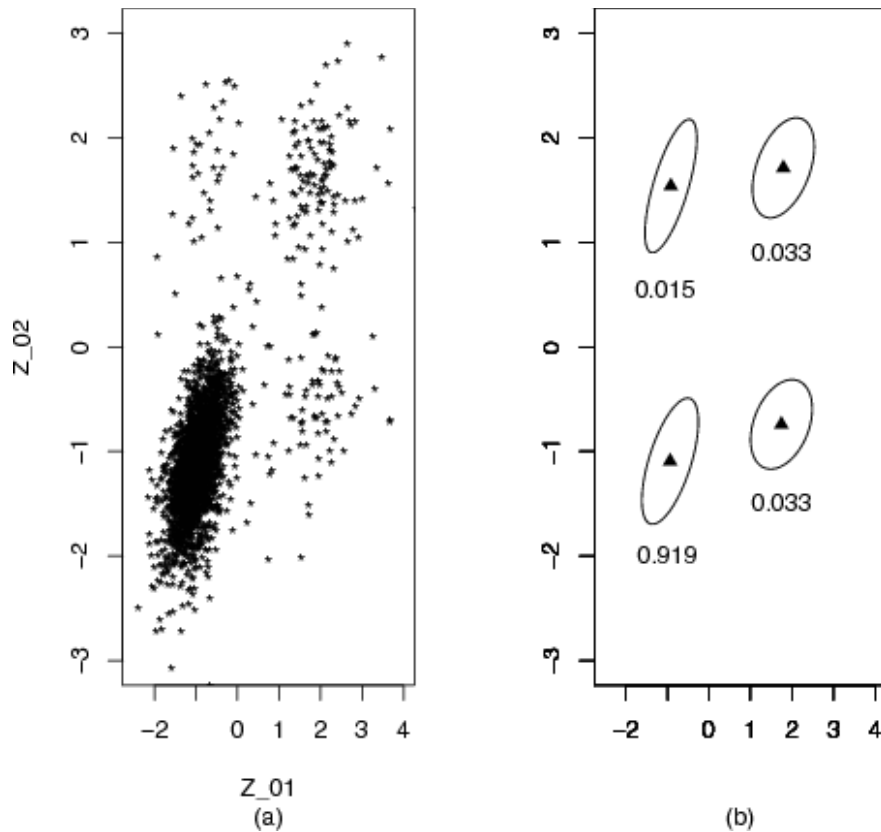


Figure 4: Interrater agreement data: Panel (a) plots draws from the latent probit scores (with a DP mixture of normal prior). Panel (b) shows one posterior draw for the moments of the mixture of normal terms (location and scale are shown as triangle and ellipse), together with the corresponding weights (number below the ellipse). Notice the varying degree of polychoric correlation across scores.

Kottas et al. (2005) propose nonparametric Bayesian inference for multivariate ordinal data, for example the rating of the extent of tumor invasion by two raters. Tumor invasion

is coded on an ordinal scale from none to extensive invasion. A feature of the data example reported in Kottas et al. (2005) is that the raters tend to agree on extreme cases, but less so on intermediate cases. This makes it inappropriate to use a bivariate ordinal probit model, as described, for example in Johnson and Albert (1999). Instead Kottas et al. (2005) propose an ordinal probit model based on a DP mixture of normal distributions for the latent variable. The mixture of normal model allows us to formalize the notion of varying degrees of interrater agreement across scores. Let  $\Sigma_j$  denote the variance-covariance matrix of the  $j$ -th term in the mixture of normal model for the latent variables. The correlation of the latent scores is known as polychoric correlation. We refer to the correlation that is implied by  $\Sigma_j$  for each term of the mixture as *local polychoric correlation*. The use of different  $\Sigma_j$  for each term in the mixture allows for varying degrees of interrater agreement, as desired. Figure 4 shows imputed ordinal probit scores and summaries of the mixture of normal model for the interrater agreement example. The plotted variables  $z_{01}$  and  $z_{02}$  are the latent ordinal probit scores. The observed scores are defined by thresholds on the latent scores. See Kottas et al. (2005) for details. For normal cases (with low scores) the two raters are in strong agreement, i.e., high polychoric correlation. For extreme cases the strength of agreement is considerably less.

