ELSEVIER

Statistics & Probability Letters 36 (1997) 69–83

STATISTICS & PROBABILITY LETTERS

A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability[1]

Sonia Petrone[a], Adrian E. Raftery[b,*]

[a] *Dipartimento di Economia Politica e Metodi Quantitativi, Universita de Pavia I-27100 Pavia, Italy*
[b] *Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322, USA*

Received July 1996; received in revised form March 1997

Abstract

We consider Bayesian nonparametric inference for continuous-valued partially exchangeable data, when the partition of the observations into groups is unknown. This includes change-point problems and mixture models. As the prior, we consider a mixture of products of Dirichlet processes. We show that the discreteness of the Dirichlet process can have a large effect on inference (posterior distributions and Bayes factors), leading to conclusions that can be different from those that result from a reasonable parametric model. When the observed data are all distinct, the effect of the prior on the posterior is to favor more evenly balanced partitions, and its effect on Bayes factors is to favor more groups. In a hierarchical model with a Dirichlet process as the second-stage prior, the prior can also have a large effect on inference, but in the opposite direction, towards more unbalanced partitions. © 1997 Elsevier Science B.V.

Keywords: Bayesian nonparametric inference; Dirichlet process; Hierarchical model; Partial exchangeability; Partition models

1. Introduction

We consider a Bayesian nonparametric model for partially exchangeable data, when the allocation of cases to groups is unknown. This includes change-point problems, mixture models and, more generally, can be seen as a nonparametric extension of hierarchical partition models (Consonni and Veronese, 1995, and references therein). As a prior, we consider a mixture of products of Dirichlet processes. This prior has been used by Cifarelli and Regazzini (1978) as a generalization of the Dirichlet process of Ferguson (1973, 1974). The peculiarity of the problem we consider is that the partition of the data into groups is unknown and, in fact, is often one of the main objects of inference.

We compute the posterior distributions of the quantities of interest in the proposed model. Our emphasis is on the effects of the discreteness of the Dirichlet process on inference about the unknown partition (posterior distributions and Bayes factors) in this situation. The phenomenon illustrated here is due to the structure of

* Corresponding author.
[1] This research was supported by ONR Grants No. N-00014-91-J-1074 and N00014-96-1-0192 and by grants from MURST, Rome.

0167-7152/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved
PII S0167-7152(97)00050-3

the distribution of a sample from a Dirichlet process and in this sense it can arise, more generally, in different contexts. For example, it is at the base of the difficulties in using the Dirichlet process prior in goodness of fit testing, pointed out by Carota and Parmigiani (1995).

More specifically, the problem discussed in the paper can be explained as follows: The Dirichlet process almost surely selects a discrete distribution function (d.f.) (Blackwell, 1973; Blackwell and McQueen, 1973). As a result, when used as a prior distribution for a sequence of exchangeable random variables (r.v.'s), it gives positive probability to ties in the data. Nevertheless, the Dirichlet process is also used as a prior in situations where the data are "continuous", i.e., when the probability of ties in the data is very small, or zero. In many situations, the discreteness of the Dirichlet process has no relevant effects. However, in the situation considered in this paper, when the data are partially exchangeable with an unknown partition, it turns out to have a large effect on inferences, in a way that is not clearly part of the prior specification.

When the observed data are all distinct, the effect of the Dirichlet process prior is to favor nearly equally sized groups and to make partitions with very unequally sized groups unlikely a posteriori. It also tends to greatly inflate Bayes factors for more groups against less groups when the number of groups is unknown a priori. Equivalently, it tends to concentrate the posterior distribution of the number of groups on higher values.

Intuitively, the reason is that, when the data are all distinct, the likelihood can be written as the product of two factors: (1) the likelihood under the parametric model that is the mean of the (product of) Dirichlet process(es) prior, and (2) the probability of having no ties in the data. The effect we have observed is due to the second factor: the probability of no ties is higher when the number of groups is bigger, and, given the number of groups, when these are of equal size rather than when they are very unequal, sometimes much more so. What is perhaps surprising is the magnitude of the effect.

This leads to suggestions and warnings about the use of a Dirichlet process prior in this situation. The warning is as follows: If the true distribution is close to the prior mean, we would prefer the posterior distribution from the nonparametric model to be close to that from the parametric model corresponding to the prior mean. However, the presence of the second factor in the likelihood shows that this will not be the case in general, and that the two posterior distributions may be very different. The point that we are making here applies to the situation where the partition into groups is unknown. The discreteness of the Dirichlet process is not so crucial if the partition of data into groups is known, in which case the probability of no ties is a constant depending only on known quantities (Cifarelli and Regazzini, 1978).

In Section 2, we discuss the likelihood of a sample from a Dirichlet process. In Section 3 we illustrate the effect discussed here in the context of a simple change-point problem. In Section 4, we give results for the more general situation of partial exchangeability where the partition is unknown. In Section 5, we make some remarks about the situation where the Dirichlet process is used as a prior at the second stage of a Bayesian hierarchical model. In this situation, the prior can also have a large effect on inference, but in the opposite direction, towards more unbalanced partitions.

2. The likelihood of a sample from a Dirichlet process

Let $X_1, \ldots, X_n$ be real valued r.v.'s, which are a sample from a Dirichlet process, i.e., $X_1, \ldots, X_n \mid F$ are conditionally i.i.d. according to the d.f. F, and F is a Dirichlet process whose parameter is a measure α: $F \sim \mathcal{D}(\alpha)$. This induces a probability P on the sample space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ (Ferguson, 1973). The predictive distribution of $X_j \mid X_1 = x_1, \ldots, X_{j-1} = x_{j-1}$ for any $j = 1, \ldots, n$, is

$$P(X_j \leqslant x \mid X_1 = x_1, \ldots, X_{j-1} = x_{j-1}) = \frac{1}{M+j-1}\alpha((-\infty, x]) + \frac{j-1}{M+j-1}F_{j-1}(x)$$

where $M = \alpha(\mathbb{R})$ and F_{j-1} is the empirical d.f. of the first $(j-1)$ observations, $(x_1,\ldots,x_{j-1})$. We will restrict our attention to the case where the parameter α is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}$, with density $\mathrm{d}\alpha/\mathrm{d}\lambda(x) = Mf_0(x)$.

From the form of the predictive distribution, proceeding as in Korwar and Hollander (1973) and Antoniak (1974, Lemma 1), it can be shown that P, i.e. the joint probability law of $(X_1,\ldots,X_n)$, gives positive probability to certain collections of hyperplanes in the sample space $\mathbb{R}^n$. Let us introduce the set

$$\mathscr{C}(D,n_1,\ldots,n_D) = \{(x_1,\ldots,x_n) : \text{ there are } D \text{ distinct values; one of them is repeated}$$

$$n_1 \text{ times, one is repeated } n_2 \text{ times, } \ldots, \text{ one is repeated } n_D \text{ times }\},$$

where $(n_1,\ldots,n_D)$ is a sequence of integers such that $n_i > 0$, $n_1 \leqslant n_2 \leqslant \cdots \leqslant n_D$ and $\sum_{i=1}^{D} n_i = n$. An ordering (o) for a sequence $(x_1,\ldots,x_n) \in \mathscr{C}(D,n_1,\ldots,n_D)$ is a subset of the integers $(1,2,\ldots,n)$ showing the positions where a value different from the previous ones appears. For example, if $(x_1,\ldots,x_5) = (1,1,9,8,1)$, the ordering is $(1,3,4)$, since new values appear in these positions.

