Bayesian Loss-based Approach to Change Point Analysis

Laurentiu C. Hinoveanu[1], Fabrizio Leisen [*1], and Cristiano Villa[1]

[1]School of Mathematics, Statistics and Actuarial Science, University of Kent

Abstract

A loss-based approach to change point analysis is proposed. In particular, the problem is looked from two perspectives. The first focuses on the definition of a prior when the number of change points is known a priori. The second contribution aims to estimate the number of change points by using a loss-based approach recently introduced in the literature. The latter considers change point estimation as a model selection exercise. The performance of the proposed approach it is shown on simulated data and real data sets.

Keywords: Change point; Discrete parameter space; Loss-based prior; Model selection.

1 Introduction

There are several practical scenarios where it is inappropriate to assume that the distribution of the observations does not change. For example, financial data sets can exhibit alternate behaviours due to crisis periods. In this case it is sensible to assume changes in the underlying distribution. The change in the distribution can be either in the value of one or more of the parameters or, more in general, on the family of the distribution. In the

[*]School of Mathematics, Statistics and Actuarial Sciences, Sibson Building, University of Kent, Canterbury, CT2 7FS, F.Leisen@kent.ac.uk

latter case, for example, one may deem appropriate to consider a normal density for the stagnation periods, while a Student t, with relatively heavy tails, may be more suitable to represent observations in the more turbulent stages of a crisis. The task of identifying if, and when, one or more changes have occurred is not trivial and requires appropriate methods to avoid detection of a large number of changes or, at the opposite extreme, seeing no changes at all. The change point problem has been deeply studied from a Bayesian point of view. Chernoff and Zacks (1964) focused on the change in the means of normally distributed variables. Smith (1975) looked into the single change point problem when different knowledge of the parameters of the underlying distributions is available: all known, some of them known or none of them known. Smith (1975) focuses on the binomial and normal distributions. In Muliere and Scarsini (1985) the problem is tackled from a Bayesian nonparametric perspective. The authors consider Dirichlet processes with independent base measures as underlying distributions. In this framework, Petrone and Raftery (1997) have showed that the Dirichlet process prior could have a strong effect on the inference and may lead to wrong conclusions in the case of a single change point. Raftery and Akman (1986) have approached the single change point problem in the context of a Poisson likelihood under both proper and improper priors for the model parameters. Carlin et al. (1992) build on the work of Raftery and Akman (1986) by considering a two level hierarchical model. Both papers illustrate the respective approaches by studying the well-known British coal-mining disaster data set. In the context of multiple change points detection, Loschi and Cruz (2005) have provided a fully Bayesian treatment for the product partitions model of Barry and Hartigan (1992). Their application focused on stock exchange data. Stephens (1994) has extended the Gibbs sampler introduced by Carlin et al. (1992) in the change point literature to handle multiple change points. Hannart and Naveau (2009) have used Bayesian decision theory, in particular 0-1 cost functions, to estimate multiple changes in homoskedastic normally distributed observations. Schwaller and Robin (2017) extend the product partition model of Barry and Hartigan (1992) by adding a graphical structure which could capture the dependencies between multivariate observations. Fearnhead and Liu (2007) proposed a filtering algorithm for the sequential multiple change points detection problem in the case of piecewise regression models. Henderson and Matthews (1993) introduced a partial Bayesian approach which involves the use of a profile likelihood, where the aim is to detect multiple changes in the mean of Poisson distributions with an application to haemolytic uraemic syndrome (HUS) data. The same data set was studied by Tian et al. (2009), who proposed a method which treats the change points as latent variables. Ko et al. (2015) have proposed an exten-

sion to the hidden Markov model of Chib (1998) by using a Dirichlet process prior on each row of the regime matrix. Their model is semiparametric, as the number of states is not specified in advance, but it grows according to the data size. Heard and Turcotte (2017) have proposed a new sequential Monte Carlo algorithm to infer multiple change points. Other contributions to the Bayesian change point literature are Harlé et al. (2016), Lai and Xing (2011), Martínez and Mena (2014) and Mira and Petrone (1995).

Whilst the literature covering change point analysis from a Bayesian perspective is vast when prior distributions are elicited, the documentation referring to analysis under minimal prior information is limited, see Moreno et al. (2005) and Girón et al. (2007). The former paper discusses the single change point problem in a model selection setting, whilst the latter paper, which is an extension of the former, tackles the multivariate change point problem in the context of linear regression models. Our work aims to contribute to the methodology for change point analysis under the assumption that the information about the number of change points and their location is minimal. First, we discuss the definition of an objective prior for change point location, both for single and multiple changes, assuming the number of changes is known a priori. Then, we define a prior on the number of change points via a model selection approach. Here, we assume that the change point coincides with one of the observations. As such, given $X_1, X_2, \ldots, X_n$ data points, the change point location is discrete. To the best of our knowledge, the sole general objective approach to define prior distributions on discrete spaces is the one introduced by Villa and Walker (2015b).

To illustrate the idea, consider a probability distribution $f(x|m)$, where $m \in \mathbb{M}$ is a discrete parameter. Then, the prior $\pi(m)$ is obtained by objectively measuring what is lost if the value m is removed from the parameter space, and it is the true value. According to Berk (1966), if a model is misspecified, the posterior distribution asymptotically accumulates on the model which is the most similar to the true one, where the similarity is measured in terms of the Kullback–Leibler (KL) divergence. Therefore, $D_{KL}(f(\cdot|m)\|f(\cdot|m'))$, where m' is the parameter characterising the nearest model to $f(x|m)$, represents the utility of keeping m. The objective prior is then obtained by linking the aforementioned utility via the self-information loss:

$$\pi(m) \propto \exp\left\{\min_{m' \neq m} D_{KL}(f(\cdot|m)\|f(\cdot|m'))\right\} - 1, \tag{1}$$

where the Kullback–Leibler divergence (Kullback and Leibler, 1951) from the sampling distribution with density $f(x|m)$ to the one with density $f(x|m')$

is defined as:

$$D_{KL}(f(\cdot|m)\|f(\cdot|m')) = \int_{\mathcal{X}} f(x|m) \cdot \log\left[\frac{f(x|m)}{f(x|m')}\right] \mathrm{d}x.$$

Throughout the paper, the objective prior defined in equation (1) will be referenced as the *loss-based* prior. This approach is used to define an objective prior distribution when the number of change points is known a priori. To obtain a prior distribution for the number of change points, we adopt a model selection approach based on the results in Villa and Walker (2015a), where a method to define a prior on the space of models is proposed. To illustrate, let us consider k Bayesian models:

$$M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\} \qquad j \in \{1, 2, \ldots, k\}, \tag{2}$$

where $f_j(x|\theta_j)$ is the sampling density characterised by θ_j and $\pi_j(\theta_j)$ represents the prior on the model parameter.

Assuming the prior on the model parameter, $\pi_j(\theta_j)$, is proper, the model prior probability $\Pr(M_j)$ is proportional to the expected minimum Kullback–Leibler divergence from M_j, where the expectation is considered with respect to $\pi_j(\theta_j)$. That is:

$$\Pr(M_j) \propto \exp\left\{\mathbb{E}_{\pi_j}\left[\inf_{\theta_i, i\neq j} D_{KL}(f_j(x|\theta_j)\|f_i(x|\theta_i))\right]\right\} \qquad j = 1, \ldots, k. \tag{3}$$

The model prior probabilities defined in equation (3) can be employed to derive the model posterior probabilities through:

$$\Pr(M_i|x) = \left[\sum_{j=1}^{k} \frac{\Pr(M_j)}{\Pr(M_i)} B_{ji}\right]^{-1}, \tag{4}$$

where B_{ji} is the Bayes factor between model M_j and model M_i, defined as

$$B_{ji} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j)\,\mathrm{d}\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i)\,\mathrm{d}\theta_i},$$

with $i \neq j \in \{1, 2, \ldots, k\}$.

This paper is structured as follows: in Section 2 we establish the way we set objective priors on both single and multiple change point locations. Section 3 shows how we define the model prior probabilities for the number of change

point locations. Illustrations of the model selection exercise are provided in Sections 4 and 5, where we work with simulated and real data, respectively. Additionally, in Section 4 we perform a comparison of the proposed method to another Bayesian approach discussed in the literature. Section 6 is dedicated to final remarks.

2 Objective Prior on the Change Point Locations

This section is devoted to the derivation of the loss-based prior when the number of change points is known a priori. Specifically, let k be the number of change points and $m_1 < m_2 < \ldots < m_k$ their locations. We introduce the idea in the simple case where we assume that there is only one change point in the data set (see Section 2.1). Then, we extend the results to the more general case where multiple change points are assumed (see Section 2.2). Note that we assume that the change in the dataset occurs after the identified point. For instance, in the case of one change point, m implies that the actual change occurs from the X_{m+1} observation.

A well-known objective prior for finite parameter spaces, in cases where there is no structure, is the uniform prior (Berger et al., 2012). As such, a natural choice for the prior on the change points location is the uniform (Koop and Potter, 2009). The corresponding loss-based prior is indeed the uniform, as shown below, which is a reassuring result as the objective prior for a specific parameter space, if exists, should be unique.

2.1 Single Change Point

As mentioned above, we show that the loss-based prior for the single change point case coincides with the discrete uniform distribution over the set $\{1, 2, \ldots, n-1\}$.

