Published in final edited form as:

Biometrika. 2008 December ; 95(4): 859–874. doi:10.1093/biomet/asn043.

Bayesian nonparametric inference on stochastic ordering

DAVID B. DUNSON, SHYAMAL D. PEDDADA

Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, U.S.A.

Summary

This article considers Bayesian inference about collections of unknown distributions subject to a partial stochastic ordering. To address problems in testing of equalities between groups and estimation of group-specific distributions, we propose classes of restricted dependent Dirichlet process priors. These priors have full support in the space of stochastically ordered distributions, and can be used for collections of unknown mixture distributions to obtain a flexible class of mixture models. Theoretical properties are discussed, efficient methods are developed for posterior computation using Markov chain Monte Carlo, and the methods are illustrated using data from a study of DNA damage and repair.

Keywords

Dependent Dirichlet process; Hypothesis testing; Mixture model; Nonparametric Bayes; Order restriction

1. Introduction

Our focus is inference on K group-specific distributions. For example, in toxicology studies, the groups may correspond to different doses of a potentially adverse chemical exposure. In such settings, it is of interest to assess whether or not the response distribution changes across groups, while also estimating the group-specific distributions. Although parametric assumptions may be difficult to justify, it is common to have prior knowledge that the magnitude of a response for a particular experimental unit would not decrease if that unit had been exposed to a higher dose. This implies stochastic ordering in the response distributions.

Focusing on the two-sample case, Arjas & Gasbarra (1996) proposed nonparametric Bayes methods for ordered hazard functions, while Gelfand & Kottas (2001) induced priors on stochastically ordered distributions through products of independent Dirichlet process (Ferguson, 1973, 1974) components. More recently, Hoff (2003b) developed general methods for estimating probability measures constrained to belong to a convex set, considering applications to mean, mode, quantile and stochastic ordering constraints. The problem of Bayesian estimation of probability measures subject to a partial stochastic order

was further considered by Hoff (2003a), relying on the theory in Hoff (2003b) to develop latent variable and rejection sampling methods for posterior computation.

It can be difficult to implement these methods routinely, particularly for moderate to large K . In addition, interest often focuses not only on estimation of the group specific distributions but also on testing hypotheses of equalities between groups against stochastically ordered alternatives. There has been surprisingly little work on nonparametric Bayes hypothesis testing and model selection: Gutierrez-Pena & Walker (2005) imbed parametric model selection problems within a Bayesian nonparametric framework; Berger & Guglielmi (2001) test the fit of a parametric model through comparison to a nonparametric alternative; Dass & Lee (2004) study consistency of Bayes factors for testing point nulls versus nonparametric alternatives; and Basu & Chib (2003) develop methods for calculating Bayes factors for comparing Dirichlet process mixture models.

There has been recent interest in Bayesian methods for unconstrained collections of probability measures. Much of this work relies on extending the Dirichlet process. Suppose P is an unknown probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, with $\mathcal{X} \subset \mathfrak{R}$ a Borel subset of Euclidean space and $\mathcal{B}(\mathcal{X})$ the Borel σ -algebra of subsets of $\mathcal{X}$. Then $P \sim \text{DP}(\alpha P_0)$ denotes that P is assigned a Dirichlet process prior with precision α and base measure P_0 . Sethuraman (1994) showed that this is equivalent to the stick-breaking representation:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\Theta_h}(\cdot), \quad \pi_h = V_h \prod_{l < h} (1 - V_l), \quad (1)$$

where $V_h \sim \text{Be}(1, \alpha)$ and $\Theta_h \sim P_0$, independently for $h = 1, \dots, \infty$. Here, δ_{Θ} is a probability measure concentrated at the atom Θ , $V = \{V_h, h = 1, \dots, \infty\}$ is an infinite sequence of stick-breaking weights, and $\Theta = \{\Theta_h, h = 1, \dots, \infty\}$ is an infinite sequence of atoms, with V and Θ mutually independent.

In order to generalize the Dirichlet process to place a prior on a collection of probability measures $(P_1, \dots, P_K)$, MacEachern (1999, 2001) proposed a dependent Dirichlet process, which represents each P_k as in (1), but incorporates dependence through dependent stick-breaking weights and/or dependent atoms. De Iorio et al. (2004) used the dependent Dirichlet process to create an analysis-of-variance-like dependence structure for random measures, assuming fixed stick-breaking weights but dependent atoms. Gelfand et al. (2005) developed a related approach for spatially-dependent probability measures. More flexible methods that also allow the stick-breaking weights to vary, with some computational expense, have been proposed by Griffin & Steel (2006) and Duan et al. (2007). Alternative strategies for modelling dependent collections of random measures through convex combinations of independent Dirichlet processes have been proposed by Müller et al. (2004), Pennell & Dunson (2006) and Dunson et al. (2007).

These methods are appealing in allowing flexible borrowing of information across groups. Our contribution is the development of a framework for testing of equalities in distributions between groups against stochastically ordered alternatives, while also allowing estimation of smooth, stochastically ordered densities. Hoff (2003a) assumes the distributions are discrete,

noting that Dirichlet process mixtures of parametric models can be used to address this problem, but without developing an approach for inference in such a case.

2. Stochastically ordered random probability measures

2.1. Formulation and background

Let $(P_1, \dots, P_K) \in \mathcal{P}^K$, with $\mathcal{P}^K$ the set of $K \times 1$ collections of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. In addition, define the following convex subset of $\mathcal{P}^K$:

$$C_E = \{(P_1, \dots, P_K) \in \mathcal{P}^K : P_i \leq P_j \text{ for all } (i, j) \in E\},$$

where $E \subset (1, \dots, K)^2$ is a prespecified $K \times K$ matrix defining a partial ordering. Here, $P_i \leq P_j$ if and only if $P_i(x, \infty) \leq P_j(x, \infty)$ for all x , so that P_j is stochastically larger than P_i . The collections of probability measures belonging to C_E satisfy the partial ordering defined by E .

