An Improved Bayesian Information Criterion for Multiple Change-Point Models

Alexis HANNART

Institut Franco-Argentin d'Etudes sur le Climat et ses Impacts
CNRS-CONICET-Universidad de Buenos Aires (UBA), Buenos Aires, Argentina
(*alexis.hannart@cima.fcen.uba.ar*)

Philippe NAVEAU

Laboratoire des Sciences du Climat et l'Environnement CNRS-CEA Gif-sur-Yvette, France
(*philippe.naveau@lsce.ipsl.fr*)

In multiple change-point analysis, inferring the number of change points is often achieved by minimizing a selection criterion that trades off data fidelity with complexity. We address the open problem of defining a selection criterion adapted to the context of multiple change-point analysis. Our approach is inspired by the Schwarz seminal formulation of the Bayesian information criterion (BIC): similarly, we introduce priors—here describing the occurrence of change points—and we use the Laplace approximation to derive a closed-form expression of the criterion. Differently from this previous work, we take advantage of the a priori information introduced, instead of asymptotically eliminating the dependence on priors. Results obtained on simulated series show a substantial gain in performance versus recent alternative criteria used in multiple change-point analysis. Results also show that the a priori information introduced in our criterion on the regularity of interevent times is the main driver of this substantial performance gain. Methods are motivated by and demonstrated on a meteorological application involving the homogenization of a temperature series.

KEY WORDS: Bayesian statistics; Change-point problem; Model selection

1. INTRODUCTION

The problem of detecting an unknown number of hidden, abrupt changes in a series, the so-called multiple change-point problem, has been extensively treated in a wide range of theoretical and applied contexts, for instance, in biology (Braun, Braun, and Muller 2000; Zhang 2007), in hydrology (Perreault et al. 2000b), and in finance (Chen and Gupta 1997; Dias and Embrechts 2004). In climatology and meteorology, the problem of detecting and correcting artificial shifts that affect long series of meteorological observations, termed "inhomogeneities," is also a field of application for change-point methods (Reeves et al. 2007, and Figure 1a). Inhomogeneities may be caused by changes in station location or environment, observing instrumentation, or practices and have the same size as the climatic variations to be expected. Hence, the homogenization of climatic series is recognized as an important step prior to the analysis of historical climate trends and variability, concerning in particular anthropogenic climate change (Abarca-Del-Rio and Mestre 2006; Kuglitsch et al. 2009). The interest of climate researchers in change-point modeling methods has therefore been raised, and the applicative context of homogenization has triggered the development of new change-point methods that may be applied to other fields (Caussinus and Mestre 2004; Hannart and Naveau 2009; Lu, Lund, and Lee 2009).

Comprehensive reviews of various approaches to the inference of multiple change points can be found, for instance, in Basseville and Nikiforov (1996), Chib (1998), and Chen and Gupta (2000). The determination of the number of change points k is often approached as a problem of model dimension choice, since k drives model dimension. Many criteria are available in a general context of model choice. They usually balance a term quantifying model fit with a term accounting for model complexity that increases with dimension and limits overfitting; otherwise, *the highest possible* k *would invariably be optimal*, to quote Schwarz (1978)—a seminal article that, with Akaike (1974), originally set the theoretical ground for model dimension choice. Since then, many works have addressed the fundamental question of studying and improving the performance of those criteria in different particular contexts of interest. In the context of the multiple change-point problem, Ninomiya (2005), for instance, proposed an adaptation of the Akaike information criterion (AIC) in a Gaussian setting. Several ad-hoc criteria for the change-point problem were also proposed, for instance, by Lavielle (2004) and Birgé and Massart (2001). On the other hand, the Schwarz criterion, often referred to as the Bayesian information criterion (BIC) because it is derived from an asymptotic expression of the Bayes factor, has been applied straightforwardly to the change-point problem, for instance, by Yao (1988) and Serbinowska (1996), and the asymptotic consistency of the resulting estimator of k has been established in the Gaussian and binomial cases, respectively. Nonetheless, while the BIC is well grounded for regular parametric models, it lacks desirable properties when applied to irregular statistical models—as defined generally, for instance, by Bickel et al. (1998, p. 12)—such as multiple change-point models. Adaptations of the Schwarz criterion, such as Zhang (2007)

TECHNOMETRICS, AUGUST 2012, VOL. 54, NO. 3
DOI: 10.1080/00401706.2012.694780

in the case of Gaussian independent and identically distributed (iid) sequences with change in the mean, are motivated by this issue.

Those adaptations of the BIC are improvements, but the assumptions grounding those criteria can be viewed as restrictive: it is often assumed, for instance, that series are independent Gaussian, and more importantly, that $n \to \infty$ with k fixed. However, in practice, in the change-point context, this particular asymptotic condition may not be verifiable since it is difficult to increase the length of an observation series without also increasing the number of change points. Rather, it may often be natural to assume that $\frac{n}{k}$ reaches a finite limit when $n \to \infty$. One can also view as a limitation the fact that some methods, for instance, Pan and Chen (2006), require the tuning of some adaptive parameters whose effect on the outcome may not always be clearly understood. Finally, in many applications such as climate series homogenization, previous information is available from past studies on the characteristics of the change points, such as their amplitude and their frequency of occurrence. One can regard as a disadvantage the fact that this information is not taken into account by those criteria to determine the number of change points.

This article presents an adaptation of the BIC to the change-point problem that attempts to circumvent those limitations while improving the performance of existing criteria. Our strategy is to formulate the problem by taking advantage of the Bayesian framework to introduce some a priori information on the relative positions and amplitude of change points. To obtain a closed-form mathematical expression for the resulting criterion, as in the original BIC approach of Schwarz (1978), we take advantage of the Laplace approximation. We also propose a few other approximations that greatly simplify implementation and avoid having to rely on computationally expensive Markov chain Monte Carlo (MCMC) methods.

The remainder of the article is organized as follows. In Section 2, we describe our general approach and introduce some modeling assumptions. In Section 3, inference issues are addressed via approximations and we derive a general expression of our criterion. In Section 4, performance is assessed through simulations generated from several change-point models. Section 5 presents an application of the method to the homogenization of temperature series at Marseille, France. Section 6 discusses some strengths and limitations of our approach and concludes.

2. A BAYESIAN MODEL FOR THE CHANGE-POINT PROBLEM

We propose the following general formulation. Denote by $\boldsymbol{x} = \{x_j \in \mathbb{R},\ j = 1, \ldots, n\}$ a multivariate time series of length n, and assume that some of its characteristics are changing at $k-1$ instants $t_1, t_2, \cdots, t_{k-1}$ such that $t_0 < t_1 < t_2 < \cdots < t_{k-1} < t_k$, with the convention $t_0 = 0$ and $t_k = n$. To reflect this situation, we assume that the distribution of $\boldsymbol{x}$ is defined piecewise on the resulting k segments by $p(x_{t_i+1}, \ldots, x_{t_{i+1}} \mid \xi_i)$, with ξ_i the d-dimensional parameter vector for the segment i such that $\xi_i \neq \xi_{i+1}$, for $i = 1, \ldots, k-1$. For instance, in the context of meteorological series homogenization introduced in Section 1 and detailed in Section 5, $\boldsymbol{x}$ is a series of yearly temperature difference between two neighboring meteorological stations and ξ is the mean of the series, that is, $d = 1$ (Figure 1a). Now, let $\boldsymbol{\delta}^{(k)} = (\delta_1, \delta_2, \ldots, \delta_k)$ be the sequence of intervals defined by $\delta_i = t_i - t_{i-1}$, $\boldsymbol{\xi}^{(k)} = (\xi_1, \xi_2, \ldots, \xi_k)$, and $\boldsymbol{\theta} = (k, \boldsymbol{\delta}^{(k)})$. It should be emphasized that quantities are indexed by (k) even though k is considered to be an unknown in the general problem that we address here. However, k is considered to be a known constant in model $\mathcal{M}_{\boldsymbol{\theta}}$, which is defined as

$$\mathcal{M}_{\boldsymbol{\theta}} : \begin{cases} p\left(\boldsymbol{x} \mid \boldsymbol{\xi}^{(k)}\right) = \Pi_{i=1}^{k}\ p(x_{t_{i-1}+1}, \ldots, x_{t_i} \mid \xi_i) \\ \pi\left(\boldsymbol{\xi}^{(k)} \mid \mu\right) = \Pi_{i=1}^{k}\ \pi(\xi_i \mid \boldsymbol{\mu}) \end{cases}, \tag{1}$$

where $\pi(. \mid \boldsymbol{\mu})$ denotes an informative prior distribution, which will be described in this section. Hence, $\mathcal{M}_{\boldsymbol{\theta}}$ models the series and its changing parameters when the number and instants of change points are known, that is, when $\boldsymbol{\theta}$ is known. Since the multiple change-point problem involves determining change-point number and instants, under this framework, it can be viewed as a problem of model choice among the set Λ of all possible models $\{\mathcal{M}_{\boldsymbol{\theta}} \mid k \in (0, 1, \ldots, n),\ \boldsymbol{\delta}^{(k)} \in \Lambda_k\}$, where Λ_k is the set of all possible segmentations into k segments. Λ_k has cardinality $\binom{n}{k}$ and Λ has cardinality 2^n. An alternative approach would be to choose between the $n+1$ models $\mathcal{M}_k$ for $k = 0, 1, \ldots, n$, which are defined similarly but with $(\boldsymbol{\delta}^{(k)}, \boldsymbol{\xi}^{(k)})$ being the vector of parameters instead of $\boldsymbol{\xi}^{(k)}$. This approach would be equally relevant as ours, yet we choose not to follow it for computational complexity reasons, which will be made clear in Section 3.

