Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 2, *pp.* 253–268

Minimally informative prior distributions for non-parametric Bayesian analysis

Christopher A. Bush

Novartis Pharmaceuticals, East Hanover, USA

and Juhee Lee and Steven N. MacEachern

Ohio State University, Columbus, USA

[Received July 2007. Final revision September 2009]

Summary. We address the problem of how to conduct a minimally informative, non-parametric Bayesian analysis. The central question is how to devise a model so that the posterior distribution satisfies a few basic properties. The concept of 'local mass' provides the key to the development of the limiting Dirichlet process model. This model is then used to provide an engine for inference in the compound decision problem and for multiple-comparisons inference in a one-way analysis-of-variance setting. Our analysis in this setting may be viewed as a limit of the analyses that were developed by Escobar and by Gopalan and Berry. Computations for the analysis are described, and the predictive performance of the model is compared with that of mixture of Dirichlet processes models.

Keywords: Bayes; Dirichlet process; Improper prior; Local mass; Mixed modes analysis; Reference prior

1. Introduction

An outstanding problem in the area of non-parametric Bayesian analysis is how to perform a non-informative or minimally informative analysis. Technical details of the models render the usual approach of simply selecting a diffuse prior distribution useless: applying the standard approaches to developing non-informative analyses in this setting leads to inference that, in important ways, does not depend on the data. We develop an alternative formulation which requires only modest input yet which provides reasonable behaviour, *a posteriori*. The analysis that we perform also provides insight into how to use models based on the Dirichlet process when we wish to use an informative prior distribution.

A new class of models lies at the heart of our analysis. The models, which we term limit of Dirichlet (LIMDIR) process models, are derived from the limit of a sequence of mixture of Dirichlet process models. We describe how to take the limit to ensure that the portion of the posterior distribution with which we are concerned has a proper limiting distribution. We present conditions that ensure this propriety.

We apply the new class of LIMDIR models to one-way analysis of variance and to the compound decision problem. We provide a prior elicitation strategy that allows us to calibrate our minimally informative Bayesian analysis with either subjective input or classical procedures.

Address for correspondence: Juhee Lee, Department of Statistics, Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, USA.
E-mail: juheelee@stat.osu.edu

 1369–7412/10/72253

We note that our approach to non-parametric Bayesian analysis of variance differs from those of Tomlinson and Escobar (1999), Müller *et al.* (2004) and DeIorio *et al.* (2004).

Section 2 describes the new class of models and presents theoretical results for them. Section 3 discusses elicitation of the prior distribution. Section 4 presents a data analysis and comparisons with other methods. The final section contains conclusions.

2. The limit of Dirichlet process model

The development of the LIMDIR model is driven by the goal of providing a sound posterior analysis. The LIMDIR process is the limit of a sequence of Dirichlet processes which, in our setting, tends to an 'improper process'. It will be used in a hierarchical model and is thus an extension of the mixture of Dirichlet processes model that has become a staple of the non-parametric Bayesian diet (Dey *et al.*, 1998; Müller and Quintana, 2004; Walker *et al.*, 1999).

The model for the data, conditional on the treatment means and variances, is the standard one-way analysis-of-variance model. There are k treatments, with n_i observations on the ith treatment. The observations are mutually independent. Thus, with θ_i and σ_i^2 representing the mean and variance of treatment i,

$$X_{i1}, \ldots, X_{in_i} | (\theta_i, \sigma_i^2) \overset{\text{ind}}{\sim} N(\theta_i, \sigma_i^2) \qquad \text{for } i = 1, \ldots, k. \tag{1}$$

For the compound decision problem, we take $n_i = 1$. The LIMDIR process is formally defined as the limit, as $t \to \infty$, of a sequence of Dirichlet processes with base measures $\{\alpha_t\}_{t=1}^{\infty}$:

$$\begin{aligned} G &\sim \text{Dir}(\alpha_t); \\ \theta_1, \ldots, \theta_k | G &\overset{\text{IID}}{\sim} G. \end{aligned} \tag{2}$$

The Dirichlet process was described in detail by Ferguson (1973). The parameter of the Dirichlet process, α, is a measure which may be split into two parts. The first is the marginal distribution for θ_i, say G_0, whereas the second is the total mass of the measure, M. Thus α is often written as MG_0. The prior (and posterior) for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a mixture of several components. Each component corresponds to a different partition of $\boldsymbol{\theta}$ into $p \leqslant k$ subsets, or clusters, where all θ_i in a cluster are equal whereas those in different clusters may differ. The k-dimensional vector $\mathbf{s}$ indicates to which cluster each of the θ_i belongs. It is defined by the relationship $s_i = j$ if and only if θ_i is in the jth cluster. When comparing more than one partition, the first subscript on $\mathbf{s}$ indicates the partition, whereas a second subscript indicates the position in the vector. For each partition, we represent the sizes of the p clusters by $\mathbf{c} = (c_1, \ldots, c_p)$ and the locations of the clusters by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$.

The Dirichlet process determines the distribution of $(p, \boldsymbol{\gamma})$. Each partition with p clusters and cluster sizes $\mathbf{c}$ has a prior probability of

$$\pi(\mathbf{s}) = M^p \prod_{i=1}^{p} \Gamma(c_i) \Big/ \prod_{i=1}^{k} (M + i - 1)$$

(Antoniak, 1974). Given p, the γ_i form a random sample from G_0. The description of $\mathbf{s}$ may be governed by many conventions. We adopt the simple convention that, for a partition with p clusters, the first occurrences of the integers from 1 to p must occur in order, as $\mathbf{s}$ is read from left to right. This is the 'no-gaps' scheme for describing $\mathbf{s}$ that was used for computational purposes in MacEachern and Müller (2000).

The division of $\boldsymbol{\theta}$ into clusters is closely tied to the description of $\boldsymbol{\theta}$ arising from a Pólya urn scheme. Sethuraman (1994) used this feature to provide a constructive definition of the Dirichlet

process. Gopalan and Berry (1998) exploited the clustering to develop multiple-comparisons procedures based on a single mixture of Dirichlet processes model: treatment means are judged equivalent if they belong to the same cluster and different if they do not. The probability distribution over partitions of $\boldsymbol{\theta}$ thus directly provides multiple-comparison inferences. Product partition models (Crowley, 1997; Quintana and Iglesias, 2003) also have this feature, as do certain models of partial exchangeability (Mallick and Walker, 1997).

