Statistica Sinica **21** (2011), 539-569

A SIMPLE BAYESIAN APPROACH TO MULTIPLE CHANGE-POINTS

Tze Leung Lai and Haipeng Xing

Stanford University and SUNY at Stony Brook

Abstract: After a brief review of previous frequentist and Bayesian approaches to multiple change-points, we describe a Bayesian model for multiple parameter changes in a multiparameter exponential family. This model has attractive statistical and computational properties and yields explicit recursive formulas for the Bayes estimates of the piecewise constant parameters. Efficient estimators of the hyperparameters of the Bayesian model for the parameter jumps can be used in conjunction, yielding empirical Bayes estimates. The empirical Bayes approach is also applied to solve long-standing frequentist problems such as significance testing of the null hypothesis of no change-points versus multiple change-point alternatives, and inference on the number and locations of change-points that partition the unknown parameter sequence into segments of equal values. Simulation studies of performance and an illustrative application to the British coal mine data are also given. Extensions from the exponential family to general parametric families and from independent observations to genearlized linear time series models are then provided.

Key words and phrases: Empirical Bayes, exponential families, generalized linear autoregressive models, multiple change-points, segmentation.

1. Introduction

We consider herein multiple change-point problems based on independent observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ such that $\mathbf{y}_i$ is a $d \times 1$ vector with density function $f_{\boldsymbol{\theta}_i}$, in which the $\boldsymbol{\theta}_i$ are unknown parameters that are piecewise constant. There is an extensive literature on the case in which the $\boldsymbol{\theta}_i$ can undergo at most one change, for which the frequentist approach dates back to the seminal works of Page (1955), Quandt (1958, 1960) and Hinkley (1970), while the Bayesian approach dates back to Shiryaev (1963). Carlin, Gelfand, and Smith (1992) review subsequent developments and propose a hierarchical Bayesian model and an associated Gibbs sampler. Extension to the multiple change-point setting has been hampered by the computational complexity of the problem. Several tractable models and computational methods have been developed in the literature to address these issues and the closely related matter of determining the number of change-points. For the frequentist approach, Bai (1997a,b), Bai and Perron

(1998, 2003), and Qu and Perron (2007) consider regression models with multiple change-points, using dynamic programming to compute the least squares estimates of the piecewise constant regression parameters when it is assumed that there are $k(\geq 2)$ change-points. An alternative approach that is computationally more convenient, especially when k is not small, is the binary segmentation procedure of Vostrikova (1981) and its recent refinement by Olshen et al. (2004). The choice of k for this approach is carried out by a model selection criterion dating back to Yao (1988), who simply applied Schwarz's Bayesian Information Criterion (BIC). However, for change-point problems, the likelihood functions do not satisfy the regularity conditions that are needed to derive the BIC, as noted by Siegmund (2004) and Zhang and Siegmund (2006), who propose modifications of the BIC for change-point problems. Earlier, Birgé and Massart (2001), Broman and Speed (2002), and Lavielle (2005) have used penalized likelihood methods that involve a shrinkage-type parameter to be chosen by the user, e.g., by cross validation. Davis, Lee, and Rodriguez-Yam (2006) use the minimum description length principle to choose p and the number and locations of change-points in their piecewise autoregressive AR(p) models.

The Bayesian approach to multiple change-points dates back to the seminal paper of Chernoff and Zacks (1964). McCulloch and Tsay (1993) extended the Chernoff-Zacks model of normal mean shift to Gaussian autoregressive models with possible changes in level and error variance, and used the Gibbs sampler to approximate the posterior distribution of the time-varying parameters. Barry and Hartigan (1992, 1993) proposed a product partition model as the prior distribution for the sequence of the piecewise constant parameters and used the Gibbs sampler to approximate the posterior means of the parameters. Subsequent developments of the Bayesian approach make use of reversible jump Markov chain Monte Carlo (MCMC) introduced by Green (1995), or Gibbs sampling used in conjunction with Metropolis-Hastings steps, as in Albert and Chib (1993), Chib (1998), Liu and Lawrence (1999), Wang and Zivot (2000), Chib, Nardari and Shephard (2002). In particular, the reversible jump MCMC extends the Metropolis-Hastings method to include jumps between parameter spaces of different dimensions. All these methods assume conjugate priors for the prior distribution of parameters and provide simulation-based inference via MCMC algorithms.

Section 2 considers a multiparameter exponential family of density functions $f_{\boldsymbol{\theta}}(\mathbf{y}) = \exp\{\boldsymbol{\theta}'\mathbf{y} - \psi(\boldsymbol{\theta})\}$ with respect to some measure ν on $\mathbb{R}^d$, and introduces a Bayesian model for multiple change-points in the multiparameter exponential family. In contrast with the aforementioned Bayesian models that require MCMC implementation, explicit recursive formulas for the Bayes estimates of the piecewise constant parameters are available for our Bayesian model and are given in

Section 2, which also describes how the hyperparameters of the Bayesian model can be estimated. Section 3 reports simulation studies of the performance, from both frequentist and Bayesian viewpoints, of these estimates of the piecewise constant parameter vectors in a multinomial distribution. In Section 4 we use the empirical Bayes estimates of the piecewise constant parameters to develop procedures with attractive statistical and computational properties for such challenging frequentist problems as segmentation and significance testing of the null hypothesis of no change-points versus multiple change-point alternatives. We also review other segmentation and testing procedures in the literature and carry out simulation studies comparing their performance with ours, from both statistical and computational viewpoints.

To illustrate the proposed methodology, we apply it in Section 5 to the time series of annual numbers of coal-mine disasters between 1851 and 1962, which has been analyzed by different methods in the change-point literature. We also compare our results with those obtained by more complex Bayesian change-point models that are implemented by MCMC, and conduct further simulation studies for sensitivity analysis of our results under various fitted change-point models. Although we have focused so far on independent random vectors from an exponential family, because of analytic tractability, we can easily extend the proposed Bayesian change-point model to much more general parametric families of time series models, thereby greatly broadening its applicability. The details are given in Section 6, where we also generalize previous work of McCulloch and Tsay (1993) and Lai, Liu, and Xing (2005) on Gaussian autoregressive models to change-point generalized linear autoregressive models.

Although our approach assumes a parametric model and superimposes on it a Bayesian change-point model, it uses the assumed model only as a *working model* to derive the Bayesian smoothers (which involve model averaging) and the frequentist segmentation procedures (which involve model selection). Quite remarkably, this working model is able to tackle both the Bayesian and the frequentist problems efficiently, as shown in Section 4. Moreover, although the working model uses a parametric formulation to perform model averaging, it leads to estimates and tests that compare favorably to nonparametric change-point procedures that do not assume parametric working models, as shown in Section 4. Further discussion of this and other issues and some concluding remarks are given in Section 7.

2. A Bayesian Model for Multiple Change-points in Exponential Families

Consider a multiparameter exponential family of densities

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = \exp\{\boldsymbol{\theta}'\mathbf{y} - \psi(\boldsymbol{\theta})\} \tag{2.1}$$

with respect to some measure ν on $\mathbb{R}^d$, and the prior density π (with respect to Lebesgue measure) on $\boldsymbol{\Theta} := \{\boldsymbol{\theta} : \int e^{\boldsymbol{\theta}'\mathbf{y}} d\nu(\mathbf{y}) < \infty\}$ given by

$$\pi(\boldsymbol{\theta}; a_0, \boldsymbol{\mu}_0) = c(a_0, \boldsymbol{\mu}_0) \exp\left\{a_0\boldsymbol{\mu}_0'\boldsymbol{\theta} - a_0\psi(\boldsymbol{\theta})\right\}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{2.2}$$

where $1/c(a_0, \boldsymbol{\mu}_0) = \int_{\boldsymbol{\Theta}} \exp\left\{a_0\boldsymbol{\mu}_0'\boldsymbol{\theta} - a_0\psi(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}$ and $\boldsymbol{\mu}_0 \in (\boldsymbol{\nabla}\psi)(\boldsymbol{\Theta})$, in which $\boldsymbol{\nabla}$ denotes the gradient vector of partial derivatives. The posterior density of $\boldsymbol{\theta}$ given the observations $\mathbf{y}_1, \ldots, \mathbf{y}_m$ drawn from $f_{\boldsymbol{\theta}}$ is

$$\pi\Big(\boldsymbol{\theta}; a_0 + m, \frac{(a_0\boldsymbol{\mu}_0 + \sum_{i=1}^m \mathbf{y}_i)}{(a_0 + m)}\Big); \tag{2.3}$$

see Diaconis and Ylvisaker (1979, p.274). Therefore, (2.2) is a conjugate family of priors and (2.3) shows that a_0 can be interpreted as an additional sample size associated with the prior and $\boldsymbol{\mu}_0$ is the prior mean of the $\mathbf{y}_i$. Moreover,

$$\int_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\mathbf{y})\pi(\boldsymbol{\theta}; a, \boldsymbol{\mu}) d\boldsymbol{\theta} = \frac{c(a, \boldsymbol{\mu})}{c(a+1, (a\boldsymbol{\mu} + \mathbf{y})/(a+1))}. \tag{2.4}$$

Suppose that, instead of being time-invariant, the parameter vector $\boldsymbol{\theta}_t$ may undergo occasional changes such that for $t > 1$, the indicator variables

$$I_t := \mathbf{1}_{\{\boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t-1}\}} \tag{2.5}$$

are independent Bernoulli random variables with $P(I_t = 1) = p$. When there is a parameter change at time t (i.e., $I_t = 1$), the changed parameter $\boldsymbol{\theta}_t$ is assumed to be sampled from π. The simplicity of the conjugate family (2.2)$-$(2.3) plays an important role in the explicit formulas for the sequential (filtering) estimates $E(\boldsymbol{\mu}_t|\mathcal{Y}_t)$ and for the fixed-sample (smoothing) estimates $E(\boldsymbol{\mu}_t|\mathcal{Y}_n)$, where $\boldsymbol{\mu}_t = \boldsymbol{\nabla}\psi(\boldsymbol{\theta}_t)$ and $\mathcal{Y}_t$ denotes $(\mathbf{y}_1, \ldots, \mathbf{y}_t)$. We also use $\mathcal{Y}_{i,j}$ to denote $(\mathbf{y}_i, \ldots, \mathbf{y}_j)$ for $i \leq j$.

2.1. Recursions for the filter $\boldsymbol{\theta}_t|\mathcal{Y}_t$

An important ingredient in the development of these explicit formulas is the most recent change-time K_t up to t, i.e., $K_t = \max\{s \leq t : I_s = 1\}$. Let $p_{it} = P(K_t = i|\mathcal{Y}_t)$. Denoting conditional densities by $f(\cdot|\cdot)$, note that

$$f(\boldsymbol{\theta}_t|\mathcal{Y}_t) = \sum_{i=1}^t p_{it} f(\boldsymbol{\theta}_t|\mathcal{Y}_{i,t}, K_t = i). \tag{2.6}$$

It follows from (2.3) that

$$f(\boldsymbol{\theta}_t|\mathcal{Y}_{i,t}, K_t = i) = \pi(\boldsymbol{\theta}_t; a_0 + t - i + 1, \bar{\mathbf{Y}}_{i,t}), \tag{2.7}$$

where $\bar{\mathbf{Y}}_{i,j} = (a_0\boldsymbol{\mu}_0 + \sum_{k=i}^{j} \mathbf{y}_k)/(a_0 + j - i + 1)$ for $j \geq i$. Combining (2.6) and (2.7) yields

$$f(\boldsymbol{\theta}_t|\mathcal{Y}_t) = \sum_{i=1}^{t} p_{it}\pi(\boldsymbol{\theta}_t; a_0 + t - i + 1, \bar{\mathbf{Y}}_{i,t}). \tag{2.8}$$

We next provide a recursive formula for p_{it} by noting that $\sum_{i=1}^{t} p_{it} = 1$ and

$$p_{it} \propto p_{it}^* := \begin{cases} pf(\mathbf{y}_t|I_t = 1) & \text{if } i = t, \\ (1-p)p_{i,t-1}f(\mathbf{y}_t|\mathcal{Y}_{i,t-1}, K_t = i) & \text{if } i \leq t-1. \end{cases} \tag{2.9}$$

Combining $f(\mathbf{y}_t|\mathcal{Y}_{i,t-1}, K_t = i) = \int f_{\boldsymbol{\theta}_t}(\mathbf{y}_t)f(\boldsymbol{\theta}_t|\mathcal{Y}_{i,t}, K_t = i)d\boldsymbol{\theta}_t$ with (2.1), (2.4), and (2.6) yields

$$p_{it}^* = \begin{cases} \frac{p\pi_{0,0}}{\pi_{t,t}} & \text{if } i = t, \\ (1-p)p_{i,t-1}\frac{\pi_{i,t-1}}{\pi_{i,t}} & \text{if } i < t, \end{cases} \tag{2.10}$$

where $\pi_{0,0} = c(a_0, \boldsymbol{\mu}_0)$ and $\pi_{i,j} = c(a_0 + j - i + 1, \bar{\mathbf{Y}}_{i,j})$.

