An ANOVA Model for Dependent Random Measures

Maria DE IORIO, Peter MÜLLER, Gary L. ROSNER, and Steven N. MACEACHERN

We consider dependent nonparametric models for related random probability distributions. For example, the random distributions might be indexed by a categorical covariate indicating the treatment levels in a clinical trial and might represent random effects distributions under the respective treatment combinations. We propose a model that describes dependence across random distributions in an analysis of variance (ANOVA)-type fashion. We define a probability model in such a way that marginally each random measure follows a Dirichlet process (DP) and use the dependent Dirichlet process to define the desired dependence across the related random measures. The resulting probability model can alternatively be described as a mixture of ANOVA models with a DP prior on the unknown mixing measure. The main features of the proposed approach are ease of interpretation and computational simplicity. Because the model follows the standard ANOVA structure, interpretation and inference parallels conventions for ANOVA models. This includes the notion of main effects, interactions, contrasts, and the like. Of course, the analogies are limited to structure and interpretation. The actual objects of the inference are random distributions instead of the unknown normal means in standard ANOVA models. Besides interpretation and model structure, another important feature of the proposed approach is ease of posterior simulation. Because the model can be rewritten as a DP mixture of ANOVA models, it inherits all computational advantages of standard DP mixture models. This includes availability of efficient Gibbs sampling schemes for posterior simulation and ease of implementation of even high-dimensional applications. Complexity of implementing posterior simulation is—at least conceptually—dimension independent.

KEY WORDS: **???**

1. INTRODUCTION

We consider dependent nonparametric models for related random probability distributions or functions. A typical application arises in modeling random effects distributions for related submodels in a hierarchical model. We propose a model that describes dependence across random distributions in an analysis of variance (ANOVA)-type fashion. Specifically, assume that random distributions F_x are indexed by a p-dimensional vector $x = (x_1, \ldots, x_p)$ of categorical covariates. For example, in a clinical trial F_{x_1,x_2} could be the random effects distribution for patients treated at levels x_1 and x_2 of two drugs. We define a nonparametric probability model for F_x in such a way that marginally, for each x, the random measure F_x follows a Dirichlet process (DP), $\mathrm{DP}(M, F_x^o)$, with total mass parameter M and base measure F_x^o (Ferguson 1973). But we introduce dependence across x, that is, dependence for $(F_x, x \in X)$, using the dependent Dirichlet process (DDP) as defined by MacEachern (1999, 2001). The random measures F_x are almost surely discrete with the point masses generated marginally from the base measure F_x^o. The DDP introduces dependence across x by imposing dependence in the distribution of these point masses. We use the DDP to define ANOVA-type dependence across related random measures by assuming ANOVA models for these point masses. The resulting probability model defines an overall average effect and offsets for each level of the categorical covariates. If desired this can be generalized to include interaction effects. We propose a Markov chain Monte Carlo scheme to implement full posterior inference in the proposed model.

Our model is based on DP prior distribution (Ferguson 1973; Antoniak 1974). The DP is a probability model for random probability distributions. It plays a central role in nonparametric Bayesian inference, and it has been successfully applied in many problems. One of the critical properties is the a.s. discreteness of a random measure $F \sim \mathrm{DP}(M, F_0)$. Letting $\delta(x)$ denote a point mass at x, we can write $F = \sum_{h=0}^{\infty} w_h\,\delta(\theta_h)$. Here w_h are the weights of point masses at locations θ_h. Sethuraman (1994) gave a constructive definition of the DP. The weights are generated from rescaled Beta distributions, $w_h / \prod_{i=1}^{h-1}(1 - w_i) \sim \mathrm{Be}(1, M)$, and the locations θ_h are iid samples from the base measure F_0.

Another property that will feature prominently in the following discussion is the Pólya urn representation for the marginal distribution of a sample from a random DP distribution. Assume $y_i \sim F, i = 1, \ldots, n$, is sampled from an unknown distribution F, which, in turn, is generated by a DP, $F \sim \mathrm{DP}(M, F_0)$. The marginal distribution of $y = (y_1, \ldots, y_n)$ is described by the following Pólya urn scheme (Blackwell and MacQueen 1973): $y_1 \sim F_0$ and

$$p(y_m|y_1, \ldots, y_{m-1}) = \begin{cases} \delta(y_i), & \text{with probability } 1/(M+m-1), \\ & \qquad i = 1, \ldots, m-1, \\ F_0, & \text{with probability } M/(M+m-1). \end{cases} \tag{1}$$

The mth sample point is either a tie with a previous sample y_i or a new draw from the base measure. The positive probability of ties in (1) is due to the discrete nature of the random distribution $F \sim \mathrm{DP}(M, F_0)$.

In many data analysis applications this discreteness is inappropriate. DP mixture (DPM) models avoid this discreteness in the sampling distribution by adding an additional convolution with a continuous kernel. The typical DPM model assumes

$$\begin{aligned} &y_i \overset{\text{iid}}{\sim} H, \qquad \text{with } H(y) = \int f(y|\mu)\,dF(\mu), \\ &F \sim \mathrm{DP}(M, F_0), \end{aligned} \tag{2}$$

that is, a mixture with a DP prior on the random mixing measure F. Many applications use a normal kernel $f(y|\mu) = N(\mu, S)$ with a common covariance matrix S, leading to a discrete mixture of normals $H(y) = \sum_{h=1}^{\infty} w_h N(\mu_h, S)$. One of the main attractions of DPM models like (2) is computational simplicity. Also, posterior simulation algorithms are dimension

Maria De Iorio is Professor, Imperial College, London, U.K. Peter Müller is Professor, Department of Biostatistics, M. D. Anderson Cancer Center, University of Texas, Houston, TX. Gary L. Rosner is Professor, Department of Statistics, M. D. Anderson Cancer Center, University of Texas, Houston, TX. Steven N. MacEachern is Professor, Department of Statistics, Ohio State University, Columbus, OH. This research was suppported by NIH/NCI grant 2 R01 CA75981-04A1 and NSF grant DMS 0072526.

© 0 American Statistical Association
Journal of the American Statistical Association
???? 0, Vol. 0, No. 0, Theory and Methods
DOI 10.1198/00

independent. See, for example, Escobar and West (1998) or MacEachern and Müller (2000) for a review of models based on (2).

Several papers considered extension of DP and DPM models to hierarchical models over related random distributions. In the context of parametric models, that is, models with finite-dimensional parameter vector, such hierarchies with submodels for related experiments are standard modeling tools. In nonparametric models such extensions are complicated by the infinite-dimensional nature of the random distribution. Some of the first developments of dependent DP models appeared in Cifarelli and Regazzini (1978), who defined dependence across related random measures by introducing a regression for the base measure of marginally DP distributed random measures. The model was used, for example, in Muliere and Petrone (1993), who defined dependent nonparametric models $F_x \sim \mathrm{DP}(M, F_x^o)$ by assuming a regression in the base measure, $F_x^o = N(\beta x, \sigma^2)$. Similar, models are discussed in Mira and Petrone (1996) and Giudici, Mezzetti, and Muliere (2002). A similar strategy of linking dependent random DP measures F_x at the level of the base measure was used in Tomlinson and Escobar (1999). They achieved increased flexibility by assuming a DPM hyperprior on the common base measure. Gelfand and Kottas (2001) defined dependent nonparametric models by considering a representation of random measures as products of DP distributed factors. This allowed them to enforce stochastic ordering. Tomlinson and Escobar (1999) and Gelfand and Kottas (2001) are appropriate to model dependence across several related random measures. However, they are not naturally extended to include regression on covariates.

MacEachern (1999) defined the dependent DP (DDP) to allow a regression on a covariate x. Consider a family of random measures $\mathcal{F} = (F_x, x \in X)$ indexed by a covariate x. MacEachern (1999) defined a probabilty model for $\mathcal{F}$ such that marginally, for each x, $F_x = \sum w_h \delta(\theta_{xh})$ follows a DP. We use an additional subindex x for the point masses θ_{xh} to indicate the point masses in the random measure F_x. In the basic DDP model the weights w_h are common to all F_x. The DDP model induces dependence across x by assuming that $\theta_h = (\theta_{xh}, x \in X)$ are iid realizations of a stochastic process (in x). For example, θ_{xh} might be assumed to be a Gaussian process. Independence across h, together with the stick-breaking prior for the weights w_h, guarantees that F_x marginally follows a DP. Dependence in the sample path of the stochastic process θ_h introduces the desired dependence across x. We use this DDP structure to develop an ANOVA-like probability model over an array of random distributions. The DDP model provides a convenient starting point for the discussion. But the proposed model is more general. It can be rewritten as a DP mixture model. With minimal changes in the computational algorithms the DP can be replaced by any nonparametric model that allows a constructive definition by a stick-breaking algorithm as in Sethuraman (1994). See, for example, Ishwaran and James (2001) for discussions of such probability distributions.