The posterior distribution on $\boldsymbol{\theta}$, for a fixed t under the model given by expressions (1)–(2) with G_{0t} an $N(0, \tau_t^2)$ measure, is a mixture of normal distributions. A component of the mixture results from the partition of $\boldsymbol{\theta}$ into p clusters. We use an asterisk to represent observations that are tied to treatments in a particular cluster. Hence $n_i^* = \Sigma_{l=1|s_l=i}^{k} n_l$, and $\Sigma_{j=1}^{n_i^*} X_{ij}^* = \Sigma_{l=1|s_l=i}^{k} \Sigma_{j=1}^{n_l} X_{lj}$ is the sum of all observations tied to cluster i. We define

$$\mathrm{SSE}_i^* = \sum_{j=1}^{n_i^*} (X_{ij}^* - \bar{X}_i^*)^2.$$

With likelihoods (1) and known variances, the location vector $\boldsymbol{\gamma}$ can be marginalized (e.g. MacEachern (1994)). The posterior distribution over the partitions is given by

$$P_t(\mathbf{s}|\mathbf{X}) \propto M_t^p \prod_{i=1}^{p} \left[\frac{\Gamma(c_i)}{(1+n_i^* \tau_t^2 \sigma^{-2})^{1/2}} \exp\left\{ \frac{\sigma^{-4} \left(\sum_{j=1}^{n_i^*} X_{ij}^* \right)^2}{2(\tau_t^{-2} + n_i^* \sigma^{-2})} - \frac{\sum_{j=1}^{n_i^*} (X_{ij}^*)^2}{2\sigma^2} \right\} \right], \tag{3}$$

when $\sigma_i^2 = \sigma^2$ for $i = 1, \ldots, k$. The conditional (on $\mathbf{s}$ and $\mathbf{X}$) distribution of the p cluster locations is a product of independent normal distributions, with means $\sigma^{-2} n_i^* \bar{X}_i^* / (\tau_t^{-2} + n_i^* \sigma^{-2})$ and variances $(\tau_t^{-2} + n_i^* \sigma^{-2})^{-1}$ respectively. The distribution on $(\mathbf{s}, \boldsymbol{\gamma})$ induces a distribution on $\theta_1, \ldots, \theta_k$. The distribution across partitions leads to the behaviour that was exploited by Berry and Christensen (1979) and Escobar (1994) for the compound decision problem, and described in a general setting by George (1986) as multiple shrinkage. We refer to the LIMDIR version of this known variance model as model 0. Formula (3) also applies to the case of known, but differing, σ_i^2. To use it, adjust the n_i to account for the differing σ_i^2.

The typical analysis-of-variance problem involves unknown σ_i^2. We concentrate on the traditional model which assumes equal variances. Conditional on the mean vector $\boldsymbol{\theta}$, and a common variance σ^2, the observations follow expression (1). This produces model 1:

$$\theta_1, \ldots, \theta_k \sim \mathrm{LIMDIR}(\{\alpha_t\});$$
$$\sigma^2 \text{ has density } \pi(\sigma^2) \propto \sigma^{-2}.$$

2.1. Limits of Dirichlet processes

The limits that we take are motivated by the posterior, conditional view (e.g. Berger (1993), section 4.4). We seek a limit for which the limiting posteriors have acceptable behaviour. We first discuss two standard limits, demonstrating their ineffectiveness for the multiple-comparisons and compound decision problems, and then present our choice of limit.

The standard limit when non-informative analyses are discussed for the Dirichlet process is to let $M \to 0$ (Ferguson, 1973). When $M \to 0$, and $\boldsymbol{\theta}$ is observed directly, the posterior estimate of G under an integrated squared error loss function is just $\hat{G}$, the empirical cumulative distribution function. However, when $\boldsymbol{\theta}$ is observed indirectly, with information about its components

filtered through the likelihood, the result is quite different. We marginalize γ and express the posterior as a sum over the partitions of $\boldsymbol{\theta}$.

$$\pi(\mathbf{s}_i|\mathbf{X}) = \frac{\pi(\mathbf{s}_i)\ f(\mathbf{X}|\mathbf{s}_i)}{\sum_{\mathbf{s}_j \in S} \pi(\mathbf{s}_j)\ f(\mathbf{X}|\mathbf{s}_j)}. \tag{4}$$

The likelihood $f(\mathbf{X}|\mathbf{s}_i)$ remains constant as $M \to 0$ whereas $\lim_{M\to 0}\{\pi(\mathbf{s}_i)/\pi(\mathbf{s}_j)\}$ is ∞ if $p_{\mathbf{s}_i} < p_{\mathbf{s}_j}$ and is $\pi(\mathbf{s}_i)/\pi(\mathbf{s}_j)$ if $p_{\mathbf{s}_i} = p_{\mathbf{s}_j}$. For our model, the likelihoods are all non-zero, and so the limiting posterior concentrates on the partition with $p = 1$ implying that all θ_i are equal. In the analysis-of-variance problem, all treatment means are equal with probability 1. In the compound decision problem, all θ_i are equal with probability 1, and so under typical loss functions all estimates of the θ_i are the same. The data have no effect on these limits!

In practice, many have taken M small but positive to represent a minimally informative analysis. However, since the limiting behaviour is degenerate, such a choice is highly informative rather than minimally informative. For discussion of this behaviour in a variety of settings, see Sethuraman and Tiwari (1982), Newton *et al.* (1996) or MacEachern (1998).

The standard limit in parametric Bayes models for normal means would take $\tau_t^2 \to \infty$, resulting in an improper uniform prior on the line for the marginal distribution of a treatment mean. The usual simultaneous limit for the distribution on the variances would result in a limiting prior density proportional to σ^{-2} (see, for example, Bernardo and Smith (2000), page 329). Discussion of this limit is complicated by the impropriety of the limiting prior. We assume that $\Sigma_{i=1}^{k} n_i \geqslant k+1$, so that the limiting posterior on $(\boldsymbol{\theta}, \sigma^2)$ is proper. Then, for each pair of partitions $\mathbf{s}_i$ and $\mathbf{s}_j$,

$$\lim_{t\to\infty}\left\{\frac{\pi(\mathbf{s}_i|\mathbf{X})}{\pi(\mathbf{s}_j|\mathbf{X})}\right\} = \lim_{t\to\infty}\left\{\frac{\pi(\mathbf{s}_i)\ f_t(\mathbf{X}|\mathbf{s}_i)}{\pi(\mathbf{s}_j)\ f_t(\mathbf{X}|\mathbf{s}_j)}\right\}. \tag{5}$$

Letting $\mathbf{s}_j$ denote the partition with $p = 1$ and $\mathbf{s}_i$ denote another arbitrary partition, we have, under model 1,

$$f_t(\mathbf{X}|\mathbf{s}_j) = \int\int \frac{1}{\tau_t}\phi\left(\frac{\gamma_1}{\tau_t}\right)\sigma^{-2}\prod_{j=1}^{n_1^*}\frac{1}{\sigma}\phi\left(\frac{X_{1j}^* - \gamma_1}{\sigma}\right)\mathrm{d}\gamma_1\,\mathrm{d}\sigma^2 \tag{6}$$

and

$$f_t(\mathbf{X}|\mathbf{s}_i) = \int\ldots\int \sigma^{-2}\prod_{l=1}^{p_{\mathbf{s}_i}}\frac{1}{\tau_t}\phi\left(\frac{\gamma_l}{\tau_t}\right)\prod_{j=1}^{n_i^*}\frac{1}{\sigma}\phi\left(\frac{X_{lj}^* - \gamma_l}{\sigma}\right)\mathrm{d}\gamma_1\ldots\mathrm{d}\gamma_{p_{\mathbf{s}_i}}\mathrm{d}\sigma^2, \tag{7}$$

where ϕ represents the standard normal probability density function. The ratio $f_t(\mathbf{X}|\mathbf{s}_i)/f_t(\mathbf{X}|\mathbf{s}_j) \to 0$, implying that the posterior probability of $\mathbf{s}_i \to 0$ also. Considering the result for each partition with $p > 1$ leads to the conclusion that $\pi(\mathbf{s}_j|\mathbf{X}) = 1$ in the limit. Thus the limiting posterior distribution concentrates on the event $p = 1$, resulting in poor inference. Again, the data have no effect on this limit!

