

An ANOVA Model for Dependent Random Measures

Maria De Iorio¹, Peter Müller², Gary L. Rosner²,
Steven N. MacEachern³

Abstract

We consider dependent non-parametric 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 combination. We propose a model that describes dependence across random distributions in an 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 (MacEachern, 2002) 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. Since the model structure follows standard ANOVA structure, interpretation and inference parallels conventions for ANOVA models. This includes the notion of main effects, interactions, contrasts, etc. 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. Since 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.

¹Oxford University, Oxford, U.K.

²Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX.

³Department of Statistics, The Ohio State University, Columbus, OH.

Research supported by NIH/NCI grant 2 R01 CA75981-04A1

1 Introduction

We consider dependent non-parametric models for related random probability distributions or functions. We propose a model which describes dependence across random distributions in an ANOVA type fashion. Specifically, assume that random distributions F_x are indexed by a p -dimensional vector $x = (x_1, \dots, x_p)$ 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 non-parametric probability model for F_x in such a way that marginally, for each x , the random measure F_x follows a Dirichlet Process (DP), $DP(M, F_x^\circ)$, with total mass parameter M and base measure F_x° (Ferguson, 1973). But we introduce dependence across x , i.e, dependence for $(F_x, x \in X)$, using the dependent Dirichlet process (DDP) as defined by MacEachern (2002). The random measures F_x are almost surely discrete with the point masses generated marginally from the base measure F_x° . MacEachern (2002) introduces dependence across random measures generated marginally by a DP 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 on the set of 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 $F \sim DP(M, F_0)$. 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) gives a constructive definition of the DP. The weights are generated from rescaled Beta distributions, $w_h / \prod_{i=1}^{h-1} (1 - w_i) \sim Be(1, M)$, and the locations θ_h are i.i.d. samples from the base measure F_0 . Another property that will feature importantly in the following discussion is the Poly Urn representation for the marginal distribution of a sample from a random DP distribution. Assume $y_i \sim F, i = 1, \dots, n$, is sampled from an unknown distribution F which, in turn, is generated by a DP, $F \sim DP(M, F_0)$. The marginal distribution of $y = (y_1, \dots, y_n)$ is described by the following Polya Urn scheme (Blackwell and MacQueen, 1973): $y_1 \sim F_0$ and

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

The m -th sample point is either a tie with a previous sample y_h , or it is a new draw from the base measure. The positive probability of ties in (1) is due to the discrete nature of the random distribution $F \sim DP(M, F_0)$. In many data analysis applications this discreteness is inappropriate. DP mixture models (MDP) avoid this discreteness by adding an additional convolution with a continuous kernel. The typical MDP model assumes

$$y_i \stackrel{iid}{\sim} G \text{ with } G(y) = \int f(y|\mu) dF(\mu), \quad F \sim DP(M, F_0), \quad (2)$$

i.e., 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 normal model $G(y) = \sum_{h=1}^{\infty} w_h N(\mu_h, S)$. One of the main attractions of MDP models like (2) is computational simplicity. Also, posterior simulation algorithms are dimension independent. See, for example, Escobar and West (1998) or MacEachern and Müller (2000) for a review of models based on (2).

Several papers have considered extension of DP and MDP models to hierarchical models over related random distributions. In the context of parametric models, i.e., models with finite dimensional parameter vector, such hierarchies with submodels for related experiments are standard modeling tools. In non-parametric models such extensions are complicated by the infinite dimensional nature of the random distribution. Some of the first developments of dependent DP (and DP mixture) models appear in Cifarelli (1979), Cifarelli, Muliere and Scarsini (1981) and Muliere and Scarsini (1983). Muliere and Petrone (1993), define dependent non-parametric models $F_x \sim DP(M, F_x^o)$ by assuming a regression in the base measure $F_x^o = N(\beta x, \sigma^2)$. A similar strategy of linking dependent random DP measures F_x at the level of the base measure is used in Tomlinson and Escobar (1999). They achieve increased flexibility by assuming a MDP hyperprior on the common base measure. Gelfand and Kottas (2001b) define dependent non-parametric models by considering a representation of random measures as products of DP distributed factors. This allows them to enforce stochastic ordering. A similar approach, but using additive decomposition and without the stochastic order constraint, is used in Müller et al. (1999). Tomlinson and Escobar (1999), Gelfand and Kottas (2001b) and Müller et al. (1999) are appropriate to model dependence across several related random measures. They are not naturally extended to include regression on covariates.

