Journal of Statistical Software

July 2012, Volume 50, Issue 4. *http://www.jstatsoft.org/*

bspmma: An R Package for Bayesian Semiparametric Models for Meta-Analysis

Deborah Burr
Gainesville, USA

Abstract

We introduce an R package, **bspmma**, which implements a Dirichlet-based random effects model specific to meta-analysis. In meta-analysis, when combining effect estimates from several heterogeneous studies, it is common to use a random-effects model. The usual frequentist or Bayesian models specify a normal distribution for the true effects. However, in many situations, the effect distribution is not normal, e.g., it can have thick tails, be skewed, or be multi-modal. A Bayesian nonparametric model based on mixtures of Dirichlet process priors has been proposed in the literature, for the purpose of accommodating the non-normality. We review this model and then describe a competitor, a semiparametric version which has the feature that it allows for a well-defined centrality parameter convenient for determining whether the overall effect is significant. This second Bayesian model is based on a different version of the Dirichlet process prior, and we call it the "conditional Dirichlet model." The package contains functions to carry out analyses based on either the ordinary or the conditional Dirichlet model, functions for calculating certain Bayes factors that provide a check on the appropriateness of the conditional Dirichlet model, and functions that enable an empirical Bayes selection of the precision parameter of the Dirichlet process. We illustrate the use of the package on two examples, and give an interpretation of the results in these two different scenarios.

Keywords: Dirichlet process, conditional Dirichlet process, meta-analysis, random effects, R package, Markov chain Monte Carlo.

1. Introduction: Bayesian random effects meta-analysis

1.1. Parametric models

In many medical studies, each of m hospitals or centers investigates the same medical issue, which we will think of as being a comparison between a new and an old treatment. Sometimes the results from the m studies are inconsistent: Some studies are favorable to the new treatment, while others are neutral or negative. The goals of a meta-analysis are then to arrive at an overall conclusion regarding the benefits of the new treatment, and also to describe and explain the heterogeneity among the different studies.

Suppose that for each i, center i gathers a summary statistic D_i together with a standard error estimate $\hat{\sigma}_i$. When heterogeneity among the studies is significant, it is now common to carry out meta-analyses that are based on random effects models, in which for each center i there is a center-specific "true effect," represented by a parameter ψ_i, and D_i has distribution $P_i(\psi_i)$. This distribution depends on ψ_i and also on other quantities, such as the sample size as well as nuisance parameters specific to the i-th center. In the most common example in epidemiological studies, ψ_i is the log of the odds ratio arising in case-control studies, and D_i is either an adjusted log odds ratio based on a logistic regression model that involves relevant covariates, or simply the usual log odds ratio based on a 2×2 table. The traditional model for dealing with this kind of situation is the following:

$$\text{conditional on } \psi_i, \quad D_i \sim \mathcal{N}(\psi_i, \sigma_i^2), \quad \text{independently, } i = 1, \dots, m, \tag{1a}$$

$$\psi_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), \quad i = 1, \dots, m. \tag{1b}$$

In (1b), μ and τ are unknown parameters. (The σ_i's are also unknown but, typically, estimation of these is of secondary interest, and sample sizes are sufficiently large so that using the $\hat{\sigma}_i$'s instead of the σ_i's does not cause any problems.) In a frequentist analysis μ and τ are estimated by maximum likelihood (DerSimonian and Laird 1986) and in a Bayesian analysis μ and τ^2 are given a prior distribution, typically a normal/inverse gamma prior, because this conjugate form results in simplifications when estimating the posterior distribution.

1.2. Bayesian nonparametric and semiparametric models

The approximation of $P_i(\psi_i)$ by a normal distribution in (1a) is typically supported by some theoretical result, for example the asymptotic normality of the observed log odds ratio or, more generally, the asymptotic normality of maximum likelihood estimates. By contrast, the normality statement in (1b) is a modelling assumption, which generally is made for the sake of convenience and does not have any theoretical justification. A number of authors have encountered meta-analyses in which the distribution of the study effects appears to be non-normal—for example because some studies appear to be outliers—and have suggested that in (1b) the normal distribution be replaced by a distribution that accommodates outliers. Sharples (1990) discusses normal contamination models in a classical Bayesian one-way random effects model. Weiss, Cho, and Yanuzzi (1999) develop a Markov chain Monte Carlo (MCMC) approach for fitting models in which either the likelihood or the prior is a mixture of two normals. Dozens of authors have considered t distributions. While t distributions accommodate outliers better than do normals, the distribution of the random effects may deviate from normality in ways that do not involve heaviness of tails. In fact, in one of the examples in the present paper, the distribution of the study effects appears to be multimodal.

This leads us to consider a model of the form

$$\text{conditional on } \psi_i, \quad D_i \sim \mathcal{N}(\psi_i, \sigma_i^2), \quad \text{independently, } i = 1, \dots, m, \tag{2a}$$

$$\text{conditional on } F, \quad \psi_i \overset{\text{iid}}{\sim} F, \quad i = 1, \dots, m, \tag{2b}$$

$$F \sim \pi, \tag{2c}$$

where π is a nonparametric prior on the family of all distribution functions. A natural approach is to take π to be a Dirichlet process (see Ferguson 1973, 1974, for a formal treatment, or the beginning of Section 2 of the present paper for an informal review) and indeed this is done by Mallick and Walker (1997). They take π to be $\mathcal{D}_\alpha$, the Dirichlet process with parameter measure $\alpha = M \cdot H$, where H is the centering distribution and M is the precision parameter, and they give a method for estimating posterior distributions based on this prior. A key drawback of this approach is that the user must specify the centering distribution H exactly. A more flexible approach involves using a model based on mixtures of Dirichlet processes as introduced by Antoniak (1974). In the context of the meta-analysis model, Model (2), one can take π to be a mixture of Dirichlet process priors with parameter consisting of the triple $(\{H_\vartheta\}_{\vartheta\in\Omega}, M, \lambda)$, where $\{H_\vartheta\}_{\vartheta\in\Omega}$ is a parametric family of distributions, M is a precision parameter, and λ is a prior on Ω. An important particular case, which we discuss further below, is to take $\{H_\vartheta\}_{\vartheta\in\Omega}$ to be the two-parameter family of normal distributions $\mathcal{N}(\mu, \tau^2)$, and take λ to be the normal/inverse gamma prior on $\vartheta = (\mu, \tau)$. Thus, (2c) becomes

$$\begin{aligned} \text{conditional on } \mu, \tau, \qquad F &\sim \mathcal{D}_{M\mathcal{N}(\mu,\tau^2)}, \\ (\mu, \tau) &\sim \lambda. \end{aligned} \tag{3}$$

Consider now Model (1). For this model, the parameter μ has a clear interpretation as the mean or median of the $\mathcal{N}(\mu, \tau^2)$ distribution (and if we replace the normal with a t then μ also has a clear interpretation as the median of the distribution). In contrast, in Model (3) the parameter μ is *not* equal to $\int x\,dF(x)$, the mean of F. (If F is chosen from a Dirichlet process with centering distribution H, then even if H is symmetric about μ, the probability that F is symmetric about μ is 0.) In fact, μ does not play the role of *any* location parameter for F. This is a problem in certain situations where there is no overwhelming evidence that the treatment is better than the control, and one is interested primarily not in estimation of the center-specific ψ_i's, but rather in resolving the basic question of whether the overall mean μ is different from 0, as this may for example justify carrying out further studies. For this reason, Burr and Doss (2005) propose using a "mixture of conditional Dirichlet processes" as the prior π in (2c). Loosely speaking, if $F \sim \mathcal{D}_{M\mathcal{N}(\mu,\tau^2)}$, then the distribution of F conditional on the event that the median of F is μ is called a *conditional Dirichlet*, and we will denote it by $\mathcal{D}^{\mu}_{M\mathcal{N}(\mu,\tau^2)}$. By construction, if $F \sim \mathcal{D}^{\mu}_{M\mathcal{N}(\mu,\tau^2)}$, then with probability one μ is the median of F, so we have a parameter μ with a clear interpretation, and we can proceed to carry out inference about this parameter.

To summarize, Burr and Doss (2005) consider the hierarchical model

$$\text{conditional on } \psi_i, \qquad D_i \sim \mathcal{N}(\psi_i, \sigma_i^2), \quad \text{independently, } i = 1, \dots, m, \tag{4a}$$

$$\text{conditional on } F, \qquad \psi_i \overset{\text{iid}}{\sim} F, \quad i = 1, \dots, m, \tag{4b}$$

$$\text{conditional on } \mu, \tau, \qquad F \sim \mathcal{D}^{\mu}_{M\mathcal{N}(\mu,\tau^2)}, \tag{4c}$$

$$\text{conditional on } \tau, \qquad \mu \sim \mathcal{N}(d_3, d_4\tau^2), \tag{4d}$$

$$\gamma = 1/\tau^2 \sim \text{Gamma}(d_1, d_2). \tag{4e}$$

In (4d) and (4e), $d_1, d_2, d_4 > 0$, $d_3 \in \mathbb{R}$, and the resulting prior on (μ, τ) is the normal/inverse gamma prior, which we will denote λ_d. For the purpose of giving a dispersed prior on μ and τ, they take $d_1 = 0.1$, $d_2 = 0.1$, $d_3 = 0$, and $d_4 = 1000$. (For this choice, the marginal distribution of μ is a t distribution with 0.2 degrees of freedom, median 0, and scale parameter $1000^{1/2}$, which is a fairly diffuse prior.) Burr and Doss (2005) develop a Markov chain Monte Carlo algorithm for estimating the posterior distribution of the vector $(\psi_1, \ldots, \psi_m, \mu, \tau)$ in this model. The purpose of this paper is to present and describe the R package **bspmma**, which implements the algorithms developed in Burr and Doss (2005), and to illustrate the use of these algorithms. The package **bspmma** is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=bspmma`.