Quintana et al. (2008) discusses semi-parametric Bayesian inference for binary sequences of indicators of loss of heterozygosity (LOH). Nonparametric inference for sequences of binary indicators subject to a partial exchangeability assumptions as defined in Quintana and Müller (2004). The exchangeability assumption is order  $\ell$  exchangeability, i.e., the assumption that the probability model is invariant with respect to any permutation of the sequence that leaves the order  $\ell$  transition counts unchanged. Quintana and Newton (2000) show that any such distribution can be represented as a mixture of order  $\ell$  Markov chains. The mixture is with respect to the transition probabilities of the Markov chain. This is implemented in Quintana and Müller (2004) by assuming a non-parametric DP prior on the mixture measure of the transition probabilities. We allow different transition probabilities for each region of the chromosome. Regions are defined as sequences of 55 to 835 SNPs (single nucleotide polymorphism). See Quintana et al. (2008) for details. The transition probabilities of the binary Markov chain with states { no LOH, LOH } imply a limiting probability of LOH. Let  $\pi_{cj}$  denote this limiting probability for region  $j$  in chromosome  $c$ . We can map the posterior distribution on the transition probabilities into  $p(\pi_{cj} | data)$ . Figure 5 shows  $I_{cj} \equiv p(\pi_{cj} > 0.01 | data)$  by region  $j$ , arranged by chromosome  $c$ .

## 6 An R Package for Nonparametric Bayesian Inference

An important impediment for the wider use of nonparametric Bayesian models was the until recently limited availability of reliable public domain software. This is in particular true for biomedical applications, where the research focus is often on the application, and only limited resources are available for the development of problem-specific software. Also reproducibility is an increasingly important issue. The impact of research publications proposing new methods and approaches remains limited unless readers can reproduce inference and

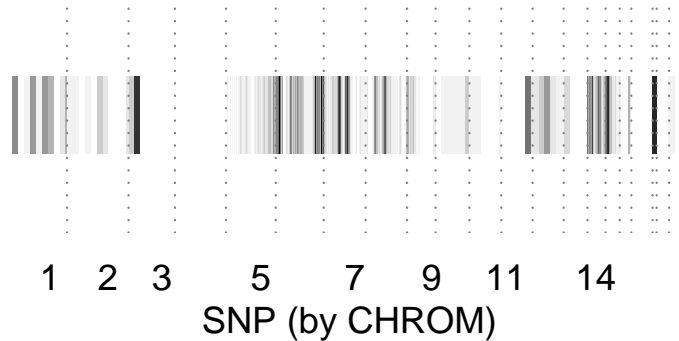


Figure 5: LOH data: grey shades show the probability of increased LOH for a given sample. SNPs are arranged by chromosomes indicated by vertical dashed lines. Darker grey shade indicates higher probability of increased LOH. The underlying model makes minimal assumptions beyond order  $\ell$  exchangeability.

implement the proposed methods for their problems.

A popular software platform for biomedical research, in particular for bioinformatics applications is the public domain R language (R Development Core Team; 2008). The R package *DPpackage* (Jara; 2007) implements inference for some of the models discussed in Dunson (2009), including DP mixture density estimation, PT priors for density estimation, non-parametric random effects models including generalized linear models. *DPpackage* is available from the package repository CRAN. We show a simple example.