MacEachern (2002) defines 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 (2002) defines a probability 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 i.i.d. realizations of a stochastic process $p(\theta_h)$. 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 non-parametric model which allows a constructive definition by a stick breaking algorithm as in Sethuraman (1994). See, for example, Green and Richardson (1998), Ishwaran and James (2001), or Neal (2000) 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. Section 4 illustrates the proposed models with two examples. We use a simulated data set and longitudinal data with non-parametric random effects distributions. 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 random effects distribution. 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 which allows to build an ANOVA type dependence structure. For example, we want 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 to incorporate prior information about the presence of interactions, i.e., whether the effect of $v = v_1$ should be allowed to depend on the level of the other covariate w , etc. The model needs to give a formal definition to notions like “main effect”, “interaction”, etc. In short, the desired process should allow to transfer interpretation and structure used for unknown normal means in a traditional ANOVA model to unknown random functions.

We achieve this 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, $w_h / \prod_{i=1}^{h-1} (1 - w_i) \sim Be(1, M)$. On the locations θ_{xh} we impose an additional structure, writing

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

with $m_h \stackrel{iid}{\sim} p_m^\circ(m_h)$, $A_{vh} \stackrel{iid}{\sim} p_{Av}^\circ(A_{vh})$, and $B_{wh} \stackrel{iid}{\sim} p_{Bw}^\circ(B_{wh})$, with independence being across h , v and w . We refer to the joint probability model on \mathcal{F} as $(F_x, x \in X) \sim \text{ANOVA DDP}(M, p^\circ)$. The model is parametrized by the total mass parameter M and the base measure p° on the ANOVA effects in (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, $A_1 = B_1 \equiv 0$. Marginally, for each $x = (v, w)$, the random distribution F_x follows a DP with mass M and base measure F_x° given by the convolution of p_m° , p_{Av}° and p_{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_m , p_{Av} and p_{Bw} . For example, consider two treatment combinations $x = (v, 1)$ and $x' = (v, 2)$ and random samples $y \sim F_x$ and $y' \sim F_{x'}$. Assuming normal priors $p_m = N(\mu_m, \sigma_m^2)$, $p_{Av} = N(\mu_{Av}, \sigma_A^2)$, and $p_{Bw} = N(\mu_{Bw}, \sigma_B^2)$ we find the marginal correlation of the two samples as $\text{corr}(y, y') = (\sigma_m^2 + \sigma_A^2) / (\sigma_m^2 + \sigma_A^2 + \sigma_B^2)$. Including hyperpriors on σ_m, σ_A and σ_B allows inference about the level of dependence, subject to the defined structure. Treating the locations μ_m, μ_{Av} and μ_{Bw} as unknown hyperparameters

allows inference on the overall location of F_x independent of inference on the dependence across F_x .

Model (3) is not constrained to univariate distributions F_x . The point masses θ_{x_h} and the ANOVA effects m_h, A_{vh}, 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 an hierarchical model. In the example discussed in Section 4.2 we use 7-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 more complex ANOVA structure. The model is easily generalized to a p -dimensional categorical covariate $x = (x_1, \dots, x_p)$:

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

where $A_i(x_i)$ is the main effect due to treatment x_i . Further extensions to include interactions A_{ij} , etc., are equally straightforward.