In Section 2 we develop the basic model as a dependent DP model. In Section 3.1 we rewrite the model as a DP mixture of ANOVA models. Building on this representation, we discuss computational implementation issues. Section 3.3 discusses the use of the ANOVA DDP model to define random effects distributions in hierarchical models and other applications. Section 4 illustrates the proposed models with three examples. Section 5 concludes with a final discussion.

2. THE ANOVA DDP

Assume $\mathcal{F} = (F_x, x \in X)$ is an array of random distributions, indexed by a categorical covariate x. For simplicity of explanation, assume for the moment that $x = (v, w)$ is bivariate with $v \in \{1, \dots, V\}$ and $w \in \{1, \dots, W\}$. The covariates (v, w) could be, for example, the levels of two treatments in a clinical trial, and the distributions F_x might be sampling distributions for recorded measurements on each patient or distributions for random effects. In the latter case, an additional layer in the model hierarchy defines a sampling distribution for the observed outcomes conditional on the random effects.

In this context we wish to develop a probability model for the random distributions F_x that will enable us to build an ANOVA-type dependence structure. For example, we want the random distributions F_x and $F_{x'}$ for $x = (v_1, w_1)$ and $x' = (v_1, w_2)$ to share a common main effect due to the common factor v_1. The model should allow us to incorporate prior information about the presence of interaction between the covariates. If interactions are present, the effect of $v = v_1$ should be allowed to depend on the level of the other covariate w. The following model gives a formal definition to notions like "main effect" and "interaction." Briefly, instead of a nonzero additive effect on the mean of the response variable in an ANOVA model, an effect is recast as a difference in distribution of some quantity that has, in turn, an impact on the distribution of the final response. Thus, the models we create allow us to transfer both the interpretation and the structure used for unknown normal means in the traditional ANOVA model to unknown random distribution functions.

Like standard ANOVA models the proposed model can be justified by a judgment of partial exchangeability for observed data. Assume y_{xi}, $i = 1, 2, 3, \dots,$ are observed data, with y_{xi} denoting the ith observation under condition x, for example, the response of the ith patient assigned treatment combination x in a clinical trial. If, for each x, the subsequence $y_x = (y_{x1}, y_{x2}, \dots)$ is judged exchangeable, then by de Finetti's representation theorem y_{xi} can be assumed to be an iid sample from some distribution F_x. Bernardo and Smith (1994, chap. 4) show how additional assumptions, including, in particular, that the sample mean and variance are sufficient predictive statistics, leads to an ANOVA model. Stopping short of the assumption about predictive sufficiency naturally leads to assuming a nonparametric prior for F_x

We achieve the desired model structure by using the DDP framework. Specifically, let $F_x = \sum w_h \delta(\theta_{xh})$ for $x = (v, w)$. We assume Sethuraman's (1994) stick-breaking prior for the common weights, $w_h / \prod_{i=1}^{h-1} (1 - w_i) \sim \mathrm{Be}(1, M)$. We impose additional structure on the locations θ_{xh}:

$$\theta_{xh} = m_h + A_{vh} + B_{wh}. \tag{3}$$

As in standard ANOVA models we need to introduce an identifiability constraint for interpretability. We may impose any of the standard constraints, for example, $A_{1h} = B_{1h} \equiv 0$. For the remaining parameters we assume $m_h \overset{\text{iid}}{\sim} p_m^o(m_h)$, $A_{vh} \overset{\text{iid}}{\sim} p_{Av}^o(A_{vh})$, and $B_{wh} \overset{\text{iid}}{\sim} p_{Bw}^o(B_{wh})$, with independence being across h, v, and w. We refer to the joint probability model on $\mathcal{F}$ as $(F_x,\ x \in X) \sim$ ANOVA DDP(M, p^o). The model is parameterized by the total mass parameter M and the base measure p^o on the ANOVA effects in (3). Marginally, for each $x = (v, w)$, the random distribution F_x follows a DP with mass M and base measure F_x^o given by the convolution of p^o_m, p^o_{Av}, and p^o_{Bw}. Model (3) defines dependence across x by defining the covariance structure of the point masses θ_{xh} across x. As in standard ANOVA the structural relationships are defined by the additive structure (3) and the level of the dependence is determined by the variances in p^o_m, p^o_{Av}, and p^o_{Bw}. For example, consider two treatment combinations $x = (v, 1)$ and $x' = (v, 2)$ and random samples $y_x \sim F_x$ and $y_{x'} \sim F_{x'}$. Assuming normal priors $p^o_m = \mathrm{N}(\mu_m, \sigma_m^2)$, $p^o_{Av} = \mathrm{N}(\mu_{Av}, \sigma_A^2)$, and $p^o_{Bw} = \mathrm{N}(\mu_{Bw}, \sigma_B^2)$, we find the marginal covariance $\mathrm{cov}(y_x, y_{x'}) = (\sigma_m^2 + \sigma_A^2)/(M+1)$. Compare this with the covariance $\mathrm{cov}(y_x, y_{x'}) = (\sigma_m^2 + \sigma_A^2)$ that would arise in a standard ANOVA with the base measure p^o as prior for the ANOVA effects. The ANOVA DDP model introduces an additional level of uncertainty by defining the random measures $(F_x, F_{x'})$. The resulting covariance structure remains unchanged except for the attenuation factor $1/(M+1)$ corresponding to the additional uncertainty about F_x. The same result remains true for arbitrary ANOVA structure, including more factors and possibly interactions. In general, the marginal covariance for the observable responses y_x is the covariance under the corresponding fully parametric ANOVA model, reduced by a factor $1/(M+1)$. The marginalization is over the random distributions F_x with respect to the ANOVA DDP prior.

Model (3) is not constrained to univariate distributions F_x. The point masses θ_{xh} and the ANOVA effects m_h, A_{vh}, and B_{wh} can be q-dimensional vectors. This is important, for example, if the random distributions F_x are used as random effects models in a hierarchical model. In the example discussed in Section 4.2 we use seven-dimensional random effects vectors. It is a critical advantage of the ANOVA DDP model that model specification and computation are dimension independent.

Another important generalization of model (3) is to a more complex ANOVA structure. The model is easily generalized to a p-dimensional categorical covariate $x = (x_1, \ldots, x_p)$:

$$\theta_{xh} = m_h + \sum_{i=1}^{p} A_i(x_i), \tag{4}$$

where $A_i(x_i)$ is the main effect due to treatment x_i. Further extensions to include interactions such as A_{ij} are equally straightforward. For example, in (3) we could introduce additional terms $C_{vw,h}$. See the examples in Section 4 for an illustration. However, the same caveat as with corresponding parametric models applies. Meaningful inference for interactions can only be achieved with sufficiently many data observed under different covariate levels x.

Like standard Bayesian ANOVA, the model allows us to incorporate differential prior information for the various levels of the covariate. This is accomplished through choice of different prior distributions p^o_{Av} for the different levels of v. In the context where $v = 1$ indicates a control and $v = 2, \ldots, V$ are exchangeable treatments, we might take p^o_{Av} to be degenerate at 0 for the control and to be an identical distribution with a larger spread for each of the treatments. As an analog of a linear contrast in standard ANOVA, we might take the distributions p^o_{Av} to have nonzero means falling along a line; including further structure on the means of these distributions lets us expand our models in a fashion similar to the classical expansion through orthogonal polynomials, though the realizations will not exactly follow the possibly lower dimensional model.

One can also place constraints on the estimated effects. Enforcing a dependence above that forces the A_{vh} to lie on a line (to do so, we need to violate the condition of independence of A_{vh} across levels of v) produces a lower-dimensional component in the model. Alternatively, a constraint such as monotonicity of the effect A_{vh} in v can be enforced. Such a constraint ensures that the random distributions F_x are stochastically ordered with respect to v. This type of constraint is meaningful, for example, if v is the toxicity level of an anticancer agent in a chemotherapy treatment.

