Developing a new BIC for detecting change-points

Gang Shen [a,*], Jayanta K. Ghosh [b,c]

[a] Department of Statistics, North Dakota State University, United States
[b] Department of Statistics, Purdue University, United States
[c] Indian Statistical Institute, India

ARTICLE INFO

Article history:
Received 7 June 2009
Received in revised form
9 April 2010
Accepted 25 October 2010

Keywords:
Change-point
Bayes factor
BIC
Laplace approximation

ABSTRACT

Usual derivation of BIC for the marginal likelihood of a model or hypothesis via Laplace approximation does not hold for a change-point which is a discrete parameter. We provide an analogue l BIC, which is a lower bound to the marginal likelihood of a model with change points and has an approximation error up to $O_p(1)$ like standard Schwartz BIC. Several applications are provided covering simulated r.v.'s and real financial figures on short-term interest rate.

1. Introduction

Changes in parameters in a model often characterize major events like a stock market crash, breakdown of quality or a suspected major latent change. Inference about the time of change, namely, a so-called change-point, has been the subject of considerable interest. For both review and original contributions based on likelihood ratio tests, see the excellent book by Csorgo and Horvath (1997) (henceforth abbreviated as [C, 1997]). Important contributions have also been made by Siegmund (1985). A major contribution to Bayesian inference on change-points is Raftery (1995), who develops Bayes Factors for tests about the null hypothesis of no change-point against an alternative of a change-point having occurred in the given data in the time interval under consideration.

The above formulation is what Moreno et al. (2005) refer to as a retrospective or offline analysis and distinguish from the sequential online problem in which one has to find an optimal stopping time to detect a change-point. Our focus in this paper is on the first problem, more specifically, on a new rigorous approximation l BIC to the Bayes Factor, valid up to $O_p(1)$, as required by Schwarz (1978) but different from a naive use of the BIC developed by Schwartz in the same paper for exponential families. Our new approximation, which we will call l BIC because it is a lower bound valid up to $O_p(1)$, differs from a naive use of Schwartz's BIC because the change-point is a discrete parameter. As Schwartz pointed out, this sort of approximation up to an error of magnitude of $O_p(1)$ does not require specification of a prior. To some extent, this is an advantage because in testing problems there is less consensus about priors than in estimation problems, in spite of major efforts by Berger and Pericchi (1996) and the considerable progress as noted in Bayarri and Garcia-Donato (2007) for normal linear models. For non-normal exponential families such priors are still not well studied. Also our approximation via l BIC is easier to calculate than the marginal likelihood of the data under H_1. Finally, such a quantification of evidence makes sense to a classical statistician, for whom this provides a penalized likelihood ratio test, making it unnecessary to specify the type one error probability. We note that the standard theory for BIC based on Schwarz (1978) or Laplace approximation, e.g., Ghosh et al. (2006, Section 4, p. 114) is not applicable because the marginal likelihood of the data under the more complex model of an unspecified change-point involves not only integrals but also a sum, whereas, Laplace approximations for sums are not known.

The theoretical problem is formulated in Section 2 and solved for independent r.v.'s with distribution belonging to an exponential family. As expected, our new approximation leads to a penalty that is different from the standard Schwartz type penalty used somewhat naively for a change-point by Yao (1993). On the other hand our penalty is similar to that of Siegmund (2004), who derived his penalty in an entirely different way. A similar approach for AIC is given in Ninomiya (2005) who modifies the usual AIC penalty for a parameter like a change-point. Our methodology, but not the proof, has a straightforward extension to independent observations with models satisfying standard regularity conditions for asymptotic theory of MLE, e.g., Serfling (1985) or to GARCH models for financial asset price and volatility. We apply the new method in Section 5 to simulated as well as real data with satisfactory results. We also calculate several Bayes factors including that due to Bayarri and Garcia-Donato (2007) for comparison. Some general remarks appear in the final section.

2. Formulation of the problem

To keep the problem simple, suppose under the no change-point hypothesis H_0, x_i's, $1 \le i \le n$, are *iid* with density $f(x|\theta)$ satisfying standard regularity conditions. Under the one change-point hypothesis H_1, with change-point at k, $1<k<n$, $x_1,\dots x_n$ are independent, $x_1,\dots,x_k$ identically distributed $f(x|\theta_1)$ and $x_{k+1},\dots,x_n$ are identically distributed $f(x|\theta_2)$. Under H_1, we have two continuous parameters θ_1,θ_2 and a discrete parameter k, whereas under H_0, there is only the continuous parameter θ. Typically a Bayesian would introduce independent priors $\pi_1(k)$, $\pi_1(\theta_1,\theta_2)$ and $\pi_0(\theta)$, and compare the marginal likelihood

$$m(H_1)=\sum_{k=1}^{n-1}\left[\int\int\prod_{i=1}^{k}f(x_i|\theta_1)\prod_{i=k+1}^{n}f(x_i|\theta_2)\ \pi_1(\theta_1,\theta_2)\,d\theta_1\,d\theta_2\right]\pi_1(k),$$

$$m(H_0)=\int\prod_{i=1}^{n}f(x_i|\theta)\ \pi_0(\theta)\,d\theta$$

via the Bayes Factor $\lambda=m(H_1)/m(H_0)$ and reject H_0 if $\lambda>3$ (mild evidence) or $\lambda>10$ (strong evidence), vide Jeffreys (1961).

Approximating the integrals by standard of BIC or Laplace approximation for $x_1,\dots,x_k$ and $x_{k+1},\dots,x_n$, respectively, and dropping off the $O_p(1)$ term, we get a partial simplification

$$m(H_1)=\sum_{k=1}^{n}\exp\left\{\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-\frac{p_1}{2}\log n\right\}\pi_1(k), \tag{1}$$

where p_1 is the dimension of the smooth parameter under H_1, and

$$\exp\{\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))\}=\sup_{\theta_1,\theta_2}\left\{\prod_{i=1}^{k}f(x_i|\theta_1)\prod_{i=k+1}^{n}f(x_i|\theta_2)\right\}.$$

Here we have used the well-known Laplace approximation for each integral separately and simplified a little, getting

$$\iint\prod_{i=1}^{k}f(x_i|\theta_1)\prod_{i=k+1}^{n}f(x_i|\theta_2)\pi_1(\theta_1,\theta_2)\,d\theta_1\,d\theta_2=\exp\left\{\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-\frac{p_1}{2}\log n\right\}O_p(1).$$

Similarly, use Laplace approximation to a single integral and drop off the $O_p(1)$ term, we get

$$m(H_0)=\exp\left\{\hat{L}(\hat{\theta})-\frac{p_0}{2}\log n\right\}, \tag{2}$$

where p_0 is the dimension of the smooth parameter under H_0 and $\hat{\theta}$ is the MLE of θ under H_0. So, the BIC for H_1 with fixed k is $\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-(p_1/2)\log n$ and the BIC for H_0 is $\hat{L}(\hat{\theta})-(p_0/2)\log n$. A naive BIC, which treats the discrete change-point k as a regular smooth parameter and hence denoted as n BIC, is

$$n\mathrm{BIC}=\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1+1}{2}\log n, \tag{3}$$

where

$$\hat{k}=\underset{1\le k\le n}{\arg\max}\,\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)). \tag{4}$$

