BAYESIAN STATISTICS 8, pp. 1–27.
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© *Oxford University Press, 2007*

Objective Bayesian Analysis of Multiple Changepoints for Linear Models

F. Javier Girón
Universidad de Málaga, Spain
fj_giron@uma.es

Elías Moreno
Universidad de Granada, Spain
emoreno@ugr.es

George Casella
University of Florida, USA
casella@stat.ufl.edu

Summary

This paper deals with the detection of multiple changepoints for independent but non identically distributed observations, which are assumed to be modeled by a linear regression with normal errors. The problem has a natural formulation as a model selection problem and the main difficulty for computing model posterior probabilities is that neither the reference priors nor any form of empirical Bayes factors based on real training samples can be employed.

We propose an analysis based on the *intrinsic priors*, which do not require real training samples and provide a feasible and sensible solution. For the case of changes in the regression coefficients very simple formulas for the prospective and the retrospective detection of changepoints are found.

On the other hand, when the sample size grows the number of possible changepoints also does and consequently the number of models involved. A stochastic search for finding only those models having large posterior probability is provided. Illustrative examples based on simulated and real data are given.

Keywords and Phrases: Bayes factors; changepoints; intrinsic priors; model selection; posterior model probabilities; stochastic search.

1. INTRODUCTION

There is an extensive literature on the changepoint problem from both the prospective and retrospective viewpoint, for single and multiple changepoints, with parametric and nonparametric sampling models, mainly from a frequentist point of view. For a review see the paper by Lai (1995).

This paper has been supported by MCyT grant SEJ2004–2447 (E. Moreno and F. J. Girón) and NSF grant DMS04–05543 (G. Casella).

The prospective, or on line, changepoint problem consists in the "sequential" detection of a change in the distribution of a set of time-ordered data. In the retrospective changepoint problem the inference of the change in the distribution of the set of time-ordered data is based on the whole data set. These are two related but different problems. Each of them can be formulated as a model selection problem, and while the latter assumes that multiple changes might occur, the former looks for the first time a change is detected.

Under the Bayesian viewpoint, the retrospective analysis have been considered by many authors, for instance Chernoff and Zack (1964), Bacon and Watts (1971), Ferreira (1975), Smith (1975), Choy and Broemeling (1980), Smith and Cook (1980), Menzefrique (1981), Raftery and Ackman (1986), Carlin *et al.* (1992), Stephens (1994), Kiuchi *et al.* (1995), Moreno, Casella and García-Ferrer (2005), among others. Except for the last one, these papers have in common that the prior distribution of the position of a single changepoint is assumed to be uniformly distributed, and for the parameters of the models before and after the change, conjugate distributions are considered. The hyperparameters are determined either subjectively or using empirical Bayes estimators. Sometimes the values of the hyperparameters are chosen to obtain flat priors. An exception is Raftery and Ackman (1986) and Stephens (1994) where objective improper priors were considered. In the former paper the arbitrary constant involved in the Bayes factor for the improper priors was determined by assigning the value one to the Bayes factor at a given sample point, which is subjectively chosen, as in Spiegelhalter and Smith (1982). In the latter paper the value one was assigned to the constant.

An alternative to these conventional methods is to use intrinsic priors (Berger and Pericchi, 1996; Moreno *et al.*, 1998) which are automatically derived from the structure of the model involved, do not depend on any tuning parameters, and have been proved to behave extremely well in a wide variety of problems, in particular for multiple testing problems involving normal linear models (Casella and Moreno, 2006; Girón *et al.*, 2006b; Moreno and Girón, 2006). We will argue in Section 3 that among the existing objective Bayesian procedures the one based on intrinsic priors seems to be the only one that can be employed for the changepoint problem.

This paper generalizes the paper by Moreno, Casella and García-Ferrer (2005) in two directions. First, multiple changepoints are deemed possible. Second, the observations between two consecutive changepoints are independent but not identically distributed. Here, a normal linear model with deterministic explanatory variables is assumed for the sample observations.

The remainder of the paper is organized as follows. Section 2 formulates the prospective and retrospective changepoint problems. For the normal linear regression model, in Section 3 we discuss the difficulties in assessing prior distributions for all the model parameters involved, propose uniform priors for the number of changepoints, and a conditional uniform prior for the position of these changes, and intrinsic priors for the model parameters. Section 4 focuses on several issues related to the objective analysis of the homoscedastic normal linear model, and its relation with maximum likelihood approach to the problem. Section 5 is devoted to computational issues, Section 6 illustrates the findings on real and simulated data, and Section 7 contains some extensions and concluding remarks.

2. FORMULATION OF THE CHANGEPOINT PROBLEM

Let Y be an observable random variable with sampling distribution $f(y|\theta)$, $\theta \in \Theta$, and $\mathbf{y}^t = (y_1, \ldots, y_n)$ a vector of sequential observations. A single changepoint in

the sample means that there is a position r in the sample, $1 \leq r \leq n-1$, such that the first r observations come from $f(x|\theta_1)$ and the rest of the observations from $f(x|\theta_2)$, where $\theta_1 \neq \theta_2$. The extension of this definition to more than a single changepoint, *i.e.*, multiple changepoints, is straightforward.

Later on this sampling distribution will be taken as a multivariate normal distribution whose mean may depend on some deterministic covariates, one of which might be a time variable.

2.1. *Retrospective formulation*

We approach the retrospective multiple changepoints problem as one of model selection in which for a fixed sample size n we admit, in the first instance, that all possible changepoints configurations might occur and want to order the associated models with respect to a given criterion. In section 3 we will discuss the case of ruling out some models *a priori*, for instance, those containing two or more successive changepoints. In our approach, the Bayesian criterion will be dictated by $0-1$ loss functions meaning that the models be compared according to their posterior probabilities.

It is convenient to think of the models involved as a hierarchy: first conditionally on a given number of changepoints, and then to allow the number of changepoints to vary.

Let p, $1 \leq p \leq n-1$, denote the number of changepoints in the sample, $\mathbf{r}_p = (r_1, \ldots, r_p)$ the positions at which the changes occur, and $S_{\mathbf{r}_p} = (\mathbf{y}_1^t, \ldots \mathbf{y}_{p+1}^t)$ the partition of the vector of observations $\mathbf{y}$ such that the order of appearance of the observations in the vector is preserved. Thus, this generic partition is given by

$$S_{\mathbf{r}_p} = (\mathbf{y}_1^t, \ldots \mathbf{y}_{p+1}^t) = \{(y_1, \ldots, y_{r_1}), (y_{r_1+1}, \ldots, y_{r_2}), \ldots, (y_{r_p+1}, \ldots, y_n)\}.$$

The sampling distribution for the partition $S_{\mathbf{r}_p}$ is

$$f(\mathbf{y}|\theta_{p+1}, \mathbf{r}_p, p) = \prod_{i=1}^{r_1} f(y_i|\theta_1) \prod_{i=r_1+1}^{r_2} f(y_i|\theta_2) \times \cdots \times \prod_{i=r_p+1}^{n} f(y_i|\theta_{p+1}),$$

where the number of changepoints p, their position $\mathbf{r}_p = (r_1, \ldots, r_p)$, and the associated parameters $\boldsymbol{\theta}_{p+1} = (\theta_1, \ldots \theta_{p+1})$ are unknown quantities. The integer vector variable $\mathbf{r}_p$ belongs to the set $\mathfrak{R}^p = \{(r_1, \ldots, r_p) : 1 \leq r_1 < r_2 < \ldots < r_p \leq n-1\}$. and the parameter $\boldsymbol{\theta}_{p+1}$ belongs to the set $\boldsymbol{\theta}_{p+1} = \Theta_1 \times \ldots \times \Theta_{p+1}$. The singular case of $p = 0$ corresponds to the case of no change; in this case we set $\mathbf{r}_0 = n$, and the corresponding partition is $S_0 = \{\mathbf{y}\}$. The sampling distribution in this case is given by

$$f(\mathbf{y}|\theta, n, 0) = \prod_{i=1}^{n} f(y_i|\theta_0).$$

In the changepoint problem, we are primarily interested in making inferences on p and $\mathbf{r}_p$ but inferences on $\boldsymbol{\theta}_{p+1}$, or some functions of θ_{p+1}, may also be of interest.

In what follows we suppress the subscripts of the $\boldsymbol{\theta}$'s and $\mathbf{r}$'s when there is no possibility of confusion due to their dependence on p.

All the possible sampling models can be classified into boxes $\{\mathfrak{B}_0, \ldots, \mathfrak{B}_{n-1}\}$, where box $\mathfrak{B}_0$ contains the model with no change, $\mathfrak{B}_1$ contains the models with one changepoint, and so on.

Assuming a prior distribution for the parameters $(\boldsymbol{\theta}, \mathbf{r}, p)$ such that $(\boldsymbol{\theta}, \mathbf{r}, p) \in \boldsymbol{\Theta}^{p+1} \times \mathfrak{R}^p \times \{0, 1, \ldots, n-1\}$, say $\pi(\boldsymbol{\theta}, \mathbf{r}, p)$ generally given by the hierarchical decomposition $\pi(\boldsymbol{\theta}|\mathbf{r})\pi(\mathbf{r}|p)\pi(p)$, we have a general Bayesian model for the changepoint

problem. Note that referring to the parameter p is equivalent to referring to the generic box $\mathfrak{B}_p$ which contains all sampling models with exactly p changepoints, so that the prior probability $\pi(p)$ assigned to the occurrence of p changepoints in the sample, is equivalently the prior probability assigned to the box $\mathfrak{B}_p$.

The main interest in our setting is making inferences on three quantities, first on the number of changepoints p, second on the configuration $\mathbf{r}$ conditionally on p and, in the third place, on the configuration $\mathbf{r}$ on the whole set of models comprising the set of all boxes $\mathfrak{B} = \mathfrak{B}_0 \cup \mathfrak{B}_1 \cup \ldots \cup \mathfrak{B}_{n-1}$. Thus, all we need is to compute $\pi(p|\mathbf{y})$, $\pi(\mathbf{r}|\mathbf{y}, p)$ and $\pi(\mathbf{r}|\mathbf{y})$.

To single out the model with changepoints at vector $\mathbf{r}$ it may be convenient to refer to it as model $M_{\mathbf{r}}$. Note that every changepoint model $M_{\mathbf{r}}$ is contained in one and only one box $\mathfrak{B}_p$; therefore, $\pi(\mathbf{r}|p) = 0$ if $\mathbf{r} \notin \mathfrak{R}^p$.

