Hyperparameter and Model Selection for Nonparametric Bayes Problems via Radon-Nikodym Derivatives

Hani Doss
Department of Statistics
University of Florida
Gainesville, FL 32611

Abstract

We consider families of semiparametric Bayesian models based on Dirichlet process mixtures, indexed by a multidimensional hyperparameter that includes the precision parameter. We wish to select the hyperparameter by considering Bayes factors. Our approach involves distinguishing some arbitrary value of the hyperparameter, and estimating the Bayes factor for the model indexed by the hyperparameter vs. the model indexed by the distinguished point, as the hyperparameter varies. The approach requires us to select a finite number of hyperparameter values, and for each get Markov chain Monte Carlo samples from the posterior distribution corresponding to the model indexed by that hyperparameter value. Implementation of the approach relies on a likelihood ratio formula for Dirichlet process models. Because we may view parametric models as limiting cases where the precision hyperparameter is infinity, the method also enables us to decide whether or not to use a semiparametric or an entirely parametric model. We illustrate the methodology through two detailed examples involving meta-analysis.

Key words and phrases: Bayes factors; Dirichlet processes; Likelihood ratio formula; Markov chain Monte Carlo; Model selection.

1 Introduction

Bayesian hierarchical models have proven very useful in analyzing random effects models, of which the following is a simple case. Suppose we have m different "centers" and at each one we gather data Y_j from the distribution $P_j(\psi_j)$. This distribution depends on ψ_j and also on other quantities, for example the sample size as well as nuisance parameters specific to the j^{th} center. In a typical example arising in biostatistics, we have the same experiment being carried out at m different centers, ψ_j represents a treatment effect for the experiment at center j (e.g. ψ_j might be a regression coefficient for an indicator of treatment vs. placebo in a Cox model, or simply an odds ratio) and Y_j is the estimate of this parameter. Because each center has its own characteristics, the ψ_j's are not assumed to be the same, but are assumed to come from some distribution.

When dealing with this kind of data it is very common to use a hierarchical model of the following sort:

$$\text{conditional on } \psi_j, \qquad Y_j \overset{\text{indep}}{\sim} \mathcal{N}(\psi_j, \sigma_j^2), \quad j = 1, \ldots, m, \tag{1.1a}$$

$$\text{conditional on } \mu, \tau, \qquad \psi_j \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), \quad j = 1, \ldots, m, \tag{1.1b}$$

$$(\mu, \tau) \sim \lambda_c. \tag{1.1c}$$

In (1.1a), the σ_j's are the standard errors that usually accompany the point estimates Y_j's. In (1.1c), λ_c is the normal/inverse gamma prior, indexed by the vector $c = (c_1, c_2, c_3, c_4)$, i.e. $1/\tau^2 \sim \text{gamma}(c_1, c_2)$, and given τ, $\mu \sim \mathcal{N}(c_3, \tau c_4)$; this prior is often used because it is conjugate to the family $\mathcal{N}(\mu, \tau^2)$. Typically one uses $c_3 = 0$, c_4 some large value, and c_1 and c_2 relatively small, giving a fairly diffuse prior.

The approximation of $P_j(\psi_j)$ by a normal distribution in (1.1a) is typically supported by some theoretical result, for example the asymptotic normality of maximum likelihood estimates. By contrast, the normality statement in (1.1b) regarding the distribution of the random effects is a modelling assumption, which generally is made for the sake of convenience and does not have any theoretical justification. While t distributions may be used in place of the normal distribution in line (1.1b) to better accommodate outliers, the distribution of the random effects may deviate from normality in ways that do not involve heaviness of tails.

It is therefore desirable to relax this assumption and consider models that allow for departures from normality. A commonly used choice is a model based on mixtures of Dirichlet processes (Antoniak 1974), and before proceeding, we give a brief review of this class of priors. Let G_ϑ, $\vartheta \in \Omega \subset \mathbb{R}^p$ be a parametric family of distributions on the real line, and let λ be a distribution on Ω. Suppose $M > 0$, and define $\alpha_\vartheta = MG_\vartheta$. If ϑ is chosen from λ, and then F is chosen from $\mathcal{D}_{\alpha_\vartheta}$, the Dirichlet process with parameter measure α_ϑ (Ferguson 1973, 1974), we say that the prior on F is a mixture of Dirichlet processes. The parameter M can be interpreted as a precision parameter that indicates the degree of concentration of the prior on F around the parametric family $\{G_\vartheta,\ \vartheta \in \Omega\}$. In a somewhat oversimplified but nevertheless useful view of this class of priors, we think of the family $\{G_\vartheta,\ \vartheta \in \Omega\}$ as a "line" (of dimension p) in the infinite-dimensional space of cdf's, and we imagine "tubes" around this line. For large values of M, the mixture of Dirichlet processes puts most of its mass in narrow tubes, while for small values of M the prior is more diffuse. Formally, the prior on F is the integral $\int \mathcal{D}_{MG_\vartheta}\,\lambda(d\vartheta)$ and is parameterized by the triple $h = (\{G_\vartheta\}_{\vartheta\in\Omega}, M, \lambda)$.

Beginning with the work of Escobar (1988, 1994), Escobar and West (1995), and West et al. (1994), a very large literature has considered Bayesian hierarchical models based on Dirichlet processes or their mixtures where, for instance, (1.1) is replaced by

$$
\begin{aligned}
&\text{conditional on } \psi_j, && Y_j \overset{\text{indep}}{\sim} \mathcal{N}(\psi_j, \sigma_j^2), \quad j = 1, \ldots, m, && \text{(1.2a)}\\
&\text{conditional on } F, && \psi_j \overset{\text{iid}}{\sim} F, \quad j = 1, \ldots, m, && \text{(1.2b)}\\
&\text{conditional on } (\mu, \tau) && F \sim \mathcal{D}_{Mt_{v,\mu,\tau}}, && \text{(1.2c)}\\
& && (\mu, \tau) \sim \lambda_c. && \text{(1.2d)}
\end{aligned}
$$

Here, $\vartheta = (\mu, \tau)$ and $G_\vartheta = t_{v,\mu,\tau}$ is the t distribution with v degrees of freedom, location μ, and scale τ.

A question that immediately arises whenever one wishes to use a model of this sort is how to choose the hyperparameters of the model. For instance in model (1.2), choosing the precision parameter M to be very large will essentially result in a parametric model, and choosing the degrees of freedom parameter v to be infinity will signify the choice of the normal distribution. Because the values of both M and v have a significant impact on the resulting inference, the question of how to choose these parameters is very important.

To explain how we can deal with this question, let us consider the specific case of model (1.2). Here, lines (1.2d), (1.2c), and (1.2b), in that order, induce a prior on $\theta = (\vartheta, \psi)$ (where $\psi = (\psi_1, \ldots, \psi_m)$). This prior depends on the hyperparameter $h = (v, M, c)$, and we shall denote it by ν_h. Model (1.2) may then be re-expressed as

$$
\begin{aligned}
&\text{conditional on } \theta, \quad Y_j \overset{\text{indep}}{\sim} \mathcal{N}(\psi_j, \sigma_j^2), \quad j = 1, \ldots, m, && \text{(1.3a)}\\
&\theta \sim \nu_h. && \text{(1.3b)}
\end{aligned}
$$

When looked at in this way, we see that choosing the hyperparameter of the prior ν_h involves not only choosing the prior on (μ, τ), but also the number of degrees of freedom v

and the precision parameter M. For a fixed value of h, the marginal distribution of Y is given by $m_h(y) = \int \ell_y(\theta)\, d\nu_h(\theta)$, in which $\ell_y(\theta)$ is the likelihood given by line (1.3a), i.e. $\ell_y(\theta) = \prod_{j=1}^m \phi_{\psi_j,\sigma_j}(y_j)$, where $\phi_{m,s}(x)$ denotes the density of the normal distribution with mean m and standard deviation s, evaluated at x. Note that $m_h(y)$ is the normalizing constant in the statement "the posterior is proportional to the likelihood times the prior," i.e. $\nu_{h,y}(d\theta) = \ell_y(\theta)\, \nu_h(d\theta)/m_h(y)$. The empirical Bayes choice of h is by definition the maximizer of $m_h(y)$ viewed as a function of h, and to obtain it we need to estimate this function. Now if h_1 is a fixed value of the hyperparameter, the information regarding h given by $m_h(y)$ and $m_h(y)/m_{h_1}(y)$ is the same, and in particular, the same value of h maximizes both functions. From a statistical and computational standpoint however, estimation of ratios of normalizing constants can be far more stable than estimation of the normalizing constants themselves (this point is discussed further in Section 4). From now on we write m_h instead of $m_h(y)$. The quantity $B(h, h_1) = m_h/m_{h_1}$ is the Bayes factor of the model with hyperparameter h relative to the model with hyperparameter h_1. If we had a method for estimating the Bayes factor, then we could fix a particular hyperparameter value h_1 and plot $B(h, h_1)$ as a function of h, and this would enable us to make reasonable choices of h. (Of course, the value of h_1 is arbitrary: if we choose a different value h_*, then $B(h, h_*)$ is a constant multiple of $B(h, h_1)$.)