Now, let us denote by $\lambda_{(D;n_1,\ldots,n_D,\mathrm{o})}$ a measure defined on $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ which is degenerate on the hyperplane

$$\{(x_1,\ldots,x_n) \in \mathbb{R}^n : (x_1,\ldots,x_n) \in \mathscr{C}(D,n_1,\ldots,n_D) \text{ and satisfies the ordering (o)}\}$$

and has D-dimensional marginal on the D distinct coordinates coincident with Lebesgue measure on $(\mathbb{R}^D,\mathscr{B}(\mathbb{R}^D))$ So, e.g., for $n = 3$, if (o) denotes the appearance of two distinct values at positions $(1,3)$, $\lambda_{(2;1,2,\mathrm{o})}$ is a measure on $(\mathbb{R}^3,\mathscr{B}(\mathbb{R}^3))$, degenerate on

$$\{(x_1,x_2,x_3) \in \mathbb{R}^3 : x_1 = x_2 \neq x_3\},$$

having Lebesgue measure on $\mathbb{R}^2$ as the marginal of the two distinct coordinates. Since the measures $\lambda_{(D;n_1,\ldots,n_D,\mathrm{o})}$ are defined on the same space, $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$, their sum can be defined as

$$\psi(B) = \sum_{D=1}^{n} \sum_{(n_1,\ldots,n_D)} \sum_{(\mathrm{o})} \lambda_{(D,n_1,\ldots,n_D,\mathrm{o})}(B), \quad B \in \mathscr{B}(\mathbb{R}^n).$$

It can be shown that P is absolutely continuous with respect to ψ, with derivative

$$\frac{\mathrm{d}P}{\mathrm{d}\psi}(x_1,\ldots,x_n) \equiv f_n(x_1,\ldots,x_n)$$

$$= \sum_{D=1}^{n} \sum_{(n_1,\ldots,n_D)} \frac{M^D}{M^{[n]}} \prod_{i=1}^{D} (n_i-1)! \prod_{i=1}^{D} f_0(x'_i)\, I_{\mathscr{C}(D,n_1,\ldots,n_D)}(x_1,\ldots,x_n),$$

where $M^{[k]} = M(M+1)\cdots(M+k-1)$; $M^{[0]} = 1$; and $(x'_1,\ldots,x'_D)$ is the vector of distinct values among $(x_1,\ldots,x_n)$, with x'_i repeated n_i times. In other words,

$$f_n(x_1,\ldots,x_n) = \frac{M^D}{M^{[n]}} \prod_{i=1}^{D} (n_i-1)! \prod_{i=1}^{D} f_0(x'_i), \tag{1}$$

if $(x_1,\ldots,x_n) \in \mathscr{C}(D,n_1,\ldots,n_D)$, having D distinct values $(x'_1,\ldots,x'_D)$ repeated $n_1,\ldots,n_D$ times.

The interpretation of the two components $M^D/M^{[n]} \prod_{i=1}^{D}(n_i-1)!$ and $\prod_{i=1}^{D} f_0(x'_i)$ of the likelihood (1) is as follows. The first is the probability that $(X_1,\ldots,X_n)$ belongs to $\mathscr{C}(D,n_1,\ldots,n_D)$ and has a given ordering. The second is the joint density of the D distinct values, given the configuration $\mathscr{C}(D,n_1,\ldots,n_D)$; indeed, Korwar and Hollander (1973) show that, given the number D of distinct values, these are i.i.d. according to f_0.

If $(X_1,\ldots,X_n)$ is a sample from a mixture of Dirichlet processes (Antoniak, 1974), i.e., $(X_1,\ldots,X_n)$ are conditionally i.i.d. according to F, $F \mid \theta \sim \mathscr{D}(\alpha(\cdot,\theta))$ and $\Theta \sim H(\theta)$, and if $\alpha(\cdot,\theta)$ is absolutely continuous with respect to Lebesgue measure for any θ, then the likelihood of $(x_1,\ldots,x_n \mid \theta)$ is given, as in (1), by

$$f_n(x_1,\ldots,x_n;\theta) = \frac{\alpha(\mathbb{R};\theta)^D}{\alpha(\mathbb{R};\theta)^{[n]}} \prod_{i=1}^{D} (n_i - 1)! \prod_{i=1}^{D} f_0(x'_i;\theta). \tag{2}$$

This differs slightly from Lemma 1 of Antoniak (1974), but only in the specification of the dominating measure.

3. A simple change-point example

3.1. Posterior distribution of the change point

To illustrate the general phenomenon, we will describe how it affects inference for a simple change-point problem. Data $(X_1,\ldots,X_n)$ are assumed to arise from the model

$$X_i \sim F_1, \quad i = 1,\ldots,C,$$
$$X_i \sim F_2, \quad i = C+1,\ldots,n,$$

where C is the change point, which is unknown. Nonparametric Bayesian inference is carried out by assigning (mixture of) Dirichlet process priors to F_1 and F_2.

The model can be written

$$P(X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_n \leqslant x_n | C, F_1, F_2) = \prod_{i=1}^{C} F_1(x_i) \prod_{i=C+1}^{n} F_2(x_i), \tag{3}$$

where C is the unknown change point. This means that the observations are conditionally i.i.d. according to F_1 up to time C and i.i.d. according to F_2 from time $(C+1)$ on. When $C = 0$ or $C = n$ there is no change point and the data are conditionally i.i.d. as F_2, or F_1, respectively.

The prior assumes that C and (F_1, F_2) are independently distributed, C with probability mass function $p(c)$ and (F_1, F_2) as a mixture of products of Dirichlet processes (Cifarelli and Regazzini, 1978). The prior distribution is, thus,

$$(F_1, F_2, C) \sim p(c) \int \mathscr{D}(\alpha_1(\cdot;\theta_1)) \; \mathscr{D}(\alpha_2(\cdot;\theta_2)) \, \mathrm{d}H(\theta_1,\theta_2). \tag{4}$$

This model has been studied by Muliere and Scarsini (1985) in the special case of independent F_1 and F_2. A Gibbs sampler algorithm for approximating the posterior distributions in the above model is provided in Mira and Petrone (1995). On the lines of Section 2, it can be shown that, when $\alpha_1(\cdot;\theta_1)$ and $\alpha_2(\cdot;\theta_2)$ have densities $M_1 f_1(\cdot;\theta_1)$ and $M_2 f_2(\cdot;\theta_2)$, respectively, with respect to Lebesgue measure, the likelihood is

$$f(x_1, x_2, \ldots, x_n \mid c, \theta_1, \theta_2,) = \frac{1}{M_1^{[r]}} \prod_{i=1}^{c\,*} M_1 f_1(x_i;\theta_1) \frac{1}{M_2^{[n-c]}} \prod_{i=c+1}^{n\,*} M_2 f_2(x_i;\theta_2), \tag{5}$$

where the $*$ indicates that the product is taken over distinct values only.