The rest of the paper is organized as follows. In Section 2, we give a review of Dirichlet and conditional Dirichlet processes and their mixtures, and give motivation for their use. Then we describe a method for determining whether the model based on mixtures of conditional Dirichlet processes fits the data as well or better than does a model based on ordinary mixtures of Dirichlet processes. A by-product of our model validation method is a procedure for carrying out an empirical Bayes selection of the Dirichlet precision parameter M. In Section 3, we show how to use the package on two examples.

Several other packages are available for conducting meta-analyses in the R language (R Development Core Team 2012). We mention **rmeta** (Lumley 2009), **meta** (Schwarzer 2010), and **metafor** (Viechtbauer 2010), which enable frequentist analyses, and assume that the effect distribution is normal; also, the **metafor** package allows meta-regression. These packages have extensive functions for plotting and producing summary statistics.

The **DPpackage** (Jara, Hanson, Quintana, Müller, and Rosner 2011; Jara 2007) is a very broad package for implementing Bayesian semiparametric models. It fits certain hierarchical models of the sort (2) where π is a mixture of ordinary Dirichlet processes, and also handles models from a variety of areas such as density estimation, survival analysis, and generalized linear mixed models. Branscum and Hanson (2008) also consider Bayesian semiparametric models for meta-analysis, and their prior on the effect distribution is a (finite) Polya tree, rather than a Dirichlet-based process. More specifically they consider a model similar to (4), except that in line (4c), the conditional Dirichlet process is replaced by a "median-constrained" Polya tree. The **DPpackage** function `PTmeta` implements the model in Branscum and Hanson (2008). In Section 4 we further discuss the Polya tree prior used in Branscum and Hanson (2008) and discuss the relationship between the model they use and the model used here. The **bspmma** package is not a general-purpose package for fitting Bayesian nonparametric or semiparametric models. It deals only with random-effects meta-analysis, and focuses specifically on inference for particular parameters of interest in meta-analysis, and on hyperparameter selection and model assessment.

2. Ordinary and conditional Dirichlet processes

2.1. Background

Let $\mathcal{P}$ be the set of all probability distributions on the real line. A parametric family H_ϑ, $\vartheta \in \Omega \subset \mathbb{R}^p$ may be viewed as a finite-dimensional subset of the infinite-dimensional space $\mathcal{P}$, and a prior on Ω induces a prior on $\mathcal{P}$ which gives all its mass to the finite-dimensional family H_ϑ, $\vartheta \in \Omega$. When there is no reason to think that any particular parametric model is exactly true, a desirable feature of a prior on $\mathcal{P}$ is that it be "nonparametric," i.e., does not give all its mass to any finite-dimensional subset of $\mathcal{P}$. (More formally, having specified a topology on $\mathcal{P}$, we would like the prior to have the "full support property," i.e., it gives positive probability to any open set.)

The most commonly used nonparametric priors are mixtures of Dirichlet processes. Standard definitions are Ferguson (1973, 1974), and Antoniak (1974), but here we give a brief intuitive review of those properties of this class of priors that are directly relevant to the present situation. Before describing mixtures of Dirichlet processes, we first discuss (single) Dirichlet processes. A Dirichlet process is a distribution on the space $\mathcal{P}$, and is parameterized by a pair (H, M), where H is a distribution function and M is a positive number. The product $\alpha = MH$ is a finite measure, and the Dirichlet process is denoted by $\mathcal{D}_\alpha$ (the measure α determines and is determined by the pair (H, M)). For the purpose of understanding the modelling assumptions when we use a Dirichlet process in the present context, the best way to explain this prior is through the construction of Sethuraman (1994). Generate $B_1, B_2, \ldots \overset{\text{iid}}{\sim} \text{Beta}(1, M)$, and independently generate $V_1, V_2, \ldots \overset{\text{iid}}{\sim} H$. Let $P_j = B_j \prod_{r=1}^{j-1}(1-B_r)$, and form the random distribution

$$F = \sum_{j=1}^{\infty} P_j \delta_{V_j}, \tag{5}$$

where δ_a denotes the probability measure giving unit mass to the point a. Sethuraman (1994) showed that F defined by (5) is distributed according to the Dirichlet process with parameter measure α, as defined in Ferguson (1973).

Key properties of the Dirichlet process are: (i) The "center" is H in the sense that for every t, we have $E(F(t)) = H(t)$; (ii) M is a precision parameter which determines the concentration of the prior around H: for large M, the distribution of $F(t)$ is tightly concentrated around $H(t)$ for every t, while for small values of M, the distribution is more diffuse; and (iii) the Dirichlet process is a nonparametric prior in the following sense: if the support of H is the entire real line, then the Dirichlet process has the full support property referred to earlier.

Note that if we first choose F from $\mathcal{D}_\alpha$ and then generate $\psi_1, \ldots, \psi_n \overset{\text{iid}}{\sim} F$, then since F is discrete, with positive probability there will be ties among the ψ_i's; that is, the ψ_i's will form clumps. When M is small, the first few P_j's add up to nearly 1, and therefore the probability of ties is higher. This leads to important consequences regarding the posterior distribution of $\psi_1, \ldots, \psi_n$ given the data $D_1, \ldots, D_n$. Consider the distribution of ψ_i. As is standard in hierarchical models, conditioning on the data and ψ_j, $j \neq i$ results in shrinkage towards D_i and toward a grand mean. But because of the propensity for clumping, the posterior is also shrunk towards those ψ_j's that are close to D_i (the extent of the shrinking is determined in part by the standard error σ_i). This last results in a way of pooling information that involves weighing results of similar studies more heavily.

As mentioned in Section 1, a mixture of Dirichlet processes involves a parametric family H_ϑ, $\vartheta \in \Omega \subset \mathbb{R}^p$, a precision parameter $M > 0$, and a distribution λ on Ω. Formally, it is the integral $\int \mathcal{D}_{MH_\vartheta} \lambda(d\vartheta)$. We think of it as arising in a two-stage process, where in the first stage we pick the parameter ϑ according to λ, and in the second stage we pick F from the Dirichlet distribution with parameter MH_ϑ. A mixture of conditional Dirichlet processes is defined in the obvious way, i.e., the Dirichlet process in the integrand is replaced by a conditional Dirichlet process. In the situation considered in this paper, $\{H_\vartheta\}_{\vartheta\in\Omega}$ is the two-parameter family of normal distributions. The role of the precision parameter may be described intuitively as follows. The parametric family $\{H_\vartheta\}_{\vartheta\in\Omega}$ is a (p-dimensional) "line" in the infinite-dimensional space of probability measures on the real line, and we imagine a "tube" around this line. Large values of M correspond to a narrow tube (which we interpret as an expression of confidence that the normal model holds), and small values of M correspond to a wider tube. (Sethuraman and Tiwari 1982 give a careful description of the proper interpretation of the Dirichlet process when M is very small.) It should be mentioned that when M is large, for any μ and τ the distributions $\mathcal{D}_{M\mathcal{N}(\mu,\tau^2)}$ and $\mathcal{D}^{\mu}_{M\mathcal{N}(\mu,\tau^2)}$ are close, since they are both nearly equal to the point mass at the $\mathcal{N}(\mu,\tau^2)$ distribution and, consequently, for large M the λ-mixture of Dirichlet processes and λ-mixture of conditional Dirichlet processes are close. A measure-theoretic discussion of this point is given in Doss (1985).

2.2. Model assessment and hyperparameter selection

In Section 1 we discussed conditional Dirichlet processes as a tool that enables us to make inference about a location parameter of the distribution of the latent effects. While there are inferential reasons for using a model based on these priors, it is nevertheless natural to ask if such a model provides a good fit for the data. More precisely, it is natural to ask if a model based on mixtures of conditional Dirichlet processes provides a better fit than does a model based on ordinary mixtures of Dirichlet processes.