Suppose that h_1 is fixed. We now explain how we can in principle estimate the entire family $B(h, h_1)$. The posterior distributions for model (1.2) cannot be calculated in closed form, but must be estimated via a simulation method, such as Markov chain Monte Carlo (MCMC). All MCMC methods for this kind of model produce, directly or indirectly, estimates of the posterior distribution of the vector of latent parameters $\theta = (\vartheta, \psi)$, which we will denote $\nu_{h,y}$. [Some methods marginalize the infinite-dimensional parameter F entirely and work specifically on $\nu_{h,y}$; for example, the original algorithm of Escobar (1994) and numerous later improvements are of this sort. Others work through a representation of F (Sethuraman 1994) that is explicit enough to enable the generation of θ from $\nu_{h,y}$; the algorithm of Doss (1994b) is of this kind.] By writing $\nu_{h,y}(d\theta) = \ell_y(\theta)\, \nu_h(d\theta)/m_h$, we easily obtain the frequently-used identity

$$\int \left[\frac{d\nu_h}{d\nu_{h_1}}\right](\theta)\, \nu_{h_1,y}(d\theta) = \frac{m_h}{m_{h_1}}, \tag{1.4}$$

where $[d\nu_h/d\nu_{h_1}]$ is the Radon-Nikodym derivative of ν_h with respect to ν_{h_1}. [This identity is normally written in terms of densities, but the distributions ν_h and ν_{h_1} are not absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{m+2}$, the reason being that when ψ is generated according to lines (1.2d), (1.2c), and (1.2b), there is positive probability that the ψ_j's are not all distinct. Although $\nu_{h,y}$ and $\nu_{h_1,y}$ will typically not have densities, they are mutually absolutely continuous if ν_h and ν_{h_1} are mutually absolutely continuous (which is the case in model (1.2)), and this is the reason why we use a Radon-Nikodym derivative instead of a ratio of densities in (1.4).] Therefore, if $\theta_1, \dots, \theta_n$ is a sample from $\nu_{h_1,y}$ (iid or ergodic Markov chain output), the average

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{d\nu_h}{d\nu_{h_1}} \right] (\theta_i) \tag{1.5}$$

converges almost surely to m_h/m_{h_1}. Of course, to use (1.5) we need to have an expression for the Radon-Nikodym derivative. Assuming we do have such an expression, one difficulty we face is that when h is not close to h_1, the estimate (1.5) may be unstable, because $[d\nu_h/d\nu_{h_1}]$ may vary greatly in the region where the θ_i's are likely to be.

This paper is organized as follows. In Section 2 we give a formula for the Radon-Nikodym derivative $[d\nu_h/d\nu_{h_1}]$ for models based on mixtures of Dirichlet processes, such as (1.2). We also discuss a way of producing estimates like (1.5), but which have small variance over a wide range of h's. In Section 3 we give two illustrations on model selection questions that arise in applications of these Dirichlet-based models. In Section 4 we discuss some implementation issues and other work on estimation of Bayes factors in nonparametric Bayes models. In the Appendix we give a proof of the result in Section 2 concerning the Radon-Nikodym derivative, and also a proof of a formula that is used in the second illustration in Section 3.

2 Theoretical Development

Consider (1.2) stated more generally:

$$\begin{aligned}
&\text{conditional on } \psi_j, \quad Y_j \overset{\text{indep}}{\sim} P_j(\psi_j), \quad j = 1, \dots, m, &\text{(2.1a)}\\
&\text{conditional on } F, \quad \psi_j \overset{\text{iid}}{\sim} F, \quad j = 1, \dots, m, &\text{(2.1b)}\\
&\text{conditional on } \vartheta \quad F \sim \mathcal{D}_{M_\vartheta G_\vartheta}, &\text{(2.1c)}\\
&\qquad\qquad\qquad\quad \vartheta \sim \lambda. &\text{(2.1d)}
\end{aligned}$$

Here, G_ϑ is a parametric family of distributions, and the precision parameter in (2.1c) is now allowed to depend on ϑ. Let $h = \big(\{G_\vartheta\}_{\vartheta\in\Omega}, \lambda, \{M_\vartheta\}_{\vartheta\in\Omega}\big)$, and as before, let ν_h denote the distribution of $\theta = (\vartheta, \psi)$. Let $h_i = \big(\{G_\vartheta^{(i)}\}_{\vartheta\in\Omega}, \lambda^{(i)}, \{M_\vartheta^{(i)}\}_{\vartheta\in\Omega}\big)$, for $i = 1, 2$, be two given instances of the model. Typically, the distributions ν_{h_1} and ν_{h_2} on θ are mutually absolutely continuous (whereas the distributions on F are not), and Theorem 1 gives a formula for the Radon-Nikodym derivative. Define also ν_{par,h_i} to be the distribution of θ under the parametric version of this model, where (2.1b) and (2.1c) are replaced simply by $\psi_j \overset{\text{iid}}{\sim} G_\vartheta^{(i)}$ [and where h_i is now just $(\{G_\vartheta^{(i)}\}_{\vartheta\in\Omega}, \lambda^{(i)})$].

2.1 The Radon-Nikodym Derivative for Some Models Based on Dirichlet Mixtures

For $\psi \in \mathbb{R}^m$, let $d = d(\psi)$ be the number of distinct values of ψ, and let $\psi_{(1)} < \psi_{(2)} < \cdots < \psi_{(d)}$ be the ordered distinct values. Let Γ denote the gamma function. We will sometimes slightly abuse notation and write ν_1, ν_2, $\nu_{\text{par},1}$, and $\nu_{\text{par},2}$ instead of ν_{h_1}, ν_{h_2}, ν_{par,h_1}, and ν_{par,h_2} in order to avoid double and triple subscripting. We will consider the following conditions:

A1 $\lambda^{(1)} \ll \lambda^{(2)}$ and $[d\lambda^{(1)}/d\lambda^{(2)}]$ is continuous;

A2 For each ϑ, $G_\vartheta^{(i)}$ has density $g_\vartheta^{(i)}$, for $i = 1, 2$;

A3 $g_\cdot^{(i)}(\cdot)$ are jointly continuous in ϑ and ψ, and $M_\vartheta^{(i)}$ are continuous in ϑ, for $i = 1, 2$;

A4 $g_\vartheta^{(1)}(\psi) = 0$ whenever $g_\vartheta^{(2)}(\psi) = 0$ for all ϑ and all ψ.

Theorem 1 *Assume the conditions stated above.*

(i) The Radon-Nikodym derivative $[d\nu_1/d\nu_2]$ *is given by*

$$\left[\frac{d\nu_1}{d\nu_2}\right](\vartheta, \psi) = \left\{\prod_{r=1}^{d} \frac{g_\vartheta^{(1)}(\psi_{(r)})}{g_\vartheta^{(2)}(\psi_{(r)})}\right\} \left(\frac{M_\vartheta^{(1)}}{M_\vartheta^{(2)}}\right)^d \left\{\frac{\Gamma\big(M_\vartheta^{(1)}\big)\Gamma\big(M_\vartheta^{(2)} + m\big)}{\Gamma\big(M_\vartheta^{(2)}\big)\Gamma\big(M_\vartheta^{(1)} + m\big)}\right\} \left[\frac{d\lambda^{(1)}}{d\lambda^{(2)}}\right](\vartheta). \quad (2.2)$$

(ii) The Radon-Nikodym derivative $[d\nu_{\text{par},1}/d\nu_2]$ *is given by*

$$\left[\frac{d\nu_{\text{par},1}}{d\nu_2}\right](\vartheta, \psi) = \begin{cases} \left[\prod_{j=1}^{m} \frac{g_\vartheta^{(1)}(\psi_j)}{g_\vartheta^{(2)}(\psi_j)}\right] \left(\prod_{j=1}^{m-1} \frac{M_\vartheta^{(2)} + j}{M_\vartheta^{(2)}}\right) \left[\frac{d\lambda^{(1)}}{d\lambda^{(2)}}\right](\vartheta) & \text{if } d = m, \\ 0 & \text{if } d < m. \end{cases} \quad (2.3)$$

A1 and A4 are absolute continuity conditions in one direction (e.g. $G_\vartheta^{(1)} \ll G_\vartheta^{(2)}$, but not $G_\vartheta^{(2)} \ll G_\vartheta^{(1)}$); however, in all typical applications we will have absolute continuity in both directions, and the choice of $\lambda^{(2)}$ and $G_\vartheta^{(2)}$ will be based on convenience.

We now discuss the role of the number of distinct observations in the formulas. Suppose α is any finite non-atomic measure, and $F \sim \mathcal{D}_\alpha$. If given F, $\psi_1, \ldots, \psi_m \overset{\text{iid}}{\sim} F$, then the ψ_j's form clusters, with the ψ_j's in the same cluster being equal. If the value of $\alpha(\mathbb{R})$ decreases, the number of distinct observations tends to decrease.

For a given cluster configuration c, let κ_c, a measure on $\mathbb{R}^m$, be "Lebesgue measure for that configuration," defined as Lebesgue measure on the hyperplane defined by the configuration. As a simple example, if $m = 2$ and the configuration is determined by the equality $\psi_1 = \psi_2$, then κ_c is Lebesgue measure on the $45°$ line in $\mathbb{R}^2$. If $\kappa = \sum_c \kappa_c$ where the sum is over all possible configurations, then the distribution of ψ is absolutely continuous with respect to κ. Therefore, returning to model (2.1), if we let Leb_Ω be Lebesgue measure on Ω (recall that Ω is the space of ϑ's), we see that if λ is absolutely continuous with respect to Leb_Ω, then ν_h is absolutely continuous with respect to $\rho = \kappa \times \mathsf{Leb}_\Omega$ (and this absolute continuity continues to hold under the parametric version of this model).

Comments on Theorem 1

1 If the parametric families $\{G_\vartheta^{(1)}\}$ and $\{G_\vartheta^{(2)}\}$ are the same, then the first term on the right side of (2.2) and (2.3) is 1, and the formulas simplify.