Let $m(\mathbf{y}|M_{\mathbf{r}})$ denote the marginal of the data $\mathbf{y}$ given model $M_{\mathbf{r}}$, that is

$$m(\mathbf{y}|M_{\mathbf{r}}) = m(\mathbf{y}|\mathbf{r}) = \int f(\mathbf{y}|\theta, \mathbf{r}, p)\pi(\theta|\mathbf{r})\, d\theta,$$

and $m(\mathbf{y}|M_0)$ the marginal under the no change model M_0, that is

$$m(\mathbf{y}|M_0) = \int f(\mathbf{y}|\theta_0)\pi(\theta_0, n, 0)\, d\theta_0.$$

If $B_{\mathbf{r}n} = m(\mathbf{y}|M_{\mathbf{r}})/m(\mathbf{y}|M_0)$ denotes the Bayes factor for comparing model $M_{\mathbf{r}}$ against M_0, straightforward probability calculations render the required posterior probabilities in terms of the Bayes factor —as will prove more convenient in the sequel, instead of the marginal $m(\mathbf{y}|M_{\mathbf{r}})$—, as follows

$$\pi(p|\mathbf{y}) = \frac{\pi(p)\sum_{\mathbf{s}\in\mathfrak{R}^p}\pi(\mathbf{s}|p)B_{\mathbf{s}n}(\mathbf{y})}{\sum_{q=0}^{n-1}\pi(q)\sum_{\mathbf{s}\in\mathfrak{R}^q}\pi(\mathbf{s}|q)B_{\mathbf{s}n}(\mathbf{y})}\ , \text{ for } p \in \{0, 1, \ldots, n-1\}, \tag{1}$$

$$P(M_{\mathbf{r}}|\mathbf{y}, p) = \frac{\pi(\mathbf{r}|p)B_{\mathbf{r}n}(\mathbf{y})}{\sum_{\mathbf{s}\in\mathfrak{R}^q}\pi(\mathbf{s}|q)B_{\mathbf{s}n}(\mathbf{y})}\ , \text{ for } M_{\mathbf{r}} \in \mathfrak{B}_p, \tag{2}$$

$$P(M_{\mathbf{r}}|\mathbf{y}) = \frac{\pi(p)\pi(\mathbf{r}|p)B_{\mathbf{r}n}(\mathbf{y})}{\sum_{q=0}^{n-1}\pi(q)\sum_{\mathbf{s}\in\mathfrak{R}^q}\pi(\mathbf{s}|q)B_{\mathbf{s}n}(\mathbf{y})}\ , \text{ for } M_{\mathbf{r}} \in \mathfrak{B}, \tag{3}$$

where, by convention, the Bayes factor for comparing the no chage model M_0 with itself is $B_{0n} = 1$.

As said above, when using 0–1 loss functions, the optimal decision on the discrete parameters p and $\mathbf{r}$ is to choose the model having the highest posterior probability of their corresponding distributions.

If inferences on the parameters θ or functions of them are required —in this case neccessarilly conditional on p—, they are made from the following posterior distribution

$$\pi(\theta_{p+1}|\mathbf{y}, p) = \sum_{\mathbf{s}\in\mathfrak{R}^p}\pi(\theta_{p+1}|\mathbf{y}, \mathbf{s}, p)\pi(\mathbf{s}|\mathbf{y}, p). \tag{4}$$

2.2. *Prospective formulation*

The prospective formulation consists in detecting the first time n in which we choose the box $\mathfrak{B}_1$ against the box $\mathfrak{B}_0$, thus indicating that an unexpected change or anomaly might occur in a neighborhood of n. In quality control, it is understood that in that case we must stop the experiment and make a decision.

2.2.1. *Stopping rules*

For a fixed sample size n, the Bayesian models in box $\mathfrak{B}_1$ are given by

$$M_{r_1} : \{f(\mathbf{y}|\theta_1, \theta_2, r_1, 1), \pi(\theta_1, \theta_2|r_1)\pi(r_1)\}, \text{ for } r_1 = 1, \ldots, n-1,$$

and the Bayesian model for the no change model in box $\mathfrak{B}_0$ is given by

$$M_0 : \{f(y|\theta_0, n, 0), \pi(\theta_0)\}.$$

From expression (1) for $p = 0$ and $p = 1$, conditional on $p \leq 1$, and assuming that $\pi(p = 0) = \pi(p = 1) = 1/2$, we have

$$\pi(p = 0|\mathbf{y}) = \frac{1}{1 + \sum_{r_1=1}^{n-1} \pi(r_1) B_{r_1 n}(\mathbf{y})}, \quad \pi(p = 1|\mathbf{y}) = \frac{\sum_{r_1}^{n-1} \pi(r_1) B_{r_1 n}(\mathbf{y})}{1 + \sum_{r_1=1}^{n-1} \pi(r_1) B_{r_1 n}(\mathbf{y})},$$

so that box $\mathfrak{B}_1$ is to be chosen if $\pi(p = 1|\mathbf{y}) > \pi(p = 0|\mathbf{y})$ for the corresponding sample size n. Therefore, the Bayesian stopping rule is to stop at time N given by

$$N = \inf \Big\{ n : \sum\nolimits_{r_1=1}^{n-1} \pi(r_1) B_{r_1 n}(\mathbf{y}) > 1 \Big\}. \tag{5}$$

Note that the general stopping rule (5) depends on a simple statistic which is a weighted sum of Bayes factors, where the weights are the prior probabilities of the models in box $\mathfrak{B}_1$. Hence, there remains the problem of eliciting these prior probabilities. If we choose a uniform distribution for r_1, that is $\pi(r_1) = 1/(n-1)$, the stopping rule becomes

$$N^U = \inf \Big\{ n : \frac{1}{n-1} \sum\nolimits_{r_1=1}^{n-1} B_{r_1 n}(\mathbf{y}) > 1 \Big\}.$$

Other choices for $\pi(r_1)$ are deemed possible; for example, for one-step ahead on line detection, if we set $\pi(r_1) = 0$ for $r_1 = 1, \ldots, n-2$ and $\pi(n-1) = 1$, then the stopping rule is

$$N^{osa} = \inf \left\{ n : B_{(n-1)n}(\mathbf{y}) > 1 \right\}.$$

This rule might be of interest for anticipating the first instance whether either an outlying observation or a possible changepoint, though it can not discriminate between both possibilities; therefore the need to develop a more comprehensive strategy to on-line detection, which we now describe.

2.2.2. *Monitoring*

For on line detection of the first true changepoint —one which lasts more than one or a small number of observation, usually larger than the dimension, say k, of the parameters θ_j— a Bayesian monitoring procedure would be required which is not only more informative than a stopping rule but it is capable of discerning on line between transient changepoints, usually outlying observation, and a permanent changepoint.

From our perspective, monitoring can be accomplished by applying the retrospective procedure sequentially to the available data at every time instant n. This produces an array of sequences of either Bayes factors $B_{r_1 n}$ for all possible values of $r_1 = 0, \ldots, n$ or, what is equivalent but much more convenient due to the use

of a common probability scale, the posterior probabilities $P(M_{r_1}|\mathbf{y})$ given by (2) conditional on $p \leq 1$. Therefore, the array we compute is

$$\begin{array}{llll} P(M_0|y_1) & & & \\ P(M_0|y_1,y_2) & P(M_1|y_1,y_2) & & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & \\ P(M_0|y_1,\ldots,y_n) & P(M_1|y_1,\ldots,y_n) & \ldots & P(M_{n-1}|y_1,\ldots,y_n), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & \end{array}$$

This monitoring procedure, when applied to real and simulated data, provides a very satisfactory solution to the problem of detecting the first changepoint because it is able to discriminate between isolated changepoints —single or patches of outlying observations— and the first true permanent changepoint. Note that for starting the monitoring process, the minimum sample size required n has to be larger than the dimension of the parameter space.

3. OBJECTIVE BAYESIAN METHODS

Once a sampling model is established, the main difficulty for both the prospective and retrospective detection of a changepoint is to assess the prior distributions for all the parameters in the models. Ideally, we would like that our subjective degree of belief on the parameters were sufficient to define the prior distributions for the problem. However, in real applications there are many parameters, as in the linear case, and to subjectively elicit the prior distributions is a very hard and demanding task. Thus, we have to rely on automatic or objective methods for deriving priors for the analysis.

3.1. *Objective priors for the discrete parameters*

On the discrete parameters p and $\mathbf{r}$ uniform priors are the common choice. We have, in principle, two ways of assessing a uniform prior on these parameters: first, a uniform prior on the set of all possible model $\mathfrak{B}$, meaning that $\pi(\mathbf{r},p) = 1/2^{n-1}$ for $\mathbf{r} \in \cup_p \mathfrak{N}^p$; and, second, using the hierarchical nature of the prior $\pi(\mathbf{r},p) = \pi(\mathbf{r}|p)\pi(p)$, assigning first a uniform prior on the set $\{0,1,\ldots,n-1\}$, and then a uniform prior on each box $\mathfrak{B}_p$, that is

$$\pi(\mathbf{r},p) = \frac{1}{n}\frac{1}{\binom{n-1}{p}} = \frac{p!(n-p-1)!}{n!}, \text{ if } \mathbf{r} \in \mathfrak{N}^p. \tag{6}$$

This second choice automatically takes into account the fact that boxes $\mathfrak{B}_p$ contains different number of models —note that the number of models in box $\mathfrak{B}_p$ is $\binom{n-1}{p}$ which also depends on the sample size n—.

On the other hand, if a uniform prior on the set of all models were used as in the first choice, the marginal of p is a Binomial distribution with parameters $n-1$ and $1/2$ instead of a discrete uniform. This means that, a priori, models with either a small or a large number of changepoints have a very small probability when compared with models with a number of changepoints of about $n/2$; and, this situation worsens when the sample size n increases. As expected, the use of this prior in simulated and real data produces paradoxical results, while the second prior produces very sensible results. Consequently, the first prior is ruled out from both theoretical and practical reasons.