From (5) the posterior distribution of (C, θ_1, θ_2) can be computed by applying Bayes' theorem. In particular, in the continuous case and if the observations $(x_1,\ldots,x_n)$ are all distinct, we have

$$p(c \mid x_1,\ldots,x_n) \propto k(c, M_1, M_2, n)\, I(c)\, p(c), \tag{6}$$

where, for $c = 0, 1, \ldots, n$, we let

$$I(c) = \int \prod_{i=1}^{c} f_1(x_i;\theta_1) \prod_{i=c+1}^{n} f_2(x_i;\theta_2)\, \mathrm{d}H(\theta_1,\theta_2),$$

and

$$k(c,M_1,M_2,n) = \frac{M_1^c\, M_2^{(n-c)}}{M_1^{[c]}\, M_2^{[n-c]}}.$$

The factor $k(c,M_1,M_2,n)$ is the probability of there being no ties in the data, given that $C = c$. The probability of the data values all being distinct is

$$\begin{aligned}\sum_{c=0}^{n} P(X_1 \neq X_2 \neq \cdots \neq X_n \mid c)\; p(c) &= \sum_{c=0}^{n} \frac{M_1^{c-1}}{(M_1+1)^{[c-1]}} \frac{M_2^{n-c-1}}{(M_2+1)^{[n-c-1]}}\; p(c) \\ &= \sum_{c=0}^{n} \frac{M_1^c}{M_1^{[c]}} \frac{M_2^{n-c}}{M_2^{[n-c]}} p(c) \\ &= \sum_{c=0}^{n} k(c,M_1,M_2,n)\; p(c).\end{aligned}$$

We note that the posterior (6) is equivalent, except for the factor $k(c,M_1,M_2,n)$, to what we obtain with the parametric model where $(X_1,\ldots,X_n)$ have the joint conditional density

$$f(x_1,\ldots,x_n \mid c,\theta_1,\theta_2) = \prod_{i=1}^{r} f_1(x_i;\theta_1) \prod_{i=r+1}^{n} f_2(x_i;\theta_2). \tag{7}$$

Here, $I(c)$ is the (integrated) likelihood of c, and the posterior is given by (6) with $k(c,M_1,M_2,n) = 1$.

The factors $k(c;M_1,M_2,n)$ are plotted in Fig. 1 for the nonparametric model with $M_1 = M_2 = 1$, and sample sizes $n = 10$ and 50.The plots show that $k(c;M_1,M_2,n)$ is much bigger for values of c close to $n/2$, and that this behavior is more marked if n is larger. For the parametric model (7), in contrast, $k(c,M_1,M_2,n) \equiv 1$, a constant. When M_1 and M_2 are very large, corresponding to a large amount of prior information, this effect is attenuated, since for $M_1, M_2 \to +\infty$, (6) converges to the parametric result.

3.2. *Bayes factors*

Suppose we want to test the hypothesis of no change-point against the alternative that one change-point occurred. In our model, these hypotheses are expressed by: $\mathrm{H}_0 : C \in \{0,n\}$ and $\mathrm{H}_1 : C \in \{1,2,\ldots,n-1\}$. For this problem, we can compute the relevant Bayes factor (see Kass and Raftery, 1995 for a review).

The Bayes factor for H_0 versus H_1 is

$$B_{01} = \frac{\sum_{\{c=0,n\}} f(\underline{x} \mid c) p_0(c)}{\sum_{c=1}^{n-1} f(\underline{x} \mid c) p_1(c)}, \tag{8}$$

where

$$p_0(c) = \begin{cases} \dfrac{p(c)}{p(0)+p(n)} & \text{for } c = 0,n, \\ 0 & \text{otherwise,} \end{cases}$$

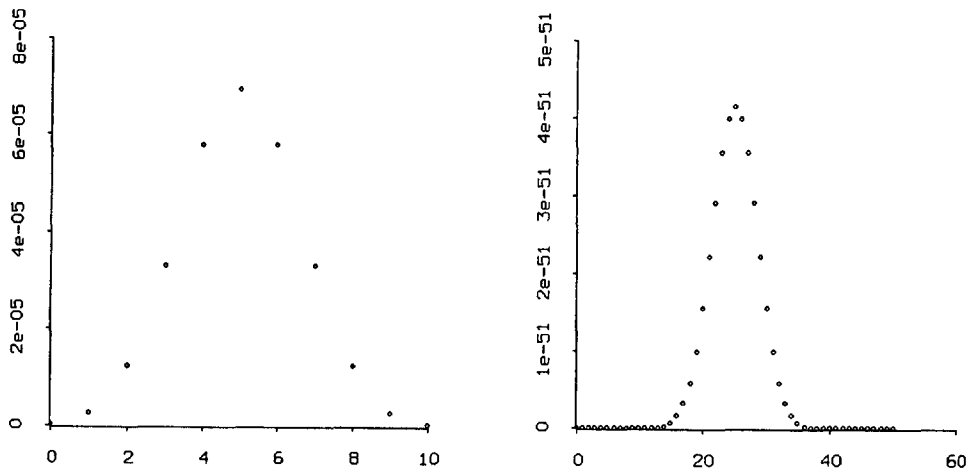

Fig. 1. Plot of factors $k(c, M_1, M_2, n)$, for $M_1 = M_2 = 1$; $n = 10$ and $n = 50$.

and

$$p_1(c) = \begin{cases} \dfrac{p(c)}{\sum_{c=1}^{n-1} p(c)} & \text{for } c = 1, \ldots, n-1, \\ 0 & \text{otherwise.} \end{cases}$$

In the continuous case, with $(x_1, \ldots, x_n)$ all distinct, the Bayes factor is given by

$$B_{01} = \frac{\sum_{\{c=0,n\}} p_0(c)\, k(c, M_1, M_2)\, I(c)}{\sum_{c=1}^{n-1} p_1(c)\, k(c, M_1, M_2)\, I(c)}. \tag{9}$$

As for the posterior distribution, the Bayes factor (9) is equivalent, except for the factor $k(c, M_1, M_2, n)$, to what we obtain with the parametric model (7); indeed, in the parametric case the Bayes factor is given by (9) with $k(c, M_1, M_2, n) = 1$. For a review of the Bayesian analysis of parametric change-point problems, see Raftery (1994).