2 Suppose furthermore that the $M_\vartheta^{(i)}$'s do not depend on ϑ, and consider now (2.2). The quantity in the second set of braces in (2.2) is then constant in ϑ, and (2.2) simplifies to

$$\left[\frac{d\nu_1}{d\nu_2}\right](\vartheta, \psi) = C\left(\frac{M^{(1)}}{M^{(2)}}\right)^d \left[\frac{d\lambda^{(1)}}{d\lambda^{(2)}}\right](\vartheta),$$

where C is a constant. Assume also that $\lambda^{(1)} = \lambda^{(2)}$. Suppose $\theta_1^{(2)}, \ldots, \theta_n^{(2)}$ is Markov chain output from the posterior $\nu_{2,y}$. If $M^{(1)} > M^{(2)}$ and the $d(\psi_i)$'s are large, then the values of $(M^{(1)}/M^{(2)})^d$ are large, and from (1.5) we see that the estimate of the Bayes factor in favor of the model indexed by h_1 is large. This is as one would expect: the model indexed by h_2 expected more ties, but didn't see them, and so the model indexed by h_1 better explains the data. Theorem 1 may be used in a number of ways, and on occasion may be applied in the following kind of situation. Suppose we are considering models indexed by $h = (\{G_\vartheta\}_{\vartheta \in \Omega}, M, \lambda)$, where we wish to let M vary, and suppose that G_ϑ is not conjugate to $P_j(\psi_j)$, making implementation of MCMC more difficult. We may then consider a model where instead of using G_ϑ we use a more convenient parametric family $G_\vartheta^{(c)}$ which is conjugate to $P_j(\psi_j)$, and run a Markov chain under this model for some value $M^{(c)}$ of the precision parameter. Equation (2.2) becomes

$$\left\{\prod_{r=1}^d \frac{g_\vartheta(\psi_{(r)})}{g_\vartheta^{(c)}(\psi_{(r)})}\right\}\left(\frac{M}{M^{(c)}}\right)^d \left\{\frac{\Gamma(M)\Gamma\big(M^{(c)} + m\big)}{\Gamma(M^{(c)})\Gamma(M + m)}\right\}.$$

Care has to be exercised when the tails of G_ϑ are heavier than those of $G_\vartheta^{(c)}$, and we may use this as long as we ascertain that the simulation size is big enough to produce adequate accuracy in our estimates.

3 Consider now (2.3), and suppose for simplicity that the parametric families $\{G_\vartheta^{(1)}\}$ and

$\{G_\vartheta^{(2)}\}$ are the same, $\lambda^{(1)} = \lambda^{(2)}$, and $M_\vartheta^{(2)} \equiv M$. Then,

$$\left[\frac{d\nu_{\text{par},1}}{d\nu_2}\right](\vartheta, \psi) = \left(\prod_{j=1}^{m-1} \frac{M+j}{M}\right) I(d(\psi) = m). \tag{2.4}$$

We recognize $\prod_{j=1}^{m-1}((M+j)/M)$ as the reciprocal of the (prior) probability that the ψ_j's are distinct, under the model indexed by h_2. Suppose Markov chain output is generated from the posterior corresponding to the model indexed by h_2. The Bayes factor in favor of the parametric model indexed by h_1 is the expectation of (2.4) [see (1.4)], which is $\nu_{2,y}\{d = m\}/\nu_2\{d = m\}$. Therefore, the parametric model is favored if and only if in the Markov chain output the event $\{d = m\}$ occurs more often than was expected a priori under the model indexed by h_2.

As $M \to \infty$, $[d\nu_{\text{par},1}/d\nu_2]$ converges monotonically downwards towards 1 if $d(\psi) = m$. We see from (2.4) that for any M, $\nu_{\text{par},1} \ll \nu_2$, but the reverse is not true.

4 Special cases of Part (i) of Theorem 1 already exist in the literature. Doss (1994a) states without proof a version for the case of censored data (which corresponds to a nonparametric Bayesian model where the likelihood function is the indicator of a censoring set). Liu (1996) states the result in a simple case which permits an elegant short proof. The method of proof that we use relies on a martingale-based calculation of the Radon-Nikodym derivative. It has the advantage that it can be applied to other nonparametric priors, for example variants of the Dirichlet such as the "symmetrized Dirichlets" that were studied by Diaconis and Freedman (1986a,b) and "conditional Dirichlets" that were studied by Doss (1985a,b), Burr et al. (2003), and Burr and Doss (2005), and are used in the illustration of Section 3.2. We know of no other method that can do this.

2.2 Estimation of the Family of Bayes Factors

Consider the estimate (1.5), where $\theta_1, \dots, \theta_n$ is Markov chain output from $\nu_{h_1,y}$. This estimate will be asymptotically normal if

(i) the Markov chain mixes fast enough, and

(ii) the random variable $[d\nu_h/d\nu_{h_1}](\theta)$ (where $\theta \sim \nu_{h_1,y}$) has a high enough moment.

Weakening condition (i) requires strengthening condition (ii) and vice versa. For example, if the chain is uniformly ergodic, then we need only a second moment in (ii) (Cogburn 1972). If the chain is only geometrically ergodic then in (ii) we need a moment of order $2 + \epsilon$, for some $\epsilon > 0$ [Theorem 18.5.3 of Ibragimov and Linnik (1971)]. See Chan and Geyer (1994) for a discussion of various sets of assumptions for a central limit theorem to hold, and Geyer (1992) for a discussion of estimation of the variance.

Unfortunately, this estimate suffers a serious defect: unless h is close to h_1, ν_h can be nearly singular with respect to ν_{h_1} over the region where the θ_i's are likely to be, and the result is that although the estimate is consistent, it can have infinite variance. From a practical point of view, this means that there is effectively a "radius" around h_1 within which one can safely move.

Let $\mathcal{H}$ be the set of hyperparameter values over which h will vary. Buta and Doss (2010) discuss a method that involves selecting k hyperparameter points $h_1, \dots, h_k \in \mathcal{H}$, and for each $j \in \{1, \dots, k\}$ getting ergodic Markov chain samples $\theta_i^{(j)}$, $i = 1, \dots, n_j$ from the posterior $\nu_{h_j,y}$. Instead of $[d\nu_h/d\nu_{h_1}]$, the Radon-Nikodym derivative $\big[d\nu_h/d\big(\sum_{s=1}^k w_s\nu_{h_s}\big)\big]$ is used, where $w_1, \dots, w_k > 0$ are appropriately chosen. The w_s's need not add up to 1. Let $n = \sum_{j=1}^k n_j$ be the total sample size, let $a_j = n_j/n$, and recall that $\ell_y(\cdot)$ is the likelihood function. If we took $w_s = a_s\, m_{h_1}/m_{h_s}$, we would have

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\frac{d\nu_h}{d\big(\sum_{s=1}^k w_s\nu_{h_s}\big)}\right](\theta_i^{(j)}) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\frac{d(\ell_y\nu_h)}{d\big(\sum_{s=1}^k w_s\ell_y\nu_{h_s}\big)}\right](\theta_i^{(j)}) \tag{2.5a}$$

$$= \frac{m_h}{m_{h_1}}\sum_{j=1}^{k}\frac{1}{n_j}\sum_{i=1}^{n_j} a_j\left[\frac{d(\ell_y\nu_h/m_h)}{d\big(\sum_{s=1}^k a_s\ell_y\nu_{h_s}/m_{h_s}\big)}\right](\theta_i^{(j)})$$

$$\xrightarrow{\text{a.s.}} \frac{m_h}{m_{h_1}}\sum_{j=1}^{k} a_j\int\left[\frac{d\nu_{h,y}}{d\big(\sum_{s=1}^k a_s\nu_{h_s,y}\big)}\right]d\nu_{h_j,y} = \frac{m_h}{m_{h_1}}. \tag{2.5b}$$

The "$\xrightarrow{\text{a.s.}}$" in (2.5b) indicates convergence in the almost sure sense. The ratios m_{h_1}/m_{h_s} are needed to form the w_s's, but in general these ratios are unknown, and must be estimated.

This estimation problem may be stated as follows. For $j = 1, \dots, k$, we have samples $\theta_i^{(j)}$, $i = 1, \dots, n_j$ from densities f_j, where $f_j = h_j/c_j$, and the h_j's are known functions, but the normalizing constants c_j are unknown. We wish to simultaneously estimate all ratios c_l/c_s, $l, s = 1, \dots, k$. This problem has been considered by many authors (Gill et al. 1988; Meng and Wong 1996; Kong et al. 2003; Tan 2004). Let r be the vector consisting of the k ratios $r = (r_1, \dots, r_k) = (c_1/c_1, \dots, c_1/c_k)$. Gill et al. (1988) show that an estimate of r may be obtained as the solution of the system of k equations

$$(r_l)^{-1} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\frac{q_l(\theta_i^{(j)})}{\sum_{s=1}^k a_s q_s(\theta_i^{(j)})r_s}, \qquad l = 1, \dots, k, \tag{2.6}$$

where, as before, $n = \sum_{j=1}^k n_j$ and $a_j = n_j/n$. They showed that under the assumption that the samples are iid, if $\hat{r}$ is the solution to (2.6), then $n^{1/2}(\hat{r} - r)$ is asymptotically normal.

Meng and Wong (1996), Kong et al. (2003), and Tan (2004) also considered this problem, and arrived at exactly the same estimate, although from rather different perspectives. Geyer (1994) established asymptotic normality of $\hat{r}$ when for each j, $\theta_i^{(j)}$, $i = 1, \dots, n_j$ are Markov chains satisfying certain mixing conditions. Our situation conforms to this setup. We can generate samples from $\nu_{h_j,y}$, the normalizing constants m_{h_j} are the marginal likelihoods $m_{h_j}(y) = \int \ell_y(\theta)\, \nu_{h_j}(d\theta)$, and we know the Radon-Nikodym derivatives $[d\nu_{h_l}/d\nu_{h_s}]$ for all l, s. (From inspection of (2.6), we see that the q_j's appear only through the ratios q_l/q_s.) Thus, we can use the method of Gill et al. (1988) to arrive at an estimate of the needed ratios m_{h_1}/m_{h_s}.