The Bayesian framework is useful for the general purpose of model selection because it offers the possibility of computing a posterior model probability on the set of all possible models, provided some prior probability can be established on this set (Robert 2001; Parent and Bernier 2007). When this condition is not met or presents difficulties, an alternative is to base model selection on the Bayes factor, which asymptotically leads to the same result. This latter strategy underlies the general approach of Schwarz (1978) in deriving the BIC. But in the present case, we are able to introduce some problem-specific assumptions that result in the straightforward derivation of informative prior model probability. Hence, we decide to follow a model selection strategy that involves choosing the model with the largest posterior probability. This posterior model probability $p(\mathcal{M}_{\boldsymbol{\theta}} \mid \boldsymbol{x})$ is obtained as

$$p(\mathcal{M}_{\boldsymbol{\theta}} | \boldsymbol{x}) \propto \int p\left(\boldsymbol{x} | \boldsymbol{\xi}^{(k)}\right) . \pi\left(\boldsymbol{\xi}^{(k)} | \mu\right)\, \mathrm{d}\boldsymbol{\xi}^{(k)} . \pi(\mathcal{M}_{\boldsymbol{\theta}}). \tag{2}$$

Note that the normalization constant is not needed for our purpose: only the right-hand term is needed for the sake of maximizing $p(\mathcal{M}_{\boldsymbol{\theta}} \mid \boldsymbol{x})$ in $\boldsymbol{\theta}$; hence, we will focus only on $\int p(\boldsymbol{x} \mid \boldsymbol{\xi}^{(k)}) . \pi(\boldsymbol{\xi}^{(k)} \mid \mu)\, \mathrm{d}\boldsymbol{\xi}^{(k)} . \pi(\mathcal{M}_{\boldsymbol{\theta}})$. This remark has considerable practical implications because we do not require to compute the value of a sum with 2^n terms—a number that grows rapidly as n increases—but we only require its maximum term; hence, practically, we move from a summation to an optimization problem.

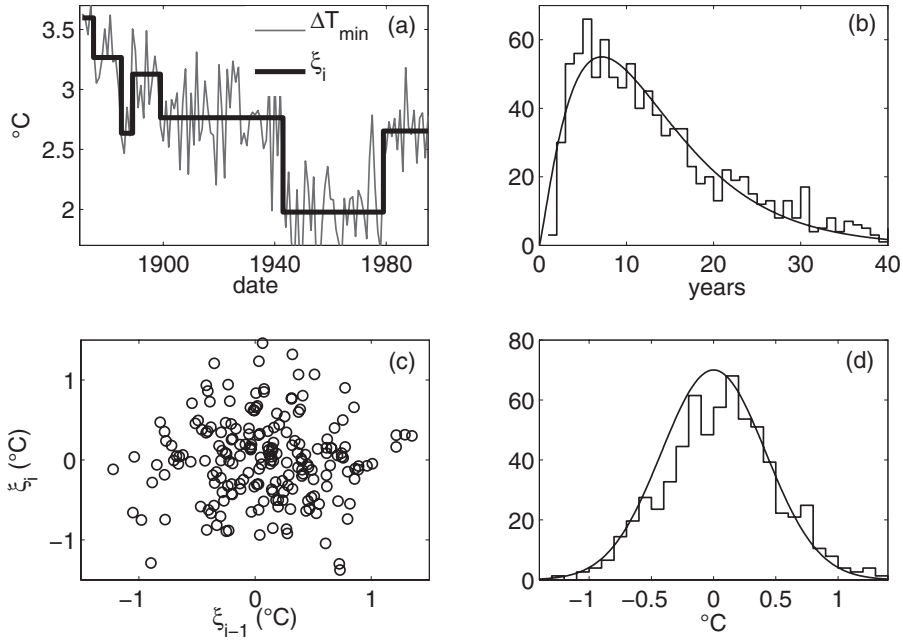

Figure 1. (a) Series of yearly temperature difference between two neighboring Météo France stations and detected instrumental changes. (b) Empirical distribution of interevent times and gamma fit. (c) Scatterplot of consecutive means showing the absence of correlation. (d) Empirical distribution of means and Gaussian fit.

Before computing this quantity, we need to introduce further modeling assumptions to characterize the prior probability distributions $\pi(\boldsymbol{\xi}^{(k)} \mid \mu)$ and $\pi(\mathcal{M}_{\boldsymbol{\theta}})$. The key idea is to assume that a discrete stochastic process simultaneously drives the number of change points k, their occurrence times $\boldsymbol{t}$ (or equivalently the interoccurrence intervals $\boldsymbol{\delta}$), and the changing parameters $\boldsymbol{\xi}$. More specifically, following usual terminology, we assume that this process is a marked renewal process, or in other words, that the successive values δ_i of interevents intervals are iid with pdf $\pi(\delta_i \mid \boldsymbol{\lambda})$ and that the values ξ_i of the parameter on the successive segments are iid with pdf $\pi(\xi_i \mid \boldsymbol{\mu})$. This marked renewal process assumption for the change-point problem is similar to earlier formulations by Yao (1984), where interevent times have a geometric distribution (i.e., a Bernoulli process) and changing parameters have a Gaussian distribution, or by Chib (1998), Lai, Liu, and Xing (2005), and Lai and Xing (2008), who extended this model to more general situations. The applicative context of climate series homogenization provides a good practical illustration for this modeling framework. Indeed, inhomogeneities are commonly modeled as changes in the mean ξ_i of a Gaussian series, and results from previous studies consistently find inhomogeneities to occur on average every 10 to 15 years, with the distribution of interevent times $\pi(\delta_i \mid \boldsymbol{\lambda})$ fitting well empirically with a negative binomial distribution (Figure 1b, and Hannart and Naveau 2009). Similarly, we can also empirically estimate $\pi(\xi_i \mid \boldsymbol{\mu})$ from previous homogenization results, and we find the successive values of the mean parameter ξ_i to be iid (Figure 1c) according to a centered Gaussian with standard deviation 0.40°C (Figure 1d).

As in this illustration, we will assume in what follows that $\pi(\delta_i \mid \boldsymbol{\lambda})$ is the discrete negative binomial distribution

$$\pi(\delta_i \mid \boldsymbol{\lambda}) = \mathcal{N}b(\delta_i \mid \boldsymbol{\lambda}) = \frac{\Gamma\left(\frac{1}{s^2} + \delta_i\right)}{\delta_i!\Gamma\left(\frac{1}{s^2}\right)} \lambda_0^{\frac{1}{s^2}} (1-\lambda_0)^{\delta_i}, \qquad (3)$$

with $\boldsymbol{\lambda} = (\lambda_0, s^2)$, where λ_0 and s are metaparameters, respectively, equal to the mean and to the ratio of standard deviation to mean of the distribution. The reason for choosing such a parameterization will become clear shortly. In what follows, for convenience, we choose to approximate this discrete distribution by its continuous version, that is, by a gamma distribution, and we use

$$\pi(\delta_i \mid \boldsymbol{\lambda}) = \mathcal{G}a(\delta_i \mid \boldsymbol{\lambda}) = \left\{(\lambda_0 s^2)^{\frac{1}{s^2}} \Gamma\left(\frac{1}{s^2}\right)\right\}^{-1} \delta_i^{\frac{1}{s^2}-1} \times \exp\left\{-\frac{\delta_i}{\lambda_0 s^2}\right\}. \qquad (4)$$

The choice of this class of distribution for interevent times does not imply a dramatic loss of generality because a wide variety of situations are covered as s varies from 0 to 1. Indeed, the renewal process ranges from the fully deterministic λ_0-periodic process when s is 0 to fully memory-less Bernoulli process when s is 1. The metaparameter s therefore reflects the a priori level of regularity on the change-point spacings.

Note that, in contrast with $\pi(\delta_i \mid \boldsymbol{\lambda})$, it would not be relevant to propose a general form for $\pi(\xi_i \mid \boldsymbol{\mu})$ as it relates to the particular nature of the changing parameter ξ_i, and thus, to the particular model considered. It is also important to emphasize that in the present approach, the actual values of $\boldsymbol{\lambda} = (\lambda_0, s)$ and $\boldsymbol{\mu}$ are assumed to be known constants—that is, they are not estimated from the data $\boldsymbol{x}$. Hence, our method is fully a priori informed Bayesian, so its execution does require the availability of some prior information (here summarized by metaparameters λ_0, s, and $\boldsymbol{\mu}$). As in any Bayesian inference based on informative priors, those values can significantly influence the outcome, and hence, they should be chosen judiciously. A general discussion on this fundamental aspect of Bayesian statistics can be found, for instance, in Robert (2001, pp. 105–165). It is also discussed

further in Section 6. One specific illustration is provided by the aforementioned application to homogenization in which a priori information is obtainable from previous homogenization studies; analyzing those previous results does indeed yield $\lambda_0 =$ 13 years, $s = 0.7$, and $\mu = 0.40°\text{C}$ (where μ represents the a priori standard deviation of the changing mean).

3. APPROXIMATIONS AND INFERENCE

We now derive an approximation of the posterior model probability $\int p(\boldsymbol{x} \mid \boldsymbol{\xi}^{(k)}) \,.\, \pi(\boldsymbol{\xi}^{(k)} \mid \mu)\, \mathrm{d}\boldsymbol{\xi}^{(k)} \,.\, \pi(\mathcal{M}_{\boldsymbol{\theta}})$. We first approximate the prior model probability $\pi(\mathcal{M}_{\boldsymbol{\theta}})$ and then the model marginal likelihood $\int p(\boldsymbol{x} \mid \boldsymbol{\xi}^{(k)}) \,.\, \pi(\boldsymbol{\xi}^{(k)} \mid \mu)\, \mathrm{d}\boldsymbol{\xi}^{(k)}$. The prior model probability can be factorized into

$$\pi(\mathcal{M}_{\boldsymbol{\theta}}) = \pi\left(\boldsymbol{\delta}^{(k)} \mid k, \boldsymbol{\lambda}\right) \,.\, \pi(k \mid \boldsymbol{\lambda}), \tag{5}$$

where $\pi(\boldsymbol{\delta}^{(k)} \mid k, \boldsymbol{\lambda})$ denotes the joint probability density function (pdf) of the interevent times, conditional on the fact that there are $k-1$ change points, and $\pi(k \mid \boldsymbol{\lambda})$ denotes the probability that there are $k-1$ change points. As a classic result of renewal analysis (Lefebvre 2005), we have $\pi(k \mid \boldsymbol{\lambda}) = p(s_k \geq n \mid \boldsymbol{\lambda}) - p(s_{k-1} \geq n \mid \boldsymbol{\lambda})$, where $s_k = \sum_{i=1}^{k} \delta_i$, and

$$\pi(k \mid \boldsymbol{\lambda}) = \bar{\Gamma}\left(n \middle| \frac{k-1}{s^2}, \lambda_0 s^2\right) - \bar{\Gamma}\left(n \middle| \frac{k}{s^2}, \lambda_0 s^2\right), \tag{6}$$

where $\bar{\Gamma}(. \mid a, b)$ stands for the cumulative density function (cdf) of the gamma distribution with parameters a and b.