2.2. Explicit formulas for $E(\boldsymbol{\theta}_t|\mathcal{Y}_n), 1 \leq t \leq n$

Following Yao (1984), who considered the case of univariate normal y_t with known variance 1, and Lai, Liu, and Xing (2005), who extended Yao's approach to the case where the variance of y_t is unknown and may also undergo jumps, and to change-point autoregressive models, we derive the posterior distribution of $\boldsymbol{\theta}_t|\mathcal{Y}_n$ by using Bayes' theorem to combine the forward filter $\boldsymbol{\theta}_t|\mathcal{Y}_t$ and the backward filter $\boldsymbol{\theta}_t|\mathcal{Y}_{t+1,n}$. The backward filter is obtained by reversing time, noting that the $\widetilde{I}_t = 1_{\{\boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t+1}\}}$ are still independent Bernoulli. Using the time-reversed counterpart $\widetilde{K}_t = \min\{s > t : \widetilde{I}_s = 1\}$ of K_t and $P(\boldsymbol{\theta}_t \in A|\mathcal{Y}_{t+1,n}) = \int P(\boldsymbol{\theta}_t \in A|\boldsymbol{\theta}_{t+1})dP(\boldsymbol{\theta}_{t+1}|\mathcal{Y}_{t+1,n})$, the backward (time-reversed) filter can be expressed as

$$f(\boldsymbol{\theta}_t|\mathcal{Y}_{t+1,n}) = p\pi(\boldsymbol{\theta}_t; a_0, \boldsymbol{\mu}_0) + (1-p)\sum_{j=t+1}^{n} q_{j,t+1}\pi(\boldsymbol{\theta}_t; a_0 + j - t, \bar{\mathbf{Y}}_{t+1,j}), \tag{2.11}$$

where $q_{jt} \propto q_{jt}^*$, $\sum_{j=t}^{n} q_{jt} = 1$ and

$$q_{j,t}^* = \begin{cases} \frac{p\pi_{0,0}}{\pi_{t,t}} & \text{if } j = t, \\ (1-p)q_{j,t+1}\frac{\pi_{t+1,j}}{\pi_{t,j}} & \text{if } j > t. \end{cases} \tag{2.12}$$

By Bayes' theorem,

$$f(\boldsymbol{\theta}_t|\mathcal{Y}_n) \propto \frac{f(\boldsymbol{\theta}_t|\mathcal{Y}_t)f(\boldsymbol{\theta}_t|\mathcal{Y}_{t+1,n})}{\pi(\boldsymbol{\theta}; a_0, \boldsymbol{\mu}_0)}. \tag{2.13}$$

Combining (2.11) with (2.8), and noting that

$$\pi\big(\boldsymbol{\theta}; a_0+t-i+1, \bar{\mathbf{Y}}_{i,t}\big)\frac{\pi\big(\boldsymbol{\theta}; a_0+j-t, \bar{\mathbf{Y}}_{t+1,j}\big)}{\pi\big(\boldsymbol{\theta}; a_0, \boldsymbol{\mu}_0\big)} = \frac{\pi_{it}\pi_{t+1,j}}{\pi_{ij}\pi_{00}}\pi\big(\boldsymbol{\theta}; a_0+j-i+1, \bar{\mathbf{Y}}_{ij}\big),$$

we obtain from (2.13) that

$$f(\boldsymbol{\theta}_t|\mathcal{Y}_n) = \sum_{1\le i\le t\le j\le n} \beta_{ijt}\pi(\boldsymbol{\theta}_t; a_0+j-i+1, \bar{\mathbf{Y}}_{i,j}), \tag{2.14}$$

where $\beta_{ijt} = \beta^*_{ijt}/P_t$, $P_t = p + \sum_{1\le i\le t<j\le n}\beta^*_{ijt}$, and

$$\beta^*_{ijt} = \begin{cases} pp_{it} & \text{if } i \le t = j, \\ (1-p)p_{it}q_{j,t+1}\frac{\pi_{it}\pi_{t+1,j}}{\pi_{ij}\pi_{00}} & \text{if } i \le t < j. \end{cases} \tag{2.15}$$

From (2.15), it follows that

$$P(I_{t+1}=1|\mathcal{Y}_n) = \frac{p}{P_t}, \qquad E(\boldsymbol{\mu}_t|\mathcal{Y}_n) = \sum_{1\le i\le t\le j\le n}\beta_{ijt}\bar{\mathbf{Y}}_{i,j}. \tag{2.16}$$

2.3. Estimation of hyperparameters for empirical Bayes approach

The Bayes estimates $E(\boldsymbol{\mu}_t|\mathcal{Y}_t)$ and $E(\boldsymbol{\mu}_t|\mathcal{Y}_n)$ involve the hyperparameters p, a_0, and $\boldsymbol{\mu}_0$, which are replaced by their estimates in the empirical Bayes approach. From the definition (2.9) of p^*_{it}, it follows that the likelihood function of p, a_0, and $\boldsymbol{\mu}_0$ is

$$\prod_{t=1}^n f(\mathbf{y}_t|\mathcal{Y}_{t-1}) = \prod_{t=1}^n\Big(\sum_{i=1}^t p^*_{it}\Big), \tag{2.17}$$

in which p^*_{it} is a function of p, a_0, and $\boldsymbol{\mu}_0$ given by (2.10). Since the $\mathbf{y}_t$ are exchangeable random vectors with mean $\boldsymbol{\mu}_0$ in the Bayesian model, we can estimate $\boldsymbol{\mu}_0$ by the sample mean $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{t=1}^n \mathbf{y}_t$. The hyperparameter a_0 is used to weight the sample mean $\widehat{\boldsymbol{\mu}}$ with the sample data between change-points in (2.8); we recommend the choice $a_0 = 1$, which can be interpreted as having an additional observation at $\widehat{\boldsymbol{\mu}}$ at a change-time when there is little information on the changed parameter. The important hyperparameter in the change-point model is the relative frequency p of change-points. Putting the above simple choice of the hyperparameters a_0 and $\boldsymbol{\mu}_0$ in (2.17), we can estimate p by maximizing the log-likelihood function $l(p) = \sum_{t=1}^n \log(\sum_{i=1}^t p^*_{it})$, which can be conveniently computed by grid search. For reasons that will be explained in the first paragraph of Section 4, the grid which we use to search for p has the form

$\{2^j/n : j_0 \leq j \leq j_1\}$, where $j_0 < 0 < j_1$ are integers; see the next section for an illustrative example.

3. Simulation Studies

This section presents two simulation studies of the performance, from both frequentist and Bayesian viewpoints, of the empirical Bayes estimates of the piecewise constant parameter vectors in a multinomial distribution $M(p_1, p_2, p_3, p_4)$ that corresponds to a 3-parameter exponential family with mean vector (p_1, p_2, p_3) and $p_4 = 1 - (p_1 + p_2 + p_3)$, and whose conjugate family of prior distributions is Dirichlet. We use the Bayesian change-point model of Section 2 as a working model. The first simulation study covers eight scenarios, only two of which are generated from the assumed Bayesian model, to evaluate the robustness of the empirical Bayes estimates. From each scenario, 500 samples of size $n = 1{,}000$ were generated to evaluate the performance of $\widehat{\mathbf{p}}_t$.

Scenario 1. The data were generated from a frequentist change-point model with two change-points at 301 and 701, and $\mathbf{p}_t = (0.1, 0.2, 0.3, 0.4)$ for $1 \leq t \leq 300$, $\mathbf{p}_t = (0.3, 0.3, 0.2, 0.2)$ for $301 \leq t \leq 700$, and $\mathbf{p}_t = (0.2, 0.1, 0.4, 0.3)$ for $701 \leq t \leq 1{,}000$.

Scenario 2. The data were generated from a frequentist change-point model with three change-points at 101, 301 and 701, and $\mathbf{p}_t = (0.1, 0.1, 0.4, 0.4)$ for $1 \leq t \leq 100$, $\mathbf{p}_t = (0.2, 0.3, 0.25, 0.25)$ for $101 \leq t \leq 300$, $\mathbf{p}_t = (0.4, 0.4, 0.1, 0.1)$ for $301 \leq t \leq 700$, $\mathbf{p}_t = (0.2, 0.1, 0.4, 0.3)$ for $701 \leq t \leq 1{,}000$.

Scenario 3. The data were generated from a frequentist change-point model with no change-points and $\mathbf{p}_t \equiv (0.2, 0.3, 0.2, 0.3)$.

Scenarios 4 *and* 5. The data were generated from the Bayesian change-point model, with $\mathbf{p}_t \sim$ Dirichlet(1,1,1,1) at each change-point t, $p = 0.002$ for Scenario 4 and $p = 0.02$ for Scenario 5.

Scenario 6. Instead of i.i.d. I_t in Scenarios 4 and 5, the actual Bayesian model assumed Markovian I_t defined by $P(I_t = 1 | I_{t-1} = 1) = 0.001$ and $P(I_t = 1 | I_{t-1} = 0) = 0.01$, and $\mathbf{p}_t \sim$ Dirichlet(1,1,1,1) at each change-point t.

Scenarios 7 *and* 8. The actual Bayesian model generating the data assumed $\mathbf{p}_t \sim$ Dirichlet$(1, 1, 1, 1)$ at each change-point t, and had two (for Scenario 7) and four (for Scenario 8) change-points that were independent and uniformly sampled from $\{1, \ldots, 1{,}000\}$.

We used the Kullback-Leibler (KL) divergence and the mean Euclidean error (EE) to assess the estimation error of the empirical Bayes estimate $\widehat{\mathbf{p}}_t$ of $\mathbf{p}_t$:

$$\mathrm{KL}(\mathbf{p}_t, \widehat{\mathbf{p}}_t) = E\Big\{ \sum_{i=1}^{4} p_{ti} \log \frac{p_{ti}}{\widehat{p}_{ti}} \Big\}, \qquad \mathrm{EE}(\mathbf{p}_t, \widehat{\mathbf{p}}_t) = E\Big\{ \sum_{i=1}^{4} (p_{ti} - \widehat{p}_{ti})^2 \Big\}^{1/2}. \quad (3.1)$$

Table 1. Smoothing estimates under eight scenarios

Scenario		1	2	3	4	5	6	7	8
KL	Bayes	0.0081	0.0120	0.0015	0.4814	0.6822	0.6528	0.0086	0.0147
		(1.40e-4)	(2.04e-4)	(5.71e-5)	(2.02e-2)	(1.62e-2)	(1.65e-2)	(1.78e-4)	(2.48e-4)
	BCMIX	0.0073	0.0120	0.0015	0.4810	0.6818	0.6528	0.0078	0.0133
		(1.61e-4)	(2.27e-4)	(5.70e-5)	(2.02e-2)	(1.62e-2)	(1.65e-2)	(1.93e-5)	(2.26e-4)
	Oracle	0.0046	0.0062	0.0015	0.5617	0.8372	0.7539	0.0047	0.0078
		(0.97e-4)	(1.14e-4)	(5.70e-5)	(2.52e-2)	(2.32e-2)	(2.16e-2)	(1.16e-4)	(1.53e-4)
EE	Bayes	0.0474	0.0556	0.0252	0.3300	0.4741	0.4551	0.0401	0.0530
		(5.03e-4)	(5.30ee-4)	(4.97e-4)	(8.18e-3)	(4.45e-3)	(5.47e-3)	(5.91e-4)	(5.78e-4)
	BCMIX	0.0482	0.0580	0.0247	0.3300	0.4742	0.4552	0.0411	0.0540
		(5.47e-4)	(6.58e-4)	(4.87e-4)	(8.17e-3)	(4.45e-3)	(5.70e-3)	(6.40e-4)	(6.28e-4)
	Oracle	0.0422	0.0463	0.0252	0.3600	0.4861	0.4637	0.0347	0.0443
		(4.76e-4)	(4.58e-4)	(4.97e-4)	(8.73e-3)	(4.32e-3)	(5.41e-3)	(4.86e-4)	(4.72e-4)

We also compared $\widehat{\mathbf{p}}_t$ with the "oracle" estimate $\mathbf{p}_t^*$ that assumes the change-points to be known and estimates the $\mathbf{p}_t$ in each known segment by maximum likelihood in Scenarios 1, 2 and 7 and by the posterior mean in the other scenarios. Table 1 gives the Monte Carlo estimates of $n^{-1}\sum_{t=1}^n \mathrm{KL}(\mathbf{p}_t, \widehat{\mathbf{p}}_t)$ and $n^{-1}\sum_{t=1}^n \mathrm{EE}(\mathbf{p}_t, \widehat{\mathbf{p}}_t)$ and their standard errors (in parentheses) for each of the eight scenarios; each result is based on 500 simulations. The results show that the Bayesian change-point working model gave estimates of the true signals that were sometimes even better than (because of the use of empirical Bayes rather than maximum likelihood estimation) and not much inferior to the oracle estimates. Table 1 also includes results for the BCMIX estimates, introduced in Section 4.2, showing that BCMIX is nearly Bayes in the Bayesian scenarios and may even be slightly better than Bayes in the frequentist scenarios. To illustrate the shape of the log-likelihood function $l(p)$, Figure 1 gives a plot of $l(p)$ for $0 < p \leq 0.03$ based on a sample of size $n = 1{,}000$ generated from Scenario 5, yielding the maximizer $\widehat{p} = 0.002$.