We now return to the quantity on the left side of (2.5a). Buta and Doss (2010) propose a two-stage procedure, whereby in Stage 1, for $j = 1, \dots, k$, we obtain Markov chain samples from $\nu_{h_j,y}$ and use these to estimate the ratios required in forming the w_s's, and in Stage 2, we generate new and independent samples and use those to calculate the quantity on the left side of (2.5a), using the w_s's formed in Stage 1. Call this estimate $\hat{B}(h, h_1)$. Buta and Doss (2010) show that under the assumption that the chains satisfy certain mixing conditions, this estimate is asymptotically normal, and they also show how to estimate its asymptotic variance. Viewed as a function of h, the estimate is far more stable (i.e. has a smaller variance for a wide range of h's). All the examples in this paper use this estimate.

Part (ii) of Theorem 1 states that if $h_1 = (v_1, M_1)$, where $M_1 < \infty$ and $h = (v, \infty)$, then $\nu_h \ll \nu_{h_1}$. But as mentioned earlier, the reverse is not true. Therefore, if we obtain a sample from the posterior corresponding to the nonparametric model (2.1), we can estimate $B(h, h_1)$ even if h corresponds to the parametric version of this model. More generally, we may use a skeleton set $h_1 = (v_1, M_1), \dots, h_k = (v_k, M_k)$ as long as there exists at least one $j \in \{1, \dots, k\}$ such that $M_j < \infty$.

3 Examples

Here we consider two examples. The distribution of latent variables is now routinely modelled through Dirichlet process mixtures, often with no justification (formal or informal), and our first example, which is simple and short, is included primarily to give an idea of how far from a normal or a t the distribution of the estimates of the latent variables has to be in order to justify using a mixture of Dirichlets. The second example is a bit more complex—it involves a variant of the Dirichlet process—and we include it to illustrate the generality of the formula for the Radon-Nikodym derivative given in Theorem 1 (and of the method used to obtain this formula). Both examples are "small scale," and the Bayes factors do not speak as loudly as in other data sets we have considered. The advantage of using these particular small data sets is that the data can be visualized more easily and this enables us to gain some insight on why the Bayes factors point to certain models.

3.1 Meta-Analysis of Studies on Decontamination of the Digestive Tract

Infections acquired in intensive care units are an important cause of mortality. One strategy for dealing with this problem involves selective decontamination of the digestive tract (henceforth DDT) through the use of antibiotics. This is designed to prevent infection by preventing carriage of potentially pathogenic micro-organisms from the oropharynx, stomach, and gut. A large number of randomized controlled trials have been carried out to investigate the potential benefits of this strategy. In each trial, patients in an intensive care unit were randomized to either a treatment or a control group. The proportion of individuals who acquired an infection was recorded for the treatment and control groups and an odds ratio was reported. An international collaborative group (Selective Decontamination of the Digestive Tract Trialists' Collaborative Group 1993) performed a meta-analysis of the 22 studies that were published during the period January 1984 to June 1992 using a fixed effects model, in which all the trials were assumed to be measuring the same quantity. However, the studies varied in many ways, for example in the kind of antibiotic used and the patient pool, and a new meta-analysis was carried out by Smith et al. (1995), using a random effects model.

Figure 1(a) displays the data for the 22 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. Also given by the figure is an estimate of the distribution of the study-specific log odds ratios, using a kernel density estimate that is based on the observed log odds ratios, with weights that reflect the estimated standard errors. (This density estimate should be viewed with caution, since it is based on the estimated log odds ratios, and not the log odds ratios themselves.) The figure suggests a non-normality, since it shows that there are a few studies with an extremely significant treatment effect. To better accommodate these outlying studies, Smith et al. (1995) suggest modelling the distribution of the random effects by a t distribution with 4 degrees of freedom.

It is natural to ask whether the apparent deviation from normality is strong enough to justify using a t distribution, and if it is, then what should be the degrees of freedom parameter. Actually, it is not clear that a t distribution would make for a good fit, and a Dirichlet-based model might be a better candidate. Therefore, we will consider the broader class of models given by (1.2) and use our methodology to select the two parameters v and M. The model used

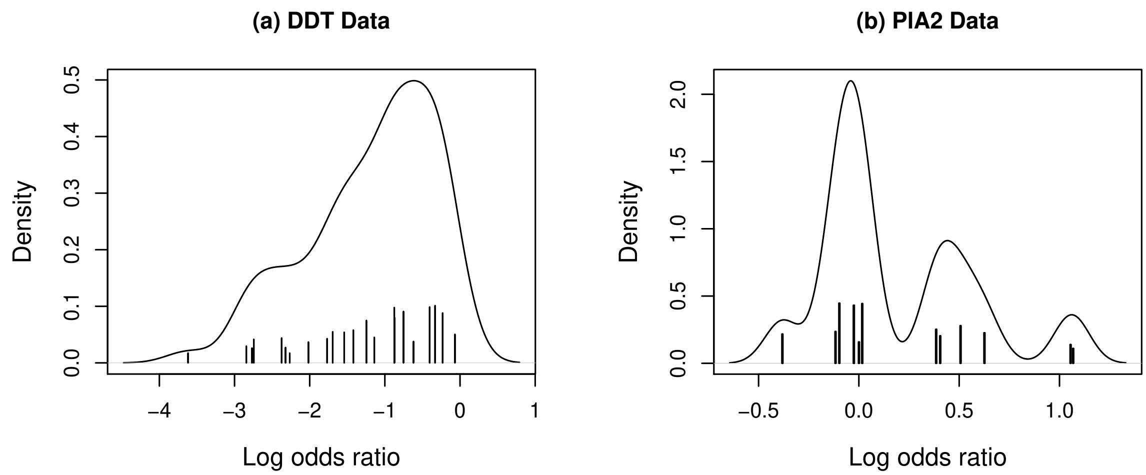

Figure 1: Estimate of distribution of the study-specific effect ψ for the DDT data and PlA2 data. Data are represented by vertical lines, whose locations are the estimates of the log odds ratios and whose heights are proportional to the reciprocals of the estimated standard errors.

by Smith et al. (1995) corresponds to the choice $(v, M) = (4, \infty)$.

Let Y_j denote the observed log odds ratio for study j, let σ_j be the corresponding standard error, and let ψ_j be the true log odds ratio, i.e. the log odds ratio that would be observed if the sample size for study j was infinite. By the well-known asymptotic normality of the observed log odds ratio, we see that (2.1a) is satisfied, where $P_j(\psi_j)$ is just the $\mathcal{N}(\psi_j, \sigma_j^2)$ distribution, and we may therefore use model (2.1). Let NIG(c_1, c_2, c_3, c_4) be the normal / inverse gamma distribution on (μ, τ) with parameters c_1, c_2, c_3, c_4, and let $m(v, M)$ be the marginal likelihood of the data when in model (2.1) we have the location/scale family of t distributions with v degrees of freedom (which gives rise to a non-conjugate model), precision parameter M, and $\lambda =$ NIG$(.1, .1, 0, 1000)$. This choice of λ gives a fairly diffuse prior on (μ, τ). For unity of notation, we will use the convention that a t distribution with infinite degrees of freedom is just the normal distribution. Also, $m(t_v, \infty)$ will denote the marginal likelihood of the data under the parametric version of the model. Define the Bayes factor

$$B(v, M) = \frac{m(v, M)}{m(\infty, 16)}, \qquad v, M \in (0, \infty].$$

We are interested in estimating $B(v, M)$ as v and M vary.

To this end, we used the estimate described in Section 2.2 in order to achieve greater stability. Specifically, we used chains of length $100{,}000$ each from the posterior corresponding to model (2.1) where G_ϑ in (2.1c) is the normal family, with M starting at 1 and increasing by factors of 2 up to 128, and also one chain of length 100,000 corresponding to the parametric version of the model, all using $\lambda = \text{NIG}(.1, .1, 0, 1000)$, for a total of nine chains. These nine chains form the Stage 1 samples which are used to form the estimate $\hat{r}$ of the vector $r = (r_1, \dots, r_9)$ discussed in Section 2.2. We then ran nine new chains, each of length 10,000, corresponding to the same skeleton grid, to form the quantity on the left side of (2.5a), via Theorem 1, using the estimate r computed in Stage 1, to estimate $B(v, M)$. (Note that we are taking $h_1 = (\infty, 16)$ to be the baseline hyperparameter value.) The estimate $\hat{r}$ obtained in Stage 1 is computed only once; by contrast, the estimate on the left side of (2.5a) needs to be computed for each value of h. So the sample sizes in Stage 1 can be taken to be quite large, whereas those in Stage 2 must be relatively small. This is the reason why we used chains of length 100,000 in Stage 1 but only 10,000 in Stage 2.

Figure 2 shows that the Bayes factors are highest for the Dirichlet centered at the normal distribution (the maximum turns out to be achieved when $M = 15$), and the difference between the Bayes factor for the Dirichlet with $M = 15$ and that for the Dirichlet with $M = \infty$ is so slight that there is hardly a justification for using a nonparametric Bayes model. The simple normal distribution suffices, a useful piece of information that would have been difficult to obtain without this formal analysis.

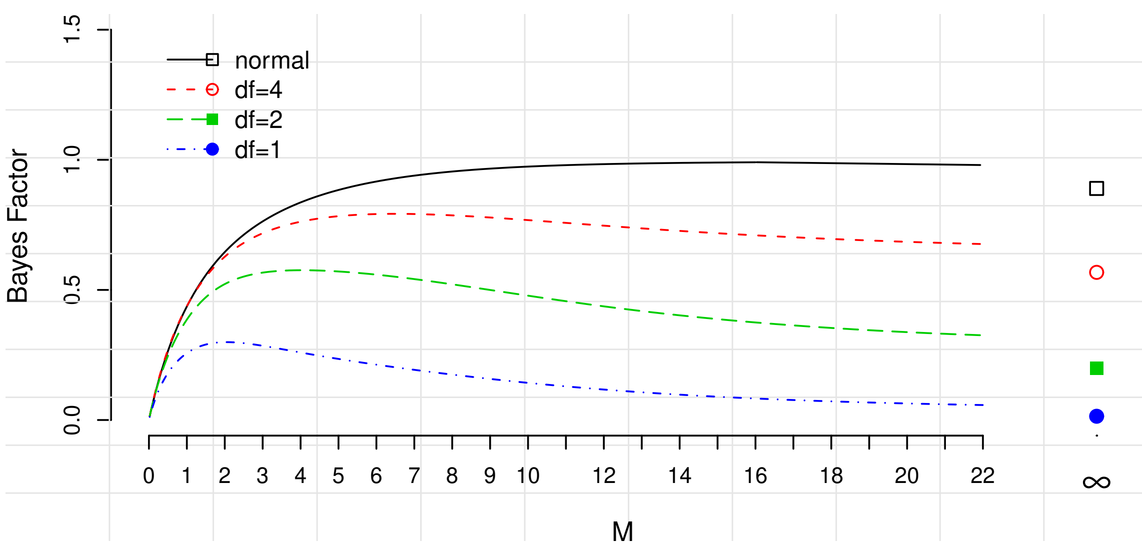

Figure 2: Model assessment for the DDT data. Shown are plots of Bayes factors for Dirichlet models centered at the location/scale families of normal and t distributions with 1, 2, and 4 degrees of freedom, as M varies.