In contrast with $\pi(k \mid \boldsymbol{\lambda})$, it is not possible to obtain a closed expression of the joint conditional pdf $\pi(\boldsymbol{\delta}^{(k)} \mid k, \boldsymbol{\lambda})$. We hence propose to approximate $\pi(\boldsymbol{\delta}^{(k)} \mid k, \boldsymbol{\lambda})$ by the product of the marginal conditional distributions $\prod_{i=1}^{k} \pi(\delta_i \mid k, \boldsymbol{\lambda})$, thus obtaining a so-called composite marginal likelihood (Varin 2008). The accuracy of this approximation is clearly a decreasing function of the degree of dependence in the joint distribution $\pi(\boldsymbol{\delta}^{(k)} \mid k, \boldsymbol{\lambda})$. Since $\rho(\delta_i, \delta_j) = -\frac{1}{k-1}$ for $i \neq j$ (Appendix A), this dependence quickly becomes negligible as k increases; hence, the approximation is justified. For gamma-distributed interevent times, the marginal conditional pdf $\pi(\delta_i \mid k, \boldsymbol{\lambda})$ is found to be equal to the beta distribution with parameters $\frac{1}{s^2}$ and $\frac{k-1}{s^2}$ (Appendix B). Remarkably, this conditional distribution is independent of λ_0, but the standard deviation to mean ratio s of the unconditional distribution remains unchanged. Hence,

$$\begin{aligned}\pi\left(\boldsymbol{\delta}^{(k)} \middle| k, \boldsymbol{\lambda}\right) = &\left[\frac{1}{n} B\left(\frac{1}{s^2}, \frac{k-1}{s^2}\right)\right]^k \left[\prod_{i=1}^{k} \frac{\delta_i}{n}\right]^{\frac{1}{s^2}-1} \\ &\times \left[\prod_{i=1}^{k}\left(1 - \frac{\delta_i}{n}\right)\right]^{\frac{k-1}{s^2}-1} \,.\, I_{\left\{\boldsymbol{\delta}^{(k)} \middle| \sum_{i=1}^{k} \delta_i = n\right\}}\left(\boldsymbol{\delta}^{(k)}\right),\end{aligned} \tag{7}$$

where $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the beta function, and where $I_A(a)$ is 1 when $a \in A$ and 0 when $a \notin A$.

The marginal likelihood $\int p(\boldsymbol{x} \mid \boldsymbol{\xi}^{(k)}) \,.\, \pi(\boldsymbol{\xi}^{(k)} \mid \mu)\, \mathrm{d}\boldsymbol{\xi}^{(k)}$ of model $\mathcal{M}_{\boldsymbol{\theta}}$, denoted $m(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\mu})$, may be obtained in a closed form in favorable situations where a conjugate prior is available for $\boldsymbol{\xi}^{(k)}$. But to remain more general, we do not want to impose this condition for the applicability of our criterion. Hence, we propose instead to evaluate the marginal likelihood using the Laplace approximation, as in the original BIC construction (Schwarz 1978) and as detailed, for instance, in Kass and Raftery (1994). By straightforward application of this approximation on each of the k segments of the series, we obtain

$$m(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\mu}) = \prod_{i=1}^{k} (2\pi)^{\frac{d}{2}} \,.\, \hat{\sigma}_i \,.\, \delta_i^{-\frac{d}{2}} \,.\, \hat{\pi}_i(\boldsymbol{\mu}) \,.\, \hat{\mathcal{L}}_i(\boldsymbol{\theta}), \tag{8}$$

where $\hat{\xi}_i$ is the maximum likelihood estimator of ξ_i on segment $(x_{t_i+1}, \ldots, x_{t_{i+1}})$, and $\hat{\mathcal{L}}_i(\boldsymbol{\theta}) = \mathcal{L}(\hat{\xi}_i \mid \boldsymbol{\theta})$ and $\hat{\pi}_i(\boldsymbol{\mu}) = \pi(\hat{\xi}_i \mid \boldsymbol{\mu})$ are the corresponding values of the likelihood function and prior pdf evaluated at $\hat{\xi}_i$. For this approximation to hold, we assume that the distribution of the series belongs to the exponential family; that is, the likelihood function $\mathcal{L}(\xi_i \mid \boldsymbol{\theta})$ can be written as $\exp(-nh(\xi_i \mid \boldsymbol{\theta}))$, where h is a function that is twice differentiable in each component of ξ_i. Denoting by $D^2h(\xi_i \mid \boldsymbol{\theta})$ the determinant of the Hessian matrix of h evaluated in ξ_i, the quantity $\hat{\sigma}_i$ in (8) is then obtained as $\hat{\sigma}_i = \mid D^2h(\hat{\xi}_i \mid \boldsymbol{\theta}) \mid^{-\frac{1}{2}}$. In the case $d = 1$, this expression simplifies to $\hat{\sigma}_i = \mid h''(\hat{\xi}_i \mid \boldsymbol{\theta}) \mid^{-\frac{1}{2}}$. Further, interevent times δ_i need to be large enough for the approximation to be accurate, a condition that is discussed in Section 6.

Combining Equations (6), (7), and (8), taking the negative logarithm for convenience, and making some arrangements (Appendix C), it follows that we determine $\boldsymbol{\theta}$, that is, the number of change points k and their positions $\boldsymbol{\delta}^{(k)}$, by minimization over $\boldsymbol{\theta}$ of the criterion:

$$\mathcal{P}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = -\hat{\ell}(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}). \tag{9}$$

In Equation (9), the first term $\hat{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \log\{\hat{\mathcal{L}}_i(\boldsymbol{\theta})\}$ is the logarithm of the maximum likelihood over $\boldsymbol{\xi}^{(k)}$ and quantifies the model fit. The second term $\mathcal{C}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda})$ is a Bayesian penalty similar in nature to the penalty term $\frac{1}{2}k \log n$ found in the BIC of Schwarz (1978). This term quantifies the divergence between model $\mathcal{M}_{\boldsymbol{\theta}}$ once fitted and the a priori knowledge. It is convenient for clarity to split $\mathcal{C}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda})$ into two components that have contrasted roles:

$$\mathcal{C}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathcal{C}_1(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) + \mathcal{C}_2(k \mid \boldsymbol{\lambda}), \tag{10}$$

in which

$$\begin{aligned}\mathcal{C}_1(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \left(\frac{d}{2} + 1 - \frac{1}{s^2}\right) \sum_{i=1}^{k} \log \delta_i - \sum_{i=1}^{k} \log\{\hat{\sigma}_i \, \hat{\pi}_i(\boldsymbol{\mu})\}, \\ \mathcal{C}_2(k \mid \boldsymbol{\lambda}) &= -\frac{1}{2} d\, k \log 2\pi \; - k \log B\left(\frac{1}{s^2}, \frac{k-1}{s^2}\right) \\ &\quad + \frac{k}{s^2}(1 + \log n) - \log \psi(k, \boldsymbol{\lambda}),\end{aligned} \tag{11}$$

and $\psi(k, \boldsymbol{\lambda}) = \bar{\Gamma}(n \mid \frac{k-1}{s^2}, \lambda_0 s^2) - \bar{\Gamma}(n \mid \frac{k}{s^2}, \lambda_0 s^2)$. The first term $\mathcal{C}_1$ is a function of k and of the series segmentation $\boldsymbol{\theta}$. It is influenced by the data $\boldsymbol{x}$ and by the prior characteristics of change-point amplitude and interevent time. In contrast, the second term $\mathcal{C}_2$ is a function of k and is not influenced by the series segmentation $\boldsymbol{\theta}$, or by the data, or by prior characteristics of change-point amplitude $\boldsymbol{\mu}$. It is also remarkable that the prior mean interevent time λ_0 does not appear in $\mathcal{C}_1$: in practice, this means that when k is fixed, λ_0 has no influence on the computation of the optimal change-point instants. Finally, the influence of the prior standard deviation to mean interevent time ratio s^2, which represents the level of determinism of the renewal process, is critical for both $\mathcal{C}_1$ and $\mathcal{C}_2$. When s approaches 0, that is, the process tends to be deterministic, the selection

criterion for change-point instants is dominated by $-\frac{1}{s^2}\sum_{i=1}^{k}\log\delta_i$, which is minimal for all δ_i's equal to $\frac{n}{k}$. The procedure thus consistently leads to a deterministic choice of equally spaced change-point instants for any data $\boldsymbol{x}$. The term $\mathcal{C}_2$ then constrains this constant interevent time to match with λ_0. Hence, it can be foreseen that the value of s strongly relates to the level of information brought by priors on interevent time: in particular, for small s, prior information overwhelms updated information brought by the data. It is intuitive from this reasoning that our a priori informed selection criterion will perform better as compared with other criteria for small s, provided the prior agrees with the truth, because the criterion then benefits from a substantial extra amount of information brought by priors. This intuition is further discussed in Section 4.

We now focus on implementation details for the minimization of the selection criterion $\mathcal{P}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\lambda})$. The minimization scheme is classic and similar to the one followed, for instance, by Hawkins (2001), Caussinus and Mestre (2004), Lavielle (2004), and Zhang (2007), and the reader is referred to these references for more details. The scheme is a two-step procedure. First, the quantity $-\hat{\ell}(k, \boldsymbol{\delta}^{(k)}) + \mathcal{C}_1(k, \boldsymbol{\delta}^{(k)} \mid \boldsymbol{\mu}, \boldsymbol{\lambda})$ is minimized in $\boldsymbol{\delta}^{(k)}$ successively for $k = 0, 1, \ldots, n$, leading to the optimal change-point instants $\hat{\boldsymbol{\delta}}^{(k)}$ for every given k. Second, the complete criterion $-\hat{\ell}(k, \hat{\boldsymbol{\delta}}^{(k)}) + \mathcal{C}_1(k, \hat{\boldsymbol{\delta}}^{(k)} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) + \mathcal{C}_2(k \mid \boldsymbol{\lambda})$ is minimized in k, leading to the optimal number of change points $\hat{k}$. The second step is straightforward and presents no computational difficulty. The first step can be solved rather easily based on a dynamic programming algorithm that has been used frequently in multiple change-point analysis (see above references). The algorithm takes advantage of the fact that the criterion to be minimized in $\boldsymbol{\delta}^{(k)}$ consists of a sum of k contributions, each associated with one of the k segments $]t_i, t_{i+1}]$. Therefore, change points can be seen as nodes of a graph and the minimization as a classic shortest-path problem (Cormen et al. 2001), where each possible segment δ_i represents an edge with a given length. The dynamic programming algorithm runs with quadratic complexity in n.