Our main theoretical goal is to further simplify the sum over k for $m(H_1)$ in Eq. (1), under H_1 with assuming the true value of the two continuous parameters are θ_1 and θ_2. Towards this end, we begin with an upper and lower bound of $m(H_1)$. First, we consider the case of the flat prior $\pi_1(k)=n^{-1}$, $\forall k$. Clearly, up to an error of magnitude $O_p(1)$,

$$\log m(H_1)\ge\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}-\log n, \tag{5}$$

$$\log m(H_1)\le\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}. \tag{6}$$

To obtain a much more tight upper bound for $\log m(H_1)$, we split $m(H_1)$ in Eq. (1) into four terms for some sufficiently large M, i.e.,

$$m(H_1) = L_1 + L_2 + L_3 + L_4$$

where

$$L_1 = \sum_{|k-k_0| < M} \exp\left[\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) - \frac{p_1}{2}\log n\right] n^{-1}, \tag{7}$$

$$L_2 = \sum_{k < M} \exp\left[\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) - \frac{p_1}{2}\log n\right] n^{-1}, \tag{8}$$

$$L_3 = \sum_{n-k < M} \exp\left[\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) - \frac{p_1}{2}\log n\right] n^{-1}, \tag{9}$$

$$L_4 = \sum_{|k-k_0|,k,n-k \geq M} \exp\left[\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) - \frac{p_1}{2}\log n\right] n^{-1}, \tag{10}$$

and k_0 is the change-point. Note that

$$L_1 \leq \exp\left[\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k})) - \frac{p_1}{2}\log n\right] n^{-1} 2M,$$

$$L_2 \leq \exp\left[\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k})) - \frac{p_1}{2}\log n\right] n^{-1} M,$$

$$L_3 \leq \exp\left[\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k})) - \frac{p_1}{2}\log n\right] n^{-1} M,$$

then $L_1 + L_2 + L_3 \leq \exp[\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k})) - (p_1/2)\log n] n^{-1} 4M$.

In Section 3, we consider the case of exponential family density $f_{\mathbf{x}}(x;\theta)$ and show that L_4 is at most in the same order of magnitude as $L_1+L_2+L_3$. Therefore, up to an error of magnitude $O_p(1)$,

$$\log m(H_1) \leq \log(L_1 + L_2 + L_3 + L_4) = \left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k})) - \frac{p_1}{2}\log n\right\} - \log n.$$

Hence, for the case of exponential family, the lower bound in Eq. (5) as an approximation to $\log m(H_1)$, has an error of magnitude $O_p(1)$ which is the same as the order of error of the BIC in regular problems with continuous parameters. We choose the lower bound as our BIC for the case of exponential family with flat prior, and denote it as l BIC to distinguish from the n BIC. Clearly, l BIC is more conservative for the test. The fact that use of l BIC is also consistent, vide Section 4, provides further justification.

The above justification holds not only for the flat prior $\pi_1(k)$, but also for a quite general class of priors $\pi_1(k)$ satisfying the assumption B1 in Section 3, which is a discretization of density function. For these reasons, we choose the lower bound in Eq. (5) as our l BIC for H_1 for the case of exponential family in all cases. It leads to a parsimonious test which is close to the full Bayes test for many priors.

3. New BIC for change-point of the exponential family

Here, as in [C, 1997], we consider the case of exponential family with density $f_{\mathbf{x}}(x;\theta) = h(x)\exp[T(x)\theta - A(\theta)]$, where θ is the natural parameter. For a change-point problem of this exponential family, without loss of generality, one may assume the natural parameter space Θ is open. It is well known that $A(\theta)$ is infinitely differentiable, e.g., Lehmann (1991, Section 2, Theorem 9). It is also well known that this family satisfies

1. $E|T(\mathbf{x})| < \infty$, for all $\theta \in \Theta$;
2. $A(\theta)$ is strictly convex;
3. $A'(\theta)$, the derivative of $A(\theta)$, has a unique inverse $(A')^{-1}(\theta)$;

vide Bickel and Doksum (2007, Theorem 1.6.2–1.6.4, Section 1). Therefore, both $\hat{\theta}_1(k)$ and $\hat{\theta}_2(k)$ have the closed form, i.e.,

$$A'(\hat{\theta}_1(k)) = k^{-1}\sum_{i=1}^{k} T(\mathbf{x}_i), \quad A'(\hat{\theta}_2(k)) = (n-k)^{-1}\sum_{i=k+1}^{n} T(\mathbf{x}_i).$$

It is convenient to use the re-parameterized notation $\hat{\theta}_1^*(k) = A(\hat{\theta}_1(k))$ and $\hat{\theta}_2^*(k) = A(\hat{\theta}_2(k))$ instead.

Elementary calculation yields $\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) = kg(\hat{\theta}_1^*(k)) + (n-k)g(\hat{\theta}_2^*(k))$, where $g(\theta) = [(A')^{-1}(\theta)]\theta - A((A')^{-1}(\theta))$. By the chain rule of derivative, $g''(\theta) = [(A')^{-1}(\theta)]'$. Hence, $g(\cdot)$ is also strictly convex as well as $A(\cdot)$, i.e.,

$$\inf_{t\in[a,b]} [tg(\theta_1) + (1-t)g(\theta_2) - g(t\theta_1 + (1-t)\theta_2)] > 0, \quad \forall [a,b] \subset (0,1).$$

Now, suppose k_0, or formally, $k_0(n)$ is the true unknown change-point for the parameter θ, i.e., $\{x_i\}_{i=1}^{k_0}$ iid $f(x;\theta_1)$ and $\{x_i\}_{i=k_0+1}^{n}$ iid $f(x;\theta_2)$, where θ_1 and θ_2 are the two distinct unknown parameters. In the change-point problem, $k_0(n)$ (sometimes simply written as k_0) is usually assumed to satisfy

A1. $0<\lim_{n\to\infty} n^{-1}k_0(n)=c_0<1$.

As for the prior $\pi_1(k)$, one may assume

B1. $\pi_1(k)=\int_{(k-1)n^{-1}}^{kn^{-1}}\phi(t)\,dt,\ k=1,2,\ldots,n,$

where $\phi(\cdot)$ is a pdf with support on [0,1] and $\phi(c_0)>0$. This guarantees $\pi_1(k_0)c>\max_k\pi_1(k)$ for sufficiently large n, where c is a constant independent of k or n. Other similar assumptions are possible.

With the assumptions A1, B1 above, one has the following basic lemma. The proof is given in the Appendix.

Lemma 1. *For $\forall\varepsilon>0$, there exists some constant $M_0(\varepsilon)$ (independent of n, k, hereafter written as M_0), such that for $\forall(n,k)\in\mathcal{D}^1(M_0)\cup\mathcal{D}^2(M_0)$, where*

$$\mathcal{D}^1(M_0)=\{(n,k):\ n-k_0>M_0,k_0-k>M_0,k>M_0\},$$

$$\mathcal{D}^2(M_0)=\{(n,k):\ k_0>M_0,k-k_0>M_0,n-k>M_0\},$$

with a probability larger than $1-\varepsilon$, one has

$$\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-\hat{L}(k_0,\hat{\theta}_1(k_0),\hat{\theta}_2(k_0))\le-\delta|k-k_0|,$$

for some constant $\delta>0$, δ independent of n, k and ε.