3.3. A numerical example

In the simple nonparametric change-point model of the previous sections, let $f_i(\cdot, \theta_i)$ be a normal density with mean θ_i and known variance Σ^2, θ_1 and θ_2 be independent and normally distributed, with $\theta_i \sim \mathrm{N}(\mu_i, \tau_i)$ $i = 1, 2;$, and let $p(c)$ be chosen so that: $p(0) + p(n) = p(1) + \ldots + p(n-1)$. This gives probability $\frac{1}{2}$ to each of the two hypotheses: (no change-point: $c \in \{0, 1\}$) and (a change-point occurred: $c \in \{1, \ldots, n-1\}$), with $p(0) = p(n)$ and $p(1) = \cdots = p(n-1)$. We fix $\sigma = 1, \tau_1 = \tau_2 = 1, \mu_1 = 5, \mu_2 = 8, M_1 = M_2 = M$ and consider varying values of M. Also, let $N_k(\boldsymbol{\mu}, \Sigma)$ denote the k-variate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, $\boldsymbol{I}_k = (1, \ldots, 1)'$ be the $(k \times 1)$ vector of ones, J_k be the $(k \times k)$ matrix every element of which is unity, and I_k be the k-dimensional identity matrix.

Then it can be shown (Muliere and Scarsini, 1984) that $I(c)$ is the product of two normal densities: $N_c(\boldsymbol{\mu}_1, \Sigma_c)$ and $N_{n-c}(\boldsymbol{\mu}_2, \Sigma_{n-c})$, computed from $(x_1, \ldots, x_c)$ and $(x_{c+1}, \ldots, x_n)$ respectively, where

$$\Sigma_c = \sigma^2 I_c + \tau_1^2 J_c$$

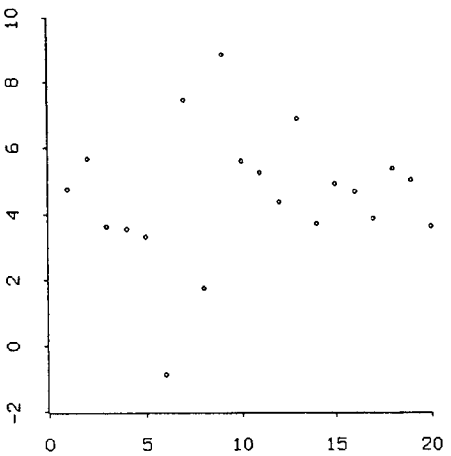

Fig. 2. Data: 20 observations from a translated Student's t distribution with two degrees of freedom.

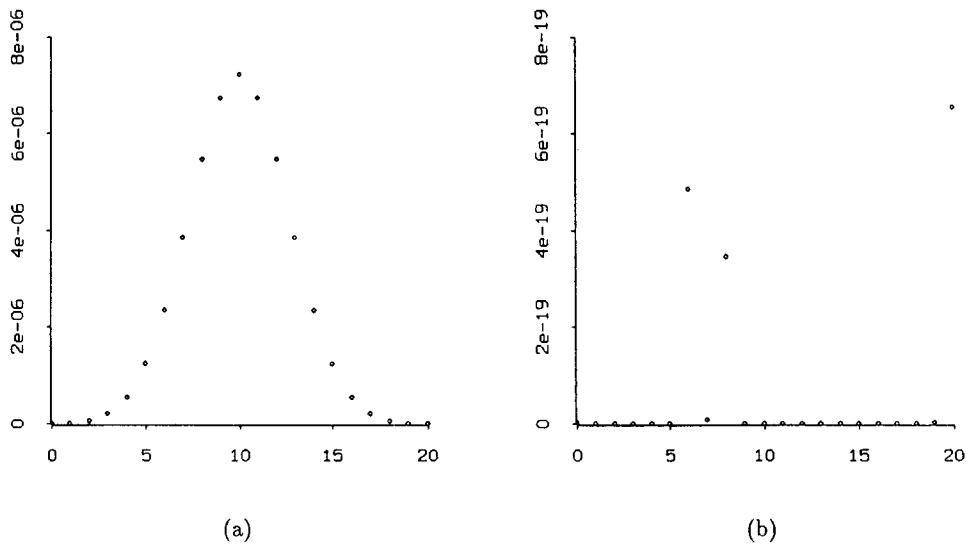

Fig. 3. Plot of factors (a)$k(c, M_1 = M_2 = 5, n = 20)$ and (b), $I(c)$.

and

$$\Sigma_{n-c} = \sigma^2 I_{n-c} + \tau_2^2 J_{n-c}.$$

We explore, for a simulated data set, the effect of the probability of ties on the posterior distribution of the change point, comparing the nonparametric result with that which we would obtain using the "dual" parametric model. The interest of a nonparametric approach lies mainly in its greater robustness with respect to misspecifications of the model. For this reason, although the prior guess on the d.f.'s is normal, we consider an example where the data are 20 values generated from a Student's t distribution with two degrees of freedom, translated by four (Fig. 2). Here there is no change point. The components $k(c, M_1, M_2, n)$ and $I(c)$ of the likelihood, for $M_1 = M_2 = 1$, are plotted in Fig. 3.

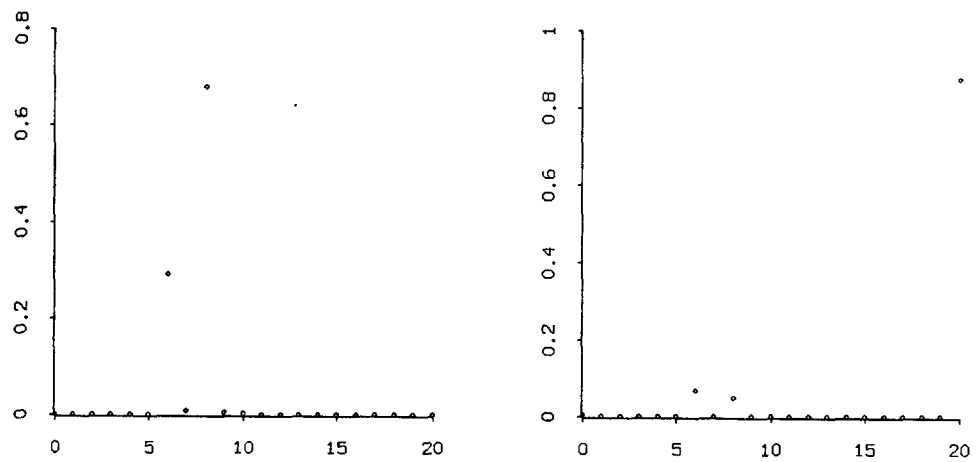

Fig. 4. Posterior distribution of the changepoint: (a) nonparametric model, $M_1 = M_2 = 1$; (b) parametric model.

Table 1
Bayes factors for H_0: no change, against H_1: a change point occurred

	Bayes factor
Parametric model	7.3
Nonparametric model, $M = 1$	0.00018
Nonparametric model, $M = 5$	0.011

The parametric (integrated) likelihood, $I(c)$, provides some evidence for the change point being at one of the times 20 (i.e. no change), 6 or 8. The parametric posterior distribution (with a prior for c that gives equal total weight to times representing no change as to times representing a change) is shown in Fig. 4(b). This is mostly concentrated on $c = 20$ (i.e. no change point). The nonparametric posterior distribution in Fig. 4(a) is very different. It suggests that there was almost certainly a change point at one of the times 6–8, and gives no weight at all to $c = 20$.