4. SIMULATION STUDY

The purpose of this section is to assess the performance of our criterion by implementing the procedure on simulated series and to compare results with the performance of some recent criteria designed for the same purpose (three criteria). We conduct this performance evaluation for a diverse range of models (five models), on a large number of simulations (10,000 simulations), and assess performance based on a diverse range of metrics (three metrics).

4.1 Models

The equations below describe the models' assumptions and corresponding estimators. Figure 2 shows simulated series for the five models. These models are intended to cover a wide range of situations: changes that affect a parameter that relates to the mean only (M1), to the variance only (M3), to both the mean and the variance (M4, M5), to the dependence structure (M2), series of variables that are continuous (M1, M2, M3) or discrete (M4, M5), and series of variables that are univariate (all except M2) or multivariate (M2).

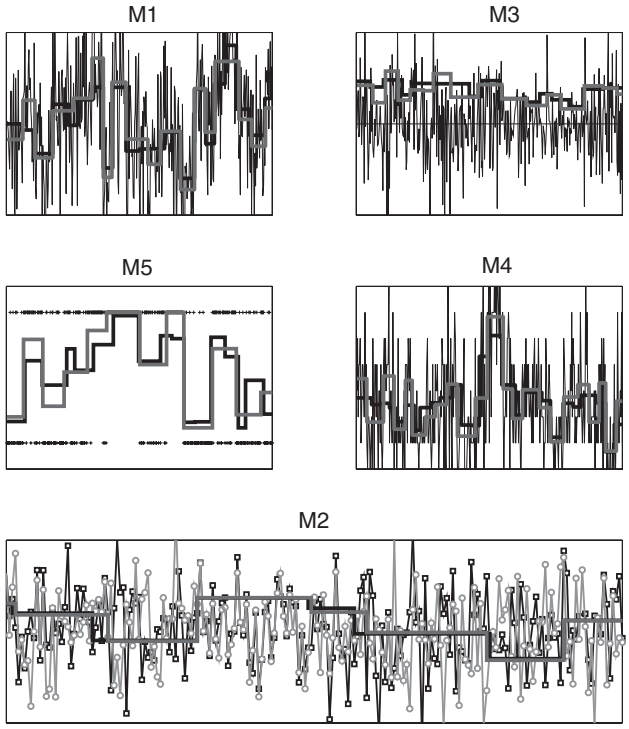

Figure 2. Examples of simulated series $\boldsymbol{x}$ in each of the five models M1 through M5, from upper-left to lower-right panel. The lighter line represents the series, the thick black line is the actual value of the changing parameter, and the thick gray line is the inferred value of the changing parameter. For model M2, which is bivariate, both series are plotted.

M1: Gaussian iid series with changes in the mean:

$$p\left(\boldsymbol{x}\middle|\boldsymbol{\xi}^{(k)}\right) = \Pi_{i=1}^{k}\ \Pi_{j=t_{i-1}+1}^{t_i}\ 2\pi^{-\frac{1}{2}}\ \exp\left\{-\frac{1}{2}(x_j-\xi_i)^2\right\}$$

$$p(\xi_i \mid \boldsymbol{\mu}) = \{2\pi\ \mu\}^{-\frac{1}{2}}\ \exp\left\{-\frac{\xi_i^2}{2\mu}\right\}, \quad \text{with} \quad \boldsymbol{\mu} = \mu,$$

$$\Rightarrow h(\xi_i) = \frac{1}{2}\left\{\log 2\pi + \hat{s}_i^2 + (\xi_i - \bar{x}_i)^2\right\} \Rightarrow \hat{\xi}_i = \bar{x}_i,\ \hat{\sigma}_i = 1,$$

M2: Gaussian bivariate iid series with changes in the correlation:

$$p\left(\boldsymbol{x}\middle|\boldsymbol{\xi}^{(k)}\right) = \Pi_{i=1}^{k}\ \Pi_{j=t_{i-1}+1}^{t_i}\left\{2\pi\sqrt{1-\xi_i^2}\right\}^{-1} \times \exp\left\{-\frac{x_{1,j}^2 - 2\,\xi_i\,x_{1,j}\,x_{2,j} + x_{2,j}^2}{2(1-\xi_i^2)}\right\}$$

$$p(\xi_i \mid \boldsymbol{\mu}) = \frac{1}{2}\,\mathcal{B}e\left(\frac{1+\xi_i}{2}\middle|\mu, \mu\right), \quad \text{with} \quad \boldsymbol{\mu} = \mu,$$

$$\Rightarrow h(\xi_i) = \frac{1}{2}\left\{\log\left(1-\xi_i^2\right) + \frac{\hat{s}_{1,i}^2 + \hat{s}_{2,i}^2 - 2\,\xi_i\,\hat{s}_{12,i}}{1-\xi_i^2}\right\}$$

$$\Rightarrow \hat{\xi}_i,\ \hat{\sigma}_i\text{: no closed expressions,}$$

M3: Gaussian iid series with changes in the variance:

$$p(\boldsymbol{x}|\boldsymbol{\xi}^{(k)}) = \Pi_{i=1}^{k}\ \Pi_{j=t_{i-1}+1}^{t_i}\ \{\sqrt{2\pi}\xi_i\}^{-1}\ \exp\left\{-\frac{x_j^2}{2\,\xi_i}\right\}$$
$$p(\xi_i \mid \boldsymbol{\mu}) = \mu^{-1}\,\mathbf{1}_{[1-\mu,1]}(\xi_i), \quad \text{with} \quad \boldsymbol{\mu} = \mu,$$
$$\Rightarrow h(\xi_i) = \frac{1}{2}\left\{\log(2\pi\,\xi_i) + \frac{\hat{s}_i^2}{\xi_i^2}\right\} \Rightarrow \hat{\xi}_i = \hat{s}_i,\ \hat{\sigma}_i = \frac{1}{\sqrt{2}}\hat{s}_i,$$

M4: Poisson iid series with changes in the parameter:

$$p(\boldsymbol{x}|\boldsymbol{\xi}^{(k)}) = \Pi_{i=1}^{k}\ \Pi_{j=t_{i-1}+1}^{t_i}\ \frac{\xi_i^{x_j}}{x_j!}\,e^{-\xi_i}$$
$$p(\xi_i \mid \boldsymbol{\mu}) = \mathcal{E}xp(\xi_i \mid \mu), \quad \text{with} \quad \boldsymbol{\mu} = \mu,$$
$$\Rightarrow h(\xi_i) = \xi_i - \bar{x}_i \log \xi_i + \hat{w}_i \Rightarrow \hat{\xi}_i = \bar{x}_i,\ \hat{\sigma}_i = \sqrt{\bar{x}_i},$$

M5: Bernoulli iid series with changes in the probability:

$$p(\boldsymbol{x} \mid \boldsymbol{\xi}^{(k)}) = \Pi_{i=1}^{k}\ \Pi_{j=t_{i-1}+1}^{t_i}\ \xi_i^{x_j}(1-\xi_i)^{1-x_j}$$
$$p(\xi_i \mid \boldsymbol{\mu}) = \mathcal{B}e(\xi_i \mid \mu, \mu) \quad \text{with} \quad \boldsymbol{\mu} = \mu,$$
$$\Rightarrow h(\xi_i) = \bar{x}_i \log \xi_i + (1-\bar{x}_i)\log(1-\xi_i)$$
$$\Rightarrow \hat{\xi}_i = \bar{x}_i,\ \hat{\sigma}_i = \sqrt{\bar{x}_i(1-\bar{x}_i)}.$$

where $\bar{x}_i = \frac{1}{\delta_i}\sum_{j=t_{i-1}+1}^{t_i} x_j$, $\hat{s}_i^2 = \frac{1}{\delta_i}\sum_{j=t_{i-1}+1}^{t_i}(x_j - \bar{x}_i)^2$, and $\hat{w}_i = \frac{1}{\delta_i}\sum_{j=t_{i-1}+1}^{t_i}\log(x_j!)$. Finally, as a reminder of Section 2, in all models, the prior distribution of interevent times is equal to $\mathcal{G}a(\delta_i \mid \boldsymbol{\lambda}) = \{(\lambda_0 s^2)^{\frac{1}{s^2}}\Gamma(\frac{1}{s^2})\}^{-1}\ \delta_i^{\frac{1}{s^2}-1}\ \exp\{-\frac{\delta_i}{\lambda_0 s^2}\}$. Also, note that in all models, the parameter ξ_i is scalar, that is, we have $d = 1$.

4.2 Simulations

We generate a pool of 10,000 simulated series (2000 per model). The objective in building the simulation pool is to cover a wide range of possible situations for each five models with respect to series length n, number, spacing, and amplitude of underlying change points driven by prior characteristics $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, and, of course, data $\boldsymbol{x}$ itself. The simulation of each series $\boldsymbol{x}$ therefore goes in steps. First, the series length n is picked randomly based on a uniform distribution on [100,1000]. Second, $\boldsymbol{\lambda}$ is picked randomly, based on a uniform distribution on [0,1] for s and [10, 40] for λ_0. Similarly, $\boldsymbol{\mu}$ is picked randomly based on a uniform distribution on a range that depends on the particular model. The upper and lower bound for the range of $\boldsymbol{\mu}$ are shown in Table 1. They were chosen to obtain a range of performance that is both large and of comparable extent in all models. Third, the interevent times and changing parameters are simulated, respectively, using $p(\delta_i \mid \boldsymbol{\lambda})$ and out of $p(\xi_i \mid \boldsymbol{\mu})$. The simulation of interevent times then drives the number k and positions of change points. Finally, the series $\boldsymbol{x}$ is simulated out of the underlying simulated change-point structure.

4.3 Criteria

We chose three criteria for performance comparison with our criterion. The original BIC criterion of Schwarz (1978) is selected because our criterion is an adaptation of this approach.