3. MIXTURE OF ANOVA MODELS

3.1 A DP Mixture of ANOVA Models

Most applications of DP models in data analysis add an additional layer in the model to convolve the discrete measure generated from a DP with a continuous, typically normal, kernel, to construct DPM models as in (2). For the same reasons we propose to add an additional normal mixture to the ANOVA DDP model. This leads to models of the form

$$\begin{aligned} (y_i | x_i = x) &\sim H_x(y_i), \qquad \text{with} \\ H_x(y) &= \int \mathrm{N}(y|\mu, S)\, dF_x(\mu), \\ (F_x, x \in X) &\sim \text{ANOVA DDP}(M, p^o), \end{aligned} \tag{5}$$

with appropriate hyperpriors for the common normal variance S and the ANOVA DDP parameters.

Implementation of posterior inference in the ANOVA DDP model (5) is most easily developed on the basis of an equivalent reformulation of the model as a mixture of ANOVA models. Consider, for example, model (3) for q-dimenensional random measures F_x. Let N denote the number of ANOVA effects in (3) and let $\alpha_h = [m_h, A_{2h}, \ldots, A_{Vh}, B_{2h}, \ldots, B_{Wh}]$ denote the $(q \times N)$ matrix of ANOVA parameters correponding to the hth point mass in the random measures. Let d_i denote a design vector to select the appropriate ANOVA effects corresponding to x_i, that is, $\theta_{xh} = \alpha_h d_i$ for $x = x_i$. Using this notation, model (5) with base measure $(p^o_m, p^o_{Av}, p^o_{Bw})$ can be rewritten as

$$\begin{aligned} (y_i \mid x) &\sim H_x(y_i), \\ H_x(y) &= \int \mathrm{N}(y|\alpha\, d_i, S)\, dF(\alpha), \\ F &\sim \mathrm{DP}(M, p^o). \end{aligned} \tag{6}$$

In words, data y_i are sampled from a mixture of ANOVA models with a DP prior on the unknown mixing measure. As usual in mixture models posterior simulation is based on breaking the mixture in (6) by introducing latent variables α_i:

$$y_i = \alpha_i\, d_i + \epsilon_i, \qquad \alpha_i \sim F, F \sim \mathrm{DP}(M, p^o), \tag{7}$$

with $\epsilon_i \sim \mathrm{N}(0, S)$. It follows from this equivalence that any Markov chain Monte Carlo (MCMC) scheme for DP mixture models can be used for posterior simulation in DDP ANOVA models of the type (5). The conjugate nature of the base measure p^o and the kernel in the error distribution $p(\epsilon_i)$ in (7) greatly simplify posterior simulation. See, for example, MacEachern and Müller (1998) for details of posterior MCMC simulation for DPM models. A description of the relevant modifications needed for the ANOVA DDP is given in Appendix A. Because the MCMC simulation proceeds by marginalizing with respect to the unknown measures F_x, inference about the unknown distribution F_x itself is not a standard part of MCMC in DP mixture models and can be difficult. See Gelfand and Kottas (2002) for a discussion and for inference about F_x in general. However, some important simplifications are possible. For any fixed y, the value of $H_x(y)$ in (6) takes the form of a linear functional of a DP random measure for which Regazzini et al. (2002) gave results for exact inference. Guglielmi and Tweedie (2001) discussed an MCMC implementation to evaluate linear functionals. Further, in many applications the desired inference on F_x is constrained to reporting the posterior mean measures $E(H_x|Y)$ and sampling from $p(H_x|Y)$. Here Y denotes the observed data. But the posterior mean $E(H_x|Y)$ is equal to the posterior predictive distribution $p(y_{n+1}|x_{n+1} = x, Y)$ for a future observation in the DPM model (6). We can exploit this to easily evaluate $E(H_x|Y)$. Proceeding with Rao–Blackwellization as usual in MCMC implementations (Casella and Robert 1996), it is possible to further simplify computations and estimate $E(H_x|Y)$ as an average of complete conditional posterior distributions that are already evaluated in the course of the MCMC simulation. Details of this estimate as well as a computationally efficient algorithm for approximate sampling from $p(H_x|Y)$ are described in Appendix A.

Although the DP mixture representation of the ANOVA DDP model is helpful to derive posterior simulation methods, the proposed model goes beyond standard DP mixture models. One crucial difference between the DDP model that we present and corresponding DPM models concerns where the models are defined. The DPM models are defined at a fixed, finite set of covariates whereas the DDP model may be defined over a much broader space. The study in Section 4.2 concerns a factorial design with two factors (doses of two drugs) studied at three and four levels, respectively. Potential doses for each drug range along a continuum. The DDP model specifies a prior distribution, which, when updated with the data, allows us to make posterior inference at an *arbitrary* combination of doses. Such a specification is consistent with the DDP model of reduced dimension that we use for computational purposes. Of course, such inference would require the prior specification for the ANOVA effects to extend over a continuum of covariates. For example, in (3) we could use a Gaussian process prior on A_{vh} as a function of a continuous dose v.

Inference for a DDP model with a typical prior specification has a particularly attractive feature, namely appropriate asymptotic degeneracy and nondegeneracy. Recall that $\alpha = [m, A_2, \ldots, A_V, B_2, \ldots, B_W]$, $\alpha \sim F$, is the matrix of ANOVA effects in (6). Suppose that an augmented set of treatment combinations includes a dose combination x^* that was not recorded in the observed data and let d^* denote the design vector corresponding to x^*. If the base measure p^o is a nonsingular multivariate normal distribution, then the conditional distribution of the mean effect αd^* at the new dose combination, conditional on the mean effect at the studied dose combinations, is a nondegenerate normal distribution. This conditional nondegeneracy is typical of a DDP prior specification. The consequence of nondegeneracy is that, even if αd_x at the studied dose combinations x were known exactly, the posterior distribution at the new dose combination would be nondegenerate. Thus, even if the posterior distribution on the mean effect at observed treatment combinations concentrates asymptotically, the posterior distribution at other dose combinations does not concentrate. This behavior is in keeping with our prior beliefs in most settings.

The parameter M induces a distribution on the number of clusters into which the observations fall. Clusters are defined by ties in the latent α_i introduced in (7). In DPM models, summaries of this distribution, such as its mean and variance, are often used to judge whether a particular prior distribution provides a match to prior beliefs. See, for example, Escobar (1994), who also used these summaries to motivate a distribution over M. The DDP model that we use in this article relies on a single mass parameter, M. For this DDP model, clusters of observations occur both within a treatment and across treatments. The number of clusters is stochastically increasing with the number of observations. See Appendix B for a formal statement and proof. Consequently, varying sample sizes across treatments imply different distributions for the number of within-treatment clusters as well as a different distribution of the number of clusters across all treatments.

For illustration, we computed summaries of the prior distribution on the number of clusters in a setup as in the case study analyzed in Section 4.2; that is, we assumed the same number of dose combinations x and the same sample sizes within each dose combination x. Let k_x denote the number of clusters of observations with covariate x and let k denote the number of clusters among all observations. Note that $k \le \sum_x k_x$ because clusters can extend over multiple values of x. Considering four alternative prior assumptions for the total mass parameter M, we evaluated prior mean and standard deviation for the number of clusters. Specifically, we computed prior summaries for k_x and k under a Gamma prior, $M \sim \mathrm{Ga}(5, .5)$ and assuming fixed values $M = 1$, 10, and 25. Let n_x denote the number of observations with covariate level x. For a treatment with $n_x =$ six observations and assuming $M \sim \mathrm{Ga}(5, .5)$, $M = 1$, 10, and 25, we find the prior mean $E(k_x)$ (standard deviation $\mathrm{SD}(k_x)$) to be 4.8 (1.0), 2.5 (.98), 4.9 (.91), and 5.5 (.68), respectively. For a treatment with $n_x =$ ten observations, we find prior moments of 6.9 (1.6), 2.9 (1.2), 7.2 (1.3), and 8.6 (1.1), respectively. For a treatment with a relatively large number of observations, $n_x = 12$, we find prior moments of 7.9 (1.9), 3.1 (1.2), 8.2 (1.5), and 10.0 (1.1), respectively. Finally, for the entire study, with $n =$ fifty-two patients, we find prior moments for k of 18.0 (5.2), 4.5 (1.7), 18.7 (3.1), and 28.5 (3.3), respectively. In summary, the impact of the different numbers of observations (6, 10, and 12) at the different treatments is relatively modest. There is, however, a substantial difference when moving from the individual treatments with $n_x = 6$, 10, and 12 to the entire experiment with $n = 52$.