Buta (1987) report velocities ( $y_i$ ) and radial position ( $x_i$ ) of galaxy NGC7531 at 323 different locations. We use the first 82 velocity measurements. The data are available as *galaxy* data set in *DPpackage*. The following R code implements density estimation using PT priors, a DP mixture model and random Bernstein polynomials (Petrone; 1999a,b). See Lijoi and Prünster (2009, Section 4.1) for a discussion of random Bernstein polynomials. Figure 6 shows a histogram of the galaxy data and the three density estimates based on DP mixture model, the PT prior and the random Bernstein polynomials.

```
library("DPpackage")

data(galaxy)          # Data
galaxy <- data.frame(galaxy,speeds=galaxy$speed/1000)
attach(galaxy)

state <- NULL        # MCMC parameters
nburn <- 10000; nsave <- 1000; nskip <- 50; ndisplay <- 10
mcmc <- list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay,
             tune1=0.15,tune2=1.1,tune3=1.1)
```

```

## tune parameters only needed for PTdensity()

## POLYA TREE
prior<-list(alpha=1,M=6) # Prior information
                        # Fitting the model
fit1 <- PTdensity(y=speeds,ngrid=1500,prior=prior,mcmc=mcmc,
                 state=state,status=TRUE)

## DIRICHLET PROCESS
                        # Prior information
prior <- list(a0=2,b0=4,m2=rep(20,1),s2=diag(100000,1),
             psiinv2=solve(diag(0.5,1)), nu1=4,nu2=4,tau1=1,tau2=100)
                        # Fitting the model
fit2 <- DPdensity(y=speeds,ngrid=1500,prior=prior,mcmc=mcmc,
                 state=state,status=TRUE)

## BERNSTEIN DIRICHLET PROCESS
                        # Prior information
prior <- list(aa0=2.01,ab0=1.01,kmax=1000,a0=1,b0=1)
                        # Fitting the model
fit3 <- BDPdensity(y=speeds,ngrid=1500,prior=prior,mcmc=mcmc,
                 state=state,status=TRUE)

rg <- range(c(fit1$dens,fit2$dens,fit3$fun))      ## Plots
hist(galaxy$speeds,xlim=c(5,40),nclass=30,ylim=rg)
lines(fit1$x1,fit1$dens,lty=1,lwd=2)
lines(fit2$x1,fit2$dens,lty=2,lwd=2)
lines(fit3$grid,fit3$fun,lty=3,lwd=2)

```

Inference for the ANOVA DDP is implemented in the *LDDPdensity()* function of *DP-package*. For the following example we generate  $n = 500$  observations each from an assumed simulation truth  $F_0^o = N(3, 0.8)$  and  $F_1^o = 0.6 N(1.5, 0.8) + 0.4 N(4, 0.6)$ . Based on the simulated data we estimate  $\{F_0, F_1\}$  under a DDP ANOVA prior for  $\{F_0, F_1\}$ . We assume  $F_x(y) = \int N(y; m, s) dG_x$  with  $\{G_0, G_1\} \sim$  DDP ANOVA. The DDP ANOVA model is based on an ANOVA model  $\mu_{xh} = M_h + xA_h$ , i.e., a main effect for  $x = 1$  and a common intercept  $M_h$ . Let  $d = (1, x)$  denote a design vector and let  $\beta_h = (M_h, A_h)$  denote the vector of ANOVA parameters. The DDP ANOVA model for  $\{F_0, F_1\}$  can be written as a mixture of DP model

$$y \mid x \sim \int N(y; \beta'd, \sigma_x) dG(\beta) \text{ with } G \sim DP(G^*, \alpha).$$

See De Iorio et al. (2004) for more details of the model. The data structure *prior* in the R code fragment below sets the parameters for the base measure  $G^*$ , a hyperprior on the residual variance  $\sigma_x$  and the total mass parameter  $\alpha$ . We assume  $1/\sigma_x^2 \sim Ga(\tau_1/2, \tau_2/2)$  and  $\alpha \sim Ga(a_0, b_0)$ . Let  $B \sim IW(\nu, A)$  denote an inverse Wishart random ( $q \times q$ ) matrix