In Model (4) the unknown parameter is effectively $\theta = (\psi, \mu, \tau)$, where $\psi = (\psi_1, \dots, \psi_m)$ is the vector of latent variables (F plays a role only in the sense that it induces a distribution on the latent variables). Different prior distributions on θ induce different models for the data vector D. Generally speaking, when deciding between two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for the data D, it is standard to consider the marginal likelihoods $m_{\mathcal{M}_1}(D)$ and $m_{\mathcal{M}_2}(D)$ which, in the framework of the present paper, are defined by

$$m_{\mathcal{M}_i}(D) = \int \ell_D(\theta)\, d\nu_i(\theta) \qquad i = 1, 2. \tag{6}$$

In (6), $\ell_D(\theta)$ is the likelihood function, and ν_i is the prior distribution of θ under model $\mathcal{M}_i$. (The marginal likelihood is just the likelihood of the data with the unknown parameter integrated out.) If these marginal likelihoods can be computed, it is common practice to select the model for which the marginal likelihood is greater. In our situation, we fix some value of the precision parameter M, some value $d = (d_1, d_2, d_3, d_4)$ of the hyperparameter vector of the normal/inverse gamma prior on (μ, τ), and $\mathcal{M}_1$ will be the mixture of conditional Dirichlet processes and $\mathcal{M}_2$ the mixture of ordinary Dirichlet processes, each based on the hyperparameter specification $h = (M, d)$. It will be more convenient to denote these models by $\mathcal{M}_h^c$ and $\mathcal{M}_h^o$, respectively. Also, ν_h^c and ν_h^o will denote the (prior) distribution of θ under $\mathcal{M}_h^c$ and $\mathcal{M}_h^o$, respectively, and m_h^c and m_h^o will denote the marginal likelihoods of the data under $\mathcal{M}_h^c$ and $\mathcal{M}_h^o$, respectively. Now selecting the model for which the marginal likelihood is greater is of course equivalent to selecting $\mathcal{M}_h^c$ if and only if the ratio $B(\mathcal{M}_h^c, \mathcal{M}_h^o) := m_h^c / m_h^o$ is greater than 1. The ratio $B(\mathcal{M}_h^c, \mathcal{M}_h^o)$ is commonly referred to as the Bayes factor of $\mathcal{M}_h^c$ relative to $\mathcal{M}_h^o$.

Unfortunately, it is impossible to calculate the two marginal likelihoods m_h^c and m_h^o exactly, and very difficult to estimate them accurately. However, as we now show, it is possible to estimate their ratio, i.e., the Bayes factor. In fact, we will show that it is possible to estimate m_h^c / m_h^o for *all* h from a single Markov chain, run under model $\mathcal{M}_{h_1}^o$, where h_1 is some prespecified value of the hyperparameter $h_1 = (M_1, d_1)$. Let us write

$$\frac{m_h^c}{m_h^o} = \frac{m_h^c}{m_{h_1}^o} \left(\frac{m_h^o}{m_{h_1}^o} \right)^{-1}. \tag{7}$$

We will estimate m_h^c/m_h^o by estimating $m_h^c/m_{h_1}^o$ and $m_h^o/m_{h_1}^o$, and taking the ratio.

Let $\nu_{h,D}^c$ and $\nu_{h,D}^o$ denote the posterior distributions of θ when the priors are ν_h^c and ν_h^o, respectively. We note that the marginal likelihood is simply the normalizing constant in the statement "the posterior is proportional to the likelihood times the prior." Proceeding as in Doss (2012), we write

$$\nu_{h,D}^c(d\theta) = \ell_D(\theta)\nu_h^c(d\theta)/m_h^c \qquad \text{and} \qquad \nu_{h_1,D}^o(d\theta) = \ell_D(\theta)\nu_{h_1}^o(d\theta)/m_{h_1}^o,$$

and using the fact that

$$\int \left[\frac{d\nu_{h,D}^c}{d\nu_{h_1,D}^o} \right] (\theta)\, \nu_{h_1,D}^o(d\theta) = 1, \tag{8}$$

we obtain the identity

$$\int \left[\frac{d\nu_h^c}{d\nu_{h_1}^o} \right] (\theta)\, \nu_{h_1,D}^o(d\theta) = \frac{m_h^c}{m_{h_1}^o}. \tag{9}$$

Note: In (8) and (9), we have used the formalism of Radon-Nikodym derivatives instead of writing the ratio of densities, and we now digress briefly to explain this. When we have the parametric model corresponding to the case $M = \infty$, i.e., model (4) except that lines (4b) and (4c) are replaced by the single line $\psi_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2)$, the distribution of θ has a density on $\mathbb{R}^{m+2}$ (with respect to Lebesgue measure), and (8) is just the trivial formula

$$\int \frac{\nu_{h,D}^c(\theta)}{\nu_{h_1,D}^o(\theta)} \nu_{h_1,D}^o(\theta)\, d\theta = 1,$$

which involves only densities. When $M < \infty$, the distributions ν_h^c, $\nu_{h,D}^c$, $\nu_{h_1}^o$, and $\nu_{h_1,D}^o$ do not have densities with respect to Lebesgue measure on $\mathbb{R}^{m+2}$, since there is positive probability that the ψ_j's have ties, and in (8) $[d\nu_{h,D}^c/d\nu_{h_1,D}^o]$ is the "Radon-Nikodym derivative of $\nu_{h,D}^c$ with respect to $\nu_{h_1,D}^o$." Loosely speaking, for $\theta \in \mathbb{R}^{m+2}$, the Radon-Nikodym derivative is given by

$$\left[\frac{d\nu_{h,D}^c}{d\nu_{h_1,D}^o} \right] (\theta) = \lim_{\epsilon \to 0} \frac{\nu_{h,D}^c(B_\theta^\epsilon)}{\nu_{h_1,D}^o(B_\theta^\epsilon)},$$

where B_θ^ϵ is the ball in $\mathbb{R}^{m+2}$ centered at θ and with radius ϵ, and this is the measure-theoretic equivalent of the "ratio of densities." A more detailed discussion is given in the Appendix of Burr and Doss (2005).

From (9) we see that if $\theta_1, \theta_2, \ldots$ is an ergodic Markov chain with stationary distribution $\nu_{h_1,D}^o$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{d\nu_h^c}{d\nu_{h_1}^o} \right] (\theta_i) \xrightarrow{\text{a.s.}} \frac{m_h^c}{m_{h_1}^o}, \tag{10}$$

where "$\xrightarrow{\text{a.s.}}$" means "converges almost surely." An explicit formula for $[d\nu_h^c/d\nu_h^o]$ was obtained in Burr and Doss (2005) and was extended in Doss (2012) to the case $[d\nu_h^c/d\nu_{h_1}^o]$, where h and h_1 are not necessarily equal.

In a similar way, we have the identity

$$\int \left[\frac{d\nu_h^o}{d\nu_{h_1}^o}\right](\theta)\, \nu_{h_1,D}^o(d\theta) = \frac{m_h^o}{m_{h_1}^o},$$

and for the same Markov chain $\theta_1, \theta_2, \ldots$ with stationary distribution $\nu_{h_1,D}^o$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d\nu_h^o}{d\nu_{h_1}^o}\right](\theta_i) \xrightarrow{\text{a.s.}} \frac{m_h^o}{m_{h_1}^o}. \tag{11}$$

An explicit formula for $[d\nu_h^o/d\nu_{h_1}^o]$ is given in Theorem 1 of Doss (2012). Combining (7), (10), and (11), we obtain that

$$\frac{\sum_{i=1}^{n}[d\nu_h^c/d\nu_{h_1}^o](\theta_i)}{\sum_{i=1}^{n}[d\nu_h^o/d\nu_{h_1}^o](\theta_i)} \xrightarrow{\text{a.s.}} \frac{m_h^c}{m_h^o}. \tag{12}$$

In the present situation, we will keep the prior λ on (μ, τ) fixed and vary M; specifically, we will fix d at $(0.1, 0.1, 0, 1000)$, and we will be interested in estimating the ratio $m_{M,d}^c/m_{M,d}^o$ which, by a slight abuse of notation, we will denote by m_M^c/m_M^o. In the illustration in Section 3.3, we give a plot of the estimate of this ratio. To summarize, the discussion above provides a method for estimating this ratio simultaneously for all M, which gives a tool for comparing the conditional Dirichlet model with the ordinary Dirichlet model.

It is also natural to ask how does one select a value for M. We consider first the unconditional Dirichlet model. As before, d is fixed at $(0.1, 0.1, 0, 1000)$ and so is not in the picture, and by slight abuse of notation we will write m_M^o instead of m_h^o, m_M^c instead of m_h^c, etc. As noted in Doss (2012), if M_1 is fixed at an arbitrary value, then maximizing $m_M^o/m_{M_1}^o$ as a function of M is equivalent to maximizing m_M^o as a function of M, and the maximizing value of M is by definition the empirical Bayes choice of M for the unconditional Dirichlet model. Therefore a plot of the estimate of $m_M^o/m_{M_1}^o$ given by the left side of (11) enables us to obtain an empirical Bayes choice of the precision parameter M for the ordinary Dirichlet model. We can also form an empirical Bayes estimate of M for the conditional Dirichlet model—again based on a run from an ergodic Markov chain with stationary distribution $\nu_{M_1,D}^o$—by proceeding as follows. We note that (i) maximizing $m_M^c/m_{M_1}^o$ with respect to M is equivalent to maximizing m_M^c, (ii) the common maximizing value is the empirical Bayes choice of M for the conditional Dirichlet model, and (iii) an estimate of the function $m_M^c/m_{M_1}^o$ from a Markov chain with stationary distribution $\nu_{M_1,D}^o$ is provided by the left side of (10). Section 3 provides an illustration of obtaining the empirical Bayes choice of the precision parameter M. To summarize, the discussion above provides a method for carrying out an empirical Bayes selection of the precision parameter M, whether we are using the unconditional or conditional Dirichlet model. In Section 3 we explain why estimates based on a single Markov chain can be unstable when the range of M is large, and discuss improvements which involve using multiple chains.