As shown by Hoff (2003a,b), C_E is a weakly closed convex set with extreme points $\{(\delta_{s_1}, \dots, \delta_{s_K}) : s \in \mathcal{S}_E\}$, where $\mathcal{S}_E = \{(s_1, \dots, s_K) \in \mathcal{X}^K : s_i \leq s_j \text{ for all } (i, j) \in E\}$. Using a corollary to Choquet's theorem, Hoff (2003a,b) shows that every $(P_1, \dots, P_K) \in C_E$ can be represented as a mixture over the extreme points. Hence, by placing an unconstrained prior on the mixing measure, one can obtain a prior for collections of stochastically ordered random measures. Let $x = (x_1, \dots, x_n)'$ denote the observed data, with $x_i \sim P_{a_i}$, where $a_i \in \{1, \dots, K\}$ is a group index. Then Hoff (2003a) induces a prior on $(P_1, \dots, P_K)$ with weak support on C_E by letting $x_i = s_{i, a_i}$, with $s_i \sim Q$, where $Q \sim \text{DP}(\alpha Q_0)$ and Q_0 is a K -variate Borel probability measure on $\mathcal{S}_E$.

2.2. Restricted dependent Dirichlet process

As a direct consequence of applying the Sethuraman (1994) formulation to Q , we obtain the following specification for $(P_1, \dots, P_K)$:

$$P_k(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\Theta_{hk}}(\cdot), \quad \pi_h = V_h \prod_{l < h} (1 - V_l), \quad k = 1, \dots, K, \quad (2)$$

where $V_h \sim \text{Be}(1, \alpha)$ and $\Theta_h = (\Theta_{h1}, \dots, \Theta_{hK})' \sim Q_0$, for $h = 1, \dots, \infty$, with $V = \{V_h\}_{h=1}^{\infty} = 1$ and $\Theta = \{\Theta_h\}_{h=1}^{\infty}$ mutually independent sequences of stick-breaking random variables and atoms, respectively. We refer to (2) as a restricted dependent Dirichlet process, since the prior implies a dependent Dirichlet process for the collection $(P_1, \dots, P_K)$, with restrictions incorporated through the atoms.

Although the prior on $P = (P_1, \dots, P_K)$ is equivalent to that induced by the Hoff (2003a) latent variable specification, so that his theoretical results apply, the reparameterization in (2) allows one to obtain additional insight into properties and to develop methods for hypothesis testing and posterior computation, as will be clear in §3 and 4.

In considering choice of Q_0 in (2), it is important to keep in mind that, unless Q_0 is chosen to assign probability mass to boundaries of S_E , $\text{pr}(\Theta_{hi} < \Theta_{hj}, h = 1, \dots, \infty, \text{for all } (i, j) \in E) = 1$. The strict order restriction on the atoms for the different groups can lead to a tendency to overestimate group differences, particularly when the true difference is small, sample sizes are small to moderate, and the number of groups is moderate to large. Similar issues arise in much simpler settings involving estimation of normal means or regression parameters subject to order restrictions (Dunson & Neelon, 2003). By choosing Q_0 to allow probability mass on the boundary of S_E , we allow atoms to be identical in the different groups with positive probability.

In considering choice of Q_0 and the impact on borrowing of information across groups, we focus on the two group case in which $E = \{(1, 2)\}$, so that $P_1 \leq P_2$. Data consist of $x_i \sim P_{a_i}$ independently, where $a_i = 1$ for $i = 1, \dots, n_1$ and $a_i = 2$ for $i = n_1 + 1, \dots, n$. We specify Q_0 as

$$f(\Theta_1, \Theta_2) = f_1(\Theta_1) \{ \pi_0 \delta_0(\Theta_2 - \Theta_1) + (1 - \pi_0) f_2(\Theta_2 - \Theta_1) \}, \quad (3)$$

where $\mathcal{X} = \mathfrak{R}$, $f_1(\cdot)$ is a density on $\mathfrak{R}$, such as Gaussian, $0 < \pi_0 < 1$ is the prior probability of $\Theta_1 = \Theta_2$, and $f_2(\cdot)$ is a density with support on $\mathfrak{R}^+$, such as truncated Gaussian.

It is clear from (3) that borrowing of information between Groups 1 and 2 is controlled by the probability π_0 placed on identical atoms, and by the choice of f_2 , which has an impact on the magnitude of the differences in the subset of paired atoms that are not identical. In order to assess how borrowing of information occurs in prediction, after updating the prior with data from the two groups, we calculate the conditional predictive distribution of x_{n+1} for a new subject in Group 2.

By a generalization of the Blackwell & MacQueen (1973) Pólya urn scheme, the conditional distribution of x_{n+1} given $a_{n+1} = 2$, x_1 and x_2 is

$$\begin{aligned} (x_{n+1} \in \cdot \mid x_2, x_1) &\sim \left(\frac{\alpha}{\alpha + n} \right) f_2^*(\cdot) + \sum_{h=1}^{n_1} \left\{ \left(\frac{\pi_0}{\alpha + n} \right) \delta_{x_h}(\cdot) + \left(\frac{1 - \pi_0}{\alpha + n} \right) f_2(\cdot; x_h) \right\} \\ &+ \sum_{h=n_1+1}^n \left(\frac{1}{\alpha + n} \right) \delta_{x_h}(\cdot), \end{aligned} \quad (4)$$

where $f_2^* = f_1 \oplus f_2$ is the convolution of f_1 and f_2 , and $f_2(x_h) = \delta_{x_h} \oplus f_2$. Hence, a new subject in Group 2 has probability $(n - n_1)/(\alpha + n)$ of being assigned to the same x value as a subject in Group 2 of the dataset, and probability $\pi_0 n_1 / (\alpha + n)$ of being assigned to the same value as a subject in Group 1. The probability of being assigned to a new value, $x_{n+1} \notin \{x_1, x_2\}$, is $\{\alpha + n_1(1 - \pi_0)\} / (\alpha + n)$.