Table 1. Simulation ranges

Model	Quantity	Range
All	n	[100,10,000]
All	λ_0	[10,40]
All	s	[0,1]
M1	μ	[0.5,3]
M2	μ	[0,1]
M3	μ	[0,1]
M4	μ	[1,10]
M5	μ	[0,1]

For the same reason, comparison with the modified BIC criterion of Zhang (2007), which constitutes, to our knowledge, the most recent adaptation of the BIC criterion to the change-point problem, appears as natural. For the third criterion, we decided to choose the criterion proposed by Ninomiya (2005) as an adaptation to the change-point problem of the general selection criterion of Akaike (1974) to obtain a comparison with a criterion based on a radically different theoretical approach. The three criteria, denoted Sc, Zh, and Ni, have the following expressions:

$$\text{Sc} = -\hat{\ell}(\boldsymbol{\theta}) + \left(k + \frac{1}{2}\right)\log n,$$
$$\text{Zh} = -\hat{\ell}(\boldsymbol{\theta}) + \frac{1}{2}\sum_{i=1}^{k}\log\delta_i + \left(k - \frac{1}{2}\right)\log n, \text{ and}$$
$$\text{Ni} = -\hat{\ell}(\boldsymbol{\theta}) + 4k,$$

which are reformulated from the reference articles based on notations and assumptions used herein. In contrast with our criterion (hereafter denoted Ha) and the BIC criterion, which are designed in the general context of the exponential family, criteria Zh and Ni are both designed in the specific context of model M1.

4.4 Performance Metrics

We chose three distinct performance metrics, denoted r_1, r_2, and r_3, defined as follows. Since the procedure aims at determining the number of change points k, performance can be measured simply by the estimation error on k and we define:

$$r_1 = |\,k - \hat{k}\,|\,/k.$$

This metric implies that the end goal of the procedure is to infer the true number of change points, that is, that it is optimal to estimate k by its actual value. However, as argued by Celisse (2008), the actual number of change points is actually not always the optimal estimate, because the corresponding estimation of change-point positions may not be accurate for changes of small amplitude, that is, hidden changes. A metric that reflects only the ability of the procedure to correctly infer the number of jumps k does not necessarily capture the performance of associated estimators $\boldsymbol{\delta}^{(k)}$ and $\boldsymbol{\xi}^{(k)}$. Therefore, we build a quadratic loss that simultaneously measures the estimation error with respect to k,

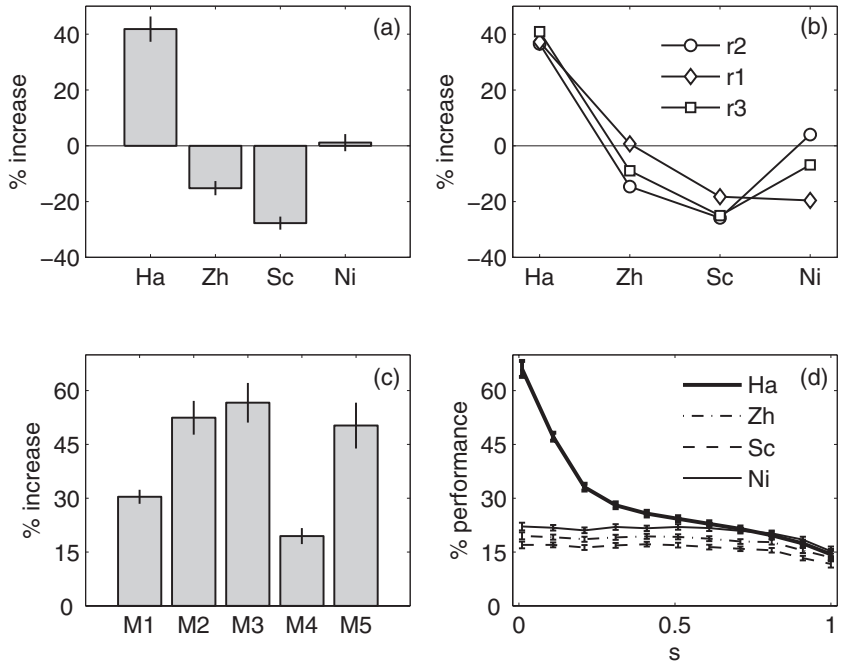

Figure 3. (a) Increase in performance versus the average by criterion, averaged over performance metrics. (b) Increase in performance versus the average by criterion and by performance metric. (c) Increase in performance versus the average by model for criterion Ha, averaged over performance metrics. (d) Absolute level of performance (r_2) by criterion and by value of metaparameter s.

$\boldsymbol{\delta}^{(k)}$, and $\boldsymbol{\xi}^{(k)}$:

$$r_2 = \sum_{j=1}^{n} (\phi_j - \hat{\phi}_j)^2 \Big/ \sum_{j=1}^{n} \phi_j^2,$$

where $\boldsymbol{\phi}$ is the piecewise constant series equal to ξ_i on the ith segment $]t_{i-1}, t_i]$, defined by $\phi_j = \sum_{i=1}^{k} \mathbf{1}_{\{j \in [t_i+1, t_{i+1}]\}} \xi_i$, $j = 1, \ldots, n$), and $\hat{\boldsymbol{\phi}}$ is its estimator. The metric r_2 can be interpreted simply as the mean squared estimation error on the overall change-point structure $\boldsymbol{\phi}$ underlying $\boldsymbol{x}$, which can be regarded as the final objective of a multiple change-point estimation procedure. It has the desirable property to be an increasing function of the estimation error attached to k, $\boldsymbol{\delta}^{(k)}$, and $\boldsymbol{\xi}^{(k)}$. Finally, we use the detection error metric r_3, which is similar to Hannart and Naveau (2009). It trades off the number of true positives $\hat{k}_{\text{tp}}$ with the number of false positives $\hat{k}_{\text{fp}}$. A detected change point is defined as a true positive whenever the estimated position matches an actual position with a $\pm\lambda_0/10$ precision, and as a false positive otherwise. The metric is defined as

$$r_3 = 1 - \left(\hat{k}_{\text{tp}} - \frac{1}{4}\hat{k}_{\text{fp}}\right) \Big/ k,$$

where the weight $\frac{1}{4}$ and the normalization by k are set to obtain $r_3 = 0$ when detection is perfect (i.e., $\hat{k}_{\text{tp}} = k$ and $\hat{k}_{\text{fp}} = 0$) and $r_3 = 1$ when detection is random. Indeed, for a $\pm\lambda_0/10$ precision level, a randomly selected position has probability 0.2 to be a true positive and 0.8 to be a false positive, that is, $\hat{k}_{\text{tp}} = 0.2\,\hat{k}$ and $\hat{k}_{\text{fp}} = 0.8\,\hat{k}$.

4.5 Main Results

The results described in this section are represented graphically in Figure 3. For each metric, we computed the average performance obtained across the five models, the four criteria, and all simulations, and derived the percent difference in performance between this global mean and the mean computed by the criterion. This percent difference reflects the overall relative position in the performance of each criterion. Based on this number, our criterion outperforms very clearly all others, with a +43% gain in performance on average as compared with the mean. The second-best criterion presents a +4% gain and the two others a −17% and −30% loss, respectively. This performance improvement is illustrated in Figure 4 on a series simulated under model M1. This massive performance gap is found to be around +40% whatever metric is used. The gap fluctuates between +15% and +60% depending on the model. These fluctuations in the performance gap are mostly driven by the overall absolute level of performance associated with each model: indeed the gap tends to reduce—but remain positive—when the average performance is high, as a result of the fact that performance is bounded. A high positive gap can therefore not be maintained when all criteria perform at a very high level. For instance, model M4 presents the lowest gap at +15% because change-point detection is easier in this case because the changing parameter affects both the mean and the variance, thus resulting in more apparent changes. Finally, it appears that the performance of our criterion is dramatically affected by the value of metaparameter s, while the performance of all other criteria is nearly insensitive to s. This key result confirms the qualitative reasoning exposed in Section 3: our a priori informed criterion does perform best for small s because it then benefits from a substantial extra amount of information brought by priors. This is not the case for the other criteria, which are not a priori informed and whose performance is nearly insensitive to s as a result. Quantitatively, the performance gap is +240% when s approaches 0, +25% for $s = 0.5$, and +5% for $s = 1$. This

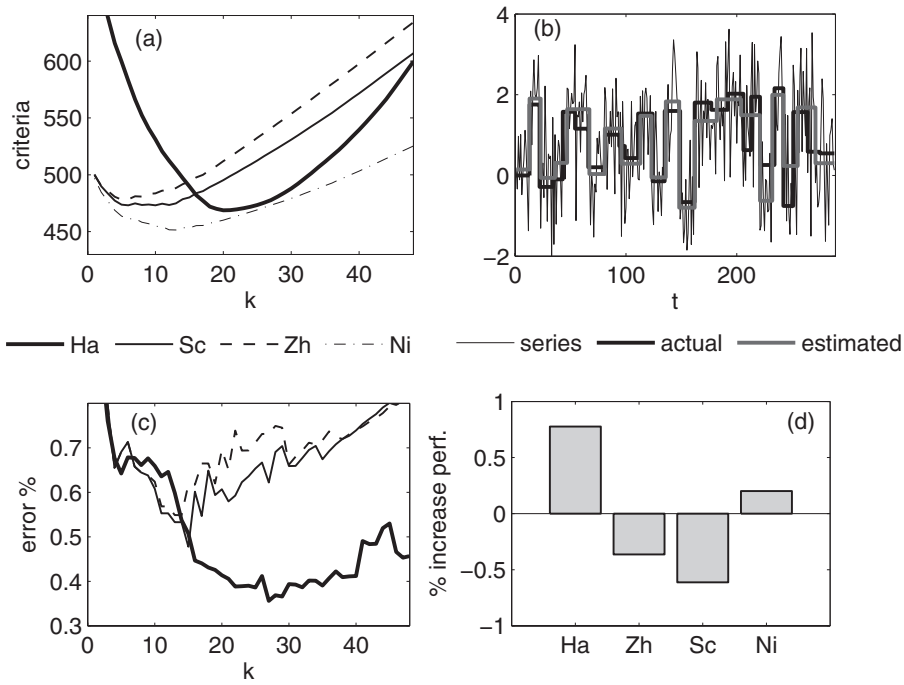

Figure 4. Results obtained for one series $\boldsymbol{x}$ of length $n = 500$, $k = 40$, and $s = 0.5$ simulated through M1: (a) value of the four criteria as a function of k, (b) series $\boldsymbol{x}$, (c) estimation error r_2 for each criterion as a function of k, and (d) performance level r_2 of each criterion.

result suggests that most of the increased performance of our criterion does actually come from the prior information that it takes into account. Finally, it is of interest to note that among the three alternative criteria tested, Ni performs better than the other two based on r_2, Zh performs better than the other two based on r_1, and Ni and Zh perform at about the same level and above Sc based on r_3. In other words, the performance rank-ordering between Ni and Zh varies depending on the performance metric used. But unsurprisingly, Sc always performs worst for any performance metric used.

