Statistica Sinica **15**(2005), 279-301

AUTOREGRESSIVE MODELS WITH PIECEWISE CONSTANT VOLATILITY AND REGRESSION PARAMETERS

Tze Leung Lai, Haiyan Liu and Haipeng Xing

Stanford University

Abstract: We introduce herein a new class of autoregressive models in which the regression parameters and error variances may undergo changes at unknown time points while staying constant between adjacent change-points. Assuming conjugate priors, we derive closed-form recursive Bayes estimates of the regression parameters and error variances. Approximations to the Bayes estimates are developed that have much lower computational complexity and yet are comparable to the Bayes estimates in statistical efficiency. We also address the problem of unknown hyperparameters and propose two practical methods for simultaneous estimation of the hyperparameters, regression parameters and error variances.

Key words and phrases: Bayesian inference, bounded complexity mixtures, change-point problems, filtering, sequential Monte Carlo, smoothing.

1. Introduction

The problem of modeling a time series whose parameters may undergo occasional changes arises in many engineering, econometric and biomedical applications, and has an extensive literature widely scattered in these fields besides statistics. In stochastic dynamical systems, if the parameters may change with time, it is more convenient to regard them as states. However, this requires specification or modeling of the dynamics of the parameters. A Bayesian approach to this problem is to take a stochastic process as the prior distribution for the time-varying parameters, whose posterior distribution then provides an estimate of the current parameter value given the current and past observations. This reduces the estimation problem to a filtering problem. Except for the special case in which the system and parameter dynamics can be represented by a linear Gaussian state-space model so that Kalman filtering can be applied, the optimal filter is typically nonlinear and infinite-dimensional. Even for the simple mean shift model $Y_t = \theta_t + \epsilon_t$, $t = 1, 2, \ldots$, in which (i) the ϵ_t are i.i.d. zero-mean normal random variables, (ii) the sequence of change-times of $\{\theta_t\}$ forms a discrete renewal process with geometric interarrival times with parameter p, and (iii) the post-change values of $\{\theta_t\}$ are i.i.d. normal, the unknown times of the occurrence of the mean shifts leads to great complexity of the Bayes estimate $\widehat{\theta}_n = E(\theta_n|Y_1, \ldots, Y_n)$. Chernoff and Zacks (1964) gave a closed-form expression of $\widehat{\theta}_n$ that requires $O(2^n)$ operations to compute. Yao (1984) later found another representation of the estimate that requires only $O(n^2)$ operations. By combining forward and backward (i.e., time-reversed) filters to solve the smoothing problem of estimating θ_t from $Y_1, \ldots, Y_n$ for $1 \le t \le n$, he also gave a formula for the Bayes estimate $E(\theta_t \mid Y_1, \ldots, Y_n)$ that requires $O(n^3)$ operations to compute.

A natural extension of the mean shift model is the regression model $Y_t = \boldsymbol{\theta}_t^T X_t + \epsilon_t$, in which the regressors X_t are random vectors that may depend on the past observations (X_i, Y_i), $i \le t-1$. In this paper we consider the special case $X_t = (1, Y_{t-1}, \ldots, Y_{t-k})^T$ that corresponds to the AR(k) model in the time series literature. Motivated by applications to financial econometrics, in which not only the levels but also the volatilities of asset returns are of basic interest, we consider a further extension to the case where changes in $\sigma_t^2 := \mathrm{Var}(\epsilon_t)$ can also occur besides changes in θ_t. These extensions, given in Section 2 for the filtering problem and in Section 3 for the smoothing problem, make use of certain formulas for Bayesian inference in normal populations, a comprehensive introduction to which can be found in Box and Tiao's (1973) classic. As pointed out in Section 3, a major difference between our and Yao's models is in the dynamics, which only involve parameter changes in Yao's model but also include autoregressive dynamics in ours. Accordingly some restrictions have to be imposed on the autoregressive parameters $\boldsymbol{\theta}_t$ to ensure a stationary or non-explosive dynamical system. These restrictions lead to considerably more complicated filters and smoothers than those in Yao's (1984) mean shift model.

Even with a Yao-type algorithm in the mean shift case, the complexity of the Bayes estimates $\widehat{\theta}_n$ becomes unmanageable when n is large, as pointed out by Lai, Liu and Xing (2004). In Section 4 we develop two recursive approximations to the Bayes estimates $E\{(\boldsymbol{\theta}_t^T, \sigma_t) \mid Y_1, \ldots, Y_n\}$ for $t \le n$ that can be updated with a fixed number (not depending on n) of operations. One approximation, called BCMIX (bounded complexity mixture), uses only a fixed number of filters in the extensions of Yao's algorithm in Sections 2 and 3. The other approximation, called SISR (sequential importance sampling with resampling), is a Monte Carlo approximation that uses a fixed and relatively small number of trajectories that are recursively simulated by importance sampling. Numerical results showing the efficiency of these approximations are provided.

Sections 2-4 assume p, the hyperparameter of the Bayesian model, to be correctly specified. Without assuming p to be known, Section 5 describes some computationally convenient (when used in conjunction with BCMIX or SISR) estimators of p that have good statistical performance. In practice the assumed change-point autoregressive model is only an approximation to the actual data generating mechanism of the observed time series, and the purpose of fitting the model is to derive forecasts of future values that are yet to be observed. The AR(k) model provides simple forecasts of future observations, and allowing the regression coefficients to change over time yields flexible non-linear predictors. Allowing also σ_t to change over time can account for relatively calm periods punctuated by highly volatile periods observed in stock price movements and other econometric time series. By using normal mixtures, the change-point model can also adapt to various distributional forms of Y_t. Markets respond to policy changes, announcements of companies' earnings and a myriad of perceived or real changes of the economy, so it is useful to incorporate uncertain (random) change-points in modeling econometric time series. Box and Tiao's (1975) seminal work on *intervention analysis* represents one of the major directions in this kind of modeling. Section 6 compares intervention analysis with change-point autoregression and gives some concluding remarks.

2. A Bayesian Change-point Model and Filters Estimating $\boldsymbol{\theta_t}$ and $\boldsymbol{\sigma_t}$

An autoregressive model with piecewise constant volatility and regression parameters has the form

$$Y_t = \mu_t + \alpha_{1,t} Y_{t-1} + \cdots + \alpha_{k,t} Y_{t-k} + \sigma_t \epsilon_t, \qquad t > k, \tag{2.1}$$

where the ϵ_t are i.i.d. unobservable random disturbances with mean 0 and variance 1, and $\boldsymbol{\theta}_t = (\mu_t, \alpha_{1,t}, \ldots, \alpha_{k,t})^T$ and σ_t are piecewise constant parameters. A Bayesian modeling approach requires also specification of the distributions of ϵ_t and $\{(\boldsymbol{\theta}_t^T, \sigma_t), t > k\}$. Following Yao (1984), we assume that the sequence of change-times of $(\boldsymbol{\theta}_t^T, \sigma_t)$ forms a discrete renewal process with parameter p or, equivalently,

$$I_t := 1_{\{(\boldsymbol{\theta}_t^T, \sigma_t) \neq (\boldsymbol{\theta}_{t-1}^T, \sigma_{t-1})\}} \text{ are i.i.d. Bernoulli random variables with } P(I_t = 1) = p \tag{2.2}$$

for $t \geq k+2$ and $I_{k+1} = 1$. Another assumption underlying Yao's closed-form expressions for Bayes estimates in the mean shift model is normal ϵ_t and $\boldsymbol{\theta}_t$, resulting in normal mixtures for the posterior distributions of the conjugate prior. We generalize this idea by assuming at change-times an inverse gamma prior distribution for σ_t^2 and a normal prior distribution for $\boldsymbol{\theta}_t$ given σ_t. Specifically, letting $\tau_t = (2\sigma_t^2)^{-1}$, we assume that

$$(\boldsymbol{\theta}_t^T, \tau_t) = (1 - I_t)(\boldsymbol{\theta}_{t-1}^T, \tau_{t-1}) + I_t(\mathbf{Z}_t^T, \gamma_t),$$

where $(\mathbf{Z}_1^T, \gamma_1), (\mathbf{Z}_2^T, \gamma_2), \ldots$ are i.i.d. random vectors such that

$$\gamma_t \sim \text{Gamma}(g, \lambda), \quad \mathbf{Z}_t \mid \gamma_t \sim \text{Normal}(\mathbf{z}, \mathbf{V}/(2\gamma_t)). \tag{2.3}$$

When there is no change-point (i.e., $p = 0$), the posterior distribution of τ_t is still gamma while that of $\boldsymbol{\theta}_t$ given τ_t is multivariate normal, and there are simple formulas for updating the shape and scale parameters of the gamma distribution and the normal mean and covariance matrices; see Section 2.7 of Box and Tiao (1973). In the presence of change-points, the posterior distribution of $(\boldsymbol{\theta}_t^T, \tau_t)$ is a mixture of gamma-normal distributions and we now extend Yao's (1984) algorithm to evaluate the parameters of the posterior distribution.