Let $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ denote an n-dimensional vector of random variables, representing the random sample, and m be our single change point location, that is $m \in \{1, 2, \ldots, n-1\}$, such that

$$\begin{aligned} X_1, \ldots, X_m | \tilde{\theta}_1 &\overset{\text{i.i.d.}}{\sim} f_1(\cdot|\tilde{\theta}_1) \\ X_{m+1}, \ldots, X_n | \tilde{\theta}_2 &\overset{\text{i.i.d.}}{\sim} f_2(\cdot|\tilde{\theta}_2). \end{aligned} \tag{5}$$

Note that we assume that there is a change point in the series, as such the space of m does not include the case $m = n$. In addition, we assume that $\tilde{\theta}_1 \neq \tilde{\theta}_2$ when $f_1 = f_2$. The sampling density for the vector of observations $\mathbf{x}^{(n)} = (x_1, \ldots, x_n)$ is:

$$f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) = \prod_{i=1}^{m} f_1(x_i|\tilde{\theta}_1) \prod_{i=m+1}^{n} f_2(x_i|\tilde{\theta}_2). \tag{6}$$

Let $m' \neq m$. Then, the Kullback–Leibler divergence between the model parametrised by m and the one parametrised by m' is:

$$D_{KL}(f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \| f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)) = \int f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \log\left(\frac{f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2)}{f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)}\right) \mathrm{d}\mathbf{x}^{(n)}. \tag{7}$$

Without loss of generality, consider $m < m'$. In this case, note that

$$\frac{f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2)}{f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)} = \prod_{i=m+1}^{m'} \frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)},$$

leading to

$$D_{KL}(f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \| f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)) = \sum_{i=m+1}^{m'} \int f_2(x_i|\tilde{\theta}_2) \log\left(\frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}\right) \mathrm{d}x_i. \tag{8}$$

On the right hand side of equation (8), we can recognise the Kullback–Leibler divergence from density f_2 to density f_1, thus getting:

$$D_{KL}(f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) || f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)) = (m' - m) D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)). \tag{9}$$

In a similar fashion, when $m > m'$, we have that:

$$D_{KL}(f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \| f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)) = (m - m') D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_2(\cdot|\tilde{\theta}_2)). \tag{10}$$

In this single change point scenario, we can consider m' as a perturbation of the change point location m, that is $m' = m \pm l$ where $l \in \mathbb{N}^*$, such

that $1 \leq m' < n$. Then, taking into account equations (9) and (10), the Kullback–Leibler divergence becomes:

$$D_{KL}(f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \| f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)) = \begin{cases} l \cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)), & \text{if } m < m' \\ l \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_2(\cdot|\tilde{\theta}_2)), & \text{if } m > m', \end{cases}$$

and

$$\begin{aligned} \min_{m' \neq m} \left[D_{KL}(f(\mathbf{x}^{(\mathbf{n})}|m, \tilde{\theta}_1, \tilde{\theta}_2) \| f(\mathbf{x}^{(\mathbf{n})}|m', \tilde{\theta}_1, \tilde{\theta}_2)) \right] &= \\ = \min_{m' \neq m} \{ l \cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)), l \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_2(\cdot|\tilde{\theta}_2)) \} & \\ = \min_{m' \neq m} \{ D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)), D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_2(\cdot|\tilde{\theta}_2)) \} \cdot \underbrace{\min_{m' \neq m} \{l\}}_{1}. & \end{aligned} \tag{11}$$

We observe that equation (11) is only a function of $\tilde{\theta}_1$ and $\tilde{\theta}_2$ and does not depend on m. Thus, $\pi(m) \propto 1$ and, therefore,

$$\pi(m) = \frac{1}{n-1} \qquad m \in \{1, \ldots, n-1\}. \tag{12}$$

This prior was used, for instance, in an econometric context by Koop and Potter (2009) with the rationale of giving equal weight to every possible change point location.

2.2 Multivariate Change Point Problem

In this section, we address the change point problem in its generality by assuming that there are $1 \leq k < n$ change points. In particular, for the data $\mathbf{x}^{(n)} = (x_1, \ldots, x_n)$, we consider the following sampling distribution

$$f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) = \prod_{i=1}^{m_1} f_1(x_i|\tilde{\theta}_1) \prod_{j=1}^{k-1} \prod_{i=m_j+1}^{m_{j+1}} f_{j+1}(x_i|\tilde{\theta}_{j+1}) \prod_{i=m_k+1}^{n} f_{k+1}(x_i|\tilde{\theta}_{k+1}), \tag{13}$$

where $\boldsymbol{m} = (m_1, \ldots, m_k)$, $1 \leq m_1 < m_2 < \ldots < m_k < n$, is the vector of the change point locations and $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_k, \tilde{\theta}_{k+1})$ is the vector of the parameters of the underlying probability distributions. Schematically:

$$\begin{array}{llll} X_1 & , \ldots , & X_{m_1}|\tilde{\theta}_1 & \overset{\text{i.i.d.}}{\sim} f_1(\cdot|\tilde{\theta}_1) \\ X_{m_1+1} & , \ldots , & X_{m_2}|\tilde{\theta}_2 & \overset{\text{i.i.d.}}{\sim} f_2(\cdot|\tilde{\theta}_2) \\ \vdots & , \ldots , & \vdots & \vdots \ldots \vdots \\ X_{m_{k-1}+1} & , \ldots , & X_{m_k}|\tilde{\theta}_k & \overset{\text{i.i.d.}}{\sim} f_k(\cdot|\tilde{\theta}_k) \\ X_{m_k+1} & , \ldots , & X_n|\tilde{\theta}_{k+1} & \overset{\text{i.i.d.}}{\sim} f_{k+1}(\cdot|\tilde{\theta}_{k+1}). \end{array}$$

If $f_1 = f_2 = \cdots = f_{k+1}$, then it is reasonable to assume that some of the θ's are different. Without loss of generality, we assume that $\tilde{\theta}_1 \neq \tilde{\theta}_2 \neq \cdots \neq \tilde{\theta}_k \neq \tilde{\theta}_{k+1}$. In a similar fashion to the single change point case, we cannot assume $m_k = n$ since we require exactly k change points.

In this case, due to the multivariate nature of the vector $\boldsymbol{m} = (m_1, \ldots, m_k)$, the derivation of the loss-based prior is not as straightforward as in the one dimensional case. In fact, the derivation of the prior is based on heuristic considerations supported by the below Theorem 1 (the proof of which is in the Appendix). In particular, we are able to prove an analogous of equations (9) and (10) when only one component is arbitrarily perturbed. Let us define the following functions:

$$\begin{aligned} d_j^{+1}(\tilde{\boldsymbol{\theta}}) &= D_{KL}(f_{j+1}(\cdot|\tilde{\theta}_{j+1}) \| f_j(\cdot|\tilde{\theta}_j)) \\ d_j^{-1}(\tilde{\boldsymbol{\theta}}) &= D_{KL}(f_j(\cdot|\tilde{\theta}_j) \| f_{j+1}(\cdot|\tilde{\theta}_{j+1})), \end{aligned}$$

where $j \in \{1, 2, \ldots, k\}$. The following Theorem is useful to understand the behaviour of the loss-based prior in the general case.

Theorem 1. *Let $f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}})$ be the sampling distribution defined in equation (13) and consider $j \in \{1, \ldots, k\}$. Let $\boldsymbol{m}'$ be such that $m'_i = m_i$ for $i \neq j$, and let the component m'_j be such that $m'_j \neq m_j$ and $m_{j-1} < m'_j < m_{j+1}$. Therefore,*

$$D_{KL}(f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \| f(\mathbf{x}^{(n)}|\boldsymbol{m}', \tilde{\boldsymbol{\theta}}) = |m'_j - m_j| d_j^S(\tilde{\boldsymbol{\theta}}),$$

where $S = \text{sgn}(m'_j - m_j)$.

Note that, Theorem 1 states that the minimum Kullback–Leibler divergence is achieved when $m'_j = m_j + 1$ or $m'_j = m_j - 1$. This result is not surprising since the Kullback–Leibler divergence measures the degree of similarity between two distributions. The smaller the perturbation caused by changes in

one of the parameters is, the smaller the Kullback–Leibler divergence between the two distributions is. Although Theorem 1 makes a partial statement about the multiple change points scenario, it provides a strong argument for supporting the uniform prior. Indeed, if now we consider the general case of having k change points, it is straightforward to see that the Kullback–Leibler divergence is minimised when only one of the components of the vector $\boldsymbol{m}$ is perturbed by (plus or minus) one unit. As such, the loss-based prior depends on the vector of parameters $\tilde{\boldsymbol{\theta}}$ only, as in the one-dimensional case, yielding the uniform prior for $\boldsymbol{m}$.

Therefore, the loss-based prior on the multivariate change point location is

$$\pi(\boldsymbol{m}) = \left\{ \binom{n-1}{k} \right\}^{-1}, \tag{14}$$

where $\boldsymbol{m} = (m_1, \ldots, m_k), 1 \leq m_1 < m_2 < \ldots < m_k < n$. The denominator in equation (14) has the above form because, for every number of k change points, we are interested in the number of k-subsets from a set of $n-1$ elements, which is $\binom{n-1}{k}$. The same prior was also derived in a different way by Girón et al. (2007).