The same phenomenon affects the Bayes factor, as shown in Table 1. For the parametric model, the Bayes factor for no change is about 7.3, providing positive evidence against a change point. For the nonparametric model with $M = 1$, the Bayes factor is about 5000 : 1 *for* a change – a very misleading conclusion, and dramatically different from that based on the parametric model. For $M = 5$ the effect is less dramatic, but still strong.

The posterior distribution of the change point also affects the estimates of the distribution functions before and after the change. The estimate of $F_i(t)$ $(i = 1,2)$ is the average, with respect to the posterior of C, of the conditional expected value of $F_i(t)$ given the data and $C = c$. In the nonparametric case, the conditional expected value of $F_i(t) \mid c, x_1, \ldots, x_n$ is a weighted average between the prior guess about $F_i(t)$ and the empirical d.f. of the first c observations. One would hope that the component due to the empirical d.f.'s would result in a more robust estimate, compared with that obtained with a strict parametric assumption. Yet, the nonparametric estimate of F_i, being an average with respect to $p(c \mid x_1, \ldots, x_n)$, will depend on the probabilities of ties.

4. The general situation: Partial exchangeability with an unknown partition

Consider a sequence of random variables $(X_1, X_2, \ldots, X_n, \ldots)$ and suppose that each X_i comes from one of S populations, so that the joint distribution function (d.f.), for $n = 1, 2, \ldots$, is

$$P(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n \mid F_1, \ldots, F_S, J_1 = j_1, \ldots, J_n = j_n) = F_{j_1}(x_1) \cdot F_{j_2}(x_2) \cdots F_{j_n}(x_n), \tag{10}$$

where $(J_1, \ldots, J_n)$ are random variables, $J_i = j$ if X_i comes from the distribution function F_j, $j = 1, 2, \ldots, S$, and $F_1, \ldots, F_S$ is a vector of random distribution functions.

We assume that $(J_1, \ldots, J_n)$ and $F_1, \ldots, F_S$ are independently distributed, $(J_1, \ldots, J_n)$ has probability function $p(j_1, \ldots, j_n)$, and that the prior distribution of $F_1, \ldots, F_S$ is a mixture of products of Dirichlet processes (Cifarelli and Regazzini, 1978). Thus,

$$F_1, \ldots, F_S \mid \theta_1, \ldots, \theta_S \sim \prod_{i=1}^{S} \mathscr{D}(\alpha_i(\cdot; \theta_i)), \tag{11}$$

$$\theta_1, \ldots, \theta_S \sim H, \tag{12}$$

where $\mathscr{D}(\alpha(\cdot; \theta))$ denotes a Dirichlet process of parameter $\alpha(\cdot; \theta)$ and H is the S-dimensional mixing d.f., defined similarly to equation (4). This prior for $(F_1, \ldots, F_S)$ is very general. As a special case, independence among the F's can be assumed: if $(\theta_1, \ldots, \theta_S)$ are independent, $(F_1, \ldots, F_S)$ are independent mixtures of Dirichlet processes. Also, if $H(\theta_1, \ldots, \theta_S)$ is degenerate on a point $(\theta_1^*, \ldots, \theta_S^*)$, then $(F_1, \ldots, F_S)$ are independent Dirichlet processes: $F_i \sim \mathscr{D}(\alpha(\cdot; \theta_i^*))$. The prior can be written as

$$(F_1, \ldots, F_S, J_1, \ldots, J_n) \sim p(j_1, \ldots, j_n) \int \prod_{i=1}^{S} \mathscr{D}(\alpha_i(\cdot; \theta_i)) \, \mathrm{d}H(\theta_1, \ldots, \theta_n) .$$

We consider the case where $\alpha_i(\cdot; \theta_i)$ is absolutely continuous with respect to Lebesgue measure λ, with density $M_i \, f_i(x; \theta_i)$. This is usually referred to as the continuous case since

$$\begin{aligned} &E(F_1(t_1), \ldots, F_S(t_S) \mid \theta_1, \ldots, \theta_S) \\ &\quad = \prod_{i=1}^{S} E(F_i(t_i) \mid \theta_1, \ldots, \theta_S) = \int_{-\infty}^{t_1} f(x; \theta_1) \, \mathrm{d}x \cdots \int_{-\infty}^{t_S} f(x; \theta_S) \, \mathrm{d}x. \end{aligned}$$

The likelihood can be computed as a generalization of (2) and is given by

$$\begin{aligned} &f(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_S, j_1, \ldots, j_n) \\ &\quad = \frac{1}{M_1^{[r_1]}} \prod_{\{i: j_i = 1\}} M_1 f(x'_{1,i}; \theta_1)(n(x'_{1,i}) - 1)! \cdots \frac{1}{M_S^{[r_S]}} \prod_{\{i: j_i = S\}} M_S f(x'_{S,i}; \theta_S)(n(x'_{S,i}) - 1)!, \end{aligned} \tag{13}$$

where $x'_{k,i}$ is the ith distinct values among the x_i's such that $j_i = k$, r_j is the number of $(j_1, \ldots, j_n)$ which are equal to j , $r_1 + \cdots + r_S = n$, and $n(x'_{k,i})$ is the number of times the value $x'_{k,i}$ occurs among the x_i's for which $j_i = k$. (The product over an empty set is defined to be 1).

In many applications, the posterior distribution of the partition $(J_1, \ldots, J_n)$ is of interest. From the Bayesian algorithm, it is given by

$$P(J_1 = j_1, \ldots, J_n = j_n \mid x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n \mid j_1, \ldots, j_n) p(j_1, \ldots, j_n),$$

where

$$f(x_1,\ldots,x_n \mid j_1,\ldots,j_n) = \int f(x_1,\ldots,x_n \mid j_1,\ldots,j_n,\theta_1,\ldots,\theta_S)\,\mathrm{d}H(\theta_1,\ldots,\theta_S).$$

From now on, we will assume that the data $(x_1,\ldots,x_n)$ are all distinct. In this case, the likelihood (13) becomes

$$f(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_S,j_1,\ldots,j_n) = \frac{M_1^{r_1}\cdots M_S^{r_S}}{M_1^{[r_1]}\cdots M_S^{[r_S]}} \prod_{\{i:j_i=1\}} f(x_i;\theta_1)\cdots \prod_{\{i:j_i=S\}} f(x_i;\theta_S) \tag{14}$$

and we have

$$\begin{aligned} &f(x_1,\ldots,x_n \mid j_1,\ldots,j_n) \\ &\quad = \frac{M_1^{r_1}\cdots M_S^{r_S}}{M_1^{[r_1]}\cdots M_S^{[r_S]}} \int \prod_{\{i:j_i=1\}} f(x_i;\theta_1)\cdots \prod_{\{i:j_i=S\}} f(x_i;\theta_S)\,\mathrm{d}H(\theta_1,\ldots,\theta_n) \qquad (15)\\ &\quad = k(r_1,\ldots,r_S;M_1,\ldots,M_S)\,I(j_1,\ldots,j_n;\boldsymbol{x}), \end{aligned}$$

where

$$k(r_1,\ldots,r_S;M_1,\ldots,M_S) = \frac{M_1^{r_1}\cdots M_S^{r_S}}{M_1^{[r_1]}\cdots M_S^{[r_S]}}, \text{ and} \tag{16}$$

$$I(j_1,\ldots,j_n;\boldsymbol{x}) = \int \prod_{\{i:j_i=1\}} f(x_i;\theta_1)\cdots \prod_{\{i:j_i=S\}} f(x_i;\theta_S)\,\mathrm{d}H(\theta_1,\ldots,\theta_n). \tag{17}$$