As in Yao's algorithm, the most recent change-time $J_n := \max\{t \le n : I_t = 1\}$ plays a basic role in computing the Bayes estimate $E\{(\boldsymbol{\theta}_n^T, \sigma_n^2) \mid Y_1, \ldots, Y_n\}$. Let $\mathbf{Y}_{t,n} = (1, Y_n, \ldots, Y_t)^T$. Recalling that $\tau_n = (2\sigma_n^2)^{-1}$, the conditional distribution of $(\boldsymbol{\theta}_n^T, \tau_n)$ given $(J_n, \mathbf{Y}_{J_n,n})$ can be described by

$$\tau_n \sim \text{Gamma}\Big(g + \frac{n - J_n + 1}{2}, \frac{1}{a_{J_n,n}}\Big), \quad \boldsymbol{\theta}_n \mid \tau_n \sim \text{Normal}\Big(\mathbf{z}_{J_n,n}, \frac{1}{2\tau_n}\mathbf{V}_{J_n,n}\Big), \tag{2.4}$$

where for $k < j \le n$,

$$\mathbf{V}_{j,n} = \Big(\mathbf{V}^{-1} + \sum_{t=j}^{n} \mathbf{Y}_{t-k,t-1}\mathbf{Y}_{t-k,t-1}^T\Big)^{-1}, \quad \mathbf{z}_{j,n} = \mathbf{V}_{j,n}\Big(\mathbf{V}^{-1}\mathbf{z} + \sum_{t=j}^{n} \mathbf{Y}_{t-k,t-1}Y_t\Big),$$
$$a_{j,n} = \lambda^{-1} + \mathbf{z}^T\mathbf{V}^{-1}\mathbf{z} + \sum_{t=j}^{n} Y_t^2 - \mathbf{z}_{j,n}^T\mathbf{V}_{j,n}^{-1}\mathbf{z}_{j,n}. \tag{2.5}$$

Note that if $(2X)^{-1}$ has a Gamma$(\widetilde{g}, \widetilde{\lambda})$ distribution, then X has the inverse gamma IG(g, λ) distribution with $g = \widetilde{g}$, $\lambda = 2\widetilde{\lambda}$, and that $EX = \lambda^{-1}(g-1)^{-1}$ when $g > 1$ and $E\sqrt{X} = \lambda^{-1/2}\Gamma(g - (1/2))/\Gamma(g)$. It then follows from (2.4) that

$$\begin{aligned} E(\boldsymbol{\theta}_n^T, \sigma_n^2 \mid \mathbf{Y}_{1,n}) &= \sum_{j=k+1}^{n} p_{j,n} E(\boldsymbol{\theta}_n^T, \sigma_n^2 \mid \mathbf{Y}_{1,n}, J_n = j) \\ &= \sum_{j=k+1}^{n} p_{j,n}\Big(\mathbf{z}_{j,n}^T, \frac{a_{j,n}}{2g + n - j - 1}\Big), \end{aligned} \tag{2.6}$$

where $p_{j,n} = P(J_n = j \mid \mathbf{Y}_{1,n})$. Moreover, $E(\sigma_n \mid \mathbf{Y}_{1,n}) = \Sigma_{j=k+1}^{n} p_{j,n}(a_{j,n}/2)^{1/2}$ Γ_{n-j}, where $\Gamma_i = \Gamma(g + i/2)/\Gamma(g + (i+1)/2)$.

The next step is to derive a recursive formula for $p_{j,n}$, as in Yao (1984) for the mean shift problem. Denoting conditional densities by $f(\cdot \mid \cdot)$, note that with the arrival of the new observation Y_n at time n, we can update the posterior density of $(\boldsymbol{\theta}_n^T, \tau_n)$ by

$$f(\boldsymbol{\theta}_n^T, \tau_n \mid \mathbf{Y}_{1,n}) = p_{n,n} f(\boldsymbol{\theta}_n^T, \tau_n \mid \mathbf{Y}_{1,n}, J_n = n) + \sum_{j=k+1}^{n-1} p_{j,n} f(\boldsymbol{\theta}_n^T, \tau_n \mid \mathbf{Y}_{1,n}, J_n = j),$$

where

$$p_{j,n} \propto p^*_{j,n} := \begin{cases} pf(Y_n \mid I_n = 1) & \text{if} \quad j = n, \\ (1-p)p_{j,n-1} f(Y_n \mid \mathbf{Y}_{j,n-1}, J_n = j) & \text{if} \quad j \leq n-1. \end{cases} \tag{2.7}$$

Since $\sum_{i=k+1}^n p_{i,n} = 1$, $p_{j,n}$ is given explicitly by $p_{j,n} = p^*_{j,n}/\sum_{i=k+1}^n p^*_{i,n}$. Moreover, noting that $Y_n = \boldsymbol{\theta}_n^T \mathbf{Y}_{n-k,n-1} + \sigma_n \epsilon_n$, we can apply the following lemma to obtain an explicit formula for the conditional density function of Y_n given $J_n = j$ and $\mathbf{Y}_{j,n-1}$.

Lemma 1. (i) *Suppose that the conditional distribution of Y given $\boldsymbol{\theta}$ and τ is* Normal$(\boldsymbol{\theta}^T \boldsymbol{\phi}, (2\tau)^{-1})$, $\tau \sim$ Gamma$(\widetilde{g}, \widetilde{\lambda})$ *and that the conditional distribution of $\boldsymbol{\theta}$ given τ is* Normal$(\mathbf{z}, \mathbf{V}/(2\tau))$. *Then*

$$(Y_t - \mathbf{z}^T \boldsymbol{\phi})/\{(1 + \boldsymbol{\phi}^T \mathbf{V} \boldsymbol{\phi})/(2\widetilde{g}\widetilde{\lambda})\}^{\frac{1}{2}} \tag{2.7}$$

has a Student-t distribution with $2\widetilde{g}$ degrees of freedom.

(ii) *Given $J_n = j$ and $\mathbf{Y}_{j,n-1}$, the conditional distribution of $(\boldsymbol{\theta}_n^T, \gamma_n)$ is*

$$\tau_n \sim \text{Gamma}(g + \frac{n-j}{2}, \frac{1}{a_{j,n-1}}), \boldsymbol{\theta}_n \mid \tau_n \sim \text{Normal}(\mathbf{z}_{j,n-1}, \frac{\mathbf{V}_{j,n-1}}{2\tau_n}) \ \textit{if} \ j < n;$$

$$\tau_n \sim \text{ Gamma } (g, \lambda), \quad \boldsymbol{\theta}_n \mid \tau_n \sim \text{Normal}(\mathbf{z}, \frac{\mathbf{V}}{2\tau_n}) \ \textit{if} \ j = n.$$

Proof. (i) follows from $f(y) = \iint f(y \mid \boldsymbol{\theta}, \tau) f(\boldsymbol{\theta} \mid \tau) f(\tau) d\boldsymbol{\theta} d\tau$ and using a change of variables to perform the integration. To prove(ii), apply (2.4) with n replaced by $n-1$ in the case $J_n < n$ (so that $J_n = J_{n-1}$). When $J_n = n$, $(\boldsymbol{\theta}_n^T, \tau_n)$ has a jump at time n and its distribution follows (2.3).

3. Bayesian Smoothers for $\boldsymbol{\theta}_t$ and $\boldsymbol{\sigma}_t$

For the simple mean shift model described in Section 1 (which corresponds to $k = 0$ and $\sigma_t = \sigma$ known), Yao's (1984) algorithm for computing the Bayes estimate $E(\theta_t \mid Y_1, \ldots, Y_n)$ with $1 \leq t \leq n$ is based on combining the forward filter involving the posterior distribution of θ_t given $Y_1, \ldots, Y_t$ and the backward filter involving the conditional distribution of θ_t given $Y_{t+1}, \ldots, Y_n$. In view of the reversibility property that $(Y_1, \ldots, Y_n)$ has the same distribution as $(Y_n, \ldots, Y_1)$, the backward filter has the same structure as the forward predictor of θ_t based on the past $n-t$ observations. For the change-point autoregressive model (2.1), reversibility cannot hold because the normal distribution for $\boldsymbol{\theta}_t$ gives positive probability to the explosive region $\{\boldsymbol{\theta} = (\mu, \alpha_1, \ldots, \alpha_k)^T : 1 - \alpha_1 z - \cdots - \alpha_k z^k$ has roots inside the unit circle$\}$. On the other hand, if we replace the normal distribution in (2.3) by a truncated normal distribution that has support in some stability region C such that

$$\inf_{|z|\leq 1} |1-\alpha_1 z - \cdots - \alpha_k z^k| > 0 \quad \text{if } \boldsymbol{\theta} = (\mu, \alpha_1, \ldots, \alpha_k)^T \in C, \tag{3.1}$$

then the following theorem shows that the Markov chain $(\tau_t, \boldsymbol{\theta}_t^T, \mathbf{Y}_{t-k+1,t})$ has a stationary distribution and is reversible if it is initialized at the stationary distribution. We use the prefix $\mathbf{T}_C$ to denote truncation of a distribution within the region C. In particular, $\mathbf{T}_C\text{Normal}(\mathbf{z}, \mathbf{V})$ denotes the conditional distribution of $\mathbf{Z}$ given $\mathbf{Z} \in C$, where $\mathbf{Z}$ is Normal$(\mathbf{z}, \mathbf{V})$.

Theorem 1. *Suppose* (2.3) *is modified as*

$$\gamma_t \sim \text{Gamma}(g, \lambda), \quad \mathbf{Z}_t \mid \gamma_t \sim \mathbf{T}_C\text{Normal}(\mathbf{z}, \mathbf{V}/(2\gamma_t)), \tag{3.2}$$

with the region C satisfying the stability condition (3.1). *Then $(\tau_t, \boldsymbol{\theta}_t^T, \mathbf{Y}_{t-k+1,t})$ has a stationary distribution under which $(\boldsymbol{\theta}_t^T, \tau_t)$ has the same distribution as $(\mathbf{Z}_t^T, \gamma_t)$ in* (3.2) *and*

$$Y_t \mid (\boldsymbol{\theta}_t^T, \tau_t) \sim \text{ Normal } (\mu_t/(1-\alpha_{1,t} - \ldots - \alpha_{k,t}),\ (2\tau_t)^{-1} v_t),$$

where $v_t = \sum_{j=0}^{\infty} \beta_{j,t}^2$ and $\beta_{j,t}$ are the coefficients in the power series representation of $1/(1-\alpha_{1,t}z - \cdots - \alpha_{k,t}z^k) = \sum_{j=0}^{\infty} \beta_{j,t} z^j$ for $|z| \leq 1$. Moreover, the Markov chain $(\tau_t, \boldsymbol{\theta}_t^T, \mathbf{Y}_{t-k+1,t})$ is reversible if it is initialized at the stationary distribution.

The proof of Theorem 1 is given in Appendix A. Since $(\tau_t, \boldsymbol{\theta}_t^T, Y_t)$ is reversible, the backward filter of $(\tau_t, \boldsymbol{\theta}_t^T)$ based on $Y_n, \ldots, Y_{t+1}$ has the same structure as the forward predictor based on the past $n-t$ observations prior to t. As in Proposition 4.2 of Yao (1984), we can apply Bayes' theorem to combine the forward and backward filters, yielding

$$f(\tau_t, \boldsymbol{\theta}_t \mid \mathbf{Y}_{1,n}) \propto f(\tau_t, \boldsymbol{\theta}_t \mid \mathbf{Y}_{1,t}) f(\tau_t, \boldsymbol{\theta}_t \mid \mathbf{Y}_{t+1,n})/\pi(\tau_t, \boldsymbol{\theta}_t), \tag{3.3}$$

where π denotes the stationary density function, which is the same as that of $(\gamma_t, \mathbf{Z}_t)$ given in (3.2).