4.6 Additional Results

While s appears to be the main driver of performance for criterion Ha, our simulation results show that the performance of all four criteria is also sensitive to other drivers. It is beyond our scope to perform a detailed study of performance sensitivity, yet we briefly and qualitatively highlight some general results related to this aspect (not shown graphically). Performance is found to be an increasing function of λ_0 in all models and for all criteria. This is unsurprising as longer interevent times (i.e., larger λ_0) lead to improved inference quality on the successive values ξ_i of the changing parameter, and thus enhanced changes. On the other hand, performance is found to be an increasing function of μ in all models and for all criteria. This is also not surprising because the value of μ is positively related to the average amplitude of changes in all models, and changes with larger amplitudes are more easily detected. In terms of performance rank-ordering between criteria, the fact that Ha performs consistently better than the other criteria, as well as other results described in the previous subsection, appears to be unchanged when varying λ_0 and μ. However, while the performance gap remains positive, its absolute value does change. For instance, large values of λ_0 and μ result in high performance for all criteria, and vice versa. In both cases, the gap between performance narrows when performance is close either to 0 or 1 for all criteria simultaneously. Finally, performance is found to be insensitive to series length n. This is because in our simulation scheme, the average segment length λ_0 is simulated independently of n, so the number of change points, and thus the number of parameters to be inferred from the series, increases linearly with series length. As a result, the quality of inference is not affected by n, but instead is affected by $n/k \approx \lambda_0$, which here is independent of n. Of course, this situation would not prevail should n and k be simulated independently. However, as was argued in Section 1, assuming independence between n and k may not be realistic in the multiple change-point context.

5. APPLICATION

The purpose of this section is to illustrate the practical relevance of our approach, by implementing it for homogenization of meteorological series. As noted in the Introduction, this particular applicative context originally motivated this work.

5.1 Data and Procedure

The data consist of yearly average series of minimum daily temperature from 16 stations of the French weather service. These stations are located in the southeast France, and the series range over a 126-year period (1882–2007). The series has already been homogenized (Caussinus and Mestre 2004; Hannart and Naveau 2009), but we naturally used the nonhomogenized, raw data record. We now recall the main principles of

homogenization—for a detailed review, the reader is referred to Menne and Williams (2008). In a nonhomogenized climate series, instrumental shifts are mixed with the climate signal and thus removing the former enhances the latter. To do so, the relative homogeneity principle is applied: *a climatological series is relatively homogeneous with respect to a synchronous series at another place if the differences of pairs of homologous averages constitute a series of random numbers* (i.e., a white noise), as stated by Conrad and Pollack (1962, p. 226). Artificial shifts are thus detected on the difference series, assuming it is iid Gaussian with changes in the mean. Following this approach, the difference series with every sufficiently correlated neighbor (we used $\rho > 0.75$ as a selection threshold) is derived for each candidate station, and the multiple change-point detection procedure is applied to each difference series. Then, two additional steps are required, attribution and reconciliation: first, each shift detected on a paired difference series may be caused by either of the two series and must hence be attributed to the culprit series; second, multiple shift locations estimated on several paired difference series must be reconciled into a unique date to be used for adjusting the candidate series. These two steps were performed by manual review of the detected shifts based on the inspection of a position × amplitude chart as described in Hannart and Naveau (2009).

5.2 Implementation Details

We further specify a few aspects related to the implementation of our procedure on pairwise difference series. Following the homogeneity principle, the procedure is applied in the context of a model that assumes changes in the mean of a Gaussian series having a fixed, known variance σ. This model is thus identical to model M1 introduced in Section 4, except that the fixed variance is no longer equal to 1. The practical incidence of this rescaling is simply that in Equation (11), $\hat{\sigma}_i$ must be taken equal to σ instead of 1. Then, the computation of the criterion requires specification of the value of the constant σ as well as of the values of metaparameters μ, λ_0, and s—where metaparameter μ represents in this model the standard deviation of the changing mean parameter. These four values are derived from the homogenization results described in Caussinus and Mestre (2004) after excluding the 16 series treated here. Based on these data, we found $\sigma = 0.26^{\circ}$C, $\mu = 0.40^{\circ}$C, $\lambda_0 = 6.5$ years, and $s = 0.8$. Note that estimates of λ_0 and s differ from the values given in Section 2 because they are obtained from difference series, as opposed to station series. Change points are, as may be expected, found to be twice as frequent (i.e., with λ_0 half as large) and less regularly spaced (i.e., larger s) in difference series than in station series.

5.3 Results

We applied this method to the Marseille station (Figure 5). We found 10 neighbors satisfying the selection criterion $\rho > 0.75$ and applied the procedure to the 10 pairwise difference series (middle panels). Between four (Marseille-Nimes) and eight change points (Marseille-Toulouse) were inferred from the pairwise difference series, leading to the detection of 58 change points in total. We plotted the Marseille series, together with its 10 selected neighbors (upper panel) and the amplitude × position chart used for visual attribution (lower panel). After attribution and reconciliation, we find eight inhomogeneities in Marseille in 1904, 1917, 1932, 1947, 1953, 1968, 1982, and 1996, which are consistent with both the metadata and the results. To evaluate the sensitivity of our results to the values of σ, μ, λ_0, and s, we modified these values by $\pm 10\%$ and ran the procedure for the 2^4 possible combinations of modified values. We found that the total number of change points detected on difference series fluctuates by $\pm 20\%$ (i.e., between 47 and 70). However, the final number of inhomogeneities was not affected by these fluctuations, because the manual attribution and reconciliation steps performed afterward lead to the same clear grouping of change points.

6. DISCUSSION AND CONCLUSION

As in the seminal approach of Schwarz (1978), our penalty stems from a Bayesian formulation that involves priors and the Laplace approximation. But while Schwarz obtained a generic expression for the BIC independently of the prior based on asymptotics for large n, we have obtained ours based on some assumptions specific to the change-point problem and by relying on a priori information regarding the number and spacing of change points. Results obtained for simulated series in different models and for a wide range of situations show that our approach systematically outperforms the three alternate criteria tested. The performance gap appears to be mostly driven by s, the metaparameter reflecting the regularity of change-point spacing. Notably, the performance gap dramatically increases when spacing is regular, that is, s is small, as a logical consequence of the fact that the informativeness of accurate priors becomes much stronger in that case. It can therefore be concluded that prior information on change points, especially on their spacing, generates a substantial benefit in inferring their number and position.

Nonetheless, as can be expected, prior information is able to bring a benefit only when it actually matches reality. Conversely, using inaccurate prior information may be toxic for inference especially when informativeness is strong. This sensitivity to the correct value of s is demonstrated in Figure 6, where performance r_2 is plotted as a function of both the true s and its assumed value. This is a limitation of our criterion in its current formulation: reliable prior information is required for its safe implementation. Yet, one can think of many situations where such a priori information may be safely extracted from previous empirical studies conducted on similar cases. For instance, this situation clearly prevails for the application treated in Section 5 of climate series homogenization. It may also prevail in a well-known field of application of change-point methods: the analysis of array-based comparative genomic hybridization (array-CGH) data. Array-CGH measures the number of chromosome copies at each genome location of a cell sample and is useful for finding the regions of genome deletion and amplification in tumor cells (Albertson et al. 2003). Due to the considerable interest raised by this application and the resulting large quantity of previous studies available on this topic, it is reasonable to speculate that some prior information could be extracted to derive values of s and λ_0 that apply specifically to this case.

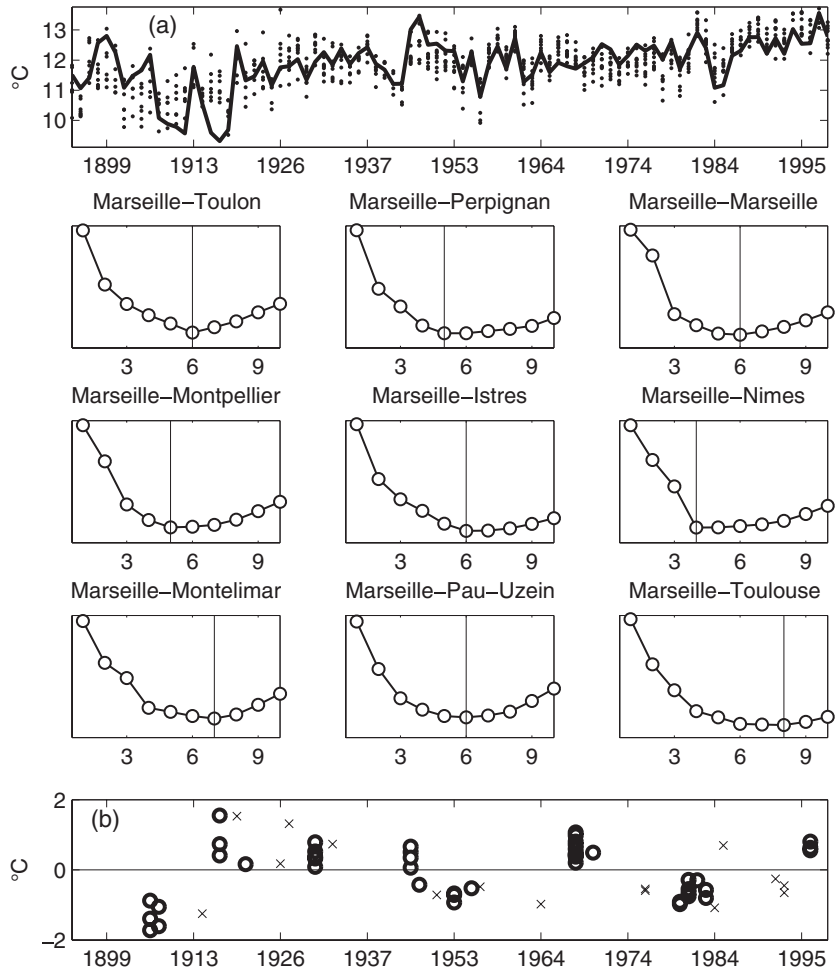

Figure 5. Application to homogenization of temperature series: (a) yearly minimum temperature in candidate series—Marseille (line) and neighboring stations (dots), and (middle panels) criterion for nine pairwise differences series. (b) Position × amplitude chart of shifts detected in difference series, attributed to Marseille (o) or to a neighbor (x).