We carried out side calculations that show that the t_1 and t_2 distributions give estimates of the mean of the predictive distribution of a future study that are quite a bit larger than estimates based on the normal, but Figure 2 shows that these estimates are not appropriate. Interestingly, the t_1 distribution provides such a bad fit ($B(1, \infty) = .014$) that using a Dirichlet with small M (around 2) improves the fit by allowing enough deviation from this t.

Since the apparent deviation from normality in Figure 1 is not strong enough to warrant using a model more complicated than a simple normal, we created a hypothetical data set in which the standard error of the most outlying study (for which $Y_j = -3.62$) is decreased by a factor of 5. Figure 3 shows the new Bayes factors, and these now indicate that the most appropriate model is the one based on the t_4 distribution (without involvement of Dirichlet processes).

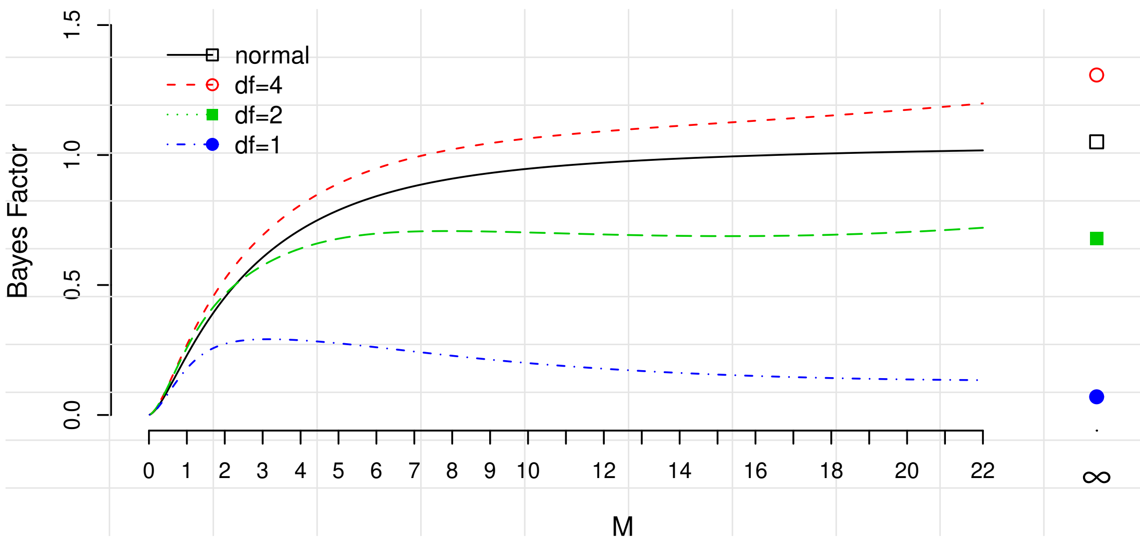

Figure 3: Model assessment for perturbed version of the DDT data in which the standard error for the observation with the smallest log odds ratio is decreased by a factor of 5.

3.2 Meta-Analysis of Data on PlA2 Polymorphism and Risk of Heart Disease

We revisit the data set analyzed by Burr et al. (2003). Very briefly, this involved a meta-analysis of studies that investigate a purported link between presence of a certain genetic trait "PlA2 polymorphism" and increased risk of coronary heart disease (CHD). There were 12 studies, each of which was a case-control study. The cases were individuals with CHD and the controls were individuals with no history of CHD, and the exposure variable was presence/absence of the PlA2 polymorphism. The studies reported an odds ratio, together with a standard error. We work on the log scale, and the framework is the same as for the DDT example. (A log odds ratio greater than 0 indicates that the polymorphism is associated with increased risk of CHD.) The data set is presented in detail in Burr et al. (2003), but the bottom part of Figure 1(b) enables a quick visual scan of the data. Because the studies gave wildly conflicting results, Burr et al. (2003) were primarily interested in the basic question of whether the mean of the distribution of the study-specific effects was greater than 0, for example to determine whether or not one should carry out further studies.

In a parametric model such as (1.1) the parameter μ has a well-defined role as the mean of the distribution of the study-specific effects, whereas in (1.2), μ is not equal to the mean of F. For this reason, Burr et al. (2003) consider model (1.2) (with a normal distribution in line (1.2c)) and replace (1.2c) with

$$\text{conditional on } \mu, \tau, \quad F \sim \mathcal{D}^{\mu}_{M\mathcal{N}(\mu,\tau^2)} \tag{1.2c$'$}$$

where $\mathcal{D}^{\mu}_{M\mathcal{N}(\mu,\tau^2)}$ is a Dirichlet conditioned on $\text{median}(F) = \mu$. In this model, the parameter μ has a well-defined role as the median of F (by construction). Additionally, the distribution of μ in this model is less affected by outliers than is the distribution of the mean of F in the model where we use ordinary (i.e. "unconditional") Dirichlets [this last is hard to handle even with Sethuraman's (1994) construction of the Dirichlet process]. Burr et al. (2003) argue that the model based on mixtures of conditional Dirichlets is therefore useful in situations where there is concern that a few poorly designed studies might have undue influence on the meta-analysis. Here we will consider model (1.2), with (1.2c) replaced by (1.2c$'$), i.e. using conditional Dirichlets, and we will in addition allow for the location-scale family to be a family of t distributions with v degrees of freedom, with $v = \infty$ corresponding to the normal family. We will consider the choice of M and v.

For $h = (v, M)$, let ν_h denote the (prior) distribution of (ψ, μ, τ) under model (1.2), where in (1.2d) $\lambda_c = \text{NIG}(.1, .1, 0, 1000)$, and let $m(v, M)$ denote the marginal likelihood of the data under this model. Also, let ν_h^c and $m^c(v, M)$ denote the corresponding quantities for the model involving conditional Dirichlets, i.e. where instead of (1.2c) we have $F \sim \mathcal{D}^{\mu}_{Mt_{v,\mu,\tau}}$ (Dirichlet conditioned on $\text{med}(F) = \mu$). We would like to estimate the Bayes factor

$$\frac{m^c(v, M)}{m^c(v_1, M_1)} \qquad v, M \in (0, \infty] \tag{3.1}$$

for some fixed $h_1 = (v_1, M_1)$, but it turns out to be equally useful, and far more convenient, to estimate the Bayes factor

$$\frac{m^c(v, M)}{m(v_1, M_1)} \qquad v, M \in (0, \infty]. \tag{3.2}$$

The family of Bayes factors equation (3.1) and (3.2) differ by a multiplicative constant, and as mentioned earlier, from the point of view of model selection, multiplicative factors are immaterial.

Estimating (3.2) is done very conveniently if we have a Markov chain sample $\theta_1, \ldots, \theta_n$ from the posterior $\nu_{h_1,y}$, which corresponds to the prior ν_{h_1} (unconditional Dirichlet) *and* we have a formula for the Radon-Nikodym derivative $[d\nu_h^c/d\nu_{h_1}]$, for in this case,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d\nu_h^c}{d\nu_{h_1}}\right](\theta_i) \xrightarrow{\text{a.s.}} \frac{m^c(v, M)}{m(v_1, M_1)}$$

(see (1.4)). In the appendix, we prove that

$$\left[\frac{d\nu_h^c}{d\nu_{h_1}}\right](\psi, \mu, \tau) = \left\{\prod_{r=1}^{d}\frac{t_v\left(\frac{\psi_{(r)}-\mu}{\tau}\right)}{t_{v_1}\left(\frac{\psi_{(r)}-\mu}{\tau}\right)}\right\}\left(\frac{M}{M_1}\right)^d \left\{\frac{\Gamma(M_1+m)\Gamma^2\left(\frac{M}{2}\right)}{\Gamma(M_1)2^m}\right\} K(\psi, \mu), \qquad (3.3a)$$

where

$$K(\psi, \mu) = \left[\Gamma\left(M/2 + \textstyle\sum_{j=1}^m I(\psi_j < \mu)\right)\Gamma\left(M/2 + \textstyle\sum_{j=1}^m I(\psi_j > \mu)\right)\right]^{-1}. \qquad (3.3b)$$

(We do not need to write a corresponding formula for $[d\nu_{\text{par},h}^c/d\nu_{h_1}]$ since it is clear that $[d\nu_{\text{par},h}^c/d\nu_{h_1}] = [d\nu_{\text{par},h}/d\nu_{h_1}]$.)