Because the truncated normal is used in lieu of the normal prior distribution in (3.1), the filtering formulas in Section 2 need to be modified somewhat. Specifically, the conditional distribution of $\boldsymbol{\theta}_n$ given τ_n needs to be replaced by $\boldsymbol{\theta}_n \mid \tau_n \sim \mathbf{T}_C\text{Normal}(\mathbf{z}_{J_n,n}, \mathbf{V}_{J_n,n}/(2\tau_n))$, while (2.5) defining $\mathbf{V}_{j,n}$, $\mathbf{z}_{j,n}$ and $a_{j,n}$ remains unchanged. Moreover, Lemma 1 that gives the conditional density of Y_n given J_n and $\mathbf{Y}_{J_n,n-1}$ (which is used in the updating formula (2.7) for the weights $p_{j,n}$) needs to be modified as follows.

Lemma 2. (i) *Suppose that the conditional distribution of* Y *given* $\boldsymbol{\theta}$ *and* τ *is* $\mathrm{Normal}(\boldsymbol{\theta}^T\boldsymbol{\phi},(2\tau)^{-1})$, $\tau\sim\mathrm{Gamma}(\widetilde{g},\widetilde{\lambda})$ *and that the conditional distribution of* $\boldsymbol{\theta}$ *given* τ is $\mathbf{T}_C\mathrm{Normal}(\mathbf{z},\mathbf{V}/(2\tau))$. *Then* (2.8) *has density function* f_C *which approaches the Student-t density with* $2\widetilde{g}$ *degrees of freedom as the truncation region* C *approaches the support* $\mathbf{R}^{k+1}$ *of the normal distribution.*

(ii) *Given* $J_n=j$ *and* $\mathbf{Y}_{j,n-1}$, *the conditional distribution of* $(\boldsymbol{\theta}_n^T,\tau_n)$ *is*

$$\tau_n\sim\mathrm{Gamma}(g+\frac{n-j}{2},\frac{1}{a_{j,n-1}}),\boldsymbol{\theta}_n\mid\tau_n\sim\mathbf{T}_C\mathrm{Normal}(\mathbf{z}_{j,n-1},\frac{\mathbf{V}_{j,n-1}}{2\tau_n})\ \textit{if}\ j<n;$$

$$\tau_n\sim\mathrm{Gamma}(g,\lambda),\ \boldsymbol{\theta}_n\mid\tau_n\sim\mathbf{T}_C\mathrm{Normal}(\mathbf{z},\frac{\mathbf{V}}{2\tau_n})\ \textit{if}\ j=n.$$

4. Bounded Complexity Approximations

For the mean shift model described in Section 1, Lai, Liu and Xing (2004) introduced a bounded complexity mixture (BCMIX) approximation to the Bayesian filter $E(\theta_t\mid\mathbf{Y}_{1,t})$, while Chen and Lai (2004) developed sequential Monte Carlo approximations to $E(\theta_t\mid\mathbf{Y}_{1,t})$ that involve a fixed and relatively small number of trajectories simulated by sequential importance sampling with resampling (SISR). In this section, we show how bounded complexity estimates of $\boldsymbol{\theta}_t$ and σ_t can be developed by extending BCMIX and SISR to the change-point model (2.1).

4.1. BCMIX filters

Although the Bayes filter uses a recursive updating formula (2.7) for the weights $p_{j,n}$ $(k<j\le n)$, the number of weights increases with n, resulting in unbounded computational complexity and memory requirements in estimating σ_n and $\boldsymbol{\theta}_n$ as n keeps increasing. A simple idea to maintain bounded complexity is to keep only a fixed number n_p of weights at every stage n (which is tantamount to setting the other weights to be 0). Lai, Liu and Xing (2004) proposed to keep the most recent m_p weights $p_{j,n}$ (with $n-m_p<j\le n$) and the largest n_p-m_p of the remaining weights, where $1\le m_p<n_p$. Specifically, the updating formula (2.7) for the weights $p_{j,n}$ is modified as follows. Let $\mathcal{K}_{n-1}$ denote the set of indices j so that $p_{j,n-1}$ is kept at stage $n-1$; thus $\mathcal{K}_{n-1}\supset\{n-1,\dots,n-m_p\}$. At stage n, define $p^*_{j,n}$ by (2.7) for $j\in\{n\}\cup\mathcal{K}_{n-1}$ and let i_n be the index not belonging to $\{n,n-1,\dots,n-m_p+1\}$ such that

$$p^*_{i_n,n}=\min\{p^*_{j,n}:j\in\mathcal{K}_{n-1}\quad\text{and}\quad j\le n-m_p\},\tag{4.1}$$

choosing i_n to be the one farthest from n if the minimizing set in (4.1) has more than one element. Define $\mathcal{K}_n=\{n\}\cup(\mathcal{K}_{n-1}-\{i_n\})$ and let $p_{j,n}=p^*_{j,n}/\sum_{i\in\mathcal{K}_n}p^*_{j,n}$ for $j\in\mathcal{K}_n$.

When the prior distribution of $\boldsymbol{\theta}_t$ is a truncated normal instead of normal, the conditional density functions f in (2.7) do not have simple closed-form expressions and the optimal filter can be implemented via Lemma 2 by numerical integration. Since the constraint set C only serves to generate non-explosive observations, but has little effect on the values of the weights $p_{j,n}$ and on the performance of the BCMIX estimates that compute the posterior means via (2.6), we propose to ignore the truncation and simply apply the formulas in Lemma 1 that are derived under the normal prior assumption (instead of Lemma 2 that entails computation of multivariate integrals). In this connection, we also use some fast algorithms to compute the Student-t density by taking logarithms and applying saddlepoint approximations when the number of degrees of freedom is large. Moreover, in view of (2.5), $\mathbf{V}_{j,n}$, $\mathbf{z}_{j,n}$ and $a_{j,n}$ $(j \leq n)$ can be updated recursively (in n) by making use of the matrix inversion lemma (cf., Caines (1988, p.824)):

$$\begin{aligned}
\mathbf{V}_{j,n} &= \mathbf{V}_{j,n-1} - \frac{\mathbf{V}_{j,n-1}\mathbf{Y}_{n-k,n-1}\mathbf{Y}^T_{n-k,n-1}\mathbf{V}_{j,n-1}}{1+\mathbf{Y}^T_{n-k,n-1}\mathbf{V}_{j,n-1}\mathbf{Y}_{n-k,n-1}}, \text{ if } j < n,\\
\mathbf{V}_{n,n} &= \mathbf{V} - \frac{\mathbf{V}\mathbf{Y}_{n-k,n-1}\mathbf{Y}^T_{n-k,n-1}\mathbf{V}}{1+\mathbf{Y}^T_{n-k,n-1}\mathbf{V}\mathbf{Y}_{n-k,n-1}},\\
\mathbf{z}_{j,n} &= \mathbf{z}_{j,n-1} + \frac{\mathbf{V}_{j,n-1}\mathbf{Y}_{n-k,n-1}(Y_n - \mathbf{Y}^T_{n-k,n-1}\mathbf{z}_{j,n-1})}{1+\mathbf{Y}^T_{n-k,n-1}\mathbf{V}_{j,n-1}\mathbf{Y}_{n-k,n-1}},\\
a_{j,n} &= a_{j,n-1} + \frac{(Y_n - \mathbf{Y}^T_{n-k,n-1}\mathbf{z}_{j,n-1})^2}{1+\mathbf{Y}^T_{n-k,n-1}\mathbf{V}_{j,n-1}\mathbf{Y}_{n-k,n-1}}.
\end{aligned} \tag{4.2}$$

4.2. SISR filters

The Bayes estimate (2.6) can be rewritten in the form

$$E(\boldsymbol{\theta}^T_n, \sigma^2_n \mid \mathbf{Y}_{1,n}) = E\{\mathbf{z}^T_{J_n,n},\ a_{J_n,n}/(2g+n-J_n-1) \mid \mathbf{Y}_{1,n}\}, \tag{4.3}$$

which can be computed by Monte Carlo simulations using the conditional distribution of J_n given $\mathbf{Y}_{1,n}$. Let $\mathbf{I}_t = (I_{k+1}, \ldots, I_t)$. It is difficult to sample $\mathbf{I}_n$ directly from its conditional distribution given $\mathbf{Y}_{1,n}$. A basic idea behind sequential importance sampling is to sample $I_1, \ldots, I_n$ sequentially from an alternative distribution Q under which $I_t \mid \mathbf{I}_{t-1}$ has the same distribution as $P(I_t = \cdot \mid \mathbf{I}_{t-1}, \mathbf{Y}_{1,t})$, which is Bernoulli assuming the values 1 and 0 with probabilities in the proportion

$$pf(Y_t \mid I_t = 1) : (1-p)f(Y_t \mid \mathbf{Y}_{J_{t-1},t-1}, I_t = 0), \tag{4.4}$$

in which the conditional densities are Student-t densities given by Lemma 1. Letting $a_t(p)$ and $b_t(p)$ denote the two terms in (4.4), note that $f(Y_t \mid \mathbf{I}_{t-1}, \mathbf{Y}_{t-1}) = a_t(p) + b_t(p)$. As in Chen and Lai (2004), we can rewrite (4.3) as

$$E(\boldsymbol{\theta}_n^T, \sigma_n^2 \mid \mathbf{Y}_{1,n}) = E_Q\{w_n(\mathbf{z}_{J_n,n}^T,\ a_{J_n,n}/(2g+n-J_n-1))\}/E_Q(w_n), \tag{4.5}$$

where the importance weights can be generated recursively by

$$w_t = w_{t-1}\{a_t(p) + b_t(p)\}, \quad t \geq k+2; \quad w_{k+1} = 1. \tag{4.6}$$

When p is small, change-points occur very infrequently. We can use importance sampling to generate more change-points by sampling instead from Q' in which p in (4.4) is replaced by $p' > p$. With Q replaced by Q' in (4.5), the weights w_t in (4.6) are changed to

$$w_t = \begin{cases} w_{t-1}\{a_t(p') + b_t(p')\}a_t(p)/a_t(p') & \text{if} \quad I_t = 1, \\ w_{t-1}\{a_t(p') + b_t(p')\}b_t(p)/b_t(p') & \text{if} \quad I_t = 0. \end{cases} \tag{4.7}$$

The weight w_n defined recursively by (4.6) (or (4.7) if Q' is used in lieu of Q) tends to have a large coefficient of variation for large n. To overcome this difficulty, the SISR filter also incorporates occasional resampling (hence the symbol R) to keep the coefficient of variation (*cv*) within certain bounds. Specifically it draws m samples $\mathbf{I}_n^{(i)}$ sequentially from the proposal distribution Q and updates the importance weights $w_t^{(i)}$, $i = 1, \ldots, n$, by the following procedure, starting with m samples $I_{t-1}^{(1)}, \ldots, I_{t-1}^{(m)}$ having weights $w_{t-1}^{(1)}, \ldots, w_{t-1}^{(m)}$ at time $t-1$.