3.2 Other Mixture of ANOVA Models

Rewriting the ANOVA DDP as (7) highlights the generality of the underlying model structure. The use of a DP prior for the discrete mixing measure is motivated by technical convenience and because the parsimonious parameterization of the DP avoids difficult prior elicitation problems. On the other hand, the fact that the DP is parameterized by a base measure and one scalar precision parameter M only could sometimes be a limitation. Green and Richardson (1998) argued for the use of more general mixture models and showed appropriate posterior simulation schemes. Muliere and Tardella (1998), Ishwaran and James (2001), and Gelfand and Kottas (2002) discussed finite truncations of DP priors. Ishwaran and James (2001) proposed alternative nonparametric priors based on similar stick-breaking representations. Any of these nonparametric priors can be substituted in (7) without changing the model structure and requiring only minimal changes in the posterior simulation schemes.

3.3 Hierarchical Models and Other Extensions

Consider a generic hierarchical model of the form

$$y_i \sim p(y_i|\theta_i), \qquad (\theta_i|x_i = x) \sim H_x(\theta|\phi). \tag{8}$$

In words, data y_i for the ith sampling unit, for example a patient, is sampled from a probability model parameterized by a random effects vector θ_i. For example, this could take the form of a nonlinear regression

$$y_{ij} = f(t_{ij}; \theta_i) + \epsilon_{ij}, \tag{9}$$

with a mean function $f(\cdot; \theta)$ parameterized by θ_i and evaluated at known times t_{ij}, $j = 1, \ldots, n_i$. The θ_i are generated from a random effects distribution H_x. The random effects distribution depends on a covariate specific to the sampling unit and possibly additional hyperparameters ϕ.

If the covariates are, for example, treatment indicators for the ith patient, then ANOVA DDP models as in (5) are appropriate prior probability models for $(H_x, x \in X)$. The random effects vector θ_i takes the place of y_i in (5), and $H_x(\cdot) = \int \mathrm{N}(\cdot|\mu, S)\, dF_x(\mu)$. In general, the ANOVA DDP model can be used whenever the random effects distributions H_x are indexed by some categorical covariates x_i specific to the ith unit with a notion of ANOVA-type dependence across the random distributions. Section 4.2 discusses a typical example. Posterior inference is implemented as in (5), with an additional step to update the random effects vectors θ_i, which now replace the data y_i in the ANOVA DDP model (5). The details of this step are problem specific. For example, if $p(y_i|\theta_i)$ is a nonlinear regression as in (9), then updating θ_i amounts to a posterior draw from the parameters in a nonlinear regression with data y_{ij}, $j = 1, \ldots, n_i$, and prior $\theta_i \sim \mathrm{N}(\alpha_i d_i, S)$. Here α_i is the latent variable introduced to break the mixture model in (7). The latent variable α_i is imputed in the course of updating the ANOVA DDP model. See, for example, MacEachern and Müller (1998) for details.

The hierarchical model (8) is only one of many diverse modeling contexts and applications that gives rise to structure as in (5). The ANOVA DDP model (5) allows a convenient formalization of regression for survival data. Consider observations y_i and censoring indicators δ_i, with $\delta_i = 1$ if y_i is a censoring time and $\delta_i = 0$ if y_i is an observed event. Assume different subpopulations are indexed by a categorical covariate x_i. Using (2) to represent $p(y_i|x_i, \delta_i = 0)$, one can represent the sampling distribution of event times under x_i as a DP mixture model. Such models are used, for example, in Gelfand and Kuo (1991), Doss (1994), and Kuo and Mallick (1997). Kuo and Mallick (1997) included a regression in an accelerated failure time model. The DDP ANOVA model allows us to introduce dependence across subpopulations indexed by x_i when stronger assumptions such as proportional hazards, additive hazards, or accelerated failure time are too stringent.

The basic ANOVA DDP model (5) is easily extended to categorical or ordinal data. Let $z_i = (z_{i1}, \ldots, z_{ip})$ denote a vector of observed categorical outcomes for experimental unit i. A multivariate probit model for z_i represents the sampling distribution for z_i by introducing a latent multivariate normal vector y_i. Specifically, assume $z_{ik} \in \{1, \ldots, R\}$ is ordinal with R possible outcomes. We introduce cutoffs $-\infty = \theta_0 < \cdots < \theta_r < \cdots < \theta_R = \infty$ and assume $z_{ik} = r$ if and only if $\theta_{r-1} < y_{ik} \le \theta_r$. See, for example, Albert and Chib (1993), Cowles, Carlin, and Connett (1996), and Chen and Dey (2000) for inference in such models. Replacing the normal distribution of the latent variable y_i by an unknown distribution with a nonparametric prior provides a natural generalization of the probit model. Kottas, Müller, and Quintana (2002) showed that without loss of generality the cutoffs θ_j can be fixed when using a nonparametric prior model for the latent variables y_i. Consider now data recorded for different subpopulations indexed by x_i. For example, x_i could be different hospitals, different treatments, or different raters. The DDP ANOVA model (5) formalizes a nonparametric regression of z_i on the covariates x_i. The categorical cell probabilities $\pi_{zx} = \Pr(z|x)$ are modeled as corresponding quantiles of H_x, that is, $\pi_{zx} = \int dH_x(y)$ with the integral extending over the range of y corresponding to categorical outcomes z.

Another interesting application is to classification. Assume we have data (x_i, y_i), $i = 1, \ldots, n$, under sampling model (5) for $p(y_i|x_i)$. Data could be repeated measurement data with the basic model being extended as in (8). Assume a new observation y_{n+1} with unknown label x_{n+1} is recorded, and interest focuses on inference about x_{n+1}. Augmented by a prior $p(x_i)$, the DDP ANOVA model allows a model-based flexible framework to implement the desired classification as $p(x_{n+1}|y_{n+1}, Y)$.

4. EXAMPLES

4.1 Simulation Example

Consider a two-way bivariate ANOVA model with two factors v and w. Assume the number of levels are $V = 2$ and $W = 2$ for v and w, respectively. Let $\alpha = [m, A_1, A_2, B_1, B_2]$ denote overall mean and main effects for $v = 1, 2$ and $w = 1, 2$, respectively. Such data might arise, for example, when investigating differential expression patterns of proteins across major subtypes of ovarian cancer. Assume expression is measured by immunohistochemical staining of tissues. The bivariate outcome could be the level of expression of the two proteins of interest, and histologically different subtypes might be defined by type v (serous, endometrioid, and mucinous) and degree of differentiation w (low and high).

We simulated 100 observations $y_i \sim \mathrm{N}(\alpha_i\, d_i, I)$ with randomly selected designs d_i and generating

$$\alpha_i = \begin{cases} \begin{bmatrix} 3 & 2 & 5 & -1 & -1 \\ 1 & 6 & 6 & 7 & 4 \end{bmatrix}, & \text{with probability } .5, \\ \begin{bmatrix} 3 & .5 & 3.5 & -2.5 & -2.5 \\ 1 & 7.5 & 7.5 & 8.5 & 5.5 \end{bmatrix}, & \text{with probability } .5, \end{cases}$$

that is, we generate from model (6) with a two-point discrete mixing measure $F(\alpha)$. We proceeded to fit the ANOVA DDP model (5) with an ANOVA structure as in (3). Let $H_{vw}(\cdot) = \int \mathrm{N}(\cdot|\mu, S)\, dF_{vw}(\mu)$ denote the unknown distribution for factor levels (v, w) and let $Y = (y_1, \ldots, y_n)$ denote the observed data. Figure 1 plots contours of the estimated distributions $E(H_{vw}|Y)$. Posterior inference correctly recovers the true mixture. For comparison, we estimated a bivariate normal regression on (v, w), that is, a standard bivariate ANOVA model. For a fair comparison the prior probability model for the bivariate ANOVA model and the hyperprior for the base measure p^o in the DDP model are matched. In both models the residual covariance matrix S is assumed to be diagonal, $S = \sigma^2 I$, with a conjugate Gamma prior on $1/\sigma^2$. Thus both models have almost the same number of hyperparameters, with the only additional hyperparameters in the ANOVA DDP model being the prior parameters for the total mass M. The posterior mean bivariate ANOVA model is shown in Figure 1 as contours overlayed on the gray shades for the estimated DDP model. Forced to fit a single normal, the bivariate ANOVA model approximates the mixture by a bivariate normal distribution centered between the locations of the two mixture terms. Failing to model the heterogeneity in the data, the bivariate ANOVA model reports the mode of the distribution in a low-density area between the two modes. The estimated residual variance is inflated to allow for the approximation. The marginal posterior means $E(\sigma^2|Y)$ are 1.01 and 3.57 under the ANOVA DDP model and the bivariate ANOVA model, respectively.