We repair the inadequacies of the two standard limits by focusing on the goal of obtaining a limiting posterior distribution which does not concentrate on the event that all treatment means are equal. Instead, we seek a posterior that assigns positive probability to each partition of $\boldsymbol{\theta}$. We also wish to have a marginal prior distribution for θ_i that corresponds to the usual parametric reference prior, as in the second limit above. Our solution is to take a limit under which the local mass of the Dirichlet measure tends to a non-zero constant for each compact, non-null measurable set. As the next subsection shows, the particulars of the sequence that takes us to

a given local mass are of little importance, and so we have the freedom to choose a convenient sequence. The conjugate form for likelihood (1) is normal, and this simplifies calculation.

We return briefly to the case of known variance for ease of exposition. Focusing purely on the sequence of measures, $\{\alpha_t\}_{t=1}^{\infty}$, we take G_{0t} to be normal with mean 0 and variance τ_t^2. Our limit will send $\tau_t^2 \to \infty$. To stabilize the local mass, we define $M_t = M\sqrt{(2\pi)}\tau_t$ for some constant $M > 0$. The end result is a sequence of measures $\{\alpha_t\}_{t=1}^{\infty}$. The measure α_t assigns to the compact interval $[a, b]$ a mass of

$$M_t\{\Phi(b/\tau_t) - \Phi(a/\tau_t)\} = M\sqrt{(2\pi)}\tau_t\{\Phi(b/\tau_t) - \Phi(a/\tau_t)\},$$

with Φ representing the standard normal cumulative density function. The limit of these masses is $M(b-a)$. Importantly, this limiting value exists, is positive when $b > a$ and is finite.

Applying this limit to expression (3),

$$P_t(\mathbf{s}|\mathbf{X}) \propto \{M\sqrt{(2\pi)}\tau_t\}^p \prod_{i=1}^{p} \left[\frac{\Gamma(c_i)}{(1+n_i^*\tau_t^2\sigma^{-2})^{1/2}} \times \exp\left\{ - \frac{\sum_{j=1}^{n_i^*} X_{ij}^{*2} - \sigma^{-2}\left(\sum_{j=1}^{n_i^*} X_{ij}^*\right)^2 \Big/ (\tau_t^{-2} + n_i^*\sigma^{-2})}{2\sigma^2} \right\} \right] \tag{8}$$

$$\to M^p(2\pi\sigma^2)^{p/2} \prod_{i=1}^{p} \Gamma(c_i)(n_i^*)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\mathrm{SSE}_i^*\right). \tag{9}$$

The conditional limiting posterior distributions on the γ_i are independent normal distributions with means $\bar{X}_i^*$ and variances σ^2/n_i^*. We note that, conditional on a partition, this simple LIMDIR model captures the standard reference analysis. The additional benefit is the distribution over the partitions. The limit that we take for model 1 parallels that of model 0. The limiting posterior distribution for model 1 is described in Appendix A.

There are many ways to generate a non-informative prior distribution for a model. With a single treatment, the model that we have written is the standard prior in a location–scale setting (Berger (1993), page 88). It preserves an *a priori* independence between the treatment mean and variance. It also gives the standard model, conditional on each partition of $\boldsymbol{\theta}$. We note that LIMDIR versions of other non-informative analyses can be created.

2.2. Local mass

The results in this subsection justify the concept of local mass. To do so, we consider a LIMDIR model based on a likelihood (to replace likelihood (1)), satisfying only minimal conditions. The results show that a wide variety of sequences $\{\alpha_t\}_{t=1}^{\infty}$ lead to the same limiting behaviour. The formal condition on α_t in the middle (a compact set that contains the region of substantial likelihood) and the tails (a co-compact set where the likelihood is negligible) is that the tail measure not contribute too much to integrals relative to the middle measure.

Throughout this subsection, we assume that θ_i lies in $\mathbf{R}$ for $i = 1, \ldots, k$ and that the $P(X_i|\theta_i)$ are mutually absolutely continuous in θ_i. These restriction can be relaxed, though at a cost of considerable complexity in the mathematics. The mutual absolute continuity condition avoids division by 0 in the proof of theorem 1.

Definition 1. Given a configuration $\mathbf{s}$ and data $\mathbf{X}$, for any fixed $0 < \varepsilon < 1$, a set $K = [-k_1, k_1] \times \ldots \times [-k_p, k_p]$ in $\mathbf{R}^p$ is said to be the middle if

$$(1-\varepsilon)\int\ldots\int P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p)\leqslant\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p)$$

where α is a positive, σ-finite (possibly infinite), measure. The set K^{c} is said to be the tail.

Fig. 1 provides intuition behind this definition of the middle and tail of the measure. Fig. 1(a) shows a set of normal likelihoods for individuals with means θ_i and $k=3$, ignoring potential clustering. Taking $k_1=10\sigma+\max_{i=1,\ldots,k}|X_i|$, and setting all other $k_i=k_1$, we note that $\boldsymbol{\theta}$ in the tail have very much smaller likelihoods than do values of $\boldsymbol{\theta}$ that are supported by the data. The middle contains all values of $\boldsymbol{\theta}$ that are even remotely supported by the data. Fig. 1(b) shows three normal measures in a sequence that we envision tending to a multiple of Lebesgue measure. The sharpest, and first in the sequence, assigns little mass in the neighbourhood of the likelihoods, and so the posterior distribution assigns great probability to a small number of large clusters. The second measure, with a modest spread, assigns more mass around the likelihoods. With the off-centre likelihoods, the posterior distribution is tipped towards a smaller number of larger clusters than is the prior distribution. The third measure is, in the region of the plot, close to a multiple of Lebesgue measure. Posterior inference on $\boldsymbol{\theta}$ under this model is well along the way to the limiting posterior inference that is obtained under the LIMDIR model. Fig. 1(c) shows three measures whose distribution functions match those in Fig. 1(b). However, the mass parameters for the three measures are identical. This sequence illustrates the breakdown of the second limit in Section 2.1. The lack of local mass in the region that is supported by the likelihood implies that the posterior assigns a high probability to a single cluster.

Convergence of the posterior distribution of $\mathbf{s}$ in a well-defined sense relies on two conditions.

(a) Condition on the middle: for each $\varepsilon>0$, there is T_1 such that, for all $t>T_1$,

$$\left|\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha_t(\mathrm{d}\gamma_1)\ldots\alpha_t(\mathrm{d}\gamma_p)-\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p)\right|$$
$$<\varepsilon\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p).$$

(b) Condition on the tail: for each $\varepsilon>0$, there is T_2 such that, for all $t>T_2$,

$$\int\ldots\int_{K^{\mathrm{c}}} P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha_t(\mathrm{d}\gamma_1)\ldots\alpha_t(\mathrm{d}\gamma_p)<\frac{\varepsilon}{1-\varepsilon}\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p).$$

The condition on the middle is very mild. The main case of interest is when α_t and α admit densities with respect to Lebesgue measure. In this setting, if a relative difference in densities tends to 0, the condition will hold. Denoting the implied product densities over K by f_t and f, the condition will be satisfied whenever a version of the densities exists such that $\sup_{\gamma\varepsilon K}|f_t(\boldsymbol{\gamma})-f(\boldsymbol{\gamma})|/f(\boldsymbol{\gamma})\to 0$. Examples of such convergence include a sequence of scaled normal measures or a sequence of bounded, uniform measures converging to a multiple of Lebesgue measure. Alternatively, if the likelihood is nicely behaved, the condition is satisfied if $\alpha_t-\alpha$ converges to the null measure on K. Such nice behaviour occurs if, conditional on each $\mathbf{s}$, we have $\sup_{\gamma,\gamma'\varepsilon K} P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})/P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma}')\leqslant B$ for some $B<\infty$. This would hold, for example, for any likelihood in a full exponential family.