We mention briefly that under moment conditions on $[d\nu_h^c/d\nu_{h_1}^o]$ and $[d\nu_h^o/d\nu_{h_1}^o]$ and mixing conditions on the chain, the pair

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d\nu_h^c}{d\nu_{h_1}^o}\right](\theta_i), \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d\nu_h^o}{d\nu_{h_1}^o}\right](\theta_i)\right)$$

satisfies a bivariate central limit theorem. Therefore, the averages on the left side of (10) and (11) satisfy a central limit theorem, and by the delta method applied to the function

$g(u, v) = u/v$, we also have a central limit theorem for the ratio on the left side of (12). Consequently we are able to form standard errors for all the estimates discussed above.

Markov chain Monte Carlo Details regarding an MCMC algorithm for estimating the posterior under Model (4) (mixture of conditional Dirichlets) are given in Burr and Doss (2005). It is important to note that the chain is a $(m+1)$-cycle Gibbs sampler which cycles through the vector of ψ_i's and the pair (μ, τ), and that the main part of the computational burden is in the first part of the cycle, the generation of the vector of ψ_i's.

3. Usage

In this section we illustrate the functions available in **bspmma** through two examples. The main function, `dirichlet.c`, carries out the MCMC algorithm to simulate data from the posterior distribution under the conditional Dirichlet model, Model (4). A corresponding function, `dirichlet.o`, simulates output when the prior on F is the mixture of ordinary Dirichlets given by Model (3). For comparison of posterior distributions for several different values of the Dirichlet precision parameter M, the package provides the functions `draw.post` (for overlaid graphs of the posteriors) and `describe.post` (for side-by-side summary statistics). For determining whether the conditional or ordinary Dirichlet model is preferred, the functions `bf1`, `bf2`, `bf.c.o`, and `draw.bf` do the computations and produce the plot of Bayes factors for the conditional Dirichlet model, Model (4) vs. the ordinary Dirichlet model, Model (3). Finally, the functions `bf.c` and `bf.o` can be used to compute Bayes factors appropriate for selecting the maximizing value of M for the conditional or ordinary Dirichlet model, respectively.

Output from the main **bspmma** functions can be analyzed further using routines in the R packages **boa** (Smith 2007) and **coda** (Plummer, Best, Cowles, and Vines 2006), which have functions for diagnosis of convergence, and for providing graphical and statistical summaries, of MCMC output. The two packages implement a similar set of published methods and can be applied to output from the **bspmma** functions `dirichlet.c` and `dirichlet.o`. The **boa** package has a menu-driven interface, and due to our focus on use of source code, we chose to illustrate application of the **coda** package here.

3.1. Example 1: Effect of NSAIDs on breast cancer

The hypothesis that use of non-steroidal anti-inflammatory drugs (NSAIDs) reduces the risk of breast cancer is currently of considerable interest and controversy in the medical literature (Harris *et al.* 2003; Terry *et al.* 2004; Zhang, Coogan, Palmer, Strom, and Rosenberg 2005). Because NSAIDs must be taken regularly for many years to have this beneficial effect, it is not possible to carry out randomized, controlled experiments in healthy populations. There have been many studies on the effect of NSAIDs on breast cancer (and other cancers) during the last 15 years; the studies that have been done are either at the epidemiological or at the cellular and molecular level, and several have strongly suggested that long-term use of NSAIDs significantly decreases the risk of breast cancer. However, this result is not seen in all studies; some suggest only a slight risk reduction and others in fact suggest no risk reduction at all. Harris, Beebe-Donk, Doss, and Burr (2005) give a review of this work and discuss the epidemiological studies that have appeared in the medical literature. Each study reports a risk ratio for NSAIDs use vs. no NSAIDs use. This risk ratio is either simply an odds ratio obtained from a case-control study or an odds ratio based on a multiple logistic regression analysis that takes into account important risk factors for breast cancer.

It is not surprising that the studies give inconsistent results, since there is heterogeneity in the subject pools (characteristics such as age, ethnicity, and health status vary across the studies), and in the way the data were obtained (covariates collected, statistical method used, etc.). It is certainly of interest to carry out a meta-analysis of these studies, and because of the heterogeneity, it seems clear that the meta-analysis should be based on a random effects model. There have been some meta-analyses in the epidemiological literature. Khuder and Mutgi (2001) find fifteen studies and point out that the studies have heterogeneous effect estimates; they then form several sub-groups of the data, and carry out fixed-effects analyses of homogeneous sub-groups and random-effects analyses of heterogeneous groups. Gonzalez-Perez, Rodriguez, and Lopez-Ridaura (2003) use the classical random-effects meta-analysis for each of six cancers including breast cancer. For ten different cancers including breast cancer, Harris *et al.* (2005) do fixed-effects meta-regression with one study-level covariate—dose of NSAID—using just the subset of studies for which dose information is available. All the random-effects meta-analyses we have seen assume normality of the distribution of effects, but without justification.

Columns 2–3 of Table 1 give the data from Harris *et al.* (2005) for each of the 17 studies that pertain to the particular NSAID aspirin. These columns give the reported risk ratio and a confidence interval for the risk ratio. Although these authors consider dose as well, and therefore consider only the 14 out of the 17 studies which contain dose information, we ignore dose in this analysis. It would be of interest to also carry out another analysis that includes dose.

For each study j, let L_j and σ_j denote the observed log risk ratio and its standard error, respectively, for that study. Let ψ_j denote the true log risk ratio, i.e., ψ_j is the log risk ratio that would be obtained if the sample size for study j were infinite. Standard asymptotic theory justifies writing $L_j \sim \mathcal{N}(\psi_j, \sigma_j^2)$. Therefore, if we let F represent the distribution of unknown shape of the ψ_j's, we are led to precisely Model (2), with $D_j = L_j$ having standard deviation σ_j. Column 4 of Table 1 gives L_j (the observed log risk ratio) and column 5 gives σ_j (the standard error of L_j). Harris *et al.* (2005) do not give the σ_j's, but these can be obtained from the confidence intervals for the risk ratios, which are given in column 3.

The package includes the dataframe `breast.17`, which gives the log risk ratios and their standard errors (columns 4 and 5 from Table 1). To prepare to run the main functions in the package, we first set up the data as a matrix with two columns, where the first column contains the log risk ratios and the second column contains the standard errors, as follows:

```
R> library("bspmma")
R> data("breast.17")
R> breast.data <- as.matrix(breast.17)
```

The next two commands show how to use the conditional Dirichlet MCMC function, first setting the precision parameter value to be $M = 5$ and then $M = 1000$. The seed is set (to 1) so that the example can be reproduced exactly. The algorithm to run MCMC for the conditional Dirichlet model, Model (4) completes 4000 cycles in about 14 seconds (as timed by R function `system.time`), on an Intel 2.8 GHz Q9550 running Linux.

```
R> set.seed(1)
```

j	RR	CI	L_j	σ_j
1	0.96	$(0.70, 1.50)$	-0.04	0.161
2	0.88	$(0.62, 1.24)$	-0.13	0.179
3	0.70	$(0.50, 0.96)$	-0.36	0.172
4	1.01	$(0.80, 1.27)$	0.01	0.119
5	0.64	$(0.45, 0.90)$	-0.45	0.180
6	0.71	$(0.58, 0.87)$	-0.34	0.103
7	0.79	$(0.66, 1.04)$	-0.24	0.092
8	0.60	$(0.37, 0.96)$	-0.51	0.247
9	0.80	$(0.60, 1.00)$	-0.22	0.147
10	0.69	$(0.46, 0.99)$	-0.37	0.207
11	0.80	$(0.35, 1.80)$	-0.22	0.422
12	0.70	$(0.50, 0.80)$	-0.36	0.172
13	0.76	$(0.63, 0.92)$	-0.27	0.096
14	0.73	$(0.61, 0.87)$	-0.31	0.092
15	1.10	$(0.92, 1.30)$	0.10	0.091
16	1.00	$(0.80, 1.10)$	0.00	0.114
17	0.40	$(0.30, 0.60)$	-0.92	0.147

Table 1: Summary data from 17 studies on aspirin and breast cancer: RR is the risk ratio for aspirin vs. no aspirin; CI is the associated confidence interval; L_j is the observed log risk ratio; and σ_j is the estimated standard error of L_j.

```
R> breast.c1 <- dirichlet.c(breast.data, ncycles = 4000, M = 5)
R> set.seed(1)
R> breast.c2 <- dirichlet.c(breast.data, ncycles = 4000, M = 1000)
```

Another argument to the function is the hyperparameter vector d of the normal/inverse gamma prior on (μ, τ), which in the above runs is taken by default to be $d = (0.1, 0.1, 0, 1000)$; there is also an argument `start`, which allows the user to supply starting values for the parameters. By default, the starting value for ψ_i is the individual study estimate D_i, for $i = 1, \ldots, m$; and the starting values for μ and τ are the mean and standard deviation of the D_i's, respectively.