3 Loss-based Prior on the Number of Change Points

Here, we approach the change point analysis as a model selection problem. In particular, we define a prior on the space of models, where each model represents a certain number of change points (including the case of no change points). The method adopted to define the prior on the space of models is the one introduced in Villa and Walker (2015a). We proceed as follows. Assume we have to select from $k+1$ possible models. Let M_0 be the model with no change points, M_1 the model with one change point and so on. Generalising, model M_k corresponds to the model with k change points. The idea is that the current model encompasses the change point locations of the previous model. As an example, in model M_3 the first two change point locations will be the same as in the case of model M_2. To illustrate the way we envision our models, we have provided Figure 1. It has to be noted that the construction of the possible models from M_0 to M_k can be done in a different way to one here described. Obviously, the approach to define the model priors stays

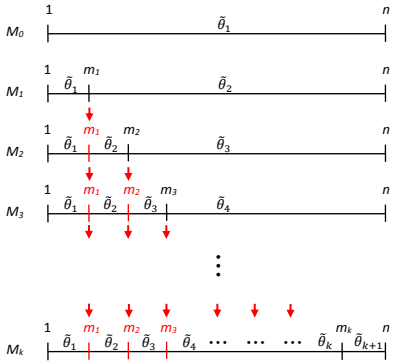

Figure 1: Diagram showing the way we specify our models. The arrows indicate that the respective change point locations remain fixed from the previous model to the current one.

unchanged. Consistently with the notation used in Section 1,

$$\theta_k = \begin{cases} \tilde{\theta}_1, \ldots, \tilde{\theta}_{k+1}, m_1, \ldots, m_k & \text{if } k = 1, \ldots, n-1 \\ \tilde{\theta}_1 & \text{if } k = 0, \end{cases}$$

represents the vector of parameters of model M_k, where $\tilde{\theta}_1, \ldots, \tilde{\theta}_{k+1}$ are the model specific parameters and $m_1, \ldots, m_k$ are the change point locations, as in Figure 1.

Based on the way we have specified our models, which are in direct correspondence with the number of change points and their locations, we state Theorem 2 (the proof of which is in the Appendix).

Theorem 2. *Let*

$$D_{KL}(M_i \| M_j) = D_{KL}(f(\mathbf{x}^{(n)}|\theta_i) \| f(\mathbf{x}^{(n)}|\theta_j)).$$

For any $0 \leq i < j \leq k$ integers, with $k < n$, and the convention $m_{j+1} = n$, we have the following:

$$D_{KL}(M_i \| M_j) = \sum_{q=i+1}^{j} \left[(m_{q+1} - m_q) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1}) \| f_{q+1}(\cdot|\tilde{\theta}_{q+1})) \right],$$

and

$$D_{KL}(M_j \| M_i) = \sum_{q=i+1}^{j} \left[(m_{q+1} - m_q) \cdot D_{KL}(f_{q+1}(\cdot|\tilde{\theta}_{q+1}) \| f_{i+1}(\cdot|\tilde{\theta}_{i+1})) \right].$$

The result in Theorem 2 is useful when the model selection exercise is implemented. Indeed, the Villa and Walker (2015a) approach requires the computation of the Kullback–Leibler divergences in Theorem 2. Recalling equation (3), the objective model prior probabilities are then given by:

$$\Pr(M_j) \propto \exp\left\{\mathbb{E}_{\pi_j}\left[\inf_{\theta_i, i\neq j} D_{KL}(M_j \| M_i)\right]\right\} \qquad j = 0, 1, \ldots, k. \tag{15}$$

For illustrative purposes, in the Appendix we derive the model prior probabilities to perform model selection among M_0, M_1 and M_2.

It is easy to infer from equation (15) that model priors depend on the prior distribution assigned to the model parameters, that is on the level of uncertainty that we have about their true values. For the change point location, a sensible choice is the uniform prior which, as shown in Section 2, corresponds to the loss-based prior. For the model specific parameters, we have several options. If one wishes to pursue an objective analysis, intrinsic priors (Berger and Pericchi, 1996) may represent a viable solution since they are proper. Nonetheless, the method introduce by Villa and Walker (2015a) does not require, in principle, an objective choice as long as the priors are proper. Given that we use the latter approach, here we consider subjective priors for the model specific parameters.

Remark 1. In the case where the changes in the underlying sampling distribution are limited to the parameter values, the model prior probabilities defined in (15) follow the uniform distribution. That is, $\Pr(M_j) \propto 1$. In the real data example illustrated in Section 5.1, we indeed consider a problem where the above case occurs.

Remark 2. As we assign a prior which depends on the number of change points, a legitimate question is how the dilution problem may affect our method, see George (2010). We would like to point out that the prior introduced in this paper implicitly takes into account the numerosity of models with the same number of change points. Indeed, the methodology used in this work builds on Villa and Walker (2015a). In particular, the approach requires to assume a prior on the change point locations and, as highlighted above, the default choice in our methodology is the uniform, which takes into account for the dilution.

3.1 A special case: selection between M_0 and M_1

Let us consider the case where we have to estimate whether there is or not a change point in a set of observations. This implies that we have to choose between model M_0 (i.e. no change point) and M_1 (i.e. one change point). Following our approach, we have:

$$\Pr(M_0) \propto \exp\left\{\mathbb{E}_{\pi_0}\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right]\right\}, \tag{16}$$

and

$$\Pr(M_1) \propto \exp\left\{\mathbb{E}_{\pi_1}\left[(n-m_1)\cdot\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}. \tag{17}$$

Now, let us assume independence between the prior on the change point location and the prior on the parameters of the underlying sampling distributions, that is $\pi_1(m_1, \tilde{\theta}_1, \tilde{\theta}_2) = \pi_1(m_1)\pi_1(\tilde{\theta}_1, \tilde{\theta}_2)$. Let us further recall that, as per equation (14), $\pi_1(m_1) = 1/(n-1)$. As such, we observe that the model prior probability on M_1 becomes:

$$\Pr(M_1) \propto \exp\left\{\left(\frac{n}{2}\right)\mathbb{E}_{\pi_1(\tilde{\theta}_1,\tilde{\theta}_2)}\left[\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}. \tag{18}$$

We notice that the model prior probability for model M_1 is increasing when the sample size increases. This behaviour occurs whether there is or not a change point in the data. We propose to address the above problem by using a non-uniform prior for m_1. A reasonable alternative, which works quite well in practice, would be the following shifted binomial as prior:

$$\pi_1(m_1) = \binom{n-2}{m_1-1}\left(\frac{n-1}{n}\right)^{m_1-1}\left(\frac{1}{n}\right)^{n-m_1-1}, 1 \leq m_1 \leq n-1. \tag{19}$$

To argument the choice of (19), we note that, as n increases, the probability mass will be more and more concentrated towards the upper end of the support. Therefore, from equations (17) and (19) follows:

$$\Pr(M_1) \propto \exp\left\{\left(\frac{2n-2}{n}\right)\mathbb{E}_{\pi_1(\tilde{\theta}_1,\tilde{\theta}_2)}\left[\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}. \tag{20}$$

For the more general case where we consider more than two models, the problem highlighted in equation (18) vanishes.

4 Change Point Analysis on Simulated Data

In this section, we present the results of several simulation studies based on the methodologies discussed in Sections 2 and 3. We start with a scenario involving discrete distributions in the context of the one change point problem. We then show the results obtained when we consider continuous distributions for the case of two change points. The choice of the underlying sampling distributions is in line with Villa and Walker (2015a).

4.1 Single sample

Scenario 1. The first scenario concerns the choice between models M_0 and M_1. Specifically, for M_0 we have:

$$X_1, X_2, \ldots, X_n | p \overset{\text{i.i.d.}}{\sim} \text{Geometric}(p),$$

and for M_1 we have:

$$\begin{aligned} X_1, X_2, \ldots, X_{m_1} | p &\overset{\text{i.i.d.}}{\sim} \text{Geometric}(p) \\ X_{m_1+1}, X_{m_1+2}, \ldots, X_n | \lambda &\overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda). \end{aligned}$$

Let us denote with $f_1(\cdot|p)$ and $f_2(\cdot|\lambda)$ the probability mass functions of the Geometric and the Poisson distributions, respectively. The priors for the parameters of f_1 and f_2 are $p \sim \text{Beta}(a, b)$ and $\lambda \sim \text{Gamma}(c, d)$.

In the first simulation, we sample $n = 100$ observations from model M_0 with $p = 0.8$. To perform the change point analysis, we have chosen the following parameters for the priors on p and λ: $a = 2$, $b = 2$, $c = 3$ and $d = 1$. Applying the approach introduced in Section 3, we obtain $\Pr(M_0) \propto 1.59$ and $\Pr(M_1) \propto 1.81$. These model priors yield the posterior distribution probabilities (refer to equation (4)) $\Pr(M_0|\mathbf{x}^{(n)}) = 0.92$ and $\Pr(M_1|\mathbf{x}^{(n)}) = 0.08$. As expected, the selection process strongly indicates the true model as M_0. Table 1 reports the above probabilities including other information, such as the appropriate Bayes factors.

The second simulation looked at the opposite set up, that is we sample $n = 100$ observations from M_1, with $p = 0.8$ and $\lambda = 3$. We have sampled 50 data points from the Geometric distribution and the remaining 50 data points from the Poisson distribution. In Figure 2, we have plotted the simulated sample, where it is legitimate to assume a change in the underlying distribution.