The condition on the tail is best understood by considering an example which violates it. Suppose that the likelihood is standard normal with $n_i=1$ and $k=2$. Define the measure α_t to be Lebesgue measure on $[-t,t]$ if t is odd, and Lebesgue measure on $[-t,t]$ with a point mass of measure $\exp\{(t+1)^4\}$ at $t+1$. The condition on the tail (and the condition on the middle) holds for the odd subsequence of measures, and the limiting measure α is Lebesgue measure on

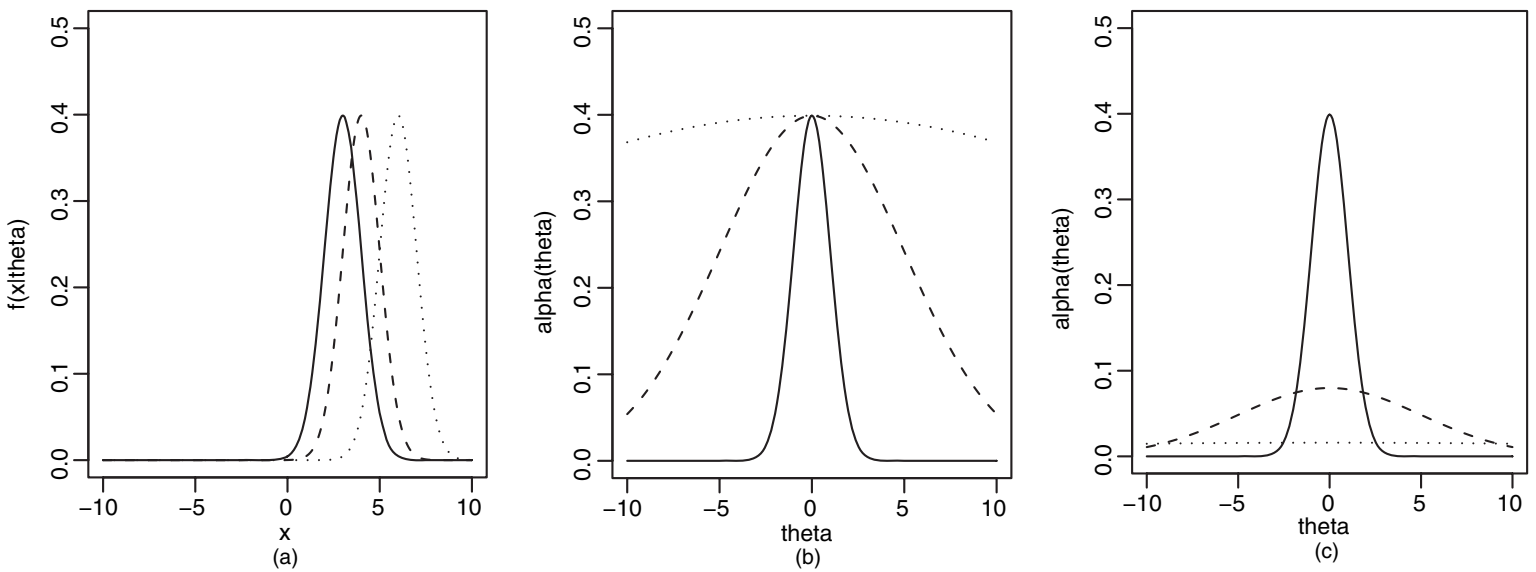

Fig. 1. Intuition behind the definition of the middle and the tails: (a) a set of normalized likelihoods for individuals with mean θ_i and $k = 3$ (———, $N(3,1)$; – – –, $N(4,1)$; · · · · · · ·, $N(6,1)$); (b) three normal measures in a sequence that we envision tending to a multiple of Lebesgue measure (———, standard deviation 1; – – –, standard deviation 5; · · · · · · ·, standard deviation 25); (c) three measures whose distribution functions match those in (b) but the mass parameters for the three are identical (———, standard deviation 1, – – –, standard deviation 5; · · · · · · ·, standard deviation 25)

the real line. The limiting posterior distribution assigns positive mass to each of the two possible values of $\mathbf{s}$. For the even subsequence, the posterior distributions are dominated by the spikes of mass. For each γ_i, these increase as $\exp(t^4)$ whereas the likelihood decreases at a mere rate of $\exp(-t^2/2)$. The resulting probabilities in the sequence of posterior distributions increase at a rate of $\exp(t^4)$ for $s_1 = s_2$ and at a rate of $\exp(2t^4)$ for $s_1 \neq s_2$. Thus, the even subsequence assigns posterior probability tending to 1 to $s_1 \neq s_2$ and so has different limiting behaviour from that of the odd subsequence. Together there is no well-defined limit. Imposition of the condition on the tail rules out sequences of measures that have enormous growing spikes of mass which move progressively further out in the tails.

Theorem 1. Suppose that there is a sequence of measures $\{\alpha_t\}_{t=1}^{\infty}$ and a measure α. If the sequence $\{\alpha_t\}_{t=1}^{\infty}$ and α satisfy the aforementioned conditions on the middle and the tail, then the posterior distribution of $\mathbf{s}$ under the mixture of Dirichlet processes models with base measures α_t converges to a well-defined limit. We describe this limit as the posterior distribution on $\mathbf{s}$ under the LIMDIR model with base measure α.

Proof. Choose a pair of partitions $\mathbf{u}$ and $\mathbf{v}$. Fix $\varepsilon > 0$ and less than 1, and find a middle and tail for $\mathbf{u}$ that apply to the α_t and α. Do the same for $\mathbf{v}$. Then, with the second subscript on T indicating the partition, $t > \max(T_{1\mathbf{u}}, T_{1\mathbf{v}}, T_{2\mathbf{u}}, T_{2\mathbf{v}})$,

$$\begin{aligned}
\frac{P_t(\mathbf{s}=\mathbf{u}|\mathbf{X})}{P_t(\mathbf{s}=\mathbf{v}|\mathbf{X})} &\leqslant \frac{(1-\varepsilon^2+\varepsilon)\int\dots\int_{K_{\mathbf{u}}} P(\mathbf{X}|\mathbf{u},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{u}}}\Gamma(c_{u_l})\,\alpha(\mathrm{d}\gamma_1)\dots\alpha(\mathrm{d}\gamma_{p_{\mathbf{u}}})}{(1-\varepsilon)\int\dots\int_{K_{\mathbf{v}}} P(\mathbf{X}|\mathbf{v},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{v}}}\Gamma(c_{v_l})\,\alpha_t(\mathrm{d}\gamma_1)\dots\alpha_t(\mathrm{d}\gamma_{p_{\mathbf{v}}})} \\
&\leqslant \frac{(1-\varepsilon^2+\varepsilon)\int\dots\int_{K_{\mathbf{u}}} P(\mathbf{X}|\mathbf{u},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{u}}}\Gamma(c_{u_l})\,\alpha(\mathrm{d}\gamma_1)\dots(\mathrm{d}\gamma_{p_{\mathbf{u}}})}{(1-\varepsilon)^2\int\dots\int_{K_{\mathbf{v}}} P(\mathbf{X}|\mathbf{v},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{v}}}\Gamma(c_{v_l})\,\alpha(\mathrm{d}\gamma_1)\dots\alpha(\mathrm{d}\gamma_{p_{\mathbf{v}}})} \\
&\leqslant \frac{(1-\varepsilon^2+\varepsilon)\,P(\mathbf{u}|\mathbf{X})}{(1-\varepsilon)^3\,P(\mathbf{v}|\mathbf{X})},
\end{aligned}$$