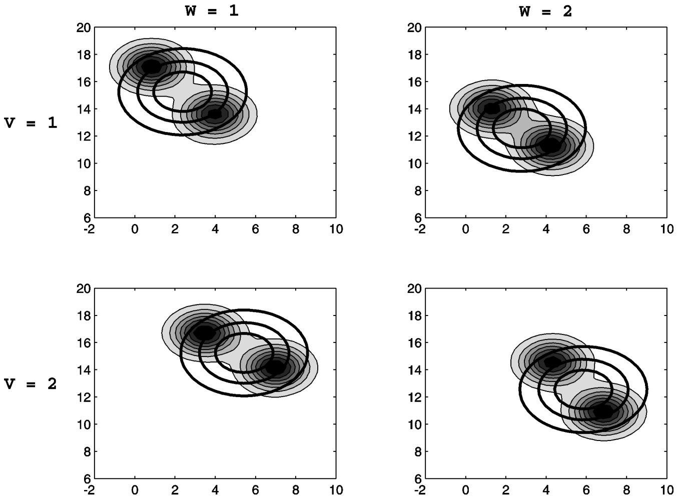

Figure 1. Posterior Estimated Distributions. The gray shades show the posterior mean distributions E[H_{vw}(·)| Y]. For comparison, the overlayed contours show the estimated distributions under a bivariate two-way ANOVA model, assuming a single bivariate normal sampling distribution for each combination of (v,w).

4.2 Hierarchical Models

Müller and Rosner (1997) described an analysis of hematologic data arising from a dose–escalation study (Lichtman et al. 1993). The data are white blood cell counts over time for each of n patients receiving relatively high doses of cancer chemotherapy. The treatment included a commonly used drug, cyclophosphamide, that is known to lower a person's white blood cell counts as the dose increases. Having very low white blood cell counts can be life-threatening because these cells are part of the human immune system. Thus, the investigators are keenly interested in knowing about the effect of dose on the white blood cell counts in order to guard their patients against severe and life-threatening toxicity.

Unfortunately, there is no standard measure in use for characterizing this form of toxicity. Is it the smallest white blood count (WBC) or is it the length of time a person's WBC is below a threshold value? Analyzing the entire profile of blood counts over time allows one to characterize depressed counts in any way one would like. The modeling approach of Müller and Rosner (1997) allows such full modeling. However, a critical limitation of the approach discussed in Müller and Rosner (1997) is the way it implements regression on patient-specific covariates, in particular, the assigned dose of chemotherapy. We reanalyze the data, using the DDP ANOVA model instead. The ANOVA structure allows us to characterize dose-specific differences in WBC profiles between the doses.

The data record white blood cell counts over time for each of $n = 52$ chemotherapy patients. Denote by y_{it} the measured response on day t for patient i. The profiles of white blood cell counts over time look similar for most patients. Figure 2 shows some typical patients. There is an initial baseline, followed by a sudden decline when chemotherapy starts, and a slow, S-shaped recovery back to approximately baseline after the end of the treatment. Profiles can be reasonably well approximated with a regression function with two changepoints corresponding to the beginning of the decline and the nadir, a horizontal line for the baseline to the left of the first changepoint, a straight line for the rapid decline between the two changepoints, and a shifted and scaled logistic for the final S-shaped recovery to the right of the second changepoint. We parameterize this piecewise linear–linear–logistic regression with a seven-dimensional parameter vector θ (Müller and Rosner 1997). But the nonlinear regression parameters differ significantly across patients. Thus we introduce a patient-specific random effects vector θ_i. Conditional on θ_i, we assume a nonlinear regression using the piecewise linear–linear–logistic regression model

$$y_{it} = f(t; \theta_i) + \epsilon_{it}. \tag{10}$$

The model is completed with a random effects model H_x. The random effects distribution H_x depends on the treatment levels for patient i. There are two treatments, the actual anticancer agent cyclophosphamide (CTX) and a second drug (GM-CSF) given to mitigate some adverse side effects of the chemotherapy. We impose an ANOVA structure on H_x with rows and columns in the two-way ANOVA indicating levels of CTX and GM-CSF. We code the levels of CTX as $v = 1, \ldots, 4$ (corresponding to dose levels 1.5, 3.0, 4.5, and 6.0 gm/m^2) and the levels for GM as $w = 1, 2, 3$ (dose levels 2.5, 5, and 10 μg/kg).

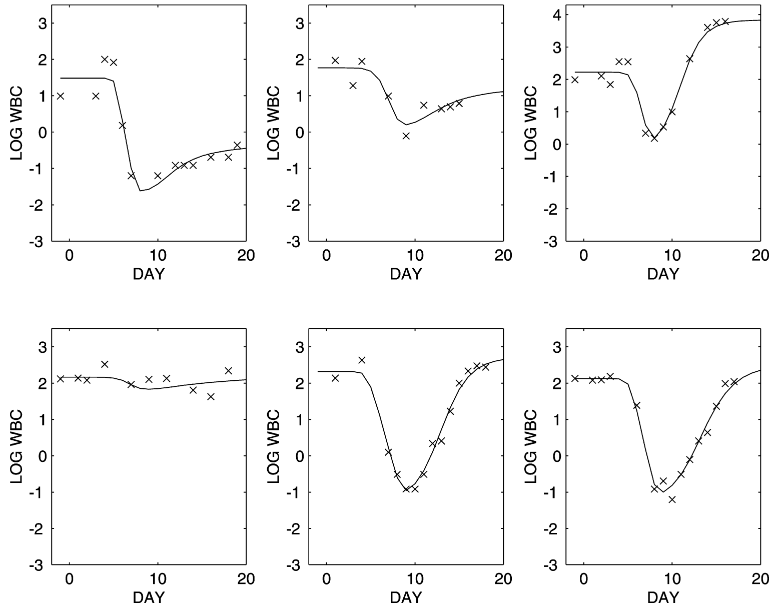

Figure 2. Some Typical Patients. The crosses plot observed log white blood cell count y_{it} for six typical patients. The curves show the posterior fitted mean response curves $f_{x_i}(t) = E\{\int f(t;\theta)\, dH_{x_i}(\theta) \mid y\}$ for these patients.

Let $x_i = (v_i, w_i)$ denote the treatment for patient i. The number of patients observed at each treatment combination x varies between $n_x = 6$ for $x \in \{(1,3), (2,1), (2,2), (2,3), (4,2)\}$, $n_x = 10$ for $x = (3,3)$, and $n_x = 12$ for $x = (3,2)$. We assume

$$
\begin{aligned}
&(\theta_i \mid x_i = x) \sim H_x(\theta),\\
&(H_x, x \in X) \sim \text{ANOVA DDP}(M, p^o).
\end{aligned}
\qquad (11)
$$

Posterior predictive inference for future patients depends on the observed historical data only indirectly through learning about the random effects distribution (11). Conditional on the random effects distributions H_x, observed and future data are independent. Thus a structured, flexible hyperprior for H_x is critical to achieve the desired learning.

Figure 3 summarizes some critical aspects of the analysis. In this example, the random effects distributions $H_x(\cdot)$ are seven-

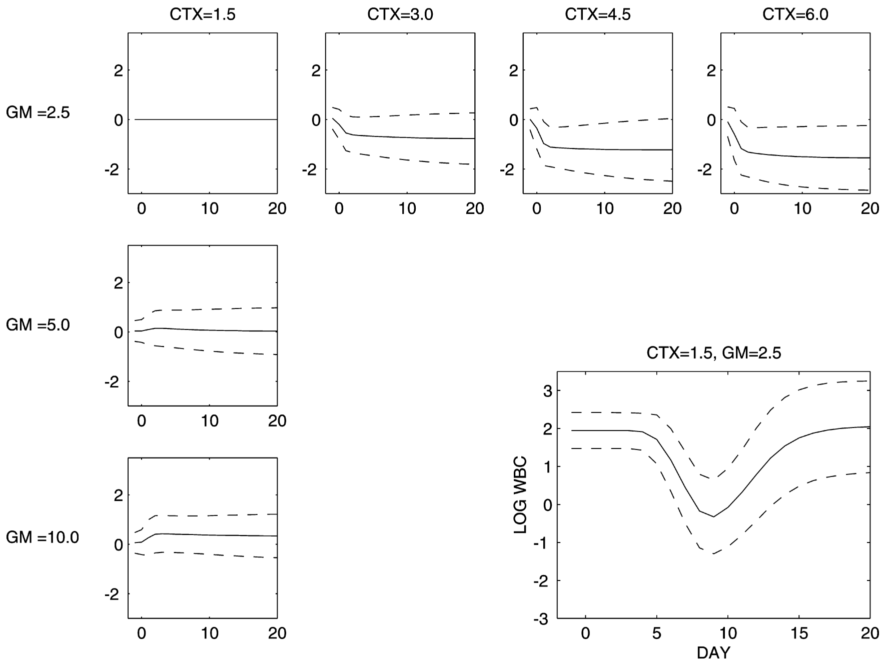

Figure 3. Estimated Main Effects A_v, B_w, and Overall Mean Effect m. For identifiability, we constrained $A_{1h} = B_{1h} \equiv 0$. The right lower panel shows the estimated hematologic profile for a patient treated with dose levels (CTX = 1.5, GM = 2.5), that is, corresponding to the overall mean m. The other figures summarize the offset corresponding to the respective main effect. For each ANOVA effect, the plot shows $E\{\int f(t;\theta = \alpha d)\, dF(\alpha) \mid Y\}$, where d is the design vector corresponding to the respective ANOVA effect. The dashed curves show corresponding pointwise one posterior standard deviation margins. See the text for a detailed explanation.