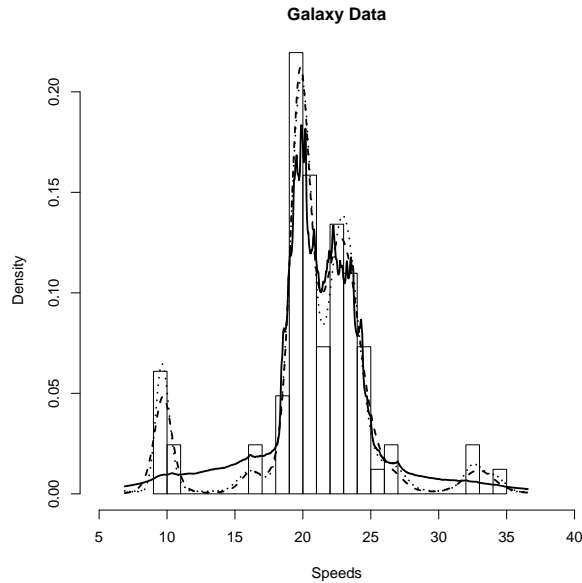


Figure 6: Galaxy data. The histogram shows the data with  $n = 82$  observations. The curves show the density estimate using a DP mixture model (dashed line), a PT prior (solid line) and a random Bernstein polynomial (dotted line).

$B$  with expectation  $E(B) = A^{-1}/(\nu - q - 1)$ . The base measure is  $G^*(\beta) = N(\mu_b, \Sigma_b)$  with conditionally conjugate hyperpriors  $\mu_b \sim N(m, S)$  and  $\Sigma_b \sim IW(\nu, \psi)$ . Figure 7 shows the density estimates  $\bar{F}_x = E(F_x | data)$ ,  $x = 0, 1$ . The estimates are produced by the R code below.

```
library(DPpackage)

## prepare simulated data: mixture of two normals
nobs <- 50;                               y1 <- rnorm(nobs, 3, .8)
y21 <- rnorm(nobs, 1.5, 0.8);             y22 <- rnorm(nobs, 4.0, 0.6)
y2 <- ifelse(runif(nobs) < 0.6, y21, y22); y <- c(y1, y2)

trt <- c(rep(0, nobs), rep(1, nobs)) # design matrix with a single factor
xpred <- rbind(c(1, 0), c(1, 1)) # design matrix for posterior predictive
m <- rep(0, 2); psiinv <- diag(1, 2); s <- diag(100, 2) # prior
prior <- list(a0=1, b0=1/5, nu=4, m=m, s=s, psiinv=psiinv, tau1=0.01, tau2=0.01)

## Fit the DDP ANOVA model
mcmc <- list(nburn=100, nsave=500, nskip=5, ndisplay=100)
fit <- LDDPdensity(y~trt, prior=prior, mcmc=mcmc, state=NULL, status=TRUE,
                  grid=seq(-1, 7, length=200), xpred=xpred)
```

```

## Estimated densities F_x (posterior predictive distributions)
plot(fit$grid,fit$dens[1,],type="l")
lines(fit$grid, dnorm(fit$grid, 3.0, 0.8), lty=2) # simulation truth
# ... and x0=(1,1)
plot(fit$grid,fit$dens[2,],type="l",xlab="Y",ylab="p",bty="l")
p2 <- 0.6*dnorm(fit$grid, 1.5, 0.8) + 0.4*dnorm(fit$grid, 4.0, 0.6)
lines(fit$grid,p2, lty=2)

```

The implementation of DDP ANOVA in *DDPpackage* is based on the R package *ddpanova* which can be downloaded from <http://odin.mdacc.tmc.edu/~pm>. The *ddpanova* package has additional options, including options for censored event time data.