$$\begin{aligned}
\frac{P_t(\mathbf{s}=\mathbf{u}|\mathbf{X})}{P_t(\mathbf{s}=\mathbf{v}|\mathbf{X})} &\geqslant \frac{(1-\varepsilon)\int\dots\int_{K_{\mathbf{u}}} P(\mathbf{X}|\mathbf{u},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{u}}}\Gamma(c_{u_l})\,\alpha_t(\mathrm{d}\gamma_1)\dots\alpha_t(\mathrm{d}\gamma_{p_{\mathbf{u}}})}{(1-\varepsilon^2+\varepsilon)\int\dots\int_{K_{\mathbf{v}}} P(\mathbf{X}|\mathbf{v},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{v}}}\Gamma(c_{v_l})\,\alpha(\mathrm{d}\gamma_1)\dots\alpha(\mathrm{d}\gamma_{p_{\mathbf{v}}})} \\
&\geqslant \frac{(1-\varepsilon)^2\int\dots\int_{K_{\mathbf{u}}} P(\mathbf{X}|\mathbf{u},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{u}}}\Gamma(c_{u_l})\,\alpha(\mathrm{d}\gamma_1)\dots\alpha(\mathrm{d}\gamma_{p_{\mathbf{u}}})}{(1-\varepsilon^2+\varepsilon)\int\dots\int_{K_{\mathbf{v}}} P(\mathbf{X}|\mathbf{v},\boldsymbol{\gamma})\prod_{l=1}^{p_{\mathbf{v}}}\Gamma(c_{v_l})\,\alpha(\mathrm{d}\gamma_1)\dots(\mathrm{d}\gamma_{p_{\mathbf{v}}})} \\
&\geqslant \frac{(1-\varepsilon)^3\,P(\mathbf{u}|\mathbf{X})}{(1-\varepsilon^2+\varepsilon)\,P(\mathbf{v}|\mathbf{X})}.
\end{aligned}$$

Thus,

$$\lim_{t\to\infty}\left\{\frac{P_t(\mathbf{u}|\mathbf{X})}{P_t(\mathbf{v}|\mathbf{X})}\right\} = \frac{P(\mathbf{u}|\mathbf{X})}{P(\mathbf{v}|\mathbf{X})}.$$

Applying this argument to all pairs $(\mathbf{u}, \mathbf{v})$ yields the result. □

To show convergence of the posterior distribution of $\mathbf{s}$, it is enough to show that $\{\alpha_t\}_{t=0}^{\infty}$ and α satisfy the conditions of theorem 1.

Theorem 2. Let $\{\alpha_t\}_{t=1}^{\infty}$ be a sequence of measures of the Dirichlet process where $M_t = M\sqrt{(2\pi)}\tau_t$, G_{0t} is normal with mean 0 and variance τ_t^2, and $\lim_{t\to\infty}(\tau_t^2) = \infty$. Define the measure of the LIMDIR process, α, to be M times Lebesgue measure on $\mathbf{R}$. Assume that there is some finite B for which $\int\ldots\int P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p) < B$ for all $\mathbf{s}$. Then the posterior distribution of $\mathbf{s}$ under the Dirichlet process model with base measure α_t converges to the posterior distribution of $\mathbf{s}$ under the LIMDIR model with base measure α.

Proof. Fix $\mathbf{s}$, $\mathbf{X}$ and ε. By assumption, $\int\ldots\int P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p) < B$, and so there is some k_1 for which the set $K = [-k_1, k_1]^p$ is a middle of α as in definition 1. Let $f(\cdot)$ denote the $N(\cdot; 0, \tau_t^2)$ density. $M_t f(k_1) < \alpha_t(\mathrm{d}\gamma)$ inside the set K whereas the inequality is reversed for K^{c}. This ensures that K is a middle for α_t for all t.

(a) Condition on the middle:

$$\left|\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha_t(\mathrm{d}\gamma_1)\ldots\alpha_t(\mathrm{d}\gamma_p) - \int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p)\right| \leqslant \{M - M_t f(k_1)\}^p \int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p).$$

The constant in front of the integral is $M^p\{1 - \exp(-k_1^2/2\tau_t^2)\}^p$. A sufficiently large t ensures that the condition holds.

(b) Condition on the tails: we use the fact that $\alpha_t(\mathrm{d}\gamma)/\alpha(\mathrm{d}\gamma) = \exp(-\gamma^2/2\tau_t^2) < 1$. Thus,

$$\int\ldots\int_{K^{\mathrm{c}}} P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha_t(\mathrm{d}\gamma_1)\ldots\alpha_t(\mathrm{d}\gamma_p) < \int\ldots\int_{K^{\mathrm{c}}} P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p) < \frac{\varepsilon}{1-\varepsilon}\int\ldots\int_K P(\mathbf{X}|\mathbf{s},\boldsymbol{\gamma})\,\alpha(\mathrm{d}\gamma_1)\ldots\alpha(\mathrm{d}\gamma_p).$$

2.3. Propriety and non-degeneracy of the limiting posteriors

To provide effective inferential procedures, we require posterior distributions that are proper for the main quantities of interest (i.e. a proper joint distribution of the treatment means and variance). We also require that the posterior that is based on a finite sample be non-degenerate. These two properties are satisfied for models 0 and 1 under mild conditions on the sampling distributions of the $X_i|(\theta_i, \sigma_i^2)$. The next result follows from formal calculation.

Theorem 3. The limiting posterior distribution of $\boldsymbol{\theta}$ calculated under model 0 with the sequence of measures that is described in theorem 2 is proper. Furthermore, all partitions of $\boldsymbol{\theta}$ receive positive posterior probability.

When σ^2 is unknown, a continuity condition and an additional sample size condition are needed to ensure that the posterior distribution on $(\boldsymbol{\theta}, \sigma^2)$ is proper. Formal calculation produces the following result for model 1. Appendix A contains the calculations.

Theorem 4. Assume that $\Sigma_{i=1}^{k}\mathrm{SSE}_i > 0$. Also assume that $n_i > 1$ for some $i = 1, \ldots, k$. Then the limiting posterior distribution of $(\boldsymbol{\theta}, \sigma^2)$ calculated under model 1 with the sequence of measures that is described in theorem 2 is proper. Furthermore, all partitions of $\boldsymbol{\theta}$ receive positive posterior probability.

To conclude the section, we note that the entire limiting posterior distribution is not proper. Under both models, the limiting posterior distribution for a new treatment, say θ_{k+1}, is uniform on the real line. The limiting posterior probability that $\theta_{k+1} = \theta_i$, $i \leqslant k$, is 0. However, if data were collected on this new treatment, the updated limiting posterior on the expanded $(\boldsymbol{\theta}, \sigma^2)$ would be proper under models 0 and 1. Inference on θ_{k+1} can be made, conditional on its joining an existing cluster. To do so, select one of the θ_i, $i = 1, \ldots, k$, at random, and set θ_{k+1} equal to it. We study the performance of this conditional method in the second predictive exercise of Section 4.