Using the prior given by (6), formulas (1), (2) and (3), respectively simplify to

$$\pi(p|\mathbf{y}) = \frac{p!(n-1-p)!\sum_{\mathbf{s}\in\mathfrak{N}^p} B_{\mathbf{s}n}(\mathbf{y})}{\sum_{q=0}^{n-1} q!(n-1-q)!\sum_{\mathbf{r}\in\mathfrak{N}^q} B_{\mathbf{r}n}(\mathbf{y})}, \text{ for } p = 0, 1, \ldots, n-1, \tag{7}$$

$$P(M_{\mathbf{r}}|\mathbf{y}, p) = \frac{B_{\mathbf{r}n}(\mathbf{y})}{\sum_{\mathbf{r}\in\mathfrak{N}^p} B_{\mathbf{r}n}(\mathbf{y})}, \text{ for } \mathbf{r} \in \mathfrak{N}^p, \tag{8}$$

$$P(M_{\mathbf{r}}|\mathbf{y}) = \frac{p!(n-1-p)!B_{\mathbf{r}n}(\mathbf{y})}{\sum_{q=0}^{n-1} q!(n-1-q)!\sum_{\mathbf{r}\in\mathfrak{N}^q} B_{\mathbf{r}n}(\mathbf{y})}, \text{ for } \mathbf{r} \in \cup_p \mathfrak{N}^p. \tag{9}$$

So far, all possible configurations $\mathbf{r}$ have been given *a priori* a positive probability. However, in practice, a consecutive pair of changepoints, which implies a partition of the data containing a single observation, should not be accepted as a *true* changepoint for this would correspond to an abrupt change caused by a single outlier between two adjacent observations.

Further, for estimating the k-dimensional parameter θ_j of the corresponding partition $\mathbf{y}_j$ we need at least a sample size larger than or equal to the dimension of θ_j. Hence, reducing the number of configurations by taking into account the preceding restrictions seems realistic in most problems. Therefore, the prior on $\mathbf{r}$ and p should now depend on a certain set of restrictions, say $\mathfrak{R}_k$, on the space of all models $\cup_p \mathfrak{N}^p$. Instead of working with the expression of the prior conditional on $\mathfrak{R}_k$ which, in some circumstances, may be difficult to specify, a much better strategy (see Box and Tiao 1992, pp. 67–69) is to restrict the posterior in the space of all models $P(M_{\mathbf{r}}|\mathbf{y})$ conditioning to the set $(\cup_p \mathfrak{N}^p) \cap \mathfrak{R}_k$, *i.e.*, , to consider

$$P(M_{\mathbf{r}}|\mathbf{y}, \mathfrak{R}_k) \propto p!(n-1-p)!B_{\mathbf{r}n}(\mathbf{y}) \quad \text{for } \mathbf{r} \in (\cup_p \mathfrak{N}^p) \cap \mathfrak{R}_k.$$

Note that, for example, if we only consider those configurations $\mathbf{r}$ that satisfy the restriction $r_j - r_{j-1} > k$ for all j, then it is easy to show, using a simple combinatorial calculation, that for all $p > (n+k-1)/(k+1)$ the sets $\mathfrak{N}^p \cap \mathfrak{R}_k$ are empty, and the remaining ones have fewer models, except boxes $\mathfrak{B}_0$ and $\mathfrak{B}_1$.

3.2. *Objective priors for the continuous parameters*

For the conditional distribution of $\theta|\mathbf{r}$ either conjugate priors or vague priors, usually a limit of conjugate priors with respect to some of the hyperparameters, are typically used. A difficulty is that conjugate prior distributions depend on many hyperparameters which need be assessed, and vague priors hide the fact that they are generally improper and consequently model posterior probabilities are not well-defined. Thus, it seems to us that, for the parameters θ of the conditional distribution $\theta|\mathbf{r}$, objective Bayesian priors might be appropriate here. Unfortunately, the objective reference priors for normal linear models are also improper.

We also note that to replace $B_{\mathbf{r}n}(\mathbf{y})$ with an empirical Bayes factor based on real training samples —for instance some sort of intrinsic Bayes factors—, is ruled out since a changepoint may occur before we have a training sample of minimal size.

To compute the Bayes factor $B_{\mathbf{r}n}(\mathbf{y})$ we propose to use as priors for θ_0 and $\boldsymbol{\theta}|\mathbf{r}$ the intrinsic priors $(\pi^N(\theta_0), \pi^I(\boldsymbol{\theta}|\mathbf{r}))$ derived from the improper reference priors $\pi^N(\theta_0)$ and $\pi^N(\boldsymbol{\theta}|\mathbf{r})$. Intrinsic priors do not use real training samples but theoretical ones and hence the difficulty due to the absence of real training samples disappear.

Further, they are completely automatic and hence there is no need to adjust any hyperparameter. Moreover, the no changepoint model M_0 is nested into any other $M_{\mathbf{r}}$, so that intrinsic priors for comparing a model with changepoints at position $\mathbf{r}$ with the no changepoint model do exist. The formal expression of the intrinsic prior for the parameters of the distribution $\boldsymbol{\theta}|\mathbf{r}$ conditional on an arbitrary but fixed point of the parameter of the no changepoint model θ_0, say $\pi^I(\boldsymbol{\theta}|\theta_0,\mathbf{r})$, is given by

$$\pi^I(\boldsymbol{\theta}|\theta_0,\mathbf{r}) = \pi^N(\theta|\mathbf{r})\; E_{\mathbf{Y}(\ell)|\theta,\mathbf{r}} \frac{f(\mathbf{Y}(\ell)|\theta_0,n)}{\int f(\mathbf{Y}(\ell)|\theta,\mathbf{r})\pi^N(\boldsymbol{\theta}|\mathbf{r})d\boldsymbol{\theta}} \;,$$

where the expectation is taken with respect to $f(\mathbf{Y}(\ell)|\theta,\mathbf{r})$, ℓ being the minimal training sample size such $0 < \int f(\mathbf{Y}(\ell)|\theta,\mathbf{r})\pi^N(\boldsymbol{\theta}|\mathbf{r})d\,\theta < \infty$, (Berger and Pericchi 1996).

This conditional intrinsic prior is a probability density, and the unconditional intrinsic prior for $\boldsymbol{\theta}|\mathbf{r}$ is given by

$$\pi^I(\boldsymbol{\theta}|\mathbf{r}) = \int \pi^I(\boldsymbol{\theta}|\theta_0,\mathbf{r})\pi^N(\theta_0)\,d\theta_0,$$

which is an improper prior if the mixing distribution $\pi^N(\theta_0)$ is also improper. However, the Bayes factor for intrinsic priors is a well-defined Bayes factor (Moreno *et al.*, 1998).

4. THE HOMOSCEDASTIC NORMAL LINEAR MODEL

Suppose that $(y_1,\ldots,y_n)$ follows the normal linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{\epsilon}|\mathbf{0},\tau^2\mathbf{I}_n),$$

where $\mathbf{X}$ is a $n\times k$ design matrix of full rank, $\boldsymbol{\beta}$ a $k\times 1$ vector of regression coefficients, and τ^2 is the common variance of the error terms. This model corresponds to the situation of no changepoint in the sample.

We assume that the variance error does not change across the sample so that the changes only affect to the regression coefficients.

4.1. *The case of one changepoint*

For clarity of exposition we consider first the case where there is only one changepoint at some unknown position r_1. Let $S_1 = (\mathbf{y}_1^t, \mathbf{y}_2^t)$ be a partition of $\mathbf{y}$ where the dimension of $\mathbf{y}_1$ is $n_1 = r_1$, the dimension of $\mathbf{y}_2$ is $n_2 = n - n_1$. We also split the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix},$$

where $\mathbf{X}_1$ has dimensions $n_1 \times k$ and $\mathbf{X}_2$ has $n_2 \times k$. In the notation of the preceding sections we now have

$$f(y|\theta_0,n,0) = N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta},\tau^2\mathbf{I}_n),$$

and

$$f(\mathbf{y}|\theta,r_1,1) = N_{n_1}(\mathbf{y}_1|\mathbf{X}_1\boldsymbol{\beta}_1,\sigma_1^2\mathbf{I}_{n_1})N_{n_2}(\mathbf{y}_2|\mathbf{X}_2\boldsymbol{\beta}_2,\sigma_1^2\mathbf{I}_{n_2}).$$

The objective intrinsic Bayesian model for the no changepoint is

$$M_0 : \{N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta},\tau^2\mathbf{I}_n), \pi^N(\boldsymbol{\beta},\tau) = \frac{c}{\tau}\},$$

and, conditional on one changepoint at position r_1, the objective intrinsic Bayesian model is

$$M_{r_1} : \{N_{n_1}(\mathbf{y}_1|\mathbf{X}_1\boldsymbol{\beta}_1, \sigma_1^2\mathbf{I}_{n_1})N_{n_2}(\mathbf{y}_2|\mathbf{X}_2\boldsymbol{\beta}_2, \sigma_1^2\mathbf{I}_{n_2}), \pi^I(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)\},$$

where π^N represents the improper reference prior for $\boldsymbol{\beta}, \tau$ (Berger and Bernardo 1992), and $\pi^I(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1)$ the intrinsic prior for the parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1$. Next theorem, stated without proof, provides the form of the intrinsic priors.

Theorem 1 *Conditional on a fix but arbitrary point $\boldsymbol{\beta}, \tau$, the conditional intrinsic prior distribution $\pi^I(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1|\boldsymbol{\beta}, \tau)$ can be shown to be*

$$\begin{aligned}\pi^I(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1|\boldsymbol{\theta}, \tau) &= \frac{2}{\pi\tau(1+\sigma_1^2/\tau^2)} \\ &\times N_k(\boldsymbol{\beta}_1|\boldsymbol{\beta}, (\tau^2+\sigma_1^2)\mathbf{W}_1^{-1}) \times N_k(\boldsymbol{\beta}_2|\boldsymbol{\beta}, (\tau^2+\sigma_1^2)\mathbf{W}_2^{-1}),\end{aligned}$$

where

$$\mathbf{W}_i^{-1} = \frac{n}{2k+1}(\mathbf{X}_i^t\mathbf{X}_i)^{-1},\ i = 1, 2.$$

Using this theorem, we get the following expression for the Bayes factor.