Remark 1. δ is explicitly defined by Eq. (18) in the proof. M_0 is implicitly defined in the proof while using the Law of Iterated Logarithm (LIL) and inequality (A.5).

Remark 2. $\mathcal{D}^1(M_0),\mathcal{D}^2(M_0)\subseteq\{k_0>M_0,n-k_0>M_0\}$, which is equivalent to $\{n:n\ge N_0\}$ for some constant N_0, by virtue of the assumption A_1. Clearly, N_0 depends on M_0 only.

To apply this lemma for estimation of the L_4 term in Eq. (10), we first let $n>N_0$, then for fixed n (so that k_0 is also fixed), we consider all the k satisfying $k_0-k>M_0$, $k>M_0$ as in $\mathcal{D}^1(M_0)$ and also all the k satisfying $k-k_0>M_0$, $n-k>M_0$ as in $\mathcal{D}^2(M_0)$, respectively.

The crux of the Laplace approximation for the integral $\int_{-\infty}^{+\infty}q(\theta)e^{nh(\theta)}\,d\theta$ is that $h(\theta)$ takes a sharp maximum within the neighborhood of θ_0, where $\theta_0=\operatorname{argmax}_\theta h(\theta)$, so that the integral over the entire domain is reduced to the integral over a neighborhood of θ_0. The lemma above is analogous to this. For Eq. (7), the lemma asserts that $\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))$ has a sharp maximum value in a neighborhood of k_0, so that the sum of $\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))$ over all k is reduced to the sum over a neighborhood of k_0.

Clearly, this lemma implies $|\hat{k}-k_0|=O_p(1)$, the conclusion of Lemma 1.5.1 in [C, 1997], Section 1. Therefore, together with assumption B1, $\forall\varepsilon>0$, $\exists N_1(\varepsilon)$, such that $\forall n>N_1(\varepsilon)$, with a probability larger than $1-\varepsilon$, one has $\pi(\hat{k})c\ge\max_k\pi(k)$. (Hereafter, we also denote $N_1(\varepsilon)$ as N_1.)

Let $N=\max\{N_0,N_1\}$. This lemma also implies that for $n>N$ and then $k\in\{k_0-k>M_0,k>M_0\}\cup\{k-k_0>M_0,n-k>M_0\}$, with a probability larger than $1-\varepsilon$

$$\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))\le-\delta|k-k_0|.$$

Therefore, by Eq. (10) and some elementary calculation, with a probability larger than $1-\varepsilon$,

$$\begin{aligned}L_4&=\left\{\sum_{k_0-k,k\ge M_0}+\sum_{k-k_0,n-k\ge M_0}\right\}\exp\left\{\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-\frac{p_1}{2}\log n\right\}\pi_1(k)\\&\le\exp\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}\pi_1(\hat{k})c\sum_{|k-k_0|\ge M_0}\exp(-\delta|k-k_0|)\\&\le\exp\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}\pi_1(\hat{k})c\exp(-\delta M_0)\delta(1-\exp(-\delta))^{-1}.\end{aligned}$$

Of course, as discussed in the flat prior case in Section 2, with a probability larger than $1-\varepsilon$, one has

$$L_1=\sum_{|k-k_0|<M_0}\exp\left\{\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))-\frac{p_1}{2}\log n\right\}\pi_1(k)\le\exp\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}\pi_1(\hat{k})2cM_0.$$

and the similar upper bound for L_2 and L_3, respectively.

All the above reasoning implies that $\forall\varepsilon>0$, $\exists N$, which depends on ε only, such that $\forall n>N$, with probability larger than $1-\varepsilon$, one has

$$\log m(H_1)\le\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}+\log\pi_1(\hat{k})+M_0^*,$$

where $M_0^*=\log\{4cM_0+c\exp(-\delta M_0)\delta(1-\exp(-\delta))^{-1}\}$. But, by Eq. (1), one has the lower bound for $m(H_1)$, i.e.,

$$\log m(H_1)\ge\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}+\log\pi_1(\hat{k})+O_p(1) \tag{11}$$

Therefore, one has a further simplification for $m(H_1)$, i.e.,

$$\log m(H_1)=\left\{\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n\right\}+\log\pi_1(\hat{k})+O_p(1). \tag{12}$$

We state this as a theorem for future reference:

Theorem 1. *With the assumption* A1, B1 *specified above, the new BIC for* H_1 *with unknown* $k_0(n)$ *for the exponential family, namely,*

$$l\mathrm{BIC}=\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1}{2}\log n+\log\pi_1(\hat{k}). \tag{13}$$

satisfies $\log m(H_1)=l\mathrm{BIC}+O_p(1)$.

As explained earlier, l is a reminder that it is an approximation to $m(H_1)$ through a lower bound. Clearly a flat prior $\pi_1(k)$ satisfies the assumption B1. If one further requires $\phi(\cdot)>0$ in assumption B1, then $\pi_1(k)$ has the same order as n^{-1}, for $\forall k=1,2,\ldots,n$. Thus, our new BIC for the exponential family with change-point in its parameter θ becomes

$$l\mathrm{BIC}=\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1+2}{2}\log n.$$

It is different from the naive BIC in this setting which is equal to

$$n\mathrm{BIC}=\hat{L}(\hat{k},\hat{\theta}_1(\hat{k}),\hat{\theta}_2(\hat{k}))-\frac{p_1+1}{2}\log n.$$

We think this is due to the fact that the change-point k_0 is not a regular parameter, as discussed in Ninomiya (2005), where the bias-corrected term for AIC in the presence of change-point was treated differently.

As for the case of multiple change-points, similar results can be achieved by replacing the assumption A1, B1 with the following two assumptions:

A2. the change-points $k_0(n)<k_1(n)<\cdots<k_{q-1}(n)$ satisfy

$$\lim_{n\to\infty}n^{-1}k_i(n)=c_i,\ \forall\ i=0,1,\ldots,q-1,$$

where $0<c_0<c_1<\cdots<c_{q-1}<1$;

B2. denote $c^*=(c_0,c_1,\ldots,c_{q-1})$, $\mathbf{l}=(l_0,l_1,\ldots,l_{q-1})$, where $1<l_0<l_1<\cdots<l_{q-1}\le n$. The prior $\pi_1(\mathbf{l})$ for the q change-points satisfies

$$\pi_1(l)=\int_{D_1}(q!)\phi(t)\,dt,$$

where $\phi(\cdot)$ is a pdf with support on $[0,1]^q$ and $\phi(c^*)>0$; $D_\mathbf{l}$ is the q-dimension hypercube $\prod_{i=0}^{q-1}((l_i-1)/n,l_i/n]$.