Using the same prior parameters as above, we obtain $\Pr(M_0|\mathbf{x}^{(n)}) = 0.06$ and $\Pr(M_1|\mathbf{x}^{(n)}) = 0.94$. Again, the model selection process is assigning heavy posterior mass to the true model M_1. These results are further detailed in Table 1.

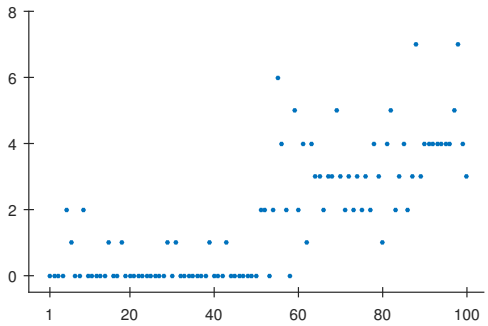

Figure 2: Scatter plot of the data simulated from model M_1 in Scenario 1.

	True model	
	M_0	M_1
$\Pr(M_0)$	0.47	0.47
$\Pr(M_1)$	0.53	0.53
B_{01}	12.39	0.08
B_{10}	0.08	12.80
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.92	0.06
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.08	0.94

Table 1: Model prior, Bayes factor and model posterior probabilities for the change point analysis in Scenario 1. We considered samples from, respectively, model M_0 and model M_1.

Scenario 2. In this scenario we consider the case where we have to select among three models, that is model M_0:

$$X_1, X_2, \ldots, X_n|\lambda, \kappa \overset{\text{i.i.d.}}{\sim} \text{Weibull}(\lambda, \kappa), \tag{21}$$

model M_1:

$$\begin{aligned} X_1, X_2, \ldots, X_{m_1}|\lambda, \kappa &\overset{\text{i.i.d.}}{\sim} \text{Weibull}(\lambda, \kappa) \\ X_{m_1+1}, X_{m_1+2}, \ldots, X_n|\mu, \tau &\overset{\text{i.i.d.}}{\sim} \text{Log-normal}(\mu, \tau), \end{aligned} \tag{22}$$

with $1 \leq m_1 \leq n-1$ being the location of the single change point, and model M_2:

$$\begin{aligned} X_1, X_2, \ldots, X_{m_1} | \lambda, \kappa &\overset{\text{i.i.d.}}{\sim} \text{Weibull}(\lambda, \kappa) \\ X_{m_1+1}, X_{m_1+2}, \ldots, X_{m_2} | \mu, \tau &\overset{\text{i.i.d.}}{\sim} \text{Log-normal}(\mu, \tau) \\ X_{m_2+1}, X_{m_2+2}, \ldots, X_n | \alpha, \beta &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta), \end{aligned} \tag{23}$$

with $1 \leq m_1 < m_2 \leq n-1$ representing the locations of the two change points, such that m_1 corresponds exactly to the same location as in model M_1. Analogously to the previous scenario, we sample from each model in turn and perform the selection to detect the number of change points.

Let $f_1(\cdot|\lambda, \kappa)$, $f_2(\cdot|\mu, \tau)$ and $f_3(\cdot|\alpha, \beta)$ represent the Weibull, Log-normal and Gamma densities, respectively, with $\tilde{\theta}_1 = (\lambda, \kappa)$, $\tilde{\theta}_2 = (\mu, \tau)$ and $\tilde{\theta}_3 = (\alpha, \beta)$. We assume a Normal prior on μ and Gamma priors on all the other parameters as follows:

$$\begin{aligned} &\lambda \sim \text{Gamma}(1.5, 1) \qquad \kappa \sim \text{Gamma}(5, 1) \qquad \mu \sim \text{Normal}(0.05, 1), \\ &\tau \sim \text{Gamma}(16, 1) \qquad \alpha \sim \text{Gamma}(10, 1) \qquad \beta \sim \text{Gamma}(0.2, 0.1). \end{aligned}$$

In the first exercise, we have simulated $n = 100$ observations from model M_0, where we have set $\lambda = 1.5$ and $\kappa = 5$. We obtain the following model priors: $\Pr(M_0) \propto 1.09$, $\Pr(M_1) \propto 1.60$ and $\Pr(M_2) \propto 1.37$, yielding the posteriors $\Pr(M_0|\mathbf{x}^{(n)}) = 0.96$, $\Pr(M_1|\mathbf{x}^{(n)}) = 0.04$ and $\Pr(M_2|\mathbf{x}^{(n)}) = 0.00$. We then see that the approach assigns high mass to the true model M_0. Table 2 reports the above probabilities and the corresponding Bayes factors. The second simulation was performed by sampling 50 observations from a Weibull with parameter values as in the previous exercise, and the remaining 50 observations from a Log-normal density with location parameter $\mu = 0.05$ and scale parameter $\tau = 16$. The data is displayed in Figure 3.

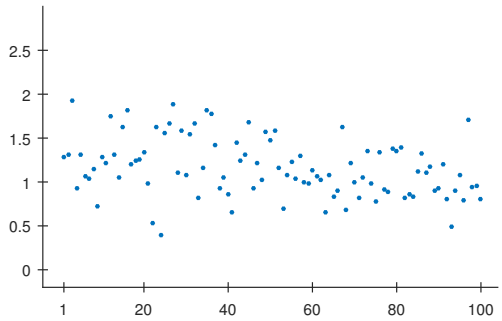

Figure 3: Scatter plot of the observations simulated from model M_1 in Scenario 2.

	True model		
	M_0	M_1	M_2
$\Pr(M_0)$	0.27	0.27	0.27
$\Pr(M_1)$	0.39	0.39	0.39
$\Pr(M_2)$	0.34	0.34	0.34
B_{01}	36.55	3.24×10^{-4}	4.65×10^{-40}
B_{02}	1.84×10^{3}	0.02	1.27×10^{-45}
B_{12}	50.44	55	2.72×10^{-6}
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.96	0.00	0.00
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.04	0.98	0.00
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.00	0.02	1.00

Table 2: Model prior, Bayes factor and model posterior probabilities for the change point analysis in Scenario 2. We considered samples from, respectively, model M_0, model M_1 and model M_2.

The model posterior probabilities are $\Pr(M_0|\mathbf{x}^{(n)}) = 0.00$, $\Pr(M_1|\mathbf{x}^{(n)}) = 0.98$ and $\Pr(M_2|\mathbf{x}^{(n)}) = 0.02$, which are reported in Table 2. In this case as well, we see that the model selection procedure indicates M_1 as the true model, as expected.

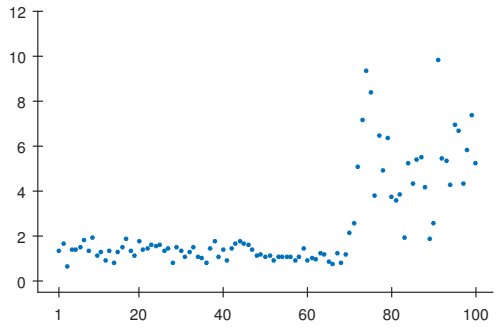

Figure 4: Scatter plot of the observations simulated from model M_2 in Scenario 2.

Finally, for the third simulation exercise we sample 50 and 20 data points from, respectively, a Weibull and a Log-normal with parameter values as defined above, and the last 30 observations are sampled from a Gamma distribution with parameters $\alpha = 10$ and $\beta = 2$. From Table 2, we note that the posterior distribution on the model space accumulates on the true model M_2.

4.2 Frequentist Analysis

In this section, we perform a frequentist analysis of the performance of the proposed prior by drawing repeated samples from different scenarios. In particular, we look at a two change points problem where the sampling distributions are Student-t with different degrees of freedom. In this scenario, we perform the analysis with 60 repeated samples generated by different densities with the same mean values.

Then, we repeat the analysis of Scenario 2 by selecting 100 samples for $n = 500$ and $n = 1500$. We consider different sampling distributions with the same mean and variance. In this scenario, where we added the further constraint of the equal variance, it is interesting to note that the change in distribution is captured when we increase the sample size, meaning that we learn more about the true sampling distributions.

We also compare the performances of the loss-based prior with the uniform prior when we analyse the scenario with different sampling distributions. Namely, Weibull/Log-normal/Gamma. It is interesting to note that the uniform prior is unable to capture the change in distribution even for a large sample size. On the contrary, the loss-based prior is able to detect the number of change points when $n = 1500$. Furthermore, for $n = 500$, even though both priors are not able to detect the change points most of the times, the loss-based prior has a higher frequency of success when compared to the uniform prior.

Scenario 3. In this scenario, we consider the case where the sampling distributions belong to the same family, that is Student-t, where the true model has two change points. In particular, let $f_1(\cdot|\nu_1)$, $f_2(\cdot|\nu_2)$ and $f_3(\cdot|\nu_3)$ represent the densities of three standard t distributions, respectively. We assume that ν_1, ν_2 and ν_3 are positive integers strictly greater than one so to have defined mean for each density. Note that this allows us to compare distributions of the same family with equal mean. The priors assigned to the number of degrees of freedom assume a parameter space of positive integers strictly larger than 1. As such, we define them as follows:

$$\nu_1 \sim 2 + \text{Poisson}(30) \qquad \nu_2 \sim 2 + \text{Poisson}(3) \qquad \nu_3 \sim 2 + \text{Poisson}(8).$$

In this experiment, we consider 60 repeated samples, each of size $n = 300$ and with the following structure:

- $X_1, \ldots, X_{100}$ from a Student-t distribution with $\nu_1 = 30$,

- $X_{101}, \ldots, X_{200}$ from a Student-t distribution with $\nu_2 = 3$,
- $X_{201}, \ldots, X_{300}$ from a Student-t distribution with $\nu_3 = 8$.