2.3. Restricted dependent Dirichlet process mixtures

Expression (2) implicitly assumes almost sure discreteness of each P_k , so does not allow modelling of continuous distributions subject to a stochastic order constraint. As a broader class of restricted dependent Dirichlet process mixture models, we propose

$$\begin{aligned}
& x_i \sim \mathcal{K}_{\sigma_i}(\cdot, \mu_i), \\
& (\mu_i, \sigma_i) \sim P_{a_i}, i = 1, \dots, n, \\
& P_k = \sum_{h=1}^{\infty} \pi_h \delta_{\psi_h}, k = 1, \dots, K,
\end{aligned} \tag{5}$$

where $\mathcal{K}_{\sigma}: \mathcal{D} \times \mathcal{X} \rightarrow \mathfrak{R}^+$ is a kernel that depends on the scale parameter σ , $\mathcal{D}$ is the domain of the data, $\Psi_h = \{\Theta_h, \Gamma_h\} \sim Q_0 \otimes R_0$, $\{\pi_h\}_{h=1}^{\infty}$, $\{\Theta_h\}_{h=1}^{\infty}$ and Q_0 are as defined in (2), $\Psi_h \in \mathfrak{R}^+$, R_0 is a base probability measure on $(\mathfrak{R}^+, \mathcal{B}(\mathfrak{R}^+))$, and $Q_0 \otimes R_0$ denotes the product measure.

Letting $g_k(x) = \int \mathcal{K}_{\sigma}(x, \mu) P_k(d\mu, d\sigma)$, for $k = 1, \dots, K$ and all $x \in \mathcal{D}$, the prior for $(P_1, \dots, P_K)$ in (5) induces a prior for the collection $g = \{g_1, \dots, g_K\} \in \mathcal{L}_E$. Suppose for all $\sigma \in \mathfrak{R}^+$ and $\mu \in \mathcal{X}$, $\mathcal{K}_{\sigma}(\cdot, \mu)$ is a nondegenerate density on $\mathcal{D}$ satisfying the monotone stochastic ordering condition

$$\int_{\mathcal{D}(z)} \mathcal{K}_{\sigma}(x, \mu) dx \geq \int_{\mathcal{D}(z)} \mathcal{K}_{\sigma}(x, \mu + \Delta) dx, \quad \text{for all } \Delta > 0, z \in \mathcal{D},$$

with $\mathcal{D}(z)$ the subset of $\mathcal{D}$ excluding values greater than z . Then $\mathcal{L}_E$ is a set of densities satisfying partial stochastic ordering E from the following result:

LEMMA 1. For all $(g_1, \dots, g_K) \in \mathcal{L}_E$,

$$\int_{\mathcal{D}(z)} g_k(x) dx \geq \int_{\mathcal{D}(z)} g_l(x) dx, \quad \text{for all } z \in \mathcal{D} \text{ and } (k, l) \in E.$$

Lemma 1 follows from the monotone stochastic ordering condition after noting that

$$\int_{\mathcal{D}(z)} g_k(x) dx = \sum_{h=1}^{\infty} \pi_h \int_{\mathcal{D}(z)} \mathcal{K}_{\sigma_h}(x, \mu_{hk}) dx, \quad k = 1, \dots, K,$$

and $\mu_{hk} \leq \mu_{hl}$ for all $(k, l) \in E$.

Note that $\mathcal{L}_E$ is a convex set with extreme points

$$[\{\mathcal{K}_{\sigma}(\cdot, \mu_1), \dots, \mathcal{K}_{\sigma}(\cdot, \mu_K)\} : \sigma \in \mathfrak{R}^+, \mu \in \mathcal{S}_E],$$

assuming that $\mathcal{K}_{\sigma}(\cdot, \mu)$ cannot be expressed as a convex combination of kernels of the same form but with different μ, σ . Every $g \in \mathcal{L}_E$ can be represented as a mixture over the extreme points, and our prior in (5) corresponds to placing a Dirichlet process prior on the mixing measure. Consider the special case in which $\mathcal{K}_{\sigma}(x, \mu) = \mathcal{K}((x - \mu)/\sigma)/\sigma$, with $\mathcal{K}(\cdot)$ an arbitrary non-degenerate density on $\mathfrak{R}$, such as the standard normal. In this case,

$g_k(x) = \int \mathcal{K}((x - \mu)/\sigma)/\sigma P_k(d\mu, d\sigma)$, with P_k having a marginal Dirichlet process prior on $(\mathfrak{R} \times \mathfrak{R}^+, \mathcal{B}(\mathfrak{R}) \otimes \mathcal{B}(\mathfrak{R}^+))$. From Lo (1984) (refer to expression 3.1) it follows that the support of the prior for g_k contains all densities on $\mathfrak{R}$ with respect to Lebesgue measure in its closure.

Note that expression (5) induces a Dirichlet process location-scale mixture on each of the densities g_k . A convenient special case corresponds to location-scale mixtures of normal densities. However, it is well known that one can approximate any density on $\mathfrak{R}$ with respect to Lebesgue measure through use of a Dirichlet process location mixture of normals having a single unknown variance (Ghosal et al., 1999, Lijoi et al., 2005). Hence, for simplicity in applications, we focus the remainder of the article on the location mixture

$$x_i \sim \mathcal{N}(\cdot, \mu_i), \quad \mu_i \sim P_{a_i}, i = 1, \dots, n, \quad (6)$$

with the prior for $(P_1, \dots, P_K)$ specified as in (2) and with σ an unknown scale.

3. Properties and hypothesis testing

3.1. Two-group case

One of our primary goals is to develop an approach for hypothesis testing of equalities in distributions between groups against stochastically ordered alternatives. We focus on the model defined in expressions (2) and (6). Under this model, differences between groups in $(g_1, \dots, g_K)$ are controlled through differences in the mixture distributions. Hence, we base inferences about differences among $(g_1, \dots, g_K)$ on tests of differences among $(P_1, \dots, P_K)$.