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 $p_{A_v}^\circ$ for the different levels of v . In the context where v indicates a control or one of a number of exchangeable treatments, we might take $p_{A_v}^\circ$ 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 $p_{A_v}^\circ$ to have non-zero 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 anti cancer 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 convolute the discrete measure generated from a DP with a continuous, typically normal, kernel. Such models are known as DP mixture (MDP) models. See, for example, MacEachern and Müller (2000) for related references. For the same reasons we propose to add an additional normal mixture to the ANOVA DDP model. This leads to models of the form

$$(y_i|x_i = x) \sim \int N(y|\mu, S) dF_x(\mu), \quad (F_x, x \in X) \sim \text{ANOVA DDP}(M, p^\circ), \quad (4)$$

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

Implementation of posterior inference in the ANOVA DDP model (4) is easiest developed on the basis of an equivalent reformulation of the model as a mixture of ANOVA models. Denote with $\alpha_h = (m_h, A_{1h}, \dots, A_{Vh}, B_{1h}, \dots, B_{Wh})$ the parameter vector in the ANOVA model (3) for the h -th point mass in the random measures. Let d_i denote a design vector to select the appropriate ANOVA effects corresponding to x_i , i.e., $\theta_{xh} = \alpha'_h d_i$ for $x_i = x$. Using this notation, model (4) with base measure $(p_m^\circ, p_{A_v}^\circ, p_{B_w}^\circ)$ can be rewritten as:

$$(y_i | x_i = x) \sim \int N(y | \alpha d_i, S) dF(\alpha), \quad F \sim DP(M, p^\circ). \quad (5)$$

In words, data y_i is 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 (5) by introducing latent variables α_i ,

$$y_i = \alpha_i d_i + \epsilon_i, \quad \alpha_i \sim F \text{ and } F \sim DP(M, p^\circ), \quad (6)$$

with $\epsilon_i \sim N(0, S)$. 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 (4). Using normal priors for the base measure p° , and an additional normal kernel as in (4) leads to a straightforward Gibbs sampling scheme. The conjugate nature of the base measure p° and the kernel in the error distribution $p(\epsilon_i)$ in (6) greatly simplify posterior simulation. See, for example, ?? for details. A brief summary is given in the appendix.

3.2 Other Mixture of ANOVA Models

Rewriting the ANOVA DDP as (6) 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 parametrization of the DP avoids difficult prior elicitation problems. On the other hand, the fact that the DP is parametrized by a base measure and one scalar precision parameter M only could sometimes be a limitation. Green and Richardson (1998) and Neal (2000) argue for the use of more general mixture models and show appropriate posterior simulation schemes. Tardella and Muliere (1998), Gelfand and Kottas (2001a) and Ishwaran and James (2001) discuss finite truncations of DP priors. Ishwaran and James (2001) propose alternative non-parametric priors based on similar stick-breaking representations. Any of these non-parametric priors can be substituted in (6) without changing the model structure, and requiring only minimal changes in the posterior simulation schemes.

3.3 Hierarchical Models

Consider a generic hierarchical model of the form

$$y_i \sim p(y_i | \theta_i), \quad \theta_i | x_i = x \sim H_x(\theta | \phi). \quad (7)$$

In words, data y_i for the i -th sampling unit, e.g., a patient, is sampled from a probability model parametrized by a random effects vector θ_i . For example, this could take the form of a non-linear regression

$$y_{ij} = f(t_{ij}; \theta_i) + \epsilon_{ij} \quad (8)$$

with a mean function $f(\cdot; \theta)$ parametrized by θ_i and evaluated at known times t_{ij} , $j = 1, \dots, n_i$. 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 i -th patient, then ANOVA DDP models as in (4) are appropriate prior probability models for $(H_x, x \in X)$. The random effects vector θ_i takes the place of y_i , and $H_x = \int N(\cdot | \mu, S) dF_x(\mu)$. 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 i -th 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 (4), with an additional step to update the random effects vectors θ_i which now replace the data y_i in the ANOVA DDP model. The details of this step are problem specific. For example, if $p(y_i | \theta_i)$ is a non-linear regression as in (8) then updating θ_i amounts to a posterior draw from the parameters in a non-linear regression with data y_{ij} , $j = 1, \dots, n_i$ and prior $\theta_i \sim N(\alpha_i d_i, S)$. Here α_i is a latent variable introduced to break the mixture model in (5). 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.