Following the same approach showed earlier in this section, one may obtain the new BIC for the multiple change-points problem in this case, i.e.,

$$l\mathrm{BIC}=\hat{L}(\hat{\mathbf{k}},\hat{\theta}_1(\hat{\mathbf{k}}),\hat{\theta}_2(\hat{\mathbf{k}}),\ldots,\hat{\theta}_{q+1}(\hat{\mathbf{k}}))-\frac{p}{2}\log n+\log\pi_1(\hat{\mathbf{k}}), \tag{14}$$

where $\hat{\mathbf{k}}=\arg\max_{\mathbf{k}}\hat{L}(\mathbf{k},\hat{\theta}_1(\mathbf{k}),\hat{\theta}_2(\mathbf{k}),\ldots,\hat{\theta}_{q+1}(\mathbf{k}))$, p is the number of smooth parameters under H_1 and q is the number of change-points under H_1. If one further requires $\phi(\cdot)>0$ in assumption B2, then $\pi_1(\mathbf{k})$ has the same order as n^{-q}, for $\forall\mathbf{k}$. Therefore,

$$l\mathrm{BIC}=\hat{L}(\hat{\mathbf{k}},\hat{\theta}_1(\hat{\mathbf{k}}),\hat{\theta}_2(\hat{\mathbf{k}}),\ldots,\hat{\theta}_{q+1}(\hat{\mathbf{k}}))-\left(\frac{p}{2}+q\right)\log n.$$

4. Consistency of the l BIC

For one change-point problem of the exponential family, under H_0, the difference between $\log m(H_0)$ and l BIC is just half of the frequentist's log likelihood ratio test statistic plus $((p_0-p_1)/2)\log n+\log\pi_1(\hat{k})$. The upper bound for this log likelihood ratio test statistic is up to the order of log log n, vide [C, 1997], Section 1, Eq (1.1.1), (1.1.19) and (1.3.6). This fact implies, under H_0, our l BIC is consistent. On the other hand, under H_1, elementary calculation yields

$$l\text{BIC}-\log m(H_0) \geq \{\hat{L}(k_0,\hat{\theta}_1(k_0),\hat{\theta}_2(k_0))-\hat{L}(\hat{\theta})\}+\frac{p_0-p_1}{2}\log n+\log\pi_1(\hat{k})$$
$$\geq n\left\{\frac{k_0}{n}g(\theta_1^*)+\left(1-\frac{k_0}{n}\right)g(\theta_2^*)-g\left(\frac{k_0}{n}\theta_1^*+\left(1-\frac{k_0}{n}\right)\theta_2^*\right)\right\}+\frac{p_0-p_1}{2}\log n+\log\pi_1(\hat{k}).$$

By the strictly convexity of $g(\cdot)$ for the exponential family and the assumption B1, this difference is $O_p(n)$ with positive coefficient. Therefore, our l BIC is consistent under H_1 as well. The consistency of our l BIC can be easily extended to multiple change-points problem for the exponential family.

The proof above also shows that, as expected, use of the exact Bayes Factor, namely $\lambda = m(H_1)/m(H_0)$, also leads to consistency of BIC. The consistency of BIC for hypotheses about exponential families was proved by Schwarz (1978). Berger et al. (2003) show BIC is inconsistent in high dimensional problems.

5. Application to simulated and real data

In this section, we extend the methodology, but not the proof, to simulated observations from exponential distribution, normal distribution, GARCH(1,1) model and to the weekly observations on the 3-month treasury bill yields. We also compared the Bayes factor via our l BIC with that via conventional prior proposed in Bayarri and Garcia-Donato (2007).

5.1. Case of simulated data

In the numerical study, we first simulated data from three distributions $\exp(\theta)$, $N(\mu,\sigma)$ and GARCH(1,1), respectively, each with n=2000 observations and with a change-point k_0= 1001. In such a case $1-k_0(n)/n$= 0.5. To be specific,

1. for $\exp(\theta)$, the first half observations *iid* exp(1) and the remaining half observations *iid* exp(2);
2. for $N(\mu,\sigma)$, the first half observations *iid* from N(1,1) and the remaining half observations *iid* from N(2,2);
3. for Garch(1,1), the first half observations from GARCH(1,1)

$$\begin{cases} x_k = 1+v_k, \\ v_k = \sigma_k\varepsilon_k, \\ \sigma_k^2 = 0.1+0.1v_{k-1}^2+0.8\sigma_{k-1}^2 \end{cases}$$

and the remaining half observations from GARCH(1,1)

$$\begin{cases} x_k = 2+v_k, \\ v_k = \sigma_k\varepsilon_k, \\ \sigma_k^2 = 0.4+0.1v_{k-1}^2+0.8\sigma_{k-1}^2, \end{cases}$$

where $\varepsilon_k \overset{iid}{\sim} N(0,1)$.

Clearly, the exponential distribution $\exp(\theta)$ has density function $f_x(x) = \exp\{-x\theta+\log\theta\}$, so it is a member of the exponential family in Section 3; the $N(\mu,\sigma)$ case with both μ and σ unknown is a member of another standard exponential family; and the GARCH(1,1) case does not belong to the exponential family in Section 3 either. As for the number of smooth parameters, $\exp(\theta)$ has p_0=1, p_1=2; $N(\mu,\sigma)$ has p_0=2, p_1=4 and the GARCH(1,1) has p_0=4, p_1=6.

In order to see the performance of our l BIC, we also simulated data from the above three change-point models with k_0= 1940, 1990 respectively. Table 1 summarizes some interesting statistics of the model fitted with the flat prior for the three cases. Clearly, our approximation of $m(H_1)$ through l BIC is quite satisfactory in each case except the case of GARCH(1,1) with the change-point very close to the end of the data, i.e., k_0= 1990. The much larger value of l BIC than log $m(H_0)$ leads support to

Table 1
Statistics of the fitted models for simulated data.

Model	k_0	$\hat{k}$	$\log m(H_0)$	$\log m(H_1)$	l BIC	n BIC
$\exp(\theta)$	1000	1000	−1459.27	−1346.44	−1347.52	−1343.72
	1940	1940	−1988.53	−1982.71	−1983.49	−1979.69
	1990	1980	−2054.06	−2057.14	−2059.41	−2055.61
$N(\mu,\sigma)$	1000	1002	−3808.52	−3505.46	−3506.97	−3503.17
	1940	1941	−2945.61	−2882.00	−2882.98	−2879.18
	1990	1989	−2877.37	−2860.36	−2861.25	−2857.45
Garch(1,1)	1000	998	−3661.49	−3559.46	−3560.77	−3557.22
	1940	1941	−2911.00	−2905.03	−2906.48	−2902.68
	1990	1393	−2775.99	−2781.37	−2786.56	−2782.76