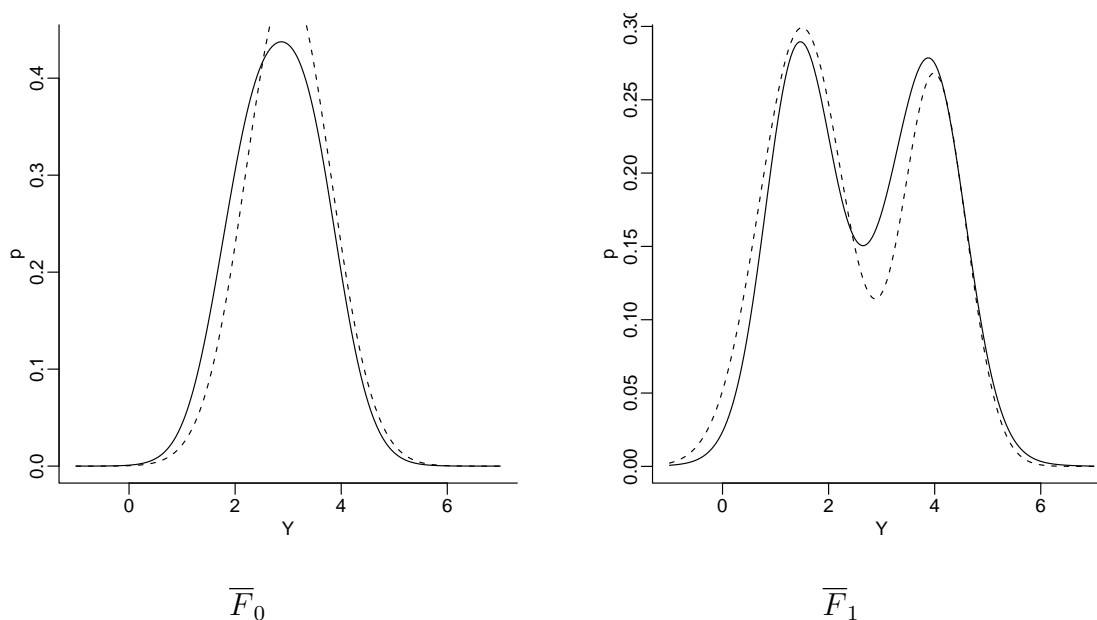


Figure 7: Simulated mixture of normal data. The figures show the simulation truth (dashed line) and the estimated distributions  $\bar{F}_x = E(F_x | data)$  (solid lines). The posterior means are equal to the posterior predictive distribution for a future observation  $\bar{F}_x = p(y_{n+1} | x_{n+1} = x, data)$ .

## 7 Discussion

We have discussed some extensions and elaborations of the models introduced in Dunson (2009). The discussion is by no means an exhaustive list of Bayesian nonparametric models. Keeping the theme of Dunson (2009) we have focused on models that find applications in biostatistics. This focus excluded, for example, interesting recent applications with spatial and spatio-temporal data.

Many more nonparametric Bayesian models and methods are reviewed in other chapters of this volume, including among many others, the Beta process, and the Indian buffet process.

## References

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems, *The Annals of Statistics* **2**: 1152–1174.
- Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**: 803–821.
- Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems, *Journal of the American Statistical Association* **88**: 309–319.
- Branscum, A. J., Johnson, W. O., Hanson, T. E. and Gardner, I. A. (2008). Bayesian semi-parametric roc curve estimation and disease diagnosis, *Statistics in Medicine* **to appear**.
- Buta, R. (1987). The structure and dynamics of ringed galaxies, iii, *The Astrophysical Journal. Supplement Series* **64**: 1–37.
- Dahl, D. B. (2003). Modal clustering in a univariate class of product partition models, *Technical Report 1085*, Department of Statistics, University of Wisconsin.
- Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model, in M. V. Kim-Anh Do, Peter Müller (ed.), *Bayesian Inference for Gene Expression and Proteomics*, Cambridge University Press, Cambridge, pp. 201–218.
- Dahl, D. B. and Newton, M. A. (2007). Multiple hypothesis testing by clustering treatment effects, *Journal of the American Statistical Association* **102**: 517–526.
- Dasgupta, A. and Raftery, A. E. (1998). Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering, *Journal of the American Statistical Association* **93**: 294–302.
- De Iorio, M., Johnson, W., Müller, P. and Rosner, G. (2008). A ddp model for survival regression, *Technical report*, M.D. Anderson Cancer Center.
- De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An anova model for dependent random measures, *Journal of the American Statistical Association* **99**: 205–215.
- De la Cruz-Mesía, R., Quintana, F. and Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers, *Applied Statistics* **56**(119–137).
- Dunson, D. (2009). Nonparametric Bayes applications to biostatistics, in N. Hjort, C. Holmes, P. Müller and S. Walker (eds), *Bayesian Nonparametrics in Practice*, CUP.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**: 209–230.
- Green, P. J. and Richardson, S. (1999). Modelling Heterogeneity with and without the Dirichlet Process, *Technical report*, University of Bristol, Department of Mathematics.