Theorem 2 *The Bayes factor for the model with a changepoint at position r_1 against the no changepoint model is*

$$B_{r_1 n}(\mathbf{y}) = \frac{2}{\pi}(2k+1)^{k/2}\int_0^{\pi/2} \frac{\sin^k\varphi\ (n+(2k+1)\sin^2\varphi)^{(n-2k)/2}}{(n\mathcal{B}_{r_1}+(2k+1)\sin^2\varphi)^{(n-k)/2}}d\varphi$$

where, if denote by $RSS_1 = \mathbf{y}_1^t(\mathbf{I}-\mathbf{H}_1)\mathbf{y}_1$, $RSS_2 = \mathbf{y}_2^t(\mathbf{I}-\mathbf{H}_2)\mathbf{y}_2$ and $RSS_0 = \mathbf{y}^t(\mathbf{I}-\mathbf{H})\mathbf{y}$ the residual sum of squares of the linear submodels induced by the partition S_1 and the no change model, then the statistic $\mathcal{B}_{r_1}$ is

$$\mathcal{B}_{r_1} = \frac{RSS_1 + RSS_2}{RSS_0}.$$

Proof. Denoting the marginal of the data under the models M_0 and M_{r_1} by $m(\mathbf{y}|M_0)$ and $m(\mathbf{y}|M_{r_1})$, respectively, it can be shown that

$$m(\mathbf{y}|M_{r_1}) = \frac{\Gamma(\frac{n-k}{2})}{\pi^{(n-k+2)/2}}\int_0^{\pi/2}\frac{d\varphi}{|\mathbf{D}_0(\varphi)|^{1/2}D_1(\varphi)D_2(\varphi)[H_1(\varphi)-H_2(\varphi))]^{(n-k)/2}},$$

where

$$\mathbf{D}_0(\varphi) = \sum_{i=1}^{2}\frac{2k+1}{n+(2k+1)\sin^2(\varphi)}\mathbf{X}_i^t\mathbf{X}_i$$

$$D_i(\varphi) = \sin^{n_i}(\varphi)\left(1+\frac{n}{(2k+1)\sin^2(\varphi)}\right)^{k/2},\ i = 1, 2,$$

$$H_1(\varphi) = \sum_{i=1}^{2}\frac{1}{\sin^2\varphi}\left(\mathbf{y}_i^t\mathbf{y}_i - \frac{n}{n+(2k+1)\sin^2\varphi}\mathbf{y}_i^t\mathbf{X}_i(\mathbf{X}_i^t\mathbf{X}_i)^{-1}\mathbf{X}_i^t\mathbf{y}_i\right),$$

$$H_2(\varphi) = (2k+1)\left(\sum_{i=1}^{2}\frac{\mathbf{y}_i^t\mathbf{X}_i}{n+(2k+1)\sin^2\varphi}\right)\left(\sum_{i=1}^{2}\frac{\mathbf{X}_i^t\mathbf{X}_i}{n+(2k+1)\sin^2\varphi}\right)^{-1}\left(\sum_{i=1}^{2}\frac{\mathbf{X}_i^t\mathbf{y}_i}{n+(2k+1)\sin^2\varphi}\right)$$

and

$$m(\mathbf{y}|M_0) = \frac{\Gamma(\frac{n-k}{2})}{\pi^{(n-k)/2}(\mathbf{y}^t(\mathbf{I_n}-\mathbf{H})\mathbf{y})^{(n-k)/2}|\mathbf{X}^t\mathbf{X}|^{1/2}},$$

with $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ the hat matrix of the no change model. After some cumbersome algebraic manipulations we finally get the simplified expression of the Bayes factor. □

From this expression we note that the Bayes factor depends on the data through the sum of square of the residuals associated to the partition of the vector of observations at the position of the changepoint. Furthermore, the partition S_1 for which the sum $RSS_1 + RSS_2$ is minimum is the partition with the highest Bayes factor. Therefore, inside the box $\mathfrak{B}_1$ the ordering provided by ranking the models according to their values of $RSS_1 + RSS_2$, and according to their model posterior probabilities

$$P(M_{r_1}|\mathbf{y}, p=1) = \frac{B_{r_1 n}(\mathbf{y})}{\sum_{r_1=1}^{n-1} B_{r_1 n}(\mathbf{y})},\ r_1 = 1,\ldots,n-1,$$

is the same.

However, in the class $\mathfrak{B} = \mathfrak{B}_0 \cup \mathfrak{B}_1$ the ordering of the models is given by the values of their models posterior probabilities conditional on $p \leq 1$, that is

$$P(M_{r_1}|\mathbf{y}, p\leq 1) = \frac{B_{r_1 n}(\mathbf{y})}{n-1+\sum_{r_1=1}^{n-1} B_{r_1 n}(\mathbf{y})},$$

and

$$P(M_0|\mathbf{y}, p\leq 1) = \frac{n-1}{n-1+\sum_{r_1=1}^{n-1} B_{r_1 n}(\mathbf{y})}.$$

From these formulas, it is clear that the new ordering of the models in the box $\mathfrak{B}_1$ is the same as before.

4.2. *The case of multiple changepoints*

For the analysis of the general case when there are p changepoints located at positions $\mathbf{r} = (r_1,\ldots,r_p)$, let the corresponding partition of the data be $S_{\mathbf{r}_p} = (\mathbf{y}_1^t,\ldots\mathbf{y}_{p+1}^t)$ and for $i = 1,\ldots,p+1$, where $r_0 = 0$ and $r_{p+1} = n$, let the dimension of each $\mathbf{y}_i$ be $n_i = r_i - r_{i-1}$. Extending the analysis of the previous subsection, it is easy to see that the Bayes factor for the corresponding intrinsic priors for comparing the model with p changepoints at $\mathbf{r}$ and the model with no changepoint, $B_{\mathbf{r}n}(\mathbf{y})$, turns out to be

$$B_{\mathbf{r}n}(\mathbf{y}) = \frac{2}{\pi}((p+1)k+1)^{pk/2}\int_0^{\pi/2}\frac{\sin^{pk}\varphi\ (n+((p+1)k+1)\sin^2\varphi)^{(n-(p+1)k)/2}}{(n\mathcal{B}_{\mathbf{r}}+((p+1)k+1)\sin^2\varphi)^{(n-k)/2}}\,d\varphi \tag{10}$$

where

$$\mathcal{B}_{\mathbf{r}} = \frac{T_{\mathbf{r}}}{RSS_0} = \frac{RSS_1 + \cdots + RSS_{p+1}}{RSS_0} \ ,$$

RSS_i is the residual sum of squares associated to the linear submodel corresponding to the data $\mathbf{y}_i$, i.e, $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$, and $T_{\mathbf{r}}$ is the total of the residual sum of squares.

We observe again that the Bayes factor (10), for fixed values of p, is a decreasing function of the total of the residual sum of squares of the partition at the positions of the changepoints. This result, together with expression (8), imply that, conditional on the occurrence of p changepoints, *i.e.*, within box $\mathfrak{B}_p$, the model $M_{\mathbf{r}}$ with the highest posterior probability $P(M_{\mathbf{r}}|\mathbf{y}, p)$ is the one that minimizes the total

$$T_{\mathbf{r}} = RSS_1 + \cdots + RSS_{p+1},$$

and both criteria render the same ordering.

On the other hand, the ordering of the models in the whole class of models $\cup_{p=0}^{n-1}\mathfrak{B}_p$ is obtained from the values of

$$P(M_{\mathbf{r}}|\mathbf{y}) = \frac{p!(n-p-1)!B_{\mathbf{r}n}(\mathbf{y})}{\sum_{q=0}^{n-1} q!(n-q-1)!\sum_{\mathbf{s}\in\mathfrak{N}^q} B_{\mathbf{s}n}(\mathbf{y})}, \text{ if } \mathbf{r} \in \cup_p \mathfrak{N}^p, \tag{11}$$

or, equivalently, from the values of $p!(n-p-1)!B_{\mathbf{r}n}(\mathbf{y})$.

Note that in the class of all models $\mathfrak{B}$ the ordering given by equation (11) restricted to any of the boxes $\mathfrak{B}_p$ is the same as the one given by minimizing $T_{\mathbf{r}}$.

One nice property of the Bayes factor for the intrinsic priors is the following. For $p \geq n/k - 1$, the minimum of $T_{\mathbf{r}_p}$ is clearly 0. It can then be shown that the Bayes factor for the intrinsic priors is a decreasing function of p, for fixed n and k, such that its value at $p = n/k - 1$ is equal to 1. Thus, for any integer p such that $p \geq n/k - 1$ the Bayes factor of any model in box $\mathfrak{B}_p$ is smaller than 1, which is the default Bayes factor of the no change model.

This property automatically penalizes models with too many changepoints p; namely, no model in any box $\mathfrak{B}_p$ such that $p \geq n/k - 1$ will ever be preferred to the no change model, and the posterior probability of p changepoints will always be smaller than that of the no change model.

On the other hand, this restriction on p is often included in the formulation of the multiple changepoint problem to avoid the problem of estimating the regression coefficients when the number of data is smaller than the number of regressor variables (see Subsection 4.4).

4.3. *Relationship with the likelihood approach*

It is straightforward to see that, conditional on p, the maximum of the profile likelihood function $L(\mathbf{r})$ is attained at the configuration $\mathbf{r}$ which minimizes the value of the statistic $T_{\mathbf{r}} = RSS_1 + \ldots + RSS_{p+1}$. Consequently, inside the box $\mathfrak{B}_p$ the optimal Bayesian model for the intrinsic priors and the one corresponding to the profile MLE of $\mathbf{r}$ are the same and further, from expression (10), the ordering of the models produced by the values of the profile likelihood is the same as the one given by the posterior probabilities.

The profile likelihood function, whose maximum as a function of $\mathbf{r}$ increases as p increases, goes to infinity when $T_{\mathbf{r}_p} = 0$. This happens for all $p \geq n/k - 1$, thus making impossible the task of estimating the number of changepoints or even

of comparing different configurations of multiple changepoints unless some penalty function is included in the profile likelihood function or some additional criterion is used for comparing among the different boxes.

When none of the contemplated models is the true model —a realistic assumption when analysing real data— and the sample size is large, the values of the profile likelihood function for models in a given box, say $\mathfrak{B}_p$ rank the models as follows: the largest likelihood corresponds to the model in $\mathfrak{B}_p$ closest to the true model, the second largest corresponds to the second closest, and so on, where *close* is understood here in terms of the Kullback-Leibler pseudo-distance. Intrinsic model posterior probabilities also share this nice property. But, as soon as we have to rank models from different boxes, the likelihood approach fails, as also does the AIC correction; however, model posterior probabilities derived from the intrinsic Bayes procedure —which also includes the objective prior distribution for $\mathbf{r}$ and p—automatically take care of the differences in size among the boxes as seen from expressions (10) and (11).

Once that we have established the relationship between the Bayesian and the likelihood approaches to the changepoint problem, the following natural question arises: Is the use of the intrinsic model posterior probabilities a really objective Bayesian procedure for changepoint problems?