In developing the methods, we initially consider the two group case in which $P_1 \preceq P_2$. Instead of requiring the distributions to be exactly identical in the two groups, interval null hypotheses are formulated based on bounding a distance metric. In particular, we focus on the total variation distance,

$$d_{12} = \max_{B \in \mathcal{B}(\mathcal{X})} |P_2(B) - P_1(B)|. \quad (7)$$

LEMMA 2. Under the prior defined in (2) and (3),

$$d_{12} = \sum_{h=1}^{\infty} \pi_h 1(\beta_h > 0) \sim \text{Be}\{\alpha(1 - \pi_0), \alpha\pi_0\},$$

The proof is in the Appendix. Lemma 2 implies that the distance between P_1 and P_2 stochastically increases with $1 - \pi_0$. In addition, $E(d_{12}) = 1 - \pi_0$, with α controlling uncertainty in this prior expectation.

Using the distance metric in (7), we formalize the null and alternative hypotheses as

$$H_0: d_{12} \leq \epsilon, \quad H_1: d_{12} > \epsilon, \quad (8)$$

where $\epsilon > 0$ is a small positive constant. In the limit as $\pi_0 \rightarrow 0$, $d_{12} \rightarrow \delta_1$ in distribution, so that $\text{pr}(d_{12} < \epsilon) \rightarrow 0$, for any $\epsilon \rightarrow 1$

Theorem 1 provides justification for using hypotheses on the mixing measures as a basis for inferences on G_1 and G_2 .

THEOREM 1. *Specify H_0 and H_1 as in (8) and let $G_k(B) = \int_B g_k(x) dx$, $k = 1, 2$, with $g_k(x) = \int \mathcal{K}(x, s) dP_k(s)$. Then H_0 implies that*

$$\max_{x \in (\mathcal{X})} |G_2(x, \infty) - G_1(x, \infty)| < \epsilon.$$

3.2. Multiple-group case

It is straightforward to generalize the approach in §3.1 to the multiple-group setting. When there are K groups subject to partial stochastic ordering indexed by E , it is first necessary to specify P_0 . For most orderings of interest in applications, one can express the h th atom in group k as $\theta_{hk} = w'_k \beta_h^*$, where w_k is a fixed $K \times 1$ vector of 0s and 1s, and $\beta_h^* = (\beta_{h1}^*, \dots, \beta_{hK}^*)'$ is a parameter vector having a subset of elements constrained to be non-negative or nonpositive.

For example, when there is a simple stochastic ordering: $P_1 \leq P_2 \leq \dots, P_K$ we let $w_k = (1'_k, 0'_{K-k})'$, which implies that $\beta_{hk}^* = \theta_{hk} - \theta_{h, k-1}$, for $k = 2, \dots, K$. Hence, constraining $\beta_{hk}^* \geq 0$, for $k = 2, \dots, K$, induces a P_0 with appropriate support. Generalizing prior (3), we let

$$f(\beta_h^*) = f_1(\beta_{h1}^*) \prod_{k=2}^K \{ \pi_{0k} \delta_0(\beta_{hk}^*) + (1 - \pi_{0k}) f_k(\beta_{hk}^*) \}, \quad (9)$$

where $\pi_{0k} = \text{pr}(\theta_{hk} = \theta_{h, k-1})$ and f_k is a density with support on $\mathfrak{R}^+$, for $k = 2, \dots, K$. To modify the prior to accommodate tree stochastic order, i.e. $P_1 \leq P_k$, $k = 2, \dots, K$, we can use the same P_0 but with $w_k = (1, 0_{k-2}, 1, 0_{K-k})'$. Umbrella orderings, with location of the peak known, and cases involving multiple factors, discussed in §5, can also be accommodated by changing $W = (w_1, \dots, w_K)'$.

For ease in exposition, we focus our discussion of multiple group hypothesis testing on the simple ordering case, though modifications to other cases are automatic. Generalizing the hypotheses in (8), we let

$$H_{0k}: d_{k, k+1} \leq \epsilon, \quad H_{1k}: d_{k, k+1} > \epsilon, \quad (10)$$

for $k = 1, \dots, K-1$, with $d_{k, k+1} = \max_{x \in \mathcal{X}} |P_{k+1}(x, \infty) - P_k(x, \infty)|$. We refer to H_{0k} as the local null hypothesis of near-equivalence in Groups k and $k+1$. One can also consider the global null hypothesis, $H_0: d_M \leq \epsilon$, where

$$d_M = \max_{(j,l) \in \{1, \dots, K\}} \max_{B \in \mathcal{B}(\mathcal{X})} |P_l(B) - P_j(B)|.$$

Hence, H_0 implies near-equivalence in all K groups. It is straightforward to show that $d_M \sim \text{Be}\{\alpha(1 - \Pi_0), \alpha\Pi_0\}$, with $\Pi_0 = \prod_{k=1}^{K-1} \pi_{0k}$. Using the Gibbs sampler of §4, we can calculate local and global hypothesis probabilities from a single chain, while also obtaining model-averaged group-specific density estimates.

4. Posterior computation

4.1 Model and background

In describing algorithms for posterior computation, we focus on the following case:

$$\begin{aligned} x_i &\sim N(w'_{a_i} \beta_i, \tau^{-1}) \\ \beta_i &\sim P = \sum_{h=1}^{\infty} V_h \prod_{l < h} (1 - V_l) \delta_{\beta_k}, \quad V_h \sim \text{Be}(1, \alpha), \beta_h^* \sim P_0, \end{aligned} \quad (11)$$

where $a_i = k$, for $i \in \{n_{k-1} + 1, \dots, n_{k-1} + n_k\}$, $k = 1, \dots, K$, with $n_0 = 0$ and n_k the number of subjects in group k , w_k chosen as discussed in §3.2, and $\beta_i = (\beta_{i1}, \dots, \beta_{iK})'$. Note that this expression is a special case of (5) with $\mathcal{X}$ corresponding to a normal kernel and with $\mu_i = w'_{a_i} \beta_i$. Assuming prior (9), we let $f_1(\beta_{h1}^*) = N(\beta_{h1}^*; \mu_0, \sigma_0^2)$ and $f_k(\beta_{hk}^*) = N_+(\beta_{hk}^*; 0, \kappa^{-1})$, for $k = 2, \dots, K$, with $N_+(\cdot)$ denoting the normal density truncated to be positive.