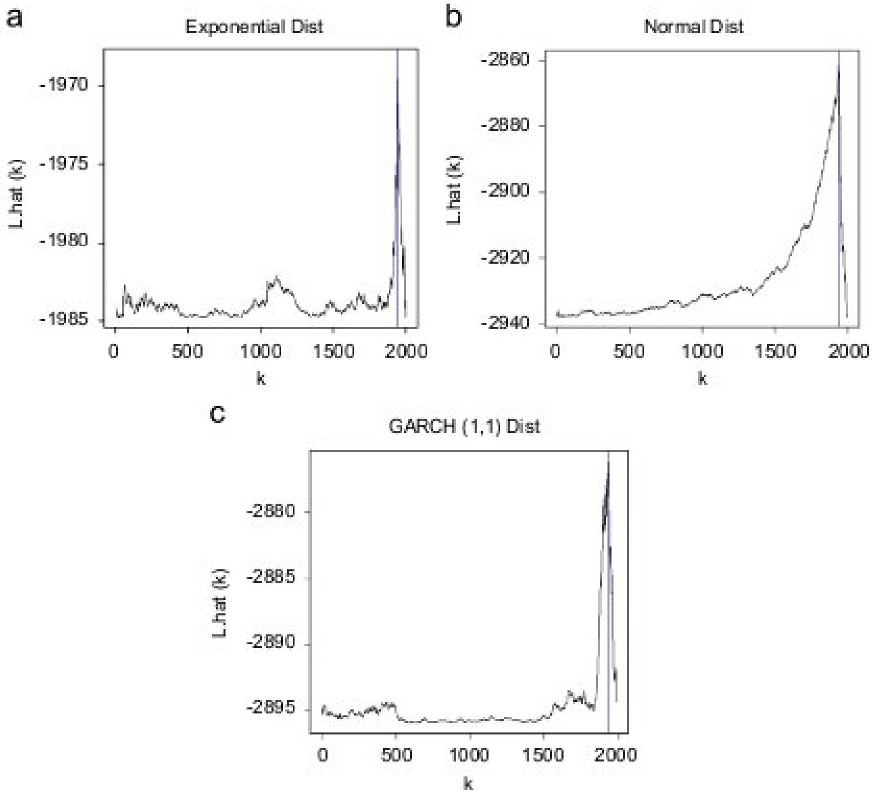

Fig. 1. The fitted log-likelihood for the models.

Table 2
Statistics of the fitted model for iid Exp(1).

$\hat{L}(\hat{k},\hat{\theta}_1,\hat{\theta}_2)$	$\log m(H_0)$	$\log m(H_1)$	l BIC	n BIC
−1991.35	−1996.98	−2000.57	−2006.55	−2002.75

the correct model H_1 in all these cases except the case of GARCH(1,1) with k_0=1990. Fig. 1 shows the fitted $\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))$ of the simulated data for all the three models with k_0=1940.

To verify the consistency of our l BIC, we also simulated 2000 observations *iid* from exp(1). Table 2 lists the statistics for the simulated data. Clearly, the much smaller value of l BIC than log $m(H_0)$ leads support to the correct model H_0.

Bayarri and Garcia-Donato (2007) suggested an objective proper prior for the change-point problem on normal linear models where a change occurs only in covariate coefficients at a fixed known change-point k_0. This setting is different from ours where the change-point k_0 is not assumed to be a known constant. However, one may also use their prior as an alternative to the Laplace approximation to compute the marginal likelihood and then, likewise, apply flat prior $\pi_1(k)$ on the change-point to derive the Bayes factor. As to compare the logBF via our l BIC with that suggested by Bayarri and Garcia-Donato (2007) via their objective proper prior, we simulated data from the model of N(μ,1), with n=500 and change-point k_0=250, 475, 490 respectively.

1. for k_0=250, the first 250 observations *iid* N(1,1) and the remaining 250 observations *iid* N(2,1);
2. for k_0=475, the first 475 observations *iid* N(1,1) and the remaining 25 observations *iid* N(2,1);
3. for k_0=250, the first 490 observations *iid* N(1,1) and the remaining 10 observations *iid* N(2,1);

For each case, we computed the log Bayes factor $\log\lambda = \log m(H_1)/m(H_0)$ via the BIC approximation of $m(H_1)$ in Eq. (1) and $m(H_0)$ in Eq. (2), we denoted it as $\log\lambda^*$. The $\log\lambda$ via l BIC as a further approximation to $m(H_1)$ in Eq. (1) is also studied. We denoted this $\log\lambda$ as $\log\lambda^{*}$. The $\log\lambda$ suggested in Bayarri and Garcia-Donato (2007) is also included in this numerical study, denoted as $\log\lambda^{\circ}$. Table 3 lists the log Bayes factor for the simulated data of each case. Clearly, in all the three cases, $\log\lambda^*$, $\log\lambda^{*}$ and $\log\lambda^{\circ}$ are pretty close to each other.

Table 3
Log Bayes factor comparison.

k_0	$\log\lambda^*$	$\log\lambda^\star$	$\log\lambda^\circ$
250	51.70	50.87	50.12
475	3.36	1.15	2.86
490	1.41	−0.20	0.86

Table 4
Statistics of the fitted models for 3-month treasury bill yields, 1973/02–1982/10.

Model	$\hat{k}$	$L(\hat{k},\hat{\theta}_1,\hat{\theta}_2)$	$m(H_0)$	$m(H_1)$	l BIC	n BIC
LEVEL	370	−152.92	−189.70	−181.24	−184.05	−180.94
GARCH	342	−113.58	−142.14	−142.58	−144.72	−141.61

5.2. Case of weekly observations on the 3-month treasury bill yields

The short-term risk-free interest rate is fundamental to much of theoretical and empirical finance. There are two popular classes of empirical models for short-term interest rate volatility, namely LEVELS models and GARCH models. Brenner et al. (1996) indicate that LEVELS models over-emphasizes the dependence of volatility on interest level and fails to capture the serial correlation in conditional variance; on the other hand, GARCH models rely too heavily on serial correlation in variance and fails to capture the relations between interest rate level and volatility. In consideration of structural break or change-point, a model is preferred if it is robust to the occurrence of major events on market. Between 10/1979 and 10/1982, Federal Reserve did a monetary targeting experiment to lower the volatility of the short-term interest rate. Empirical study of a data set of weekly observations on the 3-month treasury bill yields starting from 02/09/1973 indicates that Fed's monetary targeting experiment might cause a structural break in LEVELS models, but not GARCH models, e.g., Ball and Torous (1994) or Brenner et al. (1996). However, Chan et al. (1992) found no structural break in the LEVELS models. We may tackle this issue with the Bayesian approach as discussed above for the data set of the 506 weakly observations, 02/1973–10/1982. For this case, the change-point k_0 if any, shall be the start time of Fed's experiment, corresponding to k_0=350. Assuming we do not know k_0 and the structural break occurring at k, the LEVELS model becomes

$$\begin{cases} r_i = \alpha_1 + \beta_1 r_{i-1} + v_i, \\ v_i = \sigma_i \varepsilon_i, \\ \sigma_i^2 = \phi_1^2 r_{i-1}^{2\gamma_1}, \end{cases}$$

for $i<k$ and

$$\begin{cases} r_i = \alpha_2 + \beta_2 r_{i-1} + v_i, \\ v_i = \sigma_i \varepsilon_i, \\ \sigma_i^2 = \phi_2^2 r_{i-1}^{2\gamma_2}, \end{cases}$$