3. Elicitation of the prior distribution

Our recommended elicitation of the prior distribution consists of two stages. In the first, a limiting model is selected. In the second, a value is chosen for the limiting local mass of the Dirichlet processes' base measure. The choice of limiting local mass may be made to match classical concerns or it may be made to match subjective assessment. With classical concerns in the analysis-of-variance setting, if the main concern is the pairwise error rate, we recommend a choice of M that yields the targeted pairwise type I error rate. If the main concern is the experimentwise error rate, we recommend a choice of M that yields the targeted experimentwise type I error rate. For both types of calibration, to determine whether an error occurs, we need to set a cut-off threshold λ.

To calibrate the local mass to achieve a certain pairwise type I error rate, it is helpful to assume that there are only two treatments. Defining a to be the pairwise error rate, we set

$$a = \frac{1}{N} \sum_{i=1}^{N} I\{P^{(i)}(\theta_1 = \theta_2 | \sigma_1^2, \sigma_2^2, X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) < \lambda\},$$

where $i = 1, \ldots, N$ indexes the replicate in the simulation. The data, $X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$, are generated under the assumption that $\theta_1 = \theta_2$. Under model 1, we can write

$$P^{(i)}(\theta_1 = \theta_2 | \sigma^2, X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) = A/(A + MB),$$

where the variables A and B are easily simulated. Working with formula (10) from Appendix A, we have

$$A = (n_1 + n_2)^{-1/2} \Gamma\left(\frac{n_1 + n_2 - 1}{2}\right) \left\{\frac{1}{2}(Y_1 + Y_2)\right\}^{(n_1+n_2-1)/2}$$

and

$$B = (2\pi)^{1/2} n_1^{-1/2} n_2^{-1/2} \Gamma\left(\frac{n_1 + n_2 - 2}{2}\right) \left(\frac{1}{2} Y_1\right)^{(n_1+n_2-2)/2},$$

where Y_1 is $\sigma^2 \chi^2(n_1 + n_2 - 2)$ and Y_2 is independently $\sigma^2\ \chi^2(1)$. If the experiment was designed with a particular value of σ^2 in mind, we use that value of σ^2. Otherwise, we take σ^2 to be the mean-square error from a one-way analysis-of-variance on the actual data. Working from these expressions, a simulation is performed to find a value of M for which a is approximately the targeted type I error probability. Experimentwise calibration proceeds in a similar fashion. We generate the data from the null hypothesis that $\theta_1 = \ldots = \theta_k$ and average the experimentwise errors. The simplicity of the simulation enables us to perform the computations to an arbitrary degree of accuracy.

Subjective assessment of the local mass parameter, as with subjective elicitation of any prior distribution, can be done in many ways (e.g. Hartigan (1983), chapter 1). In our context, one

approach is to focus on $E[P(\theta_1 = \theta_2)|\sigma^2, X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}]$ under the assumption that $\theta_1 = \theta_2$. The expectation can be evaluated with the help of formulae in Appendix A and either a simulation or numerical integration.

4. The new sales insurance data

In this section, we illustrate our techniques with a new sales data set from an insurance company. The data consist of monthly counts of new policies and number of households by designated marketing area (DMA) in 2006 and 2007. A DMA is a geographic region which is covered by the same broadcast television and radio stations and the same local newspapers. A DMA typically consists of a large city and the surrounding area, and individuals within the DMA are exposed to the same advertising campaigns. Companies tailor marketing and sales efforts to the individual DMA.

Index the DMA by i. The data consist of the number of households in DMA i, HH_i, and the count of new sales in month j, NS_{ji}, $i = 1, \ldots, 200$, $j = 1, \ldots, 24$. We assume that $\mathrm{NS}_{ji} \sim \mathrm{Poisson}(p_i \mathrm{HH}_i)$. The parameter p_i describes the ith DMA's true but unknown sales rate. Following McCullagh and Nelder (1989), we use the variance stabilization transformation, $X_{ji} = \sqrt{(\mathrm{NS}_{ji}/\mathrm{HH}_i)}$, so that X_{ji} is approximately normal with mean $\theta_i = \sqrt{p_i}$ and variance $\sigma^{*2}_{ji} = 1/4\mathrm{HH}_i$. To make the normal approximation accurate, our analysis included only DMAs having $\mathrm{HH}_i > 50000$. For the aggregated analyses, the data are collapsed to quarters or years before transformation.

Insurance sales show seasonal variation and are affected by many factors, among them pricing, marketing activities and economic conditions. Preliminary analysis of the data confirmed these effects. To account for them, we limit comparisons to the same time period in successive years, we adjusted new sales in 2007 by multiplying the ratio of the means of the two years and we adjusted for overdispersion. We use a prime to denote the adjusted variables for 2007, so X'_{ji} represents the transformed, adjusted sales rate for 2007. To account for overdispersion, we estimated an overdispersion factor for each month of

$$\hat{\rho}_j = \sum_{i=1}^{n_j} (X_{ji} - X'_{ji})^2 \Big/ \sum_{i=1}^{n_j} (\sigma^{*2}_{ji} + \sigma'^{*2}_{ji})$$

and inflated σ^{*2}_{ji} and σ'^{*2}_{ji} by $\hat{\rho}_j$. We denote the inflated variances by σ^2_{ji} and σ'^2_{ji} for the first and second years respectively.

We illustrate the LIMDIR model and compare it with a mixture of Dirichlet processes model and a parametric hierarchical Bayes model through two predictive exercises. The first is prediction of each month of 2007 based on the corresponding month of 2006. The second is prediction of a DMA's new sales for each month of 2006 based on other DMAs' new sales. To study how the amount of DMA-specific information impacts prediction, we merge the monthly data sets into quarterly and full year data sets, and conduct the same predictive exercises.

The distribution of the X_{ji} appears to be unimodal, with noticeable tail decay, and with right skewness. In our judgement, the normal distribution with mean μ and variance τ^2 is a reasonable choice for a base measure of the mixture of Dirichlet processes model. For the parametric Bayes model, we placed a normal prior $N(\mu, \tau^2)$ on $\boldsymbol{\theta}$ and an improper prior on μ and τ^2 which was proportional to $1/\tau^2$.

The LIMDIR model was implemented with a base measure proportional to Lebesgue measure. The mass parameter for the mixture of Dirichlet processes model was chosen on the basis of the pairwise error rate as described in Section 3. The pairwise type 1 error rate was taken to

be 0.05 and the cut-off threshold was set at 0.5. The variance that was used in the calibration was the mean of the σ_{ji}^2. The resulting mass parameters ranged from 0.743 to 1.156. The local mass for the LIMDIR model was calibrated in a similar fashion. This led to local mass parameters ranging from 12.095 to 24.677. Note that the local mass and the mass for the mixture of Dirichlet processes model are not directly comparable, as constants have been swept into the local mass to simplify calculations.