As a result of the reparameterization used in expression (11), we effectively have a typical Dirichlet process mixture of normal linear regression models, with a constrained mixture structure used in the base measure, P_0 . Hence, we can use standard Markov chain Monte Carlo algorithms for posterior computation in Dirichlet process mixture models. Since the structure of the base measure creates some difficulties in implementing Pólya urn-based algorithms, we focus on a blocked Gibbs sampler (Ishwaran & James, 2001). This algorithm is based on updating the random weights and atoms in a truncation approximation to the infinite stick-breaking representation.

In particular, let $P = \sum_{h=1}^N V_h \prod_{l < h} (1 - V_l) \delta_{\beta_k^*}$, with the components defined as in (11) but with $V_N = 1$ so that terms $N + 1, \dots, \infty$ can be excluded. Following the approach of Ishwaran & James (2001), one can show that this approximation tends to be accurate for moderate N , such as $N = 20$, particularly if $\alpha \rightarrow 1$. One can assess whether or not N is sufficiently large by monitoring the maximum index for an occupied cluster: $N_O = \max_i \zeta_i$, with $\zeta_i = h$ if $\beta_i = \beta_h^*$. If the posterior probability assigned to N_O values close to N is not close to zero, N should be increased.

4.2. Prior specification and posterior inference

We recommend a default prior specification. To avoid sensitivity to the measurement scale, standardize the data by subtracting the Group-1 mean and dividing by the Group-1 standard

deviation. Then $\mu_0 = 0$ and $\sigma_0^2 = 1$ are chosen for the mean and variance of the Group-1 atoms. To assign high probability to a wide range of mild to moderate shifts in the response density, let $\kappa \sim \text{Ga}(1/2, 1/2)$ to induce a Cauchy prior on β_{hk}^* , for $k > 1$. The Cauchy is often used as a robust default in parametric model selection (Jeffreys, 1961). A diffuse prior can be chosen for the error precision, τ^{-1} , which is common to all the models.

In addition, we choose a $\text{Ga}(a_\alpha, b_\alpha)$ prior for α , with the gamma distribution parameterized to have mean a_α/b_α and variance a_α/b_α^2 . There is substantial information in the data about α , so a diffuse prior could be used, but we recommend letting $a_\alpha = 1$ and $b_\alpha = 1$ to favour a small-to-moderate number of clusters. Letting $\epsilon = 0.05$ and fixing $\alpha = 1$, one can assign $\text{pr}(H_{0k}) = \text{pr}(H_{1k}) = 1/2$ by choosing $\pi_{0k} = 0.792$, the solution to the equation $0.5 = \int_0^{0.05} \text{Be}(\pi; 1 - \pi_{0k}, \pi_{0k}) d\pi$. Letting $a_\pi + b_\pi = 1$ to correspond to a unit information prior, we obtain a hyperprior for π_{0k} with mean 0.792 by letting $a_\pi = 0.792$, $b_\pi = 0.208$. In the two group case, the induced prior on d_{12} is U-shaped, having a mean of 0.208 and high density near 0 and 1.

With the prior specification complete, the Gibbs sampler described in the Appendix can be run to obtain draws from the posterior distribution. These draws can be used to obtain group-specific density estimates and posterior hypothesis probabilities. The posterior distribution of the distance between any two groups can also be obtained.

5. Simulation examples

We considered two simulation cases. In both cases, $K = 2$ and data in Group 1 were simulated from the following mixture of three normals:

$$f(y) = 0.2N(y; -2.5, \tau^{-1}) + 0.7N(y; 0, \tau^{-1}) + 0.1N(y; 1.5, \tau^{-1}),$$

with $\tau = 3$. In Case 1 data in Group 2 were simulated from this same density, while in Case 2 we increased the component-specific means slightly from $(-2.5, 0, 1.5)$ to $(-2.4, 0.4, 2.2)$, resulting in a distance of $d_{12} = 1$ as each of the means varied.

For each case, we simulated 100 datasets under three sample sizes, $n_0 = 10, 25$ and 100, with n_0 denoting the sample size per group. Each dataset was analyzed using the Gibbs sampler of §4, with 2000 iterations collected after a 500 iteration burn-in. Apparent convergence was rapid and mixing was good. The set of simulations in the $n_0 = 100$ case was run using Matlab on a Mac PowerBook G4 laptop in approximately 10 hours.

Figure 1 plots the results in Case 1, showing the group-specific Bayesian density estimates for each of the 100 simulated datasets in the three sample sizes under consideration, along with a histogram of the d_{12} distance measure. For a small sample size of 10/group, the estimates were surprisingly good on average, though the primary or secondary mode was substantially overestimated in a small proportion of simulations. In addition, in many of the simulations, the density estimate exhibited shrinkage towards the normal base measure, with the primary peak and dip in the left-hand tail underestimated. The performance improved

dramatically for $n_0 = 25$, with most of the density estimates close to the truth. The $n_0 = 100$ estimates were all very close to the truth. The distribution of the posterior mean of d_{12} across the simulations was increasingly concentrated at the true value of 0 as the sample size increased. Figure 2 plots the results in Case 2. The performance was similar to that described in case 1, but with the distribution of d_{12} concentrated increasingly away from 0 as the sample size increased.