4 Examples

4.1 Simulation Example

Consider a two way ANOVA model with two factors v and w . Assume the number of levels are $V = 3$ and $W = 2$ for v and w , respectively. Let $\alpha = (m, A_1, A_2, B_1, B_2, B_3)$ denote overall mean and main effects for $v = 1, 2, 3$ and $w = 1, 2$, respectively. We simulated 100 observations with randomly selected designs and generating $\alpha_i \sim 0.5 \delta(5, 0, 3, 5, 0, -4) + 0.5 \delta(-9, 0, -3, -5, 0, 4)$, i.e., we generate from model (5) with a two point discrete mixing measure $F(\alpha)$. Define $d_{A_2} = (0, 1, 0, \dots, 0)$ as the design vector which selects the effect A_{2h} and let $F_{A_2}(\cdot) \equiv F(\alpha' d_{A_2})$ denote the unknown mixing distribution of A_{2h} , and analogous definitions for other ANOVA effects. Figure 1 shows the posterior estimated distributions $E(F_{A_2} | y)$ and $E(F_{B_2} | y)$. Compare the (correctly) bimodal posterior estimated distribution with a standard ANOVA estimate of $\hat{A}_2 = 0.59$ and $\hat{B}_2 = 0.42$, using maximum likelihood estimation and the same identifiability constraints $A_1 = B_1 = 0$.

4.2 Hierarchical Models

Müller and Rosner (1997) describe a hematologic study. The data records white blood cell counts over time for each of n chemotherapy patients. Denote with y_{it} the measured response on day t for patient i . The profiles of white blood cell counts over time look

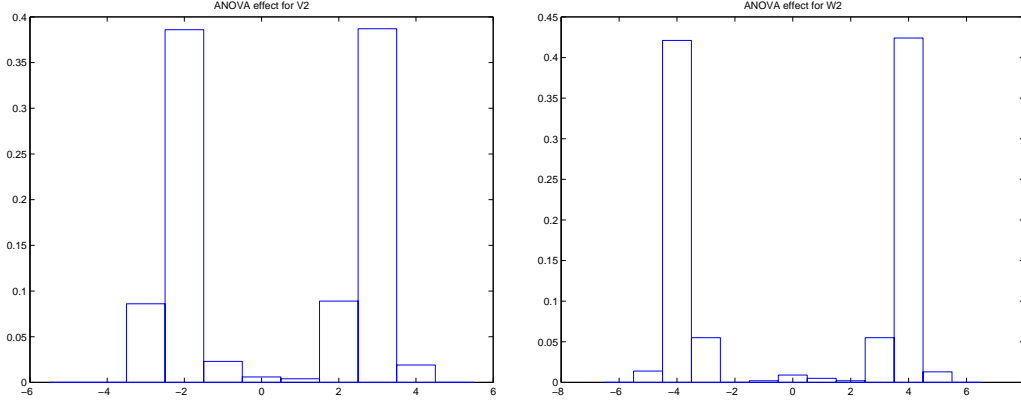


Figure 1: Posterior estimated distributions of ANOVA effects $\hat{p}(V_{2h}|data)$ and $\hat{p}(W_{2h}|data)$.

similar for most patients. Figure 2 shows some typical patients. There is an initial base line, followed by a sudden decline when chemotherapy starts, and a slow, S-shaped recovery back to approximately base line after end of the treatment. Profiles can be reasonably well approximated with a piecewise linear-linear-logistic regression, using a 7-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.