Extending our approach to the case where a priori information is unavailable is a challenging problem and an exciting opportunity for further research, which is clearly beyond the scope of this article. Indeed, it raises fundamental issues because it requires moving out of the Bayesian informative framework that is inherent in our proposal. To address these issues, two possible routes may be foreseen. A first option would be to apply the so-called empirical Bayes approach, by maximizing the obtained criterion not only in $\boldsymbol{\theta}$, but also simultaneously in $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Such an approach raises multiple difficulties, with both theoretical and implementation issues, but an initial exploration of this approach in model M1 seemed promising. As an

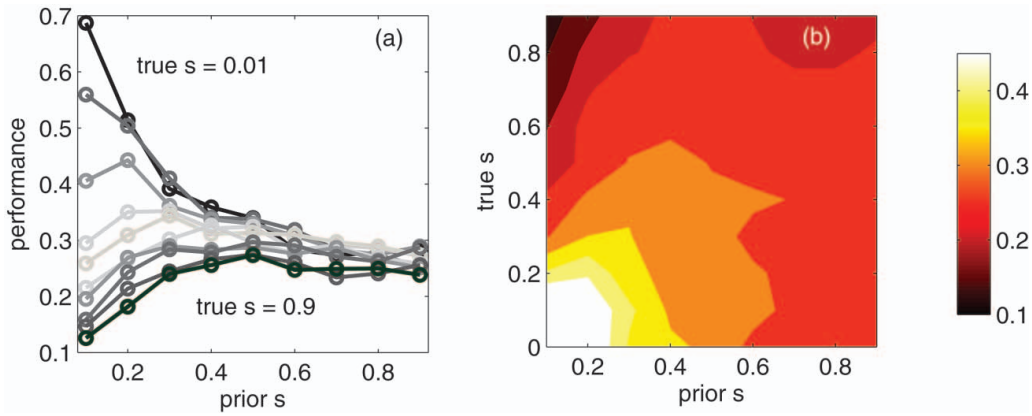

Figure 6. (a) Plot of performance (r_2) as a function of the prior value of s used for inference, for a range of fixed actual values of s. (b) Contour plot of performance as a function of prior value of s and actual values of s. (Both charts show the same data). The online version of this figure is in color.

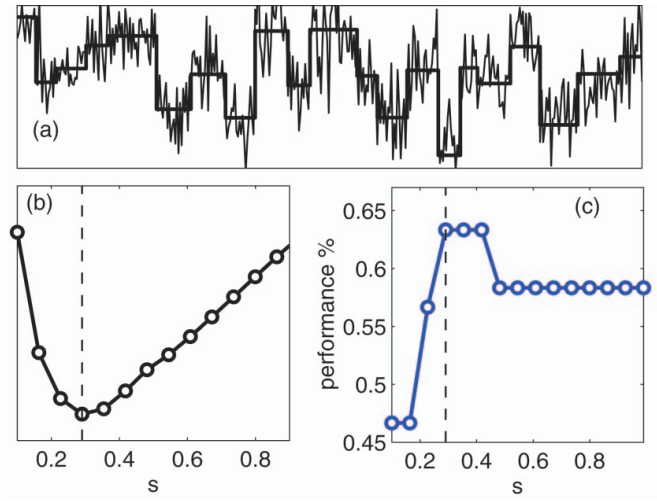

Figure 7. (a) Simulated series in model M1 with $s = 0.3$. (b) Minimal value of the criterion obtained for different values of metaparameter s. (c) Performance (r_2) obtained for different values of metaparameter s. The online version of this figure is in color.

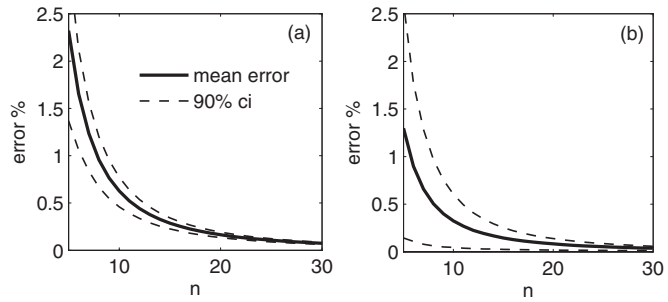

Figure 8. Average error of the Laplace approximation (thick line) and 90% confidence interval (dashed lines) as a function of segment length in model M1, for $\mu = 1$ (a) and $\mu = 2$ (b).

illustration, Figure 7 shows results obtained on a series simulated with $s = 0.3$, assuming $\boldsymbol{\mu}$ and λ_0 are known. For this series, the minimization of the criterion in s does provide a correct estimate $\hat{s} = 0.3$ of s, and its value also coincides with the one that yields maximal performance (r_2). A second option would be to employ noninformative priors. This option also raises nontrivial theoretical issues because of the lack of a unique definition for noninformativeness. Despite these difficulties, in our view, it is reasonable to expect that noninformative extensions will become available in the near future. But as long as such extensions are not available, it is our recommendation to stick restrictively to the informative case when applying the present method, that is, to seek prior information whenever possible—and to use another method otherwise.

Another useful extension of our approach concerns the situation where there exists a vector $\boldsymbol{\eta}$ of unknown parameters that are constant across the entire series (for instance, when the variance of the series $\boldsymbol{x}$ in models M1 and M3 is unknown). A rather straightforward adaptation of our general approach is sketched here to address this case. First, the Laplace approximation is applied in the same way to evaluate the marginal likelihood $m(\boldsymbol{x} \mid \boldsymbol{\theta})$ except that this time, both $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are integrated out (note that the conditionality in $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ is omitted here for clarity). This implies that an overall series-wise maximization in $\boldsymbol{\eta}$ must be performed in addition to the segment-wise maximization in $\boldsymbol{\xi}$. Of course, since $\boldsymbol{\eta}$ is constant across the series, the result of this maximization depends on the segmentation $\boldsymbol{\theta}$, and can thus be denoted $\hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}}$. The resulting Laplace approximation $m(\boldsymbol{x} \mid \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}})$ can then be inserted into Equation (9) to obtain criterion $\mathcal{P}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}})$, whose minimization over $\boldsymbol{\theta}$ yields the desired optimal segmentation. However, for a straightforward application of these two steps, the shortest-path algorithm is not applicable because $\mathcal{P}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_{\boldsymbol{\theta}})$ is no longer segment-additive, in general. Yet, other minimization schemes such as Davis, Lee, and Rodriguez-Yam (2006) are available and could be explored for inference under the same framework.

A cornerstone feature of our method is the use of the Laplace approximation, which assumes sufficiently large segments to be valid. The high performance level of the criterion achieved in simulations with average segment length λ_0 ranging from 10 to 40 and the low sensitivity of performance with respect to length in this range of values (not shown), as well as explicit evaluation of the approximation error in the case of model M1 (Figure 8), suggest that values of length in the order of 10 are sufficient for the approximation to be valid for the models considered. Nonetheless, establishing the domain of validity

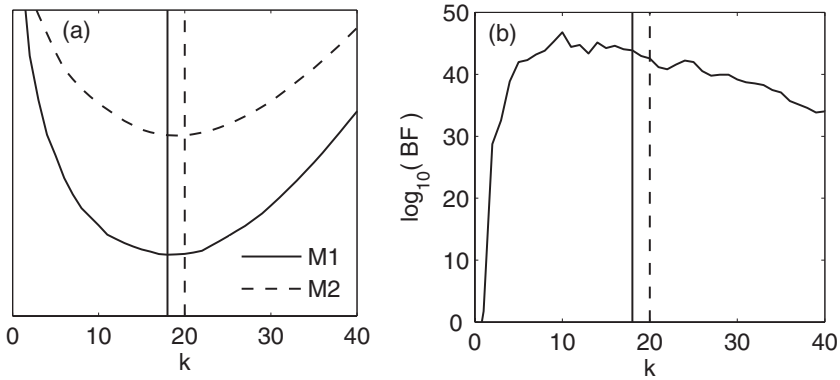

Figure 9. Results obtained for one series $\boldsymbol{x}$ of length $n = 500$ simulated through M1: (a) value of criterion by value of k for both the true model M1 (thick line) and the competing model M2 (dashed line). Minimum is obtained for $k = 18$ in M1 and $k = 20$ in M2. (b) Value of the Bayes factor M1 versus M2 ($\log_{10}$) by value of k. Model M1 is clearly selected for both $k = 18$ and $k = 20$ (value = 40, when > 3 is considered a "very strong amount of evidence").

of the approximation more precisely, as well as refining this approximation, would be a natural extension of this work. This could be done, for instance, by computing the second-order term of the Laplace approximation based on Tierney (1989).

Finally, one advantage of our approach is that it can be extended relatively easily to the more general problem of selecting a change-point model versus another, simultaneously with inference about the number of change points. This extension can be performed simply by restricting the proposed criterion to the logarithm of the marginal likelihood of Equation (8). The term obtained after approximation can be used to derive a Bayes factor for selection among different models based on an absolute scale (Kass and Raftery 1994). This approach was applied for choosing between model M1 and model M2 for a series $\boldsymbol{x}$ simulated from M1, leading to clearly selecting the right model M1 versus the wrong model M2 in the particular example used (Figure 9).

