Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2015 January 1; 77(1): 35–58. doi:10.1111/rssb.12059.

Marginally specified priors for non-parametric Bayesian estimation

David C. Kessler,
University of North Carolina, Chapel Hill, USA

Peter D. Hoff, and
University of Washington, Seattle, USA

David B. Dunson
Duke University, Durham, USA

Summary

Prior specification for non-parametric Bayesian inference involves the difficult task of quantifying prior knowledge about a parameter of high, often infinite, dimension. A statistician is unlikely to have informed opinions about all aspects of such a parameter but will have real information about functionals of the parameter, such as the population mean or variance. The paper proposes a new framework for non-parametric Bayes inference in which the prior distribution for a possibly infinite dimensional parameter is decomposed into two parts: an informative prior on a finite set of functionals, and a non-parametric conditional prior for the parameter given the functionals. Such priors can be easily constructed from standard non-parametric prior distributions in common use and inherit the large support of the standard priors on which they are based. Additionally, posterior approximations under these informative priors can generally be made via minor adjustments to existing Markov chain approximation algorithms for standard non-parametric prior distributions. We illustrate the use of such priors in the context of multivariate density estimation using Dirichlet process mixture models, and in the modelling of high dimensional sparse contingency tables.

Keywords

Contingency tables; Density estimation; Dirichlet process mixture model; Multivariate unordered categorical data; Non-informative prior; Prior elicitation; Sparse data

1. Introduction

Many real world data analysis situations do not lend themselves well to simple statistical models indexed by a finite dimensional parameter. This has led to the development of a rich class of non-parametric Bayesian (NP Bayes) methods, the general idea of which is to obtain inference under a prior that has support on the entire space of relevant probability distributions (Ferguson, 1973). These methods have been applied to a variety of problems,

Address for correspondence: Peter D. Hoff, Departments of Statistics and Biostatistics, University of Washington, Box 354322, Seattle, WA 98195, USA. pdhoff@uw.edu.

such as density estimation (Müller *et al.*, 1996), image segmentation (Sudderth and Jordan, 2009), speaker diarization (Fox *et al.*, 2011), regression and classification (Neal, 1999), functional data analysis (Petrone *et al.*, 2009) and quantitative trait loci mapping (Zou *et al.*, 2010) to name only a few. This breadth of applications reflects the utility of NP Bayes methods in modern statistical data analysis.

Many NP Bayes methods are built on either the Dirichlet distribution (DD) for finite sample spaces or the Dirichlet process (DP) (Ferguson, 1973) for infinite sample spaces. For the latter, the body of work on parameter estimation (Escobar, 1994), density estimation and inference (Escobar and West, 1995) and the steady improvement in sampling methods (Escobar, 1994; Walker, 2007; Yau *et al.*, 2011; Kalli *et al.*, 2011) have all made the DP prior an attractive choice for many applications. For a given sample space $\mathscr{Y}$, a DD or DP prior over distributions on $\mathscr{Y}$ is parameterized in terms of a 'base measure' Q_0 on $\mathscr{Y}$ and a 'concentration parameter' a. Although samples from the DP prior are discrete with probability 1, this prior is non-parametric in the sense that it has weak support on the set of all distributions having the same support as Q_0. Analogously, the DD prior is NP in the sense that it has support on the entire $(|\mathscr{Y}| - 1)$-dimensional simplex. For both the DD and the DP, a large value of a corresponds to a prior concentrated near Q_0. For the DP, a small a results in distributions with probability mass concentrated on only a few points, drawn independently from Q_0. For the DD, a small a can result in mass being concentrated near the vertices of the simplex.

For many NP Bayes methods, the DP is used as a prior for a mixing distribution in a mixture model: the data are assumed to come from a population with density $p(y|Q) = \int p(y|\psi)Q(\mathrm{d}\psi)$, where $\{p(y|\psi) : \psi \in \Psi\}$ is a simple parametric family. A DP prior on Q results in a Dirichlet process mixture model (DPMM) (Lo, 1984; Escober and West, 1995; MacEachern and Müller, 1998). As Q is discrete with probability 1, the resulting model for the population distribution is a countably infinite mixture model, where the parameters in the component measures are determined by Q_0, and the number of components with non-negligible weights is increasing in a.

Clearly, the choice of a and Q_0 will have a significant effect on the prior for the population density, and potentially on posterior inference. Many applications include priors for the base measure (Escobar and West, 1995; Müller *et al.*, 1996) and incorporate estimation of Q_0 and a in the posterior inference. Other approaches have addressed the challenge of specifying Q_0 by applying empirical Bayes techniques to develop a point estimate for Q_0 (McAuliffe *et al.*, 2006). Although it is common to give the base measure an overdispersed form in an attempt to avoid an unduly informative prior, such an approach is actually highly informative in favouring allocation to a single cluster unless a is appropriately adapted (Bush *et al.*, 2010). The particular case of the DP prior illustrates the general challenge of incorporating prior information in an NP setting. The results of Yamato (1984) and Lijoi and Regazzini (2004) can be extended to adjust a and Q_0 in normal DPMMs so that the induced prior expectation and variance of the population mean can be approximately specified (as will be discussed further in Section 3), although specification beyond the population mean is problematic. Moala and O'Hagan (2010) proposed a method to update a Gaussian process (GP) prior with expert assessments of the mean and other aspects of an unknown density. As with the DP

prior, the GP prior requires specification of the mean and covariance functions that characterize the GP. These provide a base for the prior in the same way that the Q_0 base measure does for the DP prior. In the Moala and O'Hagan approach, elicitation of these quantities is derived from expert assessments of quantiles of the unknown distributions.

In this paper, we propose a very general method that allows for the combination of an arbitrary prior on a finite set of functionals with an NP prior on the remaining aspects of the high or infinite dimensional unknown parameter. In the next section we show how such a partially informative prior distribution can be constructed from the combination of any prior distribution on the functionals of interest with the conditional distribution of the parameter given the functionals under a canonical NP prior. We show that the resulting marginally specified prior (MSP) inherits desirable features from the canonical prior: the MSP will generally share the support of the canonical prior, and posterior approximation under the MSP can typically be made via small modifications to any Markov chain Monte Carlo (MCMC) algorithm that is applicable under the canonical prior.

In Section 3 we illustrate the use of the MSP in the context of multivariate density estimation using normal DPMMs. In an example, we show that existing approaches to incorporate prior information on the mean and covariance in DPMMs lead to poor density estimates relative to marginally specified priors unless the parametric base model is an accurate approximation.

In Section 4 we examine the important problem of NP Bayes analysis of large sparse contingency tables in the presence of prior information on the margins. In this context, we develop a marginally specified prior from a canonical NP Bayes approach. In an example, we illustrate how canonical NP Bayes methods that are designed to be informative on the margins result in poor performance in terms of margin-free functionals (such as dependence functions). In contrast, an MSP accommodates prior information about the population margins while being minimally informative about other aspects of the population, resulting in strong performance in terms of both marginal and margin-free aspects of the population. A discussion of the results and directions for future research follows in Section 5.

Computer code to replicate the results can be obtained from http://wileyonlinelibrary.com/journal/rss-datasets

2. Marginally specified priors: construction and computation

We consider the general problem of Bayesian inference for a parameter f belonging to a high or infinite dimensional space $\mathscr{F}$. For example, Section 3 considers multivariate density estimation over the space of all densities on $\mathbb{R}^p$ with respect to Lebesgue measure, and Section 4 considers the high dimensional space of multiway contingency tables. In general, Bayesian inference for f is based on a posterior distribution $\pi(f \in A|y)$ derived from a sampling model $\{p(y|f) : f \in \mathscr{F}\}$ and a prior distribution π defined on a σ-algebra $\mathscr{A}$ of $\mathscr{F}$. In many high dimensional problems there are only a few classes of priors for which posterior inference is tractable. Typically, practitioners choose a member π_0 of such a class on the basis of support considerations and the feasibility of posterior approximation, rather than on how well it accurately represents any information that we have about specific

features of f. In this section, we show how to construct an NP prior π_1 that is informative about specific features of f but has the same support as π_0 and is 'close' to π_0 in terms of Kullback–Leibler divergence. We also show how MCMC approximation methods for π_0 can be modified to obtain posterior inference under π_1.

2.1. Construction of a marginally specified prior

Let $\theta = \theta(f)$ be a function of f, such as a population mean of $p(y|f)$, variance, marginal probability vectors or some finite set of functionals, and let Θ be the range of θ. Any prior distribution π_0 on $(\mathscr{Y}, \mathscr{A})$ induces a prior distribution P_0 on $(\Theta, \mathscr{B})$ defined by

$$P_0(B) = \pi_0[\{f : \theta(f) \in B\}], \quad (1)$$

for each $B \in \mathscr{B}$. If π_0 is chosen for computational convenience, the induced prior P_0 need not show substantial agreement with available prior information P_1 for the functional $\theta(f)$. In some cases a prior π_0 that is selected from a computationally feasible class will make the induced prior P_0 similar to P_1. The results of Lijoi and Regazzini (2004) and Yamato (1984) provide some guidance for DP priors if the functionals are means, but in general this will be difficult. Furthermore, depending on the structure of the NP class, selecting π_0 to match P_0 to P_1 will result in π_0 being inappropriate for other aspects of f. We present an example in Section 3 to illustrate a case where making π_0 highly informative about $\theta(f)$ also makes it highly informative about other aspects of f.

Suppose that an NP prior π_0 has been identified that is viewed as reasonable in some respects, such as being computationally feasible and having a large support, but does not represent available prior information P_1 about θ. The information in P_1 can be accommodated by replacing P_0, the θ-margin of π_0, with the desired margin P_1. Specifically, an MSP π_1 for f is obtained by combining the conditional distribution of f given θ with our desired marginal distribution P_1 for θ, so that

$$\pi_1(A) = \int \Lambda_0(A|\theta)\, P_1(\mathrm{d}\theta) \qquad \forall A \in \mathscr{A}, \quad (2)$$

where $\Lambda_0(A|\theta)$ is the conditional probability of A given θ under π_0. A prior π_1 that is constructed this way should have the desired marginal distribution P_1 over $(\Theta, \mathscr{B})$ and, if $P_1 \ll P_0$, should also have the same support as π_0, since the conditional probabilities under π_1 should match those under π_0.

Such a construction is straightforward if f is finite dimensional. Accommodation of NP problems where f is potentially infinite dimensional requires some additional mathematical detail. We consider the case where $\mathscr{A}$ are the Borel sets of a Hausdorff space $\mathscr{F}$, and $\theta : \mathscr{F} \to \Theta$ is a measurable map with respect to a σ-algebra $\mathscr{B}$ on Θ. Let the prior π_0 be a regular probability measure on $(\mathscr{F}, \mathscr{A})$, and let P_0 be the induced prior distribution on $(\Theta, \mathscr{B})$, i.e., for all $B \in \mathscr{B}$ $P_0(B) = \pi_0[\{f : \theta(f) \in B\}]$.