The estimators are compared on the basis of the sum of squared prediction error and sum of log-marginal-likelihood. When predicting the sales rate of 2007, the sum of squared prediction error is given by $\mathrm{SSPE} = \Sigma_{i=1}^{n_j}(X'_{ji} - \hat{\theta}_i)^2$, where X'_{ji} is the transformed, adjusted sales rate of the ith DMA in the jth month of 2007, and $\hat{\theta}_i$ is the posterior mean of θ_i. The sum of log-marginal-likelihoods is $\Sigma_{i=1}^{n_j}\log\{m(X'_{ji}|\mathbf{X}_j)\}$. For a single DMA, this is the log(marginal likelihood). Summing across DMAs provides an indication of the predictive ability of competing models on a log-likelihood scale. Computationally, it is far more stable than is the log(joint marginal likelihood) of all DMAs.

We evaluate $\hat{\theta}_i$ with a Gibbs sampling run for the LIMDIR model and the mixture of Dirichlet processes models. 200000 iterates were used for estimation, after a burn-in period of 10000 iterates. For the parametric Bayes model, we used numerical integration over a fine grid for τ^2 to evaluate $\hat{\theta}_i$.

Table 1 presents the results of the first predictive exercise, forecasting 2007 performance from 2006 results. The three models are essentially equivalent in terms of SSPE, with minor differences in trailing decimal places. In terms of log(marginal likelihood), the LIMDIR model is the best overall performer. Performance varies with the amount of data that is used to create the forecasts. When a full year's data are used to forecast the next year's performance, σ_{ji}^2 is quite small and the likelihood for the DMA is quite strong, and so there is little benefit in pooling information across DMAs. When the likelihood is weaker, as for the quarterly and monthly data, the benefit of pooling information is greater. This, coupled with the skewed distribution of the X_{ij}, leads to superior performance for the non-parametric methods. We examined which DMAs contribute to the difference in evaluation. There is little difference

Table 1. Predictive performance for sales rates in 2007 (X'_j) based on sales rates in 2006 (X_j)†

Model	*Parameter or criterion*	*Results for the following data sets:*		
		Yearly	*Quarterly*	*Monthly*
LIMDIR process	M	12.095	17.631	24.675
	SSPE	0.0047	0.0022	0.0011
	$\log\{m(\mathbf{X}'_j\|\mathbf{X}_j)\}$	740.826	821.450	912.296
Mixture of Dirichlet processes	M	1.156	0.875	0.743
	SSPE	0.0047	0.0022	0.0011
	$\log\{m(\mathbf{X}'_j\|\mathbf{X}_j)\}$	718.728	821.283	908.773
Parametric Bayes	SSPE	0.0047	0.0022	0.0011
	$\log\{m(\mathbf{X}'_j\|\mathbf{X}_j)\}$	744.724	819.290	909.154

†Performance is measured by SSPE (the sum of squared prediction error) and the sum of log(marginal likelihoods). The mass parameter M was recalibrated for each quarter and each month. For the quarterly data and the monthly data, the mean across quarter (month) is presented.

Table 2. Predictive performance of the models when a DMA is held out†

Model	*Parameter or criterion*	*Results for the following data sets:*		
		Yearly	*Quarterly*	*Monthly*
LIMDIR process	M	12.095	17.632	24.677
	SSPE	0.144	0.037	0.013
	$\log\{m(\mathbf{X}'_j\|\mathbf{X}_j)\}$	2901.124	5095.572	7964.198
Mixture of Dirichlet processes	M	1.155	0.876	0.748
	SSPE	0.144	0.037	0.013
	$\log\{m(\mathbf{X}'_j\|\mathbf{X}_j)\}$	2965.649	5093.092	7957.310
Parametric Bayes	SSPE	0.144	0.037	0.013
	$\log\{m(\mathbf{X}'_j\|\mathbf{X}_j)\}$	2046.349	4042.740	6866.515

†Evaluation is specific to the full year, each quarter or each month of 2006. Performance is measured by SSPE (the sum of squared prediction error) and sum of log(marginal likelihoods). For the quarterly data and the monthly data, the mean across quarter (month) is presented.

between the three models for the DMAs with middling X_{ji}. The main difference is in the treatment of DMAs with extreme values of X_{ji}. In the mixture of Dirichlet processes model, these DMAs tend to fall in larger clusters than in the LIMDIR model owing to the much smaller local mass in the tails. This excessive tail clustering leads to poorer predictive performance.

The second predictive exercise examines prediction for a new DMA. We hold out one DMA at a time and predict the sales rate of the DMA on the basis of the rest of the DMAs. The prediction under the LIMDIR model is conditional, requiring the new DMA to join an existing cluster. The results are shown in Table 2. As in the first predictive exercise, there is very little difference in SSPE. This follows from the fact that we are not using any data to distinguish between DMAs. The point predictions and hence SSPE are essentially the mean for the remaining DMAs under all three models. There are, however, dramatic differences in log(marginal likelihood). The two non-parametric models have the flexibility to capture the non-normality of the distribution of the X_{ji} and so have far greater log(marginal likelihood) than the parametric model. The mixture of Dirichlet processes model is the best performer for the yearly data, whereas the LIMDIR model is the best performer when the likelihoods are weaker in the quarterly and monthly data.

Further analyses that illustrate use of the LIMDIR model and that contrast its performance with that of mixture of Dirichlet processes models are available in Bush *et al.* (2007). Their investigation focused on multiple comparisons in the one-way analysis of variance. They developed a suite of models for a variety of formulations of the problem. In particular, they developed models for the Behrens–Fisher problem (different σ_i^2) for independent mean and variance structures that allows some treatments to have the same variance, and for a joint mean and variance structure that focuses on identical or differing (θ_i, σ_i^2). The last two types of model require the use of LIMDIR models on a half-plane (for the variance structure) and in a two-dimensional space (for the combined mean and variance structure) respectively.

5. Conclusions

The LIMDIR model that was developed in this paper extends the popular mixture of Dirichlet processes model. It allows us to use an improper base measure for the Dirichlet process, thus creating a sort of non-informative, non-parametric Bayesian analysis. The non-informa-

tive analysis provides insight into how to use the proper mixture of Dirichlet processes model, and it suggests several interesting extensions.

First, the development of the LIMDIR models that was presented in this paper has been tied to normal locations. The strategy is first to identify a non-informative analysis for a single-component problem, then to create a sequence of base measures that preserve local mass and that lead to this analysis for the single component, and then to describe explicitly the limiting local mass. That this strategy applies quite generally is indicated by the results in Section 2.2. This strategy allows us to create Dirichlet-based Jeffreys analyses, to adapt Bernardo's (1979) reference priors, etc.

A traditional advantage of non-informative analyses is that they simplify calculation of the posterior distribution. This advantage is less important in the modern environment of simulation-based fits, but it is still relevant when it comes to calibration of a prior distribution and also for problems involving a large number of treatments. The advantages carry over from the parametric setting to the LIMDIR setting, enabling us to implement very efficient computational algorithms to fit the models. The simplicity of the calculations should also prove useful in developing extensions of fractional Bayes factors (O'Hagan, 1995) and intrinsic Bayes factors (Berger and Pericchi, 1993) to a non-parametric Bayesian setting. Such extension would allow routine use of the models for automated Bayesian analysis. The LIMDIR models provide a springboard from which these techniques can be developed.

The development of the LIMDIR models carries a strong message for mixture of Dirichlet processes modelling. The main message is the importance of local mass. In selecting the mass parameter of the Dirichlet process, the relevant quantity is local mass—not the entire mass of the base measure. We recommend replacing the standard independent prior distributions on G_0 and M (Escobar and West, 1995) with a distribution on G_0 and a distribution on $M|G_0$. The latter distribution would be chosen to keep the local mass towards the centre of the distribution approximately constant.