The output of `dirichlet.c` is a list with several components; the main component, called `chain`, is a matrix such that each row contains the output of a single iteration of the Monte Carlo simulation. In the above runs the matrix is of dimension 4001 × 19. The first row contains initial parameter values; each of the remaining 4000 rows contains a set of parameter values for one iteration of the chain. The columns contain the simulated values of the parameters of Model (4). Columns 1–17 contain the individual ψ_i's, that is, the log risk ratios for studies 1 through 17. Column 18 contains μ, the median of the distribution of the ψ_i's (and the mean of the centering normal distribution of the conditional Dirichlet model, Model (4)), and column 19 contains τ (the standard deviation of the centering normal distribution of the conditional Dirichlet model).

For each of the runs above taken separately, the MCMC output can be checked for convergence by any of the diagnostic functions in **coda**. This requires using the `chain` component of the **bspmma** output as an argument to the desired **coda** function, after first converting it to an object of class `mcmc` in **coda**. For the first chain above, with $M = 5$, we use **coda** to get the autocorrelation plot for the last three ψ_i's and for μ and τ. The reason for examining the autocorrelation plots for these particular ψ_i's is that they include the two most unusual studies. The study by "Langman" (in column 15) had the only positive estimate of the log odds ratio, and the study by "Moorman" (in column 17) had the estimate farthest from the mean, the most extreme negative estimate.

```
R> library("coda")
R> breast.coda <- mcmc(breast.c1$chain)
R> autocorr.plot(breast.coda[, 15:19])
```

In the resulting plots (not shown), the autocorrelations are 0 (for practical purposes) for all lags greater than three; thus this particular diagnostic does not provide evidence of a problem with the chain. There are many other functions in **coda**, for posterior inference as well as for convergence diagnosis, which can be applied in a similar manner.

In addition, there are two user-accessible functions in **bspmma** which are useful for exploration of models corresponding to different values of the Dirichlet precision parameter M. The function `draw.post` may be used to produce superimposed graphs for several M values, of the distributions of the hyperparameters μ and τ and, optionally, of the individual ψ_i's. The function uses the R function `density` to compute the kernel density estimate of the posteriors and the function `matplot` to produce the superimposed plots. The function has one required input argument, which is a list object, each element of which is the `chain` component of the output from `dirichlet.c` or `dirichlet.o`. These `chain` components are matrices of MCMC output and are assumed to correspond to different values of the Dirichlet precision parameter M. The names of the list elements will go into legend labels. An optional argument, `burnin`, with default value 1000, specifies how many of the initial runs to omit. Below is sample code which produces plots of the distributions of μ and τ, but not the ψ_i's, for the breast cancer data. The graphs produced by this code are shown in Figure 1.

```
R> breast.c1c2 <- list("5" = breast.c1$chain, "1000" = breast.c2$chain)
R> draw.post(breast.c1c2, burnin = 100)
```

The value of $M = 1000$ corresponds closely to a parametric model, whereas the value $M = 5$ is a typical value that would be used in practice. From the shapes of the two distributions in both of the above graphs, for μ and τ, we see that as expected, there is greater certainty or precision expressed in the parametric analysis. In addition, the centers of the posterior distributions arising from the parametric and semiparametric models, are different. To look at a brief comparison of the quantitative conclusions for different M values, we can get descriptive statistics of the posterior distributions using the function `describe.post` as follows:

```
R> describe.post(breast.c1c2, burnin = 100)
```

The output is the posterior means, and the probabilities that the risk ratios are less than 1:

```
 Table of Posterior Means
     Paganin Thun et Schrein Egan et Harris  Johnson Harris  Harris  Rosenbu
5      -0.11   -0.18   -0.32  -0.045   -0.37   -0.33   -0.25   -0.38   -0.24
1000   -0.11   -0.18   -0.32  -0.045   -0.38   -0.33   -0.24   -0.38   -0.23
```

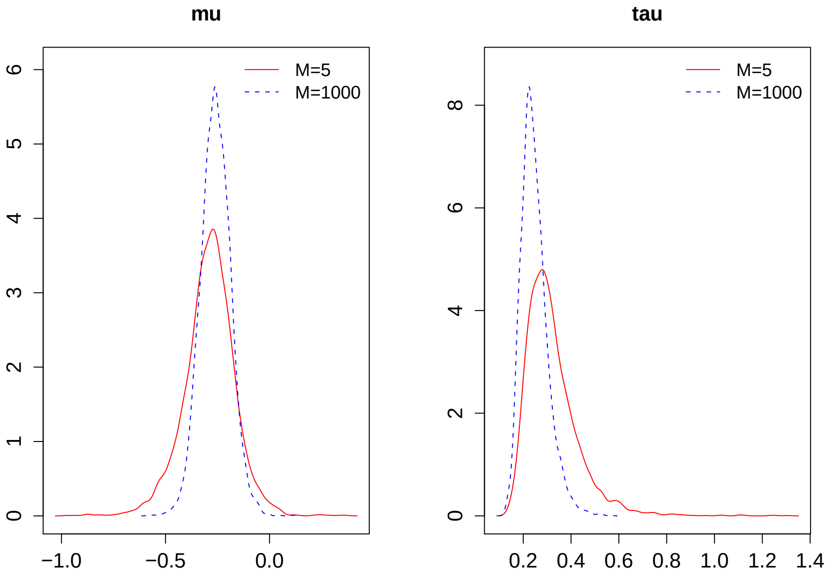

Figure 1: Posterior distributions of parameters of conditional Dirichlet model for the log risk ratios of breast cancer for NSAIDS vs. no NSAIDS. Left panel is for the median μ, and the right panel is for the standard deviation τ.

```
     Harris  Neuget  Coogan  Sharpe  Cotterc Langman Meier e Moorman    mu
5     -0.32   -0.27   -0.32   -0.28   -0.31   0.030  -0.050   -0.76 -0.29
1000  -0.32   -0.24   -0.32   -0.27   -0.31   0.048  -0.051   -0.74 -0.26
      tau
5    0.32
1000 0.25
```

```
 Table of Posterior P(RR < 1)
     Paganin Thun et Schrein Egan et Harris  Johnson Harris  Harris  Rosenbu
5       0.76    0.86    0.99    0.63       1       1       1    0.98    0.96
1000    0.79    0.89    0.99    0.66       1       1       1    0.99    0.96
     Harris  Neuget  Coogan  Sharpe  Cotterc Langman Meier e Moorman   mu
5       0.97    0.87    0.99       1       1    0.37    0.65       1 0.99
1000    0.98    0.88    0.99       1       1    0.30    0.69       1 1.00
```

For the overall conclusion about the parameter μ, both the parametric and the semiparametric model show very high significance of the result so that with either model one concludes that long-term use of NSAIDs appears to be associated with reduction of the risk of breast cancer, at least at the study level.

3.2. Example 2: Decontamination of the digestive tract

Burr and Doss (2005) summarize the background for this dataset, which appeared in Selective Decontamination of the Digestive Tract Trialists' Collaborative Group (1993). The dataset in the package, `ddtm.s`, consists of fourteen rows corresponding to fourteen different studies of the effect of a combined regimen of topical and systemic antibiotics on mortality from infection in intensive care units. The dataset has four columns, giving counts of deaths and total sample sizes, first for the treatment group and then for the control group. Each row of counts must be converted to an odds ratio and its standard error. This data is accessed (the first four rows are shown), and then converted to the appropriate form for doing the meta-analysis, as follows:

```
R> library("bspmma")
R> data("ddtm.s")
R> ddtm.s

   treat.deaths treat.total cont.deaths cont.total
1            14          45          23         46
2            22          55          33         57
3            27          74          40         77
4            11          75          16         75
...

R> ddtm.s$treat.deaths <- ddtm.s$treat.deaths + 0.5
R> ddtm.s$treat.total <- ddtm.s$treat.total + 1
R> ddtm.s$cont.deaths <- ddtm.s$cont.deaths + 0.5
R> ddtm.s$cont.total <- ddtm.s$cont.total + 1
R> attach(ddtm.s)
R> or <- (treat.deaths / (treat.total - treat.deaths)) /
+    (cont.deaths / (cont.total - cont.deaths))
R> lor <- log(or)
R> se.lor <- ((treat.total / (treat.deaths * (treat.total - treat.deaths))) +
+    (cont.total / (cont.deaths * (cont.total - cont.deaths))))^0.5
R> ddtm.14 <- data.frame(psi.hat = lor, se.psi.hat = se.lor)
```