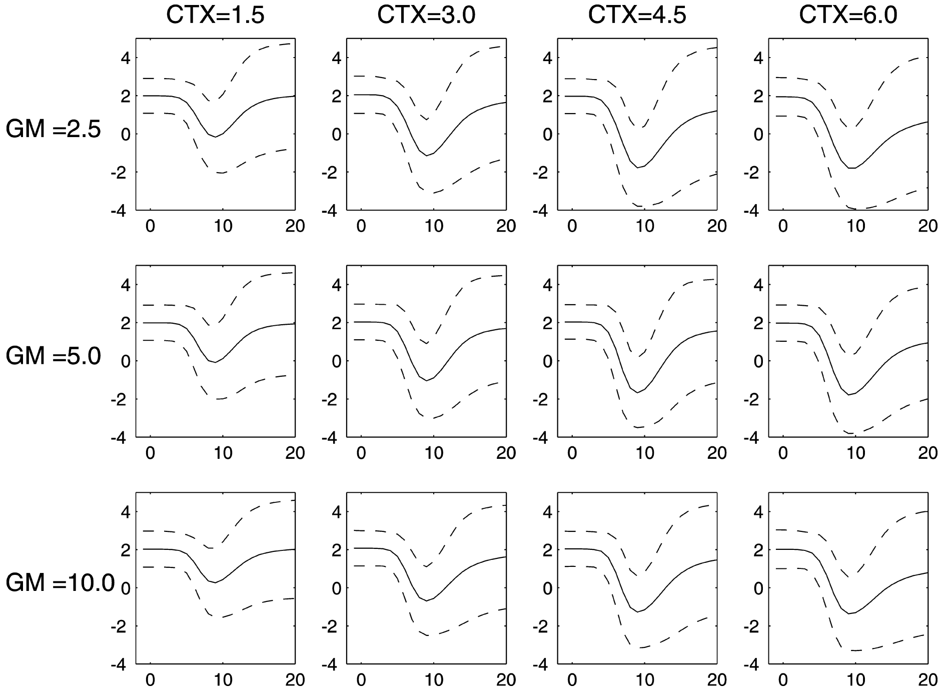

Figure 4. Posterior Predictive Distributions p(y_{it}|x_i,y) for a Future Patient $i = n + 1$ Assuming Treatment Combination $x_i = $ (v,w), Arranged by v and w. The solid line shows the pointwise posterior predictive mean E(y_{it}|x_i,y), plotted against t. The dashed lines show pointwise one posterior standard deviation margins.

dimensional. We summarize inference on $H_x(\cdot)$ by showing implied profiles. Let Y denote the observed data. For example, for $x = (1, 1)$ we show $f_x(t) = E\{\int f(t; \theta)\, dH_x(\theta)|Y\}$, the posterior mean hematologic profile over time t for a patient treated with doses $x = (1, 1)$. Imposing identifiability constraints $A_{1h} = B_{1h} \equiv 0$, the corresponding design vector in (6) is $d = (1, 0, \dots, 0)$, including a main effect only. The figure shows the estimated mean curve $f_x(t)$ in the right lower panel.

For patients treated at other doses, we display corresponding offsets in a familiar ANOVA fashion. For example, for $x = (2, 1)$ the figure plots the posterior expected mean curve corresponding to design vector $d = (0, 1, \dots, 0)$ with an offset for $v = 2$. The second panel in the top row shows $f_x(t) = E\{\int f(t; \theta = \alpha\, d)\, dF(\alpha)|Y\}$. The other panels in the same figure have analogous interpretations. Interpretation of Figure 3 needs to take into account that the additive ANOVA structure applies only for the ANOVA effects in (3). The plotted mean profiles are only convenient summaries of the seven-dimensional distributions and are not additive. Because inference is based on posterior simulation, any other desired posterior summary can be derived. For example, Figure 4 shows posterior predictive distributions for the likely responses of a future patient treated at any of the 12 possible treatment combinations.

4.3 Categorical Outcome

Fluorescence-activated cell-sorter (FACS) analysis is used to measure properties of cells in flow (Melamed, Lindmo, and Mendelsohn 1990). Freedman et al. (2002) used FACS to measure the proportion of monocytes in a blood sample that express certain surface markers. Using multicolor FACS analysis, up to four different surface markers can be recorded simultaneously. The instrument reports the number of cells that express each of the identified surface markers beyond a certain preset threshold. If data are reported for blood samples collected under different treatment conditions $x_i \in X$, the resulting data format fits the categorical data model described in Section 3.3, with $z_{ik} \in \{0, 1\}$ reporting presence and absence of the kth surface marker.

We set up a simulation study mimicking the setup in a currently ongoing clinical trial. The trial proposes chemoimmunotherapy for ovarian cancer. Two agents, denoted here as G and I, are used for the immunotherapy. FACS analysis is used to measure the expression of a variety of surface markers in blood samples collected under different treatment conditions. The first agent, G, has a growth factor stimulating effect. It can expand and prime monocytes, a type of white blood cell. The hypothesized effect of the second agent, I, is to serve as an activator to mature and fully activate these cells. As part of the protocol, blood samples will be analyzed for the number and activation status of monocytes. FACS analysis will be used to record expression of selected markers that are modified on activated monocytes. For this simulation study we consider two markers and samples under four different treatment conditions x_i, $x_i = (\mathrm{G}, \mathrm{I}) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. A pair of binary outcomes, $z_i = (z_{i1}, z_{i2})$ reports whether markers 1 and 2, respectively, are upregulated.

The conjectured role of I as an activator of cell populations expanded by G makes it important to include interactions in the model. We proceed with the ANOVA DDP model for ordinal data proposed in Section 3.3. We include main effects for G and I, plus an additional interaction I $*$ G. We generated hypothetical data for four blood samples, one under each treatment combination. We assume 100 cells are measured in each sample, with observed frequences reported in Table 1(a). Let

Table 1. Observed and Fitted Cell Probabilities $\pi_{zx} = Pr(z \mid x)$.

	$z = (z_1, z_2)$			
(G, I)	*(1, 1)*	*(1, 0)*	*(0, 0)*	*(0, 1)*
(a) Observed frequencies (in %)				
(0, 0)	29	7	59	5
(0, 1)	26	8	64	2
(1, 0)	10	26	4	60
(1, 1)	7	25	39	29
(b) Estimated probabilities (in %)				
(0, 0)	27 (5)	8 (3)	59 (6)	6 (3)
(0, 1)	27 (5)	8 (4)	62 (6)	3 (2)
(1, 0)	10 (4)	25 (6)	5 (3)	60 (6)
(1, 1)	7 (4)	25 (5)	39 (6)	29 (6)
(c) Estimated probabilities (in %) (bivariate probit)				
(0, 0)	16	21	45	18
(0, 1)	12	22	51	15
(1, 0)	28	8	24	40
(1, 1)	15	17	46	21

NOTE: Part (b) reports in parentheses the marginal posterior standard deviations. For comparison, part (c) reports the posterior mean cell probabilities under a bivariate probit model.

$\pi_{zx} = \Pr(z|x)$ denote the probability of binary outcomes z under subpopulation x. Using the DDP ANOVA model, the cell probabilities π_{zx} are expressed as the corresponding quantiles under H_x. For example, $\pi_{00} = \int_{-\infty}^{0} \int_{-\infty}^{0} dH_x(y_1, y_2)$, using the cutoff $\theta_1 = 0$.