Thus, the likelihood $f(x_1,\ldots,x_n \mid j_1,\ldots,j_n)$ factorizes into two components: $k(r_1,\ldots,r_S;M_1,\ldots,M_S)$ and $I(j_1,\ldots,j_n;\boldsymbol{x})$. Here, $I(j_1,\ldots,j_n;\boldsymbol{x})$ is the likelihood of $(x_1,\ldots,x_n \mid j_1,\ldots,j_n)$ that one would obtain with a parametric model where $(x_1,\ldots,x_n)$ have joint conditional density:

$$h(x_1,\ldots,x_n \mid j_1,\ldots,j_n,\theta_1,\ldots,\theta_n) = \prod_{\{i:j_i=1\}} f(x_i;\theta_1)\cdots \prod_{\{i:j_i=S\}} f(x_i;\theta_S), \tag{18}$$

which is the "parametric-dual" of our model (10). So, in the "continuous case" and when the data are all distinct, the distinguishing feature of the likelihood (15) with respect to the parametric model (18) is the factor $k(r_1,\ldots,r_S;M_1,\ldots,M_S)$. This factor is due to the discrete nature of the Dirichlet process; indeed, we have

Proposition 1. $k(r_1,\ldots,r_S;M_1,\ldots,M_S) = P(X_1 \neq X_2 \neq \cdots \neq X_n \mid r_1,\ldots,r_S)$.

Proof.

$$\begin{aligned} &P(X_1 \neq X_2 \neq \cdots \neq X_n \mid r_1,\ldots,r_S)\\ &\quad = E(P(X_1 \neq X_2 \neq \cdots \neq X_n \mid r_1,\ldots,r_S,F_1,\ldots,F_S,j_1,\ldots,j_n))\\ &\quad = E(P(\text{all } X_i : j_i = 1 \text{ are distinct } \mid r_1,\ldots,r_S,F_1,\ldots,F_S,j_1,\ldots,j_n)\\ &\qquad \cdots P(\text{all } X_i : j_i = S \text{ are distinct } \mid r_1,\ldots,r_S,F_1,\ldots,F_S,j_1,\ldots,j_n))\\ &\quad = P(\text{all } X_i : j_i = 1 \text{ are distinct } \mid r_1)\cdots P(\text{all } X_i : j_i = S \text{ are distinct } \mid r_S)\\ &\quad = \frac{M_1^{r_1}\cdots M_S^{r_S}}{M_1^{[r_1]}\cdots M_S^{[r_S]}} = k(r_1,\ldots,r_S;M_1,\ldots,M_S), \end{aligned}$$

which completes the proof. □

Table 2
Values of $k(r_1,\ldots,r_S;M_1,\ldots,M_S)$, Eq. (16), for $M_1 = \cdots = M_S = M$

	$M = 1$	$M = 10$	$M = 50$
$n = 10, S = 2$			
$k(5, 5; M)/k(10, 0; M)$	252	5.8	1.6
$n = 50, S = 2$			
$k(25, 25; M)/k(50, 0; M)$	1.3×10^{14}	5.8×10^{8}	5188
$n = 50, S = 10$			
$k(5,\ldots,5;M)/k(50,0,\ldots,0;M)$	4.9×10^{43}	6.0×10^{20}	2.5×10^{7}

The posterior distribution of the partition $(J_1,\ldots,J_n)$ is

$$P(J_1 = j_1,\ldots,J_n = j_n \mid x_1,\ldots,x_n) = \frac{k(r_1,\ldots,r_S;M_1,\ldots,M_S)\, I(j_1,\ldots,j_n;\boldsymbol{x})p(j_1,\ldots,j_n)}{\sum_{j_1,\ldots,j_n} k(r_1,\ldots,r_S;M_1,\ldots,M_S)\, I(j_1,\ldots,j_n;\boldsymbol{x})p(j_1,\ldots,j_n)}.$$

It differs from the parametric result only in the presence of the factors $k(r_1,\ldots,r_S;M_1,\ldots,M_S)$. Now, these are clearly bigger when (if possible) $r_1 = r_2 = \cdots = r_S$ and smaller for unequal sequences, of the kind: $(r_j = n, r_i = 0$ for all $i \neq j)$. As exemplified in Table 2, the values of $k(r_1,\ldots,r_S;M_1,\ldots,M_S)$ can be numerically very different for the parametric and nonparametric models, regardless of the data, with a consequent, possibly serious, bias in the posterior distribution of $(J_1,\ldots,J_n)$.

5. Dirichlet process mixing distribution in hierarchical models

Recently, there has been interest in using the Dirichlet process as a prior at the second stage of Bayesian hierarchical models (Antoniak, 1974; Berry and Christensen, 1978; Padgett and Tsokos, 1979; Ghorai and Susarla, 1982; Deely and Lindley, 1981; Ferguson, 1983; Escobar, 1994; Escobar and West, 1995; Liu, 1996). The hierarchical model is basically as follows: Given $\theta_1,\theta_2,\ldots$, the r.v.'s $X_1,X_2,\ldots$ are independent, with $X_i \sim f(x \mid \theta_i)$, where f is a continuous density; $\theta_1,\theta_2,\ldots$ are a sample from an unknown d.f. G, i.e. given G they are i.i.d. according to G; G is a Dirichlet process $\mathscr{D}(MG_0)$.

In this section we make some remarks on the probability of ties among the θ's in this model. First, note that ties in the θ's reflect the structure of dependence among the data: $(\theta_1 = \cdots = \theta_n)$ shows that the data are exchangeable; $(\theta_1 = \cdots = \theta_{n/2} \neq \theta_{n/2+1} = \cdots = \theta_n)$ shows that the data can be split into two exchangeable groups of equal size, and so on. Therefore, one might think of using this model for the problem of partially exchangeable data with unknown partition, discussed in the previous sections. For example, focusing on the simple change-point problem, the fact that one change point occurred at c can be expressed as the event $\{\theta_1 = \theta_2 = \cdots = \theta_c \neq \theta_{c+1} = \cdots = \theta_n\}$.