(a) Draw $\widehat{I}_t^{(j)}$ from (4.4) and update the weight $w_t^{(j)}$ by (4.6), $j = 1, \ldots, m$. If Q' is used as the proposal distribution instead, replace p in (4.4) by p' and (4.6) by (4.7).

(b) If the *cv* of $\{w_t^{(1)}, \ldots, w_t^{(m)}\}$ exceeds or equals a certain bound, resample from $\{\widehat{I}_t^{(1)}, \ldots, \widehat{I}_t^{(m)}\}$ with probabilities proportional to $\{w_t^{(1)}, \ldots, w_t^{(m)}\}$ to produce a random sample $\{I_t^{(1)}, \ldots, I_t^{(m)}\}$ with equal weights. Otherwise let $\{I_t^{(1)}, \ldots, I_t^{(m)}\} = \{\widehat{I}_t^{(1)}, \ldots, \widehat{I}_t^{(m)}\}$ and return to step (a).

4.3. Bounded complexity smoothers

As pointed out in Section 3, if the prior normal distribution for $\boldsymbol{\theta}_t$ is truncated within a stability region C, then $(\tau_t, \boldsymbol{\theta}_t^T, \mathbf{Y}_{t-k+1,t})$ is reversible and we can form the Bayesian smoother by combining the forward and backward filters. We ignore the truncation in implementing these forward and backward filters. Specifically, let $\widetilde{I}_n = 1, \widetilde{I}_t = 1_{\{(\boldsymbol{\theta}_t^T, \sigma_t) \neq (\boldsymbol{\theta}_{t+1}^T, \sigma_{t+1})\}}$ and $\tilde{J}_t = \min\{j > t \mid \widetilde{I}_j = 1\}$. Let $p_{i,t} = P(J_t = i \mid \mathbf{Y}_{1,t})$, $\tilde{p}_{j,t} = P(\tilde{J}_t = j \mid \mathbf{Y}_{t+1,n})$, and note that $\sum_{i=k+1}^{t} p_{i,t} = 1 =$

$\sum_{j=t+1}^{n+1} \tilde{p}_{j,t}$. The backward weights $\tilde{p}_{j,t}$ can be determined by backward induction on t using an analog of (2.7).

Analogous to (2.4)-(2.5), it is shown in Appendix B that for $i \le t < j \le n$, the conditional distribution of $(\boldsymbol{\theta}_t, \tau_t)$ given $J_t = i, \widetilde{J}_t = j$ and $\mathbf{Y}_{i,j}$ can be described by

$$\tau_t \sim \text{Gamma}\Big(g + \frac{j-i+1}{2}, \frac{1}{a_{i,j,t}}\Big),\ \boldsymbol{\theta}_t \mid \tau_t \sim \text{Normal}\Big(\mathbf{z}_{i,j,t}, \frac{1}{2\tau_t}\mathbf{V}_{i,j,t}\Big) \tag{4.8}$$

if we ignore the truncation in the truncated normal distribution in (3.2), where

$$\begin{aligned}
\mathbf{V}_{i,j,t} &= (\mathbf{V}_{i,t}^{-1} + \tilde{\mathbf{V}}_{j,t}^{-1} - \mathbf{V}^{-1})^{-1},\\
\mathbf{z}_{i,j,t} &= \mathbf{V}_{i,j,t}(\mathbf{V}_{i,t}^{-1}\mathbf{z}_{i,t} + \tilde{\mathbf{V}}_{j,t}^{-1}\tilde{\mathbf{z}}_{j,t+1} - \mathbf{V}^{-1}\mathbf{z}),\\
a_{i,j,t} &= a_{i,t} + \tilde{a}_{j,t+1} - \lambda^{-1} + \mathbf{z}_{i,t}^T\mathbf{V}_{i,t}^{-1}\mathbf{z}_{i,t} + \tilde{\mathbf{z}}_{j,t+1}^T\tilde{\mathbf{V}}_{j,t}^{-1}\tilde{\mathbf{z}}_{j,t+1} - \mathbf{z}^T\mathbf{V}^{-1}\mathbf{z} - \mathbf{z}_{i,j,t}^T\mathbf{V}_{i,j,t}^{-1}\mathbf{z}_{i,j,t}\\
&= \lambda^{-1} + \mathbf{z}^T\mathbf{V}^{-1}\mathbf{z} + \sum_{l=i}^{j} Y_l^2 - \mathbf{z}_{i,j,t}^T\mathbf{V}_{i,j,t}^{-1}\mathbf{z}_{i,j,t},
\end{aligned}$$

in which $\mathbf{V}_{i,t}$, $\mathbf{z}_{i,t}$ and $a_{i,t}$ are defined in (2.5) and $\widetilde{\mathbf{V}}_{j,t}$, $\tilde{\mathbf{z}}_{j,t}$ and $\tilde{a}_{j,t}$ are defined similarly by reversing time. Let $|\cdot|$ denote the determinant of a matrix,

$$\begin{aligned}
b_{i,j,t} &= \Big(\frac{|\mathbf{V}_{i,t}||\tilde{\mathbf{V}}_{j,t}|}{|\mathbf{V}||\mathbf{V}_{i,j,t}|}\Big)^{-\frac{1}{2}}\Big\{\frac{\Gamma(g)\Gamma(g+\frac{1}{2}(j-i+1))}{\Gamma(g+\frac{1}{2}(t-i+1))\Gamma(g+\frac{1}{2}(j-t))}\Big\}\frac{a_{i,t}^{g+(t-i+1)/2}\tilde{a}_{j,t}^{g+(j-t)/2}}{a^g a_{i,j,t}^{g+(j-i+1)/2}},\\
B_t &= p + (1-p)\sum_{k+1\le i\le t<j\le n} p_{i,t}\tilde{p}_{j,t}b_{i,j,t}.
\end{aligned}$$

Using (3.3) and (4.8), it is shown in Appendix B that analogous to (2.6),

$$\begin{aligned}
E(\sigma_t^2 \mid \mathbf{Y}_{1,n}) &\doteq \frac{p}{B_t}\sum_{i=k+1}^{t}\frac{p_{i,t}a_{i,t}}{2g+t-i-1} + \frac{1-p}{B_t}\sum_{k+1\le i\le t<j\le n} p_{i,t}\tilde{p}_{j,t}b_{i,j,t}\frac{a_{i,j,t}}{2g+j-i+1},\\
E(\boldsymbol{\theta}_t \mid \mathbf{Y}_{1,n}) &\doteq \frac{p}{B_t}\sum_{i=k+1}^{t} p_{i,t}\mathbf{z}_{i,t} + \frac{1-p}{B_t}\sum_{k+1\le i\le t<j\le n} p_{i,t}\tilde{p}_{j,t}b_{i,j,t}\mathbf{z}_{i,j,t},
\end{aligned} \tag{4.9}$$

in which the approximation ignores truncation within C. The BCMIX smoother further approximates (4.7) by allowing at most n_p weights $p_{i,t}$ and n_p weights $\tilde{p}_{j,t}$ to be nonzero.

The SISR smoother can be formed in a similar way. Define $\mathbf{I}_t$ and Q as in Section 4.2. Let $\widetilde{\mathbf{I}}_t = (\widetilde{I}_t, \ldots, \widetilde{I}_n)$ and define $\widetilde{Q}$ similarly so that $I_t \mid \widetilde{\mathbf{I}}_{t+1}$ has the Bernoulli distribution assuming the value 1 and 0 with probabilities in the proportion

$$pf(Y_t \mid \widetilde{I}_t = 1) : (1-p)f(Y_t \mid \mathbf{Y}_{t+1,\tilde{J}_t}, \widetilde{I}_t = 0). \tag{4.10}$$

The SISR forward filter is sampled from Q and the backward filter from $\widetilde{Q}$, yielding importance weights w_t defined recursively by (4.6) and $\widetilde{w}_t$ defined similarly by backward induction, with occasional resampling to keep the coefficients of variation of w_t and $\widetilde{w}_t$ within certain bounds. With m forward and m backward trajectories simulated in this way, the SISR smoother for $(\boldsymbol{\theta}_t^T, \sigma_t^2)$ can be expressed as

$$\sum_{i=1}^{m} \zeta_{i,t}\Big(\mathbf{z}_{J_t^{(i)},t}, \ \frac{a_{J_t^{(i)},t}}{2g+t-J_t^{(i)}-1}\Big) + \sum_{i,j=1}^{m} \beta_{i,j,t}\Big(\mathbf{z}_{J_t^{(i)},\tilde{J}_t^{(j)},t}, \frac{a_{J_t^{(i)},\tilde{J}_t^{(j)},t}}{2g+\widetilde{J}_t^{(j)}-J_t^{(i)}+1}\Big), \tag{4.11}$$

where $\zeta_{i,t} = pw_t^{(i)}/A_t$, $\beta_{i,j,t} = (1-p)w_t^{(i)}\widetilde{w}_t^{(j)} b_{J_t^{(i)},\tilde{J}_t^{(j)},t}/A_t$ and

$$A_t = p\sum_{i=1}^{m} w_t^{(i)} + (1-p)\sum_{i,j=1}^{m} w_t^{(i)} w_t^{(j)} b_{J_t^{(i)},\tilde{J}_t^{(j)},t};$$

see Appendix B for the derivation. Note the analogy of (4.11) to (4.9).