Next we run the Markov chain for the conditional Dirichlet model, Model (4), using several values of the precision parameter M. Below we show the code for values of $M = 5$, $M = 20$, and $M = 100$, plot the posterior distributions of μ and τ (see Figure 2), and compute the summary statistics. The algorithm to run MCMC for the conditional Dirichlet model, Model (4), completes 4000 cycles in about 11 seconds (as timed by the R function `system.time`), on an Intel 2.8 GHz Q9550 running Linux.

```
R> ddtm.s.data <- as.matrix(ddtm.14)
R> set.seed(1)
R> ddtm.s.c1 <- dirichlet.c(ddtm.s.data, ncycles = 4000, M = 5)
R> set.seed(1)
R> ddtm.s.c2 <- dirichlet.c(ddtm.s.data, ncycles = 4000, M = 20)
R> set.seed(1)
R> ddtm.s.c3 <- dirichlet.c(ddtm.s.data, ncycles = 4000, M = 100)
R> ddtm.s.l1 <- list("5" = ddtm.s.c1$chain, "20" = ddtm.s.c2$chain,
+    "100" = ddtm.s.c3$chain)
R> draw.post(ddtm.s.l1, burnin = 100)
R> describe.post(ddtm.s.l1, burnin = 100)
```

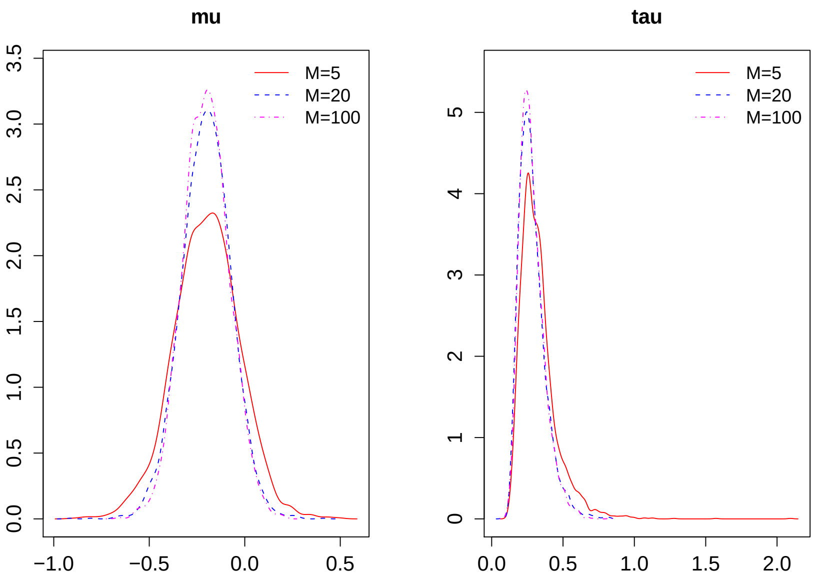

Figure 2: Posterior distributions of parameters of conditional Dirichlet model for the log risk ratios of mortality for treatment vs. control groups in the decontamination of the digestive tract dataset. Left panel is for the median μ, right panel is for the standard deviation τ.

```
 Table of Posterior Means
        1     2     3     4     5     6     7     8     9    10    11     12
5   -0.37 -0.37 -0.37 -0.27 -0.23 -0.25 -0.25 -0.23 -0.18 -0.13 -0.12 -0.047
20  -0.35 -0.36 -0.36 -0.26 -0.22 -0.25 -0.24 -0.23 -0.18 -0.13 -0.13 -0.059
100 -0.37 -0.37 -0.37 -0.28 -0.23 -0.26 -0.24 -0.24 -0.19 -0.13 -0.13 -0.061
         13    14    mu  tau
5   -0.0033 0.015 -0.20 0.33
20  -0.0174 0.010 -0.20 0.28
100 -0.0117 0.018 -0.20 0.28

 Table of Posterior P(RR < 1)
       1    2    3    4    5    6    7    8    9   10   11   12   13   14
5   0.92 0.93 0.94 0.84 0.79 0.89 0.86 0.85 0.75 0.69 0.67 0.59 0.51 0.47
20  0.93 0.94 0.95 0.86 0.80 0.91 0.87 0.87 0.77 0.71 0.71 0.62 0.55 0.47
100 0.93 0.95 0.95 0.87 0.80 0.92 0.88 0.87 0.77 0.72 0.71 0.62 0.52 0.47
      mu
5   0.88
20  0.94
100 0.95
```

For $M = 5$ the posterior probability that μ is less than 1 is 0.88, whereas for $M = 20$ and $M = 100$, the probabilities are 0.94 and 0.95, respectively. Thus whereas the parametric model gives significant results, the semiparametric model suggests that there is not enough evidence to conclude that the combined antibiotic treatment reduces the risk of mortality.

3.3. Bayes factors for breast cancer data

Package functions `bf1` and `bf2` estimate the Bayes factors for conditional vs. ordinary Dirichlet models, for a series of M values, by applying (12). This requires formulas for the Radon-Nikodym derivatives on the left-hand side of (12), and it requires generation of MCMC samples from the posterior distribution of θ under the ordinary Dirichlet model with specified hyperparameter h_1.

Formulas for the Radon-Nikodym derivatives $[d\nu_h^o/d\nu_{h_1}^o](\theta)$ and $[d\nu_h^c/d\nu_{h_1}^o](\theta)$ are taken from Doss (2012). They are computed by functions `lr` and `pnew.pold` for ordinary Dirichlet vs. ordinary Dirichlet, and by functions `lr.c.o` and `pnew.pold.c.o` for conditional Dirichlet vs. ordinary Dirichlet. These functions are provided in the package, but are not user-accessible.

Regarding the hyperparameter value $h_1 = (M_1, d_1)$ under which the Markov chain will be run to estimate the Bayes factors for a range of M's, in principle, any value of h_1 can be used, but in practice one would like to use a value of M_1 that is "close" to all values of M for which the Bayes factor m_h^c/m_h^o will be estimated, keeping in mind that in the practical sense, $M = 1$ is farther away from $M_1 = 4$ than is $M = 10$. In fact, for better accuracy of the estimates, it is preferable to run multiple Markov chains corresponding to several values of h_1. Buta and Doss (2011) motivate and develop the use of multiple chains in a general context, and Doss (2012) gives a discussion focused on the present situation involving Dirichlet processes; in particular, he gives guidelines regarding the selection of the multiple values of h_1. We do not give a theoretical discussion of these issues in the present paper, but mention only that doing importance sampling with respect to multiple chains results in a very significant increase in the accuracy of the estimates. We follow the recommendations given in Doss (2012) and use values of the Dirichlet precision parameter M starting from $2^{-2} = 0.25$ and increasing to $2^6 = 64$ in multiples of 2, for a total of nine values. These values of M_1 should yield accurate estimates unless the user wishes to estimate the Bayes factor for values of M that are very small (M less than 0.25), which is generally considered dubious (Sethuraman and Tiwari 1982). Very large values of M are "covered" since for most data sets, the posteriors corresponding to $M = \infty$ and $M = 64$ are close. The nine simulations are carried out by the user-accessible function `bf1`. The only required argument is the data, in the form of a two-column matrix, as for the function `dirichlet.c`, illustrated in Section 3.1 and Section 3.2.

There are two steps controlled by the user in the package implementation of the multi-chain algorithm. First, we need the constants for the denominator on the left side of Equation 2.5 in Doss (2012). This requires output from the nine Markov chains produced by `bf1`, and then, a call to the function `bf2` to use the MCMC output to compute the constants. It is not necessary to understand the role of these constants nor the method needed to compute them in order to use `bf1` and `bf2`, and the user can view these functions as a "black box." The necessary commands for this preliminary step are illustrated below for the breast cancer data. In our analysis we fix the hyperparameter vector of the normal/inverse gamma prior on (μ, τ) to be $(0.1, 0.1, 0, 1000)$, which is the default value in the function `bf1`. The total number of cycles for each of the nine chains is indicated by the argument `ncycles`, and the number of simulations dropped for each chain is given by the argument `burnin`; in this example, the object returned by `bf1` is a list of nine matrices, each having $5000 - 1000 = 4000$ rows of MCMC output. Calls to `bf1` with `ncycles` set at 5000 take about 1.5 minutes to run on an Intel 2.8 GHz Q9550 running Linux; calls to `bf2` take about 7 seconds in examples this size.

```
R> data("breast.17")
```

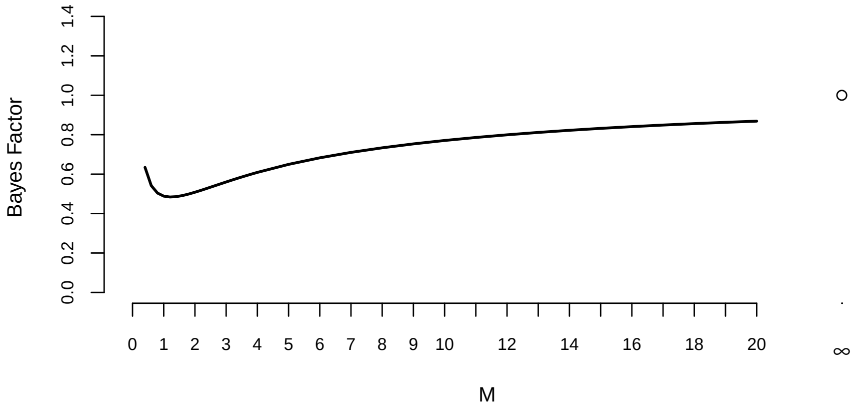

Figure 3: Bayes factors for conditional Dirichlet vs. ordinary Dirichlet for breast cancer data.

```
R> breast.data <- as.matrix(breast.17)
R> chain1.list <- bf1(breast.data, ncycles = 5000, burnin = 1000)
R> cc <- bf2(chain1.list)
```