In practice, many users of mixture of Dirichlet process models deliberately choose an overdispersed base measure in the mistaken belief that this reduces the effect of the prior distribution on the analysis. An overdispersed analysis which ignores the concept of local mass results in an artificially small number of clusters.

For analyses with an overdispersed base measure, calibration via local mass is essential. However, the posterior distribution of θ_{k+1} remains unrealistic. The entire mass of an overdispersed base measure will be large, leading to too much weight given to θ_{k+1} initiating a new cluster. For LIMDIR models with an improper marginal distribution, we have shown that use of the conditional distribution of θ_{k+1}, given that it joins an already established cluster, performs well. This fix can be used with proper overdispersed distributions as well. In addition, a second fix is available: we can average this conditional distribution with a downweighted and less dispersed substitute for the base measure.

We believe that the LIMDIR models will find a useful place in the arsenal of non-parametric Bayesian techniques. We also believe that consideration of the models will improve the practice of non-parametric Bayesian analysis, and that the models open a valuable collection of research problems.

Acknowledgements

This work was supported by the National Science Foundation under grants SES-0437251 and DMS-0605041. The views in this paper are not necessarily those of the National Science Foundation.

Appendix A

This section contains the derivation of the posterior probability (up to the normalizing constant) of a particular configuration under model 1. For this particular configuration, we assume that there are p clusters, of size $c_1, \ldots, c_p$. The asterisk notation denotes observations within a particular cluster. n denotes $\Sigma_{i=1}^{k} n_i$. For clarity, we write v in place of σ^2.

To compute the posterior probabilities under model 1, we extend the previous calculation under model 0 to account for an unknown variance. Carrying that portion of the posterior dealing with the variance throughout the model 0 calculation, we write, with P_a denoting the probability under some fixed $a > 0$,

$$P_a(\mathbf{s}|X) \propto M^p \int (2\pi v)^{-(n-p)/2} \prod_{i=1}^{p} \Gamma(c_i)(n_i^*)^{-1/2} \exp\left(-\frac{\mathrm{SSE}_i^*}{2v}\right) \frac{b^{-a}}{\Gamma(a)} v^{-(a+1)} \exp\left(-\frac{1}{vb}\right) \mathrm{d}v$$

$$\to M^p (2\pi)^{-(n-p)/2} \prod_{i=1}^{p} \Gamma(c_i)(n_i^*)^{-1/2} \int v^{-(n-p)/2-1} \exp\left(-\frac{1}{v} \sum_{i=1}^{p} \frac{\mathrm{SSE}_i^*}{2}\right) \mathrm{d}v$$

$$\propto M^p (2\pi)^{p/2} \prod_{i=1}^{p} \Gamma(c_i)(n_i^*)^{-1/2} \Gamma\left(\frac{n-p}{2}\right) \left(\sum_{i=1}^{p} \frac{\mathrm{SSE}_i^*}{2}\right)^{-(n-p)/2}. \tag{10}$$

Conditional on a particular cluster vector $\mathbf{s}$, the posterior distribution of v is inverse gamma with parameters $a' = (n-p)/2$ and $b' = (\Sigma_{i=1}^{p} \mathrm{SSE}_i^*/2)^{-1}$. For this inverse gamma distribution to exist, we must have $a' > 0$ (and hence $n > p$) and we must also have $b' > 0$. This leads to the conditions for the existence of a proper prior distribution that were given in Section 2. Since the final expression is positive and finite for all partitions, and since the same constant of proportionality applies to all partitions, each partition receives positive posterior probability. Given $\mathbf{s}$ and v, the posterior distributions of the γ_i are independent normal distributions with means $\bar{X}_i^*$ and variances v/n_i^*.

The simplicity of the formulae for models 0 and 1 allows us to evaluate directly the posterior distribution on $\mathbf{s}$ for experiments with a modest number of treatments. For larger experiments, following MacEachern's (1994) refinement of Escobar's (1994), we can run a Gibbs sampler on $\mathbf{s}$.

References

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Ann. Statist.*, **2**, 1152–1174.

Berger, J. O. (1993) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.

Berger, J. O. and Pericchi, L. R. (1993) The intrinsic Bayes factor for model selection. *J. Am. Statist. Ass.*, **91**, 109–122.

Bernardo, J. M. (1979) Reference posterior distributions for Bayesian inference (with discussion). *J. R. Statist. Soc.* B, **41**, 113–147.

Bernardo, J. M. and Smith, A. F. M. (2000) *Bayesian Theory*. Chichester: Wiley.

Berry, D. and Christensen, R. (1979) Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.*, **7**, 558–568.

Bush, C. A., Lee, J. and MacEachern, S. N. (2007) Minimally informative nonparametric Bayesian analysis, with application to multiple comparisons. *Technical Report*. Department of Statistics, Ohio State University, Columbus.

Crowley, E. M. (1997) Product partition models for normal means. *J. Am. Statist. Ass.*, **92**, 192–198.

DeIorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An ANOVA model for dependent random measures. *J. Am. Statist. Ass.*, **99**, 205–215.

Dey, D., Müller, P. and Sinha, D. (eds) (1998) *Practical Nonparametric and Semiparametiric Bayesian Statistics*. New York: Springer.

Escobar, M. D. (1994) Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Ass.*, **89**, 268–277.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, **90**, 577–588.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

George, E. I. (1986) A formal Bayes multiple shrinkage estimator. *Communs Statist. Theory Meth.*, **15**, 2099–2114.

Gopalan, R. and Berry, D. A. (1998) Bayesian multiple comparisons using Dirichlet process priors. *J. Am. Statist. Ass.*, **93**, 1130–1139.

Hartigan, J. A. (1983) *Bayes Theory*. New York: Springer.

MacEachern, S. N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communs Statist. Simuln Computn*, **23**, 727–741.
MacEachern, S. N. (1998) Computational methods for mixture of Dirichlet process models. in *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds D. Dey, P. Müller and D. Sinha), pp. 23–44. New York: Springer.
MacEachern, S. N. and Müller, P. (2000) Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis* (eds D. Rios Insua and F. Ruggeri), pp. 295–315. New York: Springer.
Mallick, B. K. and Walker, S. G. (1997) Combining information from several experiments with nonparametric priors. *Biometrika*, **84**, 697–706.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. New York: Chapman and Hall.
Müller, P. and Quintana, F. A. (2004) Nonparametric Bayesian data analysis. *Technical Report*. Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston.
Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related nonparametric Bayesian models. *J. R. Statist. Soc.* B, **66**, 735–749.
Newton, M. A., Czado, C. and Chappell, R. (1996) Bayesian inference for semiparametric binary regression. *J. Am. Statist. Ass.*, **91**, 142–153.
O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.
Quintana, F. A. and Iglesias, P. L. (2003) Bayesian clustering and product partition models. *J. R. Statist. Soc.* B, **65**, 557–574.
Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.
Sethuraman, J. and Tiwari, R. C. (1982) Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics III*, vol. 2, pp. 305–315. New York: Academic Press.
Tomlinson, G. and Escobar, M. (1999) Analysis of densities. *Technical Report*. University of Toronto, Toronto.
Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion). *J. R. Statist. Soc.* B, **61**, 485–527.