for $i \geq k$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $i=1,\ldots,506$. Clearly, p_0=4 and p_1=8 in this case. The summary of the statistics of the model fitted with the flat prior is given in Table 4. $\hat{k}=370$ is a little bit far away from the value of k_0. This might be due to the fact that the volatility in LEVELS model relies too much on the level of interest rate and the interest rate level was relatively high when the Fed's experiment started, thus the alarming signal is delayed. However, our approximation for $m(H_1)$ through l BIC is quite satisfactory. The larger value of l BIC than $m(H_0)$ supports that there is a structural break in the LEVELS model for this case. Now we proceed to check whether there is a structural break in the ARMA-GARCH model for this data set, i.e.,

$$\begin{cases} r_i = \alpha_1 + \beta_1 r_{i-1} + v_i, \\ v_i = \sigma_i \varepsilon_i, \\ \sigma_i^2 = \omega_1 + a v_{i-1}^2 + b\sigma_{i-1}^2, \end{cases}$$

for $i<k$ and

$$\begin{cases} r_i = \alpha_2 + \beta_2 r_{i-1} + v_i, \\ v_i = \sigma_i \varepsilon_i, \\ \sigma_i^2 = \omega_2 + a v_{i-1}^2 + b\sigma_{i-1}^2, \end{cases}$$

for $i \geq k$, with $\varepsilon_i \overset{iid}{\sim} N(0,1)$. Clearly, for this case, p_0= 5 and p_1=8. The summary of the statistics of the model fitted with the flat prior is also given in Table 4. Our approximation for $m(H_1)$ through l BIC is quite satisfactory. The smaller value of l BIC or

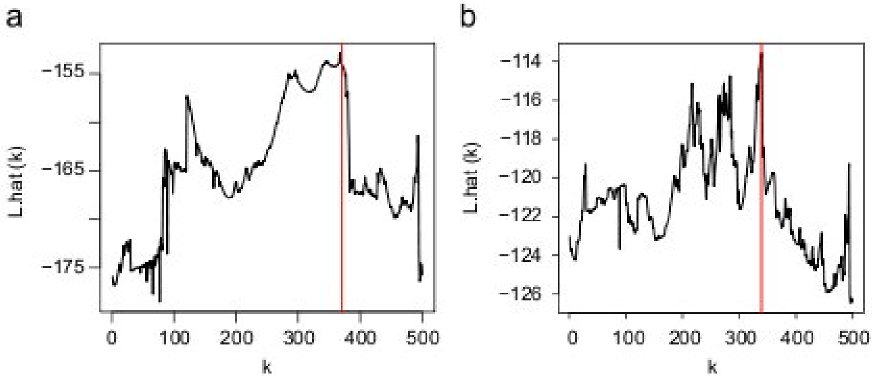

Fig. 2. The fitted log-likelihood for 3-month treasury bill yields, 02/1973–10/1982.

$m(H_1)$ than $m(H_0)$ suggests no structural break in the ARMA-GARCH model for this case. In contrast, one would conclude $m(H_1)$ if using n BIC instead. Our result confirms the conclusion of structural break about LEVELS models and GARCH models in Ball and Torous (1994) and Brenner et al. (1996). Fig. 2 shows the fitted $\hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k))$ of the real data for both models.

6. Concluding remarks

For the exponential family, our l BIC is a lower bound to the marginal likelihood of a model with change-points, hence more conservative. It has an approximation error up to $O_p(1)$ like standard Schwartz BIC. Like BIC approximation to the integral, marginal likelihood, l BIC holds for almost any nice priors for smooth parameters there, i.e., so long as the priors are continuous and positive. The assumption for our l BIC on the prior for the discrete change-point is also fairly weak, which requires a uniform prior or the discretization of a positive density. For this reason, our approximation is valid for a large families for priors. The numerical examples provided above also suggest that our approximation of BIC in the presence of change-point works very well in many cases, in addition to the exponential family as we proved. However, the theoretical justification of these extensions requires much more challenging theory.

Appendix A. Proof of Lemma 1

Like the approach in [C, 1997], Section 1.5.1, p. 41, our approach for Lemma 1 is to use Taylor expansion and the Law of Iterated Logarithm (LIL) to approximate the remainder

$$\begin{aligned}\Delta_k &= \hat{L}(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) - \hat{L}(k_0,\hat{\theta}_1(k_0),\hat{\theta}_2(k_0)).\\ &= [kg(\hat{\theta}_1(k)) + (n-k)g(\hat{\theta}_2(k))] - [k_0 g(\hat{\theta}_1(k_0)) + (n-k_0)g(\hat{\theta}_2(k_0))].\end{aligned} \tag{A.1}$$

Proof. Without loss of generality, we consider $k < k_0$ and let

$$T(\xi) = \frac{k_0-k}{n-k}T(\theta_1) + \frac{n-k_0}{n-k}T(\theta_2). \tag{A.2}$$

For convenience in notation, hereinafter we denote

$$\mathbf{x}_i^* = T(\mathbf{x}_i), \quad \xi^* = T(\xi), \quad \theta_1^* = T(\theta_1), \quad \theta_2^* = T(\theta_2) \tag{A.3}$$

and $\mathcal{D}(\theta_1^*,\theta_2^*) = [\min(\theta_1^*,\theta_2^*), \max(\theta_1^*,\theta_2^*)]$, then $\xi^* \in \mathcal{D}(\theta_1^*,\theta_2^*)$.

First of all, we give the explicit form of δ. Let

$$C_1 = \inf_{t\in[a,b]} \{tg(\theta_1^*) + (1-t)g(\theta_2^*) - g(t\theta_1^* + (1-t)\theta_2^*)\},$$

$$C_2 = g(\theta_1^*) - g(\theta_2^*) - g'(\theta_2^*)(\theta_1^* - \theta_2^*),$$

$$C_3 = \max_{\mathcal{D}(\theta_1^*,\theta_2^*)} \{|g'(\theta^*)|, |g''(\theta^*)|, |g'''(\theta^*)|\},$$

$$C_4 = \max\{C_3, C_3|\theta_1^* - \theta_2^*|\}.$$

Clearly, due to the strict convexity of $g(\cdot)$, $C_1, C_2 > 0$. Now we may define our constant

$$\delta = \min\{3^{-1}C_1, 6^{-1}C_2\}. \tag{A.4}$$

The constant C_4 above will later be used in defining M_0.