2.1.1. Example (Dirichlet process mixture model)—Recall the DPMM prior, defined by

$$p(y|Q) = \int p(y|\psi)\, Q(\mathrm{d}\psi),$$
$$Q \sim DP(\alpha, Q_0),$$

where $\{p(y|\psi), \psi \in \mathbb{R}^p\}$ is a collection of absolutely continuous probability densities over some Euclidean space $\mathscr{Y}$ and Q_0 is an absolutely continuous probability measure over $\mathbb{R}^p$. The random mixing measure Q has a representation as an infinite weighted sum of point mass measures, $Q = {}^{\mathrm{d}}\Sigma w_k \delta_{\psi k}$, where $\psi = \{\psi_1, \psi_2, \ldots\}$ are an infinite independent and identically distributed (IID) sample from Q_0, and $w_k = v_k \Pi_{j<k}(1 - v_j)$, with $v = \{v_1, v_2, \ldots\}$, are an infinite IID sample from a beta(1, α) distribution. Therefore the prior over Q can be represented as a prior over $\mathbb{R}^p$. This space, with the usual product topology, is Hausdorff. Now let θ be a moment of $p(y|Q)$, so that

$$\begin{aligned}\theta(f) &= \int g(y)\, p(y|Q)\, \mathrm{d}y \\ &= \sum_{k=1}^{\infty} \left\{ v_k \prod_{j<k} (1 - v_j) \right\} \int g(y)\, p(y|\psi_k)\, \mathrm{d}y.\end{aligned}$$

The function θ is Borel measurable as long as $p(y|\psi)$ is measurable in ψ for each $y \in \mathscr{Y}$.

Returning to the marginally specified prior given by equation (2), note that $\pi_0(\mathrm{A}|\theta)$ is not well defined on null sets of P_0. To make equation (2) meaningful, we restrict attention to informative prior distributions such that P_1 is dominated by P_0. Under this condition and the conditions on $(\mathscr{F}, \mathscr{A})$ and θ given above, the measure π_1 on $\mathscr{A}$ is well defined and the θ-marginal of π_1 is given by P_1.

Theorem 1: Let $\Lambda_0(\cdot|\cdot) : \mathscr{A} \times \Theta \to [0, 1]$ be a conditional probability function for π_0 given θ and let P_1 be a probability measure on $(\Theta, \mathscr{B})$ such that $P_1 \ll P_0$. Then $\pi_1 : \mathscr{A} \to [0, 1]$, defined by

$$\pi_1(A) = \int \Lambda_0(A|\theta)\, P_1(\mathrm{d}\theta),$$

- **a.** is a probability measure over $\mathscr{A}$,
- **b.** satisfies $\pi_1(\{f : \theta \in B\}) = P_1(B)$ for each $B \in \mathscr{Y}$ and
- **c.** is dominated by π_0 with Radon–Nikodym derivative

$$\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_0}(f) = \frac{\mathrm{d}P_1}{\mathrm{d}P_0}\{\theta(f)\}.$$

For notational economy, we have used θ to represent both an element of Θ and as the function mapping $\mathscr{F}$ to Θ, depending on the context. A proof of theorem 1 is provided in Appendix A.

The MSP π_1 that was constructed above is dominated by π_0, but ideally we would like it to have the same support as π_0. Since π_1 and π_0 share conditional distributions, intuitively it seems that π_1 should have reduced support relative to π_0 only if P_1 has reduced support relative to P_0. This result can be shown with the aid of the Radon–Nikodym derivative given above, which implies that $\pi_1(A)$ can be computed as

$$\pi_1(A) = \int_A \frac{p_1\{\theta(f)\}}{p_0\{\theta(f)\}} \pi_0(\mathrm{d}f) = E_{\pi_0}\left[1(f \in A)\frac{p_1(\theta)}{p_0(\theta)}\right],$$

where p_1 and p_0 are densities of P_1 and P_0 with respect to some common dominating measure (which could be taken to be P_0, for example). On the basis of this identity, we have the following result.

Lemma 1: Suppose that $P_1 \ll P_0 \ll P_1$. Then $\pi_1 \ll \pi_0 \ll \pi_1$.

Proof: It is clear from the definition of π_1 that $\pi_1 \ll \pi_0$. To show $\pi_0 \ll \pi_1$, let $A \in \mathscr{Y}$ be a set such that $\pi_1(A) = 0$. We shall show that $P_0 \ll P_1$ implies that $\pi_0(A) = 0$. Let $B_j = \{\theta : p_j(\theta) > 0\}$ and $A_j = \{f : \theta(f) \in B_j\}$ so that $\pi_j(A_j) = P_j(B_j) = 1$ for $j \in \{0, 1\}$. We have

$$\begin{aligned} 0 = \pi_1(A) &= \pi_1(A \cap A_1) \\ &= E_{\pi_0}[1(A \cap A_1)\, p_1/p_0] \\ &= E_{\pi_0}[1(A \cap A_0 \cap A_1)\, p_1/p_0]. \end{aligned} \tag{3}$$

Since $p_1/p_0 > 0$ on $A_0 \cap A_1$, equation (3) implies that $\pi_0(A \cap A_0 \cap A_1)$ Since $\pi_0(A_0) = 1$, we have $\pi_0(A \cap A_1) = \pi_0(A) - \pi_0(A \cap A_1^c) = 0$. Since $0 = \pi_1(A_1^c) = P_1(B_1^c)$ and $P_0 \ll P_1$, we must have $0 = P_0(B_1^c) = \pi_0(A_1^c)$, and so $\pi_0(A) = 0$.

We also note that π_1 has a characterization as the prior distribution that is closest to π_0 in terms of Kullback–Leibler divergence, among priors with θ-marginal density equal to p_1. This follows from re-expressing the probability measures π_1 and π_0 in terms of densities with respect to a common dominating product measure, so that

$$\pi_k\left(A \cap \theta^{-1}B\right) = \int_B \int_A \lambda_k(f|\theta)\, p_k(\theta)\, \mu(\mathrm{d}f) \times \nu(\mathrm{d}\theta)$$

for $k \in \{0, 1\}$. The Kullback–Leibler divergence is then

$$D(\pi_1 \| \pi_0) = E_{\pi_1}\left[ln\left\{\frac{\lambda_1(f|\theta)\, p_1(\theta)}{\lambda(f|\theta)\, p_0(\theta)}\right\}\right] = E_{\pi_1}\left[ln\left\{\frac{\lambda_1(f|\theta)}{\lambda_0(f|\theta)}\right\}\right] + E_{\pi_1}\left[ln\left\{\frac{p_1(\theta)}{p_0(\theta)}\right\}\right].$$

Fixing p_1, the divergence is minimized by setting $\lambda_1(f|\theta) = \lambda_0(f|\theta)$ for θ almost everywhere P_1, i.e. matching the conditional distributions, giving $D(\pi_1 \| \pi_0) = D(P_1 \| P_0)$.

Lemma 2: Let $P_1 \ll P_0$. Then, among probability measures π_1 on $(\mathscr{F}, \mathscr{A})$ with θ-marginal equal to P_1, the Kullback–Leibler divergence of π_0 from π_1 is minimized when $\pi_1(A) = \int \Lambda_0(A|\theta) P_1(\mathrm{d}\theta)$ for all $A \in \mathscr{A}$ and θ almost everywhere P_1.

A more detailed derivation of this result is given in Appendix A.

2.2. Posterior approximation under marginally specified priors

Let $\{p(y|f) : f \in \mathscr{F}\}$ be a dominated statistical model, i.e. a family of probability densities with respect to a common measure. Given a prior distribution π, inference for $f \in \mathscr{F}$ proceeds via the conditional probability distribution $\pi(\cdot|y) : \mathscr{A} \to [0,1]$, or alternatively the conditional density $\pi(f|y)$, given by

$$\pi(f|y) = \frac{p(y|f)\,\pi(f)}{p(y)} \equiv \frac{p(y|f)\,\pi(f)}{\int p\left(y|f'\right)\pi\left(f'\right)\mu\left(\mathrm{d}f'\right)},$$

where $\pi(f)$ denotes the density of π with respect to a dominating measure μ. This represents the conditional measure in that $\int_A \pi(f|y)\mu(\mathrm{d}f)$ is a version of the conditional probability $\pi(A|y)$ for each $A \in \mathscr{A}$.

For practical reasons the most commonly used priors are those for which there are straightforward Gibbs samplers or Metropolis–Hastings algorithms for posterior approximation. In many cases, simple modifications to these algorithms will allow for the incorporation of informative priors over functionals of interest. To illustrate, suppose that under prior π_0 we have a Gibbs sampler for a high dimensional parameter f. Recall that the Gibbs sampler can be viewed as a Metropolis–Hastings algorithm for which the proposals are accepted with probability 1. From this perspective, a Gibbs sampler for approximating the posterior density $\pi_0(f|y)$ is constructed from proposal distributions with densities $\boldsymbol{J}(f^*|f, y)$ that are proportional to the posterior density, so that

$$\frac{J(f^*|f,y)}{J(f|f^*,y)} = \frac{\pi_0(f^*|y)}{\pi_0(f|y)}. \quad (4)$$

For example, decomposing f as $\{f_1, \ldots, f_K\}$, the full conditional distribution $\pi_0(f_k|f_{-k}, y)$ is one such proposal distribution.