Table 1(b) shows the posterior estimated cell probabilities, together with marginal posterior standard deviations. Marginal posterior distributions for the categorical cell probabilities are shown in Figure 5. For comparison, Table 1(c) shows inference under a bivariate probit regression. To allow a fair comparison, we use the base measure p^o as prior for the parametric probit model. Although the two models have a comparable number of hyperparameters, the nonparametric mixture in the ANOVA DDP model allows signficantly more flexibility in fitting the observed frequencies.

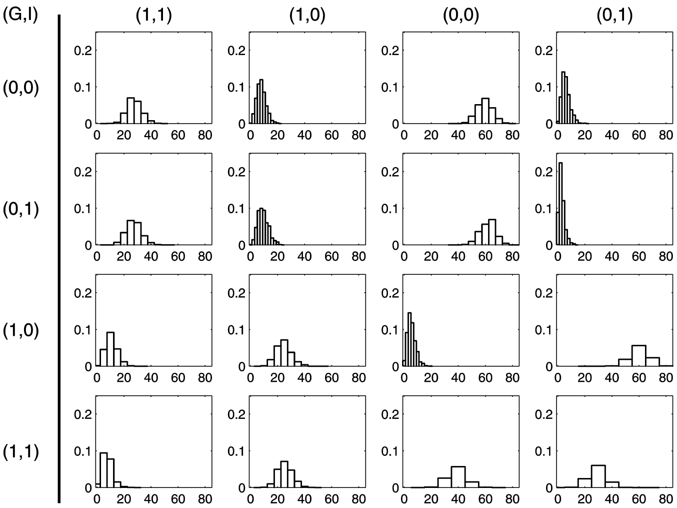

Figure 5. Posterior Distribution for the Unknown Cell Probabilities π_{zx} Arranged by Covariate x = (G, I) (rows) and by Outcome z (columns).

5. DISCUSSION

We have introduced a probability model for random distributions arranged in an ANOVA-like array. The main features of the proposed model are ease of interpretation, facility to impose structure in the usual ANOVA-like fashion, and efficient computation.

Limitations of the ANOVA DDP model are the need for MCMC simulation for posterior inference and the practical limitation to stick-breaking priors for the nonparametric mixing measure. Also, the model inherits limitations inherent in the DP prior. For example, the model includes only one scalar precision parameter for the random mixing measure. This makes it impossible to express different levels of prior precision across the sample space of the base measure. On the other hand, the parsimonious prior parameterization facilitates prior elicitation.

The ANOVA model that we have developed and illustrated in the examples relies on a particularly simple version of the DDP model. In this simple model, the weights of the point masses, w_h, do not depend on the level of the treatment. This model has the advantage of allowing us to rely on computational strategies developed for DPM models. More complex DDP models allow the w_h to vary across the treatments, but would necessitate more complex computational strategies that include additional parameters (the treatment-specific w_h or equivalent) in the Gibbs sampling steps. Inclusion of these additional parameters can, in other words, be described as forsaking a possible marginalization of them. The result of not marginalizing parameters when they can be marginalized is a poorer mixing Markov chain (Liu 1994; MacEachern 1994). An alternative approach to fitting the more complex DDP models is to use the model we have fit as an importance sampler. The two complex and simple models may be matched so as to provide the same marginal prior distributions at each level of the treatment. They would differ only in how the structure is connected across treatments, with this difference appearing in the prior distribution on the configuration vector, s (see Appendix B for a description of s). As a consequence, the important sampling weights would be determined by the two distributions on s.

APPENDIX A: POSTERIOR MCMC SIMULATION

We briefly describe the implementation of posterior simulation in the ANOVA DDP model (6). Because F is almost surely discrete (see Ferguson 1973), there is a positive probability for ties among the α_i. Write $\{\alpha_1^*, \dots, \alpha_k^*\}$ for the of $k \le n$ distinct elements in $\{\alpha_1, \dots, \alpha_n\}$. Set $s_i = j$ iff $\alpha_i = \alpha_j^*$. Let n_j be the number of s_i equal to j, that is, n_j is the size of the jth cluster, and let $\Gamma_j = \{i : s_i = j\}$. Let η denote possibly unknown hyperparameters, including hyperparameters in the base measure p^o, the total mass parameter M, and the covariance matrix S in (5). See, for example, Bush and MacEachern (1996) or MacEachern and Müller (1998) for a full description of Gibbs sampling scheme to estimate DPM models.

1. *Resampling s_i given all the other parameters.* Marginalize over α_i and sample s_i from

$$\Pr(s_i = j|s_{-i}, y, \eta, S)$$

$$\propto \begin{cases} n_j^- p(y_i|s_i = j, s_{-i}, S, \eta, y_{-i}), & j = 1, \dots, k^-, \\ M \int \mathrm{N}(y_i; \alpha d_i, S)\, dp^o(\alpha), & j = k^- + 1, \end{cases}$$

where $y_{-i} = \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n\}$, $s_{-i} = \{s_1, \dots, s_{i-1}, s_{i+1}, \dots s_I\}$,

$$n_j^- = \begin{cases} n_j - 1 & \text{if } j = s_i \\ n_j & \text{otherwise,} \end{cases}$$

and k^- = the number of clusters with α_i removed. If $n_{s_i}^- = 0$, we relabel the remaining clusters $j = 1, \dots, k^- = k - 1$. After sampling s_i, set

$$k = \begin{cases} k^- & \text{if } s_i \le k^- \\ k^- + 1 & \text{if } s_i = k^- + 1. \end{cases}$$

Note that, by conditional independence, in $p(y_i|s_i = j, s_{-i}, S, \eta, y_{-i})$ the data y_{-i} can be replaced by $\{y_l, l \in \Gamma_j, l \notin i\}$.

2. *Resampling* α_j^*. The posterior distribution $p(\alpha_j^*|s, y, \eta, S)$ is

$$p(\alpha_j^*|s, y, \eta, S) \propto \left[\prod_{i \in \Gamma_j} \mathrm{N}(y_i; \alpha_j^* d_i, S)\right] p^o(\alpha_j^*|\eta).$$

With a normal base measure p^o the posterior distribution of α_j^* is normal with mean and covariance matrix that can be found by standard calculations.

3. Resampling η conditional on the currently imputed values s, α^*, and k follows standard posterior simulation for DP mixture models.

To report posterior inference on H_x, the MCMC simulation can be augmented as follows. Let θ denote the vector of all model parameters and let $\theta^{(i)}$ denote the parameters imputed after i iterations of the MCMC simulation. We evaluate $E(H_x|Y) = p(y_{n+1}|x_{n+1} = x, Y)$ as

$$\begin{aligned} p(y_{n+1}|x_{n+1}, Y) &= E\big[p(y_{n+1}|x_{n+1}, Y, \theta)|Y\big] \\ &\approx \frac{1}{T}\sum_{i=1}^{T} p\big(y_{n+1}\big|x_{n+1}, \theta^{(i)}, Y\big) \\ &= \frac{1}{T}\sum_{i=1}^{T} p\big(y_{n+1}\big|x_{n+1}, \theta^{(i)}\big). \end{aligned}$$

The terms in the last average are easily computed. We use a superindex $^{(i)}$ to identify the imputed parameter values after i iterations of the MCMC simulation. Let d_x denote the design vector for $x = x_{n+1}$. We find

$$\begin{aligned} p\big(y_{n+1}\big|x_{n+1}, \theta^{(i)}\big) &\propto \sum_{j=1}^{k^{(i)}} n_j^{(i)} \mathrm{N}\big(y_{n+1}; \alpha_j^{*(i)} d_x, S^{(i)}\big) \\ &\quad + M^{(i)} \int \mathrm{N}\big(y_i; \alpha d_x, S^{(i)}\big)\, dp^o(\alpha). \qquad \text{(A.1)} \end{aligned}$$

Uncertainty in H_x is illustrated through posterior draws of H_x. For the following argument we consider augmenting the imputed parameter vector $\theta^{(i)}$ with the random distribution F defined in (6). Given $\theta^{(i)}$, the conditional posterior for F is a DP with updated parameters,

$$\big(F\big|\theta^{(i)}, Y\big) \sim \mathrm{DP}\big(q, M^{(i)} + n\big),$$

$$\text{with } q \propto M^{(i)} p^o + \sum_{j=1}^{k^{(i)}} n_j^{(i)} \delta_{\alpha_j^{*(i)}}. \qquad \text{(A.2)}$$

The large total mass parameter $M^{(i)} + n$ implies that the random measure F is close to the conditional expectation q, the DP base measure in (A.2). We exploit this to approximate a posterior draw $F \sim p(F \mid \theta^{(i)}, Y)$ as $F \approx q$, and thus a posterior draw for H_x as $\int \mathrm{N}(\alpha d_x, S^{(i)})\, dq(\alpha)$, that is,

$$H_x(y) \propto M^{(i)} \int \mathrm{N}\big(y; \alpha\, d_x, S^{(i)}\big)\, dp^o(\alpha) + \sum_{j=1}^{k^{(i)}} n_j^{(i)} \mathrm{N}\big(y; \alpha_j^{*(i)} d_x\, S^{(i)}\big).$$

The latter is simply the predictive distribution conditional on $\theta^{(i)}$ in (A.1).

