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, marginalized, for each $x$, the random measure $F_x$ follows a Dirichlet process (DP), $DP(M, F_0)$, with total mass parameter $M$ and base measure $F_0$ (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 $F_0$. 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 $F \sim DP(M, F_0)$. Letting $\delta(x)$ denote a point mass at $x$, we can write $F = \sum_{h=0}^{\infty} w_\theta \delta(\theta_h)$. Here $w_\theta$ are the weights of point masses at locations $\theta_h$.

Sethuraman (1994) gave a constructive definition of the DP. The weights are generated from rescaled Beta distributions, $w_\theta/\prod_{i=0}^{\theta_h-1} (1 - w_i) \sim Be(1, M)$, and the locations $\theta_h$ 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 $y_i \sim F_i$, $i = 1, \ldots, n$, 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 $y = (y_1, \ldots, y_m)$ is described by the following Pólya urn scheme (Blackwell and MacQueen 1973): $y_1 \sim F_0$ and

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

The $m$th 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 $F \sim DP(M, F_0)$.

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

$$y_i \sim H, \quad \text{with } H(y) = \int f(y | \mu) dF(\mu),$$

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 $H(y) = \sum_{h=1}^{\infty} w_\mu N(\mu_h, S)$. One of the main attractions of DPM 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 considered extension of DP and DPM mod-
elors to hierarchical models over related random distributions. In
the context of parametric models, that is, models with finite-
dimensional parameter vector, such hierarchies with submod-
els 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 mod-
els \( F_x \sim \text{DP}(M, F_0) \) by assuming a regression in the base
measure, \( F_0^x = N(\beta x, \sigma^2) \). 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 Tom-
linson 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 natu-
rally 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 probability model for \( \mathcal{F} \) such
that marginally, for each \( x \), \( F_x = \sum w_h \delta(\theta_h) \) follows a DP.
We use an additional subindex \( x \) for the point masses \( \theta_h \)
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 stochas-
tic process (in \( x \)). For example, \( \theta_{xh} \) might be assumed to be a
Gaussian process. Independence across \( h \), together with the
stick-breaking prior for the weights \( w_h \), guarantees that \( F_x \) mar-
ginally follows a DP. Dependence in the sample path of the
stochastic process \( \theta_h \) introduces the desired dependence across \( x \).
We use this DDP structure to develop an ANOVA-like probabil-
ity 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 computa-
tional 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 distribu-
tions.
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 distrib-
utions 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 distrib-
utions, indexed by a categorical covariate \( x \). For simplicity of
explanation, assume for the moment that \( x = (v, w) \) is bivari-
te with \( v \in \{1, \ldots, V\} \) and \( w \in \{1, \ldots, W\} \). 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 dis-
tributions for recorded measurements on each patient or distrib-
utions 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 \( x' =
(v_2, w_2) \) to share a common main effect due to the common fac-
tor \( v_1 \). The model should allow us to incorporate prior informa-
tion 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 re-
sponse. 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 \( y_{xi}, i = 1, 2, 3, \ldots, \) are observed data, with
\( y_{xi} \) denoting the \( i \)th observation under condition \( x \), for exam-
ple, the response of the \( i \)th patient assigned treatment com-
bination \( x \) in a clinical trial. If, for each \( x \), the subsequence
\( y_{x1}, y_{x2}, \ldots \) 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 non-
parametric prior for \( F_x \).
We achieve the desired model structure by using the DDP
framework. Specifically, let \( F_x = \sum w_{bh} \delta(\theta_{bh}) \) for \( x = (v, w) \).
We assume Sethuraman’s (1994) stick-breaking prior for the
common weights, \( w_{bh}/\prod_{i=1}^{n-1}(1 - w_i) \sim \text{Be}(1, M) \). We impose
additional structure on the locations \( \theta_{bh} \):

\[
\theta_{bh} = m_h + A_{vh} + B_{wh}.
\]

(3)
As in standard ANOVA models we need to introduce an iden-
tifiability constraint for interpretability. We may impose any
of the standard constraints, for example, \( A_{1h} = B_{1h} = 0 \). For
the remaining parameters we assume \( m_h \sim \text{p}_{n}^{\text{id}}(m_h), A_{vh} \sim
\text{p}_{A}^{\text{id}}(A_{vh}) \), and \( B_{wh} \sim \text{p}_{B}^{\text{id}}(B_{wh}) \), with independence being