The function K is interesting: viewed as a function of μ, $K(\psi, \mu)$ has a maximum when μ is at the median of the ψ_j's, and as μ moves away from the median in either direction, it is constant between the ψ_j's, and decreases by jumps at each ψ_j. This effect is stronger when M is small. The interpretation is as follows. Suppose that $\theta^{(i)} = (\psi^{(i)}, \mu^{(i)}, \tau^{(i)})$, $i = 1, \ldots, n$ is Markov chain output from the posterior corresponding to model (1.2) (mixture of ordinary Dirichlets). If the $\theta^{(i)}$'s are such that $\mu^{(i)}$ is close to the median of $(\psi_1^{(i)}, \ldots, \psi_m^{(i)})$, which is likely under the model based on conditional Dirichlets (since μ is the median of F, and is therefore expected to be close to the median of the ψ_j's, as these are a sample from F), then the estimate of the Bayes factor in favor of the model based on conditional Dirichlets is large.

As before, an estimate based on several Markov chains enables us to reliably estimate the Bayes factors over a wider range of values of h, and this is what we will use. We ran Markov chains exactly as in Section 3.1 (i.e. under the model that involves ordinary Dirichlets as opposed to conditional Dirichlets, and with the same λ) except that we used values of M starting at $1/4$ and increasing by factors of 2 up to 32. Define the Bayes factor

$$B^c(v, M) = \frac{m^c(v, M)}{m(\infty, 4)}. \qquad (3.4)$$

Figure 4 gives a plot of the Bayes factor (3.4) as M varies, for $v = 1, 2, 4, \infty$ (bottom four lines). Note that the denominator in (3.4) refers to a point outside of the model. We may write

$$\frac{m^c(v, M)}{m^c(\infty, 4)} = B^c(v, M)\frac{m(\infty, 4)}{m^c(\infty, 4)} \tag{3.5}$$

and note that the second term on the right side of (3.5) is available from Figure 4. So we may rescale the plots in Figure 4 to get estimates of $m^c(v, M)/m^c(\infty, 4)$, but as mentioned earlier, there is no need to do so. Figure 4 suggests that the model based on conditional Dirichlets centered at the family of normal distributions provides a better fit than does the parametric model. For example, $m^c(\infty, 1/4)/m^c(\infty, \infty) = 3.16$ which, according to the scale of Jeffreys, is "substantial" evidence in favor of the nonparametric model with $M = 1/4$ as compared with the parametric model. In view of the usual aversion to using very small values of M, Figure 4 suggests that it is reasonable to use the value $M = 1$. This is the value used by Burr et al. (2003) (with no justification), who show that the analysis based on this choice gives conclusions that are more conservative than those based on a parametric model.

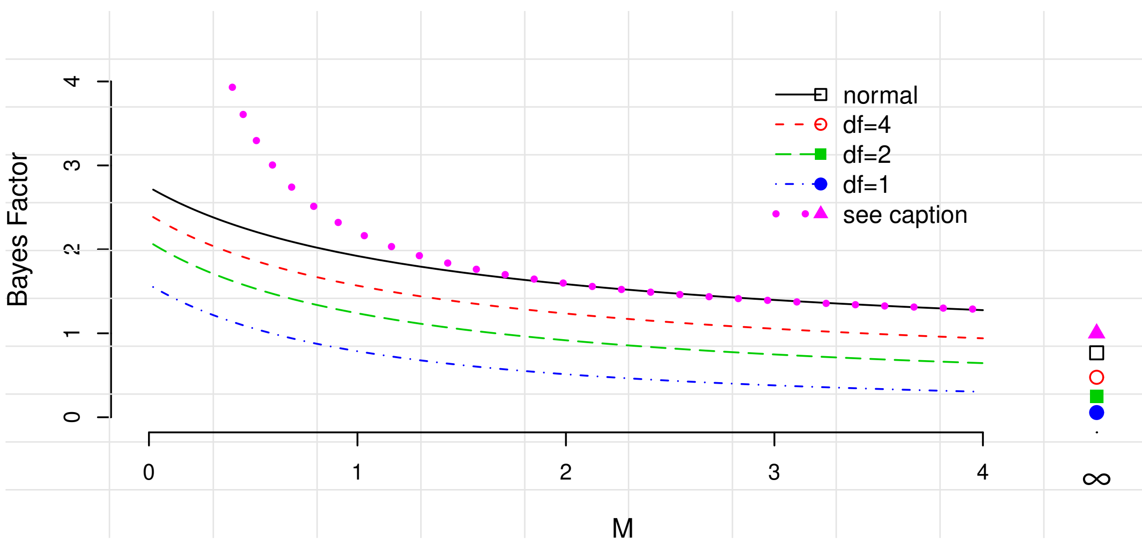

Figure 4: Model assessment for the PlA2 data. Bottom four lines are plots of Bayes factors for the conditional Dirichlet models centered at the location/scale families of normal and t distributions with 1, 2, and 4 degrees of freedom, as M varies, relative to the Dirichlet models centered at the normal family with $M = 4$. Top line is a plot of the Bayes factor for the conditional Dirichlet vs. Dirichlet models, with M starting at .4.

Figures 2 and 3 suggest that for the DDT data there is no need for a Bayesian nonparametric model, while 4 suggests that for the PlA2 data there is. A look at Figure 1 reveals why. Even if we form the perturbed version of the DDT data, the non-normality is essentially due to outliers, and these are handled by a t distribution. In contrast, Figure 1(b) suggests that for the PlA2 data we have a tri-modal distribution; using a t distribution doesn't address this problem, whereas using a nonparametric Dirichlet-based model with a small value of M does. (In experiments not shown here, we artificially increased the "clumping" by merging together the observations into three groups and then moved the three groups apart from each other. The Bayes factors then speak more strongly in favor of the nonparametric Bayes model and in fact the plots need to be shown on a log scale.)

The analysis involving the model based on conditional Dirichlets would be incomplete if we didn't see whether this model provides a good fit. The top line in Figure 4 is a plot of the Bayes factor for the conditional Dirichlet vs. the Dirichlet models as M varies, i.e. a plot of $m^c(v, M)/m(v, M)$. This quantity does not depend on v. It may be estimated via (3.3) separately for each M, but we may estimate it for a range of M's using the Markov chains run for $M = 1/4, 1/2, \ldots, 32$, using the methods of this paper. The plot shows that for this data set, the model based on conditional Dirichlets provides a uniformly better fit than does the one based on Dirichlets, with the advantage being greater for small values of M. (Of course, as $M \to \infty$, the Bayes factor converges to 1, since the models both converge to the parametric model.)

`R` functions for calculating the Radon-Nikodym derivatives, the Bayes factor estimates, and for producing plots such as those in Figures 2–4 are available from the author upon request. There are many programs that run the Markov chains for Dirichlet-based models already in existence [see Neal (2000) and Jain and Neal (2007)], and the present author's implementation is also available upon request. Functions for making plots of Bayes factors of conditional Dirichlet vs. unconditional Dirichlet models are given as part of the `R` package `bspmma` (Burr 2010).

4 Discussion and Relation to Other Approaches

Choice of the Design Points Buta and Doss (2010) show how to estimate the variance of the Bayes factor estimate discussed in Section 2.2. This variance depends on the choice of the hyperparameters $h_1, \ldots, h_k$. To emphasize this dependence, let us denote it by $V(h, h_1, \ldots, h_k)$. In the models we are considering, $h = (v, M, c)$ and we will typically be interested primarily in the degrees of freedom parameter v and the precision parameter M. So by a slight abuse of notation, let us write $h = (v, M)$, i.e. we ignore the hyperparameter of the prior on ϑ. If we fix a range over which h is to vary, e.g. $\mathcal{H} = [a, \infty) \times [b, \infty)$ where $a, b > 0$, and fix k, then we will face the problem below.

Design Problem Find the values of $h_1, \ldots, h_k$ that minimize $\max_{h \in \mathcal{H}} V(h, h_1, \ldots, h_k)$.

Unfortunately, it is not possible to calculate $V(h, h_1, \ldots, h_k)$ analytically—even if $k = 1$, and even if h and h_1 differ only in the precision parameter M—let alone minimize it with respect to $2k$ variables, and solving the design problem is hopeless.

In our experience, we have found that the following method works reasonably well. Having specified the range $\mathcal{H}$, we select trial values $h_1, \ldots, h_k$ and plot the estimated variance as a function of h, using the variance estimate obtained in Buta and Doss (2010). If we find a region in $\mathcal{H}$ where this variance is unacceptably large, we "cover" this region by moving some h_l's closer to the region, or by simply adding new h_l's in that region, which increases k. Of course, we note that a region of the form $[a, \infty) \times [b, \infty)$ is not "unbounded" in the practical sense, as it is easy to see that $B(h, h_1)$ converges as $d \to \infty$ or $M \to \infty$, or both.

Consequences of Extrapolation Beyond Coverage Provided by the Design Points We now give some indication of the extent to which the variability of the Bayes factor curves, say those in Figure 2, depends on the choice of the points $h_1, \ldots, h_k$. Suppose we wish to estimate $B = m(\infty, \infty)/m(\infty, M)$, i.e. the Bayes factor of the parametric normal model vs. the Dirichlet model centered at the normal family, with precision parameter M, and suppose we run a single chain under the latter model. Let ν_M and $\nu_{M,y}$ denote the prior and posterior distributions, respectively, of θ under this model. As mentioned in Comment 3 of Section 2.1, the Bayes factor is $\nu_{M,y}(d = 22)/\nu_M(d = 22)$, and its estimate is $\hat{\nu}_{M,y}(d = 22)/\nu_M(d = 22)$, where $\hat{\nu}_{M,y}(d = 22)$ is the observed proportion of θ's for which the ψ_j's are all distinct. If we had an iid sample of size n from the posterior (a best-case scenario), the variance of this estimate would be

$$\frac{\nu_{M,y}(d = 22)(1 - \nu_{M,y}(d = 22))}{n\,(\nu_M(d = 22))^2}. \tag{4.1}$$

From accurate experiments involving Markov chains run under models with a wide range of M's, we know that the Bayes factor is about 1 for $M \geq 7$ (see Figure 2), i.e.

$$\nu_{M,y}(d = 22) \doteq \nu_M(d = 22) = \prod_{j=1}^{21} (M/(M + j)), \qquad M \geq 7,$$

so that the variance in (4.1) essentially reduces to $(1/n) \prod_{j=1}^{21} ((M + j)/M)$ for this range of M's [the term $(1 - \nu_{M,y}(d = 22))$ is nearly 1 for the range of M's we are interested in]. If $M = 7$, this is about $(1.1 \times 10^8)/n$, and if $M = 15$, this is about $(5.7 \times 10^4)/n$, and only for $M \geq 60$ do we get reasonable numbers. The point of this discussion is to stress that when the prior ν_h is nearly singular with respect to all ν_{h_s}, $s = 1, \dots, k$, over the region where the $\theta_i^{(s)}$'s are likely to be, the estimate will be unstable. In our particular case this means that it is essential that we run chains under a wide spectrum for the precision parameter M, if we want to produce accurate estimates of the Bayes factor over a wide range of hyperparameter values.