The second simulation study used 500 simulated samples of size $n = 1{,}000$ under Scenario 1, which is a frequentist rather than a Bayesian model, to investigate the sensitivity of the Bayes estimates that assume fixed p, to different choices of p. Here we considered Bayes rather than empirical Bayes estimates and assumed $\mathbf{p}_t \sim \mathrm{Dirichlet}(1, 1, 1, 1)$ at each change-point t for the Bayesian model. Table 2 shows that as p changes from 0.00025 to 0.032 over the grid $\{2^j/1{,}000\colon -2 \leq j \leq 5\}$, the KL divergence and the Euclidean error between true parameter $\mathbf{p}_t$ and the Bayes estimate $\widehat{\mathbf{p}}_t$ change little.

4. Applications to Segmentation and a Bootstrap Test

As noted in the second paragraph of Section 1, the frequentist approach to multiple change-point problems involves minimizing the sum of squared residu-

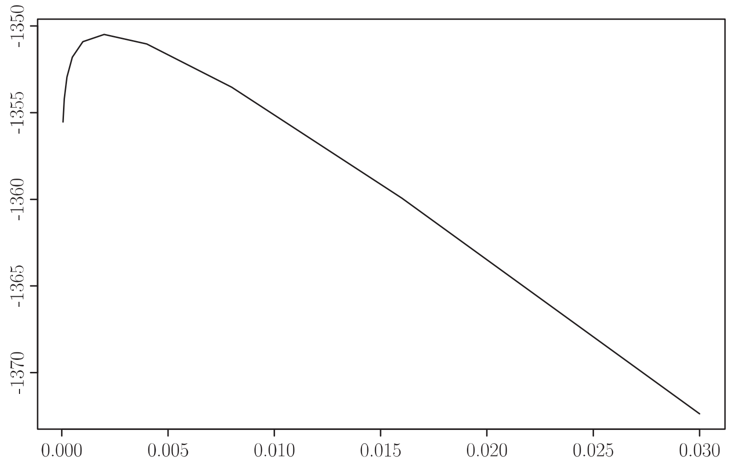

Figure 1. Log-likelihood a function of p.

Table 2. Smoothing estimates in Scenario 1 with different p's.

$p \times 10^3$	2^{-2}	2^{-1}	1	2	2^2	2^3	2^4	2^5
KL	0.0079	0.0077	0.0076	0.0076	0.0076	0.0079	0.0086	0.0107
	(2.10e-4)	(1.90e-4)	(1.74e-4)	(1.65e-4)	(1.62e-4)	(1.70e-4)	(1.95e-4)	(2.57e-4)
EE	0.0499	0.0494	0.0491	0.0491	0.0491	0.0497	0.0511	0.0553
	(6.25e-4)	(5.95e-4)	(5.68e-4)	(5.50e-4)	(5.41e-4)	(5.36e-4)	(5.52e-4)	(5.96e-4)

als or maximizing the log-likelihood over the locations of the change-points and the piecewise constant parameters when it is assumed that there are k change-points. This optimization problem can be solved by dynamic programming and constitutes only the inner loop of an algorithm whose outer loop is another minimization, over k, of a model selection criterion to determine k. Besides the computational complexity, there are additional complications in the frequentist approach to inference on change-points because the usual χ^2-approximations and other asymptotic properties of generalized likelihood ratio statistics or residual sum of squares no longer hold. The relative simplicity of the posterior distribution of $\boldsymbol{\theta}_t$ given $\mathcal{Y}_n$ in our Bayesian model opens up new possibilities in resolving some long-standing difficulties in the frequentist problems of testing for change-points and determining the segmentation. In this section we use an appropriately chosen hyperparameter p in our Bayesian model to tackle these frequentist problems. Note that the frequentist approach typically assumes that k is small relative to n and that adjacent change-points are sufficiently far apart so that the segments are identifiable except for relatively small neighborhoods of the change-points; see e.g., Bai and Perron (1998). Motivated by this assumption, in our Bayesian approach we restrict p to an interval $[c_1/n, c_2/n]$ for some positive constants $c_1 < c_2$ so that the arrival of change-points is approximately Poisson. The reason why we choose a grid of the form $\{2^j/n : j_0 \le j \le j_1\}$ instead of $\{j/n : j_0 \le j \le j_1\}$, say, is that as $p \to 0$, the behavior of $\widehat{\boldsymbol{\mu}}_{t|n}(ap) - \boldsymbol{\mu}_t$ is asymptotically equivalent to that of $\widehat{\boldsymbol{\mu}}_{t|n}(p) - \boldsymbol{\mu}_t$ for any given $a > 1$, as can be shown by an argument similar to that used in the Appendix to prove Theorem 2.

4.1. A bootstrap test for no change-points

To begin with, consider the simpler problem in which the $\mathbf{y}_t$ are i.i.d. with common density function $f_{\boldsymbol{\theta}}$. There is a one-to-one correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta})$, whose inverse function will be denoted by $\boldsymbol{\theta}(\boldsymbol{\mu})$, i.e., $\nabla\psi(\boldsymbol{\theta}(\boldsymbol{\mu})) = \boldsymbol{\mu}$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})$, where $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{t=1}^{n} \mathbf{y}_t$ as in Section 2.3. The classical generalized likelihood ratio (GLR) test of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ rejects this null hypothesis if the GLR statistic

$$\Lambda_n = \sum_{t=1}^{n} \log f_{\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})}(\mathbf{y}_t) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \sum_{t=1}^{n} \log f_{\boldsymbol{\theta}}(\mathbf{y}_t) \tag{4.1}$$

exceeds c, where the threshold c can be determined from the prescribed type I error probability by using the χ^2-approximation, with $d - \dim(\boldsymbol{\Theta}_0)$ degrees of freedom, of the null distribution of $2\Lambda_n$.

We next consider the more general setting in which $\mathbf{y}_t$ has parameter $\boldsymbol{\theta}_t$, and the null hypothesis assumes the $\boldsymbol{\theta}_t$ to be time-invariant, i.e., $H_0 : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_n$. In the univariate case $y_t \sim N(\theta_t, 1)$, James, James, and Siegmund (1987) studied the GLR test of H_0 versus the single change-point alternative $\theta_1 = \cdots = \theta_m \ne \theta_{m+1} = \cdots = \theta_n$ for some $m_0 \le m \le m_1$, with $m_0 \ge 1$ and $m_1 < n$ with m unknown. The GLR statistic involves $\max_{m_0 \le m \le m_1}$ and consequently its null distribution no longer has the χ^2-approximation. By developing a new approximation to the null distribution of the GLR statistic under certain assumptions on m_0 and m_1, they implemented the test and compared its performance to a score test proposed by Pettitt (1980) and another test proposed by Brown, Durbin and Evans (1975) based on recursive residuals. Bai and Perron (1998) considered the GLR test of H_0 versus the alternative that assumes k change-points at unknown locations $t_1^{(n)} < \cdots < t_k^{(n)}$. Computation of the GLR statistic in this case, denoted by GLR(k), involves dynamic programming, details of which are given by Bai and Perron (2003), who also extended the GLR statistic to $\max_{1 \le k \le K}$ GLR(k) for the more general alternative in which the number k of change-points is unknown but bounded above by K.

The computational complexity of the preceding GLR tests, even in the simple univariate $N(\theta_t, 1)$ case, is due to the complexity of the GLR statistics and the determination of the critical values, since standard χ^2 approximations are no longer applicable. We propose to use the Bayesian change-point model of Section

2 to provide a Bayesian counterpart of the GLR statistic for general multiple change-point alternatives. Note that the counterpart of the second term on the right hand side of (4.1) for the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_n (= \boldsymbol{\theta})$ is $\sup_{\boldsymbol{\theta}} \sum_{t=1}^n \log f_{\boldsymbol{\theta}}(\mathbf{y}_t) = \sum_{t=1}^n \log f_{\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})}(\mathbf{y}_t)$. Hence the Bayesian change-point model in Section 2, with the hyperparameter p estimated by maximum likelihood and $\boldsymbol{\mu}_0$ estimated by the sample mean $\widehat{\boldsymbol{\mu}}$ as in Section 2.3, suggests the statistic

$$L_n = \sup_{p \in \{2^j/n : j_0 \leq j \leq j_1\}} \sum_{t=1}^n l_t(p; a_0, \widehat{\boldsymbol{\mu}}) - \sum_{t=1}^n \log f_{\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})}(\mathbf{y}_t) \tag{4.2}$$

for testing H_0 versus a Bayesian change-point alternative parameterized by p, where $l_t(p; a_0, \widehat{\boldsymbol{\mu}}) = \log(\sum_{i=1}^t p_{it}^*)$, with p_{it}^* given by (2.10) that involves p, a_0, and $\boldsymbol{\mu}_0$ which is estimated by $\widehat{\boldsymbol{\mu}}$. Instead of maximizing $\sum_{t=1}^n l_t(p; a_0, \widehat{\boldsymbol{\mu}})$ over $0 < p < 1$ in (4.2), we maximize it over a grid of the form $\{2^j/n : j_0 \leq j \leq j_1\}$ as in Section 2.3, where $j_0 < 0 < j_1$ are integers. An important advantage of (4.2) over the GLR statistic $\max_{1 \leq k \leq K} \mathrm{GLR}(k)$ is its computational simplicity. The high-dimensional parameter space $\{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) : t \mapsto \boldsymbol{\theta}_t \text{ is piecewise constant}\}$ suggests that some regularization is needed to estimate the parameters, and putting a constraint K on the number of change-points, as in Bai and Perron (1998), can be regarded as regularization. Our Bayesian model "regularizes" by putting a stochastic structure, involving the parameter p, on the sequence $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$.

Although the null hypothesis H_0 is composite, it only involves the common value $\boldsymbol{\theta}$ of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, whose maximum likelihood estimate is $\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})$. Therefore we can use the parametric bootstrap to test H_0. Specifically, generate B bootstrap samples of independent random vectors $\mathbf{y}_{1,b}^*, \ldots, \mathbf{y}_{n,b}^*$ from $f_{\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})}$, and compute from each bootstrap sample the test statistic $L_{n,b}^*$, $b = 1, \ldots, B$. Letting $\widehat{\alpha} = B^{-1} \sum_{b=1}^B I_{\{L_{n,b}^* \geq L_n\}}$, the bootstrap test rejects H_0 if $\widehat{\alpha} \leq \alpha$. Note that $\widehat{\alpha}$ can be regarded as the p-value of the bootstrap test. In the Appendix we prove the following result on the type I error of the test.

Theorem 1. *As $n \to \infty$ and $B \to \infty$, $P_{\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_n}$(Bootstrap test rejects H_0)= $\alpha + O(n^{-1/2}) + O(B^{-1/2})$.*

Example 1. We compared the above bootstrap test (abbreviated by BOOT) with the CUSUM test and the tests of H_0 proposed by James, James, and Siegmund (1987, abbreviated by JJS), Pettitt (1980, abbreviated by P), and Brown, Durbin and Evans (1975, abbreviated by BDE) in a simulation study for the univariate normal setting $y_t \sim N(\theta_t, 1)$, $1 \leq t \leq n = 80$. The nominal significance level was $\alpha = 0.05$. For the bootstrap test, we used $B =$ 1,000 bootstrap samples and $j_1 = 5 = -j_0$ in (4.2). The results are given in Table 3, which considers the type I error of the tests at $\theta_1 = \theta_2 = \cdots = \theta_{80} = 0$,

Table 3. Type I error at (a) the null hypothesis $\theta_1 = \cdots = \theta_{80} = 0$, and power at the alternatives given by (b) single change-point at 41, $\theta_1 = -0.5$, $\theta_{41} = 0.5$, (c) two change-points at 33 and 65, $\theta_1 = -0.7$, $\theta_{33} = 0$, $\theta_{65} = 0.7$, (d) two change-points at 33 and 65, $\theta_1 = 0$, $\theta_{33} = 0.8$, $\theta_{65} = -0.8$.

Case	BOOT	JJS	P	BDE	CUSUM
(a)	0.0470 (0.0067)	0.0490 (0.0068)	0.0500 (0.0068)	0.0480 (0.0067)	0.0440 (0.0065)
(b)	0.9790 (0.0045)	0.9910 (0.0030)	0.9940 (0.0024)	0.8310 (0.0118)	0.9710 (0.0053)
(c)	0.9830 (0.0041)	0.9970 (0.0017)	0.9960 (0.0020)	0.8880 (0.0010)	0.9780 (0.0046)
(d)	0.9890 (0.0033)	0.1710 (0.0119)	0.2050 (0.0128)	0.5300 (0.0158)	0.8900 (0.0100)

and their power at three parameter configurations, two with 2 change-points and the other with only 1 change-point. Each result in Table 3 is based on 1,000 simulations and the standard errors are given in parentheses. The bootstrap test had high power (0.82 and above) at alternatives labeled (b)$-$(d), whereas JJS, P and BDE had low power (between 0.17 and 0.53) at the alternative (d), for which the first mean change is an increase and the second mean change is a decrease of the baseline mean. In this connection, note that for the alternative labeled (c), the mean increases at both change-points. The last column of Table 3 is about the CUSUM test that will be decribed more fully in the next example. Here, we use the fact that the variance of ϵ_t is known to be 1 to modify the usual CUSUM statistic in Example 2 to $n^{-1/2}\max_{1\le k\le n}|\sum_{t=1}^k(y_t-\widehat{\mu})|$, in which $\widehat{\mu}=n^{-1}\sum_{t=1}^n y_t$ as in Section 2.3.