The answer —we believe— is yes. Our argument runs as follows: Conditional on the number of changepoints p the total of the residual sum of squares $T_{\mathbf{r}_p} = \sum_{i=1}^{p+1} RSS_{r_i}$ is the minimal sufficient statistic for estimating the changepoint configuration $\mathbf{r}$. Therefore, the vector statistic $(T_{\mathbf{r}_0}, \ldots, T_{\mathbf{r}_{n-1}})$ is the minimal sufficient statistic for making inferences on the set of all sampling models $\mathfrak{B}$. But, as the Bayesian procedure based on intrinsic model posterior probabilities depends on the Bayes factor for the intrinsic priors given by expression (10), and this, in turn, depends on $(T_{\mathbf{r}_0}, \ldots, T_{\mathbf{r}_{n-1}})$, and, further, does not either depend on any hyperparameters nor statistic but the ancillary n and k, we conclude that it is an objective procedure.

4.4. *Estimating the magnitude of the changes*

Sometimes, conditional on the ocurrence of p changepoints, the interest might be on estimating the variations produced by the changepoints in the parameters $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_{p+1}$, which can inform us on the magnitude of these changes. The use of the intrinsic priors for estimating the parameters does not provide simple analytical expressions and even the numerical computation of the estimates are cumbersome. Instead, and recalling that the intrinsic priors are also improper priors based on the reference priors, the use of these priors for estimation produces a simple and sensible posterior for the whole parameter set $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_{p+1}$.

Conditional on the occurrence of p changepoints in the sample, the reference prior for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p+1}, \sigma_p$, where we want to remind that σ_p also depends on p, is

$$\pi^N(\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_{p+1}, \sigma_p) = c/\sigma_p$$

Multiplying this prior by the likelihood function

$$f_n(\mathbf{y}|\boldsymbol{\beta}, \sigma_p, \mathbf{s}) = \prod_{i=1}^{p+1} N_{n_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_i, \sigma_p^2\mathbf{I}_{n_i})$$

conditional on p changepoints at position vector $\mathbf{r} = (r_1, r_2 \ldots, r_p)$ and normalizing, after some algebra, it renders the posterior of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p+1}, \sigma_p$ conditional on $\mathbf{y}$, $\mathbf{r}$ and p, which is a multivariate normal-sqrt-inverted-gamma distribution.

Integrating out σ_p in this posterior, the resulting posterior distribution of the $p+1$ regression parameters, conditional on the data $\mathbf{y}$, and the occurrence of p changepoints at position vector $\mathbf{r}$, is the following multivariate Student t distribution

$$\begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_{p+1} \end{pmatrix} \sim t_{(p+1)k} \left[\begin{pmatrix} \tilde{\boldsymbol{\beta}}_1 \\ \vdots \\ \tilde{\boldsymbol{\beta}}_{p+1} \end{pmatrix}, s^2 \begin{pmatrix} (\mathbf{X_1}^t\mathbf{X_1})^{-1} & \dots & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & \dots & (\mathbf{X_{p+1}}^t\mathbf{X_{p+1}})^{-1} \end{pmatrix}; \nu \right] \tag{12}$$

where $\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_{p+1}$ denote the least squares estimates of the corresponding regression coefficients, conditional on the partition induced by $\mathbf{r}$, $s^2 = T_{\mathbf{r}_p}/\nu$ is the usual estimate of the variance σ_p^2, and $\nu = n - (p+1)k$ are the degrees of freedom. Incidentally, note that for this posterior distribution to be proper and neither singular nor degenerate, the conditions $\nu > 0$, $s^2 > 0$ and $|\mathbf{X_i}^t\mathbf{X_i}| > 0$ must hold, and for this it is neccessary that the number p of changepoints be strictly smaller than $n/k - 1$ and that the partitions $\mathbf{y}_i$ of $\mathbf{y}$ should have a minimum size $n_i \geq k$.

This is the —small— price to be paid in order to estimate the regression parameters using the reference prior instead of the intrinsic one. Note that these constraints imply a reduction in the number of boxes $\mathfrak{B}_p$ and in the models within each box, except when $k = 1$.

Finally, the conditional posterior of the regression parameters on p is the following mixture of multivariate Student t distributions

$$\pi(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{p+1}|\mathbf{y}, p) \sim \sum\nolimits_{\mathbf{s}} \pi(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{p+1}|\mathbf{y}, \mathbf{s}, p) P(M_{\mathbf{s}}|\mathbf{y}, p), \tag{13}$$

where $\pi(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{p+1}|\mathbf{y}, \mathbf{s}, p)$ are given by equation (12) for configuration $\mathbf{s}$, the weights of the mixture are $P(M_{\mathbf{r}}|\mathbf{y},p) = B_{\mathbf{r}n}(\mathbf{y}) / \sum_{\mathbf{s}} B_{\mathbf{s}n}(\mathbf{y})$, and the number of mixture terms and the sum are restricted to those configurations $\mathbf{s}$ satisfying the constraints $n_i > 0$ for $i = 1, \dots, p+1$.

The most useful parameters of interest in changepoint problems are the successive differences $\delta_i = \boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_i$ for $i = 1, \dots, p$, the corresponding distributions of which can be easily obtained from equations (12) and (13), using well known properties of the multivariate Student t distribution. In fact,

$$\delta_i|\mathbf{y}, \mathbf{r}, p \sim t_k(\tilde{\boldsymbol{\beta}}_{i+1} - \tilde{\boldsymbol{\beta}}_i, s^2((\mathbf{X_i}^t\mathbf{X_i})^{-1} + (\mathbf{X_{i+1}}^t\mathbf{X_{i+1}})^{-1}); \nu)$$

and

$$\delta_i|\mathbf{y}, p \sim \sum\nolimits_{\mathbf{s}} t_k(\tilde{\boldsymbol{\beta}}_{i+1} - \tilde{\boldsymbol{\beta}}_i, s^2((\mathbf{X_i}^t\mathbf{X_i})^{-1} + (\mathbf{X_{i+1}}^t\mathbf{X_{i+1}})^{-1}); \nu) P(M_{\mathbf{s}}|\mathbf{y}, p).$$

Thus, as a conclusion, for parameter estimation using the reference priors, the simplicity of working with well known distributions, from which it is also easy to sample if Montecarlo estimates are needed for more complex functions of the parameters, compensates the very small numerical differences between the intrinsic and the reference Bayesian approaches. However, when model comparison is involved, as in making inferences on $\mathbf{r}$ or p, reference priors on the papameters can not be used: they simply do not work. Intrisic priors for changepoint problems involving normal linear models, on the other hand, have nice theoretical properties, behave as true objective priors, and, as we will see, they work very well in both simulated and real data sets.

5. COMPUTATIONAL ISSUES

By far, the main difficulty in the analysis of changepoint problems is the large number of possible models, which is 2^{n-1}. Computing Bayes factors for the intrinsic priors for all models is thus unfeasible. Unless the number of changepoints p be known and small, the computation of Bayes factors for all models can be very time consuming. Therefore, we need to devise estrategies to detect the most probable models in each box $\mathfrak{B}_p$ and also estimating the total or the mean of the Bayes factors within each box.

5.1. *Forward Search*

A simple procedure to tackle the first problem, which can be named or described as sequential forward search, is to visit sequentialy all boxes starting from box $\mathfrak{B}_0$, which contains a single model —the no change model, whose Bayes factor is always equal to 1— then box $\mathfrak{B}_1$ which contains $n-1$ models and selecting the changepoint with highest Bayes factor and so on, retaining at each step the preceding model and adding the new changepoint with highest Bayes factor within the correspoding box. In this way, and in the case of examining all boxes, we only have to compute $1+(n-1)+(n-2)+\ldots+1 = n(n-1)/2+1$ Bayes factors at most.

The proposed forward changepoint search is a fast algorithm for finding a relatively good changepoint configuration within each box, as will be shown in the examples, without the need to resort to an exhaustive all models search, which turns out to be unfeasible even for small sample sizes. On the other hand, it provides no answer to the problem of estimating the number p of changepoints. In addition, this algorithm may have the same similar drawbacks as the classical stepwise algorithms (forward selection and backward elimination) used for variable selection in regression, in its Bayesian counterpart, see Girón, Moreno and Martínez (2006a), as it is just a conditional sequential search estrategy. Notwithstanding these weaknesses, this algorithm usually locates within each box models with high Bayes factors.

By adapting the Gibbs sampling procedure proposed by Stephens (1994) in subsection 3.1, pp. 166–167, sampling from the discrete distribution of $M_{\mathbf{r}}|\mathbf{y}, p$ seems very easy as the model posterior probabilities are proportional to the Bayes factor for intrinsic priors $B_{\mathbf{r}n}$. On the other hand, sampling from the posterior $M_{\mathbf{r}}|\mathbf{y}$ seems a much more demanding task.

All this prompted us to devise an efficient stochastic search algorithm for the multiple changepoint problem.

5.2. *Retrospective search*

For the retrospective search, we find that a random walk Metropolis-Hastings algorithm works very well. We choose a symmetric random walk, and use the posterior probability (11) as the objective function. This insures that, at convergence, the resulting Markov chain is a sample from the posterior probability surface. Hence, states of high posterior probability will be visited more often.

To now choose the "best" model, or to examine a range of good models, we would like to rank the models by their posterior probabilities, but, as mentioned above, this is not possible, as the number of models can be prohibitively large. Moreover, it is also the case that calculation of the denominator in (11) is prohibitive. The solution is to construct an MCMC algorithm with (11) as the stationary distribution. Such an algorithm, if properly constructed, would not only visit every model, but would

visit the better models more often. Thus, a frequency count of visits to the models is directly proportional to the posterior probabilities.

For the regression model, we keep track of changepoints with a $n \times 1$ vector

$$\mathbf{c} = (0, 0, 1, 0, 1, 0, \ldots, 0, 1, 0, 0, 0)^t$$

where "1"indicates a changepoint. At each step of the stochastic search we select an observation at random (actually an index $1, 2, \ldots, n$). If it is a 1 we evaluate whether to change it to a 0, and if it is a 0 we evaluate whether to change it to a 1. This is done with a Metropolis step as follows:

Corresponding to a vector $\mathbf{c}$ there is a vector $\mathbf{r}$ of changepoints and a model $M_{\mathbf{r}}$.

(i) Generate a new $\mathbf{c}'$ and $\mathbf{r}'$, and $U \sim \text{Uniform}(0, 1)$.