APPENDIX A: CORRELATION BETWEEN INTEREVENT TIMES CONDITIONAL ON THEIR SUM

Since $\sum_{i=1}^{k} \delta_i = n$ is fixed, its variance is zero; hence, $\sum_{i=1}^{k} \mathrm{V}(\delta_i \mid \boldsymbol{\lambda}, k) + \sum_{i \neq j} \mathrm{cov}(\delta_i, \delta_j \mid \boldsymbol{\lambda}, k) = 0$. Since the constraint is on their sum, the δ_i's are interchangeable; hence, $\mathrm{V}(\delta_i \mid \boldsymbol{\lambda}, k)$ and $\mathrm{cov}(\delta_i, \delta_j \mid \boldsymbol{\lambda}, k)$ depend only on $\boldsymbol{\lambda}$ and k. Therefore, $k + k(k-1)\,\rho(\delta_i, \delta_j \mid \boldsymbol{\lambda}, k) = 0$ and $\rho(\delta_i, \delta_j \mid \boldsymbol{\lambda}, k) = -\frac{1}{k-1}$.

APPENDIX B: MARGINAL PDF OF INTEREVENT TIMES CONDITIONAL ON k

Noting that δ_i's are interchangeable, it suffices to derive $p(\delta_1 \mid \boldsymbol{\lambda}, k)$. Since the δ_i are iid $\mathcal{G}a(\delta_i \mid \frac{1}{s^2}, s^2\lambda_0)$, the sum $s_k = \sum_{i=2}^{k} \delta_i$ also follows a gamma distribution $\mathcal{G}a(s_k \mid \frac{k-1}{s^2}, s^2\lambda_0)$. Thus, the marginal conditional distribution $p(\delta_1 \mid \boldsymbol{\lambda}, k)$ is proportional to $\mathcal{G}a(\delta_1 \mid \frac{1}{s^2}, s^2\lambda_0) \times \mathcal{G}a(n - \delta_1 \mid \frac{k-1}{s^2}, s^2\lambda_0) \propto \delta_1^{\frac{1}{s^2}-1}(n-\delta_1)^{\frac{k-1}{s^2}-1}$, which is the beta distribution with mean $\frac{n}{k}$ and variance $s^2\,(\frac{n}{k})^2$.

APPENDIX C: DERIVATION OF $\mathcal{C}_1$ AND $\mathcal{C}_2$

The expression of $\mathcal{C}_1$ and $\mathcal{C}_2$ found in Equation (11) results from straightforward combination of Equations (6), (7), and (8) except for one step, which we detail here. Taking the logarithm of $\pi(\boldsymbol{\delta}^{(k)} \mid k, \boldsymbol{\lambda})$ in Equation (7) results in a term $(\frac{k-1}{s^2} - 1)\sum_{i=1}^{k} \log(1 - \frac{\delta_i}{n})$, which can be simplified based on a first-order approximation, provided the δ_i's are small compared with n, that is, that k is sufficiently large. In that case, $(\frac{k-1}{s^2} - 1)\sum_{i=1}^{k} \log(1 - \frac{\delta_i}{n}) \simeq -(\frac{k-1}{s^2} - 1)\sum_{i=1}^{k} \frac{\delta_i}{n} = -(\frac{k-1}{s^2} - 1)$. Remarkably, under this approximation, this term no longer depends on $\boldsymbol{\delta}^{(k)}$ but only on k; hence, it can be removed from $\mathcal{C}_1$ for the purpose of determining the optimal $\boldsymbol{\delta}^{(k)}$ and can be inserted into $\mathcal{C}_2$, leading to Equation (11). Note that this approximation is actually critical for computational complexity because it makes the minimization over $\boldsymbol{\delta}^{(k)}$ independent of k, thus making possible the direct implementation of the dynamical programming algorithm with quadratic complexity.

ACKNOWLEDGMENTS

The authors thank the Editor, Associate Editor, and two referees for their careful reading of earlier versions of this article. Their thoughtful comments and suggestions greatly improved the article. The authors also gratefully acknowledge the Centre National de la Recherche Scientifique (CNRS), the Consejo Nacional de Investigacion Cientifica y Tecnologica (CONICET), and the University of Buenos Aires (UBA) for its support in this collaboration. Part of this work has been supported by the EU-FP7 ACQWA Project (*www.acqwa.ch*) under contract no. 212250, by the PEPER-GIS project, by the ANR-MOPERA project, by the ANR-McSim project, and by the MIRACCLE-GICC project.

[Received July 2010. Revised March 2012.]

REFERENCES

Abarca-Del-Rio, R., and Mestre, O. (2006), "Decadal to Secular Time Scales Variability in Temperature Measurements Over France," *Geophysical Research Letters*, 33(13), L13705. [256]

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control* 19(6), 716–723. [256,261]

Albertson, D. J., Collins, C., McCormick, F., and Gray, J. W. (2003), "Chromosome Aberrations in Solid Tumors," *Nature Genetics*, 34, 369–376. [264]

Basseville, M., and Nikiforov, I. V. (1996), *Detection of Abrupt Changes: Theory and Application*, Englewood Cliffs, NJ: Prentice Hall. [256]

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer. [256]

Birgé, L., and Massart, P. (2001), "Gaussian Model Selection," *Journal of the European Mathematical Society*, 15, 203–268. [256]

Braun, J., Braun, R., and Muller, H. (2000), "Multiple Change Point Fitting via Quasilikehood With Application to DNA Sequence Segmentation," *Biometrika*, 87, 301–314. [256]

Caussinus H., and Mestre, O. (2004), "Detection and Correction of Artificial Shifts in Climate Series," *Journal of the Royal Statistical Society,* Series C, 53, 405–425. [256,260,264]

Celisse, A. (2008), "Model Selection via Cross-Validation in Density Estimation, Regression and Change-Points Detection," Ph.D. dissertation, University Paris-Sud 11. Available online at *http://tel.archives-ouvertes.fr/tel-00346320/en/* [261]

Chen, J., and Gupta, A. K. (1997), "Testing and Locating Variance Changepoints With Application to Stock Prices," *Journal of the American Statistical Association*, 92, 739–747. [256]

——— (2000), *Parametric Statistical Change Point Analysis*, New York: Birkhauser. [256]

Chib, S. (1998), "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86, 221–241. [256,258]

Conrad, V., and Pollack, L. W. (1962), *Methods in Climatology*, Boston, MA: Harvard University Press. [263]

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001), *Introduction to Algorithms* (2nd ed.), Boston, MA: MIT Press and McGraw-Hill. [260]

Davis, A. R., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006), "Structural Break Estimation for Nonstationary Time Series Models," *Journal of the American Statistical Association*, 101, 223–239. [266]

Dias, A., and Embrechts, P. (2004), "Change-Point Analysis for Dependence Structures in Finance and Insurance," in *Risk Measures for the 21st Century* (Wiley Finance Series), New York: Wiley, pp. 321–335. [256]

Hannart, H., and Naveau, P. (2009), "Bayesian Multiple Change-Points and Segmentation: Application to Homogenization of Climatic Series," *Water Resources Research*, 45, W10444. [256,258,262,263]

Hawkins, D. M. (2001), "Fitting Multiple Change-Points to Data," *Statistical Data Analysis*, 37, 323–341. [260]

Kass, R. E., and Raftery, A. E. (1994), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [259,267]

Kuglitsch, F. G., Toreti, A., Xoplaki, E., Della-Marta, P. M., Luterbacher, J., and Wanner, H. (2009), "Homogenization of Daily Maximum Temperature

Series in the Mediterranean," *Journal of Geophysical Research*, 114, D15108. [256]

Lai, T. L., Liu, H., and Xing, H. (2005), "Autoregressive Models With Piecewise Constant Volatility and Regression Parameters," *Statistica Sinica*, 15, 279–301. [258]

Lai, T. L., and Xing, H. (2011), "A Simple Bayesian approach to multiple change-points." *Statistica Sinica*, 21, 539–569. [258]

Lavielle, M. (2004), "Using Penalized Contrasts for the Change-Point Problem," *Signal Processing*, 85, 1501–1510. [256,260]

Lefebvre M. (2005), *Processus Stochastiques Appliqués* [Stochastic Processes Applied], Paris: Hermann. [259]

Lu, Q., Lund, R., and Lee, T. (2009), "An MDL Approach to the Climate Segmentation Problem," *Annals of Applied Statistics*, 4, 299–319. [256]

Menne, M. J., and Williams, C. N., Jr., (2008), "Homogenization of Temperature Series via Pairwise Comparisons," *Journal of Climate*, 22, 1700–1717. [263]

Ninomiya, Y. (2005), "Information Criterion for Gaussian Change-Point Model," *Statistics and Probability Letters*, 72, 237–247. [256,261]

Pan, J., and Chen, J. (2006), "Application of Modified Information Criterion to Multiple Change Point Problems," *Journal of Multivariate Analysis*, 97, 2221–2241. [257]

Parent, E., and Bernier, J. (2007), *Le Raisonnement Bayésien: Modélisation et Inférence* [Bayesian Reasoning: Inference and Modelization], Paris: Springer. [257]

Perreault, L., Parent, É ., Bernier, J., and Bobée, B. (2000b), "Retrospective Multivariate Bayesian Change-Point Analysis: A Simultaneous Single Change in the Mean of Several Hydrological Sequences," *Stochastic Environmental Research and Risk Assessment*, 14, 243–261. [256]

Robert, C. (2001), *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, New York: Springer. [257,258]

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. (2007), "A Review and Comparison of Change-Point Detection Techniques for Climate Data," *Journal of Applied Meteorology and Climatology*, 46, 900–915. [256]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [256,257,259,261,264]

Serbinowska, M. (1996), "Consistency of an Estimator of the Number of Changes in Binomial Observations," *Statistics and Probability Letters*, 29, 337–344. [256]

Tierney, M. (1989), "Fully Exponential Laplace Approximations," *Journal of the American Statistical Association*, 84, 710–717. [266]

Varin, C. (2008), "On Composite Marginal Likelihoods," *Advances in Statistical Analysis*, 92, 1–28. [259]

Yao, Y. C. (1984), "Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches," *The Annals of Statistics*, 12, 1434–1447. [258]

——— (1988), "Estimating the Number of Change Points by Schwarz Criterion," *Statistics and Probability Letters*, 6, 181–189. [256]

Zhang, N. R., and Siegmund, D. O. (2007), "A Modified Bayes Information Criterion With Applications to the Analysis of Comparative Genomic Hybridization Data," *Biometrics*, 63, 22–32. [256,260,261]