Posterior approximation of $\pi_1(f|y)$ can proceed by using the proposal distributions of the Gibbs sampler for $\pi_0(f|y)$, but adjusting the acceptance probability. Specifically, the algorithm for approximating $\pi_1(f|y)$ proceeds by iteratively simulating proposals f^* from distributions of the form $\boldsymbol{J}(f^*|f, y)$ which satisfy equation (4), and accepting each proposal f^* with probability $1 \wedge r_{\mathrm{MH}}$, where

$$\begin{aligned} r_{MH} &= \frac{\pi_1(f^*|y)}{\pi_1(f|y)}\frac{J(f|f^*,y)}{J(f^*|f,y)} \\ &= \frac{\pi_1(f^*|y)}{\pi_1(f|y)}\frac{\pi_0(f|y)}{\pi_0(f^*|y)} \\ &= \frac{p(y|f^*)\pi_1(f^*)}{p(y|f)\pi_1(f)}\frac{p(y|f)\pi_0(f)}{p(y|f^*)\pi_0(f^*)} = \frac{\pi_1(f^*)/\pi_0(f^*)}{\pi_1(f)/\pi_0(f)}. \end{aligned}$$

Let the θ-marginal distribution of π_0 be P_0, and let π_1 be an MSP based on π_0 and a θ-marginal distribution $P_1 \ll P_0$. Let p_0 and p_1 be the densities of P_0 and P_1 with respect to a

common dominating measure. By theorem 1, $\pi_1(f) = \pi_0(f) = p_1(\theta) = p_0(\theta)$ and the acceptance ratio simplifies to

$$\frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)}.$$

Similarly, an approximation algorithm for $\pi_1(f|y)$ can be constructed from a Metropolis–Hastings algorithm for $\pi_0(f|y)$ via the same adjustment. Suppose that we have a proposal distribution $\boldsymbol{J}(f^*|f, y)$ such that the acceptance ratio r^0_{MH} for π_0 is computable:

$$r^0_{MH} = \frac{\pi_0(f^*|y)}{\pi_0(f|y)} \frac{J(f|f^*, y)}{J(f^*|f, y)}.$$

The Metropolis–Hastings algorithm for approximating $\pi_1(f|y)$ using $\boldsymbol{J}(f^*|f, y)$ has acceptance ratio

$$\begin{aligned} r_{MH} &= \frac{\pi_1(f^*|y)}{\pi_1(f|y)} \frac{J(f|f^*,y)}{J(f^*|f,y)} \\ &= \frac{\pi_1(f^*|y)}{\pi_1(f|y)} \frac{\pi_0(f|y)}{\pi_0(f^*|y)} r^0_{MH} \\ &= \frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)} r^0_{MH}. \end{aligned}$$

These results show that an MCMC approximation to $\pi_1(f|y)$ can be constructed from an MCMC algorithm for $\pi_0(f|y)$ long as the ratio $p_1(\theta)/p_0(\theta)$ can be computed. The value of $p_1(\theta)$ for each $\theta \in \Theta$ is presumably available as p_1 is our desired prior distribution for θ. In contrast, obtaining a formula for $p_0(\theta)$ will be difficult in some settings. In situations where the dimension of θ is moderate, one simple solution is to obtain a Monte Carlo estimate of p_0 based on samples of f from π_0. Specifically, we can obtain an IID sample $\{\theta^{(s)} = \theta(f^{(s)}), s = 1, \ldots, S\}$ from $f^{(1)}, \ldots, f^{(S)} \sim_{\text{IID}} \pi_0$, and then approximate p_0 with a kernel density estimate or flexible parametric family. The method of approximation will depend on the nature of θ; the approaches just described are appropriate when $p_0(\theta)$ is absolutely continuous with respect to Lebesgue measure. Note that this can be done before the Markov chain is run, so that the same estimate of p_0 is used for each iteration of the algorithm.

In situations where obtaining a reliable estimate of p_0 is not feasible, it is still possible to induce a prior p_1 that is approximately equal to a target prior $\tilde{p}_1$, as long as p_0 is chosen to be flat compared with $\tilde{p}_1$. This can be done by replacing p_0, the θ-marginal density of π_0, with $p_1(\theta) \propto p_0(\theta)\tilde{p}_1(\theta) = Kp_0(\theta)\tilde{p}_1(\theta)$. This defines a valid probability density as long as $p_0\tilde{p}_1$ is integrable, which is so, for example, if either density is bounded. In terms of the MCMC approximation to the resulting marginally specified prior π_1, the adjustment to the acceptance ratio is then

$$\frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)} = \frac{\tilde{p}_1(\theta^*)}{\tilde{p}_1(\theta)},$$

which is presumably computable as $\tilde{p}_1$ is the desired prior density. In this setting, $\tilde{p}_1$ contains the marginal prior information and p_1 takes on a form with computational convenience.

The algorithm proposed is closely related to importance sampling methods described in the literature. Besag *et al.* (1995) detailed an importance-sampling-based approach for assessing prior sensitivity. In this development, an existing MCMC chain $\{\theta^{(t)}\}$ is weighted by using the ratios |, where $h(\cdot)$ is the original prior used to produce the sample and $\tilde{h}(\cdot)$ is an alternative prior. The similiarity with our proposed method and its use of ratios of the marginally specified prior p_1 to the induced prior p_0 is clear; one important distinction is that our method replaces an induced prior on functionals with an elicited prior on those functionals, rather than substituting a prior in the main specification.

3. Density estimation with marginally adjusted Dirichlet process mixture model

Perhaps the most commonly used NP Bayes procedure is the DPMM (Lo, 1984; Escobar and West, 1995; MacEachern and Müller, 1998). The DPMM consists of a mixture model along with a DP prior for the mixing distribution. The population density to be estimated and the prior can be expressed as

$$p(y|Q) = \int p(y|\psi)\, Q(\mathrm{d}\psi),$$
$$Q \sim DP(\alpha Q_0),$$

where α and Q_0 are hyperparameters of the DP prior, with Q_0 typically chosen to be conjugate to the parametric family of mixture component densities, $p(y|\psi) : \psi \in \Psi\}$, to facilitate posterior calculations. In this section we show how to obtain posterior approximations under an MSP π_1 on the basis of a DPMM. The approach is illustrated with the specific case of multivariate density estimation, for which we take the parametric family to be the class of multivariate normal densities. In an example analysis of the well-known bivariate data set on eruption times of the Old Faithful Geyser, we construct a prior distribution π_1 based on the multivariate normal DPMM with a marginally specified informative prior on the marginal means and variances. Here, we use a parametric approximation for the induced joint distribution p_0 of these specific functionals θ. Inference under π_1 is compared with inference under two standard DPMMs: one where the hyperparameters are chosen to be informative about θ and another where the hyperparameters are non-informative.

3.1. Posterior approximation

Given a sample $y_1, \ldots, y_n \overset{\text{IID}}{\sim} p(y|Q)$, posterior approximation for conjugate DPMMs is often made with a Gibbs sampler that iteratively simulates values of a function that associates data indices with the atoms of Q. In a DPMM, since Q is discrete with probability 1, a given mixture component (atom of Q) may be associated with multiple observations. Let $g : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be the unknown mixture component membership function, so that $g_i = g_j$ means that y_i and y_j came from the same mixture component. Note that g can always be expressed as a function that maps $\{1, \ldots, n\}$ onto $\{1, \ldots, K\}$, where $K \leq n$.

Inference for conjugate DPMMs often proceeds by iteratively sampling each g_i from } its full conditional distribution $p(g_i|y_1, \ldots, y_n, g_{-i})$ (Bush and MacEachern, 1996). Additional features of Q and $p(y|Q)$ can be simulated given $g_1, \ldots, g_n$ and the data.

This standard algorithm for DPMMs can be modified to accommodate an MSP distribution on a parameter $\theta = \theta(Q)$. Let $f = \{g, \theta\}$ and let π_0 be the prior density on f induced by the DP on Q. Our marginally specified prior is given by $\pi_1(f) = \pi_0(f)p_1(\theta)/p_0(\theta)$, where p_0 is the density for θ induced by π_0, and p_1 is the informative prior density. An MCMC approximation to $\pi_1(f|y_1, \ldots, y_n)$ can be obtained via the procedure that was outlined in Section 2.2. Given a current state of the Markov chain $f = \{\theta, g_i, g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n\} = \{\theta, g_i, g_{-i}\}$, the next state is determined as follows.

Step 1: generate a proposal $f^* = \{\theta^*, g_i^*, g_{-i}\}$ from $\pi_0(\theta, g_i|g_{-i}, y) = \pi_0(g_i|g_{-i}, y)\pi_0(\theta|g, y)$ by

- **a.** generating $g_i^* \sim \pi_0(g_i|g_{-i}, y)$ and
- **b.** generating $\theta^* \sim \pi(\theta|g_i^*, g_{-i}, y)$.

Step 2: set the value of the next state of the chain to f^* with probability

$$1 \wedge \frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)};$$

otherwise let the next state equal the current state.

This procedure is iterated over values of $i \in \{1, \ldots, n\}$, possibly in random order, and repeated until the desired number of simulations of f is obtained. Note that steps 1(a) and 1(b) compose a standard Gibbs sampler for the DPMM in which posterior inference for θ is provided, although typically we would simulate θ only once per complete update of $g_1, \ldots, g_n$. The algorithm for the marginally specified prior π_1 requires that θ be simulated with each proposed value of g_i so that the acceptance probability in step 2 can be calculated.

Implementing the steps of this MCMC algorithm involves two non-trivial computations: simulation of θ from $\pi_0(\theta|g, y)$, and calculation of $p_0(\theta)$ to obtain the acceptance probability. General methods for the latter were discussed in Section 2.2. For the former, we suggest the use of a Monte Carlo approximation to Q based on a representation of DPs due to Pitman (1996). Let K be the number of unique values of $g_1, \ldots, g_n$ and let n_k be the number of observations i for which $g_i = k$. If Q_0 is conjugate, then the parameter values $\psi_{(1)}, \ldots, \psi_{(K)}$ corresponding to the mixture components can generally be easily simulated. Corollary 20 of Pitman (1996) gives the conditional distribution of Q given $\psi_{(1)}, \ldots, \psi_{(K)}$ and counts $n_1, \ldots, n_K$ as

$$\left\{Q(H)|\psi_{(1)}, \ldots, \psi_{(K)}, n_1, \ldots, n_K\right\} \stackrel{\mathrm{d}}{=} \gamma \sum_{k=1}^{K} 1\left(\psi_{(k)} \in H\right) w_k + (1-\gamma)\tilde{Q}(H),$$

where $\gamma \sim$ beta (n, a), $w \sim$ Dirichlet(n_1, … , n_K) and $\tilde{Q} \sim DP(aQ_0)$. A Monte Carlo approximation to Q, and therefore any functional of Q, can be obtained via simulation of a large number S of ψ-values from Q. To do this, we first simulate γ and w_1, … , w_K from their beta and Dirichlet full conditional distributions. From these values we sample cluster memberships for a sample of size S from Q by using a multinomial(S, {γw_1, … , γw_K, 1 − γ}) distribution. Note that the count s for the (K + 1)th category represents the number of ψ-values that must be simulated from $\tilde{Q}$. To obtain the sample from $\tilde{Q}$ we run a Chinese restaurant process of length s and then generate the unique ψ-values from Q_0 for each partition. This can generally be done quickly for two reasons: first, the expected number of samples that is needed from $\tilde{Q}$ is only $Sa/(a + n)$. For example, with S = 1000, n = 30 and a = 1, we expect to need only about s = 32 simulations from $\tilde{Q}$. Second, the number of unique values in a sample of size s from $\tilde{Q}$ is only of order log(s), which will generally be manageably small.