(ii) Calculate

$$\rho = \min\left\{1, \frac{P(M_{\mathbf{r}'}|\mathbf{y})}{P(M_{\mathbf{r}}|\mathbf{y})}\right\}.$$

(iii) Move to $\mathbf{c}'$ if $U < \rho$, otherwise remain at $\mathbf{c}$.

(iv) Return to 1.

6. ILLUSTRATIVE EXAMPLES: REAL AND SIMULATED DATA

The exact procedure described above, and the Metropolis algorithm of Section 5.2 were tested and compared on a number of examples, both real and simulated.

6.1. *Simulated data*

Example 1. We first tested the search algorithm on the data given in Figure 1, which were simulated with the following model

$$y_i = \begin{cases} \frac{1}{4}x + \varepsilon & \text{if } x = 1, \ldots, 6; \\ 4 + \varepsilon & \text{if } x = 7, \ldots, 12; \\ 12 - \frac{1}{2}x + \varepsilon & \text{if } x = 13, \ldots, 18, \end{cases} \tag{14}$$

where $\varepsilon \sim N(0, 1)$. We ran the simulations for $\sigma = 0.25, 0.5, 0.75$, and typical data are shown in Figure 1.

Note that it is very difficult to see the three different models for $\sigma = 0.75$. In fact, it is sometimes the case that, for large σ, the true model does not have the highest intrinsic posterior probability. In such cases, there are many competing models that are candidates for the "best", as we will now describe.

For each of $\sigma = 0.25, 0.5, 0.75$, the algorithm was run on 100 datasets, with 20,000 iterations of the Metropolis algorithm. The performance is summarized in Table 1, where we measured the number of times that the true model was in the top 5, top 10, or top 25, ranked on posterior probabilities.

Comparing the performance to the typical data sets, we see that the procedure always finds the true model when the error is reasonable, and does worse as the error term increases. But it is quite surprising that for $\sigma = 0.5$, where our eye cannot see the changepoint at 6, the true model is in the top 5 34% of the time, and in the top 10 almost 60% of the time.

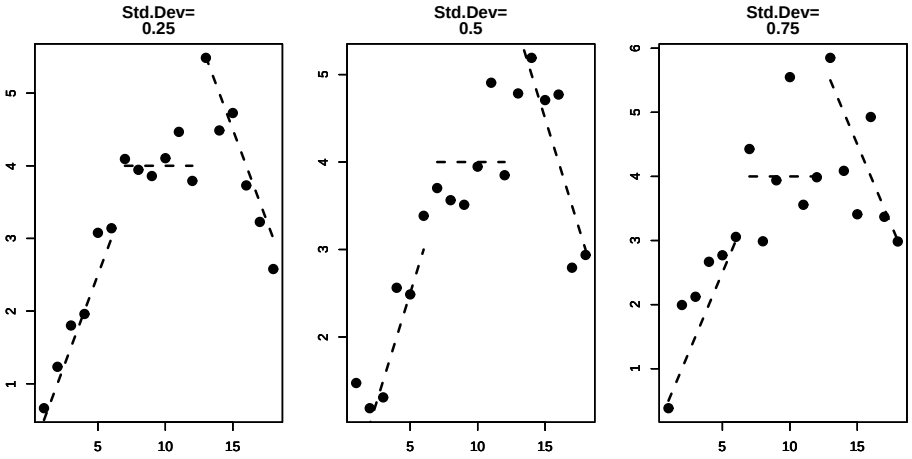

Figure 1: *Typical datasets and true model for data from Model (14). Note that the changepoint at $x = 6$ is barely discernible for $\sigma \geq 0.5$.*

Table 1: *Results of the simulations of model (14). Proportion of times that the true model is in the corresponding top category. The models are ranked by their intrinsic posterior probabilities.*

σ	Top 5	Top 10	Top 25
0.25	94	99	99
0.50	34	59	88
0.75	14	28	69

Example 2. Quandt's Data. A second simulated data set that we look at is from Quandt (1958), which consists of two simulated linear regressions with a slight change at time 12. That data are given in Figure 2, and it is clear by looking at this Figure that the changepoint is barely discernible.

The simulated data come from the two regression models

$$y_i = \begin{cases} 2.5 + 0.7x_i + \varepsilon_i & i = 1, 2, \ldots, 12; \\ 5 + 0.5x_i + \varepsilon_i & i = 13 \ldots, 20. \end{cases}$$

The results from the exact analysis conditional on there being up to five changepoints, *i.e.*, , $p \leq 5$ can be summarized as follows: As seen from Table 2, the posterior mode of $\pi(p|\mathbf{y})$ is $\tilde{p} = 1$, pointing out to the existence of a single changepoint. This agrees with the true origin of the simulated data. Notice that this posterior probability does not have a very pronounced mode and that posterior probabilites to the right of the mode decrease very slowly. This behavior is typical of small data sets where, in addition, the models before and after the changepoint do not differ much in the range where the covariates lay as seen from Figure 2.

On the other hand, in the space of all models, the six most probable models, are displayed in Table 3 in decreasing order.

The analysis of these data reveals that there is a —not clearcut— single changepoint at position 12 but, on the other hand, the no change model has higher posterior probability than that model. The rest of models have smaller posterior probability, but most of them are single changepoint models in the neiborhood of M_{12} as seen from Table 4. It is worthwhile remarking that, for these data, the forward selection procedure computes the best models within each box $\mathfrak{B}_p$ as the exact method for all $p = 0, 1, 2 \ldots, 5$. These results are also confirmed by the monitoring procedure, showing that the first permanent changepoint occurs at position 12.

Table 2: *Posterior probabilities of the number of changepoints p.*

p	0	1	2	3	4	5
$\pi(p\|\mathbf{y})$	0.162	0.229	0.180	0.154	0.142	0.132

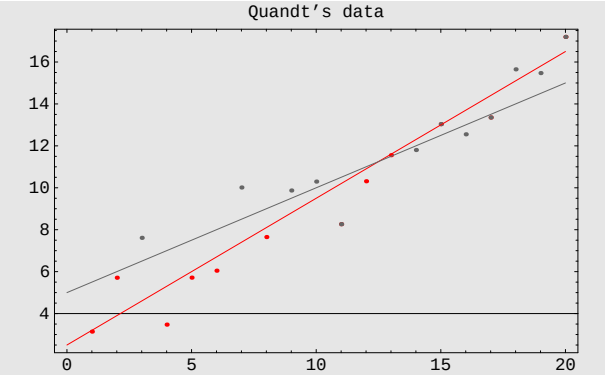

Figure 2: *Quandt's simulated data from two regression lines.*

Table 3: *Posterior probabilities of the most probable models $M_{\mathbf{r}}$.*

Models $\mathbf{r}$	0	12	10	11	9,12	9
$P(M_{\mathbf{r}}\|\mathbf{y})$	0.162	0.116	0.017	0.017	0.013	0.013

6.2. *Real data*

The following example refers to a famous real data set, which is also a favorite in the changepoint literature.

Example 3. The Nile River Data. The data appearing in Figure 4 are measurements of the annual volume of discharge from the Nile River at Aswan for the years 1871 to 1970.

This series was examined by Cobb (1978), Carlstein (1988), Dümbgen (1991), Balke (1993), and Moreno *et al.* (2005) among others, and the plot of the data reveals a marked and long-recognized steady decrease in annual volume after 1898.

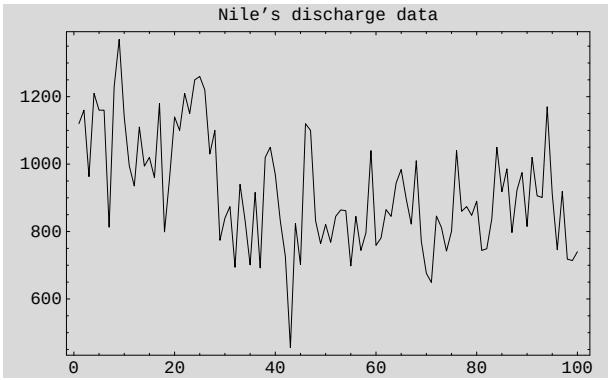

Figure 3: *Nile's discharge data for the years 1871 to 1970.*

Some authors have associated the drop to the presence of a dam that began operation in 1902, but Cobb (1978) cited independent evidence on tropical rainfall records to support the decline in volume. A Bayesian analysis using the jump diffusions, is found in Phillips and Smith (1996), though their results are not very accurate probably due to the small number of iterations of the algorithm. Denison *et al.* (2002) also consider the Bayesian analysis of these data. Both analysis use broadly uninformative conjugate priors. Another Bayesian analysis utilizing fractional Bayes factors and allowing for correlation in the error terms, is that of Garish and Groenewald (1999), which produces results closer to ours, and provides further references on other analysis of these data.

The results from the exact analysis conditional on there being up to three changepoints, *i.e.*, , $p \leq 3$ can be summarized as follows: As seen from Table 4, the posterior mode of $\pi(p|\mathbf{y})$ is $\tilde{p} = 1$, clearly pointing out to the existence of a single changepoint.

On the other hand, in the space of all models, the eight most probable models are displayed in Table 5, in decreasing order.

Table 4: *Posterior probabilities of the number of changepoints p.*

p	0	1	2	3
$\pi(p\|\mathbf{y})$	0.000	0.615	0.258	0.127

Table 5: *Posterior probabilities of the most probable models $M_{\mathbf{r}}$.*

Models $\mathbf{r}$	28	27	26	29	19, 28	21, 28	20, 28	28, 97
$P(M_{\mathbf{r}}\|\mathbf{y})$	0.466	0.076	0.036	0.029	0.006	0.006	0.006	0.005

The forward search produces the same results as the exact procedure for $p = 0, 1, 2$. For $p = 3$, the forward search finds model $M_{10,19,28}$ with posterior probability 0.00052 which is the fifth in position within box $\mathfrak{B}_3$, the first being model $M_{28,83,95}$ with posterior probability 0.00082, which is slightly more probable than the former. The stochastic search mostly confirms the above results as the most visited model occurs at position 28.

Summarizing our findings in analyzing these data: there is a single clearcut changepoint at position 28. The four most favored models are single changepoint models in the neighborhood of 28, and the probability of models with two or three changepoints are very small. The analysis of the remaining data, once the first 28 data points are deleted, clearly confirms that the point detected by the monitoring procedure at position 28 is the only one changepoint.