$$y_{it} = f_{\theta_i}(t) + \epsilon_{it}. \quad (9)$$

The model is completed with a random effects model H_x . The random effects distribution H_x depends on the treatment levels (v_i, w_i) . There are two treatments, the actual anti-cancer agent cyclophosphamide (CTX), and a second drug (GM-CSF) which is 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 level of CTX and GM-CSF. Let $x_i = (v_i, w_i)$ denote the treatment for patient i . We assume

$$\theta_i|x_i = x \sim H_x(\theta), \quad (H_x, x \in X) \sim \text{ANOVA DDP}(M, p^\circ) \quad (10)$$

Posterior predictive inference for future patients depends on the observed historical data only indirectly through learning about the random effects distribution (10). 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 effect the desired learning.

Figure 3 summarizes some critical aspects of the analysis. In this example, the random effects distributions $H_x(\cdot)$ are 7-dimensional. We summarize inference on $H_x(\cdot)$ by showing implied profiles. For example, for $x = (CTX = 6, GM = 5)$ we show implied inference for $f(t) = E \{ \int f_\theta(t) dH_x(\theta) | data \}$, the estimated distribution of hematologic profiles over time t for a patient treated with doses $x = (6, 5)$. To define design vectors, as in (5) we index the levels of CTX as $v = 1, \dots, 4$ and the levels for GM as $w = 1, 2$ and use $A_{1h} = B_{1h} \equiv 1$ as identifiability constraint. Thus the corresponding design vector in (5) is $d = (1, 0, \dots, 0)$, including a main effect only. The figure shows the estimated mean curve $f(t)$ in the right lower panel.

For patients treated at other doses, we display corresponding offsets in a familiar ANOVA fashion. For example, for $x = (1.5, 5)$ the figure plots the posterior expected mean curve corresponding to design vector $d = (0, 1, 0, 0, 1)$ with offsets for $v = 2$ and $w = 2$. The second panel in the top row shows $df(t) = E \{ \int f_{\alpha_d}(t) dF(\alpha) | data \}$. The other panels in the same figure have analogous interpretations.