Now, the structure of the probabilities of ties among the θ's is automatically fixed once we choose the value of M. Indeed, from (1) it follows that, if G_0 has density g_0 with respect to Lebesgue measure, then the likelihood of $(\theta_1,\ldots,\theta_n)$ is

$$g_n(\theta_1,\ldots,\theta_n) = \frac{M^D}{M^{[n]}} \prod_{i=1}^{D}(n_i - 1)! \prod_{i=1}^{D} g_0(\theta'_i),$$

where D is the number of distinct values among $(\theta_1,\dots,\theta_n)$, $(\theta'_1,\dots,\theta'_D)$ is the vector of distinct values and n_i is the number of times the value θ'_i occurs. The probability of a sequence in the subset $\mathscr{C}(D,n_1,\dots,n_D)$ of the parameter space, having a given ordering, is

$$\frac{M^D}{M^{[n]}}\prod_{i=1}^{D}(n_i-1)!. \tag{19}$$

The number of such sequences is

$$\frac{n!}{n_1!\cdots n_D!}\,\frac{1}{m_1!\cdots m_n!},$$

where m_i = number of $(n_1,\dots,n_D)$ which are equal to i. It follows that the probability of the configuration $\mathscr{C}(D,n_1,\dots,n_D)$ is

$$P(\mathscr{C}(D,n_1,\dots,n_D))=\frac{M^D}{M^{[n]}}\,\frac{n!}{n_1\dots n_D}\,\frac{1}{m_1!\dots m_n!}. \tag{20}$$

The probability (20) corresponds to formula (4) in Antoniak (1974), noting that $n_1\,n_2\cdots n_D=1^{m_1}\,2^{m_2}\cdots n^{m_n}$.

Given the number of distinct values, the prior tends to favor unbalanced partitions over partitions in groups of equal size. For example, in the change-point problem, the prior probability that a change point occurred at c, $c=1,2,\dots,n-1$, is, from (19),

$$P(\theta_1=\cdots=\theta_c\neq\theta_{c+1}=\cdots=\theta_n)=\frac{M^2}{M^{[n]}}\,(c-1)!\,(n-c-1)! \tag{21}$$

and, e.g., we have (for even sample size n)

$$\frac{P(\theta_1=\cdots=\theta_{n/2}\neq\theta_{n/2+1}=\cdots=\theta_n)}{P(\theta_1=\cdots=\theta_{n-1}\neq\theta_n)}=\frac{(n/2-1)!\,(n/2-1)!}{(n-2)!}.$$

Fixing $(D,n_1,\dots,n_D)$ and the order (o) according to which the D distinct values are located, is equivalent to fixing a partition of $\{1,2,\dots,n\}$ into D groups $G_j=G_j(D,n_1,\dots,n_D,\mathrm{o})$, such that the θ_i's with $i\in G_j$ share the same value θ'_j, $j=1,2,\dots,D$. Then, we can let

$$\begin{aligned}L(x_1,\dots,x_n;D,n_1,\dots,n_D,\mathrm{o})&\equiv\int\prod_{i=1}^{n}f(x_i\mid\theta_i)\,\mathrm{d}g_n(\theta_1,\dots,\theta_n)\\&=\frac{M^D}{M^{[n]}}\prod_{j=1}^{D}(n_i-1)!\int\prod_{i\in G_j(D,n_1,\dots,n_D,\mathrm{o})}f(x_i\mid\theta'_j)\,g_0(\theta'_j)\,\mathrm{d}(\theta'_j),\end{aligned} \tag{22}$$

where the first integral is over the set

$$\{(\theta_1,\dots,\theta_n)\in\mathscr{C}(D,n_1,\dots,n_D)\ \text{which are ordered according to (o)}\}.$$

It follows that

$$\begin{aligned}f(x_1,\dots,x_n)&\equiv\int\prod_{i=1}^{n}f(x_i\mid\theta_i)\,\mathrm{d}g_n(\theta_1,\dots,\theta_n)\\&=\sum_{D=1}^{n}\sum_{(n_1,\dots,n_D)}\sum_{(\mathrm{o})}L(x_1,\dots,x_n;D,n_1,\dots,n_D,\mathrm{o}).\end{aligned} \tag{23}$$

The posterior distribution of $(\theta_1,\ldots,\theta_n \mid x_1,\ldots,x_n)$ is

$$g(\theta_1,\ldots,\theta_n \mid x_1,\ldots,x_n) \propto \prod_{i=1}^{n} f(x_i \mid \theta_i)\; g_n(\theta_1,\ldots,\theta_n).$$

Therefore, it depends on the probability of ties, which appears in $g_n(\theta_1,\ldots,\theta_n)$ and in the normalizing constant (23). In particular, the posterior distribution of the number D of distinct values among $(\theta_1,\ldots,\theta_n)$ is

$$\begin{aligned} P(D=d \mid x_1,\ldots,x_n) &= \int_{\{(\theta_1,\ldots,\theta_n):D=d\}} \mathrm{d}g(\theta_1,\ldots,\theta_n \mid x_1,\ldots,x_n) \\ &= \frac{\sum_{(n_1,\ldots,n_d)} \sum_{(o)} L(x_1,\ldots,x_n; d, n_1,\ldots,n_d, o)}{\sum_{D=1}^{n} \sum_{(n_1,\ldots,n_D)} \sum_{(o)} L(x_1,\ldots,x_n; D, n_1,\ldots,n_D, o)} \end{aligned} \tag{24}$$

Expressions (23) and (24) generalize results provided by Liu (1996, Theorems 1 and 2) for the beta-binomial case.

In the change-point problem, we can compute the posterior probability that one change point occurred at c, as

$$\begin{aligned} &P(\theta_1 = \cdots = \theta_c \neq \theta_{c+1} = \cdots = \theta_n \mid x_1,\ldots,x_n) \\ &\propto \left\{ \int \prod_{i=1}^{c} f(x_i \mid \theta'_1)\, g_0(\theta'_1)\, \mathrm{d}\theta'_1 \quad \int \prod_{i=c+1}^{n} f(x_i \mid \theta'_2) g_0(\theta'_2)\, \mathrm{d}\theta'_2 \right\} \frac{M^2}{M^{[n]}} (c-1)!\,(n-c-1)!. \end{aligned}$$

The part in brackets is proportional to the posterior of $(C, \theta'_1, \theta'_2)$ that one would obtain for a parametric model of the kind (7), with θ'_1 and θ'_2 i.i.d. according to g_0 and $p(c)$ uniform. The second factor is the