4.4. Simulation studies

The top panel of Figure 1 plots a time series of $n = 3,000$ observations generated from the change-point AR(2) model with

$$p = 0.001, \ \gamma_t \sim \text{Gamma } (3,4), \ \mathbf{Z}_t \mid \gamma_t \sim \mathbf{T}_C \text{ Normal } (\mathbf{0}, \mathbf{I}), \tag{4.12}$$

where $C = \{(\mu, \alpha_1, \alpha_2)^T : |\alpha_1| + |\alpha_2| < 1\}$. There are two change-times in the dataset and the piecewise constant parameter values are

$$(\mu_t, \sigma_t, \alpha_{1,t}, \alpha_{2,t}) = \begin{cases} (0.5019, -0.2171, -0.8360, 0.0629) & \text{if} \quad 1 \le t < 943, \\ (0.8723, \quad 1.0373, -0.0328, 0.2855) & \text{if} \quad 943 \le t < 1,623, \\ (0.5970, \quad 0.1043, -0.1115, 0.4333) & \text{if} \quad 1,623 \le t \le 3,000. \end{cases} \tag{4.13}$$

The Bayes estimates, $E(\sigma_t^2 \mid \mathbf{Y}_{1,t})$ and $E(\mu_t + \alpha_{1,t}Y_{t-1} + \alpha_{2,t}Y_{t-2} \mid \mathbf{Y}_{1,t})$, of the variance and regression function are also plotted in the middle and bottom panels, together with the corresponding BCMIX estimates (with $n_p = 25, \ m_p = 10$) and SISR estimates (based on $m = 100$ SISR trajectories).

Table 1(a) reports simulation results on the sum of squared errors

$$\text{SSE} := \sum_{t=3}^{n} \{(1, Y_{t-1}, Y_{t-2})(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)\}^2 \tag{4.14}$$

and the Kullback-Leibler divergence between the true and estimated parameter values, defined by

$$\text{KL} = \sum_{t=3}^{n} \Big\{ \frac{[(1, Y_{t-1}, Y_{t-2})(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2}{\widehat{\sigma}_t^2} + \Big(\frac{\sigma_t^2}{\widehat{\sigma}_t^2} - 1 - \log \frac{\sigma_t^2}{\widehat{\sigma}_t^2}\Big) \Big\}, \tag{4.15}$$

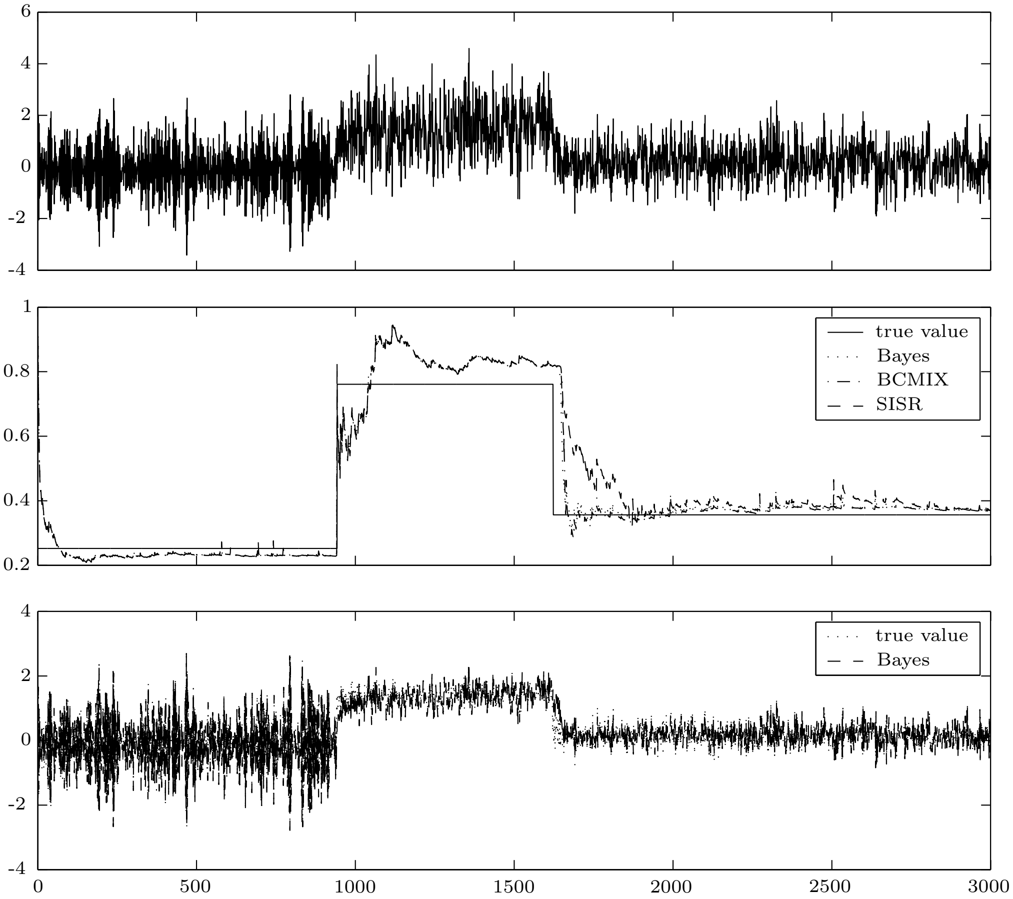

Figure 1. Top panel: time series of 3,000 observations generated from a change-point AR(2) model. Middle panel: true and estimated vlaues of σ_t^2. Bottom panel: true value of $\mu_t + \alpha_{1,t} Y_{t-1} + \alpha_{2,t} Y_{t-2}$ and its Bayes estimate, which are visually indistinguishable at most places where only the Bayes estimates is shown. Not shown in the bottom panel are the BCMIX and SISR estimates which are visually indistingushable from the Bayes estimate.

with $(\widehat{\boldsymbol{\theta}}_t, \widehat{\sigma}_t)$ being the Bayes, BCMIX and SISR filters in the respective columns. Note that the quantity inside the curly brackets of (4.15) is $E[\log\{f_{\boldsymbol{\theta}_t,\sigma_t}(Y_t^*)/ f_{\widehat{\boldsymbol{\theta}}_t,\widehat{\sigma}_t}(Y_t^*)\}]$, where the expectation is conditional on $(\widehat{\boldsymbol{\theta}}_t, \widehat{\sigma}_t)$ and taken under the true parameter $(\boldsymbol{\theta}_t, \sigma_t)$, and Y_t^* has density function $f_{\boldsymbol{\theta}_t,\sigma_t}$ and is independent of $(\widehat{\boldsymbol{\theta}}_t, \widehat{\sigma}_t)$. Each result in Table 1(a) is based on 100 simulations generated from the change-point AR(2) model whose parameters are given in (4.12), and also for several other values of p listed in the table. For $p = 0.0005$ we choose $n = 10,000$, whereas for the larger values of p we take $n = 5,000$, noting that the expected number of change-points in each simulated sequence is np. Note that although SSE for BCMIX can be more then twice that for the Bayes filter, the KL for BCMIX remains within 1.2 of that for the Bayes filter. Moreover, the SISR filter has a smaller SSE but larger KL than BCMIX. Whereas the Bayes filter at time t consists of a mixture of t components and t can increase up to 5,000 (or 10,000), the BCMIX filter involves a mixture of only $n_p = 25$ components. It is also worth noting that when the SSE of BCMIX exceeds twice that of the Bayes filter in Table 1, the SSE values for both the Bayes and BCMIX filters are small (< 0.15) when divided by $n - 2$ (yielding SSE per observation).

Table 1. Sum of squared errors (SSE) and Kullback-Leibler divergence (KL) for Bayes, BCMIX and SISR filters. Standard errors are given in parentheses.

		SSE			KL		
n	p	Bayes	BCMIX	SISR	Bayes	BCMIX	SISR
(a) Bayesian setting							
10,000	.0005	137.6	233.4	397.9	94.4	108.9	262.6
		(7.72)	(43.83)	(24.90)	(3.67)	(4.75)	(10.60)
5,000	0.001	122.4	305.8	213.0	84.6	97.0	146.6
		(7.10)	(104.00)	(13.59)	(3.29)	(4.23)	(7.30)
5,000	0.003	290.3	746.8	432.3	190.6	216.8	283.7
		(11.29)	(260.47)	(16.71)	(4.25)	(5.00)	(6.38)
5,000	0.01	659.3	1,082.0	812.0	437.1	493.3	524.7
		(12.22)	(98.91)	(15.26)	(5.46)	(6.37)	(6.28)
5,000	0.02	1,044.7	1614.1	1319.8	693.2	761.9	842.5
		(15.64)	(75.04)	(19.88)	(6.57)	(6.41)	(8.40)
(b) Frequentist setting ($n = 3,000$)							
Case 1		32.6	34.2	54.8	41.2	42.7	70.4
		(0.72)	(0.77)	(1.62)	(0.77)	(0.83)	(2.02)
Case 2		27.5	82.6	39.05	36.5	47.4	48.7
		(0.89)	(6.95)	(3.61)	(0.94)	(0.98)	(2.53)
Case 3		32.2	34.2	77.02	41.4	43.6	99.43
		(0.70)	(0.73)	(4.86)	(0.77)	(0.85)	(6.15)

While Table 1(a) considers the Bayes risks of various estimators of $\boldsymbol{\theta}_t$ and σ_t^2 and generates each simulation from the Bayesian model, Table 1(b) considers the frequentist risks for three fixed piecewise constant specifications of $(\boldsymbol{\theta}_t, \sigma_t)$. The first specification, called Case 1, is the same as that in Figure 1; see (4.13). The other two specifications allow unit-root nonstationarity in the AR(2) model.