Example 2. While Example 1 considers the case of known σ for normal mean shifts, nonparametric tests such as the CUSUM test have been introduced to test for mean shifts without assuming normality and prespecified σ; see Csörgö and Horvath (1998). The CUSUM statistic $n^{-1/2}\max_{1\le k\le n}|\sum_{t=1}^k(y_t-\widehat{\mu})|/\widehat{\sigma}$ has limiting null distribution $\max_{0\le t\le 1}|B_t|$ as $n\to\infty$, where B_t, $t\ge 0$, is Brownian motion, $\widehat{\mu}=n^{-1}\sum_{t=1}^n y_t$ as in Section 2.3, and $\widehat{\sigma}^2=n^{-1}\sum_{t=1}^n(y_t-\widehat{\mu})^2$ is a consistent estimate of σ^2 under the null hypothesis. The same weak convergence theory can also be applied to derive the limiting null distribution of (4.2) that assumes a Bayesian normal mean shift model as the alternative hypothesis, with p restricted to a range between $n^{-1}2^{j_0}$ and $n^{-1}2^{j_1}$. Since this limiting null distribution is the same as that when the y_t are normal, the parametric bootstrap test is still asymptotically valid even though the y_t may be non-normal. The common variance σ^2 of the y_t in the mean shift model can also be unknown and consistently estimated by $\widehat{\sigma}^2$, noting that the y_t are exchangeable random variables under the Bayesian mean shift model and under the null hypothesis. Table 4 compares the type I error and the power of this bootstrap test, abbreviated by BOOT, with those of the CUSUM test when the y_t were normal (left panel, denoted by N$(\cdot)$) and when the y_t were exponentially distributed with means θ_t

Table 4. Left panel (normal case): Type I error and power at the parameter configurations (a) and (b) of Table 3 when σ^2 is unknown and estimated by $\widehat{\sigma}^2$. Right panel (exponential case): Type I error at (a′) $\theta_1 = \cdots = \theta_{80} = 1$, and power at (b′) $\theta_1 = 0.5, \theta_{41} = 1.5$ for change-point at 41, (c′) $\theta_1 = 1, \theta_{34} = 1.5, \theta_{65} = 2$ for change-points at 34 and 65, (d′) $\theta_1 = 1, \theta_{34} = 2, \theta_{65} = 0.6$ for change-points at 34 and 65, (e′) $\theta_1 = 1, \theta_{10} = 1.5$ and $\theta_{72} = 0.5$ for change-points at 10 and 72.

Test	N(a)	N(b)	Exp(a′)	Exp(b′)	Exp(c′)	Exp(d′)	Exp(e′)
BOOT	0.0450	0.9790	0.0540	0.9540	0.8500	0.7460	0.8770
	(0.0066)	(0.0045)	(0.0072)	(0.0066)	(0.0113)	(0.0137)	(0.0104)
CUSUM	0.0440	0.9710	0.0460	0.9700	0.7980	0.4250	0.3330
	(0.0065)	(0.0053)	(0.0067)	(0.0053)	(0.0127)	(0.0156)	(0.0149)

(right panel, denoted by Exp(·)). As in Table 3, the nominal significance level was $\alpha = 0.05$ and each result was based on 1,000 simulations, with the standard error given in parentheses. Table 4 shows that BOOT and CUSUM had type I error near α even when the y_t were non-normal and that BOOT had substantially higher power than CUSUM for the last two columns of Table 4.

4.2. BCMIX smoothers

Although the Bayes filter uses a recursive updating formula (2.10) for the weights $p_{it} \propto p^*_{it}$ $(1 \le i \le t)$, the number of weights increases with t, resulting in unbounded computational complexity and memory requirements in estimating $\boldsymbol{\theta}_t$ as t keeps increasing. A *bounded complexity mixture* (*BCMIX*) approximation, having $M(p)$ components and keeping the most recent $m(p)$ weights $p_{j,n}$ (with $n - m(p) < j \le n$ and $m(p) < M(p)$) of the posterior density (2.8) can be obtained as follows. Let $\mathcal{K}_{t-1}(p)$ be the set of indices i for which $p_{i,t-1}$ is kept at stage $t-1$; thus, $\mathcal{K}_{t-1}(p) \supset \{t-1, \ldots, t-m(p)\}$. At stage t, define $p^*_{i,t}$ as in (2.10) for $i \in \{t\} \cup \mathcal{K}_{t-1}(p)$, and let i_t be the index not belonging to $\{t, \ldots, t-m(p)+1\}$ such that

$$p^*_{i_t,t} = \min\{p^*_{j,t} : j \in \mathcal{K}_{t-1}(p) \quad \text{and} \quad j \le t - m(p)\},$$

choosing i_t to be the minimizer farthest from t if the above set has two or more minimizers. Define $\mathcal{K}_t(p) = \{t\} \cup (\mathcal{K}_{t-1}(p) - \{i_t\})$, and let

$$p_{i,t} = \Big(\frac{p^*_{i,t}}{\sum_{j \in \mathcal{K}_t(p)} p^*_{j,t}}\Big), \quad i \in \mathcal{K}_t(p).$$

Similarly, to obtain a BCMIX approximation to the backward filter $\boldsymbol{\theta}_t | \mathcal{Y}_{t+1,n}$, let $\widetilde{\mathcal{K}}_{t+1}(p)$ denote the set of indices j for which $q_{j,t+1}$ in (2.11) is kept at stage $t+1$; thus, $\widetilde{\mathcal{K}}_{t+1}(p) \supset \{t+1, , \ldots, t+m\}$. At stage t, define $q^*_{j,t}$ as in (2.12) for

$j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}(p)$, and let j_t be the index not belonging to $\{t, \ldots, t+m(p)-1\}$ such that

$$q^*_{j_t,t} = \min\{q^*_{j,t} : j \in \widetilde{\mathcal{K}}_{t+1}(p) \quad \text{and} \quad j \geq t+m(p)\},$$

choosing j_t to be the minimizer farthest from t if the above set has two or more minimizers. Define $\widetilde{\mathcal{K}}_t(p) = \{t\} \cup (\widetilde{\mathcal{K}}_t(p) - \{j_t\})$ and let $q_{j,t} = \left(q^*_{j,t} \Big/ \sum_{j \in \widetilde{\mathcal{K}}_t(p)} q^*_{j,t}\right)$, $j \in \widetilde{\mathcal{K}}_t(p)$, which yields a BCMIX approximation to the backward filter $\boldsymbol{\theta}_t | \mathcal{Y}_{t+1,n}$.

The BCMIX approximation to the smoother can be obtained by combining the forward and backward BCMIX filters via Bayes' theorem:

$$f(\boldsymbol{\theta}_t | \mathcal{Y}_n) \approx \sum_{i \in \mathcal{K}_t(p),\ j \in \widetilde{\mathcal{K}}_{t+1}(p)} \widetilde{\beta}_{ijt} \pi(\boldsymbol{\theta}_t; a_0 + j - i + 1, \bar{\mathbf{Y}}_{i,j}), \tag{4.3}$$

in which $\widetilde{\beta}_{ijt} = \beta^*_{ijt} / \widetilde{P}_t$, $\widetilde{P}_t = p + \sum_{1 \leq t \leq n, i \in \mathcal{K}_t(p), j \in \widetilde{\mathcal{K}}_{t+1}(p)} \beta^*_{ijt}$, and β^*_{ijt} given by (2.15) for $i \in \mathcal{K}_t(p)$ and $j \in \widetilde{\mathcal{K}}_{t+1}(p)$. The BCMIX approximation to $E(\boldsymbol{\mu}_t | \mathcal{Y}_n)$ is therefore

$$\widehat{\boldsymbol{\mu}}_{t|n}(p) = \sum_{i \in \mathcal{K}_t(p),\ j \in \widetilde{\mathcal{K}}_{t+1}(p)} \widetilde{\beta}_{ijt} \bar{\mathbf{Y}}_{i,j}. \tag{4.4}$$

The following theorem, whose proof is given in the Appendix, assumes conditions similar to those of Yao (1988) for piecewise constant normal means:

(C1) The true change-points occur at $t_1^{(n)} < \cdots < t_k^{(n)}$ such that $\liminf_{n \to \infty} n^{-1} (t_i^{(n)} - t_{i-1}^{(n)}) > 0$ for $1 \leq i \leq k+1$, with $t_0^{(n)} = 0$ and $t_{k+1}^{(n)} = n$.

(C2) There exists $\delta > 0$, which does not depend on n, such that $\min_{1 \leq i \leq k} ||\boldsymbol{\mu}_{t_i^{(n)}} - \boldsymbol{\mu}_{t_{i-1}^{(n)}}|| \geq \delta$ for all large n.

Theorem 2. *Assume* (C1), (C2), *and that* $m(p) \sim |\log p|^{1+\epsilon}$ *and* $1 \leq M(p) - m(p) = O(1)$ *as* $p \to 0$, *for some* $\epsilon > 0$. *Then*

$$\max_{1 \leq t \leq n \ :\ \min_{1 \leq i \leq k} |t - t_i^{(n)}| \geq m(p)} ||\widehat{\boldsymbol{\mu}}_{t|n}(p) - \boldsymbol{\mu}_t|| \xrightarrow{P} 0 \ \textit{as}\ n \to \infty,$$

uniformly in $a_1/n \leq p \leq a_2/n$.

As noted in Section 2.3, the hyperparameter $\boldsymbol{\mu}_0$ can be estimated by the sample mean and the important hyperparameter p can be estimated by maximum likelihood. We can use the BCMIX approximation of the forward filter to replace $\sum_{i=1}^t p^*_{it}$ by $\sum_{i \in \mathcal{K}_t(p)} p^*_{it}$ in the likelihood function (2.17), and thereby estimate p by maximizing the approximate likelihood function over a grid $\{2^j/n : j_0 \leq$

$j \le j_1\}$. Putting this estimated p in $\widehat{\boldsymbol{\mu}}_t(p)$ in (4.4) yields the empirical Bayes estimator $\widehat{\boldsymbol{\mu}}_t$, which by Theorem 2 also satisfies the consistency property

$$\max_{1 \le t \le n\,:\, \min_{1\le i\le k} |t - t_i^{(n)}| \ge m(p)} ||\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t|| \xrightarrow{P} 0 \quad \text{as } n \longrightarrow \infty. \tag{4.5}$$

4.3. A new approach to choosing the number of segments

The second paragraph of Section 1 has reviewed frequentist methods in the literature for determining the number of change-points. These methods involve the log-likelihood statistic (under the assumption of k change-points)

$$l_n(k) = \sup_{1\le t_1 < \cdots < t_k < n,\ \boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(k+1)}} \sum_{j=1}^{k+1} \sum_{t=t_{j-1}}^{t_j - 1} \log f_{\boldsymbol{\theta}^{(j)}}(\mathbf{y}_t), \tag{4.6}$$

in which $t_0 = 1$ and $t_{k+1} - 1 = n$, together with a *penalized likelihood criterion* that subtracts some penalty term from $l_n(k)$. For example, the BIC penalty term is $(1/2)(k+1)d \log n$, which is used by Yao (1988) for the univariate normal mean shift model. The optimization problem involved in (4.6) can be solved by dynamic programming as in Bai and Perron (1998, 2003), but it is computationally intensive when k is not small. Zhang and Siegmund (2006) propose to approximate $l_n(k)$ by using the circular binary segmentation procedure of Olshen et al. (2004) to determine the maximizer $(t_1, \ldots t_k)$ in (4.6). Earlier Davis, Lee, and Rodriguez-Yam (2006) approximate it by using genetic algorithms.

Let $\widehat{\boldsymbol{\mu}}_t$ be a BCMIX approximation of $E(\boldsymbol{\mu}_t | \mathcal{Y}_n)$ and let

$$\Delta_t = ||\widehat{\boldsymbol{\mu}}_{t+b(p)} - \widehat{\boldsymbol{\mu}}_{t-b(p)}||^2, \tag{4.7}$$

where $b(p)$ is a "bandwidth" whose choice will be given later. Instead of using dynamic programming to determine the k change-points by maximizing the likelihood function (4.6) in a model with $k+1$ segments, we estimate the k change-points sequentially by making use of $\{\Delta_t : b(p) < t < n - b(p)\}$. Let $\widehat{\tau}_1$ be the maximizer of Δ_t over $b(p) < t < n - b(p)$. After $\widehat{\tau}_1, \ldots, \widehat{\tau}_{j-1}$ have been defined, we can define $\widehat{\tau}_j$ as the maximizer of Δ_t over t that lies outside the $b(p)$-neighborhoods of $\widehat{\tau}_i$ for $1 \le i \le j-1$, i.e.,

$$\Delta_{\widehat{\tau}_j} = \max\{\Delta_t : b(p) < t < n - b(p),\ \min_{1\le i \le j-1} |t - \widehat{\tau}_i| \ge b(p)\}. \tag{4.8}$$

Note that whereas (C1) orders the change-points $t_1^{(n)}, \ldots, t_k^{(n)}$, the estimates $\widehat{\tau}_j$ of the locations of the change-points in (4.8) are unordered and do not depend on k. Under the model of k change-points, we can take $\widehat{\tau}_1, \ldots, \widehat{\tau}_k$ and order them as $\widehat{t}_{(1),k} < \cdots < \widehat{t}_{(k),k}$ to provide estimates of the k change-points. We can then estimate the common parameter $\boldsymbol{\theta}^{(j)}$ over $[t_{j-1}^{(n)}, t_j^{(n)} - 1]$ by the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}^{(j)}$ over the estimated segment $[\widehat{t}_{(j-1),k}, \widehat{t}_{(j),k} - 1]$, $1 \le j \le k{+}1$, where $\widehat{t}_{(0),k} = 1$ and $\widehat{t}_{(k+1),k}{-}1 = n$. This yields the following approximation to (4.6):

$$\Lambda_n(k) = \sum_{j=1}^{k+1} \sum_{t=\widehat{t}_{(j-1),k}}^{\widehat{t}_{(j),k}-1} \log f_{\widehat{\boldsymbol{\theta}}^{(j)}}(\mathbf{y}_t). \tag{4.9}$$

With an upper bound K on the number k of change-points in (C1), we propose to estimate k by

$$\widehat{k}_n = \text{argmax}_{1 \le k \le K} \{\Lambda_n(k) - (k+1)C_n\}, \tag{4.10}$$

where C_n is the common penalty for each segment and satisfies

$$C_n \to \infty \text{ and } C_n/n \to 0 \text{ as } n \to \infty. \tag{4.11}$$

Theorem 3. *Under* (4.11) *and the assumptions of Theorem* 2*, together with* $b(p) \sim \beta m(p)$ *for some* $\beta > 0$*,* $\widehat{k}_n \xrightarrow{P} k$.