APPENDIX B: STOCHASTIC MONOTONICITY OF *k*

The following two results are stated for the ANOVA DDP model that we develop in this article. The results apply more generally, with substitution of essentially any DP or DDP model for the model in (5). The first result describes the impact of the mass parameter on the distribution of the number of clusters in prior and posterior.

Result B.1. Consider a family of DDP models of the form given in (5), indexed by M. That is, the models have identical prior specifications except for the mass parameter of the DDP. Then both the prior distribution and the posterior distribution on the number of clusters is stochastically increasing in M.

Proof. The probability of obtaining k clusters from a sample of size n when the mass parameter of the Dirichlet process is M is given by

$$P_M(k) = \frac{M^k \Gamma(M)}{\Gamma(M+n)} \sum \prod_{i=1}^{k} \Gamma(n_i),$$

where the sum runs over vectors of length k having components n_i, with $n = \sum_{i=1}^{k} n_i$. This expression follows from the Pólya urn scheme and holds for $k = 1, \dots, n$.

Consider two values of the mass parameter, $M_1 < M_2$. Then

$$\frac{P_{M_1}(k)}{P_{M_2}(k)} = \left(\frac{M_1}{M_2}\right)^k \frac{\Gamma(M_1)}{\Gamma(M_1+n)} \frac{\Gamma(M_2+n)}{\Gamma(M_2)}$$

for $k = 1, \dots, n$. This ratio of probabilities is monotone decreasing in k, and so the prior distribution of the number of clusters for mass M_2 is stochastically greater than that for mass M_1.

Conditional on the number of clusters, k, the distribution on the configuration vector, s, does not depend on M. Consequently, with matched priors on hyperparameters and the shape of the base measure, the likelihood for the data depends on M only through k. Thus we may write

$$\frac{P_{M_1}(k|Y)}{P_{M_2}(k|Y)} \propto \frac{P_{M_1}(Y|k) P_{M_1}(k)}{P_{M_2}(Y|k) P_{M_2}(k)} \propto \left(\frac{M_1}{M_2}\right)^k.$$

As before, monotonicity of these ratios in k leads to the conclusion that the *posterior* distribution for the number of clusters is stochastically increasing in M.

This result describes the impact of the sample size on the number of clusters. Although the qualitative result follows from the Pólya urn scheme (a new draw from the DDP may either join an already existing cluster or begin a new cluster—it cannot remove an existing cluster), the recursion in the following proof is useful for quick computation of the prior distribution on the number of clusters. This distribution is useful for marginalizing the mass parameter through preintegration, as described in MacEachern (1998).

Result B.2. Consider the model in (5). The number of clusters is stochastically increasing in n.

Proof. Let $p^{(n,M)}$ represent the vector of length n with components $P_M(k)$. $p^{(1,M)} = (1)$ for all M. To find the corresponding distribution for a sample of size $n+1$, given the distribution for a sample of size n, note that

$$p^{(n+1,M)} = \frac{n}{M+n}\big(p^{(n,M)},0\big) + \frac{M}{M+n}\big(0,p^{(n,M)}\big).$$

This yields the result.

[Received October 2002. Revised October 2003.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

Bush, C. A., and MacEachern, S. N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–285.

Casella, G., and Robert, C. P. (1996), "Rao–Blackwellization of Sampling Schemes," *Biometrika*, 83, 81–94.

Chen, M.-H., and Dey, D. K. (2000), "Bayesian Analysis for Correlated Ordinal Data Models," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, W. S. Ghosh, and B. Mallick, New York: Marcel Dekker, pp. 135–162.

Cifarelli, D., and Regazzini, E. (1978), "Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative," technical report, Quaderni Istituto Matematica Finanziaria, Torino.

Cowles, M. K., Carlin, B. P., and Connett, J. E. (1996), "Bayesian Tobit Modeling of Longitudinal Ordinal Clinical Trial Compliance Data With Nonignorable Missingness," *Journal of the American Statistical Association*, 91, 86–98.

Doss, H. (1994), "Bayesian Nonparametric Estimation for Incomplete Data via Successive Substitution Sampling," *The Annals of Statistics*, 22, 1763–1786.

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and West, M. (1998), "Computing Nonparametric Hierarchical Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, New York: **???**, pp. 1–22.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

Freedman, R., **???** (2002), "Pilot Study of FLT3 Ligand in Patients With Peritoneal Carcinomatosis Associated With Müllerian and Gastrointestinal Carcinomas," technical report, University of Texas, M. D. Anderson Cancer Center.

Gelfand, A. E., and Kottas, A. (2001), "Nonparametric Bayesian Modeling for Stochastic Order," *Annals of the Institute of Statistical Mathematics*, 53, 865–876.

——— (2002), "A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11, 289–305.

Gelfand, A. E., and Kuo, L. (1991), "Nonparametric Bayesian Bioassay Including Ordered Polytomous Response," *Biometrika*, 78, 657–666.

Giudici, P., Mezzetti, M., and Muliere, P. (2002), "Mixtures of Dirichlet Process Priors for Variable Selection in Survival Analysis," *Journal of Statistical Planning and Inference*, to appear.

Green, P., and Richardson, S. (1998), "Modelling Heterogeneity With and Without the Dirichlet Process," technical report, University of Bristol.

Guglielmi, A., and Tweedie, R. L. (2001), "Markov Chain Monte Carlo Estimation of the Law of the Mean of a Dirichlet Process," *Bernoulli*, 7, 573–592.

Ishwaran, H., and James, L. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96.

Kottas, A., Müller, P., and Quintana, F. (2002), "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," technical report, University of California, Santa Cruz.

Kuo, L., and Mallick, B. (1997), "Bayesian Semiparametric Inference for the Accelerated Failure-Time Model," *The Canadian Journal of Statistics*, 25, 457–472.

Lichtman **???** (1993), **???**

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966.

MacEachern, S. N. (1994), "Estimating Normal Means With a Conjugate Style Dirichlet Process Prior," *Communications in Statistics, Part B—Simulation and Computation*, 23, 727–741.

——— (1998), "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, New York: **???**, pp. 23–44.

——— (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

——— (2001), "Decision Theoretic Aspects of Dependent Nonparametric Processes," in *Bayesian Methods With Applications to Science, Policy and Official Statistics*, ed. E. George, **???**: ISBA, pp. 551–560.

MacEachern, S., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–239.

——— (2000), "Efficient MCMC Schemes for Robust Model Extensions Using Encompassing Dirichlet Process Mixture Models," in *Robust Bayesian Analysis*, eds. F. Ruggeri and D. R. Insua, New York: **???**.

Melamed, M. R., Lindmo, T., and Mendelsohn, M. L. (eds.) (1990), *Flow Cytometry and Sorting*, **???**: Wiley–Liss.

Mira, A., and Petrone, S. (1996), "Bayesian Hierarchical Nonparametric Inference for Change-Point Problems," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press.

Muliere, P., and Petrone, S. (1993), "A Bayesian Predictive Approach to Sequential Search for an Optimal Dose: Parametric and Nonparametric Models," *Journal of the Italian Statistical Society*, 2, 349–364.

Muliere, P., and Tardella, L. (1998), "Approximating Distributions of Random Functionals of Ferguson–Dirichlet Priors," *The Canadian Journal of Statistics*, 26, 283–297.

Müller, P., and Rosner, G. (1997), "A Bayesian Population Model With Hierarchical Mixture Priors Applied to Blood Count Data," *Journal of the American Statistical Association*, 92, 1279–1292.

Regazzini, E., et al. (2002), "Theory and Numerical Analysis for Exact Distributions of Functionals of a Dirichlet Process," *The Annals of Statistics*, 30, 1376–1411.

Sethuraman, J. (1994), "A Constructive Defnition of the Dirichlet Process Prior," *Statistica Sinica*, 2, 639–650.

Tomlinson, G., and Escobar, M. (1999), "Analysis of Densities," technical report, University of Toronto.