Specifically, Case 2 (or 3) has the same $(\boldsymbol{\theta}_t, \sigma_t)$ values as in Case 1 except that $(\mu_t, \alpha_{1,t}, \alpha_{2,t})$ takes the value $(0, 1, 0)$ for $943 \leq t < 1,623$ in Case 2 (or for $t \geq 1,623$ in Case 3). The KL of BCMIX ranges between 1.05 to 1.30 times that of the Bayes filter, although BCMIX has a markedly larger SSE. For Case 3, both the KL and the SSE of SISR are over two times those of the Bayes filter. To handle the unit-root nonstationarity for $t \geq 1,623$, it seems necessary to increase the number of simulated SISR trajectories by ten or more times the number 100 used in Table 1.

We next consider the performance of the bounded complexity smoothers. For n of the size considered in Table 1, it is difficult to compute the Bayesian smoother $E(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,n})$ of Section 3. Instead of this computationally prohibitive benchmark for comparison with BCMIX smoothers, we consider a much simpler benchmark in which the change-points are known so that the Bayes estimates of $(\boldsymbol{\theta}_t, \sigma_t^2)$ between two change-points are given by the standard Bayesian formulas for normal populations (cf. Section 2.7 of Box and Tiao (1973)). Table 2 compares this "fictitious Bayes" smoother with the BCMIX smoother in terms of the SSE and KL (for which the sums in (4.14) and (4.15) are now replaced by $\Sigma_{t=3}^{n-2}$ to allow for backward filtering). Comparison of Table 2 with Table 1 shows the substantially smaller SSE and KL for BCMIX smoothers than the BCMIX filters.

Table 2. Sum of squared errors (SSE) and Kullback-Leibler divergence (KL) for "fictitious Bayes" (fBayes) and BCMIX smoothers. Standard errors are given in parentheses.

		SSE		KL	
n	p	fBayes	BCMIX	fBayes	BCMIX
10,000	0.0005	77.7 (30.55)	164.7 (41.96)	44.4 (7.96)	55.8 (4.34)
5,000	0.001	80.8 (31.96)	158.6 (32.67)	65.7 (16.76)	47.7 (3.61)
5,000	0.003	229.7 (49.12)	661.6 (194.39)	142.1 (24.26)	130.3 (4.70)
5,000	0.01	526.2 (36.05)	1390.7 (126.4)	304.0 (12.89)	372.4 (5.97)
5,000	0.02	1347.5 (528.36)	2382.0 (419.14)	467.0 (9.58)	641.0 (8.10)

We have not included SISR smoothers in the simulation study of Table 2 because of the computational cost of many simulated replicates of such smoothers, which involve another layer of SISR simulations. Simulating SISR smoothers for some particular cases yields results similar to those for BCMIX smoothers. For the simulated data in Figure 1, we evaluated the BCMIX smoother and the results for the estimates of σ_t^2 are shown in Figure 2.

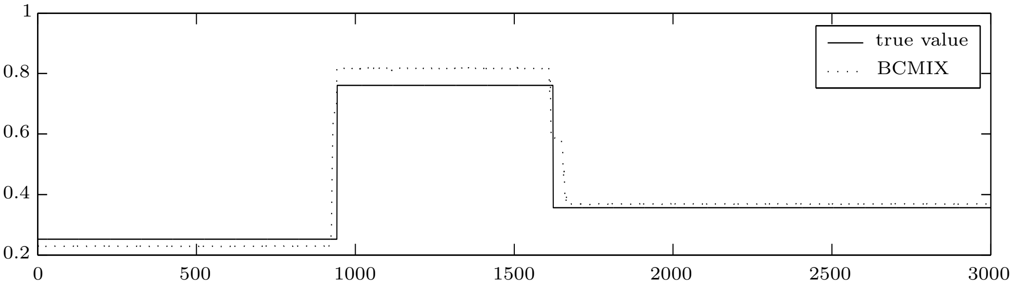

Figure 2. True value of σ_t^2 and its BCMIX estimate based on 3,000 observations.

5. Choice of Hyperparameters

In practice the frequency of change-points in an observed time series is unknown and one has to estimate the hyperparameter p, and possibly also other hyperparameters of the Bayesian change-point autoregressive model, from the data. For the mean shift model in Section 1, Yao (1984) considered the maximum likelihood approach. He developed an algorithm to compute the likelihood function with $O(n^2)$ operations but found it impractical to use in conjunction with an iterative search for the maximum likelihood estimate. Accordingly he developed an approximation to the likelihood function by collapsing the mixture distribution for $\theta_t \mid \mathbf{Y}_{1,t-1}$ to a single normal and maximized this pseudo-likelihood instead. In this section we describe two methods to estimate the hyperparameters, one based on the accumulated prediction error (APE) criterion to be used in conjunction with the BCMIX algorithm, and the other assuming a prior distribution on the hyperparameters to be used in conjunction with the SISR estimates.

5.1. APE criterion and BCMIX estimates

The accumulated prediction error (APE) criterion was introduced by Rissanen (1986) and applied to order determination in classical time series models by Hemerly and Davis (1989), Wei (1992) and Lai and Lee (1997). Because APE(ν) can be computed recursively, it is well suited to recursive estimators in time series models that involve an unspecified order ν to be determined from the data, and to BCMIX filters in change-point autoregressive models involving an unspecified hyperparameter ν that consists of p, g, λ, $\mathbf{z}$ and $\mathbf{V}$.

For the change-point AR(k) model, the accumulated prediction error at time n is defined by

$$\mathrm{APE}_n(\nu) = \sum_{t=k+1}^{n} \{Y_t - \widehat{Y}_{t|t-1}(\nu)\}^2, \tag{5.1}$$

where $\widehat{Y}_{t|t-1}(\nu)$ is the one-step-ahead predictor of Y_t given by

$$\widehat{Y}_{t|t-1}(\nu) = \{(1-p)\widehat{\boldsymbol{\theta}}_{t-1}(\nu) + p\mathbf{z}\}^T \mathbf{Y}_{t-k,t-1}, \tag{5.2}$$

in which $\widehat{\boldsymbol{\theta}}_i(\nu)$ is the BCMIX estimate (assuming the hyperparameter ν) of $\boldsymbol{\theta}$ based on $Y_1, \ldots, Y_i$. Following Lai, Liu and Xing (2004), we use the APE criterion to choose the hyperparameter at every stage n from a given set $\mathcal{H}$ of possible values. Specifically, let $\widehat{\nu}_n$ be the minimizer of $\mathrm{APE}_n(\nu)$ over $\nu \in \mathcal{H}$, and estimate $\boldsymbol{\theta}_t$ and σ_t^2 by $\widehat{\boldsymbol{\theta}}_t(\widehat{\nu}_{t-1})$ and $\widehat{\sigma}_t^2(\widehat{\nu}_{t-1})$. These estimates are called BCMIX-APE and Table 3 compares them with BCMIX(ν) that assumes the hyperparameter ν to be known.

When $\mathcal{H}$ is a finite set with h elements, we can compute BCMIX-APE by h parallel recursions. Specifically, for each $\nu \in \mathcal{H}$, $\widehat{\theta}_{t-1}(\nu)$ and $\widehat{\sigma}_{t-1}^2(\nu)$ can be updated recursively in view of (2.7) and (4.2), and therefore we can easily update $\mathrm{APE}_t(\nu)$ by $\mathrm{APE}_t(\nu) = \mathrm{APE}_{t-1}(\nu) + \{Y_t - \widehat{Y}_{t|t-1}(\nu)\}^2$. To choose a finite $\mathcal{H}$, one can start by using prior information to come up with a range $(0 <) p' \le p \le p''(< 1)$. Typically one has an upper bound (horizon) N for the length of past and future observations to be considered and can take $p' \ge 1/(2N)$, noting that the expected number of change-points for the series is Np. As will be explained in Appendix C, we propose to replace the interval $[p', p'']$ by a finite set of points

$$2^\ell p' \ (0 \le \ell \le L), \text{ where } L = \max\{\ell : 2^\ell p' \le p''\}. \tag{5.3}$$

Whereas the relative frequency p of change-points is an essential feature of our Bayesian model, the other hyperparameters $g, \lambda, \mathbf{z}$ and $\mathbf{V}$ become significant only around change-points, where one does not have enough post-change observations and relies on prior information to estimate $\boldsymbol{\theta}_t$ and σ_t. Accordingly, including a few plausible values of $(g, \lambda, \mathbf{z}, \mathbf{V})$ should suffice. In particular, estimates of baseline values of $\boldsymbol{\theta}$ and σ^2 (assumed not to be time-varying) from some historical data prior to $t \ge 0$ can provide estimated values of $(g, \lambda, \mathbf{z}, \mathbf{V})$ via the method of moments. Thus the set $\mathcal{H}$ chosen in this way typically has manageable cardinality h for the implementation of parallel recursions.

As an illustration, we consider the change-point AR(2) model whose hyperparameters are given in (4.12) and also for several other values of p, as in Table 1(a). For each value of p, we let $p' = p/10$ and $p'' = 10p$ and define $\mathcal{H}$ with cardinality $2L$, where L is defined above and each $(p, g, \lambda, \mathbf{z}, \mathbf{V}) \in \mathcal{H}$ has the following form: $p = 2^\ell p'$ and either (i) $g = 4, \lambda = 1/10, \mathbf{z} = (1/2)\mathbf{1}, \mathbf{V} = (3/2)\mathbf{I}$, or (ii) $g = 5/2, \lambda = 1/3$, $\mathbf{z} = -(1/2)\mathbf{1}$, $\mathbf{V} = 2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a vector of 1's. Note that σ_t^2 has a prior mean of 5/3 under (i) and 1 under (ii). The BCMIX-APE estimate of $(\boldsymbol{\theta}_t^T, \sigma_t^2)$ constructed in this way is compared with BCMIX(ν) in terms of the Kullback-Leibler divergence (4.15) in Table 3, each result of which is based on 100 simulations.

Table 3. Kullback-Leibler divergence of BCMIX(ν) and BCMIX-APE filters with $n_p = 35$. Standard errors are given in parentheses.