Summary statistics of d_{12} across the simulations are shown in Table 1. The posterior probability of H_0 increased as the sample size increased under Case 1, and decreased as the sample size increased under Case 2. For the two small sample sizes, there was not strong evidence in favour of either hypothesis. This is as expected, given that we have purposely chosen a simulation case in which the difference between the groups is subtle. Repeating the simulation in a case with a larger difference, results not shown, showed that we can obtain strong evidence in favour of H_1 even in sample sizes of $n_0 = 10$. Table 1 illustrates that the posterior hypothesis probabilities are somewhat robust to ϵ , with changes of less than 0.10 even with ϵ varying from 0.01 to 0.1. To assess sensitivity of hypothesis testing to the hyperprior on π_0 , we repeated simulation Cases 1 and 2 for sample size $n_0 = 100$ with (i) π_0 fixed at 0.5 and (ii) $a_\pi = b_\pi = 1$. In each case, density estimation performance was the same as that shown in Figs 1 and 2. However, when $\pi_0 = 0.5$, there was a tendency to overestimate d_{12} , leading to substantially lower posterior probabilities of H_0 in Case 1 than those shown in Table 1. This is as expected given that fixing π_0 does not allow borrowing of information across the atoms in the restricted dependent Dirichlet process specification. Although prior (ii) was also centred on $\pi_0 = 0.5$, there was less of a tendency to overestimate d_{12} under H_0 .

6. Genotoxicology application

6.1. Data structure and the scientific problem

We applied the approach to data from a study of DNA damage and repair. Batches of cells were exposed to 0, 5, 20, 50 or 100 micromoles H_2O_2 and DNA damage was then measured in individual cells after allowing a repair time of 0, 60 or 90 minutes. With $i = 1, \dots, n$ indexing the cells under study, the measured response x_i for cell i was the Olive tail moment, which is a surrogate of the frequency of DNA strand-breaks obtained using the comet assay.

The goal of the study is to assess the sensitivity of the comet assay to detecting damage induced by the known genotoxic agent H_2O_2 , while also investigating how rapidly damage is repaired. Let $a_i \in \{1, \dots, K\}$ be a group index denoting the level of H_2O_2 and repair time for cell i . The value of a_i for each dose $\times$ repair time value is shown in Fig. 3. The total sample size is 1400, with 100 cells per group except for groups 9 and 13, which had 50.

Among cells with zero repair time, DNA damage should be nondecreasing with dose of H_2O_2 . In addition, within a given dose level, DNA damage should be nonincreasing with repair time. Hence, we make the ordering assumption illustrated in Fig. 3 using a directed graph, with arrows pointing towards stochastically larger groups. We wish to assess whether or not DNA damage continues to increase at higher levels of H_2O_2 exposure and to investigate whether or not damage is significantly reduced across each increment of the repair time.

6.2. Analysis and results

Previous authors analyzed the data in Groups 1–5 using a dynamic mixture of Dirichlet processes (Dunson, 2006; Pennell & Dunson, 2006), which allows for dependence in the distributions within adjacent dose groups but does not enforce stochastic ordering restrictions. In addition, that model does not allow the incorporation of both dose of H_2O_2 and repair time, though extensions are possible.

We implemented the approach described in §4. The Gibbs sampler was run for 20,000 iterations after a 1,000-iteration burn-in. As in the simulation study, the chain appeared to converge rapidly and mix efficiently based on standard diagnostics. In fact, 20,000 iterations was really more than sufficient, and we obtained essentially indistinguishable results based on a 500-iteration burn-in and a 2,000-iteration collection interval, as used in the simulation studies.

Figure 4 plots group-specific Bayesian density estimates and 95% pointwise credible intervals. Although the Bayes estimates have been constrained to follow a stochastic ordering, there is no evidence of systematic deviations from frequentist kernel density estimates obtained using only the data in a given group. Figure 3 provides posterior probabilities of local alternative hypotheses for different group comparisons. The results suggest highly significant increases in DNA damage between the 0, 5 and 20 micromole H_2O_2 dose groups given a repair time of 0 minutes, with the evidence of further increases at higher doses less clear. As expected, there is no evidence of a change in the distribution between Groups 1, 6 and 11, since there was no induced damage to be repaired. However, there were highly significant decreases in DNA damage in each of the exposed groups after a repair time of 60 minutes. Allowing an additional 30 minutes of repair did not significantly alter the distribution. These conclusions were the same for $\epsilon = 0.01$ and $\epsilon = 0.10$.

These results are consistent with the raw data and plots shown in Fig. 4, and are both scientifically reasonable and interesting. It is known that the type of damage induced by hydrogen peroxide can be repaired quickly by base excision and repair mechanisms. The result that there is no further improvement after 60 minutes suggests that it may be unnecessary to collect data for repair times exceeding an hour in future molecular epidemiology studies using the comet assay to identify genotypes predictive of DNA repair rates.

To assess sensitivity, we repeated analyses using a variety of alternative hyperprior settings. In particular, we tried the cases $a_\pi + b_\pi = 5$ instead of 1 to correspond to a more informative prior, $a_\alpha = 1$, $b_\alpha = 1$ to favour more clusters, $\kappa = 1$, and $\kappa = 2$. The estimated densities did not change noticeably across these analyses and the conclusions were robust.

7. Discussion

The proposed approach for posterior computation focused on order restrictions that can be expressed in terms of sign constraints on a $K \times 1$ vector of regression coefficients specific to each Dirichlet process component. This encompasses simple order, tree orders, umbrella orders with known changepoints, and certain multi-factor cases, such as that considered in

Section 6. However, orders having more than $K - 1$ restrictions cannot be accommodated. For example, considering the order restriction illustrated graphically in Fig. 3, a different computational approach is needed if the graph were modified to include additional arrows, e.g., from Groups 6 to 7.

Although we have focused on a relatively simple setting, it is straightforward to imbed our prior for stochastically ordered mixture distributions within a larger hierarchical model. For example, one can incorporate a parametric adjustment for covariates within the kernel. In addition, one can allow for stochastically ordered latent variable distributions. Ongoing work focuses on the extension to incorporate continuous predictors, which can conceptually be accomplished by replacing the atoms with nondecreasing stochastic processes.

Acknowledgement

This research was supported by the Intramural Research Program of the U.S., National Institutes of Health, National Institute of Environmental Health Sciences. The authors thank an anonymous referee for important comments.

Appendix

Technical details

Proof of Lemma 2.