\[
\text{p}_{A}^{\text{id}}(A_{vh}), \text{p}_{B}^{\text{id}}(B_{wh}) \), with independence being
across $h$, $v$, and $w$. We refer to the joint probability model on $\mathcal{F}$ as $(F_x, x \in X) \sim \text{ANOVA DDP}(M, p^0)$. The model is parameterized by the total mass parameter $M$ and the base measure $p^0$ 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 $F^0$ given by the convolution of $p^0_m$, $p^0_{Av}$, and $p^0_{Bw}$. Model (3) defines dependence across $x$ by defining the covariance structure of the point masses $\theta_{xh}$ 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 $p^0_m$, $p^0_{Av}$, and $p^0_{Bw}$. For example, consider two treatment combinations $x = (v, 1)$ and $x' = (v, 2)$ and random samples $y_{xv} \sim F_x$ and $y_{x'v} \sim F_{x'}$. Assuming normal priors $p^0_m = N(\mu_m, \sigma^2_m), p^0_{Av} = N(\mu_{Av}, \sigma^2_{Av})$, and $p^0_{Bw} = N(\mu_{Bw}, \sigma^2_{Bw})$, we find the marginal covariance $\text{cov}(y_{x}, y_{x'}) = (\sigma^2_m + \sigma^2_{Av})/(M + 1)$. Compare this with the covariance $\text{cov}(y_{x}, y_{x'}) = (\sigma^2_m + \sigma^2_{Av})$ that would arise in a standard ANOVA with the base measure $p^0$ 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 $\theta_{xh}$ and the ANOVA effects $m_n, A_vh$, 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 $x = (x_1, \ldots, x_p)$:

\[ \theta_{xh} = m_h + \sum_{i=1}^{p} A_i(x_i), \]

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 $p^0_{Av}$ for the different levels of $v$. In the context where $w = 1$ indicates a control and $w = 2, \ldots, V$ are exchangeable treatments, we might take $p^0_{Av}$ 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 $p^0_{Av}$ 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 $A_{xh}$ to lie on a line (to do so, we need to violate the condition of independence of $A_{xh}$ across levels of $v$) produces a lower-dimensional component in the model. Alternatively, a constraint such as monotonicity of the effect $A_{xh}$ 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

\[ (y_i | x_i = x) \sim H_x(y_i), \quad \text{with} \quad H_x(y) = \int N(y | \mu, \Sigma) dF_x(\mu), \]

\[ (F_x, x \in X) \sim \text{ANOVA DDP}(M, p^0), \]

with appropriate hyperpriors for the common normal variance $\Sigma$ 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 $q$-dimensional random measures $F_x$. Let $N$ denote the number of ANOVA effects in (3) and let $a_h = [m_h, A_{xh}, \ldots, A_{yv}, B_{wh} \ldots, B_{wh}]$ denote the $(q \times N)$ matrix of ANOVA parameters corresponding to the $h$th point mass in the random measures. Let $d_i$ denote a design vector to select the appropriate ANOVA effects corresponding to $x_i$, that is, $\theta_{xh} = a_hd_i$. Using this notation, model (5) with base measure $(p^0_m, p^0_{Av}, p^0_{Bw})$ can be rewritten as

\[ (y_i \mid x_i = x) \sim H_x(y_i), \]

\[ H_x(y) = \int N(y | \alpha d_i, \Sigma) dF(\alpha), \]

\[ F \sim \text{DP}(M, p^0). \]

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 $\alpha_i$:

\[ y_i = \alpha_i d_i + \epsilon_i, \quad \alpha_i \sim F, \quad F \sim \text{DP}(M, p^0), \]
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 (5). The conjugate nature of the base measure \( \rho^0 \) 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 means \( E(H_x(y)) \) 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 \( p(Y_{\text{new}} | x, Y) \) 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 from 5 to 10, and 25, we find the prior mean \( \bar{x} \) and standard deviation \( \sigma_x \) for the number 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 \( k \leq \sum k_x \) 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, \( M \sim \text{Ga}(\alpha, \beta) \) 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 \( n_x = six observations and assuming M \sim \text{Ga}(\alpha, \beta) \), we find the prior mean \( E(k_x) \) (standard deviation SD) to be 4.8 (1.0), 2.5 (1.2), 4.9 (0.91), and 5.5 (0.68), 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 \( \alpha \) 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. Miliere and Tardella (1998), Ishwaran and James (2001), and Gelfand and Kottas (2002) discussed fine 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

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

(8)

In words, data \( y_i \) for the \( i \)th sampling unit, for example a patient, is sampled from a probability model parameterized by a random effects vector \( \theta_i \). For example, this could take the form of a nonlinear regression

\[
y_{ij} = f(t_{ij}; \theta_i) + \epsilon_{ij},
\]

(9)

with a mean function \( f(\cdot; \theta) \) parameterized by \( \theta_i \) and evaluated at known times \( t_{ij}, j = 1, \ldots, n_i \). The \( \theta_i \) 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 \( \phi \).