	$m_p = 5$		$m_p = 0$	
p	BCMIX(ν)	BCMIX-APE	BCMIX(ν)	BCMIX-APE
0.0005	96.0 (3.68)	105.8 (4.24)	96.1 (3.68)	106.8 (4.56)
0.001	86.1 (3.33)	94.6 (3.52)	86.2 (3.34)	97.8 (4.03)
0.003	192.0 (4.30)	209.8 (4.51)	192.2 (4.31)	210.4 (4.51)
0.01	439.2 (5.51)	478.1 (6.09)	439.3 (5.50)	478.3 (6.12)
0.02	696.1 (6.57)	770.5 (7.25)	696.2 (6.57)	770.9 (7.25)

Extending the APE criterion to the smoothing problem yields the usual cross validation criterion

$$\mathrm{CV}(\nu) = \sum_{t=3}^{n-2} \{Y_t - \widehat{Y}_{t,n}(\nu)\}^2, \tag{5.4}$$

where $\widehat{Y}_{t,n}(\nu)$ is a predictor of Y_t based on $(\mathbf{Y}_{1,t-1}, \mathbf{Y}_{t+1,n})$ obtained by combining the forward and backward BCMIX estimates in a way similar to that in Section 4.3; see Appendix D for details. The hyperparameter ν can be estimated for BCMIX smoothers by $\widetilde{\nu}_n$ that minimizes $\mathrm{CV}(\nu)$ over $\nu \in \mathcal{H}$.

5.2. Conjugate hyperprior distributions and SISR estimates

Suppose that in the change-point AR(k) model the probability p of change is unknown and is specified by a prior Beta(a, b) distribution with mean $a/(a+b)$, where a and b are positive integers. Using the proposal distribution Q for which $I_t \mid \mathbf{I}_{t-1}$ has the same distribution as $P(I_t = \cdot \mid \mathbf{I}_{t-1}, \mathbf{Y}_{1,t})$, it can be shown that $I_t \mid \mathbf{I}_{t-1}$ is Bernoulli assuming the values 1 and 0 with probabilities in the proportion

$$(n_{t-1,1} + a) f(Y_t \mid I_t = 1) : (n_{t-1,0} + b) f(Y_t \mid \mathbf{Y}_{J_{t-1},t-1}, I_t = 0), \tag{5.5}$$

where the conditional densities are Student-t densities given by Lemma 1, and $n_{t-1,1}$ and $n_{t-1,0}$ are the number of 1's and 0's in $\{I_1, \ldots, I_{t-1}\}$. Letting a_t and b_t denote the two terms in (5.5), note that $f(Y_t \mid \mathbf{I}_{t-1}, \mathbf{Y}_{t-1}) = a_t + b_t$ and that the importance weights can be generated recursively by $w_t = w_{t-1}(a_t + b_t)$, as in (4.6). By combining these adaptive forward filters with the corresponding backward filters, we obtain adaptive smoothers.

6. Conclusion

The change-point AR(k) model introduced herein is a simple Bayesian model that captures structural changes in both the volatility and regression parameters. It is a hidden Markov model (HMM), with the unknown regression and volatility parameters $\boldsymbol{\theta}_t$ and σ_t undergoing Markovian jump dynamics, so that estimation of $\boldsymbol{\theta}_t$ and σ_t can be treated as filtering and smoothing problems in the HMM. Making use of the special structure of the HMM that involves gamma-normal conjugate priors, we have been able to develop two approximations, with relatively low computational complexity, to the Bayesian filters and smoothers. The first is the simulation-based SISR which uses recent advances in sequential Monte Carlo methods. The second is BCMIX, which is developed from explicit formulas for the Bayesian filters.

The Bayesian model has certain hyperparameters among which is the relative frequency p of change-points in the time series. We have described two general approaches to determining hyperparameters from the data. Omitted from the discussion in the preceding sections is that one often has for a particular application some external information, and the strength of Bayesian modeling is that one can conveniently incorporate it into the Bayesian model. For example, since p is a hyperparameter that is sequentially determined from the data in Section 5, we can incorporate external information besides using a criterion like APE (which depends solely on the observed time series) to determine p. Such external information plays a fundamental role in the *intervention analysis* of Box and Tiao (1975). They noted that certain external events (e.g., the diversion of traffic by the opening of a new freeway, or a new law on the allowable proportion of reactive hydrocarbons in gasoline) might produce structural changes in an observed time series (such as hourly readings of oxidant pollution level in a town). Their approach is to model a known intervention at a certain time point t_0 by an input of the form $1_{\{t \geq t_0\}}$ in an ARMAX model. The ARMAX model, however, only involves the dynamics of the level of Y_t but not its volatility. If we use the change-point AR(k) model instead to model both the level and volatility, we can adjust the hyperparameter p at different time points to incorporate knowledge of external events such as interventions at these times.

Acknowledgement

This research was supported by the National Science Foundation.

Appendix A. Proof of Theorem 1

Lemma 3. (i) *If $\{Y_t, 1 \le t \le n\}$ is a stationary Gaussian sequence, then it is reversible, i.e., it has the same distribution as the time-reversed sequence $\tilde{Y}_t := Y_{n+1-t}$.*

(ii) *Let $Y_t = \mu + \alpha_1 Y_{t-1} + \cdots + \alpha_k Y_{t-k} + \sigma\epsilon_t$, $t > k$, in which the ϵ_t are i.i.d. standard normal and $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ satisfies the stability assumption* (3.1)*. Then $\{\mathbf{Y}_{t-k,t-1}, t > k\}$ is a geometrically ergodic Markov chain having a normal stationary distribution. In particular, Y_n has a limiting normal distribution with mean $\mu/(1 - \alpha_1 - \cdots - \alpha_k)$ and variance $v = \sigma^2 \sum_{j=0}^{\infty} \beta_j^2$, where β_j are the coefficients in the power series representation of $1/(1 - \alpha_1 z - \cdots - \alpha_k z^k) = \sum_{j=0}^{\infty} \beta_j z^j$ for $|z| \le 1$.*

(iii) *For the* AR(k) *model in* (ii)*, if $\mathbf{Y}_{1,k}$ is initialized at the stationary distribution, then $\{Y_t, k < t < n - k\}$ is reversible.*

Proof. (i) follows easily from the covariance function of the stationary Gaussian sequence. To prove (ii), note that $\mathbf{Y}_{t-k+1,t} = A\mathbf{Y}_{t-k,t-1} + (0, \mu + \sigma\epsilon_t, 0, \ldots, 0)^T$, where the first row of A is $(1, 0, \ldots, 0)$, the second row is $(0, \alpha_1, \ldots, \alpha_k)$, the third row is $(0, 1, 0, \ldots, 0)$, etc. It is a geometrically ergodic Markov chain (cf. Meyn and Tweedie (1993)). If the chain is initialized at its stationary distribution, then Y_n can be written as an infinite moving average $\sum_{j=0}^{\infty} \beta_j(\mu + \sigma\epsilon_{n-j})$, so (iii) follows from (i).

Proof of Theorem 1. First note that the probability measure Q, under which $(\tau_{k+2}, \boldsymbol{\theta}_{k+2}^T)$ has the same distribution as (3.2) and is independent of the Bernoulli random variable I_{k+2}, is an invariant measure (stationary distribution) for the Markov chain $\{(\tau_t, \boldsymbol{\theta}_t^T, I_t), t \ge k+2\}$. Moreover, the chain clearly satisfies the geometric drift condition (V4) of Meyn and Tweedie (1993, p.367) and is reversible (since the I_t are i.i.d.). Combining this with parts (ii) and (iii) of the preceding lemma gives the desired conclusion.

Appendix B. Proof of (4.8), (4.9) and (4.11)

From (3.3) it follows that

$$\begin{aligned} & f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,n}) \\ \propto\ & \frac{f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,t}) f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{t+1,n})}{f(\boldsymbol{\theta}_t^T, \sigma_t^2)} \\ \propto\ & \frac{f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,t})}{f(\boldsymbol{\theta}_t^T, \sigma_t^2)} [f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \tilde{I}_t{=}1) P(\tilde{I}_t{=}1) + f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{t+1,n}, \tilde{I}_t{=}0) P(\tilde{I}_t{=}0)] \end{aligned}$$

$$\propto p f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,t}) + (1-p)\frac{f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,t}) f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{t+1,n}, \tilde{I}_t = 0)}{f(\boldsymbol{\theta}_t^T, \sigma_t^2)}. \tag{B.1}$$

Note that given $J_t = i$ and $\mathbf{Y}_{i,t}$, the conditional distribution of $(\boldsymbol{\theta}_t^T, \sigma_t^2)$ can be described by

$$\tau_t \sim \text{Gamma}\Big(g + \frac{t-i+1}{2}, \frac{1}{a_{i,t}}\Big), \quad \boldsymbol{\theta}_t \mid \tau_t \sim \text{Normal}\Big(\mathbf{z}_{i,t}, \frac{1}{2\tau_t}\mathbf{V}_{i,t}\Big).$$

Similarly, given $\tilde{J}_t = j \geq t+1$ and $\mathbf{Y}_{t+1,j}$, the conditonal distribution of $(\boldsymbol{\theta}_t^T, \sigma_t^2)$ can be described by

$$\tau_t \sim \text{Gamma}\Big(g + \frac{j-t}{2}, \frac{1}{\tilde{a}_{j,t}}\Big), \quad \boldsymbol{\theta}_t \mid \tau_t \sim \text{Normal}\Big(\tilde{\mathbf{z}}_{j,t}, \frac{1}{2\tau_t}\tilde{\mathbf{V}}_{j,t}\Big).$$