Relation to Previous Work One approach for dealing with the choice of the precision parameter M is to simply put a prior on it, as in the early paper by West (1992). He considers a model of the form (2.1), except with a single Dirichlet, as opposed to a mixture of Dirichlets [i.e. the hyperparameter is $h = M$, as opposed to $h = (M, v, c)$]. He uses a gamma prior on M, which itself is indexed by two parameters, which must then be specified. One can consider instead a flat prior, in which case the posterior distribution of M is proportional to the marginal likelihood of the data, i.e. $m_h(y)$, and the mode of the posterior is then the point at which the Bayes factor is maximized. But as is well known, for certain parameters, flat priors can be very informative. Here, putting a flat prior on M and v in effect skews the results in favor of $M = \infty$ (the parametric version of the model), and in favor of $v = \infty$ (the normal distribution). Thus, while the approach of putting a prior on the hyperparameters can be useful, there are problems with it. One is that, as mentioned above, the choice of prior can have great influence on the analysis. Another is that, in broad terms, the general interest in empirical Bayes methods arises in part from a desire to select specific values of the hyperparameters because these give a model that is more parsimonious and interpretable. These points are discussed more fully (in a general context) in George and Foster (2000) and Robert (2001, Chapter 7).

Estimation of Bayes factors for nonparametric Bayes problems has been considered by several other authors. Berger and Guglielmi (2001) consider the case where the prior is a mixture of Polya trees. Polya trees can offer more modelling flexibility than do Dirichlet priors in that their specification involves parameters that control the smoothness of the random distribution. Berger and Guglielmi (2001) consider what is in essence the "case of complete data;" that is, they consider the case where there are observations $\psi_1, \dots, \psi_j$ which are iid from a distribution F whose distribution is a mixture of Polya trees. Our situation [model (2.1)] is different: although we assume that $\psi_1, \dots, \psi_j \overset{\text{iid}}{\sim} F$, in our model we don't observe the ψ_j's. Rather, for each j, we observe a random variable Y_j which gives us partial information on ψ_j. (The case where the distribution P_j in (2.1a) is degenerate at ψ_j would reduce to the "complete data problem.") The present author does not know if the calculation of Radon-Nikodym derivatives can be extended to hierarchical models involving mixtures of Polya trees.

Basu and Chib (2003) consider a Dirichlet-based hierarchical model similar to ours. They consider the situation where there are specifications $h_1, \dots, h_k$ of the model and they wish to calculate Bayes factors for all possible pairs. Their method involves calculating the marginal likelihood for each model. Casting their approach into our framework and notation to facilitate our description of their idea, in their approach they express the marginal likelihood $m_{h_j}(y)$ via the identity

$$\log m_{h_j}(y) = \log m_{h_j}(y \mid \vartheta^*) + \log \lambda^{(j)}(\vartheta^*) - \log \lambda_y^{(j)}(\vartheta^*). \tag{4.2}$$

Here, $\lambda^{(j)}$ and $\lambda_y^{(j)}$ are the prior and posterior densities of ϑ. In (4.2), ϑ^* is selected as a point of high posterior density. They estimate the first term using a sequential importance sampling scheme and the third term using the output of a Markov chain run under the specification h_j. Their approach has the advantage that they can actually estimate the marginal likelihood. On the other hand, it requires running a separate Markov chain for each model under consideration. In our approach, once Markov chains have been run under the models indexed by $h_1, \dots, h_k$, we may estimate Bayes factors for a continuum of indexing values h as long as these are not too far from all the h_j's. Another advantage of our approach is that estimation of ratios of normalizing constants tends to be far more stable than estimation of the normalizing constants themselves. For example, if we wish to estimate m_h/m_{h_1}, then a procedure that involves estimating m_h and m_{h_1} separately and then taking the ratio is not guaranteed to provide accurate estimates even when $h = h_1$, whereas in this case the simple estimate (1.5) gives an unbiased estimate with zero variance.

Acknowledgements

I am grateful to the reviewers, whose careful reading and constructive criticism improved the paper. This work was supported by NSF Grant DMS-08-05860.

Appendix

Appendix A: Proof of Theorem 1

We first prove Part (i). Let $\vartheta^{(0)} \in \Omega$ and $\psi^{(0)} = (\psi_1^{(0)}, \dots, \psi_m^{(0)}) \in \mathbb{R}^m$ be fixed. For $\eta > 0$, let $C_{\psi^{(0)}}^{\eta}$ be the cube

$$C_{\psi^{(0)}}^{\eta} = (\psi_1^{(0)} - \eta/2, \psi_1^{(0)} + \eta/2) \times \cdots \times (\psi_m^{(0)} - \eta/2, \psi_m^{(0)} + \eta/2),$$

and similarly let $B^{\eta}_{\vartheta^{(0)}}$ be the cube in $\mathbb{R}^p$ centered at $\vartheta^{(0)}$ and with sides of width η. Our plan is to calculate the probability of the set $\bigl\{\vartheta \in B^{\eta}_{\vartheta^{(0)}},\, \psi \in C^{\eta}_{\psi^{(0)}}\bigr\}$ under the two distributions ν_1 and ν_2, and take the ratio. The limit as $\eta \to 0$ is the Radon-Nikodym derivative at $(\vartheta^{(0)}, \psi^{(0)})$, by a martingale argument [see, e.g. pp. 209–210 of Durrett (1991)].

Let π_i be the distribution of (ϑ, F, ψ) under the model indexed by h_i, for $i = 1, 2$. Let $\psi^{(0)}_{(1)} < \psi^{(0)}_{(2)} < \cdots < \psi^{(0)}_{(d)}$ be the distinct values of $\psi^{(0)}_1, \ldots, \psi^{(0)}_m$, and let $m_1, \ldots, m_d$ be their multiplicities. Denoting the set of all probability measures on $\mathbb{R}$ by $\mathcal{P}$, we have

$$\frac{\nu_1\bigl\{\vartheta \in B^{\eta}_{\vartheta^{(0)}},\, \psi \in C^{\eta}_{\psi^{(0)}}\bigr\}}{\nu_2\bigl\{\vartheta \in B^{\eta}_{\vartheta^{(0)}},\, \psi \in C^{\eta}_{\psi^{(0)}}\bigr\}} = \frac{\pi_1\bigl\{\vartheta \in B^{\eta}_{\vartheta^{(0)}},\, F \in \mathcal{P},\, \psi \in C^{\eta}_{\psi^{(0)}}\bigr\}}{\pi_2\bigl\{\vartheta \in B^{\eta}_{\vartheta^{(0)}},\, F \in \mathcal{P},\, \psi \in C^{\eta}_{\psi^{(0)}}\bigr\}} \tag{A.1a}$$

$$= \frac{\displaystyle\int_{B^{\eta}_{\vartheta^{(0)}}} \int_{\mathcal{P}} \prod_{j=1}^{m} \bigl[F(\psi^{(0)}_j + \eta/2) - F(\psi^{(0)}_j - \eta/2)\bigr]\, \mathcal{D}_{M^{(1)}_{\vartheta} G^{(1)}_{\vartheta}}(dF)\, \lambda^{(1)}(d\vartheta)}{\displaystyle\int_{B^{\eta}_{\vartheta^{(0)}}} \int_{\mathcal{P}} \prod_{j=1}^{m} \bigl[F(\psi^{(0)}_j + \eta/2) - F(\psi^{(0)}_j - \eta/2)\bigr]\, \mathcal{D}_{M^{(2)}_{\vartheta} G^{(2)}_{\vartheta}}(dF)\, \lambda^{(2)}(d\vartheta)}$$

$$= \frac{\displaystyle\int_{B^{\eta}_{\vartheta^{(0)}}} \frac{\int_{\mathcal{P}} \prod_{j=1}^{m} \bigl[F(\psi^{(0)}_j + \eta/2) - F(\psi^{(0)}_j - \eta/2)\bigr]\, \mathcal{D}_{M^{(1)}_{\vartheta} G^{(1)}_{\vartheta}}(dF)}{\eta^d \prod_{l=1}^{d}(m_l - 1)!}\, \lambda^{(1)}(d\vartheta)}{\displaystyle\int_{B^{\eta}_{\vartheta^{(0)}}} \frac{\int_{\mathcal{P}} \prod_{j=1}^{m} \bigl[F(\psi^{(0)}_j + \eta/2) - F(\psi^{(0)}_j - \eta/2)\bigr]\, \mathcal{D}_{M^{(2)}_{\vartheta} G^{(2)}_{\vartheta}}(dF)}{\eta^d \prod_{l=1}^{d}(m_l - 1)!}\, \lambda^{(2)}(d\vartheta)}$$

$$= \frac{\displaystyle\int_{B^{\eta}_{\vartheta^{(0)}}} f^{1,\eta}_{\psi^{(0)}}(\vartheta)\, \lambda^{(1)}(d\vartheta)}{\displaystyle\int_{B^{\eta}_{\vartheta^{(0)}}} f^{2,\eta}_{\psi^{(0)}}(\vartheta)\, \lambda^{(2)}(d\vartheta)}, \tag{A.1b}$$