Before applying Taylor expansion on Δ_k in Eq. (A.1), we consider the upper bound for the following terms, which are closed related with the remainders in Taylor expansion later:

$$t_k = k^{-1}\sum_{1}^{k}(\mathbf{x}_i^* - \theta_1^*),$$

$$t_{k_0} = k_0^{-1}\sum_{1}^{k_0}(\mathbf{x}_i^* - \theta_1^*),$$

$$r_k = (n-k)^{-1}\sum_{k+1}^{n}(\mathbf{x}_i^* - \xi^*),$$

$$r_{k_0} = (n-k_0)^{-1}\sum_{k_0+1}^{n}(\mathbf{x}_i^* - \theta_2^*),$$

$$s_k = (k_0-k)^{-1}\sum_{k+1}^{k_0}(\mathbf{x}_i^* - \theta_1^*).$$

By LIL, $\forall \varepsilon > 0$, $\exists M_0, C > 0$, such that $\forall (n,k) \in \mathcal{D}^1(M_0)$, where

$$\mathcal{D}^1(M_0) = \{(n,k) : k > M_0, k_0-k > M_0, n-k_0 > M_0\},$$

with a probability larger than $1-\varepsilon$, one has

$$\max_{\mathcal{D}^1(M_0)} |[k_0-k]^{1/2}[2\log\log(k_0-k)]^{-1/2} s_k| \le C,$$

$$\max_{\mathcal{D}^1(M_0)} |[n-k_0]^{1/2}[2\log\log(n-k_0)]^{-1/2} r_{k_0}| \le C.$$

Clearly, the above result implies

$$\max_{\mathcal{D}^1(M_0)} |s_k|, \quad \max_{\mathcal{D}^1(M_0)} |r_{k_0}| \le C M_0^{-1/2}(2\log\log M_0)^{1/2}.$$

Note that $r_k = ((k_0-k)/(n-k))s_k + ((n-k_0)/(n-k))r_{k_0}$, therefore, $\forall k,n$ satisfying $k_0-k > M_0$ and $n-k_0 > M_0$, with a probability larger than $1-\varepsilon$, one has

$$\max_{\mathcal{D}^1(M_0)} |r_k| \le C M_0^{-1/2}(2\log\log M_0)^{1/2}.$$

Similarly, $\forall k,n$ satisfying $k > M_0$, with a probability larger than $1-\varepsilon$, one has

$$\max_{\mathcal{D}^1(M_0)} |t_k|, \ \max_{\mathcal{D}^1(M_0)} |t_{k_0}| \le C M_0^{-1/2}(2\log\log M_0)^{1/2}.$$

Actually, one may take M_0 sufficiently large such that further it satisfies

$$C M_0^{-1/2}(2\log\log M_0)^{1/2} \le 1, \quad 4C M_0^{-1/2}(2\log\log M_0)^{1/2} C_4 \le \delta. \tag{A.5}$$

Now Taylor expanding Δ_k in Eq. (A.1). For $\forall (n,k) \in \mathcal{D}^1(M_0)$, with a probability larger than $1-\varepsilon$, one has

$$g\left(k_0^{-1}\sum_{1}^{k_0}\mathbf{x}_i^*\right) = g(\theta_1^*) + g'(\theta_1^*)t_{k_0} + \frac{1}{2}g''(\theta_1^*)t_{k_0}^2 + o(k_0^{-1}),$$

$$g\left(k^{-1}\sum_{1}^{k}\mathbf{x}_i^*\right) = g(\theta_1^*) + g'(\theta_1^*)t_k + \frac{1}{2}g''(\theta_1^*)t_k^2 + o(k^{-1}),$$

$$g\left((n-k_0)^{-1}\sum_{k_0+1}^{n}\mathbf{x}_i^*\right) = g(\theta_2^*) + g'(\theta_2^*)r_{k_0} + \frac{1}{2}g''(\theta_2^*)r_{k_0}^2 + o((n-k_0)^{-1}),$$

$$g\left((n-k)^{-1}\sum_{k+1}^{n}\mathbf{x}_i^*\right) = g(\xi^*) + g'(\xi^*)r_k + \frac{1}{2}g''(\xi^*)r_k^2 + o((n-k)^{-1}).$$

Clearly, the approximations above are uniform in k, for $\forall(n,k)\in\mathcal{D}^1(M_0)$. Now plug the Taylor expansions above into Δ_k, then $\Delta_k = I_1+I_2+I_3+o(1)$, where

$$I_1=(n-k)g(\xi^*)-(n-k_0)g(\theta_2^*)-(k_0-k)g(\theta_1^*), \tag{A.6}$$

$$I_2=(n-k)g'(\xi^*)r_k-(n-k_0)g'(\theta_2^*)r_{k_0}-(k_0-k)g'(\theta_1^*)s_k, \tag{A.7}$$

$$I_3=\tfrac{1}{2}\{g''(\xi^*)(n-k)r_k^2+g''(\theta_1^*)kt_k^2-g''(\theta_2^*)(n-k_0)r_{k_0}^2-g''(\theta_1^*)k_0t_{k_0}^2\}. \tag{A.8}$$

First, we check I_1 in Eq. (A.6), the deterministic term of Δ_k. Clearly, by Taylor expansion and mean-value theorem, $\exists\eta^*$ such that

$$g(\xi^*)=g(\theta_2^*)+g'(\theta_2^*)(\xi^*-\theta_2^*)+\frac{g''(\eta^*)}{2}(\xi^*-\theta_2^*)^2.$$

Hence, by Eqs. (A.2) and (A.3)

$$g(\xi^*)=g(\theta_2^*)+g'(\theta_2^*)\left[\frac{k_0-k}{n-k}\right](\theta_1^*-\theta_2^*)+\frac{g''(\eta^*)}{2}\left[\frac{k_0-k}{n-k}\right]^2(\theta_1^*-\theta_2^*)^2.$$

Therefore, plug $g(\xi^*)$ in the above equation into Eq. (A.6), one has

$$I_1=-(k_0-k)\left[g(\theta_1^*)-g(\theta_2^*)-g'(\theta_2^*)(\theta_1^*-\theta_2^*)-\frac{g''(\eta^*)}{2}\frac{k_0-k}{n-k}(\theta_1^*-\theta_2^*)^2\right].$$

Note that, due to the convexity of $g(\cdot)$, $g(\theta_1^*)-g(\theta_2^*)-g'(\theta_2^*)(\theta_1^*-\theta_2^*)>0$, then by assumption A1, one may choose $\varepsilon_0>0$ such that for $\forall k$ further satisfying $(k_0-k)/(n-k)\le\varepsilon_0$, one has

$$\frac{g''(\eta^*)}{2}\frac{k_0-k}{n-k}(\theta_1^*-\theta_2^*)^2<\frac{1}{2}[g(\theta_1^*)-g(\theta_2^*)-g'(\theta_2^*)(\theta_1^*-\theta_2^*)].$$

Hence

$$[(n-k)g(\xi^*)-(n-k_0)g(\theta_2^*)-(k_0-k)g(\theta_1^*)]\le-\tfrac{1}{2}(k_0-k)[g(\theta_1^*)-g(\theta_2^*)-g'(\theta_2^*)(\theta_1^*-\theta_2^*)]. \tag{A.9}$$

Clearly, for the rest of k's, i.e., k satisfying $(k_0-k)/(n-k)>\varepsilon_0$, one has

$$\begin{aligned}(n-k)g(\xi^*)-(n-k_0)g(\theta_2^*)-(k_0-k)g(\theta_1^*)&=(n-k)\left[g(\xi^*)-\frac{n-k_0}{n-k}g(\theta_2^*)-\frac{k_0-k}{n-k}g(\theta_1^*)\right]\\&\le-(n-k)\inf_{t\in[a,b]}[tg(\theta_1^*)+(1-t)g(\theta_2^*)-g(t\theta_1^*+(1-t)\theta_2^*)].\end{aligned} \tag{A.10}$$

