ELSEVIER Journal of Econometrics 86 (1998) 221–241

JOURNAL OF
Econometrics

Estimation and comparison of multiple change-point models

Siddhartha Chib*

John M. Olin School of Business, Washington University, 1 Brookings Drive, Campus Box 1133,
St. Louis, MO 63130, USA

Received 1 January 1997; received in revised form 1 August 1997; accepted 22 September 1997

Abstract

This paper provides a new Bayesian approach for models with multiple change points. The centerpiece of the approach is a formulation of the change-point model in terms of a latent discrete state variable that indicates the regime from which a particular observation has been drawn. This state variable is specified to evolve according to a discrete-time discrete-state Markov process with the transition probabilities constrained so that the state variable can either stay at the current value or jump to the next higher value. This parameterization exactly reproduces the change point model. The model is estimated by Markov chain Monte Carlo methods using an approach that is based on Chib (1996). This methodology is quite valuable since it allows for the fitting of more complex change point models than was possible before. Methods for the computation of Bayes factors are also developed. All the techniques are illustrated using simulated and real data sets. © 1998 Published by Elsevier Science S.A. All rights reserved.

JEL classification: C1; C4

Keywords: Bayes factors; Change points; Gibbs sampling; Hidden Markov model; Marginal likelihood; Markov mixture model; Markov chain Monte Carlo; Monte Carlo EM

1. Introduction

1.1. Change-point model

This paper is concerned with the problems of estimating and comparing time-series models that are subject to multiple change points. Suppose

* E-mail: chib@simon.wustl.edu

PII S0304-4076(97)00115-2

$Y_n = \{y_1, y_2, \dots, y_n\}$ is a time series such that the density of y_t given Y_{t-1} (with respect to some σ-finite measure) depends on a parameter ξ_t whose value changes at unknown time points $\Upsilon_m = \{\tau_1, \dots, \tau_m\}$ and remains constant otherwise, where $\tau_1 > 1$ and $\tau_m < n$. Upon setting

$$\xi_t = \begin{cases} \theta_1 & \text{if} \quad t \leqslant \tau_1, \\ \theta_2 & \text{if} \quad \tau_1 < t \leqslant \tau_2, \\ \vdots & \vdots \quad \vdots \\ \theta_m & \text{if} \quad \tau_{m-1} < t \leqslant \tau_m, \\ \theta_{m+1} & \text{if} \quad \tau_m < t \leqslant n, \end{cases}$$

where $\theta_k \in \Re^d$, the problems of inference include the estimation of the parameter vector $\Theta = (\theta_1, \dots, \theta_{m+1})$, the detection of the unknown change points $\Upsilon_m = (\tau_1, \dots, \tau_m)$, and the comparison of models with different numbers of change points.

This multiple change-point model has generated an enormous literature. One question concerns the specification of the jump process for the ξ_t. In the Bayesian context this is equivalent to the joint prior distribution of the τ_k and the θ_k, the parameters of the different 'regimes' induced by the change points. The model is typically specified through a hierarchical specification in which (for every time point t) one first models the probability distribution of a change, given the previous change points; then, the process of the parameters in the current regime, given the current change points and previous parameters; and finally, the generation of the data, given the parameters and the change points.

Chernoff and Zacks (1964) propose a special case of this general model in which there is a constant probability of change at each time point (not dependent on the history of change points). Then, given that the process has experienced changes at the points $\Upsilon_{k-1} = (\tau_1, \dots, \tau_{k-1})$, the parameter vector θ_k in the new regime, conditioned on the parameters Θ_{k-1} of the previous regimes, is assumed to be drawn from some distribution $\theta_k | \theta_{k-1}$ that depends only on θ_{k-1}. The parameters of this distribution, referred to as hyperparameters, are either specified or estimated from the data. For instance, one may let $\theta_k | \Theta_{k-1}$, $\Upsilon_{k-1} \sim \mathcal{N}(\theta_{k-1}, \Sigma_0)$, where $\mathcal{N}$ denotes the normal distribution and Σ_0 (the variance matrix) is the hyperparameter of this distribution.