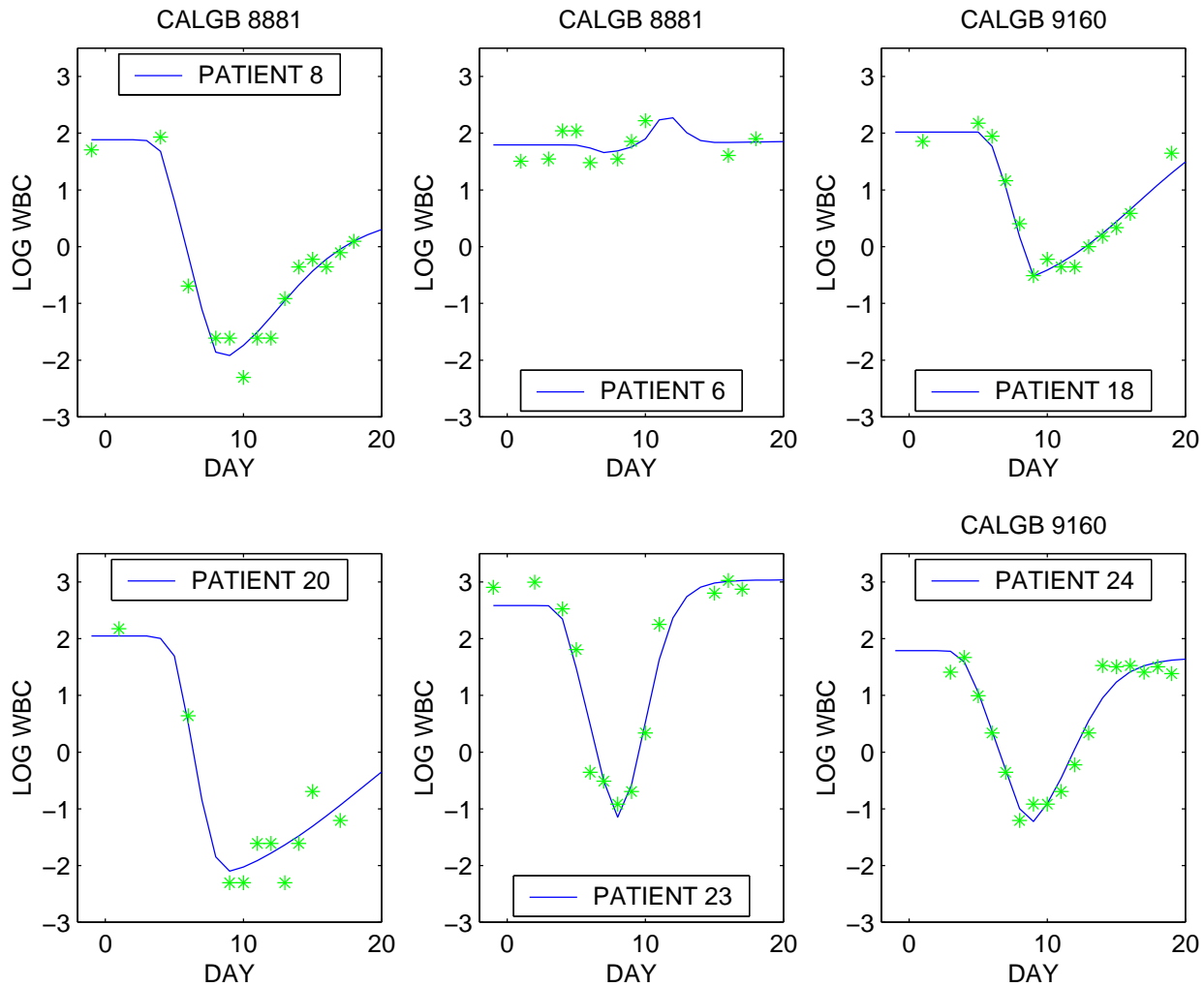


Figure 2: Some typical patients.

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 of MCMC simulation for posterior inference, and the practical limitation to stick-breaking priors for the non parametric mixing