7. FURTHER DISCUSSION

In this paper we regard the detection of changepoints in the distribution of a sequential sample of observations as a model selection problem or, equivalently, as a multiple testing problem; consequently, we feel that a Bayesian formulation is the simplest and accuratest way to deal with it. Changepoint problems can be analysed from two different but otherwise complementary perspectives: sequential and retrospective. The former focuses on the first time a changepoint occurs and the second investigates the number and position of the changes in the sample based on the knowledge of the whole sample. However, by treating the first problem as a retrospective analysis at each step, we get a monitoring procedure which seems to be an excellent tool for constructing a sensible stopping rule.

We feel that monitoring should be preferred to the two stopping rules described in subsection 2.2.1 based on Bayes factors, which generalize the classical ones based on the likelihood estimators (Moreno *et al.*, 2005). Nevertheless, we want to stress the main difference between the sequential analysis —-monitoring— in which the class of sampling models consists of the no change model and the set of models with exactly one changepoint, and the retrospective analysis where we consider the class of all possible models.

The main difficulties to carry out a retrospective analysis come from two sources: i) the fact that for the continuous parameters involved in the problem —the regression coefficients and the variance errors—, the usual objective priors are unsuitable for computing model posterior probabilities as they are improper, and ii) the huge number of models involved for large, or even moderate, sample size makes unfeasible the computation of all model posterior probabilities.

These difficulties are solved by using intrinsic priors for the continuous model parameters, and a Metropolis-like stochastic search algorithm for selecting those models with highest posterior probabilities. For the integer parameters, we have assumed a uniform prior for the number of changepoints and, conditional on them, a uniform prior for their models, justified by both theoretical and practical arguments. When some locations are subjectively excluded, for instance we do not allow consecutive changepoints, the posterior distribution on the models or configurations is truncated accordingly; this is the simplest computational strategy.

The above priors provide an automatic and quite simple Bayesian analysis in which there are no hyperparameters to be adjusted. Numerical examples indicate that this objective Bayesian formulation behaves extremely well for detecting changepoints and finding the distribution of their locations. We have also realized that when the number of possible models is very large the Metropolis algorithm is able to find those models with highest posterior probability in a reasonable computing time.

An additional property of the intrinsic priors is that —as they are proper distributions conditionally on the parameters of the no change model and, further, have a close form in terms of multivariate normal and half-Cauchy distributions— they are also amenable to the implementation of a simple Gibbs algorithm involving sampling

from standard distributions plus a simple additional Metropolis step. Although this property has not been exploited in the paper, it will be of the utmost importance for the extension of our model to the heteroscedastic case.

REFERENCES

Bacon, D. W. and Watts, D. G. (1971). Estimating the transition between two intersecting straight lines. *Biometrika* **58**, 525–534.

Balke, N. S. (1993). Detecting level shifts in time series. *J. Business Econ. Statist.* **11**, 81–92.

Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

Box, G. and Tiao, G. (1992). *Bayesian Inference in Statistical Analysis. 2nd Edition*. New York: Wiley.

Carlstein, E. (1988). Nonparametric change-point estimation. *Ann. Statist.* **16**, 188–197.

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection, *J. Amer. Statist. Assoc.* **101**, 157–167.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian Analysis of changepoint Problems. *Applied Statist.* **41**, 389–405.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.* **35**, 999–1018.

Choy, J. H. and Broemeling, L. D. (1980). Some Bayesian inferences for a changing linear model. *Technometrics* **22**, 71–78.

Cobb, G. W. (1978). The problem of the Nile: conditional solution to a change-point problem. *Biometrika* **65**, 243–251.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods fro Nonlinear Classification and Regression.* Wiley: Chichester.

Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.* **19**, 1471–1495.

Ferreira, P.E. (1975). A Bayesian analysis of a switching regression model: known number of regimes. *J. Amer. Statist. Assoc.* **70**, 370–374.

Garisch, I. and Groenewald, P. C. (1999). The Nile revisited: Changepoint analysis with autocorrelation. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 753–760.

Girón, F. J., Moreno, E. and Martínez, M. L. (2006a). An objective Bayesian procedure for variable selection in regression. *Advances on Distribution Theory, Order Statistics and Inference.* (N. Balakrishnan *et al.*, eds.) Boston: Birkhauser, 393–408.

Girón, F. J., Martínez, M. L., Moreno, E. and Torres, F. (2006b). Objective Testing Procedures in Linear Models: Calibration of the *p*-values. *Scandinavian J. Statist.* (to appear).

Kiuchi, A. S., Hartigan, J. A., Holford, T. R., Rubinstein, P. and Stevens, C. E. (1995). changepoints in the series of T4 counts prior to AIDS. *Biometrics* **51**, 236–248.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *J. Roy. Statist. Soc. B* **57**, 613–658.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Amer. Statist. Assoc.* **93**, 1451–1460.

Moreno, E., Casella, G. and García-Ferrer, A. (2005). Objective Bayesian analysis of the changepoint problem. *Stochastic Environ. Res. Risk Assessment* **19**, 191–204.

Moreno, E. and Girón, F.J. (2006). On the frequentist and Bayesian approaches to hypothesis testing (with discussion). *Sort* (to appear).

Menzefrike, U. (1981). A Bayesian analysis of a change in the precision for a sequence of independent normal random variables at an unknown time point. *Applied Statist.* **30**, 141–146.

Phillips, D. V. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.) London: Chapman and Hall, 215–239.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeys two separate regimes. *J. Amer. Statist. Assoc.* **53**, 873–870.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73**, 85–89.

Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* **62**, 407–416.

Smith, A. F. M. and Cook, D. G. (1980). Straight lines with a changepoint: a Bayesian analysis of some renal ransplant data. *Applied Statist.* **29**, 180–189.

Spiegelhalter, D. J. and Smith A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. B* **44**, 377–387.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statist.* **43**, 159–168.

DISCUSSION

RAÚL RUEDA (*IIMAS, UNAM, Mexico*)

This paper is one more of a series of papers by the authors and their collaborators about model selection using intrinsic Bayes factors and related concepts. It is well written and easy to follow. It was a pleasure to read it.

The paper deals with the following situation:

Sampling from a normal linear model,

$$y_1, \ldots, y_n \sim N(y|\boldsymbol{\beta}, \tau^2 \mathbf{I}_n),$$

the problem is to infer the number p of potential changepoints and their corresponding positions in the sample, $\mathbf{r}_p = \{r_1, \ldots, r_p\}$, as well as to make inferences on the regression parameters $\boldsymbol{\beta}$ or transformations thereof.

There are three main tasks in this model selection problem:

i). Prior assignation. Intrinsic Bayes factors cannot be used here, since a changepoint can occur in the *training* sample; besides, the order is relevant.

ii). Implementation. The number of models, and therefore, the number of comparisons in retrospective analysis increases with the sample size.

iii). Monitoring. In prospective analysis, where the search for a changepoint is carried out sequentially, the proposed strategy for on-line detection has advantages over a stopping rule.

Concerning the *implementation* and *monitoring* issues, the authors propose a forward searching and a Metropolis-Hasting algorithm for the retrospective search which at the same time, provide a monitoring plan for the prospective case without any additional effort.

The need for automatic procedures for model selection, including hypotheses testing, has produced a variety of "objective Bayes factors", which in some cases behave like real Bayes factors, at least asymptotically. Almost everybody agrees that the use of improper priors must be avoided when computing a Bayes factor.

This is because improper priors do not, in general, define unambiguously the Bayes factor. Unfortunately, "automatic" priors are typically improper.

In almost all the proposals, the basic idea is "to pay" with some sample information in order to obtain proper posterior distribution from the improper prior and use this posterior as a prior to define a real Bayes factor. Thus, we have *training samples, minimum training samples, partial samples* and so on. But the intrinsic priors, introduced by Berger and Pericchi (1996), use *imaginary samples*, so in the end we do not have to pay for anything! An advantage of the automatic Bayes factor resulting from using intrinsic priors is that "it is close to an 'actual' Bayes factor as desired".

However, it seems that the need for automatic or objective procedures has produced precisly what we criticise about frequentist statisitcs: a collection of *ad hoc* methods out of Bayesian principles. As an example of this, the authors use *intrinsic* priors for model selection and *reference* priors to estimate the magnitude of the changes. Morever, Bayes factors, including *actual* Bayes factors, can produce incoherent results, as Lindley's examples in the discussion of Aitkin's paper show (Aitkin, 1991).

A question for the authors: what about robustness? It is known that Bayes factors, as a device for model selection, arise from the $\mathcal{M}$-closed perspective, so what happens if, for example, the errors are Student-t instead of normal?. On a related note, I understand that the *intrinsic* prior depends on a "standard" noninformative prior. How sensitive is it to this choice?

I congratulate the authors for a very interesting paper. I thank the authors because they have forced me to take an intensive course on Bayes factors and all their *automatic* extensions. When I first read his paper, a thought appeared almost immediately in my mind...I hope that The Beatles will forgive me:

When I was younger, so much younger than today,
I only needed random samples to use factors of Bayes
But now these days are gone and I feel so insecure
Now I am also confused with all this new stuff

Help me if you can, with the training samples
And before they become imaginary
Help me get my feet back on the ground
Won't you please, please help me.

NICOLAS CHOPIN and PAUL FEARNHEAD
(*University of Bristol, UK and Lancaster University, UK*)

Change point modelling is a 'treat' for Bayesians, as frequentist methods are particularly unsatisfactory in such settings: asymptotics do no make sense if a given model is assumed to be true only for a finite interval of time. Moreover, practitioners are increasingly turning their attention to simple models, which are allowed to change over time, as an alternative to overly complicated time series models (e.g. long memory processes) that are difficult to interpret. Any contribution to this interesting field is therefore most welcome.

Our first comments relate to the choice of prior. The objective prior chosen for the number of changepoints appears slightly inconsistent in the following sense. Imagine analysing a data set with $2n$ observations, but where you are first given

the first n observations to analyse, then the second half later. The uniform prior on the number of changepoints within the first n observations, together with the same prior on the number of changepoints in second n observation implies a prior on the number of changepoints within the full data set that is not uniform. In more practical terms, the choice of a uniform prior appears unrealistic: we cannot think of any application where length of segments (distance between consecutive changepoints) should be as small as 2. A more coherent approach is to model directly the length of each segment (resulting in a product-partition prior structure, see Barry and Hartigan, 1993). This produces both consistent priors (in the sense described above), and is modelling directly an important feature of the data.