Table 3 reports the frequentist results of the simulation study. First, note that $P(M_1) = P(M_2) = P(M_3) = 1/3$ as per the Remark in Section 3. For all the simulated samples, the loss-based prior yields a posterior with the highest probability assigned to the true model M_2. We also note that the above posterior is on average 0.75 with a variance 0.02, making the inferential procedure extremely accurate.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.01	3.84×10^{-4}	0/60
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.24	0.0160	0/60
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.75	0.0190	60/60

Table 3: Average model posterior probabilities, variance and frequency of true model for the Scenario 3 simulation exercise.

Scenario 4. In this scenario, we perform repeated sampling from the setup described in scenario 2 above, where the true model has two change points. In particular, we draw 100 samples with $n = 500$ and $n = 1500$. For $n = 500$, the loss-based prior probabilities are $P(M_0) = 0.18$, $P(M_1) = 0.16$ and $P(M_2) = 0.66$. For $n = 1500$, the loss-based prior probabilities are $P(M_0) = 0.015$, $P(M_1) = 0.014$ and $P(M_2) = 0.971$. The simulation results are reported, respectively, in Table 4 and in Table 5. The two change point locations for $n = 500$ are at the 171st and 341st observations. For $n = 1500$, the first change point is the 501st observation, while the second is at the 1001st observation. We note that there is a sensible improvement in detecting the true model, using the loss-based prior, when the sample size increases. In particular, we move from 30% to 96%.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	9.88×10^{-4}	2.60×10^{-5}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.63	0.0749	70/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.37	0.0745	30/100

Table 4: Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 500$ and the loss-based prior.

To compare the loss-based prior with the uniform prior we have run the

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	1.33×10^{-13}	1.76×10^{-24}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.08	0.0200	4/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.92	0.0200	96/100

Table 5: Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 1500$ and the loss-based prior.

simulation on the same data samples used above. The results for $n = 500$ and $n = 1500$ are in Table 6 and in Table 7, respectively. Although we can observe an improvement when the sample size increases, the uniform prior does not lead to a clear detection of the true model for both sample sizes.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	16×10^{-4}	7.15×10^{-5}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.82	0.0447	91/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.18	0.0443	9/100

Table 6: Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 500$ and the uniform prior.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	8.64×10^{-12}	7.45×10^{-21}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.501	0.1356	49/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.499	0.1356	51/100

Table 7: Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 1500$ and the uniform prior.

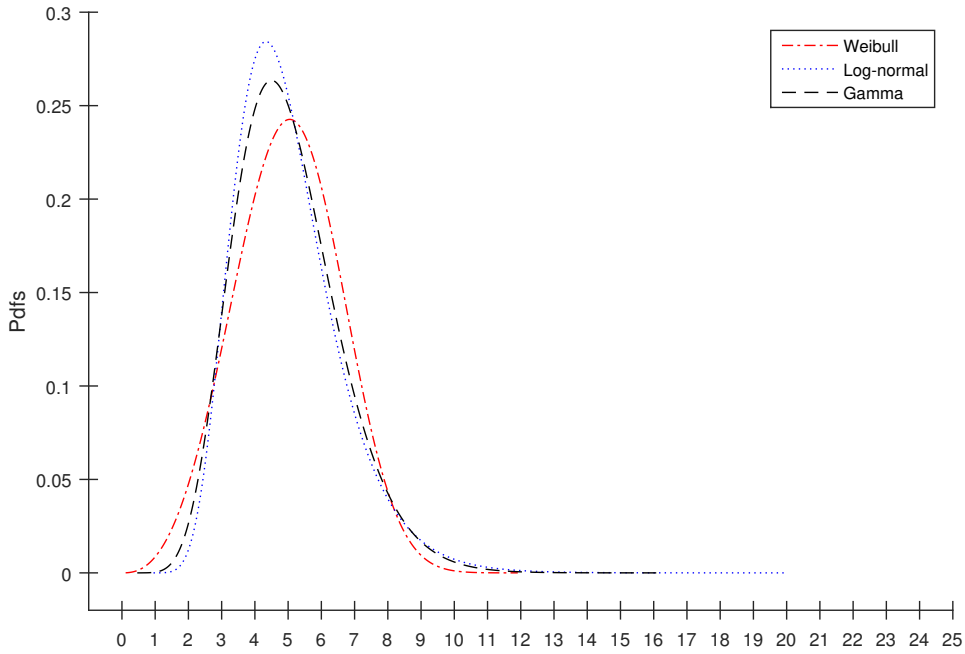

Figure 5: The densities of Weibull(λ, κ), Log-normal(μ, τ) and Gamma(α, β) with the same mean (equal to 5) and the same variance (equal to 2.5).

Finally, we conclude this section with a remark. One may wonder why the change point detection requires an increasing in the sample size, and the reply can be inferred from Figure 5, which displays the density functions of the distributions employed in this scenario. As it can be observed, the densities are quite similar, which is not surprising since these distributions have the same means and the same variances. The above similarity can also be appreciated in terms of Hellinger distance, see Table 8. In other words, from Figure 5 we can see that the main differences in the underlying distributions are in the tail areas. It is therefore necessary to have a relatively large number of observations in order to be able to discern differences in the densities, because in this case only we would have a sufficient representation of the whole distribution.

	Hellinger distances		
	Weibull(λ, κ)	Log-normal(μ, τ)	Gamma(α, β)
Weibull(λ, κ)		0.1411996	0.09718282
Log-normal(μ, τ)			0.04899711

Table 8: Hellinger distances between all the pairs formed from a Weibull(λ, κ), Log-normal(μ, τ) and Gamma(α, β). The six hyperparameters are such that the distributions have the same mean=5 and same variance=2.5.

4.3 Comparison to Barry and Hartigan's method

In this section we perform a comparison of the proposed change point method to the one described in Barry and Hartigan (1993). The simulation study is performed by considering three different scenarios: we simulate data, assumed to be normally distributed, and which exhibits, respectively, one, two and three change points.

Barry and Hartigan (1993) proposal is based on a product partition approach. In particular, product models on partitions represent a framework for Bayesian inference on change points. The authors highlight that, even if the initial probability model for partitions and parameters is not a product model, under specific conditions it represents a suitable approximation for the analysis.

To make the results comparable, we assume normality as the Barry and Hartigan (1993) method is based on this assumption. As described in detail below, we consider for each scenario normal distributions with variance 1 (assumed as known) and differences in the mean (at each change point) of, respectively, 1, 2.5 and 3. In addition, in each scenario we consider as a possible model the no-change point model. The prior distribution for the means is a normal with zero mean and large variance (i.e. 10^6).

To perform the simulations, for the Barry and Hartigan (1993) method, we employ the R package `bcp`, developed by Erdman and Emerson (2007), and assume a change point when the posterior probability is at least 0.5. All simulations have a burnin of 10000, with a total number of samples of 100000.

One change point. We consider the following model for the case with one change point:

$$X_1, X_2, \ldots, X_{m_1} | \mu_{11} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{11}, 1)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_n | \mu_{21} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{21}, 1)$$

Model M_0 corresponds to no changes in the mean of the data. We set $\mu_{21} = \mu_{11} + \Delta_1$ with $\Delta_1 \in \{0, 1, 2.5, 3\}$ and $\mu_{11} = 0$. In Table 9, we see the frequency of identifying the true model amongst 100 repeated samples for different sampling scenarios.

n	Frequency of identifying the true model							
	$\Delta_1 = 0$ (M_0)		$\Delta_1 = 1$ (M_1)		$\Delta_1 = 2.5$ (M_1)		$\Delta_1 = 3$ (M_1)	
	Our method	bcp	Our method	bcp	Our method	bcp	Our method	bcp
100	100/100	95/100	43/100	17/100	100/100	86/100	100/100	94/100
250	100/100	99/100	100/100	7/100	100/100	86/100	100/100	99/100
500	100/100	100/100	100/100	7/100	100/100	91/100	100/100	98/100

Table 9: Frequency of identifying the true model (the one within the nearby parentheses to the Δ_1 values) amongst 100 repeated samples for different sampling scenarios. The change point location is in $m_1 = 70, 175, 350$ for, respectively, $n = 100, 250, 500$.