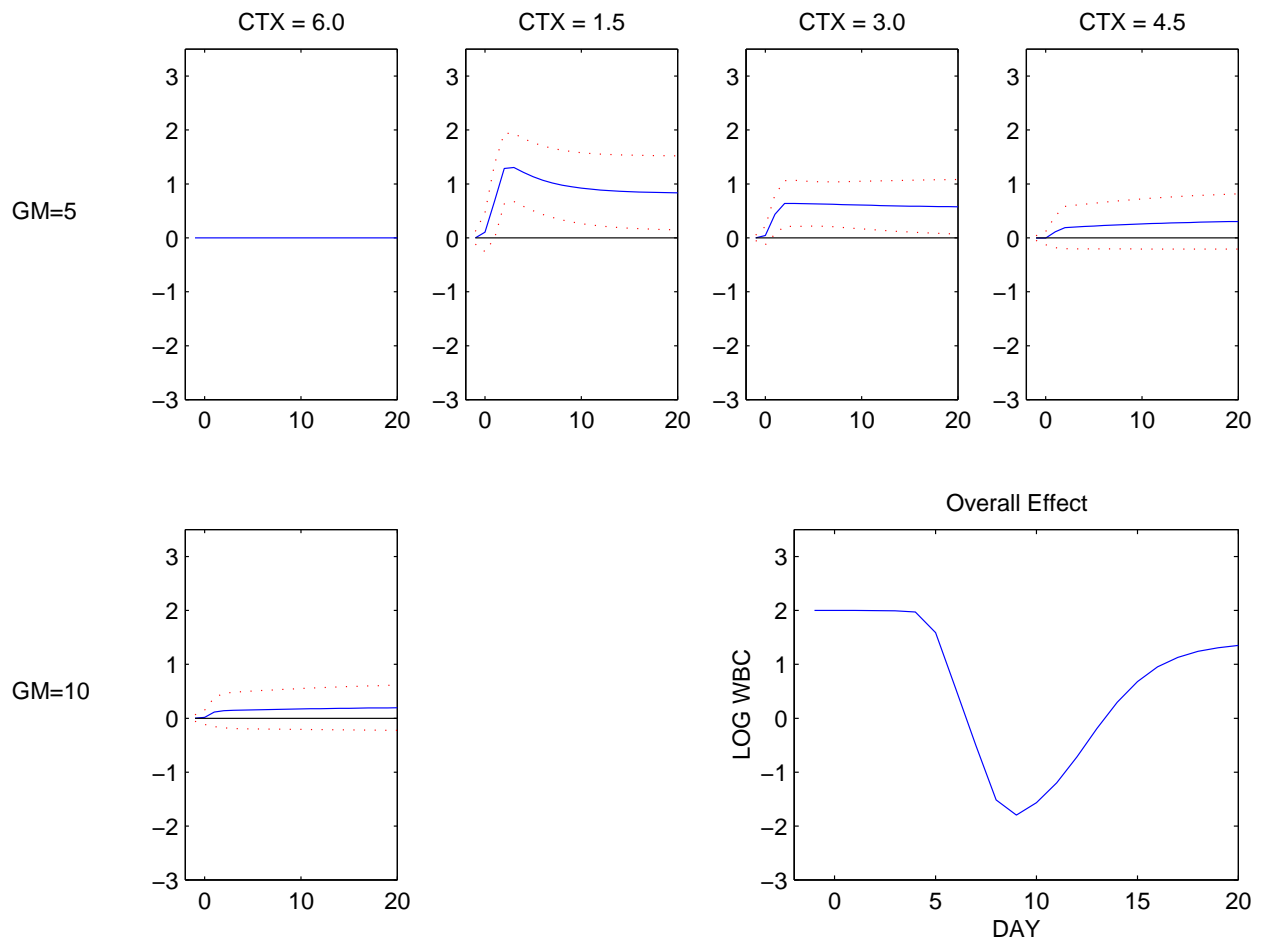


Figure 3: Estimated main effects A_v and overall mean effect m . For identifiability, we constrained $A_{1h} = B_{1h} \equiv 0$. The right lower panel shows the estimated hematologic profile for a patient treated with dose levels ($CTX = 6, GM = 5$). The other figures summarize the offset corresponding to the respective main effect. See the text for a detailed explanation.

measure. Also, the model inherits limitations inherent in the DP prior. For example, the weights in the mixture are *a priori* stochastically ordered. 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 parametrization facilitates prior elicitation.

Appendix

We briefly describe the implementation of posterior simulation in the ANOVA DDP model (5). Since 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 $k \leq n$ 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 , i.e. n_j is the size of the j th cluster and let $\Gamma_j = \{i : s_i = j\}$. MacEachern and Müller propose a Gibbs sampling scheme to estimate MDP models':

1. *Resampling s_i given all the other parameters:*

marginalize over α_i and sample s_i from

$$\Pr(s_i = j \mid \alpha_i, s_{-i}, y, \eta, \Sigma) \propto \begin{cases} n_j^- p(y_i \mid s_i = j; \Sigma; y_l : l \in \Gamma_j, l \neq i) & j = 1, \dots, k^- \\ M \int N(y_i; \alpha d_i, \Sigma) dF_0(\alpha) & j = k^- + 1 \end{cases}$$

where

$$\begin{aligned} \alpha_{-i} &= \{\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_n\} \\ s_{-i} &= \{s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n\} \\ n_j^- &= \begin{cases} n_j - 1 & \text{if } j = s_i \\ n_j & \text{otherwise} \end{cases} \\ k^- &= \# \text{ of clusters with } \alpha_i \text{ removed} \end{aligned}$$