The marginal sampler that we described above has advantages in terms of efficiency and convergence rates (MacEachern, 1994). However, because it does marginalize out the random measure Q, we must use the embedded Pitman method to draw samples from $\theta(Q)$ to evaluate the Metropolis–Hastings ratio. An alternative approach is to use a stick breaking representation that does not integrate out the random measure. We can then use a slice sampler (Kalli *et al.*, 2011) or exact block Gibbs sampler (Yau *et al.*, 2011) and compute $\theta(Q)$ without needing an embedded sampling step, but at the possible expense of lower efficiency in the sampler.

3.2. Example: Old Faithful eruption times

The Old Faithful data set consists of 272 bivariate observations of eruption times and waiting times between eruptions, both measured in minutes. To illustrate and evaluate the MSP methodology we construct two subsets of these data: a random sample of size n_0 = 30 from which we obtain prior information and a second, non-overlapping random sample of size n = 30 representing our observed data. The random samples were obtained by setting the random seed in R (version 2.14.0) to 1, sampling the prior data set, and then sampling the observed data set from the remaining observations. The observed sample had marginal means (2.97, 64.2) and marginal variances (1.29, 206.7). The prior sample had marginal means (3.54, 71.9) and marginal variances (1.24, 134.9). For the purpose of this example, we view the full data set of 272 observations as the true population. A scatter plot of the observed data and marginal density estimates are shown in Fig. 1. The observed data set consisting of n = 30 observations clearly captures the bimodality of the population. However, the marginal plots indicate that the sample has overrepresented one of the modes.

Suppose that our knowledge of the prior sample is limited to the bivariate marginal sample means $m_0 \in \mathbb{R}^2$ and sample variances $v_0 \in \left(\mathbb{R}^+\right)^2$. In such a situation it would be desirable to construct a prior density p_1 over the unknown population marginal means m and variances v on the basis of the values of m_0, v_0 and n_0, and to combine this information with the information in our fully observed sample to improve our inference about the population. Incorporating this information with conjugate priors would be straightforward if our

sampling model were bivariate normal, but it is difficult in the context of a DPMM. Proposition 5 of Yamato (1984) indicates that, if the base measure Q_0 in the DP prior is multivariate normal(μ_0, Σ_0), then the induced prior distribution on the mean $\int xQ(\mathrm{d}x)$ is approximately multivariate normal$\{\mu_0, \Sigma_0/(a+1)\}$. This result is not directly applicable to the multivariate normal DPMM for two reasons, one being that Q represents the mixing distribution and not the population distribution, and the other being that in the conjugate multivariate normal DPMM the parameter ψ in the mixture component consists not just of a mean μ but also a covariance matrix Σ. Specifically, in the conjugate p-variate normal DPMM, the density q_0 of the base measure Q_0 for $\psi = (\mu, \Sigma)$ is given by

$$q_0(\mu, \Sigma) = \mathrm{normal}_p(\mu : \mu_0, \Sigma/\kappa_0)\,\mathrm{inverse-Wishart}\left(\Sigma : S_0^{-1}, \nu_0\right) \quad (5)$$

where the functions on the right-hand side are the multivariate normal and inverse Wishart densities, the latter being parameterized so that $E[\Sigma] = S_0/(\nu_0 - p - 1)$. Given a choice for a it is possible to obtain values of the hyperparameters (μ_0, κ_0, S_0, ν_0) so that the induced prior distributions on the population mean

$$m(Q) = \int\int y p(y|\psi)\, Q(\mathrm{d}\psi)\, \mathrm{d}y$$

and variance

$$V(Q) = \int\int y y^{\mathrm{T}} p(y|\psi)\, Q(\mathrm{d}\psi)\, \mathrm{d}y - m(Q)\, m(Q)^{\mathrm{T}}$$

have the following properties:

$$\left.\begin{array}{c} E[m(Q)] = m_0, \\ var\{m(Q)\} = \frac{V_0}{n_0 - p - 1} \approx \frac{V_0}{n_0}, \\ E[V(Q)] = \frac{n_0 + \alpha + 1}{n_0}\frac{n_0 V_0}{n_0 - p - 1} \approx V_0. \end{array}\right\} \quad (6)$$

Here, m_0 is the desired prior mean and V_0 is the desired prior covariance matrix, derived from the marginal prior information. Within the context of the DPMM, it is difficult to specify the prior on $V(Q)$ separately from that on $m(Q)$. We construct three different NP prior distributions for a comparative analysis of the Old Faithful data.

a. Informative DPMM π_0^{I}: the base measure density q_0 is as in expression (5) with ($\mu_0 = m_0$, $\kappa_0 = n_0/(a+1)$, $\nu_0 = n_0$, $S_0 = \nu_0 V_0$), where the diagonal of V_0 is v_0, the marginal variances from the prior sample, and the correlation is equal to the sample correlation from the observed data. This results in a prior on Q satisfying properties (6), thereby utilizing the prior information.

b. Non-informative DPMM π_0^{N}: the base measure density q_0 is as in expression (5) with $\left(\mu_0 = \bar{y}, \kappa_0 = 1/10, \nu_0 = p + 2 = 4, S_0 = S_y\right)$, where $\bar{y}$ is the sample mean from the $n = 30$ values in the observed sample and S_y is sample covariance matrix. This prior does not use information from the prior sample and is designed to promote relative

diffuseness of the induced prior on the marginal population means and variances. Using sample moments for the hyperparameters weakly centres the prior near the observed data. We can view this as a type of 'unit information' prior (Kass and Wasserman, 1995).

c. MSP π_1: letting $\theta = (m_1, m_2, v_1, v_2)$ be the unknown population means and marginal variances, we construct an MSP by replacing the θ-margin of π_0^{N} with $p_1(\theta)$, a product of two univariate normal and two inverse gamma densities, chosen to match the prior on θ induced by π_0^{I} as closely as possible.

Fig. 2 compares p_1 with kernel density estimates of the marginal priors induced by π_0^{I}, showing that π_0^{I} and π_1 have very similar θ-margins, but otherwise π_1 matches the more diffuse prior π_0^{N}. We can give π_1 any θ-margin we wish, but matching the margins of π_0^{I} and π_1 facilitates comparison. The hyperparameter a was set to 1 for all the above prior distributions. To evaluate the Metropolis–Hastings ratios when approximating the posterior distribution under π_1, we found that a skewed multivariate *t*-distribution provided a very accurate approximation to the joint distribution of the marginal means and log-variances induced by π_0^{N}. Via a change of variables, this provides an accurate approximation to $p_0(\theta)$, with which the acceptance probability is computed for approximation of $\pi_1(f|y)$. Fig. 3 gives an assessment of the adequacy of this approximation, comparing a smoothed density estimate of random draws from the approximated p_0 with a smoothed density estimate of random draws from the true p_0 induced by π_0^{N}.

We ran Markov chains of length 25000 under each prior, with parameter values being saved every 10th iteration, resulting in 2500 simulated values of each parameter with which to make posterior approximations. The chains showed no evidence of non-stationarity and mixed well under each prior: on the basis of the dependent MCMC sequences of length 2500, the equivalent numbers of independent observations of θ (i.e. the effective sample sizes) were estimated as above 2000 for each element of θ and under each prior. We did sample from the posterior under π_1 by using a stick breaking representation and a slice sampler. The results were not markedly different from those obtained by using the marginal sampler. This slice sampling approach required dependent MCMC sequences of length 550000 to achieve an effective sample size of 2500; computational time per independent sample was 95% that of the marginal sampler.

Posterior predictive distributions under the three priors are shown in Fig. 4. The informative DPMM provides a poor representation of the population distribution, given in light grey contours. This is primarily a result of having to set the κ_0-hyperparameter to be moderately large ($\kappa_0 = 15$) to obtain the desired informative prior variance for the population mean $m = (m_1, m_2)$. Unfortunately, setting this parameter so high means that values of μ in the mixture model are tightly concentrated around m_0, and so the multimodality is not captured. In contrast, the posteriors under the non-informative DPMM π_0^{I} and the MSP π_1 can capture the multimodality of the population.

Fig. 5 gives marginal density estimates under the various priors. Fig. 5 suggests that the posterior under π_1 is better at representing the underlying population than the posteriors under the other priors. Recall that the observed sample contains an unrepresentative number of low valued observations. The posterior under the non-informative prior π_0^{N} uses only the observed data and thus is equally unrepresentative of the population. In contrast, π_1 can use some information from the prior sample and is therefore more representative of the population.

Finally, the marginal posterior distributions of the marginal parameters m and $\log(v)$ are given in Fig. 6. The priors are given in grey and the resulting posterior distributions are given in black. The population values based on the full set of 272 observations are given by grey vertical lines. Across all parameters, π_1 gives posteriors that are most concentrated around the population means. Note that the difference between the priors and the posteriors under π_0^{I} is not that large. We conjecture that this is primarily a result of the fact that, under π_0^{I}, most observations are estimated as coming from the same mixture component, thereby overestimating the entropy, when in fact the data are bimodal. In contrast, π_1 can recognize the bimodality and obtain improved estimates of the marginal densities.

In this example, we have shown that efforts to make the canonical DPMM informative in terms of marginal means and variances leads to poor density estimates, whereas a non-informative DPMM leads to suboptimal estimates of functionals because of its inability to incorporate prior information. In contrast, a marginally specified prior can both incorporate prior information and provide accurate density estimation.

4. Marginally specified priors for contingency table data

Even when multivariate categorical data include only moderate numbers of variables and categories, large or full models that allow for complex or arbitrary multivariate dependence can involve a very large number of parameters. For example, a full model for the $(2 \times 3 \times 2 \times 8 \times 12)$-way contingency table data that we consider later in this section requires a 1151-dimensional parameter. One Bayesian approach to the analysis of such data is via model selection between reduced log-linear models (Dawid and Lauritzen, 1993; Dobra and Massam, 2010). However, model selection can be difficult even for moderate numbers of variables and categories, owing to the large number of models with low posterior probability and the resulting difficulty in completely exploring the model space. An alternative NP Bayes approach was provided by Dunson and Xing (2009), who developed a prior based on a DP mixture of product multinomial distributions. Such a prior has full support on the parameter space but concentrates prior mass near simple submodels. One drawback to this approach is the lack of a straightforward method for incorporating the type of marginal prior information that is frequently available for categorical data.

In this section we consider an alternative NP Bayes approach based on a marginal adjustment to a standard Dirichlet prior distribution. This approach is computationally straightforward and allows for the incorporation of prior information on specific functionals of the unknown population distribution, such as the univariate marginals.