The proof of Theorem 3 is given in the Appendix. The choice of C_n that corresponds to BIC is $C_n = (d/2)\log n$, which clearly satisfies (4.11). The following simulation study considers the finite-sample performance of $\widehat{k}_n$ with this choice of C_n, and also of the BCMIX estimator $\widehat{\boldsymbol{\mu}}_t$ introduced in the preceding section.

Example 3. Figure 2 illustrates the performance of the BCMIX estimator $\widehat{\mu}_{t|n}$ in the normal mean shift model $y_t \sim N(\mu_t, 1)$, $1 \le t \le n = 2{,}500$. The top left panel shows a random sample generated from the model with four change-points:

$$\begin{aligned}\mu_t = {} & I_{\{1 \le t \le 500\}} + 1.8 I_{\{501 \le t \le 1{,}000\}} + 0.5 I_{\{1{,}001 \le t \le 1{,}500\}} \\ & + I_{\{1{,}501 \le t \le 1{,}750\}} + 0.6 I_{\{1{,}751 \le t \le 2{,}500\}}.\end{aligned} \tag{4.12}$$

The top right panel, which has 22 change-points, shows a random sample generated from the Bayesian model $\mu_t = (1 - I_t)\mu_{t-1} + I_t z_t$, in which I_t is Bernoulli(p) independent of z_t, where $p = 0.006$ and z_t has the conditional distribution of a standard normal random variable given that it exceeds 1 in absolute value; this restriction is introduced to avoid small jumps in the Bayesian model. The bottom panels plot μ_t (solid line) and the BCMIX estimate $\widehat{\mu}_t$ (dotted line), and show that $\widehat{\mu}_t$ is close to μ_t in both cases. We use the criterion (4.10) with $C_n = (1/2)\log n$ to choose the number k of change-points. The preceding method estimates that there are 4 change-points located at 5,00, 1,004, 1,495, 1,775 for the left panel. The estimated number of change-points for the right panel is the actual number 22.

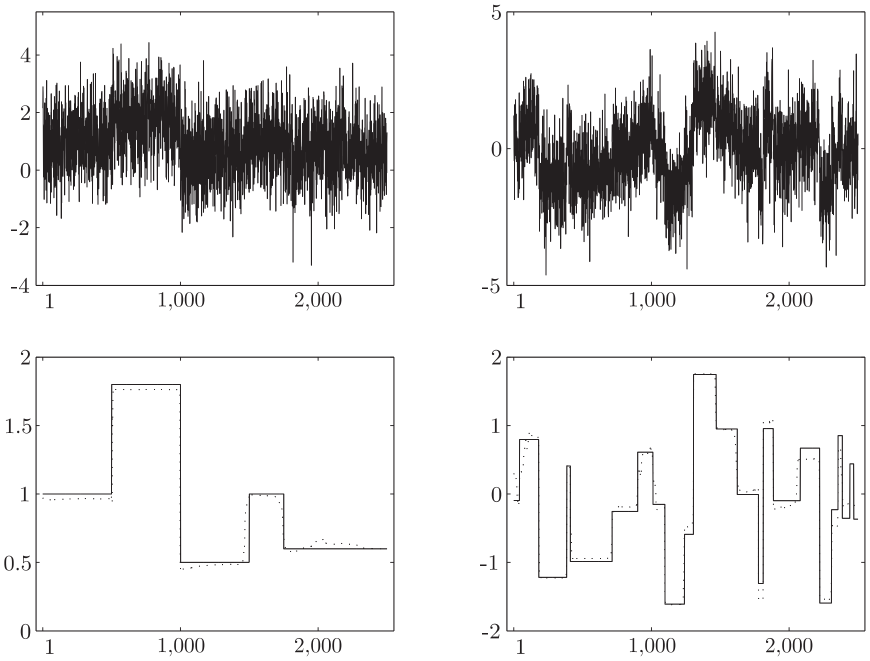

Figure 2. Top: Time series generated from the mean-shift models in the frequentist (left panel) and the Bayesian framework (right panel). Bottom: True (solid line) and estimated means (dotted line).

Table 5 compares the BCMIX procedure with three other methods in the literature. The first is the circular binary segmentation (CBS) method of Olshen et al. (2004), which determines k via the type I error of a change-point test and certain pruning procedures; it is widely regarded as one of the fastest segmentation methods in genomic studies. The CBS method can be implemented by using the DNAcopy package from the Bioconductor Web site

`http://www.bioconductor.org/packages/2.3/bioc/html/DNAcopy.html`.

The second method, due to Davis, Lee, and Rodriguez-Yam (2006), and denoted by MDL_g, uses the minimum description length (MDL) criterion to choose k and a variant of the genetic algorithm to approximate (4.6). We have used in Table 5 the canonical genetic algorithm instead of its faster variant used by Davis et al. The third method, due to Bai and Perron (1998, 2003), estimates k by minimizing BIC and uses dynamic programming (DP) to evaluate (4.6). We implement it by using the strucchange package from the R Project Web site

`http://cran.r-project.org/web/packages/strucchange/index.html`.

Table 5. Segmentation using different methods.

	Model (4.12)						Bayesian Model	
Method	MSE	$\widehat{k}=2$	$\widehat{k}=3$	$\widehat{k}=4$	$\widehat{k}=5$	$\widehat{k}=6$	MSE	$E\lvert\widehat{k}-k\rvert$
BCMIX	15.93 (0.27)	10%	4.1%	81.2%	4.7%	0%	33.71 (0.39)	0.079 (0.009)
CBS	33.71 (0.39)	2.5%	0.1%	91.4%	1.4%	4%	48.53 (0.71)	0.850 (0.027)
MDL$_g$	22.13 (0.48)	15.6%	1.1%	83.3%	0%	0%	175.65 (2.72)	0.401 (0.023)
DP	18.97 (0.42)	7.1%	2.5%	90.4%	0%	0%	379.95 (5.66)	1.574 (0.043)

Let MSE$=E\{\sum_{t=1}^n(\widehat{\mu}_t-\mu_t)^2\}$. Based on 1,000 simulations, Table 5 gives the MSE of BCMIX, CBS, MDL$_g$, and DP for model (4.12), and for the Bayesian mean shift model described in the preceding paragraph. Also given in Table 5 are the percentages of $\widehat{k}=2,3,4,5,6$ of the 1,000 simulations in model (4.12), and the mean of $|\widehat{k}-k|$ in the Bayesian model for each method.

We ran all simulations on an Intel(R) Pentium (R) 4 CPU 2.80GHz computer with 2GB memory. The average CPU times per simulation of BCMIX, CBS, and MDL$_g$ were 0.43 seconds, 0.87 seconds, and 10 minutes, respectively. To implement the DP method, we set the minimum distance between two adjacent change-points to be 250 for model (4.12), and to be 125 for the Bayesian model. The corresponding average CPU times per simulation were 67 and 98 minutes, respectively. Note that choosing a smaller minimum distance in the DP method for the Bayesian model would reduce the MSE but increase the lengthy CPU time.

5. Application to British Coal Mine Data

In this section we use the method developed in Sections 2 and 4.3 to analyze the time series of annual numbers Y_t of British coal mine disasters from 1850 to 1962. Previous works that used change-point methods to analyze these data assumed that Y_t were independent Poisson random variables with means θ_t; see Green (1995, p.723) for a summary of previous Bayesian analyses. We also assume that Y_t are independent Poisson random variables with means θ_t, in conjunction with the change-point model in Section 2, for which a conjugate prior Gamma(γ,λ) is assumed for θ_t at the time of a jump, which occurs with probability p. As shown in Section 2.2, the posterior distribution of θ_t given $Y_1,\dots,Y_n$ is a mixture of Gamma$(\gamma_{ij},\lambda_{ij})$ distributions, where

$$\gamma_{ij}=\gamma+\sum_{k=i}^{j}Y_k, \qquad \lambda_{ij}=(\lambda^{-1}+j-i+1)^{-1}.$$

Moreover, the $\pi_{i,j}$ in (2.15) can be expressed as $\pi_{i,j}=[\Gamma(\gamma_{ij})]^{-1}\lambda_{ij}^{-\gamma_{ij}}$. To estimate the hyperparameters p, γ and λ, we calculate the likelihood function (2.17)

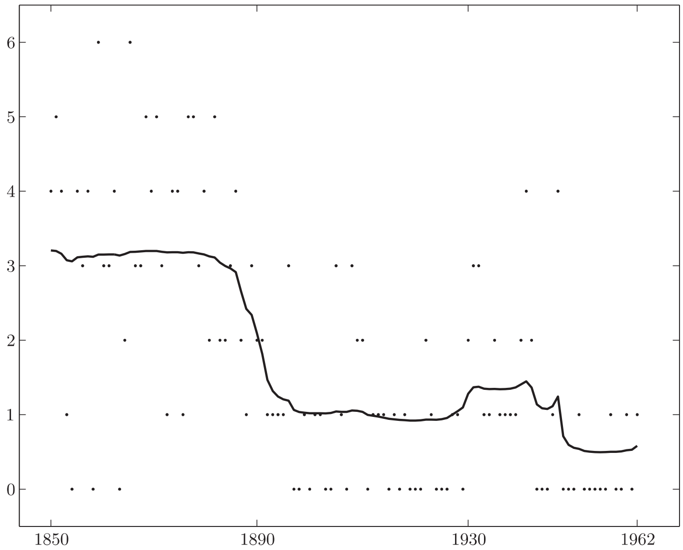

Figure 3. The British coal mine data and posterior means of the Poisson rates.

for $p = 2^i/n$ $(-10 \leq i \leq 5)$, $\lambda = 0.5j$, and $\gamma = 0.1 + 0.2k$ $(1 \leq j, k \leq 10)$. The maximum likelihood estimates (over this grid) are $\widehat{p} = 0.0357$, $\widehat{\lambda} = 1.0$, and $\widehat{\gamma} = 1.7$, which we use as the estimated hyperparameters. Here $n = 112$, so $n\widehat{p}$ is still small and we can use the BCMIX smoother $\widehat{\theta}_t$ (for which we choose $m = 10$ and $M = 20$) to estimate θ_t. Figure 3 plots Y_t and $\widehat{\theta}_t$; the plot of $\widehat{\theta}_t$ is similar to that of Green (1995) based on daily (instead of yearly) data and a different Bayesian model. Figure 4 plots the posterior probability that $\theta_{t+1} \neq \theta_t$ for $1 \leq t < n$; see (2.16). Combining both figures suggests three change-points around 1891, 1929 and 1947. We also apply (4.10) with $C_n = (1/2) \log n$, which yields 3 as the estimated number of change-points. Moreover, the method in Section 4.3 estimates the locations of the three change-points to be 1891, 1929 and 1947.

Assuming $y_t \sim \text{Poisson}(\theta_t)$, $1 \leq t \leq n = 112$, and θ_t to be piecewise constant, we performed a simulation study of our method to estimate θ_t generated from three frequentist and Bayesian models that differ from our working Bayesian model. The first, denoted by W, is Worsley's (1986) fitted change-point model that has a single change-point at the 45th observation and estimates θ_1 and θ_{45} by maximum likelihood. We simulated the y_t from the fitted model. The second model is the Bayesian model for a single change-point considered by Raftery and Akman (1986) and is denoted by RA. We simulated the θ_t from the posterior distribution of the change-point and the pre- and post- change values of θ_t. The third model, denoted by BH, is Barry and Hartigan's (1992; 1993) product partition model, in which we use the cohesion $c_{ij} = (j - i + 1)^{-3}$,

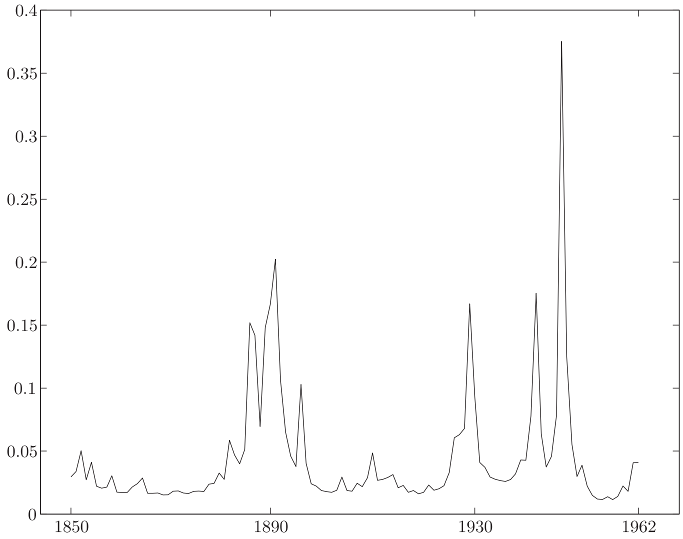

Figure 4. Posterior probability of occurrence of change-point.