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

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

2. *Resampling α_j^* :*

the posterior distribution given all the other parameters, $p(\alpha_j^* \mid s, y, \eta, \Sigma)$ is

$$p(\alpha_j^* \mid s, y, \eta, \Sigma) \propto \left[\prod_{i \in \Gamma_j} N(y_i; \alpha_j^* d_i, \Sigma) \right] F_0(\alpha_j^* \mid \eta)$$

If $F_0(\alpha_j^* \mid \eta)$ is Normal then also the posterior distribution of α_j^* is Normal with mean θ and covariance matrix C that can be calculated through recursive equations. For $i \in \Gamma_j$, set

$$F_i = d_i \otimes I_N$$

where I_N is the $N \times N$ identity matrix and \otimes denotes the kronacker product. Let $\tilde{\alpha}_j^*$ be the column vector obtained by writing each column of α_j^* one after the other. We obtain the linear regression model defined by

$$\begin{aligned} y_i &= F_i \tilde{\alpha}_j^* + \omega & \omega &\sim N(\omega; \mathbf{0}, S), \quad i \in \Gamma_j \\ \tilde{\alpha}_j^* &\sim N(\tilde{\alpha}_j^*; a; R) \end{aligned}$$

for some prior mean a and prior covariance matrix R . Therefore the posterior mean and covariance matrix of $\tilde{\alpha}_j^*$ can be updated sequentially using standard result of Normal theory. The same holds for $p(y_i | s_i = j; \Sigma; y_l : l \in \Gamma_j, l \neq i)$, but in this case we want to evaluate a predictive distribution.

3. We might wish to express uncertainty about the total mass parameter M (see West, 1992 for prior elicitation and posterior inference on M), η and Σ . In this last two cases, conditioning on all other parameters leaves a standard Bayes model.

References

- Antoniak, C. E. (1974), “Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems,” *Annals of Statistics*, 2, 1152–1174.
- Blackwell, D. and MacQueen, J. B. (1973), “Ferguson distributions via Pólya urn schemes,” *The Annals of Statistics*, 1, 353–355.
- 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, 1–22, New York, NY, USA.
- Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, 1, 209–230.
- Gelfand, A. E. and Kottas, A. (2001a), “A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models,” *Journal of Computational and Graphical Statistics*, to appear.
- (2001b), “Nonparametric Bayesian Modeling for Stochastic Order,” *Annals of the Institute of Statistical Mathematics*, to appear.
- Green, P. and Richardson, S. (1998), “Modelling heterogeneity with and without the Dirichlet process,” Technical report, University of Bristol.
- Ishwaran, H. and James, L. (2001), “Gibbs sampling methods for stick-breaking priors,” *Journal of the American Statistical Association*, 96.
- MacEachern, S. (2002), “Dependent Nonparametric Processes,” *Journal of the American Statistical Association*.
- MacEachern, S. and Müller, P. (1998), “Estimating mixture of Dirichlet process models,” *Journal of Computational and Graphical Statistics*, 7, 223–239.
- MacEachern, S. N. and Müller, P. (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, NY, USA.

- 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.
- Müller, P., Quintana, F., and Rosner, G. (1999), “Hierarchical Meta-Analysis over Related Non-parametric Bayesian Models,” Technical report, Duke University ISDS, USA.
- 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.
- Neal, R. M. (2000), “Markov chain sampling methods for Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, 9.
- Sethuraman, J. (1994), “A constructive definition of the Dirichlet process prior,” *Statistica Sinica*, 2, 639–650.
- Tardella, L. and Muliere, P. (1998), “Approximating distributions of random functionals of Ferguson-Dirichlet priors,” *The Canadian Journal of Statistics*, 26.
- Tomlinson, G. and Escobar, M. (1999), “Analysis of densities,” Technical report, University of Toronto.