In that respect, trying to be non-informative may not always be relevant or useful for change point models. As for any parameter with a physical dimension, there is almost always prior information on these durations which can easily extracted: for daily data for instance, common sense is enough to rule out durations larger than a few years. This kind of consideration is essential in on line applications, where the sample size n of the complete data is not known in advance.

Secondly, an alternative approach to using intrinsic priors is to have a hierarchical model structure where you introduce hyperpriors on the hyperparameters of the priors for the parameters of each segment. This enables the data to inform the choice of prior for the segment parameters. It is possible to choose the hyperpriors to be improper in many situations (see Punskaya *et al.*, 2002 and Fearnhead, 2006). One advantage of this is that the priors on the parameters of each segment can be interpreted in terms of describing the variation in features of the model (e.g. means) across the different segments. It also seems mathematically and computationally simpler than using intrinsic priors.

Finally we would like to point out some other work on Bayesian analysis of multiple changepoint models. Perfect (iid) sampling from the posterior is possible for certain classes of changepoint models (Barry and Hartigan, 1993; Liu and Lawrence, 1998; Fearnhead , 2005, 2006). Otherwise, standard and trans-dimensional MCMC algorithms can be derived (Green, 1995; Gerlach *et al.*, 2000), as well as particle filters (Fearnhead and Clifford, 2003; Fearnhead and Liu, 2006; Chopin, 2006) for on-line inference. These particle filters can also be used for off-line inference, and at a smaller cost than MCMC, *i.e.*, $O(n)$ instead of $O(n^2)$. The MCMC algorithm proposed in the discussed paper is also implicitly $O(n^2)$: each iteration is $O(n)$, but proposing at random a new change means that $O(n)$ iterations are required to visit a given location. In fact, since the considered linear model allows for conjugacy, it would be interesting to see if one of the exact method mentioned above could be used as an importance sampling proposal.

REPLY TO THE DISCUSSION

Rueda. We thank Raúl Rueda for his comments on the paper and answer the questions he raises in his discussion.

We note that the conditional intrinsic prior is proper, and the unconditional intrinsic prior is improper. Therefore, this impropriety is inherited from the impropriety of the objective reference prior that we start with. However, the Bayes factor for intrinsic priors is a well defined limit of Bayes factors for proper priors.

The use of intrinsic priors for testing and reference priors for estimation is not at all *ad hoc*. We know that reference priors are perfectly reasonable objective priors for estimation, but they are not well-calibrated so cannot be used for testing. The

intrinsic approach calibrates the reference prior for testing so, in fact, these two priors are connected in this way, with each being suited for its particular task.

We have taken another look at the Lindley/Aitkin discussion, and find that we neither agree with Lindley nor with Rueda. As Aitkin points out in his rejoinder, it is not Bayes factors, but Lindley's criterion, "that is ridiculous". To illustrate, Lindley asserts that if we prefer model $\mathcal{M}_1$ to $\mathcal{M}_2$ and $\mathcal{M}_3$ to $\mathcal{M}_4$, then we should prefer $\mathcal{M}_1 \cup \mathcal{M}_3$ to $\mathcal{M}_2 \cup \mathcal{M}_4$. But this does not always follow. Suppose that in a four-variable regression problem, the best model with two regressors is $\{x_2, x_4\}$, and the univariate models are ordered, from best to worst, $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$. Then we prefer $\{x_1\}$ to $\{x_2\}$ and $\{x_3\}$ to $\{x_4\}$, but $\{x_2, x_4\}$ beats any other model with two regressors. So Lindley's preference ordering is not self-evident.

In our approach to changepoint problems we have assumed that the underlying regression models are normal. This assumption allows —in the monitoring procedure— for distinguishing between outliers and permanent changepoints. If, instead, we would consider Student errors —apart from the problem of deriving the corresponding intrinsic priors and Bayes factors, as there is no sufficient statistic of fixed dimension— it is not at all clear that outliers and changepoints could be distinguised.

The question about how sensitive the Bayes factor is to the choice of the non-informative prior when intrinsic priors are used is discussed at length in subsection 3.5 of Girón *et al.* (2006b) where, in the normal regression setting, we compare the influence of using priors of the form

$$\pi^N(\theta, \tau) \propto \frac{1}{\tau^q} \quad \text{for} \quad q = 1 \ldots, k,$$

in the posterior probability of the null.

Note that this class of priors includes the reference, when $q = 1$, and the Jeffreys' prior when $q = k$. The conclusion was that the choice of the usual reference prior results in a a much more stable procedure for comparing models. Box and Tiao (subsection 2.4.6, pp. 101-102, 1992) also discuss the sensitivity of the posterior with respect to the choice of the prior within this class in the standard normal setting.

Finally, we thank Raúl Rueda for his thoughtful comments, and add, in a spirit similar to his:

Hey Raúl, don't make it bad
We take a sad prior, and make it better
Remember to let it into your heart
Then you can start to make it better.

Chopin and Fearnhead. We thank Chopin and Fearnhead for their thoughtful comments. We will reply to each separately.

Choice of Prior. You are right in saying that the prior on the set of models should not be uniform. A uniform prior penalizes the model with small number of changes and this is not reasonable. However, we have not used a uniform prior on the set of models, but rather have used a uniform prior on the set of boxes, and then spread mass uniformly among the models within each box. Moreover, This prior only depend on the ancillary sample size statistic n, so that we do not believe it suffers from the slight inconsistency that you mention.

An objective analysis of a statistical problem is appropriate when you do not have subjective prior information. If you have such information, by all means use it!

For example, in analyzing the famous Coal-Mining Disaster Data, Fearnhead (2006) chooses a Gamma prior (with specific values for the hyperparameters) for the mean of the Poisson sampling distribution, and a Poisson distribution with mean 3 for the number of possible change points. We suspect that there are good reasons for choosing such specific priors.

Consecutive Changepoints. Although there may be cases when there is reason to model changepoints through segmental length, recall that we are considering objective methods that do not take into account special features of the data.

However, if there is reason to impose the constraints on the length of the segments between two changepoints, this can be accomplished through restrictions on some configurations. For example, if $\mathfrak{R}_k$ is a set of restrictions requiring segments to have length greater than k, a simple combinatorial calculation shows that, once those configurations have been deleted, the objective prior $\pi(p|\mathfrak{R}_k)$ is

$$\pi(p|\mathfrak{R}_k) \propto \frac{\binom{n-1-k(p-1)}{p}}{\binom{n-1}{p}} \text{ for } p = 0, 1, \ldots, n-1.$$

It then follows that $\pi(p|\mathfrak{R}_k) = 0$ for $p \geq (n-1)/k$ and all such boxes $\mathfrak{B}_p$ are empty. Furthermore, $\pi(p|\mathfrak{R}_k)$ is a decreasing function of p with $\pi(0|\mathfrak{R}_k) = \pi(1|\mathfrak{R}_k)$ for all k. This last property implies that for monitoring, where we only consider boxes $\mathfrak{B}_0$ and $\mathfrak{B}_1$, each box has probability 1/2.

Lastly, we have also commented in Section 3.1 about the possibility of incorporating additional information which results in constraints on the *posterior* instead of on the prior, which may be technically easier in some cases.

Hierarchical Models. Models based on intrinsic priors are inherently hierarchical, and we believe that, in general, hierarchical models are very useful. However, using improper priors in the last stage of a hierarchy results in improper marginals and can result in improper p osterior distributions (Hobert and Casella, 1996). This means that the Bayes factors are not well defined. However, intrinsic priors always provide well defined Bayes Factors.

We are convinced that the natural formulation of the change point problem is as a model selection problem, that is, it is a *testing problem* rather than an estimation problem. As such, improper priors cannot be used in this context. However, intrinsic priors can be used for both testing and estimation. From the details in Section 4.4, and using the hierarchical structure of the intrinsic priors, it is easy to set up a Gibbs Sampler to estimate all of the posterior parameters.

Computational Issues. You are right to point out that the algorithm used in the paper is $O(n^2)$, however, it should also be pointed out that the calculations are very fast in `R`, and large scale searches are feasible. Also, we have also developed another algorithm, b ased on an independent Metropolis-Hastings scheme that exploits the box structure, (see Casella and Moreno 2006) that is uniformly ergodic.

The exact methods that you discuss are indeed faster, but we have decided to use methods that are based on simpler algorithms in order to allow our methods to apply to more complex models. It seems that, for example, the faster recursions in Fearnhead (2006) do not apply to very general models without embedding them in a Gibbs sampler.

We also point out that our search algorithm is driven by a very specific objective function, the posterior probability of the models. In searching such spaces, we have

found that a Metropolis-Hastings algorithm with a well-chosen candidate is hard to bea t. Even with isolated modes, where we may need transition kernels based on a tempering structure, the Metropolis-Hastings Algorithm is typically an excellent choice (Jerrum and Sinclair, 1996).

ADDITIONAL REFERENCES IN THE DISCUSSION

Aitkin, M. (1991). Posterior Bayes factors. *J. Roy. Statist. Soc. B* **53**, 131 - 138. (with discussion).

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88**, 309–319.

Chopin, N. (2006). Dynamic detection of change points in long time series. *Ann. Inst. Statist. Math.* (to appear).

Fearnhead, P. (2005). Exact Bayesian curve fitting and signal segmentation. *IEEE Trans. Signal Processing* **53** ,2160–2166.

Fearnhead, P. (2006). Exact and efficient inference for multiple changepoint problems. *Statist. Computing* **16**, 203–213.

Fearnhead, P. and Clifford, P. (2003). On-line inference for hidden Markov models. *J. Roy. Statist. Soc. B* **65**, 887–899.

Fearnhead, P. and Liu, Z. (2006) On-line inference for multiple changepoint problems. *Submitted*; available from `www.maths.lancs.ac.uk/~fearnhea/publications`.

Gerlach, R., Carter, C., and Kohn, R. (2000) Efficient Bayesian inference for dynamic mixture models. *J. Amer. Statist. Assoc.* **88** ,819–828.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. Amer. Statist. Assoc.* **91**, 1461–1473.

Jerrum, M. and Sinclair, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. *Approximation Algorithms for NP-hard Problems* (D. S. Hochbaum, ed.) Boston: PWS Publishing, 482–520.

Liu, J. S. and Lawrence, C. E. (1998). Bayesian inference on biopolymer models.*Bioinformatics* **15**, 38–52.

Punskaya, E., Andrieu, C., Doucet , A. and Fitzgerald, W. J. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Trans. Signal Processing* **50**, 747–758.