Table 6. Performance of $\widehat{\theta}_t$ and $\widehat{k}$ in change-point Poisson models.

Model	KL	EE	$E(\Delta)$ ($\Delta = \lvert\widehat{k} - k\rvert$)	$\Delta = 0$	$\Delta = 1$	$\Delta = 2$
W	0.0724 (7.27e-4)	0.2946 (2.70e-3)	0.0470 (0.008)	96.4%	2.5%	1.1%
RA	0.0269 (5.73e-4)	0.2034 (2.49e-3)	0.0450 (0.007)	96.1%	3.3%	0.6%
BH	0.0287 (5.87e-4)	0.2094 (2.50e-3)	0.1640 (0.012)	84.3%	15%	0.7%

$1 \le i < t \le j \le n$, and impose an additional constraint that there are at most 3 change-points. Using the posterior distribution of $(\theta_1, \ldots, \theta_n)$ from these data, we simulated data from the fitted BH model and used the simulated data to compute the BCMIX estimate based on our working Bayesian change-point model that differs from the actual BH model. We use the Kullback-Leibler divergence $\mathrm{KL}(\theta_t, \widehat{\theta}_t) = E\big\{\theta_t(\log\theta_t - \log\widehat{\theta}_t) - (\theta_t - \widehat{\theta}_t)\big\}$ and the mean absolute error $E|\widehat{\theta}_t - \theta_t|$ to assess the estimation error of the empirical Bayes estimate $\widehat{\theta}_t$ of θ_t. Table 6 gives the results of KL$= n^{-1}\sum_{t=1}^n \mathrm{KL}(\theta_t, \widehat{\theta}_t)$ and EE$= n^{-1}\sum_{t=1}^n E|\widehat{\theta}_t - \theta_t|$ based on 1,000 simulations. Besides the estimation error of $\widehat{\theta}_t$, it also gives $E|\widehat{k} - k|$ and the percentages of $|\widehat{k} - k| = 0, 1, 2$ in the 1,000 simulations.

6. General Parametric Families and Change-point Generalized Linear Models

We can apply the same change-point model and use the same ideas to develop recursive estimators for more general parametric families $f_{\boldsymbol{\theta}}(\mathbf{y}_t)$ than the exponential family (2.1). In particular, corresponding to a prior density function that is proportional to $g(\boldsymbol{\theta})$, the conditional density function $g_{i,t}$ of $\boldsymbol{\theta}_t$ given $K_t = i$ and $\mathcal{Y}_t$ is

$$g_{i,t}(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta}) \prod_{k=i}^{t} f_{\boldsymbol{\theta}}(\mathbf{y}_i),$$

and therefore (2.8) can be generalized to $f(\boldsymbol{\theta}_t|\mathcal{Y}_t) = \sum_{i=1}^{t} p_{it} g_{i,t}(\boldsymbol{\theta}_t)$, with $p_{it} = p_{it}^* \big/ \sum_{k=i}^{t} p_{kt}^*$, where p_{it}^* is given by (2.10) but with

$$\frac{1}{\pi_{0,0}} = \int g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \qquad \frac{1}{\pi_{ij}} = \int \Big[\prod_{k=i}^{j} f_{\boldsymbol{\theta}}(\mathbf{y}_k)\Big] g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{6.1}$$

The argument of Section 2.2 yields $f(\boldsymbol{\theta}_t|\mathcal{Y}_n) = \sum_{1 \le i \le t \le j \le n} \beta_{ijt} g_{i,j}(\boldsymbol{\theta}_t)$, in which β_{ijt} is proportional to β_{ijt}^* given by (2.15), with the modification for p_{it}, q_{jt}, and $\pi_{i,j}$ given by (6.1). An advantage of the exponential family and the conjugate prior density function is that the integrals in (6.1) can be explicitly evaluated. For general parametric families that do not have explicit formulas for (6.1), we can use Laplace's asymptotic formula to evaluate (6.1) when $j - i$ is sufficiently large; see Lemma 1 in the Appendix for details.

Another important extension of these ideas is to generalized linear models with piecewise constant parameters. By incorporating lagged covariates in the generalized linear model, we can extend the Gaussian autoregressive model with piecewise constant regression parameters and error variances, introduced by Lai, Liu, and Xing (2005), to generalized linear models. Suppose that at time t, $\mathbf{y}_t$ has density function

$$f_{\boldsymbol{\theta}_t,\phi_t}(\mathbf{y}_t) = c(\boldsymbol{\theta}_t, \phi_t) \exp\Big\{ \frac{[\boldsymbol{\theta}_t' \mathbf{y}_t - \psi(\boldsymbol{\theta}_t)]}{a(\phi_t)} \Big\}, \tag{6.2}$$

with respect to some measure ν on $\mathbb{R}^d$, and that $\boldsymbol{\mu}_t = \boldsymbol{\beta}_t' \mathbf{x}_t$ in which $\boldsymbol{\mu}_t = \boldsymbol{\nabla} \psi(\boldsymbol{\theta}_t)$ is the mean of $\mathbf{y}_t$ and $\mathbf{x}_t$ is the covariate vector at time t. The dynamics of $(\boldsymbol{\beta}_t, \phi_t)$ are the same as in Section 2. At time t when a parameter change occurs, ϕ_t has density function $\pi(\phi)$ and the conditional density function of $\boldsymbol{\beta}_t$ given $\phi_t = \phi$ is $\pi(\boldsymbol{\beta}|\phi)$, where

$$\begin{aligned} \pi(\boldsymbol{\beta}|\phi) &= c_1(a_0, \boldsymbol{\mu}_0) \exp\Big\{ a_0 \boldsymbol{\beta}' \boldsymbol{\mu}_0 - a_0 \psi(\boldsymbol{\beta}) \Big\}, \\ \pi(\phi) &= c_2(b_0, \nu_0) \exp\{ b_0 \nu_0 \phi - b_0 \lambda(\phi) \}, \end{aligned} \tag{6.3}$$

in which $c_1(a_0, \boldsymbol{\mu}_0)$ and $c_2(b_0, \nu_0)$ are normalizing constants, similar to (2.2). Note that the arguments in Section 2 can be applied here with obvious modifications; in particular, the $\pi_{i,j}$ can be computed by numerical integration or Laplace's approximation.

Table 7. Performance of BCMIX and Oracle estimates under two scenarios.

	Estimate	Scenario A	Scenario B
KL	BCMIX	0.0051 (3.56e-4)	0.0066 (1.54e-5)
	Oracle	0.0030 (5.60e-5)	0.0040 (6.36e-5)
EE	BCMIX	0.1176 (3.56e-3)	0.1073 (1.24e-3)
	Oracle	0.0979 (1.15e-3)	0.0915 (8.71e-4)

Example 4. Consider a Poisson autoregressive model for counts data $y_t \sim \text{Poisson}(\mu_t)$, with

$$\log(\mu_t) = \alpha_t + \gamma_t \log(y^*_{t-1}), \tag{6.4}$$

where $y^*_{t-1} = \max(y_{t-1}, c)$ and $c = 0.5$ is used to determine the probability that $y_t > 0$ given $y_{t-1} = 0$; see Zeger and Qaqish (1988, p.1021). This is a special case of the above change-point generalized linear model with $a(\phi_t) \equiv 1$, $\boldsymbol{\beta}_t = (\alpha_t, \gamma_t)'$, and $\mathbf{x}_t = (1, \log y^*_{t-1})'$. We consider two scenarios for the piecewise constant $\boldsymbol{\beta}_t$ with $1 \le t \le n = 1{,}000$.

$$\begin{aligned} \textit{Scenario A}: \ \boldsymbol{\beta}'_t &= (0.1, 0.1) I_{\{1 \le t \le 200\}} + (-0.5, 0.3) I_{\{201 \le t \le 600\}} \\ &\quad + (0.5, -0.5) I_{\{601 \le t \le 1{,}000\}}. \\ \textit{Scenario B}: \ \boldsymbol{\beta}'_t &= (0.5, -0.2) I_{\{1 \le t \le 200\}} + (-0.2, 0.3) I_{\{201 \le t \le 500\}} \\ &\quad + (0.3, -0.3) I_{\{501 \le t \le 800\}} + (0.3, 0.7) I_{\{801 \le t \le 1{,}000\}}. \end{aligned}$$

We use the Kullback-Leibler divergence and the mean Euclidean error to assess the estimation error of the empirical Bayes estimate $\widehat{\boldsymbol{\beta}}_t$ of $\boldsymbol{\beta}_t$:

$$\text{KL}_t = E\Big\{\mu_t(\log \mu_t - \log \widehat{\mu}_t) - (\mu_t - \widehat{\mu}_t)\Big\}, \quad \text{EE}_t = E\Big\{(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t)'(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t)\Big\}^{1/2}.$$

We also compare $\widehat{\boldsymbol{\beta}}_t$ with the "oracle" estimate $\boldsymbol{\beta}^*_t$ that assumes the change-points to be known and estimates the common $\boldsymbol{\beta}_t$ in each segment by maximizing the likelihood function. Table 7 gives the Monte Carlo estimates of KL$=n^{-1}\sum_{t=1}^n \text{KL}_t$ and EE$=n^{-1}\sum_{t=1}^n \text{EE}_t$ and their standard errors (in parentheses) for scenarios A and B; each result is based on 1,000 simulations. The table shows that the BCMIX estimates are not much inferior to the oracle estimates.

7. Conclusion

By making use of K_t and $\widetilde{K}_t$ in Sections 2.1 and 2.2, we have derived explicit recursive formulas for the posterior estimates of $\boldsymbol{\theta}_t$ given $\mathcal{Y}_t$ or $\mathcal{Y}_n$ in our Bayesian model of occasional parameter jumps in the multiparameter exponential family (2.1). The BCMIX smoothers (4.4) and the associated empirical Bayes estimators $\widehat{\boldsymbol{\mu}}_t$ in Section 4.2 are shown to provide simple approximations that perform nearly as well as their fully Bayesian counterparts. We have also shown in Sections 4.1 and 4.3 how this empirical Bayes approach, with its computationally attractive recursive estimators, can be used to address challenging frequentist problems. A commonly used method to estimate piecewise constant signals $\boldsymbol{\theta}_t$ is to first segment the data and then to estimate the constant signal for each segment. Since our empirical Bayesian approach already provides a relatively simple estimate of the signal without segmentation, it seems that this approach makes segmentation superfluous. Although this is the case when there are many possible but no clear-cut segments, there are sometimes subject-matter reasons for segmentation, especially if the signal actually consists of a few long segments as prescribed by conditions (C1) and (C2); see Siegmund (2004), Olshen et al. (2004) and Lai, Xing, and Zhang (2008). Determining the number and locations of change-points in these situations is an important problem even though we can estimate the signal $\boldsymbol{\theta}_t$ well by using the empirical Bayes approach.

The comparative study in Section 4 shows the computational and statistical advantages of the BCMIX smoother and the associated segmentation method over the widely used CBS algorithm in genomic studies. For on-line (sequential) estimation and detection problems, the BCMIX filter can be extended beyond the exponential family to provide recursive estimators and detection schemes. In connection with ongoing work in this direction, Lai, Liu, and Xing (2009) have made use of BCMIX filters to develop efficient sequential surveillance schemes for detecting multiple parameter changes in multivariate exponential families. In Section 6 we extended the change-point model and its associated methodology from the independent exponential family to generalized linear autoregressive models whose regression and dispersion parameters may undergo occasional jumps. This extension greatly broadens the scope of previous work by Bai, Perron, and Qu on regression models with multiple change-points, and by McCulloch and Tsay, Lai, Liu, and Xing on Gaussian autoregressive models with piecewise constant regression and volatility parameters (see Section 1).

Although the relatively simple Bayesian model we have used seems to miss certain features incorporated in previous more complicated Bayesian change-point models that require MCMC methods for their implementation, the simulation studies in Sections 3 and 5 suggest that our method based on the simple Bayesian change-point model still performs well when the data are generated from these more complicated Bayesian models. A heuristic explanation is that there is not much information around the unknown change-points, where the Bayesian model serves as a working model to smooth the data. On the other hand, the constant parameter vector over a long segment between change-points can be estimated well. Moreover, our empirical Bayes approach adjusts the hyperparameters of the Bayesian change-point model to the data and thereby accounts for the robustness of the working model. This also explains the results of Example 2 in which the bootstrap test assuming a Bayesian normal mean shift model for the alternative hypothesis is compared to the nonparametric CUSUM test. Our test statistic can be viewed as a general type of the scan statistics considered in Chan and Lai (2003, 2006), to which functional central limit theorems and moderate deviation approximations can be used to establish the "invariance principle" that the y_t can be treated as normal, similar to the arguments for the CUSUM and other nonparametric test statistics for mean shifts (Csörgö and Horvath (1998)).