The next step is another set of nine MCMC simulations independent of the previous set (we use a different seed value). These will be used for the final computation of the sequence of Bayes factors.

```
R> chain2.list <- bf1(breast.data, seed = 2, ncycles = 5000, burnin = 1000)
```

In general, this method of estimating Bayes factors produces accurate estimates for M ranging from a little less than 0.25 to infinity; however, for this breast cancer dataset, after experimentation, we determined that the only real information in the Bayes factor plot occurs for $M \leq 20$; after that, the graph levels off. The commands to produce the plot for values of M from 0.8 to 20 are shown below; this command required about 3 minutes on an Intel 2.8 GHz Q9550 running Linux.

```
R> breast.bfco <- bf.c.o(from = 0.8, incr = 0.2, to = 20, cc = cc,
+    mat.list = chain2.list)
R> draw.bf(breast.bfco)
```

The resulting graph is shown in Figure 3. Note that at the right end of the graph, the value of the Bayes factor is 1 for $M = \infty$, since the two models are the same: They are both equal to the parametric model; thus $M = \infty$ serves as a reference point. For this particular data set, the Bayes factors are always less than 1, and thus the ordinary Dirichlet model is always preferred to the conditional model, although the preference is hardly strong, particularly when $M \geq 8$.

For selection of the value of M, the package has the function `bf.o` for the ordinary model and `bf.c` for the conditional model. Since for this dataset the ordinary Dirichlet model is preferred to the conditional, we next illustrate how to select the value of M in the ordinary model. The steps to carry out the multi-chain algorithm are the same as given above, and for convenience, we will use the same sets of chains and constants as before; the only change is

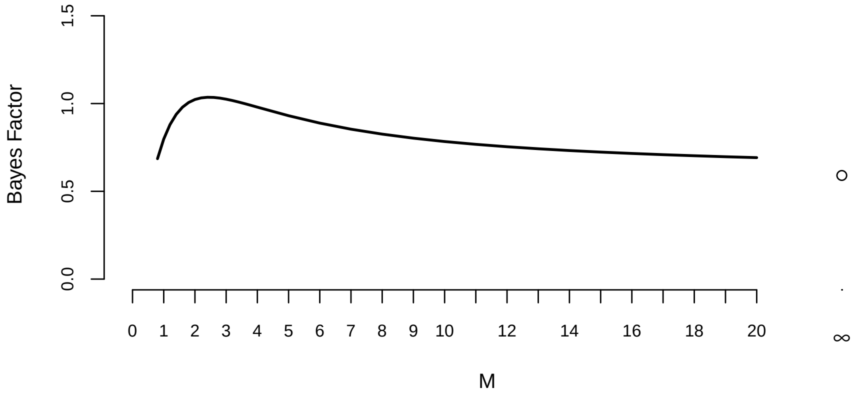

Figure 4: Bayes factors for selection of M in the ordinary Dirichlet model for breast cancer data.

in the final two commands, which are shown below. The call to `bf.o` takes about 1.5 minutes to run on an Intel 2.8 GHz Q9550 running Linux.

```
R> breast.bfo <- bf.o(from = 0.8, incr = 0.2, to = 20,cc = cc,
+    mat.list = chain2.list)
R> draw.bf(breast.bfo)
```

The resulting graph is shown in Figure 4. The relevant features are the shape of the graph, and the point of the maximum, which occurs in this case at $M = 2.4$. Then, the Bayes factor for $M = 2.4$ vs. $M = \infty$ may be obtained by taking the ratio of the appropriate elements in the object `breast.bfo`, as shown below.

```
R> breast.bfo$y[9] / breast.bfo$yinfinity
```

```
[1] 1.7541
```

The value of 1.75 for the Bayes factor suggests a slight preference for the nonparametric model over the parametric for the breast cancer dataset.

Since for this dataset the conditional model is not preferred to the ordinary, we created a hypothetical dataset by changing three of the points in such a way as to make the data more clumped. Figure 5 displays the original and modified versions of the breast cancer data for the 17 studies. The locations of the vertical lines are the observed log odds ratios, and their heights are proportional to the reciprocals of the estimated standard errors. The distribution is estimated using a kernel density approach (as implemented in the R function `density`) based on the observed log odds ratios, with weights proportional to the estimated standard errors. This density estimate is exploratory and should be viewed with caution, since the observed log odds ratios are only estimates of the true log odds ratios. Figure 6 shows the new Bayes factors, which now indicate that the conditional Dirichlet model is preferred to the ordinary model. From a wide range of experiments we have carried out, we have noticed that the Bayes factor plot seems to favor the model based on conditional Dirichlets for data sets which have outliers and also have clusters (although it is very difficult to come up with

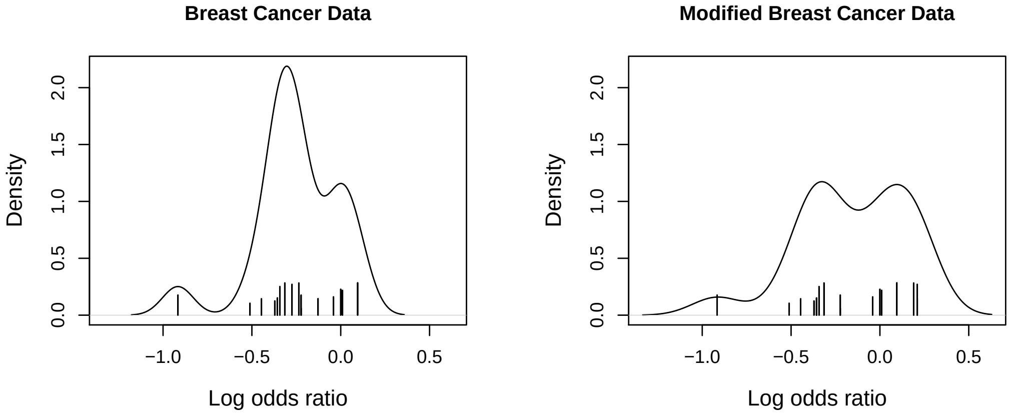

Figure 5: Estimate of distribution of the study-specific effect ψ for the breast cancer data, original and modified. Data are represented by vertical lines, whose locations are the estimates of the log odds ratios and whose heights are proportional to the reciprocals of the estimated standard errors.

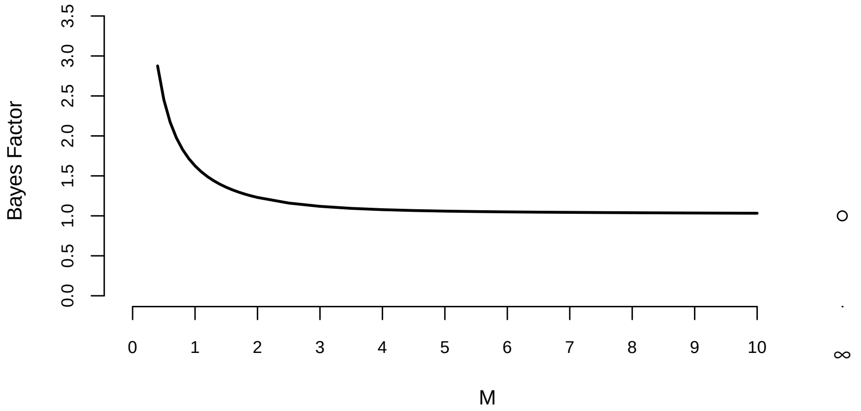

Figure 6: Bayes factors for conditional Dirichlet vs. ordinary Dirichlet for breast cancer data with two points modified to make clumps in the distribution.

a mathematical explanation for this phenomenon). In general it is hard to tell by visual inspection of a data set which model will be favored, and the only way to tell for sure is to look at the Bayes factor plot.

4. Discussion

The core contributions of **bspmma** are: (i) a suite of functions which implement meta-analysis when the effect distribution has as prior a mixture of conditional Dirichlet processes, (ii) functions to carry out a comparison between a model based on a mixture of conditional Dirichlets vs. a model based on a mixture of ordinary Dirichlets, and (iii) functions to carry out an empirical Bayes selection of the precision parameter M of the Dirichlet process. It should be noted that (iii) effectively enables a decision on whether to use a semiparametric model or use a parametric model: an empirical Bayes choice of $M = \infty$ indicates that there is no compelling reason to use a more complicated semiparametric model, and that the parsimonious choice of a parametric model is adequate.