4.1. The canonical Dirichlet prior

Multivariate categorical data consist of observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$, for which $y_{ij} \in \{1, 2, \ldots, d_j\}$ for $j = 1, \ldots, p$. A p-way contingency table is a common representation for such data, in which each cell of the table indicates the count of observations $\boldsymbol{y}_i$ such that $y_{i1} = c_1, \ldots$ and $y_{ip} = c_p$ for a specific response vector $c = (c_1, \ldots, c_p)$. The sampling model for a contingency table can be expressed as a multinomial distribution, where for each cell $c \in \mathscr{C} = \{c : 1 \leq c_j \leq d_j, j = 1, \ldots, p\}$ we define $f_c \equiv \Pr(y_{i1} = c_1, \ldots, y_{ip} = c_p)$. The full model of all distributions for the data can then be indexed by the parameter $f = \{f_c : c \in \mathscr{C}\}$, which lies in the $(\Pi d_j - 1)$-dimensional simplex. Given n IID observations, the likelihood is

$$L(f|y_1, \ldots, y_n) = \prod_{c_1=1}^{d_1} \times \ldots \times \prod_{c_p=1}^{d_p} f_c^{\sum_i 1(y_{i1}=c_1, \ldots, y_{ip}=c_p)},$$

for which a standard conjugate prior is the Dirichlet distribution with hyperparameter $\alpha \in \left(\mathbb{R}^+\right)^{\Pi d_j}$. This is an NP prior in the sense that it gives full support on the space of possible values of f.

The Dirichlet prior is an appealing choice computationally because of its conjugacy, but this convenience can have undesirable side effects. In particular, choosing an uninformative Dirichlet prior for f induces substantial informativeness about the marginals $\{\theta_1, \ldots, \theta_p\}$, where $\theta_j = \{\theta_{j1}, \ldots, \theta_{jd_j}\} = \{\Pr(y_{ij} = 1|f), \ldots, \Pr(y_{ij} = d_j|f)\}$. For example, setting $a_c = 1$ for each cell $c \in \mathscr{C}$ results in a uniform prior distribution for f, which is often used as a default prior distribution in the absence of prior information. However, the induced prior on the marginals $\theta_1, \ldots, \theta_p$ is highly informative: the marginalization properties of the Dirichlet distribution result in $\theta_j \sim \text{Dirichlet}(\Pi_{k \neq j} d_k, \ldots, \Pi_{k \neq j} d_k)$, which is generally highly concentrated around the uniform distribution on $\{1, \ldots, d_j\}$. However, it is reasonably straightforward to choose values of a_c to induce particular marginal Dirichlet priors on the θ_js, although each marginal prior must have the same concentration, but this approach to constructing an informative prior for the margins necessarily induces a prior over the remaining aspects of f, such as the dependence structure, that could be undesirably informative.

4.2. A marginally specified prior

To overcome these undesirable features of the Dirichlet prior, we construct an NP prior on f based on a Dirichlet distribution with a low total concentration, but with the induced marginal priors for $\theta_1, \ldots, \theta_p$ replaced with informative priors to reflect known information. Specifically, our prior for f takes the form

$$\begin{aligned} \pi_1(f) &= \pi_0(f|\theta)\, p_1(\theta) \\ &= \pi_0(f|\theta) \prod_{j=1}^{p} p_{1j}(\theta_j), \end{aligned}$$

where $\pi_0(f)$ is a Dirichlet($\alpha_0, \ldots, \alpha_0$) distribution on the ($\Pi d_j - 1$)-dimensional simplex and p_{1j} is an informative Dirichlet distribution on a ($d_j - 1$)-dimensional simplex. Recall from Section 2 that the MSP π_1 is the closest distribution in Kullback–Leibler divergence to π_0 that has the desired priors on $\theta_1, \ldots, \theta_p$. Also note that the methodology does not require that these induced priors be Dirichlet, although making them so will facilitate comparison with an informative Dirichlet prior distribution on f in the example data analysis that follows.

Estimation of f via the posterior distribution $\pi_1(f|\mathbf{y})$ can proceed via an MCMC algorithm. As in the previous section, we modify an MCMC algorithm for simulating from $\pi_0(f|\mathbf{y})$, the posterior under the canonical NP prior, to obtain simulations from $\pi(f|\mathbf{y})$, the posterior under the marginally specified prior. Our particular MCMC scheme relies on the representation of a Dirichlet-distributed random variable as a set of independent gamma variables scaled to sum to 1, i.e., if $Z_c \sim \text{gamma}(\alpha_c, 1)$ and $f_c = Z_c/\Sigma Z_c$', then $f \sim \text{Dirichlet}\left(\alpha_1, \ldots, \alpha_{|\mathscr{C}|}\right)$. We employ an MCMC algorithm that is based on simulating proposed values of $\{ln(Z_c) : c \in \mathscr{C}\}$ from a normal distribution centred at the current values. Because of the high dimension of the parameter f, proposing changes to every element of f simultaneously results in low rates of acceptance. To avoid this problem, at each iteration of the algorithm we propose changes to randomly chosen subvectors of f. The steps in a single iteration of the MCMC algorithm are then as follows.

Step 1: generate a proposal $\left\{f^*, \theta_1^*, \ldots, \theta_p^*\right\}$,

- **a.** randomly sample a set of cells $\mathscr{C}' \subset \mathscr{C}$;
- **b.** simulate proposals $\left\{log(Z_c^*) : c \in \mathscr{C}'\right\} = \left\{log(Z_c) : c \in \mathscr{C}'\right\} + \varepsilon, \varepsilon \sim N(0, \delta I)$;
- **c.** compute the corresponding f^* and marginal probabilities $\theta_1^*, \ldots, \theta_p^*$.

Step 2: compute the acceptance ratio $r = r_0 r_1$ from r_0, the acceptance ratio for f under π_0, and r_1, the marginal prior ratio:

$$r_0 = \frac{p(y|f^*)\pi_0(Z^*)}{p(y|f)\pi_0(Z)} \prod_c \frac{Z_c^*}{Z_c},$$
$$r_1 = \frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)}.$$

Step 3: accept $f^*, \theta_1^*, \ldots, \theta_p^*$ with probability $1 \wedge r$.

Note that the ratio r_0 includes the Jacobian of the transformation from Z to ln(Z), as the proposal distribution is symmetric on the log-scale. The number of cells $|\mathscr{C}'|$ to update at each step and the variance parameter δ in the proposal distribution can be adjusted to achieve target acceptance rates.

As mentioned above, we take p_1 to be a product of Dirichlet densities representing prior information about the margins $\theta_1, \ldots, \theta_p$. To calculate r_1 we must also compute the corresponding joint distribution p_0 of $\theta_1, \ldots, \theta_p$ under the Dirichlet distribution π_0 on f. We

approximate p_0 by the product of the prior marginal densities of $\theta_1, \ldots, \theta_p$ under π_0, each of which are Dirichlet. However, we note that the θ_js are only approximately independent of each other under π_0.

4.3. Example: North Carolina public use microdata sample data

We evaluate the performance of the MSP and several associated priors in terms of their performance under the scenario of a researcher with accurate prior information about the marginal distributions of the p categorical variables. Our scenario is based on data from the public use microdata sample of the American Community Survey, which is a yearly demographic and economic survey. We consider data on gender (male or female: $d_1 = 2$), citizenship (native, nat uralized or non-citizen: $d_2 = 3$), primary language spoken (English or other: $d_3 = 2$), class of worker ($d_4 = 8$) and mode of transportation to work ($d_5 = 12$), from 40769 survey participants. The last two variables are each dominated by a single category: 'employee of private company' (63.75%) for worker class and 'car, truck or van' (91.97%) for transportation. These classifications yield a five-way contingency table with $|\mathscr{C}| = 1152$ cells. From these data we constructed a true joint distribution $\tilde{f}$ and marginal frequencies $\tilde{\theta}$ by filling out the multiway contingency table with the public use microdata sample data, replacing zero counts in the contingency table with small fractional counts, and normalizing the resulting counts to produce a probability distribution over $|\mathscr{C}|$. We then simulated smaller data sets of various sample sizes from $\tilde{f}$ and obtained posterior estimates for each under three different prior distributions.

a. The informative Dirichlet prior π_0^{I} is a Dirichlet distribution with parameter $\alpha_1 f_0^{\mathrm{I}}$, where $\alpha_1 |\mathscr{C}|$ and f_0^{I} is in the $(|\mathscr{C}| - 1)$-simplex. Using the method of Csiszár (1975), the prior mean f_0^{I} of f was chosen to be the frequency vector that was closest in Kullback–Leibler divergence to the uniform distribution on $|\mathscr{C}|$ among those with margins equal to $\tilde{\theta}$. The induced marginal prior on each θ_j is then Dirichlet$\left(|\mathscr{C}|\tilde{\theta}_j\right)$ which has prior expectation $\tilde{\theta}_j$ as desired. Note that the concentration hyperparameter α_{I} is the same as that for a uniform prior on the simplex.

b. The non-informative Dirichlet prior π_0^{N} is a Dirichlet distribution with parameter $\alpha_{\mathrm{N}} f_0^{\mathrm{N}}$, where $\alpha_{\mathrm{N}} \sqrt{|\mathscr{C}|}$ and $f_0^{\mathrm{N}} = \{1/|\mathscr{C}|, \ldots, 1/|\mathscr{C}|\}$. This prior has the same prior expectation as the uniformprior on the $(|\mathscr{C}| - 1)$-simplex, but a smaller prior concentration by a factor of $\sqrt{|\mathscr{C}|}$.

c. The MSP π_1 was constructed by replacing the marginal prior for θ induced by π_0^{N} with the marginal prior under π_0^{I}. (To compute acceptance ratios,we have used a product of independent Dirichlet distributions corresponding to the marginal distributions induced by π_0^{N} to approximate p_0. The adequacy of the approximation to p_0 is assessed in Fig. 7 through a comparison of smoothed density estimates of random draws from the approximated p_0 with smoothed density estimates of random draws from the true p_0 induced by π_0^{N}.

We used the true joint distribution $\tilde{f}$ to generate 200 replicate data sets of sizes $n \in \{100, 1000, 5000, 10000, 20000, 40000\}$. The π_0^{I}- and π_0^{N}-priors are conjugate to themultinomial likelihood, and so their posterior distributions are available in closed form. For estimation under π_1, the MCMC algorithm that was described above was run for 3×10^6 iterations for each simulated data set. The acceptance rate varied with the sample size *n*, from 89% at $n = 100$ down to 63% at $n = 10000$. Effective sample sizes corresponding to thinned Markov chains based on every 500th iterate were obtained and were found to be around 1000 (based on thinned chains of length 6000).