Acknowledgement

The first part of this paper and a synopsis of Lai, Xing, and Zhang (2008) were presented by Lai in a plenary lecture at the Cross Straits Statistics Conference in Miaoli, Taiwan, in 2006. He thanks Anges Hsiung, chair of the Organizing Committee, for the invitation, and the participants, particularly Hongzhe Li and Yi-Ching Yao, for helpful comments. He also gratefully acknowledges partial support of this research by the National Science Foundation under grant DMS-0805879. Xing's research is supported by the National Science Foundation under grant DMS-0906593. The authors are grateful to an associate editor and the referees for their careful reading of the paper and for valuable comments.

Appendix

Proof of Theorem 1. Let A_c denote the event $\{L_n \geq c\}$ and let $\boldsymbol{\theta}^0$ denote the common value of $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_n$ under H_0. In view of the asymptotic normality of $\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})$, we can write

$$\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}}) = \boldsymbol{\theta}^0 + \frac{\mathbf{z}_n}{\sqrt{n}}, \tag{A.1}$$

where $\mathbf{z}_n$ has a limiting normal distribution with mean $\mathbf{0}$. By Lemma II.1.1 and Theorem II.1.2 of Ibragimov and Has'minskii (1981)

$$\Big|P_{\boldsymbol{\theta}^0+\mathbf{d}_n/\sqrt{n}}(A_c) - P_{\boldsymbol{\theta}^0}(A_c)\Big| = \Big|E_{\boldsymbol{\theta}^0}\Big\{\Big[\frac{\prod_{i=1}^n f_{\boldsymbol{\theta}^0+\mathbf{d}_n/\sqrt{n}}(y_i)}{f_{\boldsymbol{\theta}^0}(y_i)}\Big] - 1\Big\} I_{A_c}\Big| = O(n^{-1/2}) \tag{A.2}$$

uniformly in $\mathbf{d}_n \in D$, for any compact subset D of $\mathbb{R}^d$ such that $\boldsymbol{\theta}_0 + n^{-1/2}D \subset \boldsymbol{\Theta}_0$. Combining (A.2) with (A.1), and noting that the bootstrap test uses B independent samples drawn from $P_{\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})}$ to estimate $P_{\boldsymbol{\theta}(\widehat{\boldsymbol{\mu}})}(A_c)$, we obtain the desired conclusion.

To prove Theorems 2 and 3, we first prove the following lemmas, which make the same assumptions as those in Theorem 2. The first three lemmas are used to analyze the weights p_{it}, $i \in \mathcal{K}_t(p)$, in the forward BCMIX filter. A similar analysis yields corresponding results for the weights $q_{j,t}$, $j \in \widetilde{\mathcal{K}}_t(p)$, in the backward BCMIX filter. Lemma 4 combines these results to provide the asymptotic behavior of the weights $\widetilde{\beta}_{ijt}$ in (4.4), which we use to prove Theorem 3.

Lemma 1. *Let* $\boldsymbol{\theta}_{\boldsymbol{\mu}} = (\nabla\psi)^{-1}(\boldsymbol{\mu})$, $I(\boldsymbol{\mu}) = \boldsymbol{\theta}'_{\boldsymbol{\mu}}\boldsymbol{\mu} - \psi(\boldsymbol{\theta}_{\boldsymbol{\mu}})$, $h(\boldsymbol{\mu}) = det\big(\boldsymbol{\nabla}^2\psi(\boldsymbol{\theta}_{\boldsymbol{\mu}})\big)$, *where* $\boldsymbol{\nabla}^2(\psi)$ *denotes the Hessian matrix of second partial derivative* $\partial^2\psi/\partial\theta_i\partial\theta_j$. *Define* $\bar{\mathbf{Y}}_{i,j}$ *and* $\pi_{i,j}$ *for* $i \le j$ *as in Section* 2.1. *Then as* $j - i \to \infty$,

$$\pi_{i,j}^{-1} \sim \frac{(2\pi)^{d/2} e^{(a_0+j-i+1)I(\bar{\mathbf{Y}}_{i,j})}}{\Big\{(a_0+j-i+1)^d h(\bar{\mathbf{Y}}_{i,j})\Big\}^{1/2}} \tag{A.3}$$

uniformly in $\bar{\mathbf{Y}}_{i,j} \in \Gamma$, *for every compact subset* Γ *of* $\boldsymbol{\nabla}\psi(\boldsymbol{\Theta})$.

Proof. Note that

$$\pi_{i,j}^{-1} = \int_{\boldsymbol{\Theta}} \exp\Big\{(a_0+j-i+1)\big[\boldsymbol{\theta}'\bar{\mathbf{Y}}_{i,j} - \psi(\boldsymbol{\theta})\big]\Big\} d\boldsymbol{\theta}. \tag{A.4}$$

Moreover, $I(\boldsymbol{\mu}) = \max_{\boldsymbol{\theta}} \big(\boldsymbol{\theta}'\boldsymbol{\mu} - \psi(\boldsymbol{\theta})\big)$. Hence (A.3) is simply Laplace's asymptotic formula for the above integral.

The BCMIX weights p_{it}, $i \in \mathcal{K}_t(p)$, are difficult to analyze directly because they are defined recursively via (2.9) for which there is renormalization (from p^*_{it} to p_{it}) at every stage t. We approximate them by a more tractable version in Lemmas 2 and 3. To fix the ideas, first assume that $M(p) - m(p) = 1$, which is tantamount to allowing one more change-point prior to t besides the most recent possibilities $t, \ldots, t - m(p) + 1$. Our approximation is similar to the AMOC ("at most one change") estimator of Chernoff and Zacks (1964, Sec. 6) and assumes that at most one change can occur before the filter removes the "ancient" one of the $m(p) + 1$ components from the mixture. Denoting $m(p)$ and $M(p)$ simply by m and M, we now describe the AMOC filter more precisely. First note that for the Bayesian model in Section 2, for $t - m + 1 \le i \le t$,

$$\begin{aligned} P\{I_j = 0 \text{ for } 1 < j \le t|\mathcal{Y}_t\} &\propto (1-p)^{t-1}\frac{\pi_{0,0}}{\pi_{1,t}}, \\ P\{I_i = 1, I_j = 0 \text{ for } 1 < j \le t \text{ and } j \ne i|\mathcal{Y}_t\} &\propto p(1-p)^{t-2}\frac{\pi_{0,0}^2}{\pi_{1,i-1}\pi_{i,t}}. \end{aligned} \tag{A.5}$$

Note that the sum of the above probabilities is $P\{t_1^{(n)} \ge t - m + 1, t_2^{(n)} > t|\mathcal{Y}_t\}$, i.e., the posterior probability of at most one parameter change up to time t, which change can only occur at times $t-m+1,\dots,t$. We can use the AMOC filter to estimate $t_1^{(n)}$ by

$$\tau_1=\inf\Big\{t:\frac{(1-p)}{\pi_{1,t}}<\frac{p\pi_{0,0}}{(\pi_{1,t-m}\pi_{t-m+1,t})}\Big\},\tag{A.6}$$

and then repeat the same procedure, with $\mathcal{Y}_t$ replaced by $\mathcal{Y}_{\tau_1,t}$, to estimate $t_2^{(n)}$ from $\{\mathbf{y}_t, t>\tau_1\}$. Proceeding inductively in this way yields the change-time estimates $\tau_1<\tau_2<\dots$. In view of (A.5), the AMOC filter weights for $t<\tau_1$ are

$$p_{1,t}^A=\frac{1-p}{\pi_{1,t}P_t^A},\quad p_{i,t}^A=\frac{p\pi_{0,0}}{\pi_{1,i-1}\pi_{i,t}P_t^A}\text{ for }t-m+1\le i\le t,\tag{A.7}$$

where P_t^A is the normalizing constant to make the $m+1$ weights in (A.7) add up to 1. While keeping the most recent m indices as in BCMIX, (A.6) basically compares $P\{I_j=0 \text{ for } 1\le j\le t|\mathcal{Y}_t\}$ with $P\{I_{t-m}=1, I_j=0 \text{ for } 1<j\le t \text{ and } j\ne t-m|\mathcal{Y}_t\}$ and keeps the index with the larger posterior probability, analogous to BCMIX. The AMOC filter weights for $\tau_i\le t<\tau_{i+1}$ are defined similarly, with $\mathcal{Y}_{\tau_i,t}$ taking the place of $\mathcal{Y}_t$. The following lemma gives the asymptotic properties of the AMOC filter.

Lemma 2. *As* $n\to\infty$, $P\{\max_{1\le i\le k}|t_i^{(n)}-\tau_i|\le m\}\to 1$. *Moreover,* $\max_{t<t_1^{(n)}}|p_{1,t}^A-1|\xrightarrow{P}0$ *and* $\max_{\tau_i<t<t_{i+1}^{(n)}}|p_{\tau_i,t}^A-1|\xrightarrow{P}0$ *for* $1\le i\le k+1$.

Proof. Let $\mathcal{I}(\boldsymbol{\mu},\boldsymbol{\gamma})=(\boldsymbol{\theta}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_{\boldsymbol{\gamma}})'\boldsymbol{\mu}-\big(\psi(\boldsymbol{\theta}_{\boldsymbol{\mu}})-\psi(\boldsymbol{\theta}_{\boldsymbol{\gamma}})\big)$ denote the Kullback-Leibler information number. We first show that as $j\to\infty$ and $t/j\to\infty$,

$$\begin{aligned}&(a_0+t-j)I(\bar{\mathbf{Y}}_{1,t-j})+(a_0+j)I(\bar{\mathbf{Y}}_{t-j+1,t})-(a_0+t)I(\bar{\mathbf{Y}}_{1,t})\\&\quad=j\big\{\mathcal{I}(\bar{\mathbf{Y}}_{t-j+1,t},\bar{\mathbf{Y}}_{1,t-j})+O(\frac{j}{t})\big\},\end{aligned}\tag{A.8}$$

in which the $O(j/t)$ term is uniform over $||\bar{\mathbf{Y}}_{t-j+1,t}||+||\bar{\mathbf{Y}}_{1,t-j}||\le B$, for every given B. To prove the asymptotic relation (A.8) for which the prior distribution has a negligible effect, we drop a_0 (by letting it approach 0) for notational simplicity so that $\bar{\mathbf{Y}}_{1,t}$ becomes the sample mean, which can be written as $\big\{(t-j)\bar{\mathbf{Y}}_{1,t-j}+j\bar{\mathbf{Y}}_{t-j+1,t}\big\}/t$. Since $\boldsymbol{\nabla}I(\boldsymbol{\mu})=\boldsymbol{\theta}_{\boldsymbol{\mu}}$, it follows that

$$I(\bar{\mathbf{Y}}_{1,t})=I(\bar{\mathbf{Y}}_{1,t-j})+\frac{j}{t}\big(\bar{\mathbf{Y}}_{t-j+1,t}-\bar{\mathbf{Y}}_{1,t-j}\big)'\boldsymbol{\theta}_{\bar{\mathbf{Y}}_{1,t-j}}+O\Big((\frac{j}{t})^2\Big)\tag{A.9}$$

uniformly over $||\bar{\mathbf{Y}}_{t-j+1,t}||+||\bar{\mathbf{Y}}_{1,t-j}||\le B$. Putting $I(\boldsymbol{\mu})=\boldsymbol{\mu}'\boldsymbol{\theta}_{\boldsymbol{\mu}}-\psi(\boldsymbol{\theta}_{\boldsymbol{\mu}})$, with $\boldsymbol{\mu}=\bar{\mathbf{Y}}_{1,t-j}$ and $\boldsymbol{\mu}=\bar{\mathbf{Y}}_{t-j+1,t}$, into (A.9) yields (A.8).

First consider $t \le t_1^{(n)} + 2m$. By applying the Law of Large Numbers together with Lemma 1 and (A.8) to $\pi_{1,t}$, $\pi_{1,i-1}$, and $\pi_{i,t}$ for $t-m+1 \le i \le t$, we obtain that $\max_{t<t_1^{(n)}} |p_{1,t}^A - 1| \xrightarrow{P} 0$, $P\{\tau_1 \le t_1^{(n)}\} \to 0$ and

$$P\{\tau_1 \le t_1^{(n)} + m\} \to 1 \text{ as } p \to 0. \tag{A.10}$$

In particular, to derive (A.10), first apply Lemma 1 to obtain

$$\begin{aligned} &\log(\pi_{1,t-m}^{-1}\pi_{t-m+1,t}^{-1}) - \log(\pi_{1,t}^{-1}) \\ &\quad = (a_0 + t - m) I(\bar{\mathbf{Y}}_{1,t-m}) + (a_0 + m) I(\bar{\mathbf{Y}}_{t-m+1,t}) \\ &\qquad -(a_0 + t) I(\bar{\mathbf{Y}}_{1,t}) - \log\Big(\big[\frac{(a_0+t-m)(a_0+m)}{(a_0+t)}\big]^{d/2}\Big) + O_P(1). \end{aligned} \tag{A.11}$$

For $t > t_1^{(n)}$, combining (A.11) with (A.8) yields

$$\begin{aligned} &\log(\pi_{1,t-m}^{-1}\pi_{t-m+1,t}^{-1}) - \log(\pi_{1,t}^{-1}) + \log p \\ &\quad = m\{\mathcal{I}(\bar{\mathbf{Y}}_{t-m+1,t}, \bar{\mathbf{Y}}_{1,t-m}) + O(\frac{m}{t})\} - \log(\frac{1}{p}) + O_P(\log m). \end{aligned} \tag{A.12}$$

Recalling (C1) and (C2) and applying the Law of Large Numbers to (A.12), we obtain that, for $t \ge t_1^{(n)} + m$,

$$\log\Big(\frac{p\pi_{1t}}{\pi_{1,t-m}\pi_{t-m+1,t}}\Big) = m\Big\{\mathcal{I}(\boldsymbol{\mu}_{t_1^{(n)}}, \boldsymbol{\mu}_1) + o_P(1)\Big\} - \log\big(\frac{1}{p}\big) + O_P(\log m). \tag{A.13}$$

In view of (C2) and $m \sim |\log p|^{1+\epsilon}$, (A.13) shows that its left hand side becomes infinite in probability as $p \to 0$. Hence, in view of the definition (A.6) of τ_1, (A.10) follows.