We first show that

$$d_{12} = \max_{B \in \mathcal{B}(\mathcal{X})} \left\{ \sum_{h=1}^{\infty} \pi_h 1(\theta_{h1} \in B, \theta_{h2} \notin B) \right\} = \sum_{h=1}^{\infty} \pi_h 1(\beta_h > 0).$$

This follows by first noting that the expression in $\{\cdot\}$ is maximized when B is chosen as the union of small neighbourhoods around the atoms $\{\theta_{h1}\}_{h=1}^{\infty}$, with these neighbourhoods sufficiently small to exclude all elements of $\{\theta_{h2}\}_{h=1}^{\infty}$ that do not belong to $\{\theta_{h1}\}_{h=1}^{\infty}$. Then, under (3), it is clear that, for this choice of B , $1(\theta_{h1} \in B, \theta_{h2} \notin B) = 1(\beta_h > 0)$. The final part of Lemma 1 follows from noting that $G \sim \text{DP}(\alpha G_0)$ implies

$$G(B) = \sum_{h=1}^{\infty} \pi_h 1(\theta_h \in B) \sim \text{Be}[\alpha G_0(B), \alpha \{1 - G_0(B)\}], \text{ for all } B \in \mathcal{B}(\mathcal{X}),$$

where $V_h \sim \text{Be}(1, \alpha)$, $\theta_h \sim G_0$, $h = 1, \dots, \infty$. In addition, $1(\theta_h \in B) \sim \text{Ber}\{G_0(B)\}$.

Proof of Theorem 1.

The conditions in Theorem 1 imply that

$$G_1(x, \infty) = \sum_{h=1}^{\infty} \pi_h \mathcal{K}^*(x, \theta_{h1}) \quad \text{and} \quad G_2(x, \infty) = \sum_{h=1}^{\infty} \pi_h \mathcal{K}^*(x, \theta_{h1} + \beta_h), \quad \forall x \in \mathcal{X},$$

where $\mathcal{K}^*(x, \Theta) = \int_x^\infty \mathcal{K}(z, \Theta) dz$. Letting $\mathcal{H}_0 = \{h: \beta_h = 0, h = 1, 2, \dots, \infty\}$ and $\overline{\mathcal{H}}_0 = \{1, 2, \dots, \infty\} \setminus \mathcal{H}_0$, we have

$$G_2(x, \infty) - G_1(x, \infty) = \sum_{h \in \overline{\mathcal{H}}_0} \pi_h \{ \mathcal{K}^*(x, \Theta_{h1} + \beta_h) - \mathcal{K}^*(x, \Theta_{h1}) \}.$$

Since $\sum_{h \in \overline{\mathcal{H}}_0} \pi_h < \epsilon$ under H_0 and, for any fixed x , $0 \leq \mathcal{K}^*(x, \Theta) \leq 1$ is a monotone increasing function of Θ , Theorem 1 follows directly.

Gibbs sampling steps.

Let $\zeta_i = h$ denote that individual i is sampled from component h , so that $\beta_i = \beta_h^*$, for $h = 1, \dots, T$. Then the blocked Gibbs sampler proceeds through the following steps.

Step 1. Update ζ_i , for $i = 1, \dots, n$, by sampling from the multinomial conditional distribution with

$$\text{pr}(\zeta_i = h) = \frac{\pi_h N(x_i; w'_{a_i} \beta_h^*, \tau^{-1})}{\sum_{l=1}^N \pi_l N(x_i; w'_{a_i} \beta_l^*, \tau^{-1})}, \quad h = 1, \dots, N.$$

Step 2. Assuming $\tau \sim \text{gamma}(a_\tau, b_\tau)$, update τ by sampling from the full conditional,

$$\text{gamma}\left(a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2} \sum_{i=1}^n (x_i - w'_{a_i} \beta_i)^2\right).$$

Step 3. Update V_h , for $h = 1, \dots, N-1$, by sampling from the full conditional distribution,

$$\text{Be}\left(1 + \sum_{i=1}^n 1(\zeta_i = h), \alpha + \sum_{i=1}^n 1(\zeta_i > h)\right).$$

Step 4. Update β_h^* , for $h = 1, \dots, N$, via the following Gibbs sub-steps:

- a.** update β_{h1}^* from $\mathcal{N}(E_{h1}, V_{h1})$, where

$$E_{h1} = V_{h1} \left[\sigma_0^{-2} \mu_0 + \tau \sum_{i: \zeta_i = h} \left(x_i - \sum_{k=2}^K w_{a_i, k} \beta_{ik} \right) \right], \quad V_{h1} = \left[\sigma_0^{-2} + \tau \sum_{i=1}^n 1(\zeta_i = h) \right]^{-1}$$

- b.** update β_{hk}^* , for $k = 2, \dots, K$, from the full conditional,

$$\hat{\pi}_{hk} \delta_0 + (1 - \hat{\pi}_{hk}) N_+(E_{hk}, V_{hk}), \text{ where the conditional probability of } \beta_{hk}^* = 0 \text{ is}$$

$$\hat{\pi}_{hk} = \pi_{0k} \left(\pi_{0k} + (1 - \pi_{0k}) \frac{\sqrt{2\kappa} \int_0^\infty N(z; E_{hk}, V_{hk}) dz}{\sqrt{\pi} N(0; E_{hk}, V_{hk})} \right)^{-1},$$

and the mean and variance in the normal component are, respectively,

$$E_{hk} = V_{hk} \left\{ \tau \sum_{i: \zeta_i = h} \left(x_i - \sum_{l \neq k} w_{a_l, i} \beta_{il} \right) \right\}, \quad V_{hk} = \left\{ \kappa + \tau \sum_{i: \zeta_i = h} w_{a_l, k}^2 \right\}^{-1}.$$

Step 5. Assuming $\alpha \sim \text{Ga}(a_\alpha, b_\alpha)$, update α from its full conditional distribution,

$$\text{Ga} \left(a_\alpha + N - 1, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h) \right).$$

Step 6. Assuming $\pi_{0k} \sim \text{Be}(a_\pi, b_\pi)$, for $k = 2, \dots, K$, update π_{0k} from its full conditional distribution,

$$\text{Be} \left(a_\pi + \sum_{h=1}^N 1(\beta_{hk}^* = 0), b_\pi + \sum_{h=1}^N 1(\beta_{hk}^* \neq 0) \right).$$