For each simulated data set and prior we obtain posterior mean estimates $\left(\hat{f}, \hat{\theta}\right)$ which we compare with the true values $\left(\tilde{f}, \tilde{\theta}\right)$ that were used to generate the simulated data. To evaluate $\hat{\theta}$, we use an average of the absolute value of the Kullback–Leibler divergence between the true marginal distributions $\left\{\tilde{\theta}_1, \ldots, \tilde{\theta}_p\right\}$ and the estimated marginal distributions $\left\{\hat{\theta}_1, \ldots, \hat{\theta}_p\right\}$:

$$M = \frac{1}{p}\sum_{j=1}^{p}\left|\sum_{c=1}^{d_j}\tilde{\theta}_{jc} ln\left(\frac{\hat{\theta}_{jc}}{\tilde{\theta}_{jc}}\right)\right|.$$

Smaller values of *M* indicate better performance with respect to this marginal metric.

To assess the performance of $\hat{f}$ on aspects of *f* other than the marginal distributions, we compared the true and estimated values of the local dependence functions (LDFs) of the $\binom{p}{2}$ separate two-way marginal distributions. These LDFs describe the two-way dependences between the variables and are invariant to changes in the marginal distributions (Goodman, 1969). The LDFs are formed from cross-product ratios of *f* as follows: letting $f_{c_1,c_2}^{j_1,j_2} = Pr\left(y_{j_1} = c_1, y_{j_2} = c_2 | f\right)$, we define

$$LDF_{c_1,c_2}^{j_1,j_2}(f) = ln\left(\frac{f_{c_1,c_2}^{j_1,j_2} f_{c_1+1,c_2+1}^{j_1,j_2}}{f_{c_1,c_2+1}^{j_1,j_2} f_{c_1+1,c_2}^{j_1,j_2}}\right).$$

For each simulated data set and prior distribution, we computed the average squared error between $LDF_{c_1,c_2}^{j_1,j_2}\left(\hat{f}\right)$ and $LDF_{c_1,c_2}^{j_1,j_2}\left(\tilde{f}\right)$ as

$$L = \binom{p}{2}^{-1}\sum_{j_1<j_2}\frac{1}{\left(d_{j_1}-1\right)\left(d_{j_2}-1\right)}\sum_{c_1=1}^{d_{j_1-1}d_{j_2-1}}\left\{LDF_{c_1,c_2}^{j_1,j_2}\left(\hat{f}\right) - LDF_{c_1,c_2}^{j_1,j_2}\left(\tilde{f}\right)\right\}^2.$$

Smaller values of *L* indicate better performance in terms of representing the two-way dependence structure of the true distribution $\tilde{f}$.

Fig. 8 shows the M and L performance metrics for each prior and simulated data set, with the averages over simulations at each sample size joined by lines. The sample sizes are displayed ordinally, with a slight horizontal shift for each prior so that the results under different priors can be distinguished. Not surprisingly, the estimates of θ under π_0^{I} and π_1 outperform those under π_0^{N}, as these former two priors were designed to have correct prior expectations for θ. The initial non-monotonic trend in the performance of π_0^{I} with sample size is because π_0^{I} has exactly correct prior expectation. If the sample size were 0, then M would be 0 as well. In contrast, Fig. 8(b) indicates that π_0^{I} provides poor estimates of the dependence functions: at all sample sizes, this prior underperforms compared with the other two, demonstrating the cost of making π_0^{I} directly informative about the marginals. In contrast, π_0^{N} and π_1 have very comparable performance in terms of estimation of the dependence functions. These comparisons, using both the marginal and the margin-free performance metrics, highlight the desirable properties of the marginally specified prior formulation: an MSP π_1 can represent prior information about specific functionals θ of the high dimensional parameter f without being overly informative about other aspects of the parameter.

5. Discussion

NP priors for a high dimensional parameter f based on DPs or Dirichlet distributions do not easily facilitate partial prior information about arbitrary functionals $\theta = \theta(f)$. Attempts to make such priors informative about θ can make the prior undesirably informative = about other aspects of f.

In this paper, we have presented a simple solution to this problem, via construction of an MSP that can induce a target marginal prior on a functional θ but is otherwise as close as possible to a given canonical ‘non-informative’ NP prior. We have provided general posterior approximation schemes for such priors, based on simple modifications to standard MCMC routines for canonical NP priors. In two examples we have shown that the MSP behaves as expected: given accurate prior information, the MSP provides improved estimation for θ compared with non-informative priors, while providing similar or better estimation performance for other aspects of the unknown parameter f.

One barrier to the adoption of MSPs is that the posterior approximation schemes that we have presented require that the ratio $p_1(\theta)/p_0(\theta)$ be computable, where p_1 is the desired informative prior for θ and p_0 is the prior induced on θ by a canonical prior π_0. Generally, p_0 will not have a closed form and so must be approximated numerically or otherwise. If the dimension of θ is small, it should generally be feasible to approximate p_0 with a kernel density estimate, or by a simple parametric family. If θ is high dimensional, then other approximation strategies will be required, such as approximating the joint density of θ as a product density (i.e. assuming independence of subvectors of θ) or perhaps by using mixture models. The latter strategy is more flexible than the former, but it doubles the modelling efforts in any given problem by requiring one to estimate p_0 before estimating f.

Replication code is available at the second author's Web site: http://www.stat.washington.edu/hoff/Code/kessler_hoff_dunson_2012.

Acknowledgements

David Kessler's work was partially supported by National Institute of Environmental Health Sciences training grant T32ES007018. Peter Hoff's work was partially supported by National Institute of Child Health and Development grant 1R01HD067509-01A1. David Dunson's work was supported by award R01ES017436 from the National Institute of Environmental Health Sciences.

Appendix A: Proofs for Section 2

A.1. Proof of theorem 1

Let $\mathscr{A}$ be the Borel sets of a Hausdorff space $\mathscr{F}$ and let θ be a measurable map from $(\mathscr{F}, \mathscr{A})$ to the measurable space $(\Theta, \mathscr{B})$. Let P_0 be the probability measure over $\mathscr{B}$ defined by $P_0(B) = \pi\left(\theta^{-1}B\right) \forall B \in \mathscr{B}$. The results of Hoffman-Jørgensen (1971) give the existence of a regular conditional probability function $\Lambda_0(A|\theta)$ such that $\Lambda_0(A|\cdot)$ is $\mathscr{B}$ measurable for each $(\mathscr{A} \in \mathscr{A}, \Lambda_0(\cdot|\theta)$ is a probability distribution over $\mathscr{A}$ for each $\theta \in \Theta$, and that $\Lambda_0\{A|\theta(f)\}$ is a version of the conditional probability of A given $\mathscr{B}$, in that

$$E_0\left[1\left\{\theta(f) \in B\right\} \times \pi_0\left\{A|\theta(f)\right\}\right] \equiv \int_B \Lambda_0(A|\theta) P_0(\mathrm{d}\theta) = \pi_0\left(A \cap \theta^{-1}B\right) \qquad \forall B \in \mathscr{B},$$

where θ represents either the function mapping $\mathscr{F}$ or a point in Θ, depending on the context.

Let P_1 be a probability measure on $\mathscr{B}$ such that $P_1 \ll P_0$. Define $\pi_1 : \mathscr{A} \to [0, 1]$ by

$$\pi_1(A) = \int \Lambda_0(A|\theta) P_1(\mathrm{d}\theta).$$

Then clearly $0 = \pi_1(\varnothing) \leq \pi_1(A) \leq \pi_1(\mathscr{F}) = 1$ for all $A \in \mathscr{A}$ Additionally, for a countable disjoint collection of sets $\{A_1, A_2, \ldots\} \subset \mathscr{A}$ with $A = \bigcup A_i$, we have

$$\begin{aligned}
\pi_1(A) &= \int \Lambda_0(A|\theta) P_1(\mathrm{d}\theta) \\
&= \int \sum_{i=1}^{\infty} \Lambda_0(A_i|\theta) P_1(\mathrm{d}\theta) \\
&= \sum_{i=1}^{\infty} \int \Lambda_0(A_i|\theta) P_1(\mathrm{d}\theta) \\
&= \sum_{i=1}^{\infty} \pi_1(A_i),
\end{aligned}$$

where the second-to-last line follows from the monotone convergence theorem. Therefore, $\pi_1(A)$ is a probability measure on $(\mathscr{F}, \mathscr{A})$. To compute the marginal distribution of π_1, let $B \in \mathscr{B}$ and $h(\theta) = \mathrm{d}P_1 / \mathrm{d}P_0$. Then

$$\begin{aligned}
\pi_1\left(\theta^{-1}B\right) &= \int \Lambda_0\left(\theta^{-1}B|\theta\right) P_1(\mathrm{d}\theta) \\
&= \int \Lambda_0\left(\theta^{-1}B|\theta\right) h(\theta) P_0(\mathrm{d}\theta).
\end{aligned}$$

The Radon–Nikodym derivative $h(\theta)$ is positive and measurable, and so we can express $h(\theta)$ as the limit of simple functions, $h(\theta) = lim_{n\to\infty} \Sigma_{k=1}^{n} h_{n,k} 1(\theta \in B_{n,k})$. By the monotone convergence theorem we have

$$\begin{aligned}\pi_1(\theta^{-1}B) &= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \int \Lambda_0(\theta^{-1}B|\theta)\, 1(\theta \in B_{n,k})\, P_0(\mathrm{d}\theta) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \Lambda_0\{\theta^{-1}(B \cap B_{n,k})\} \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} P_0(B \cap B_{n,k}) = \int_B h(\theta)\, P_0(\mathrm{d}\theta) = \int_B \frac{\mathrm{d}P1}{\mathrm{d}P_0}(\theta)\, P_0(\mathrm{d}\theta) = P_1(B).\end{aligned}$$

Finally, the Radon–Nikodym derivative of π_1 with respect to π_0 can be found via a similar calculation: for any $A \in \mathscr{A}$,

$$\begin{aligned}\pi_1(A) &= \int \Lambda_0(A|\theta)\, P_1(\mathrm{d}\theta) \\ &= \int \Lambda_0(A|\theta)\, h(\theta)\, P_0(\mathrm{d}\theta) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \int \pi_0(A|\theta)\, 1(\theta \in B_{n,k})\, P_0(\mathrm{d}\theta) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \pi_0(A \cap \theta^{-1}B_{n,k}) \\ &= \int_A \left[\lim_{n\to\infty} \sum_{k=1}^{n} h_k 1\{\theta(f) \in B_{n,k}\}\right] \pi_0(\mathrm{d}f) \\ &= \int_A h\{\theta(f)\}\, \pi_0(\mathrm{d}f) = \int_A \frac{\mathrm{d}P_1}{\mathrm{d}P_0}\{\theta(f)\}\, \pi_0(\mathrm{d}f).\end{aligned}$$

A.2. Proof of lemma 2

Let $\mathscr{A}$ be the Borel sets of a Hausdorff space $\mathscr{F}$. For $k \in \{0, 1\}$ let π_k be a probability measure on $(\mathscr{F}, \mathscr{A})$ and let P_k be the measure on $(\Theta, \mathscr{B})$ induced by the measurable map $\theta: \mathscr{F} \to \Theta$. Recall that, if $\pi_1 \ll \pi_0$, then the Kullback–Leibler divergence $D(\pi_1 \| \pi_0)$ is infinite. In contrast, we shall show that, if $\pi_1 \ll \pi_0$ and $P_1 \ll P_0$, then the Kullback–Leibler divergence $D(\pi_1 \| \pi_0)$ of π_0 from π_1 can be expressed in terms of marginal and conditional densities with respect to a common dominating measure, and that, if P_1 and P_0 are fixed, the divergence is minimized by matching the conditional distributions of π_0 and π_1.