By Eqs. (A.9), (A.10) and (A.4), with a probability larger than $1-\varepsilon$, one has

$$I_1=\{(n-k)g(\xi^*)-(n-k_0)g(\theta_2^*)-(k_0-k)g(\theta_1^*)\}\le-3\delta|k-k_0|.$$

Second, we check I_2 in Eq. (A.7). By Taylor expansion and mean-value theorem, $\exists\tau^*$ such that

$$g'(\xi^*)=g'(\theta_2^*)+g''(\tau^*)(\xi^*-\theta_2^*).$$

Hence, by Eqs. (A.2) and (A.3)

$$(n-k)g'(\xi^*)r_k=(n-k)g'(\theta_2^*)r_k+(k_0-k)g''(\tau^*)(\theta_1^*-\theta_2^*)r_k.$$

Plug the above equation into Eq. (A.7) and apply the upper bound for $|s_k|,|r_k|$, with a probability larger than $1-\varepsilon$, one has

$$\begin{aligned}|I_2|&=|k_0-k|\,|[g'(\theta_2^*)-g'(\theta_1^*)]s_k-[g''(\tau^*)(\theta_2^*-\theta_1^*)]r_k|\le|k_0-k|\{|s_k|2C_4+|r_k|C_4\}\\&\le|k_0-k|\{3CM_0^{-1/2}(2\log\log M_0)^{1/2}C_4\}\le\delta|k_0-k|.\end{aligned}$$

The last step is by virtue of (A.5).

Finally, we check I_3 in Eq. (A.8). By Taylor expansion and mean-value theorem, $\exists\zeta^*$ such that

$$g''(\xi^*)=g''(\theta_2^*)+g'''(\zeta^*)(\xi^*-\theta_2^*).$$

Hence, by Eqs. (A.2) and (A.3)

$$(n-k)g''(\xi^*)r_k^2=(n-k)g''(\theta_2^*)r_k^2+(k_0-k)g'''(\zeta^*)(\theta_1^*-\theta_2^*)r_k^2.$$

Plug the above equation into Eq. (A.8), then

$$I_3=\tfrac{1}{2}g''(\theta_2^*)[(n-k)r_k^2-(n-k_0)r_{k_0}^2]+\tfrac{1}{2}g''(\theta_1^*)[kt_k^2-k_0t_{k_0}^2]+\tfrac{1}{2}g'''(\zeta^*)(\theta_1^*-\theta_2^*)(k_0-k)r_k^2. \tag{A.11}$$

Note that, by definition

$$(n-k)r_k-(n-k_0)r_{k_0}=(k_0-k)s_k,$$

$$k_0 t_{k_0} - k t_k = (k_0 - k) s_k,$$

$$|(n-k) r_k^2 - (n-k_0) r_{k_0}^2| \le |(n-k) r_k - (n-k_0) r_{k_0}||r_k + r_{k_0}| + |r_k r_{k_0}||k-k_0| = |s_k||r_k + r_{k_0}||k-k_0| + |r_k r_{k_0}||k-k_0|,$$

$$|k t_k^2 - k_0 t_{k_0}^2| \le |k t_k - k_0 t_{k_0}||t_k + t_{k_0}| + |t_k t_{k_0}||k-k_0| = |s_k||t_k + t_{k_0}||k-k_0| + |t_k t_{k_0}||k-k_0|.$$

Therefore, plug the above result in Eq. (A.11) and apply the upper bound for $|s_k|$, etc., with a probability larger than $1-\varepsilon$, one has

$$|I_3| \le |k-k_0|\{|s_k||r_k + r_{k_0}| + |r_k r_{k_0}|\}C_4 + \tfrac{1}{2}|k-k_0| r_k^2 C_4 \le |k-k_0|\{4[CM_0^{-1/2}(2\log\log M_0)^{1/2}]^2 C_4\} \le \delta|k-k_0|.$$

The last step is by virtue of (A.5).

The same conclusion holds for $\forall (n,k) \in \mathcal{D}^2(M_0)$, where

$$\mathcal{D}^2(M_0) = \{(n,k) : k_0 > M_0,\ k-k_0 > M_0,\ n-k > M_0\}.$$

Hence, for $\forall (n,k) \in \mathcal{D}^1(M_0) \cup \mathcal{D}^2(M_0)$, with a probability larger than $1-\varepsilon$, one has

$$\Delta_k = \hat{L}(k, \hat{\theta}_1(k), \hat{\theta}_2(k)) - \hat{L}(k_0, \hat{\theta}_1(k_0), \hat{\theta}_2(k_0)) \le -\delta|k-k_0|. \qquad \square \tag{A.12}$$

References

Ball, C.A., Torous, W.N., 1994. A stochastic volatility model for short term interest rates and the detection of structural shifts. Unpublished Manuscript, Anderson School of Management, University of California at Los Angeles.

Bayarri, M.J., Garcia-Donato, G., 2007. Extending conventional priors for testing general hypotheses in linear models. Biometrika 94 (1), 135–152.

Berger, J., Pericchi, L., 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the Acoustical Society of America 91 (433), 109–122.

Berger, J., Ghosh, L., Mukhopadhyay, N., 2003. Approximations and consistency of the Bayes factors as model dimension grows. Journal of Statistical Planning and Inference 112, 241–258.

Bickel, P.J., Doksum, K.A., 2007. Mathematical Statistics: Basic Ideas and Selected Topics, vol. 1, second ed. Pearson Education, Inc.

Brenner, R.J., Harjes, R.H., Kroner, K.F., 1996. Another look at models of the short-term interest rate. The Journal of Finance and Quantitative Analysis 31 (1), 85–107.

Chan, K.C., Karolyi, G.A., Longstaff, F.A., Sanders, A.B., 1992. An empirical comparison of alternative models of the short-term interest rate. Journal of Finance 47, 1209–1227.

Csorgo, M., Horvath, L., 1997. Limit Theorems in Change-Point Analysis. John Wiley & Sons.

Ghosh, J., Delampady, M., Samanta, T., 2006. An Introduction to Bayesian Analysis: Theory and Methods. Springer.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press, New York.

Lehmann, E.L., 1991. Testing Statistical Hypotheses, second ed. Wadsworth, Inc., Belmont, CA.

Moreno, E., Casella, G., Garsia-Ferrer, A., 2005. An objective Bayesian analysis of the change point problem. Stochastic Environmental Research and Risk Assessment 19, 191–204.

Ninomiya, Y., 2005. Information criterion for Gaussian change-point model. Statistics & Probability Letters 72, 237–247.

Raftery, A.E., 1995. Change point and change curve modeling in stochastic processes and spatial statistics. Journal of Applied Statistical Science 1 (4), 403–424.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Serfling, R.J., 1985. Approximation Theorems of Mathematical Statistics. John Wiley, New York.

Siegmund, D., 1985. Sequential Analysis: Tests and Confidence Intervals. Springer-Verlag GmbH.

Siegmund, D., 2004. Model selection in irregular problems: applications to mapping quantitative trait loci. Biometrika 91 (4), 785–800.

Yao, Q., 1993. Tests for change-points with epidemic alternatives. Biometrika 80 (1), 179–191.