Replacing $\mathcal{Y}_t$ in the preceding argument by $\mathcal{Y}_{\tau_1,t}$ then proves the corresponding results for τ_2 and $\max_{\tau_1<t<t_2^{(n)}} |p_{\tau_1,t}^A - 1|$. Proceeding inductively in this way then completes the proof.

Note that the AMOC filter weights (A.7) can be represented recursively by using $K_t = \max\{s \le t : I_t = 1\}$, as in (2.9) and (2.10) but with p_{it}^A in place of p_{it}. The analog of the set $\mathcal{K}_{t-1}(p)$ for the AMOC filter is $\mathcal{K}_{t-1}^A(p) = \{t-1, \ldots, t-(m \vee \tau(t)), \tau(t)\}$, where $\tau(t)$ is the largest τ_j that is $\le t-1$. Thus, the main difference between AMOC and the more flexible BCMIX is that AMOC allows one additional index $\tau(t)$ to be included in $\mathcal{K}_{t-1}^A(p)$ besides the most recent $t-1, \ldots, t-m$, while BCMIX allows $M-m$ more previous indices that need not be $\tau(t)$, thereby removing the "at most one change" requirement. Whereas AMOC filter weights have the explicit formula (A.7), which plays an important role in the proof of Lemma 2, the recursive representation (2.9)−(2.10) of BCMIX does not have a similar explicit formula. On the other hand, in view of (C1), (C2), and that $p = O(1/n)$, "at most one change" dominates "more than one change" in probability, and Lemma 2 and its proof can be used to prove the following lemma for BCMIX, in which we also weaken the assumption $M - m = 1$ for AMOC to $1 \le M - m = O(1)$.

Lemma 3. *As $n \to \infty$,*

$$\max_{t < t_1^{(n)}} |p_{1,t} - 1| \xrightarrow{P} 0. \tag{A.14}$$

Moreover, for $1 \le j \le k$ and $1 < \eta < 1 + \epsilon$,

$$\max_{t_j^{(n)} + |\log p|^\eta \le t < t_{j+1}^{(n)}} \Big| \sum_{i \in \mathcal{K}_t(p),\ t_j^{(n)} \le i \le t_j^{(n)} + M(p)} p_{i,t} - 1 \Big| \xrightarrow{P} 0. \tag{A.15}$$

Proof. First note that (A.14) basically says that $p_{1,t-1}$ behaves asymptotically like $p^A_{1,t-1}$ so that even though BCMIX allows $M - m \ge 1$ (instead of $M = m + 1$), the additional weights are negligible compared with $p_{1,t-1}$. Since the $p_{i,t}$ are defined recursively, (A.14) and (A.15) can be proved by induction on t. Concerning (A.15) with $j = 1$, we can use an argument similar to that in (A.11)-(A.13) to show that the weight $p_{1,t}$ is eliminated by time $t_1^{(n)} + |\log p|^\eta$, with probability approaching 1 as $n \to \infty$. For $t \ge t_1^{(n)} + |\log p|^\eta$, the weight $p^A_{\tau_1,t}$ in Lemma 2 is now replaced by the sum of weights $p_{i,t}$ in the set $\{i \in \mathcal{K}_t(p) : t_1^{(n)} \le i \le t_1^{(n)} + M(p)\}$. We can then modify the induction proof of (A.14) to prove (A.15) with $j = 1$ for the range $t_1^{(n)} + m(p) \le t < t_2^{(n)}$, and then proceed to $j = 2, \dots, k$.

For the backward BCMIX filter, a time-reversal argument establishes the analogs of (A.14) and (A.15) for $q_{i,t}$. In particular, the analog of (A.14) is $\max_{t > t_k^{(n)}} |q_{n,t} - 1| \xrightarrow{P} 0$. Combining these results on the forward and backward BCMIX filter weights via (2.13) yields the following.

Lemma 4. *Let $1 \le \nu \le k$ and $1 < \eta < 1 + \epsilon$. Then as $n \to \infty$,*

$$\max_{t < t_1^{(n)}} \Big| \sum_{j \in \widetilde{\mathcal{K}}_t(p),\ t_1^{(n)} - M(p) \le j \le t_1^{(n)}} \widetilde{\beta}_{1jt} - 1 \Big| \xrightarrow{P} 0,$$

$$\max_{t > t_k^{(n)}} \Big| \sum_{i \in \mathcal{K}_t(p),\ t_k^{(n)} \le i \le t_k^{(n)} + M(p)} \widetilde{\beta}_{int} - 1 \Big| \xrightarrow{P} 0,$$

$$\max_{t_\nu^{(n)} + |\log p|^\eta \le t \le t_{\nu+1}^{(n)} - |\log p|^\eta} \Big| \sum_{i \in \mathcal{K}_t(p),\ j \in \widetilde{\mathcal{K}}_t(p),\ t_\nu^{(n)} \le i \le t_\nu^{(n)} + M(p),\ t_{\nu+1}^{(n)} - M(p) \le j \le t_{\nu+1}^{(n)}} \widetilde{\beta}_{ijt} - 1 \Big| \xrightarrow{P} 0.$$

Proof of Theorem 2. From large deviation bounds in the exponential family, it follows that for $1 \le \nu \le k$,

$$\max_{t_\nu^{(n)} \le i \le t_\nu^{(n)}+M(p),\ t_{\nu+1}^{(n)}-M(p) \le j \le t_{\nu+1}^{(n)}} ||\bar{\mathbf{Y}}_{i,j} - \boldsymbol{\mu}_{t_\nu^{(n)}}|| \xrightarrow{P} 0,$$

and that $\max_{t_1^{(n)}-M(p) \le j < t_1^{(n)}} ||\bar{\mathbf{Y}}_{1,j} - \boldsymbol{\mu}_1|| \xrightarrow{P} 0$. Combining this with Lemma 4, (C1) and (4.4) then yields the desired conclusion.

Proof of Theorem 3. Let $\delta_i^{(n)} = ||\boldsymbol{\mu}_{t_i^{(n)}} - \boldsymbol{\mu}_{t_{i-1}^{(n)}}||$ and order them as $\delta_{[1]}^{(n)} \ge \delta_{[2]}^{(n)} \ge \cdots \ge \delta_{[k]}^{(n)}$. This ordering induces a corresponding ordering $t_{[j]}^{(n)}$ of the $t_i^{(n)}$; in case of ties with $\delta_{[j_1]} = \cdots = \delta_{[j_l]}$, we can choose an appropriate permutation of $[j_1], \ldots, [j_l]$ to order the corresponding $t_{[j]}^{(n)}$. We next show that

$$\max_{1 \le j \le k} \frac{\left|\widehat{\tau}_j - t_{[j]}^{(n)}\right|}{m(p)} \xrightarrow{P} 0. \tag{A.16}$$

Recall that the $\widehat{\tau}_j$ are defined by (4.8) which involves $\Delta_t = ||\widehat{\boldsymbol{\mu}}_{t+b(p)} - \widehat{\boldsymbol{\mu}}_{t-b(p)}||^2$. By (4.5) (which follows from Theorem 2), $\Delta_t = ||\boldsymbol{\mu}_{t+b(p)} - \boldsymbol{\mu}_{t-b(p)}||^2 + o_p(1)$, in which the $o_p(1)$ term is uniform in $t \in \{1, \ldots, n\}$ such that $\min_{1 \le i \le k} |t - t_i^{(n)}| \ge |\log p|^\eta$. Since $I(\boldsymbol{\mu}) = \boldsymbol{\theta}_{\boldsymbol{\mu}}' \boldsymbol{\mu} - \psi(\boldsymbol{\theta}_{\boldsymbol{\mu}})$, we can write

$$\Sigma_{t=\widehat{t}_{(j-1),k}}^{\widehat{t}_{(j),k}-1} \log f_{\widehat{\boldsymbol{\theta}}^{(j)}}(\mathbf{y}_t) = \left(\widehat{t}_{(j),k} - \widehat{t}_{(j-1),k}\right) I\left(\frac{\Sigma_{t=\widehat{t}_{(j-1),k}}^{\widehat{t}_{(j),k}-1} \mathbf{y}_t}{\left(\widehat{t}_{(j),k} - \widehat{t}_{(j-1),k}\right)}\right), \tag{A.17}$$

Putting (A.17) into the right hand side of (4.9) that defines $\Lambda_n(j)$, apply the Law of Large Numbers to $\Lambda_n(j)$ with $j < k$, and the Functional Central Limit Theorem to $\Lambda_n(j)$ with $j \ge k$, to conclude from (A.16) and a variant of (A.8) that $\widehat{k}_n \xrightarrow{P} k$.

References

Albert, J. H. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Bus. Econ. Statist.* **11**, 1-15.

Bai, J. (1997a). Estimation of a changepoint in multiple regression models. *Rev. Econ. Statist.* **79**, 551-563.

Bai, J. (1997b). Estimation multiple breaks one at a time. *Econometric Theory* **13**, 315-352.

Bai, J. and Perron, P. (1998). Testing for and estimation of multiple structural changes. *Econometrica* **66**, 817-858.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18**, 1-22.

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20**, 260-279.

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

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. European Math. Soc.* **3**, 203-268.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experiment crosses. *J. Roy. Statist. Soc. Ser. B* **64**, 641-656.

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time (with discussion). *J. Roy. Statist. Soc. Ser. B* **37**, 149-192.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.* **41**, 389-405.

Chan, H. P. and Lai, T. L. (2003). Saddlepoint approximations for Markov random walks and nonlinear boundary crossing probabilities for Markov random walks. *Ann. Appl. Probab.* **13**, 395-429.

Chan, H. P. and Lai, T. L. (2006). Maxima of asymptotically Gaussian random fields and moderate deviation approximations to boundary crossing probabilities for sums of random variables with multidimensional indices. *Ann. Probab.* **34**, 80-121.

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

Chib, S. (1998). Estimation and comparison of multiple change-point models. *J. Econometrics* **86**, 221-241.

Chib, S., Nardari, F. and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *J. Econometrics* **108**, 281-316.

Csörgö, M. and Horvath, L. (1998). *Limit Theorems in Change-Point Analysis.* Wiley, Chichester, England.

Davis, R., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101**, 223-239.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269-281.

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

Hinkley, D. (1970). Inference about the change point in a sequence of random variables. *Biometrika* **57**, 1-17.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

James, B., James, K. L. and Siegmund, D. (1987). Tests for a change-point. *Biometrika* **74**, 71-83.

Lai, T. L., Liu, H. and Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statist. Sinica* **15**, 279-301.

Lai, T. L., Liu, T. and Xing, H. (2009). A Bayesian approach to sequential surveillance in exponential families. To appear in *Comm. Statist. Theory Methods*, Special Issue in honor of S. Zacks.

Lai, T. L., Xing, H. and Zhang, N. (2008). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics* **9**, 290-307.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85**, 1501-1510.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15**, 38-52.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *J. Amer. Statist. Assoc.* **88**, 968-978.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* **42**, 523-527.

Pettitt, A. N. (1980). A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika* **67**, 79-84.

Qu, Z. and Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica* **75**, 459-502.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53**, 873-880.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *J. Amer. Statist. Assoc.* **55**, 324-330.

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

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.* **8**, 22-46.

Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika* **91**, 785-800.

Vostrikova, L. J. (1981). Detecting disorder in multidimensional random processes. *Sov. Math. Doklady* **24**, 55-59.

Wang, J. and Zivot, E. (2000). A Bayesian time series model of multiple structural changes in level, trend, and variance. *J. Bus. Econ. Statist.* **18**, 374-386.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* **73**, 91-104.

Yao, Y. (1984). Estimation of a noisy discrete-time step functions: Bayes and empirical Bayes approach. *Ann. Statist.* **12**, 1434-1447.

Yao, Y. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6**, 181-189.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrika* **44**, 1019-1031.

Zhang, N. and Siegmund, D. (2006). A modified Bayes information criterion with applications to comparative genomic hybridization data. *Biometrics* **63**, 22-32.

Department of Statistics, Stanford University, Stanford, CA 94305-4065, U.S.A.
E-mail: lait@stanford.edu

Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794, U.S.A.
E-mail: xing@ams.sunysb.edu

(Received March 2009; accepted December 2009)