Let μ be a dominating measure for π_0 and π_1, and let ν be a dominating measure for P_1 and P_0. The results of Hoffman-Jørgensen (1971) give the existence of a regular conditional probability function $\tilde{\Lambda}_0(\cdot|\cdot): \mathscr{A} \times \Theta \to [0, 1]$ with the properties that were described in the proof of theorem 1. Now, for each $A \in \mathscr{A}$ and $\theta \in \Theta$, define

$\Lambda_0(A|\theta) = \tilde{\Lambda}_0(A|\theta) \times 1\{\pi_0(A) > 0\}$. It is easy to check that this is measurable in θ for each $A \in \mathscr{A}$, is a version of the conditional probability of A given θ and is dominated by π_0, and therefore by μ, for each $\theta \in \theta$. Therefore, the measures $\{\Lambda_0(\cdot|\theta): \theta \in \Theta\}$ form a dominated class with densities $\{\lambda_0(\cdot|\theta): \theta \in \Theta\}$ with respect to μ. By Tonelli's theorem we can write

$$Pr_0(\{f, \theta\} \in A \times B) \equiv \pi_0\left(A \cap \theta^{-1}B\right) = \int_B \int_A \lambda_0(f|\theta)\, p_0(\theta)\, \mu(\mathrm{d}f) \times \nu(\mathrm{d}\theta),$$

and so π_0 has a density $\lambda_0(f|\theta)p_0(\theta)$ with respect to the product measure $\mu \times \nu$. The same construction can be made for π_1, giving the existence of a conditional probability density $\lambda_1(f|\theta)$ for which

$$Pr_1(\{f,\theta\} \in A \times B) \equiv \pi_1\left(A \cap \theta^{-1}B\right) = \int_B \int_A \lambda_1(f|\theta)\, p_1(\theta)\, \mu(\mathrm{d}f) \times \nu(\mathrm{d}\theta).$$

Letting $B = \{\theta: p_0(\theta) > 0\}$, the Kullback–Leibler divergence is

$$\begin{aligned} D(\pi_1 \| \pi_0) &= \int_\Theta \int_{\mathscr{F}} log\left\{\frac{\lambda_1(f|\theta)p_1(\theta)}{\lambda_0(f|\theta)p_0(\theta)}\right\} \lambda_1(f|\theta)\, p_1(\theta)\, \mu(\mathrm{d}f) \times \nu(\mathrm{d}\theta) \\ &= \int_B \int_{\mathscr{F}} log\left\{\frac{\lambda_1(f|\theta)p_1(\theta)}{\lambda_0(f|\theta)p_0(\theta)}\right\} \lambda_1(f|\theta)\, p_1(\theta)\, \mu(\mathrm{d}f) \times \nu(\mathrm{d}\theta) \\ &= \int_B \int_{\mathscr{F}} log\left\{\frac{\lambda_1(f|\theta)}{\lambda_0(f|\theta)}\right\} \lambda_1(f|\theta)\, p_1(\theta)\, \mu(\mathrm{d}f) \times \nu(\mathrm{d}\theta) + \int_B log\left\{\frac{p_1(\theta)}{p_0(\theta)}\right\} p_1(\theta)\, \nu(\mathrm{d}\theta) \\ &= \int_\Theta D\{\Lambda_1(\cdot|\theta) \| \Lambda_0(\cdot|\theta)\} P_1(\mathrm{d}\theta) + D(P_1 \| P_0), \end{aligned} \tag{7}$$

where the last line follows from the assumption that $P_1 \ll P_0$ and so $P_1(B) = P_0(B) = 1$. Since the integrand in equation (7) is always greater than or equal to 0, we have $D(\pi_1 \| \pi_0) \geq D(P_1 \| P_0)$ with equality when $\Lambda_1(\cdot|\theta) = \Lambda_0(\cdot|\theta)$ for θ almost everywhere P_1. Of course, if $D(P_1 \| P_0) = \infty$ then equation (7) is 'minimized' at ∞ for all π_1.

References

Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems. Statist. Sci. 1995; 10:3–41.

Bush CA, Lee J, MacEachern SN. Minimally informative prior distributions for non-parametric Bayesian analysis. J. R. Statist. Soc. B. 2010; 72:253–268.

Bush CA, MacEachern SN. A semiparametric Bayesian model for randomised block designs. Biometrika. 1996; 83:275–285.

Csiszar I. I-divergence geometry of probability distributions and minimization problems. Ann. Probab. 1975; 3:146–158.

Dawid AP, Lauritzen SL. Hyper Markov laws in the statistical-analysis of decomposable graphical models. Ann. Statist. 1993; 21:1272–1317.

Dobra A, Massam H. The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Statist. Methodol. 2010; 7:240–253.

Dunson DB, Xing C. Nonparametric Bayes modeling of multivariate categorical data. J. Am. Statist. Ass. 2009; 104:1042–1051.

Escobar MD. Estimating normal means with a Dirichlet process prior. J. Am. Statist. Ass. 1994; 89:268–277.

Escobar MD, West M. Bayesian density estimation and inference using mixtures. J. Am. Statist. Ass. 1995; 90:577–588.

Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1973; 1:209–230.

Fox EB, Sudderth EB, Jordan MI, Willsky AS. A sticky HDP-HMM with application to speaker diarization. Ann. Appl. Statist. 2011; 5:1020–1056.

Goodman LA. How to ransack social mobility tables and other kinds of cross-classification tables. Am. J. Sociol. 1969; 75:1–40.

Hoffmann-Jørgensen J. Existence of conditional probabilities. Math. Scand. 1971; 28:257–264.

Kalli M, Griffin J, Walker SG. Slice sampling mixture models. Statist. Comput. 2011; 21:93–105.

Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J. Am. Statist. Ass. 1995; 90:928–934.

Lijoi A, Regazzini E. Means of a Dirichlet process and multiple hypergeometric functions. Ann. Probab. 2004; 32:1469–1495.

Lo AY. On a class of Bayesian nonparametric estimates: I, Density estimates. Ann. Statist. 1984; 12:351–357.

MacEachern SN. Estimating normal means with a conjugate-style Dirichlet process prior. Communs Statist. Simuln Computn. 1994; 23:727–741.

MacEachern SN, Müller P. Estimating mixture of Dirichlet process models. J. Computnl Graph. Statist. 1998; 7:223–238.

McAuliffe JD, Blei DM, Jordan MI. Nonparametric empirical Bayes for the Dirichlet process mixture model. Statist. Comput. 2006; 16:5–14.

Moala FA, O'Hagan A. Elicitation of multivariate prior distributions: a nonparametric Bayesian approach. J. Statist. Planng Inf. 2010; 140:1635–1655.

Müller P, Erkanli A, West M. Bayesian curve fitting using multivariate normal mixtures. Biometrika. 1996; 83:67–79.

Neal, RM. Regression and classification using Gaussian process priors. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics 6. Oxford University Press; Oxford: 1999. p. 475-501.

Petrone S, Guindani M, Gelfand AE. Hybrid Dirichlet mixture models for functional data. J. R. Statist. Soc. B. 2009; 71:755–782.

Pitman, J. Statistics, Probability and Game Theory. Institute of Mathematical Statistics; Hayward: 1996. Some developments of the Blackwell-MacQueen urn scheme; p. 245-267.

Sudderth, EB.; Jordan, MI. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In: Koller, D.; Schuurmans, D.; Bengio, Y.; Bottou, L., editors. Advances in Neural Information Processing Systems. Vol. 21. 2009. p. 1585-1592.

Walker SG. Sampling the Dirichlet mixture model with slices. Communs Statist. Simuln Computn. 2007; 36:45–54.

Yamato H. Characteristic functions of means of distributions chosen from a Dirichlet process. Ann. Probab. 1984; 12:262–267.

Yau C, Papaspiliopoulos O, Roberts GO, Holmes C. Bayesian non-parametric hidden Markov models with applications in genomics. J. R. Statist. Soc. B. 2011; 73:37–57.

Zou F, Huang H, Lee S, Hoeschele I. Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction. Genetics. 2010; 186:385–394. [PubMed: 20551445]

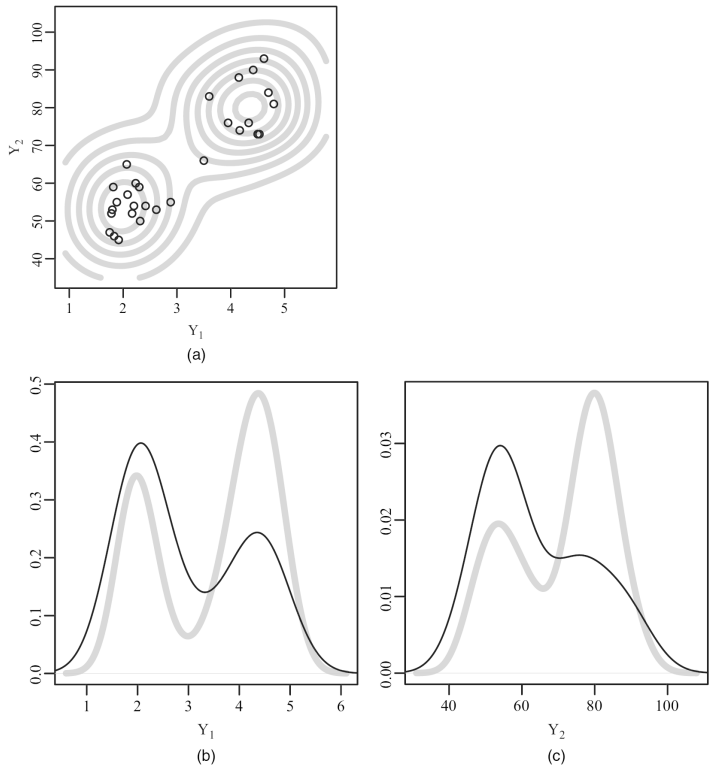

Fig. 1.