If the covariates are, for example, treatment indicators for the \( i \)th patient, then ANOVA DDP models as in (5) are appropriate prior probability models for \( (H_x, x \in X) \). The random effects vector \( \theta_i \) takes the place of \( y_i \) in (5), and \( H_x(\cdot) = \int N(\cdot | \mu, S) \, dF_x(\mu) \). In general, the ANOVA DDP model can be used whenever the random effects distributions \( H_x \) are indexed by some categorical covariates \( x \) 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 (5), with an additional step to update the random effects vectors \( \theta_i \), 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 \( \theta_i \) amounts to a posterior draw from the parameters in a nonlinear regression with data \( y_{ij}, j = 1, \ldots, n_i \), and prior \( \theta_i \sim N(\alpha_i d_i, S) \). Here \( \alpha_i \) is the latent variable introduced to break the mixture model in (7). The latent variable \( \alpha_i \) 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 \( \delta_i \), 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 \( x \). 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 \( -\infty = \theta_0 < \cdots < \theta_R = \infty \) and assume \( z_{ik} = r \) if and only if \( \theta_{r-1} < y_{ik} \leq \theta_r \). See, for example, Albert and Chib (1993), Cowles, Carlin, and Connell (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 \( \theta_r \) 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 rats. The DDP ANOVA model (5) formalizes a nonparametric regression of \( z_i \) on the covariates \( x_i \). The categorical cell probabilities \( \pi_{z|x} = Pr(z|x) \) are modeled as corresponding quantities of \( H_x \), that is, \( \pi_{z|x} = \int dH_x(y) \) 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 \( x_{n+1} \) with unknown label \( y_{n+1} \), 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 \( \alpha = [\alpha_1, \alpha_2, \beta_1, \beta_2] \) 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 \( y_i \sim N(\alpha_i, d_i, I) \) with randomly selected designs \( d_i \) and generating
\[
\alpha_i = \begin{bmatrix} 3 & 2 & 5 & -1 & -1 \\ 1 & 6 & 6 & 7 & 4 \end{bmatrix}, \quad \text{with probability .5},
\begin{bmatrix} 3 & .5 & 3.5 & -2.5 & -2.5 \\ 1 & 7.5 & 7.5 & 8.5 & 5.5 \end{bmatrix}, \quad \text{with probability .5},
\]
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 \( H_{vw}(\cdot) = \int N(\cdot|\mu, S) dF_{vw}(\mu) \) 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 \( p^\theta \) 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 overlaid 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.

4.2 Hierarchical Models

Müller and Rosner (1997) described an analysis of hematology 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 that 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 \( \theta \) (Müller and Rosner 1997). But the nonlinear regression parameters differ significantly across patients. Thus we introduce a patient-specific random effects vector \( \theta_i \). Conditional on \( \theta_i \), we assume a nonlinear regression using the piecewise linear–linear–logistic regression model
\[
y_{it} = f(t; \theta_i) + \epsilon_{it}.
\]

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 \( \mu \)g/kg).
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

$$
(\theta_i | x_i = x) \sim H_x(\theta),
$$

(11)

$$
(H_x, x \in X) \sim \text{ANOVA DDP}(M, p^*).
$$

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 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 $f_{xi}(t) = E\{f(t; \theta) \mid \theta_{H}(\theta) | y\}$ for these patients.
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 $f_x(t) = E\{f(t; \theta) \mid dH_x(\theta)|Y\}$, the posterior mean hematologic profile over time $t$ for a patient treated with doses $x = (1, 1)$. Imposing identifiability constraints $A_{1b} = B_{1b} \equiv 0$, the corresponding design vector in (6) is $d = (1, 0, \ldots, 0)$, 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 $d = (0, 1, \ldots, 0)$ with an offset for $v = 2$. The second panel in the top row shows $f_x(t) = E\{f(t; \theta = \alpha d) \mid Y\}$. 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, where $z_i \in [0, 1]$ reporting presence and absence of the $k$th surface marker.

We set up a simulation study mimicking the setup in a currently ongoing clinical trial. The trial proposes chemoinmunotherapy 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 activation of monocytes. For this simulation study we consider two markers and samples under four different treatment conditions $x_i, x_i \in \{G, I\} \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. 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 $I \ast G$. We generated hypothetical data for four blood samples, one under each treatment combination. We assume 100 cells are measured in each sample, with observed frequencies reported in Table 1(a). Let...
\( \pi_{zx} = \Pr(z | x) \) denote the probability of binary outcomes \( z \) under subpopulation \( x \). Using the DDP ANOVA model, the cell probabilities \( \pi_{zx} \) 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 \( p^0 \) 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 significantly more flexibility in fitting the observed frequencies.