Step 7. Assuming $\kappa \sim \text{Ga}(a_\kappa, b_\kappa)$, update κ from its full conditional distribution,

$$\text{Ga} \left(a_\kappa + \frac{1}{2} \sum_{h=1}^N \sum_{k=2}^K 1(\beta_{hk}^* \neq 0), b_\kappa + \frac{1}{2} \sum_{h=1}^N \sum_{k=2}^K (\beta_{hk}^*)^2 \right).$$

REFERENCES

- ARJAS E & GASBARRA D (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *J. Am. Statist. Assoc* 91, 1101–9.
- BASU S & CHIB S (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Am. Statist. Assoc* 98, 224–35.
- BERGER JO & GUGLIELMI A (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *J. Am. Statist. Assoc* 96, 174–81.
- DASS SC & LEE J (2004). A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *J. Statist. Plan. Infer* 119, 143–52.
- DE IORIO M, MULLER P, ROSNER GL & MACEACHERN SN (2004). An ANOVA model for dependent random measures. *J. Am. Statist. Assoc* 99, 205–15.
- DUAN JA, GUIDANI M & GELFAND AE (2007). Generalized spatial Dirichlet process models. *Biometrika* 94, 809–25.
- DUNSON DB & NEELON B (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics* 59, 286–95. [PubMed: 12926713]
- DUNSON DB, PILLAI N & PARK J-H (2007). Bayesian density regression. *J. Royal Statist. Soc. B* 69, 163–183.
- GELFAND AE & KOTTAS A (2001). Nonparametric Bayesian modeling for stochastic order. *Ann. Inst. Statist. Math* 53, 865–76.

- GELFAND AE, KOTTAS A, & MACEACHERN SN (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Am. Statist. Assoc* 100, 1021–35.
- GHOSAL S, GHOSH JK & RAMAMOORTHY RV (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist* 27, 143–58.
- GRIFFIN JE & STEEL MFJ (2006). Order-based dependent Dirichlet processes. *J. Am. Statist. Assoc* 101, 179–94.
- GUTIERREZ-PENA E & WALKER SG (2005). Statistical decision problems and Bayesian nonparametric methods. *Int. Statist. Rev* 73, 309–30.
- HANSEN MB & LAURITZEN SL (2002). Nonparametric Bayes inference for concave distribution functions. *Statist. Neer* 56, 110–27.
- HOFF PD (2003a). Bayesian methods for partial stochastic orderings. *Biometrika* 90, 303–17.
- HOFF PD (2003b). Nonparametric estimation of convex models via mixtures. *Ann. Statist* 31, 174–200.
- ISHWARAN H & JAMES LF (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc* 96, 161–73.
- JEFFREYS H (1961). *Theory of Probability*, 3rd ed Oxford University Press.
- LIJOI A, PRUNSTER I & WALKER SG (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Am. Statist. Ass* 100, 1292–6.
- LO AY (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist* 12, 351–7.
- MACEACHERN SN (1999). Dependent nonparametric processes In *Proc. Sect. Bayes. Statist. Sci*, pp. 50–5., Alexandria, VA: American Statistical Association.
- MACEACHERN SN (2001). Decision theoretic aspects of dependent nonparametric processes In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, Ed. George E, pp. 551–60. Creta: International Society for Bayesian Analysis.
- MULLER P, QUINTANA F & ROSNER G (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Statist. Soc. B* 66, 735–49.
- PENNELL ML & DUNSON DB (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* 62, 1044–52. [PubMed: 17156278]
- SETHURAMAN J (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* 4, 639–50.

Figure 1:

Results from simulation study Case 1. The rows correspond to the sample sizes of 10, 25 and 100 per group. The first two columns show the posterior mean density estimates for Groups 1 and 2 from each simulation in dotted curves, with the true densities as solid curves. The third column shows a histogram of the posterior means of the distance d_{12} between the groups, with vertical lines for the prior mean.

Figure 2:

Results from simulation study Case 2. The rows correspond to the sample sizes of 10, 25 and 100 per group. The first two columns show the posterior mean density estimates for groups 1 and 2 from each simulation in dotted curves, with the true densities as solid curves. The third column shows a histogram of the posterior means of the distance d_{12} between groups, with vertical lines for the prior mean.

Figure 3: Genotoxicity application. Directed graph illustrating order restriction. Arrows point towards stochastically larger groups. Posterior probabilities of H_{1k} are shown.

Figure 4: Genotoxicity application. Estimated densities of the Olive tail moment in a subset of the H_2O_2 dose $\times$ repair-time groups. Solid curves are posterior mean density estimates and dashed curves provide 95% pointwise credible intervals.

Table 1.

Simulation study. Summary statistics of d_{12} across the simulations for each sample size. Results shown are means with 95% empirical confidence limits in parentheses.

Case	Common sample size	d_{12}	$\Pr(d_{12} < \epsilon \mid \text{data})$		
			$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.1$
1	10	0.140 (0.028, 0.614)	0.67 (0.17, 0.86)	0.72 (0.18, 0.91)	0.75 (0.22, 0.93)
1	25	0.085 (0.011, 0.299)	0.72 (0.19, 0.89)	0.78 (0.22, 0.94)	0.81 (0.28, 0.96)
1	100	0.044 (0.003, 0.215)	0.79 (0.28, 0.94)	0.85 (0.30, 0.99)	0.89 (0.42, 1.00)
2	10	0.233 (0.029, 0.660)	0.55 (0.10, 0.82)	0.60 (0.10, 0.91)	0.63 (0.10, 0.93)
2	25	0.278 (0.032, 0.765)	0.46 (0.02, 0.84)	0.51 (0.02, 0.87)	0.54 (0.06, 0.91)
2	100	0.560 (0.091, 0.884)	0.13 (0.00, 0.74)	0.15 (0.00, 0.79)	0.18 (0.00, 0.83)