Population and sample: (a) contours of the population density and a scatter plot of the $n = 30$ randomly sampled observations; (b), (c) marginal densities for the population (——) and sample (——)

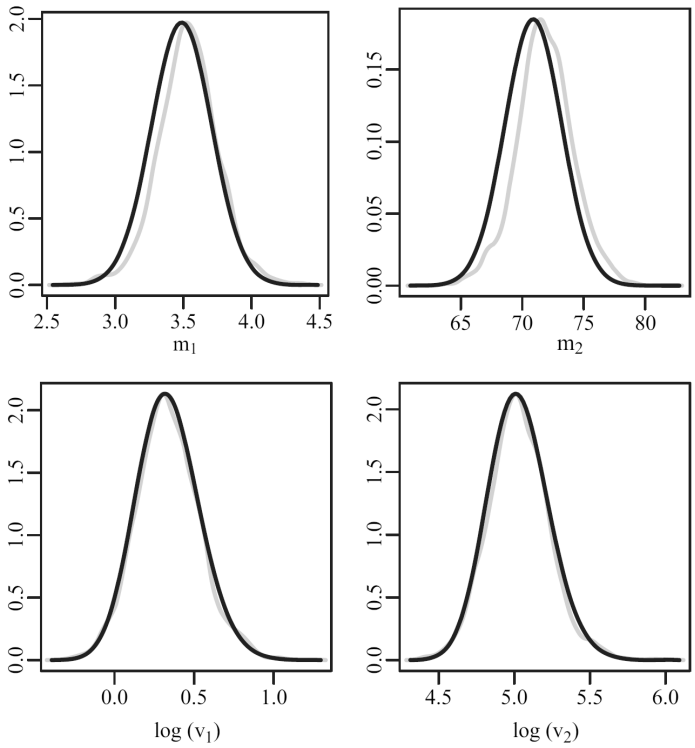

Fig. 2.

p_1-priors (——) and kernel density estimates of priors induced by π_0^{I} (——)

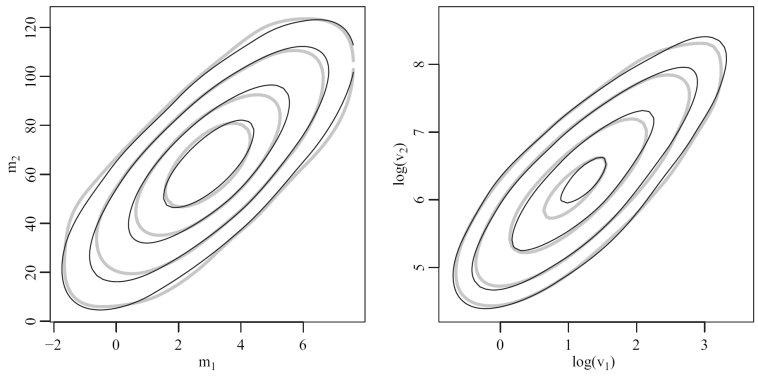

Fig. 3.

Comparison of approximated p_0 (▬▬) and p_0 induced by π_0^N (——)

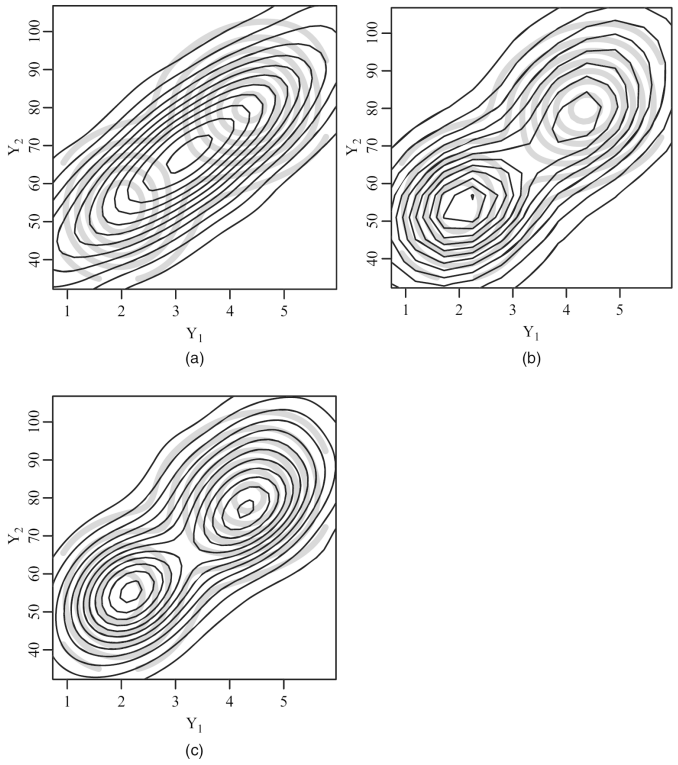

Fig. 4.

Contour plots of the posterior predictive density (————) and the population density (▬▬▬), under (a) π_0^{I}, (b) π_0^{N} and (c) π_1

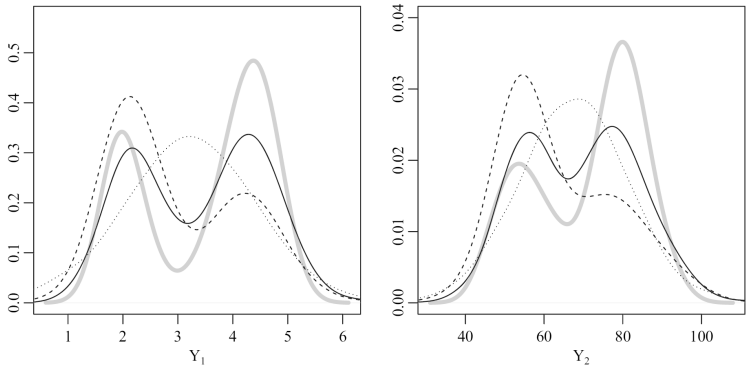

Fig. 5.

Marginal population densities and estimates from the three priors: ·····, informative DPMM; – – – – –, non-informative DPMM; ————, MSP; ————, population

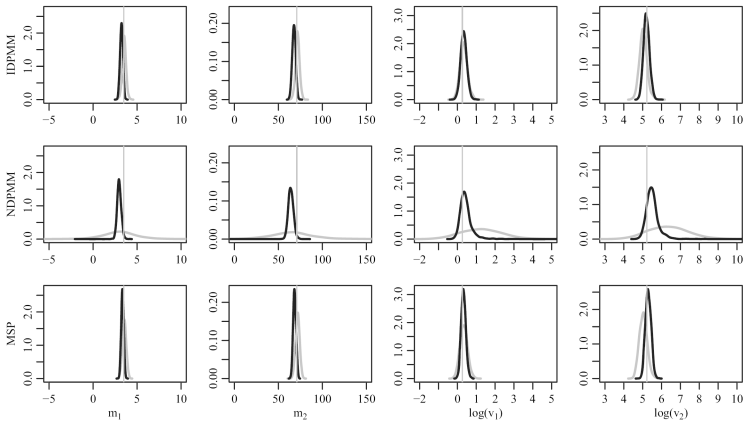

Fig. 6.

Priors (——) and posteriors (——) for the marginal means and log-variances: |, corresponding population values derived from the full (n = 272) data set

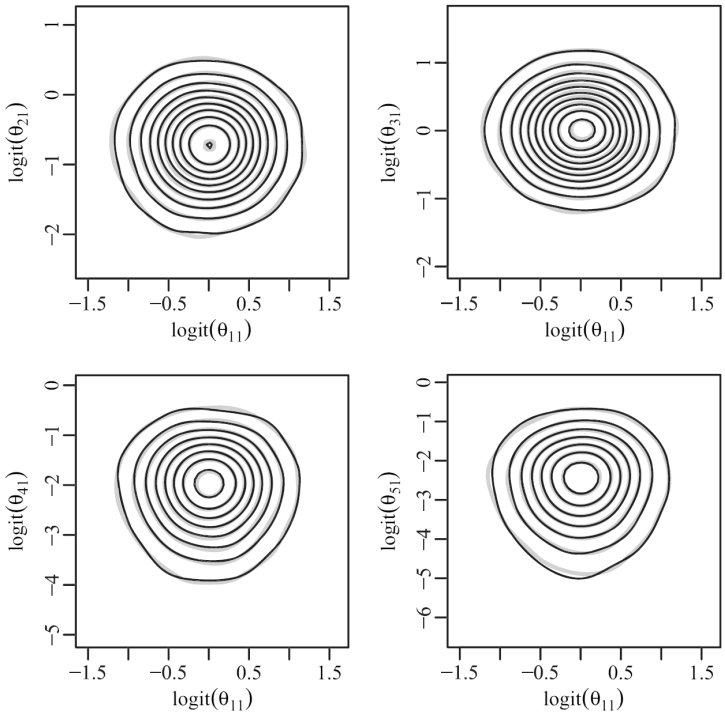

Fig. 7.

Comparison of approximated p_0 (——) and p_0 induced by π_0^N (——) for a subset of the margins: to facilicate comparison, a logit transform was used

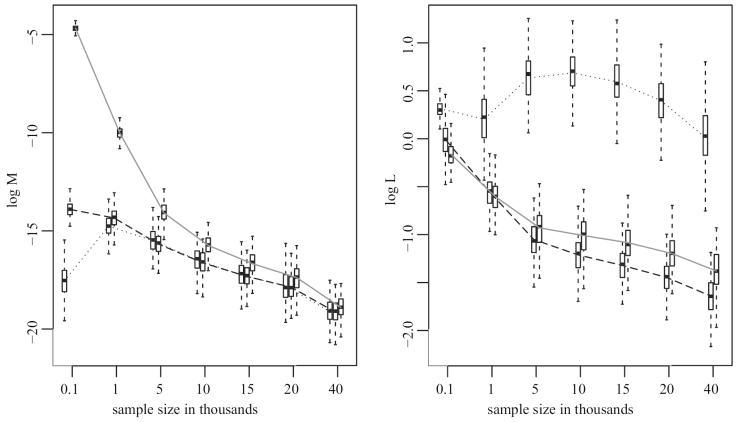

Fig. 8.

Comparison of the *M*- and *L*-metrics on the log-scale for π_0^{I}(• • • • • •), π_0^{N} (————) and π_1 (— — —) at various sample sizes