- Hanson, T. E. (2006). Inference for mixtures of finite Polya tree models, *Journal of the American Statistical Association* **101**: 1548–1564.
- Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of polya trees, *Journal of the American Statistical Association* **97**: 1020–1033.
- Hanson, T. and Yang, M. (2007). Bayesian Semiparametric Proportional Odds Models, *Biometrics* **63**(1): 88–95.
- Hartigan, J. A. (1990). Partition models, *Communications in Statistics, Part A – Theory and Methods* **19**: 2745–2756.
- Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves, *Journal of the American Statistical Association* **101**(473): 18–29.
- Jara, A. (2007). Applied bayesian non- and semi-parametric inference using dppackage, *Rnews* pp. 17–26.
- Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*, New York: Springer.
- Kottas, A., Müller, P. and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data, *Journal of Computational and Graphical Statistics* **14**: 610–625.
- Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling, *The Annals of Statistics* **20**: 1222–1235.
- Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling, *The Annals of Statistics* **22**: 1161–1176.
- Lijoi, A. and Prünster, I. (2009). Models beyond the Dirichlet process, in N. Hjort, C. Holmes, P. Müller and S. Walker (eds), *Bayesian Nonparametrics in Practice*, CUP.
- MacEachern, S. (1999). Dependent nonparametric processes, *ASA Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA.
- Paddock, S., Ruggeri, F., Lavine, M. and West, M. (2003). Randomised Polya Tree Models for Nonparametric Bayesian Inference, *Statistica Sinica* **13**: 443–460.
- Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials, *Canadian Journal of Statistics* **27**: 105–126.
- Petrone, S. (1999b). Random Bernstein polynomials, *Scandinavian Journal of Statistics* **26**: 373–393.

- Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme, in T. S. Ferguson, L. S. Shapely and J. B. MacQueen (eds), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Haywar, California: IMS Lecture Notes - Monograph Series, pp. 245–268.
- Quintana, F. A. (2006). A predictive view of Bayesian clustering, *Journal of Statistical Planning and Inference* **136**(8): 2407–2429.
- Quintana, F. A. and Iglesias, P. L. (2003). Bayesian Clustering and Product Partition Models, *Journal of The Royal Statistical Society Series B* **65**: 557–574.
- Quintana, F. A. and Newton, M. A. (2000). Computational aspects of Nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences, *Journal of Computational and Graphical Statistics* **9**(4): 711–737.
- Quintana, F. and Müller, P. (2004). Nonparametric Bayesian assessment of the order of dependence for binary sequences, *Journal of Computational and Graphical Statistics* **13**: 213–231.
- Quintana, F., Müller, P., Rosner, G. and Relling, M. (2008). A semiparametric Bayesian model for repeatedly repeated binary outcomes, *Applied Statistics* p. to appear.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.  
**URL:** <http://www.R-project.org>
- Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, Series B* **59**: 731–792.
- Somoza, J. L. (1980). Illustrative analysis: infant and child mortality in colombia, *World Fertility Survey Scientific Reports* **10**.
- Tadesse, M. G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data, *Journal of the American Statistical Association* **100**(470): 602–617.