### 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 \( \alpha_i \). Write \( \{\alpha_1^{*}, \ldots, \alpha_k^{*}\} \) for the \( k \leq n \) distinct elements in \( \{\alpha_1, \ldots, \alpha_n\} \). Set \( s_j = j \) if \( \alpha_i = \alpha_j^{*} \). Let \( n_j \) be the number of \( s_j \) equal to \( j \), that is, \( n_j \) is the size of the \( j \)th cluster, and let \( G_j = \{ i : s_i = j \} \). Let \( \eta \) denote possibly unknown hyperparameters, including hyperparameters in the base measure \( p^0 \), the total mass parameter \( M \), and the covariance matrix \( \Sigma \) 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 \( s_j \) given all the other parameters.** Marginalize over \( \alpha_i \) and sample \( s_j \) from

\[
\Pr(s_j = j | s_{-j}, \eta, S, \eta) \propto \begin{cases} 
\int N(\gamma_i; \alpha d_i, S) d\gamma_i, & j = 1, \ldots, k^- \\
M & j = k^- + 1,
\end{cases}
\]

where

\[
\begin{array}{l}
\gamma_i | s_j = j \sim \mathcal{N}(\alpha d_i, S) \\
\gamma_i | s_j = j \sim \mathcal{N}(0, \Sigma) \\
\end{array}
\]
where \( y_{-i} = \{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\} \), \( s_{-i} = \{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_j\} \),

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

and \( k^- \) is the number of clusters with \( \alpha_i \) removed. If \( n_{si} = 0 \), we relabel the remaining clusters \( j = 1, \ldots, k^- = k - 1 \). After sampling \( s_i \), set

\[
k = \begin{cases} 
  k^- & \text{if } j \leq k^- \\
  k^- + 1 & \text{if } j = k^- + 1.
\end{cases}
\]

Note that, by conditional independence, in \( p(y_j | s_j = j, s_{-j}, S, \eta, y_{-j}) \) the data \( y_{-i} \) can be replaced by \( (y_j, i \in \Gamma_j, j \notin \Gamma \).

2. Resampling \( \alpha^*_j \). The posterior distribution of \( p(\alpha^*_j | s, \eta, S) \) is

\[
p(\alpha^*_j | s, \eta, S) \propto \left[ \prod_{j \in \Gamma_j} N(y_j; \alpha^*_j d_x, S) \right] p(\alpha^*_j | \eta).
\]

With a normal base measure \( p^0(\alpha) \) the posterior distribution of \( \alpha_j^* \) is normal with mean and covariance matrix that can be found by standard calculations.

3. Resampling \( \eta \) conditional on the currently imputed values \( s, \alpha^* \), and \( k \) follows standard posterior simulation for DP mixture models.

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

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

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

\[
p(y_{n+1} | x_{n+1}, \theta^{(i)}) \propto \sum_{j=1}^{k^{(i)}} n_j^{(i)} \cdot N(y_{n+1}; \alpha_j^{(i)} d_x, S^{(i)}) \\
+ M^{(i)} \int N(y; \alpha d_x, S^{(i)}) \, dp^0(\alpha).
\]

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

\[
F \sim F(N(\alpha d_x, S^{(i)}) \, dq(\alpha)), \text{ as } F \approx q, \text{ and thus a posterior draw for } H_k \text{ as } \\
\int N(\alpha d_x, S^{(i)}) \, dq(\alpha). \text{ that is,} \\
H_k(Y) \propto M^{(i)} \int N(y; \alpha d_x, S^{(i)}) \, dp^0(\alpha) + \sum_{j=1}^{k^{(i)}} n_j^{(i)} \cdot N(y; \alpha_j^{(*)} d_x, S^{(i)})
\]

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

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

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 \( k = 1, \ldots, n \).

Consider two values of the mass parameter, \( M_1 < M_2 \). Then

\[
\frac{P_{M_1}(k)}{P_{M_2}(k)} = \frac{M_1}{M_2} \frac{\Gamma(M_1)}{\Gamma(M_1+n)} \frac{\Gamma(M_2+n)}{\Gamma(M_2)}
\]

for \( k = 1, \ldots, n \). 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 \( q^{(n,M)} \) represent the vector of length \( n \) with components \( P_M(k, p^{(n,M)} = 1) \) 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
\[ p^{(n+1,M)} = \frac{n}{M+n} p^{(n,M)}(0) + \frac{M}{M+n} (0, p^{(n,M)}). \]
This yields the result.

[Received October 2002. Revised October 2003.]

REFERENCES


Lichtman ??? (1993), ???