Two change points We consider the following model for the case with two change points:

$$X_1, X_2, \ldots, X_{m_1} | \mu_{12} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{12}, 1)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_{m_2} | \mu_{22} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{22}, 1)$$
$$X_{m_2+1}, X_{m_2+2}, \ldots, X_n | \mu_{32} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{32}, 1)$$

As before, model M_0 corresponds to no changes in the mean. In the simulations we set $\mu_{22} = \mu_{12} + \Delta_2$ and $\mu_{32} = \mu_{12}$ with $\Delta_2 \in \{0, 1, 2.5, 3\}$ and $\mu_{12} = 0$. In Table 10, we see the frequency of identifying the true model amongst 100 repeated samples for different sampling scenarios.

n	Frequency of identifying the true model							
	$\Delta_2 = 0$ (M_0)		$\Delta_2 = 1$ (M_2)		$\Delta_2 = 2.5$ (M_2)		$\Delta_2 = 3$ (M_2)	
	Our method	bcp	Our method	bcp	Our method	bcp	Our method	bcp
100	100/100	96/100	3/100	9/100	100/100	67/100	100/100	89/100
250	100/100	100/100	86/100	5/100	100/100	78/100	100/100	90/100
500	100/100	98/100	100/100	1/100	100/100	76/100	100/100	90/100

Table 10: Frequency of identifying the true model (the one within the nearby parentheses to the Δ_2 values) amongst 100 repeated samples for different sampling scenarios. The location of the first change point is $m_1 = 30, 75, 150$, respectively, for $n = 100, 250, 500$, and for the second change point is $m_2 = 70, 175, 350$.

Three change points Finally, we consider the following model for the case with three change points:

$$\begin{aligned} X_1, X_2, \ldots, X_{m_1} | \mu_{13} &\overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{13}, 1) \\ X_{m_1+1}, X_{m_1+2}, \ldots, X_{m_2} | \mu_{23} &\overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{23}, 1) \\ X_{m_2+1}, X_{m_2+2}, \ldots, X_{m_3} | \mu_{33} &\overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{33}, 1) \\ X_{m_3+1}, X_{m_3+2}, \ldots, X_n | \mu_{43} &\overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{43}, 1) \end{aligned}$$

We set $\mu_{23} = \mu_{13} + \Delta_3$, $\mu_{33} = \mu_{13} + 2\Delta_3$ and $\mu_{43} = \mu_{13} + 3\Delta_3$ with $\Delta_3 \in \{0, 1, 2.5, 3\}$ and $\mu_{13} = 0$. In Table 11, we see the frequency of identifying the true model amongst 100 repeated samples for different sampling scenarios.

n	Frequency of identifying the true model							
	$\Delta_3 = 0$ (M_0)		$\Delta_3 = 1$ (M_3)		$\Delta_3 = 2.5$ (M_3)		$\Delta_3 = 3$ (M_3)	
	Our method	bcp	Our method	bcp	Our method	bcp	Our method	bcp
100	100/100	97/100	0/100	1/100	100/100	65/100	100/100	88/100
250	100/100	98/100	30/100	0/100	100/100	69/100	100/100	88/100
500	100/100	100/100	100/100	0/100	100/100	73/100	100/100	80/100

Table 11: Frequency of identifying the true model (the one within the nearby parentheses to the Δ_3 values) amongst 100 repeated samples for different sampling scenarios. The location of the first change point is $m_1 = 25, 62, 125$ for, respectively, $n = 100, 250, 500$; the location of the second change point is $m_2 = 50, 125, 250$ and the location of the third change point is $m_3 = 75, 188, 375$.

By looking at the above tables, we note the following. In general, both methods improve the detection of the change points as n increases, which is an expected result as the information about change points in the sample

increases. For the cases where $\Delta = 2.5, 3$, our method appears to have a better performance than the one in Barry and Hartigan (1993). This is more obvious for the smaller Δ. Furthermore, the proposed approach seems to select the model with the true number of change points when this number increases. A noteworthy aspect is that the Barry and Hartigan (1993) method diminishes its performance as n increases when the difference between the means is relatively small (i.e. $\Delta = 1$). A possible explanation is due to a degenerate behaviour of the product partition model; however, we did not investigate further as it does not impact the performance of our method.

5 Change Point Analysis on Real Data

In this section, we illustrate the proposed approach applied to real data. We first consider a well known data set which has been extensively studied in the literature of the change point analysis, that is the British coal-mining disaster data (Carlin et al., 1992). The second set of data we consider refers to the daily returns of the S&P 500 index observed over a period of four years. The former data set will be investigated in Section 5.1, while the latter in Section 5.2.

5.1 British Coal-Mining Disaster Data

The British coal-mining disaster data consists of the yearly number of deaths for the British coal miners over the period 1851-1962. It is believed that the change in the working conditions, and in particular, the enhancement of the security measures, led to a decrease in the number of deaths. This calls for a model which can take into account a change in the underlying distribution around a certain observed year. With the proposed methodology we wish to detect if the assumption is appropriate. In particular, if a model with one change point is more suitable to represent the data than a model where no changes in the sampling distribution are assumed. Figure 6 shows the number of deaths per year in the British coal-mining industry from 1851 to 1962. As in Chib (1998), we assume a Poisson sampling distribution with a possible change in the parameter value. That is

$$\begin{aligned} X_1, X_2, \ldots, X_m | \phi_1 &\overset{\text{i.i.d.}}{\sim} \text{Poisson}(\phi_1) \\ X_{m+1}, X_{m+2}, \ldots, X_n | \phi_2 &\overset{\text{i.i.d.}}{\sim} \text{Poisson}(\phi_2), \end{aligned} \tag{24}$$

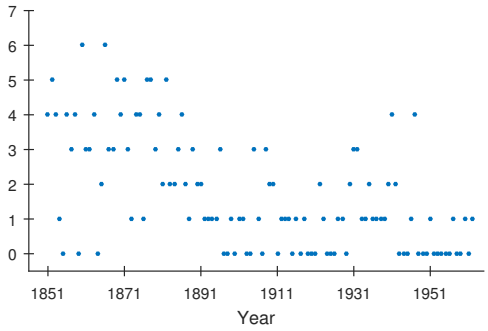

Figure 6: Scatter plot of the British coal-mining disaster data.

where m is the unknown location of the single change point, such that $1 \leq m \leq n$, and a Gamma$(2, 1)$ is assumed for ϕ_1 and ϕ_2. The case $m = n$ corresponds to the scenario with no change point, that is model M_0. The case $m < n$ assumes one change point, that is model M_1.

Let $f_1(\cdot|\phi_1)$ and $f_2(\cdot|\phi_2)$ be the Poisson distributions with parameters ϕ_1 and ϕ_2, respectively. Then, the analysis is performed by selecting between model M_0, that is when the sampling distribution is f_1, and model M_1, where the sampling distribution is f_1 up to a certain $m < n$ and f_2 from $m + 1$ to n.

As highlighted in the Remark at the end of Section 3, the prior on the model space is the discrete uniform distribution, that is $\Pr(M_0) = \Pr(M_1) = 0.5$. The proposed model selection approach leads to the Bayes factors $B_{01} = 1.61 \times 10^{-13}$ and $B_{10} = 6.20 \times 10^{12}$, where it is obvious that the odds are strongly in favour of model M_1. Indeed, we have $\Pr(M_1|\mathbf{x}^{(n)}) \approx 1$.

5.2 Daily S&P 500 Absolute Log-Return Data

The second real data analysis aims to detect change points in the absolute value of the daily logarithmic returns of the S&P500 index observed from the 14/01/2008 to the 31/12/2011 (see Figure 7). As underlying sampling distributions we consider the Weibull and the Log-normal (Yu, 2001), and the models among which we select are as follows. M_0 is a Weibull(λ, κ), M_1 is formed by a Weibull(λ, κ) and a Log-normal(μ_1, τ_1) and, finally, M_2 is formed by a Weibull(λ, κ), a Log-normal(μ_1, τ_1) and a Log-normal(μ_2, τ_2). An interesting particularity of this problem is that we will consider a scenario where the changes are in the underlying distribution or in the parameter values of the same distribution. As suggested in Section 4.1.3 of Kass and Raftery (1995), due to the large sample size of the data set, we could approximate

the Bayes factor by using the Schwartz criterion. Therefore, in this case the specification of the priors for the parameters of the underlying distributions is not necessary. From the results in Table 12, we see that the model indicated

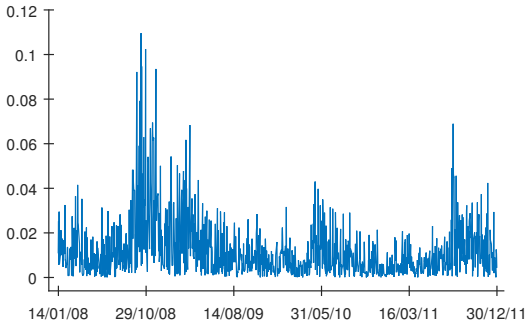

Figure 7: Absolute daily log-returns of the S&P500 index from 14/01/08 to 30/12/11.

by the proposed approach is M_2. In other words, there is very strong indication that there are two change points in the data set. From Table 12, we note

$\Pr(M_0)$	0.36
$\Pr(M_1)$	0.32
$\Pr(M_2)$	0.32
B_{01}	7.72×10^{18}
B_{02}	3.30×10^{-3}
B_{12}	4.28×10^{-22}
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.00
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.00
$\Pr(M_2\|\mathbf{x}^{(n)})$	1.00

Table 12: Model prior, Bayes factor and model posterior probabilities for the S&P500 change point analysis.

that the prior on model M_1 and M_2 assigned by the proposed method are the same. This is not surprising as the only difference between the two models is an additional Log-normal distribution with different parameter values.