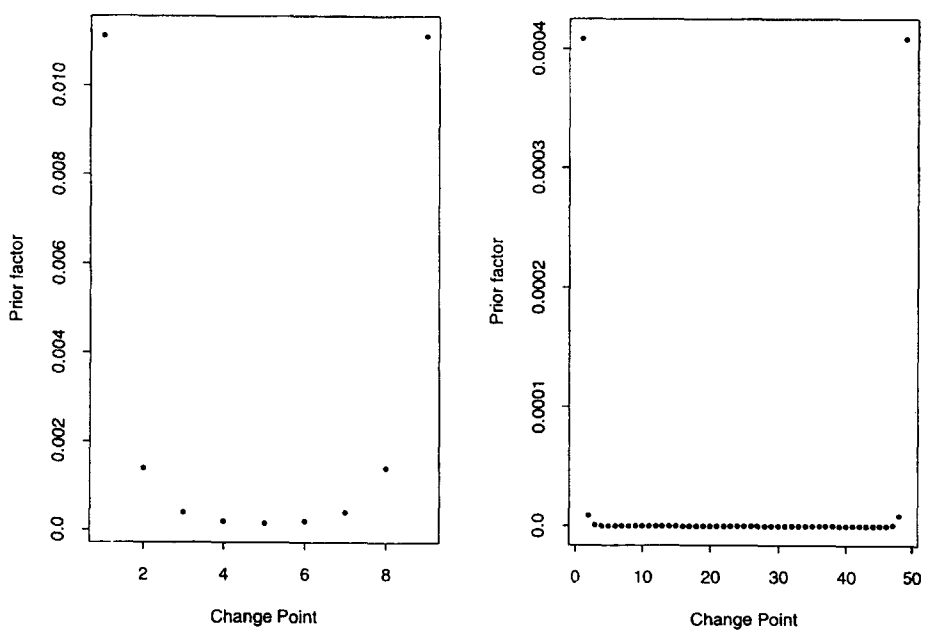

Fig. 5. Plot of the factors introduced by the prior in the hierarchical model for $M_1 = M_2 = 1$; $n = 10$ and $n = 50$.

prior probability (21) of one change point at c. In this sense, the posterior probabilities of the change point have a structure similar to (6), with a factor (here $M^2/M^{[n]}\ (c-1)!\ (n-c-1)!$) which is part of the prior and can have a large effect on inference and Bayes factors.

To see this, consider Fig. 5, which shows examples of these factors contributed by the prior. There are for the same situations as those considered in Fig. 1, with which they can be compared. Clearly, the prior can bias estimation results strongly towards unbalanced partitions.

When using a Dirichlet process for the mixing distribution, the researcher needs to be aware of this aspect. This is also discussed by Escobar and West (1995), who consider, in addition, the possibility of including uncertainty about M in the model; see also Liu (1996).

6. Discussion

We have shown that when the data are continuous and partially exchangeable with an unknown partition, use of a Dirichlet process prior can have a large effect on inferences in an unanticipated way.

This is an example of a case where nonparametric analysis is not robust to the form of the prior. The problem, in our context, seems due to the discrete nature of the Dirichlet process. Presumably, a more robust nonparametric model would be obtained using a prior incorporating the available information that ties in data will not happen, i.e., a prior that selects a continuous distribution. Unfortunately, in the Bayesian literature there are not many proposals of priors which are well suited for nonparametric analysis (i.e., which have "large" support on the class of d.f.'s on the sample space) and which select continuous distributions. In the previous section, we discussed a hierarchical model with $X_i \mid \theta_i \sim f(x \mid \theta_i)$, and with a Dirichlet process as a prior for the parameters at the second stage. However, as we argued, this solution can also not be robust. Other possibilities include Polya trees priors (Mauldin et al., 1992; Lavine, 1992, 1994), which can be specified to select an absolutely continuous distribution, or, for data in a closed bounded interval, Bernstein priors (Petrone, 1995).

Acknowledgements

The authors are grateful to Peter Green for interesting discussions about the problem treated in Section 5 and to an anonymous referee for helpful comments.

References

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist. 2 (6), 1152–1174.

Berry, D.A., Christensen, R., 1979. Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. Ann. Statist. 3, 558–568.

Blackwell, D., 1973. Discreteness of Ferguson selection. Ann. Statist. 1, 356–358.

Blackwell, D., Mac Queen J.B., 1973. Ferguson distributions via Polya scheme. Ann. Statist. 1, 353–355.

Carota, C., Parmigiani, G., 1996. On Bayes factors for nonparametric alternatives. In: Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M., (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, Oxford, pp. 507–511.

Cifarelli, D.M., Regazzini, E., 1978. Problemi statistici nonparametrici in condizioni di scambiabilitaà parziale. Impiego di medie associative Quaderni dell'Istituto di Matematica Finanziaria. dell' Università di Torino Serie III, pp. 1–36.

Consonni, G., Veronese, P., 1995. A Bayesian method for combining results from sevaral binomial experiments. J. Amer. Statist. Assoc. 90, 935–944.

Deely, J.J., Lindley, D.V., 1981. Bayes empirical Bayes. J. Amer. Statist. Assoc. 76, 833–841.

Escobar, M.D., 1994. Estimating normal means with a Dirichlet process prior. J. Amer. Statist. Assoc. 89, 268–277.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc. 90, 577–587.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1 (2) 209–230.

Ferguson, T.S., 1974. Prior distributions on spaces of probability measures. Ann. Statist. 2, (4) 615–629.

Ferguson, T.S., 1983. Bayesian density estimation by mixtures of normal distributions. In: Rizvi H., Rustagi J., (Eds.), Recent Advances in Statistics. Academic Press, New York, 287–302.

Ghorai, J.K., Susarla, V., 1982. Empirical Bayes estimation of probability density function with Dirichlet process prior. In: Grossmann, W., et al. (Eds.), Probability and Statistical Inference. Reidel, Dordrecht, pp. 101–114.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

Korwar, R.M., Hollander, M., 1973. Contribution to the theory of Dirichlet processes. Ann. Probab. 1, 705–711.

Lavine, M., 1992. Some aspects of Polya tree distributions for statistical modelling. Ann. Statist. 20, 1222–1235.

Lavine, M., 1994. More aspects of Polya tree distributions for statistical modelling. Ann. Statist. 22, 1161–1176.

Liu, J.S., 1996. Nonparametric hierarchical Bayes via sequential imputation. Ann. Statist. 24 (3), 911–930.

Mauldin, R.D., Sudderth, W.D., Williams, S.C., 1992. Polya trees and random distributions. Ann. Statist. 20 (3), 1203–1221.

Mira, A., Petrone, S., 1996. Bayesian hierarchical nonparametric inference for changepoint problems. Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M., (Eds.), Bayesian Statistics vol. 5. Oxford University Press, Oxford, 693–703.

Muliere, P., Scarsini, M., 1984. Bayesian inference for change-point problems. Riv. Statist. Appl. 17, 93–102.

Muliere, P., Scarsini, M., 1985. Changepoints problems: a Bayesian nonparametric approach. Appl. Math., 30, 397–402.

Padget, W.J., Tsokos, C. T., 1979. Bayes estimation of reliability using an estimated prior distribution. Oper. Res. 27, 1142–1157.

Petrone, S., 1995. Random Bernstein polynomials. Quaderni di Dipartimento 30 (12-95), Dipartimento di Economia Politica e Metodi Quantitativi, Università di Pavia.

Raftery, A.E., 1994. Change point and change curve modeling in stochastic processes and spatial statistics. J. Appl. Statist. Sci. 1, 403–424.