where

$$f^{1,\eta}_{\psi^{(0)}}(\vartheta) = \frac{\int_{\mathcal{P}} \prod_{j=1}^{m} \bigl[F(\psi^{(0)}_j + \eta/2) - F(\psi^{(0)}_j - \eta/2)\bigr]\, \mathcal{D}_{M^{(1)}_{\vartheta} G^{(1)}_{\vartheta}}(dF)}{\eta^d \prod_{l=1}^{d}(m_l - 1)!} \tag{A.2}$$

and $f^{2,\eta}_{\psi^{(0)}}(\vartheta)$ is defined similarly. We may rewrite (A.2) as

$$f^{1,\eta}_{\psi^{(0)}}(\vartheta) = \frac{\int_{\mathcal{P}} \prod_{l=1}^{d} \bigl[F(\psi^{(0)}_{(l)} + \eta/2) - F(\psi^{(0)}_{(l)} - \eta/2)\bigr]^{m_l}\, \mathcal{D}_{M^{(1)}_{\vartheta} G^{(1)}_{\vartheta}}(dF)}{\eta^d \prod_{l=1}^{d}(m_l - 1)!}, \tag{A.3}$$

and we have a similar expression for $f^{2,\eta}_{\psi^{(0)}}(\vartheta)$. Let $A^{(1)}_l(\eta) = M^{(1)}_{\vartheta} G^{(1)}_{\vartheta}\bigl\{(\psi^{(0)}_{(l)} - \eta/2, \psi^{(0)}_{(l)} + \eta/2)\bigr\}$, $l = 1, \ldots, d$, and also define $A^{(1)}_{d+1}(\eta) = M^{(1)}_{\vartheta} - \sum_{l=1}^{d} A^{(1)}_l(\eta)$. Assume that η is so small that the sets $(\psi^{(0)}_{(l)} - \eta/2, \psi^{(0)}_{(l)} + \eta/2)$, $l = 1, \ldots, d$ are disjoint. Note that calculation of (A.3) is routine since it involves only the finite-dimensional Dirichlet distribution. The integral on the right side of (A.3) is $E(U_1^{m_1} \cdots U_d^{m_d})$ where $(U_1, \ldots, U_d, U_{d+1}) \sim$

Dirichlet$\big(A_1^{(1)}(\eta), \dots, A_{d+1}^{(1)}(\eta)\big)$, and a simple calculation shows that this expectation is equal to

$$\frac{\Gamma\big(M_\vartheta^{(1)}\big)}{\Big(\prod_{l=1}^d \Gamma\big(A_l^{(1)}(\eta)\big)\Big)\Gamma\big(A_{d+1}^{(1)}(\eta)\big)} \frac{\Big(\prod_{l=1}^d \Gamma\big(A_l^{(1)}(\eta)+m_l\big)\Big)\Gamma\big(A_{d+1}^{(1)}(\eta)\big)}{\Gamma\big(M_\vartheta^{(1)}+m\big)}.$$

Let

$$f^1_{\psi^{(0)}}(\vartheta) = \left(\prod_{l=1}^d g_\vartheta^{(1)}(\psi_{(l)}^{(0)})\right)\frac{\big(M_\vartheta^{(1)}\big)^d \Gamma\big(M_\vartheta^{(1)}\big)}{\Gamma\big(M_\vartheta^{(1)}+m\big)}.$$

Using the recursion $\Gamma(x+1) = x\Gamma(x)$ and the definition of the derivative, we see that for each $\vartheta \in \Theta$, $f^{1,\eta}_{\psi^{(0)}}(\vartheta) \to f^1_{\psi^{(0)}}(\vartheta)$. Under the regularity conditions listed just prior to the statement of the theorem, we see that this implies that (A.1b) converges to

$$\left(\prod_{l=1}^d \frac{g^{(1)}_{\vartheta^{(0)}}(\psi^{(0)}_{(l)})}{g^{(2)}_{\vartheta^{(0)}}(\psi^{(0)}_{(l)})}\right)\left(\frac{M^{(1)}_{\vartheta^{(0)}}}{M^{(2)}_{\vartheta^{(0)}}}\right)^d \frac{\Gamma\big(M^{(1)}_{\vartheta^{(0)}}\big)\Gamma\big(M^{(2)}_{\vartheta^{(0)}}+m\big)}{\Gamma\big(M^{(2)}_{\vartheta^{(0)}}\big)\Gamma\big(M^{(1)}_{\vartheta^{(0)}}+m\big)}\left[\frac{d\lambda^{(1)}}{d\lambda^{(2)}}\right](\vartheta^{(0)}), \tag{A.4}$$

and expression (A.4) thus gives $[d\nu_1/d\nu_2]$.

To prove Part (ii), we reconsider the calculation. In (A.1a), the numerator is replaced by

$$\begin{aligned}\nu_{\text{par},1}\big\{\vartheta \in B^\eta_{\vartheta^{(0)}},\, \psi \in C^\eta_{\psi^{(0)}}\big\} &= \int_{B^\eta_{\vartheta^{(0)}}} \prod_{j=1}^m \big[G_\vartheta^{(1)}(\psi_j^{(0)}+\eta/2) - G_\vartheta^{(1)}(\psi_j^{(0)}-\eta/2)\big]\,\lambda^{(1)}(d\vartheta) \\ &= \int_{B^\eta_{\vartheta^{(0)}}} \eta^m \prod_{j=1}^m g_\vartheta^{(1)}(\psi_j^{(0)})\,\lambda^{(1)}(d\vartheta) + o_p(\eta^m),\end{aligned}$$

where the last equality follows from Assumption A3. If $d < m$, the ratio in (A.1a) is $O_p(\eta^{m-d})$. If $d = m$, the limit is the expression given in (2.3).

Appendix B: Proof of Equation (3.3)

We may write

$$\left[\frac{d\nu_h^c}{d\nu_{h_1}}\right](\psi,\mu,\tau) = \left[\frac{d\nu_h^c}{d\nu_h}\right](\psi,\mu,\tau)\left[\frac{d\nu_h}{d\nu_{h_1}}\right](\psi,\mu,\tau),$$

since (i) $\nu_h^c \ll \nu_h$, as established in Burr and Doss (2005), who show that

$$\left[\frac{d\nu_h^c}{d\nu_h}\right](\psi,\mu,\tau) = \left[\frac{\Gamma^2\big(\frac{M}{2}\big)\Gamma(M+m)}{2^m\Gamma(M)}\right]K(\psi,\mu)$$

(see Proposition 2 and expression (A.7) of their paper) and (ii) $\nu_h \ll \nu_{h_1}$ (Theorem 1 of the present paper). In our particular case, the distribution λ on (μ,τ) is the same under ν_h and ν_{h_1}, and Theorem 1 gives simply

$$\left[\frac{d\nu_h}{d\nu_{h_1}}\right](\psi,\mu,\tau) = \left\{\prod_{r=1}^d \frac{t_v\big(\frac{\psi_{(r)}-\mu}{\tau}\big)}{t_{v_1}\big(\frac{\psi_{(r)}-\mu}{\tau}\big)}\right\}\left(\frac{M}{M_1}\right)^d \left\{\frac{\Gamma\big(M\big)\Gamma\big(M_1+m\big)}{\Gamma\big(M_1\big)\Gamma\big(M+m\big)}\right\}.$$

The result now follows after some algebraic simplifications.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2** 1152–1174.

Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association* **98** 224–235.

Berger, J. O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association* **96** 174–184.

Burr, D. (2010). `bspmma`: An R package for Bayesian semi-parametric models for meta-analysis. Tech. rep., Department of Statistics, University of Florida.

Burr, D. and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association* **100** 242–251.

Burr, D., Doss, H., Cooke, G. and Goldschmidt-Clermont, P. (2003). A meta-analysis of studies on the association of the platelet PlA polymorphism of Glycoprotein IIIa and risk of coronary heart disease. *Statistics in Medicine* **22** 1741–1760.

Buta, E. and Doss, H. (2010). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. Tech. rep., Department of Statistics, University of Florida.

Chan, K. S. and Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions". *The Annals of Statistics* **22** 1747–1758.

Cogburn, R. (1972). The central limit theorem for Markov processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2*. University of California Press, Berkeley.

Diaconis, P. and Freedman, D. (1986a). On inconsistent Bayes estimates of location. *The Annals of Statistics* **14** 68–87.

Diaconis, P. and Freedman, D. (1986b). On the consistency of Bayes estimates (c/r: P26–67). *The Annals of Statistics* **14** 1–26.

Doss, H. (1985a). Bayesian nonparametric estimation of the median: Part I: Computation of the estimates. *The Annals of Statistics* **13** 1432–1444.

Doss, H. (1985b). Bayesian nonparametric estimation of the median: Part II: Asymptotic properties of the estimates. *The Annals of Statistics* **13** 1445–1464.

Doss, H. (1994a). Bayesian estimation for censored data: An experiment in sensitivity analysis. In *Statistical Decision Theory and Related Topics, V* (S. S. Gupta and J. O. Berger, eds.). Springer-Verlag, New York.

Doss, H. (1994b). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics* **22** 1763–1786.

Durrett, R. (1991). *Probability: Theory and Examples*. Brooks/Cole Publishing Co.

Escobar, M. (1988). *Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means*. Ph.D. thesis, Yale University.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89** 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90** 577–588.

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

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2** 615–629.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7** 473–511.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, Department of Statistics, University of Minnesota.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* **16** 1069–1112.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis* **2** 445–472.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society, Series B* **65** 585–618.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics* **24** 911–930.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6** 831–860.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9** 249–265.

Robert, C. P. (2001). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, New York.

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.

Smith, T. C., Spiegelhalter, D. J. and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* **14** 2685–2699.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association* **99** 1027–1036.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

West, M., Müller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty. A Tribute to D. V. Lindley*.