6 Conclusion

Bayesian inference in change point problems under the assumption of minimal prior information has not been deeply explored in the past, as the limited literature on the matter shows.

We contribute to the area by deriving an objective prior distribution to detect change point locations, when the number of change points is known a priori. As a change point location can be interpreted as a discrete parameter, we apply recent results in the literature (Villa and Walker, 2015b) to make inference. The resulting prior distribution, which is the discrete uniform distribution, it is not new in the literature (Girón et al., 2007), and therefore can be considered as a validation of the proposed approach.

A second major contribution is in defining an objective prior on the number of change points, which has been approached by considering the problem as a model selection exercise. The results of the proposed method on both simulated and real data, show the strength of the approach in estimating the number of change points in a series of observations. A point to note is the generality of the scenarios considered. Indeed, we consider situations where the change is in the value of the parameter(s) of the underlying sampling distribution, or in the distribution itself. For the simulation study we have compared the proposed method with an existing Bayesian approach for detection of change points discussed in Barry and Hartigan (1993). Of particular interest is the last real data analysis (S&P 500 index), where we consider a scenario where we have both types of changes, that is the distribution for the first change point and on the parameters of the distribution for the second.

The aim of this work was to set up a novel approach to address change point problems. In particular, we have selected prior densities for the parameters of the models to reflect a scenario of equal knowledge, in the sense that model priors are close to represent a uniform distribution. Two remarks are necessary here. First, in the case prior information about the true value of the parameters is available, and one wishes to exploit it, the prior densities will need to reflect it and, obviously, the model prior will be impacted by the choice. Second, in applications it is recommended that some sensitivity analysis is performed, so to investigate if and how the choice of the parameter densities affects the selection process.

Acknowledgements

We thank the Associate Editor and the reviewers for the thoughtful suggestions which have significantly improved the paper. Fabrizio Leisen was supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no: 630677. Cristiano Villa was supported by the Royal Society Research Grant no: RG150786.

References

Barry, D., Hartigan, J. A., 1992. Product Partition Models for Change Point Problems. The Annals of Statistics 20 (1), 260–279.

Barry, D., Hartigan, J. A., 1993. A Bayesian Analysis for Change Point Problems. Journal of the American Statistical Association 88 (421), 309–319.

Berger, J. O., Bernardo, J. M., Sun, D., 2012. Objective Priors for Discrete Parameter Spaces. Journal of the American Statistical Association 107 (498), 636–648.

Berger, J. O., Pericchi, L. R., 1996. The Intrinsic Bayes Factor for Model Selection and Prediction. Journal of the American Statistical Association 91 (433), 109–122.

Berk, R. H., 1966. Limiting Behavior of Posterior Distributions when the Model is Incorrect. The Annals of Mathematical Statistics 37 (1), 51–58.

Carlin, B. P., Gelfand, A. E., Smith, A. F. M., 1992. Hierarchical Bayesian Analysis of Changepoint Problems. Journal of the Royal Statistical Society. Series C (Applied Statistics) 41 (2), 389–405.

Chernoff, H., Zacks, S., 1964. Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time. The Annals of Mathematical Statistics 35 (3), 999–1018.

Chib, S., 1998. Estimation and Comparison of Multiple Change-point Models. Journal of Econometrics 86 (2), 221–241.

Erdman, C., Emerson, J., 2007. bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. Journal of Statistical Software 23 (3), 1–13.

Fearnhead, P., Liu, Z., 2007. On-line Inference for Multiple Changepoint Problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (4), 589–605.

George, E. I., 2010. Dilution Priors: Compensating for Model Space Redundancy. Vol. 6 of Collections. Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp. 158–165.

Girón, F. J., Moreno, E., Casella, G., 2007. Objective Bayesian Analysis of Multiple Changepoints for Linear Models (with discussion). In: Bernardo, J.M.., Bayarri, M.J.., Berger, J.O.., Dawid, A.P.., Heckerman, D., Smith, A.F.M.., West, M. (Eds.), Bayesian Statistics 8. Oxford University Press, London, pp. 227–252.

Hannart, A., Naveau, P., 2009. Bayesian Multiple Change Points and Segmentation: Application to Homogenization of Climatic Series. Water Resources Research 45 (10), W10444.

Harlé, F., Chatelain, F., Gouy-Pailler, C., Achard, S., 2016. Bayesian Model for Multiple Change-Points Detection in Multivariate Time Series. IEEE Transactions on Signal Processing 64 (16), 4351–4362.

Heard, N. A., Turcotte, M. J. M., 2017. Adaptive Sequential Monte Carlo for Multiple Changepoint Analysis. Journal of Computational and Graphical Statistics 26 (2), 414–423.

Henderson, R., Matthews, J. N. S., 1993. An Investigation of Changepoints in the Annual Number of Cases of Haemolytic Uraemic Syndrome. Journal of the Royal Statistical Society. Series C (Applied Statistics) 42 (3), 461–471.

Kass, R. E., Raftery, A. E., 1995. Bayes Factors. Journal of the American Statistical Association 90 (430), 773–795.

Ko, S. I. M., Chong, T. T. L., Ghosh, P., 2015. Dirichlet Process Hidden Markov Multiple Change-point Model. Bayesian Analysis 10 (2), 275–296.

Koop, G., Potter, S. M., 2009. Prior Elicitation in Multiple Change-point Models. International Economic Review 50 (3), 751–772.

Kullback, S., Leibler, R. A., 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22 (1), 79–86.

Lai, T. L., Xing, H., 2011. A Simple Bayesian Approach to Multiple Change-Points. Statistica Sinica 21 (2), 539–569.

Loschi, R.H., Cruz, F.R.B., 2005. Extension to the Product Partition Model: Computing the Probability of a Change. Computational Statistics & Data Analysis 48 (2), 255 – 268.

Martínez, A. F., Mena, R. H., 2014. On a Nonparametric Change Point Detection Model in Markovian Regimes. Bayesian Analysis 9 (4), 823–858.

Mira, A., Petrone, S., 1995. Bayesian hierarchical nonparametric inference for change-point problems.

Moreno, E., Casella, G., Garcia-Ferrer, A., 2005. An Objective Bayesian Analysis of the Change Point Problem. Stochastic Environmental Research and Risk Assessment 19 (3), 191–204.

Muliere, P., Scarsini, M., 1985. Change-point Problems: A Bayesian Nonparametric Approach. Aplikace matematiky 30 (6), 397–402.

Petrone, S., Raftery, A. E., 1997. A Note on the Dirichlet Process Prior in Bayesian Nonparametric Inference with Partial Exchangeability. Statistics & Probability Letters 36 (1), 69 – 83.

Raftery, A. E., Akman, V. E., 1986. Bayesian Analysis of a Poisson Process with a Change-point. Biometrika 73 (1), 85–89.

Schwaller, L., Robin, S., 2017. Exact Bayesian Inference for Off-line Change-point Detection in Tree-structured Graphical Models. Statistics and Computing 27 (5), 1331–1345.

Smith, A. F. M., 1975. A Bayesian Approach to Inference about a Change-point in a Sequence of Random Variables. Biometrika 62 (2), 407–416.

Stephens, D. A., 1994. Bayesian Retrospective Multiple-Changepoint Identification. Journal of the Royal Statistical Society. Series C (Applied Statistics) 43 (1), 159–178.

Tian, G.-L., Ng, K. W., Li, K.-C., Tan, M., 2009. Non-iterative Sampling-based Bayesian Methods for Identifying Changepoints in the Sequence of Cases of Haemolytic Uraemic Syndrome. Computational Statistics & Data Analysis 53 (9), 3314–3323.

Villa, C., Walker, S., 2015a. An Objective Bayesian Criterion to Determine Model Prior Probabilities. Scandinavian Journal of Statistics 42 (4), 947–966.

Villa, C., Walker, S. G., 2015b. An Objective Approach to Prior Mass Functions for Discrete Parameter Spaces. Journal of the American Statistical Association 110 (511), 1072–1082.

Yu, J., 2001. Chapter 6 - Testing for a Finite Variance in Stock Return Distributions. In: Knight, J., Satchell, S. E. (Eds.), Return Distributions in Finance (Quantitative Finance). Butterworth-Heinemann, Oxford, pp. 143 – 164.