Hence we can write the last term in (B.1) as $(1-p)\sum_{i=k+1}^{t}\sum_{j=t+1}^{n} p_{i,t}\tilde{p}_{j,t}e_{i,j,t}$, where for $i \leq t < j \leq n$,

$$\begin{aligned} e_{i,j,t} &= \frac{f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{1,t}, J_t = i) f(\boldsymbol{\theta}_t^T, \sigma_t^2 \mid \mathbf{Y}_{t+1,n}, \tilde{J}_t = j)}{f(\boldsymbol{\theta}_t^T, \sigma_t^2)} \\ &= \frac{\text{Normal}(\mathbf{z}_{i,t}, \mathbf{V}_{i,t}/(2\tau_t))\text{Normal}(\tilde{\mathbf{z}}_{j,t}, \tilde{\mathbf{V}}_{j,t}/(2\tau_t))}{\text{Normal}(\mathbf{z}, \mathbf{V}/(2\tau_t))} \\ &\quad \times \frac{\text{Gamma}(g + (t-i+1)/2, 1/a_{i,t})\text{Gamma}(g + (j-t)/2, 1/\tilde{a}_{j,t})}{\text{Gamma}(g, \lambda)} \\ &= b_{i,j,t}\text{Normal}(\mathbf{z}_{i,j,t}, \mathbf{V}_{i,j,t}/(2\tau_t)) \times \text{Gamma}(g + \frac{j-i+1}{2}, \frac{1}{a_{i,j,t}}). \end{aligned} \tag{B.2}$$

In (B.2), we have used Gamma$(\cdot,\cdot)$ to denote the gamma density function of $\tau_t := (2\sigma_t^2)^{-1}$, with the indicated shape and scale parameters, and Normal$(\cdot,\cdot)$ to denote the normal density function of $\boldsymbol{\theta}_t$ given τ_t, with the indicated mean and convariance matrices. Note that by Bayes' theorem, (B.1) also gives the conditional distribution of $(\boldsymbol{\theta}_t, \tau_t)$ given $J_t = i, \tilde{J}_t = j$ and $\mathbf{Y}_{i,j}$, thus proving (4.8). Combining (B.1) with (B.2) and (2.6) yields (4.9). For the SISR smoother, changing $p_{i,t}$ and $\tilde{p}_{j,t}$ to $w_t^{(i)}$ and $\tilde{w}_t^{(j)}$ in (4.9) gives (4.11).

Appendix C. Rationale behind the choice (5.3) for $\mathcal{H}$

For the case $\alpha_{1,t} = \cdots = \alpha_{k,t} = 0$, Lai, Liu and Xing (2004) proposed to use (5.3) as the candidate set for the hyperparameter p to be chosen sequentially via the APE criterion. They based their choice on the following asymptotic property of hierarchical Bayes estimators that put a prior distribution on the unknown hyperparameter p. Take $\beta > 1$ and let G be any probability distribution with support $[\beta^{-1}p_0, \beta p_0]$. This includes the prior disribution that is degenerate at p_0, corresponding to the case of a known hyperparameter. Let $(\hat{\mu}_t(G), \hat{\sigma}_t^2(G))$ be the Bayes estimate of (μ_t, σ_t^2) assuming the hyperprior G on the unknown p. Lai, Liu and Xing (2004) have shown that as $n \to \infty$ and $p_0 \to 0$ such that $np_0/|\log p_0| \to \infty$,

$$E_G\{\sum_{t=1}^{n}[(\hat{\mu}_t(G) - \mu_t)^2 + (\hat{\sigma}_t^2(G) - \sigma_t^2)^2]\} \sim Anp_0|\log p_0|$$

for some constant A that does not depend on G. Note that this result also covers the case of a known hyperparameter (corresponding to degenerate G at p_0). With $\beta = 2$, this suggests that specifying the hyperparameter p only within a range from 50% to 200% of its exact value yields Bayes estimates of (μ_t, σ_t^2) that are asymptotically as efficient, for small p, as those that assume exact knowledge of p. Simulation studies confirm this and show that misspecifying p within a factor of 2 does not lead to substantial proportional increase in the cumulative mean squared error even when p is not small, not only in the special case considered by Lai, Liu and Xing (2004) but also more generally in change-point autoregressive models (2.1) in which the autoregressive parameters $\alpha_{1,t}, \ldots, \alpha_{k,t}$ are restricted to some stability region.

Appendix D. Derivation of $\widehat{\mathbf{Y}}_{\mathbf{t,n}}(\boldsymbol{\nu})$ in (5.4)

Anaolgous to (3.3), Bayes' theorem yields

$$f(Y_t \mid \mathbf{Y}_{1,t-1}, \mathbf{Y}_{t+1,n}) \propto f(Y_t \mid \mathbf{Y}_{1,t-1}) f(Y_t \mid \mathbf{Y}_{t+1,n})/f(Y_t). \tag{D.1}$$

Let $\tilde{Y}_t = Y_{n+1-t}$,

$$\mu_{i,t} = \begin{cases} \mathbf{z}_{i,t-1}^T \mathbf{Y}_{t-k,t-1} & \text{if} \quad i \le t-1, \\ \mathbf{z}^T \mathbf{Y}_{t-k,t-1} & \text{if} \quad i = t, \end{cases}$$

$$\tilde{\mu}_{j,t} = \begin{cases} \tilde{\mathbf{z}}_{j,t}^T \tilde{\mathbf{Y}}_{n-t-k+1,n-t} & \text{if} \quad j \ge t+1, \\ \mathbf{z}^T \tilde{\mathbf{Y}}_{n-t-k+1,n-t} & \text{if} \quad j = t, \end{cases}$$

$$\sigma_{i,t}^2 = \begin{cases} a_{i,t-1}(1 + \mathbf{Y}_{t-k,t-1}^T \mathbf{V}_{i,t-1} \mathbf{Y}_{t-k,t-1})/(2g + t - i) & \text{if} \quad i \le t-1, \\ (1+\mathbf{Y}_{t-k,t-1}^T \mathbf{V} \mathbf{Y}_{t-k,t-1})/(2g\lambda) & \text{if} \quad i = t, \end{cases}$$

$$\tilde{\sigma}_{j,t}^2 = \begin{cases} \tilde{a}_{j,t+1}(1+\tilde{\mathbf{Y}}_{n-t-k+1,n-t}^T \tilde{\mathbf{V}}_{j,t+1} \tilde{\mathbf{Y}}_{n-t-k+1,n-t})/(2g+n-t-j) & \text{if} \quad j \ge t+1, \\ (1 + \tilde{\mathbf{Y}}_{n-t-k+1,n-t}^T \mathbf{V} \tilde{\mathbf{Y}}_{n-t-k+1,n-t})/(2g\lambda) & \text{if} \quad j = t. \end{cases}$$

Then by Lemma 1(i), conditional on $J_t = i$ and $\mathbf{Y}_{i,t-1}$.

$$(Y_t - \mu_{i,t})/\sigma_{i,t} \sim \text{Student-t } (2g + t - i), \tag{D.2}$$

in which the quantity $2g+t-i$ in parentheses denotes the degrees of freedom. Using time reversal, it follows similarly from Lemma 1(i) that, conditional on $\tilde{J}_t = j$ and $\mathbf{Y}_{t+1,j}$,

$$(Y_t - \tilde{\mu}_{j,t})/\tilde{\sigma}_{j,t} \sim \text{Student-t } (2g+n-t-j). \tag{D.3}$$

Although both factors in the numerator of (D.1) have the simple forms (D.2) and (D.3), the denominator is considerably more complicated. We ignore the denominator as our goal is to combine the forward and backward predictors of Y_t in (D.2) and (D.3) in a simple way. Using this approximation in (D.1), we can regard $\mathbf{Y}_{1,t-1}$ and $\mathbf{Y}_{t+1,n}$ as two independent scources of information on Y_t since the numerator in the right hand side of (D.1) factors into these two components. Accordingly we weight the forward and backward predictors given by (D.2) and (D.3) by their respective variances, leading to the estimate

$$\widehat{E}(Y_t \mid J_t = i, \tilde{J}_t = j, \mathbf{Y}_{i,t-1}, \mathbf{Y}_{t+1,j}) = \{\frac{\mu_{i,t}}{\sigma^2_{i,t}} + \frac{\tilde{\mu}_{j,t}}{\tilde{\sigma}^2_{j,t}}\}/\{\frac{1}{\sigma^2_{i,t}} + \frac{1}{\tilde{\sigma}^2_{j,t}}\}. \tag{D.4}$$

There is little loss of information in ignoring the denominator in (D.1) unless $\max(t-i, j-t)$ is small, but (i,j) pairs with both i and j near t do not have much predictive value for Y_t. We therefore have for any fixed value of the hyperparameter $\nu = (p, g, \lambda, \mathbf{z}, \mathbf{V})$ the estimate

$$\hat{Y}_{t,n}(\nu) = \sum_{i=1}^{t}\sum_{j=t}^{n} p_{i,t}\tilde{p}_{j,t}\{\frac{\mu_{i,t}}{\sigma^2_{i,t}} + \frac{\tilde{\mu}_{j,t}}{\tilde{\sigma}^2_{j,t}}\}/\{\frac{1}{\sigma^2_{i,t}} + \frac{1}{\tilde{\sigma}^2_{j,t}}\}.$$

References

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *J. Amer. Statist. Assoc.* **70**, 70-79.

Caines, P. E. (1988). *Linear Stochastic Systems.* Wiley, New York.

Chen, Y. and Lai, T. L. (2004). Identification and adaptive control of ARX models with occasional parameter jumps via fast particle filters. Technical Report, Department of Statistics, Stanford University.

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

Hemerly, E. M. and Davis, M. H. A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.* **17**, 941-946.

Lai, T. L. and Lee, C. P. (1997). Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statist. Sinica* **7**, 285-309.

Lai, T. L., Liu, T. and Xing, H. (2004). Efficient sequential change-point detection and estimation of parameters undergoing occasional changes in exponential families. Technical Report, Department of Statistics, Stanford University.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag, New York.

Rissanen, J. (1986). Order estimation by accumulated prediction errors. In *Essays in Time Series and Applied Processes*, *J. Appl. Probab.* **23A**, 55-61.

Tsay, R. S. (2002). *Analysis of Financial Time Series.* Wiley, New York.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

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

Department of Statistics, Sequoia Hall, 390 Serra Hall, Stanford University, Stanford, CA 94305-4065, U.S.A.

E-mail: lait@stat.stanford.edu

Department of Statistics, Sequoia Hall, 390 Serra Hall, Stanford University, Stanford, CA 94305-4065, U.S.A.

Department of Statistics, Sequoia Hall, 390 Serra Hall, Stanford University, Stanford, CA 94305-4065, U.S.A.

E-mail: xing@stanford.edu

(Received November 2003; accepted September 2004)