As mentioned in Section 1, it is possible to use mixtures of finite Polya trees instead of mixtures of conditional Dirichlets in model (4). This is done in Branscum and Hanson (2008), which also gives a comparison of Polya tree and conditional Dirichlet models on the digestive tract dataset of Section 3.2. In order to be able to comment on their approach we very briefly review the construction they use. Let $\{H_\vartheta\}_{\vartheta \in \Omega}$ be a parametric family of distributions on $\mathbb{R}$, let ρ be a function mapping the positive integers into the positive real numbers, and let $c > 0$. We initially view ϑ as fixed. The real line is first split into two intervals, with the split point being the median of H_ϑ; then each interval is further split into two subintervals, with split points being $H_\vartheta(1/4)$ and $H_\vartheta(3/4)$; and this process is continued indefinitely. At the j-th level, when the k-th interval is split ($k = 1, \ldots, 2^j$), we form a random variable $p(j, k)$ with the beta distribution $\text{Beta}(c\rho(j), c\rho(j))$. These random variables are all independent. At any level j, an interval at that level has (random) probability given by the product of all the beta random variables along the branch of the tree leading to that set. It can be shown that this process specifies a distribution on the set of cumulative distributions on the real line, and that if F is distributed according to this distribution, then for every $t \in \mathbb{R}$, $E(F(t)) = H_\vartheta(t)$. This distribution is called a Polya tree, and a mixture of Polya trees results when ϑ is random, and a probability distribution λ is assigned to it. In principle, the function ρ can be very general but, Branscum and Hanson (2008) consider the interesting one-parameter family given by $\rho_\nu(j) = 2^{-\nu j}$. Roughly speaking, small values of ν give rise to distributions which are smooth, while large values of ν give rise to distributions which are discrete. For instance, if $\nu < 0$, the Polya tree gives probability 1 to the set of distributions which are absolutely continuous with respect to Lebesgue measure, while the case $\nu = 1$ gives exactly the Dirichlet process. By taking the random variable associated with the first split to be deterministically equal to 1/2 (i.e., $p(1, 1) = p(1, 2) = 1/2$), one guarantees that if F is distributed according to the Polya tree, the median of F is equal to the median of H_ϑ, which gives a direct analogue to the conditional Dirichlet process.

Of course a Polya tree is a probability distribution on an infinite-dimensional space, and for computational purposes it is necessary to take the number of levels J to be finite. Hanson (2006) gives recommendations for the value of J based on empirical evidence, but there are no theoretical results for the choice of J. Inference based on Polya trees can depend on the sequence of binary partitions used to define the tree. It is worth mentioning that the construction of the Dirichlet process does not depend on a sequence of partitions of the real line, and the Markov chain algorithm used in **bspmma** is exact, i.e., the stationary distribution of $\theta = (\psi, \mu, \tau)$ is exactly the conditional distribution of θ given the data for model (4), and the only error incurred is the Monte Carlo error associated with the rate of convergence of the chain.

The class of Polya trees is overwhelmingly large, and this makes it necessary—and difficult—to make a choice of a particular Polya tree to use. The choice $\rho_\nu(j) = 2^{-\nu j}$ makes a reduction to a one-parameter family, which contains the conditional Dirichlet processes used in this paper. It would be interesting to develop methods for creating Bayes factor plots similar to those given in Section 2, for the purpose of enabling an empirical Bayes estimate of the hyperparameter ν. More specifically, let m_ν be the marginal likelihood of the data when we use a Polya tree with parameter ν, and let $B(\nu) = m_\nu / m_1$. Suppose we can develop an estimate $\hat{B}(\nu)$ for a range of values of ν that includes $\nu = 1$. If the maximum of the estimated function is achieved at or near 1, then this would be viewed as evidence that the Dirichlet-based model holds, while if the maximum occurs at a point that is far away from 1, the Dirichlet-based model would be viewed as inadequate.

Acknowledgments

The author would like to thank the two anonymous reviewers for their valuable remarks and suggestions.

References

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

Branscum A, Hanson T (2008). "Bayesian Nonparametric Meta-Analysis Using Polya Tree Mixture Models." *Biometrics*, **64**, 825–833.

Burr D, Doss H (2005). "A Bayesian Semiparametric Model for Random-Effects Meta-Analysis." *Journal of the American Statistical Association*, **100**, 242–251.

Buta E, Doss H (2011). "Computational Approaches for Empirical Bayes Methods and Bayesian Sensitivity Analysis." *Annals of Statistics*, **39**, 2658–2685.

DerSimonian R, Laird N (1986). "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials*, **7**, 177–188.

Doss H (1985). "Bayesian Nonparametric Estimation of the Median. Part I: Computation of the Estimates." *The Annals of Statistics*, **13**, 1432–1444.

Doss H (2012). "Hyperparameter and Model Selection for Nonparametric Bayes Problems via Radon-Nikodym Derivatives." *Statistica Sinica*, **22**, 1–26.

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

Ferguson TS (1974). "Prior Distributions on Spaces of Probability Measures." *The Annals of Statistics*, **2**, 615–629.

Gonzalez-Perez A, Rodriguez LAG, Lopez-Ridaura R (2003). "Effects of Non-Steroidal Anti-Inflammatory Drugs on Cancer Sites Other than the Colon and Rectum: A Meta-Analysis." *BMC Cancer*, **3**(28).

Hanson T (2006). "Inferences for Mixtures of Finite Polya Tree Models." *Journal of the American Statistical Association*, **101**, 1548–1565.

Harris R, Beebe-Donk J, Doss H, Burr D (2005). "Aspirin, Ibuprofen and Other Non-Steroidal Anti-Inflammatory Drugs in Cancer Prevention: A Critical Review of Non-Selective COX-2 Blockade." *Oncology Reports*, **13**, 559–584.

Harris RE, Chlebowski RT, Jackson RD, Frid DJ, Ascenseo JL, Anderson G, Loar A, Rodabough RJ, White E, McTiernan A (2003). "Breast Cancer and Nonsteroidal Anti-Inflammatory Drugs: Prospective Results from the Women's Health Initiative." *Cancer Research*, **63**, 6096–6101.

Jara A (2007). "Applied Bayesian Non- and Semi-parametric Inference Using **DPpackage**." *R News*, **7**(3), 17–26. URL http://CRAN.R-project.org/doc/Rnews/.

Jara A, Hanson T, Quintana F, Müller P, Rosner G (2011). "**DPpackage**: Bayesian Semi- and Nonparametric Modeling in R." *Journal of Statistical Software*, **40**(5), 1–30. URL http://www.jstatsoft.org/v40/i05/.

Khuder SA, Mutgi AB (2001). "Breast Cancer and NSAID Use: A Meta-Analysis." *British Journal of Cancer*, **84**(9), 1188–1192.

Lumley T (2009). ***rmeta**: Meta-Analysis.* R package version 2.16, URL http://CRAN.R-project.org/package=rmeta.

Mallick BK, Walker SG (1997). "Combining Information from Several Experiments with Nonparametric Priors." *Biometrika*, **84**, 697–706.

Plummer M, Best N, Cowles K, Vines K (2006). "**coda**: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL http://CRAN.R-project.org/doc/Rnews/.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Schwarzer G (2010). ***meta**: Meta-Analysis with R.* R package version 1.6-1, URL http://CRAN.R-project.org/package=meta.

Selective Decontamination of the Digestive Tract Trialists' Collaborative Group (1993). "Meta-Analysis of Randomised Controlled Trials of Selective Decontamination of the Digestive Tract." *British Medical Journal*, **307**, 525–532.

Sethuraman J (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, **4**, 639–650.

Sethuraman J, Tiwari RC (1982). "Convergence of Dirichlet Measures and the Interpretation of Their Parameter." In *Statistical Decision Theory and Related Topics III*, volume 2, pp. 305–315. Academic, London.

Sharples LD (1990). "Identification and Accommodation of Outliers in General Hierarchical Models." *Biometrika*, **77**(3), 445–453.

Smith BJ (2007). "**boa**: An R Package for MCMC Output Convergence Assessment and Posterior Inference." *Journal of Statistical Software*, **21**(11), 1–37. URL http://www.jstatsoft.org/v21/i11/.

Terry MB, Gammon MD, Zhang FF, Tawfik H, Teitelbaum SL, Britton JA, Subbaramaiah K, Dannenberg AJ, Neugut AI (2004). "Association of Frequency and Duration of Aspirin Use and Hormone Receptor Status with Breast Cancer Risk." *Journal of the American Medical Association*, **291**, 2433–2440.

Viechtbauer W (2010). "Conducting Meta-Analyses in R with the **metafor** Package." *Journal of Statistical Software*, **36**(3), 1–48. URL http://www.jstatsoft.org/v36/i03/.

Weiss R, Cho M, Yanuzzi M (1999). "On Bayesian Calculations for Mixture Likelihoods and Priors." *Statistics in Medicine*, **18**, 1555–1570.

Zhang Y, Coogan PF, Palmer JR, Strom BL, Rosenberg L (2005). "Use of Nonsteroidal Anti-Inflammatory Drugs and Risk of Breast Cancer: The Case-Control Surveillance Study Revisited." *American Journal of Epidemiology*, **162**, 165–170.

Affiliation:

Deborah Burr
Gainesville, FL, United States of America
E-mail: burr@stat.ufl.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 50, Issue 4 *Submitted:* 2010-08-13
July 2012 *Accepted:* 2012-05-08