Appendix

A Model prior probabilities to select among models M_0, M_1 and M_2

Here, we show how model prior probabilities can be derived for the relatively simple case of selecting among scenarios with no change points (M_0), one change point (M_1) or two change points (M_2). First, by applying the result in Theorem 2, we derive the Kullback–Leibler divergences between any two models. That is:

- the prior probability for model M_0 depends on the following quantities:

$$\begin{aligned} D_{KL}(M_0\|M_1) =& (n-m_1)\cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) \\ D_{KL}(M_0\|M_2) =& (m_2-m_1)\cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) \\ &+ (n-m_2)\cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3)) \end{aligned}$$

- the prior probability for model M_1 depends on the following quantities:

$$\begin{aligned} D_{KL}(M_1\|M_2) =& (n-m_2)\cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_3(\cdot|\tilde{\theta}_3)) \\ D_{KL}(M_1\|M_0) =& (n-m_1)\cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)) \end{aligned}$$

- the prior probability for model M_2 depends on the following quantities:

$$\begin{aligned} D_{KL}(M_2\|M_1) =& (n-m_2)\cdot D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_2(\cdot|\tilde{\theta}_2)) \\ D_{KL}(M_2\|M_0) =& (m_2-m_1)\cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)) \\ &+ (n-m_2)\cdot D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_1(\cdot|\tilde{\theta}_1)) \end{aligned}$$

The next step is to derive the minimum Kullback–Leibler divergence computed at each model:

- for model M_0:

$$\begin{aligned}
\inf_{\theta_1} D_{KL}(M_0\|M_1) &= \underbrace{\left[\inf_{m_1\neq n}(n-m_1)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right] \\
&= \inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) \\
\inf_{\theta_2} D_{KL}(M_0\|M_2) &= \underbrace{\left[\inf_{m_1\neq m_2}(m_2-m_1)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right] \\
&\quad + \underbrace{\left[\inf_{m_2\neq n}(n-m_2)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_3} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3))\right] \\
&= \inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) + \inf_{\tilde{\theta}_3} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3))
\end{aligned}$$

- for model M_1:

$$\begin{aligned}
\inf_{\theta_2} D_{KL}(M_1\|M_2) &= \underbrace{\left[\inf_{m_2\neq n}(n-m_2)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_3} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_3(\cdot|\tilde{\theta}_3))\right] \\
&= \inf_{\tilde{\theta}_3} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_3(\cdot|\tilde{\theta}_3)) \\
\inf_{\theta_0=\tilde{\theta}_1} D_{KL}(M_1\|M_0) &= (n-m_1)\cdot\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))
\end{aligned}$$

- for model M_2:

$$\begin{aligned}
\inf_{\theta_1} D_{KL}(M_2\|M_1) &= (n-m_2)\cdot\inf_{\tilde{\theta}_2} D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_2(\cdot|\tilde{\theta}_2)) \\
\inf_{\theta_0=\tilde{\theta}_1} D_{KL}(M_2\|M_0) &= (m_2-m_1)\cdot\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)) \\
&\quad + (n-m_2)\cdot\inf_{\tilde{\theta}_1} D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_1(\cdot|\tilde{\theta}_1))
\end{aligned}$$

Therefore, the model prior probabilities can be computed through equation (15), so that:

- the model prior probability $\Pr(M_0)$ is proportional to the exponential of the minimum between:

$$\left\{\mathbb{E}_{\pi_0}\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right], \mathbb{E}_{\pi_0}\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) + \inf_{\tilde{\theta}_3} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3))\right]\right\}$$

- the model prior probability $\Pr(M_1)$ is proportional to the exponential of the minimum between:

$$\left\{ \mathbb{E}_{\pi_1} \left[\inf_{\tilde{\theta}_3} D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_3(\cdot|\tilde{\theta}_3)) \right], \right.$$
$$\left. \mathbb{E}_{\pi_1} \left[(n - m_1) \cdot \inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)) \right] \right\}$$

- the model prior probability $\Pr(M_2)$ is proportional to the exponential of the minimum between:

$$\left\{ \mathbb{E}_{\pi_2} \left[(n - m_2) \cdot \inf_{\tilde{\theta}_2} D_{KL}(f_3(\cdot|\tilde{\theta}_3) \| f_2(\cdot|\tilde{\theta}_2)) \right], \right.$$
$$\mathbb{E}_{\pi_2} \left[(m_2 - m_1) \cdot \inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)) + (n - m_2) \right.$$
$$\left. \left. \cdot \inf_{\tilde{\theta}_1} D_{KL}(f_3(\cdot|\tilde{\theta}_3) \| f_1(\cdot|\tilde{\theta}_1)) \right] \right\}$$

B Proofs

Proof of Theorem 1

We distinguish two cases: $S = +1$ and $S = -1$. When $S = +1$, equivalent to $m_j < m'_j$:

$$D_{KL}(f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \| f(\mathbf{x}^{(n)}|\boldsymbol{m}', \tilde{\boldsymbol{\theta}})) = \int f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \cdot \ln\left(\frac{f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}})}{f(\mathbf{x}^{(n)}|\boldsymbol{m}', \tilde{\boldsymbol{\theta}})}\right) \mathrm{d}\mathbf{x}^{(n)}$$

$$= \int f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \cdot \left[\sum_{i=m_j+1}^{m'_j} \ln\left(\frac{f_{j+1}(x_i|\tilde{\theta}_{j+1})}{f_j(x_i|\tilde{\theta}_j)}\right)\right] \mathrm{d}\mathbf{x}^{(n)}$$

$$= \sum_{i=m_j+1}^{m'_j} \int f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \cdot \left[\ln\left(\frac{f_{j+1}(x_i|\tilde{\theta}_{j+1})}{f_j(x_i|\tilde{\theta}_j)}\right)\right] \mathrm{d}\mathbf{x}^{(n)}$$

$$= \sum_{i=m_j+1}^{m'_j} \left\{1^{n-1} \cdot \int f_{j+1}(x_i|\tilde{\theta}_{j+1}) \cdot \left[\ln\left(\frac{f_{j+1}(x_i|\tilde{\theta}_{j+1})}{f_j(x_i|\tilde{\theta}_j)}\right)\right] \mathrm{d}x_i\right\}$$

$$= \sum_{i=m_j+1}^{m'_j} D_{KL}(f_{j+1}(x_i|\tilde{\theta}_{j+1}) \| f_j(x_i|\tilde{\theta}_j))$$

$$= (m'_j - m_j) \cdot D_{KL}(f_{j+1}(\cdot|\tilde{\theta}_{j+1}) \| f_j(\cdot|\tilde{\theta}_j))$$

$$= (m'_j - m_j) \cdot d_j^{+1}(\tilde{\boldsymbol{\theta}}). \tag{25}$$

When $S = -1$, equivalent to $m_j > m'_j$, in a similar fashion, we get

$$D_{KL}(f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \| f(\mathbf{x}^{(n)}|\boldsymbol{m}', \tilde{\boldsymbol{\theta}})) = (m_j - m'_j) \cdot d_j^{-1}(\tilde{\boldsymbol{\theta}}) \tag{26}$$

From equations (25) and (26), we get the result in Theorem 1.

Proof of Theorem 2

We recall that the model parameter θ_i is the vector $(m_1, m_2, \ldots, m_i, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{i+1})$, where $i = 0, 1, \ldots, k$. Here, $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{i+1}$ represent the parameters of the underlying sampling distributions considered under model M_i and $m_1, m_2, \ldots, m_i$ are the respective i change point locations. In this setting,

$$f(\mathbf{x}^{(n)}|\theta_i) = \prod_{r=1}^{m_1} f_1(x_r|\tilde{\theta}_1) \prod_{t=1}^{i-1} \prod_{r=m_t+1}^{m_{t+1}} f_{t+1}(x_r|\tilde{\theta}_{t+1}) \prod_{r=m_i+1}^{n} f_{i+1}(x_r|\tilde{\theta}_{i+1}) \tag{27}$$

We proceed to the computation of $D_{KL}(M_i\|M_j)$, that is the Kullback–Leibler divergence introduced in Section 3. Similarly to the proof of Theorem 1, we obtain the following result.

$$\begin{aligned} D_{KL}(M_i\|M_j) = &\sum_{r=m_{i+1}+1}^{m_{i+2}} \int f(\mathbf{x}^{(n)}|\theta_i) \ln\left(\frac{f_{i+1}(x_r|\tilde{\theta}_{i+1})}{f_{i+2}(x_r|\tilde{\theta}_{i+2})}\right) \mathrm{d}\mathbf{x}^{(n)} \\ &+ \sum_{r=m_{i+2}+1}^{m_{i+3}} \int f(\mathbf{x}^{(n)}|\theta_i) \ln\left(\frac{f_{i+1}(x_r|\tilde{\theta}_{i+1})}{f_{i+3}(x_r|\tilde{\theta}_{i+3})}\right) \mathrm{d}\mathbf{x}^{(n)} + \\ &\ldots + \sum_{r=m_j+1}^{n} \int f(\mathbf{x}^{(n)}|\theta_i) \ln\left(\frac{f_{i+1}(x_r|\tilde{\theta}_{i+1})}{f_{j+1}(x_r|\tilde{\theta}_{j+1})}\right) \mathrm{d}\mathbf{x}^{(n)}. \end{aligned}$$

Given equation (27), if we integrate out the variables not involved in the logarithms, we obtain

$$\begin{aligned} D_{KL}(M_i\|M_j) = &(m_{i+2} - m_{i+1}) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{i+2}(\cdot|\tilde{\theta}_{i+2})) \\ &+ (m_{i+3} - m_{i+2}) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{i+3}(\cdot|\tilde{\theta}_{i+3})) + \\ &\ldots + (n - m_j) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{j+1}(\cdot|\tilde{\theta}_{j+1})). \end{aligned}$$

In a similar fashion, it can be shown that

$$\begin{aligned} D_{KL}(M_j\|M_i) = &(m_{i+2} - m_{i+1}) \cdot D_{KL}(f_{i+2}(\cdot|\tilde{\theta}_{i+2})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1})) \\ &+ (m_{i+3} - m_{i+2}) \cdot D_{KL}(f_{i+3}(\cdot|\tilde{\theta}_{i+3})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1})) + \\ &\ldots + (n - m_j) \cdot D_{KL}(f_{j+1}(\cdot|\tilde{\theta}_{j+1})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1})) \end{aligned}$$