Yao (1984) specified the same model for the change points but assumed that the joint distribution of the parameters $\{\theta_k\}$ is exchangeable and independent of the change points. Similar exchangeable models for the parameters have been studied by Carlin et al. (1992) in the context of a single change point and by Inclan (1993) and Stephens (1994) in the context of multiple change points. Barry and Hartigan (1993) discussed alternative formulations of the change points in terms of product partition distributions.

One purpose of this paper is to show that it is possible to extend the literature and fit models in which the change-point probability is not a constant but depends on the regime. In this case, the probability distribution of the change points is characterized not by just one parameter but by a set of parameters (say) P. The precise definition of these parameters is explained in Section 2 below.

Consider the distribution of the data given the parameters. Let $Y_i=(y_1,\dots,y_i)$ denote the history through time i, and $Y^{i,j}=(y_i,y_{i+1},\dots,y_j)$ the history from time i through j. Then the joint density of the data conditional on (Θ, Υ_m) is given by

$$f(Y_n|\Theta,\Upsilon_m)=\prod_{k=1}^{m+1} f(Y^{\tau_{k-1}+1,\tau_k}|Y_{\tau_{k-1}},\theta_k,\tau_k), \tag{1}$$

where $\tau_0=0$, $\tau_{m+1}=n$, and $f(\cdot)$ is a generic density or mass function. This is obtained by applying the law of total probability to the partition induced by Υ_m. A feature of this problem is that the density $f(Y_n|\Theta,P)$, obtained by marginalizing $f(Y_n|\Theta,\Upsilon_m)$ over all possible values of $\{\tau_j\}$ with respect to the prior mass function on Υ_m, is generally intractable.

1.2. An existing computational approach

The intractability of $f(Y_n|\Theta, P)$ has led to some interesting approaches based on Markov chain Monte Carlo simulation methods. One idea due to Stephens (1994) (see Barry and Hartigan, 1993 for an alternative approach) is to sample the unknown change points Υ_m and the parameters from the set of full conditional distributions

$$\Theta,\ P|Y_n,\ \Upsilon_m;\ \tau_k|Y_n,\ \Theta,\ P,\ \Upsilon_{m\backslash k},\quad k\leqslant m, \tag{2}$$

where P denotes (generically) the parameters of the change-point process and

$$\Upsilon_{m\backslash k}=(\tau_1,\dots,\tau_{k-1},\ \tau_{k+1},\dots,\tau_m)$$

is the set of change points excluding the kth. The essential point is that the conditional distribution Θ, $P|Y_n$, Υ_m is usually simple, whereas each of the conditional distributions $\tau_k|Y_n$, Θ, P, $\Upsilon_{m\backslash k}$ depends on the two neighboring change points (τ_{k-1},τ_{k+1}) and on the data $Y^{\tau_{k-1}+1,\tau_{k+1}}$. Specifically, $\tau_k|Y_n$, Θ, P, $\Upsilon_{m\backslash k}$ is given by the mass function

$$\begin{aligned}\Pr(\tau_k=j|Y_n,\ \Theta,\ P,\ \Upsilon_{m\backslash k}) &= \Pr(\tau_k=j|Y_{\tau_{k+1}},\ \theta_k,\ \theta_{k+1},\ P,\ \tau_{k-1},\ \tau_{k+1})\\ &\propto f(Y^{\tau_{k-1}+1,j}|Y_{\tau_k},\ \theta_k)\times f(Y^{j+1,\ \tau_{k+1}-1}|Y_j,\ \theta_{k+1})\times\Pr(\tau_k=j|\tau_{k-1}),\end{aligned}$$

for $\tau_{k-1}<j<\tau_{k+1}$. The normalizing constant of this mass function is the sum of the right-hand side over j.

This approach suffers from two weaknesses. Both are connected to the simulation of the change points. The first arises from the fact that the change points Υ_m are simulated one at a time from the m full conditional distributions $\tau_k|Y_n, \Theta, P, \Upsilon_{m\backslash k}$ rather than from the joint distribution

$$\Upsilon_m|Y_n, \Theta, P.$$

Sampling the latter distribution directly, say through a Metropolis–Hastings step, is not practical because of the difficulty of developing appropriate proposal generators for the change points. Liu et al. (1994) in theoretical work and Carter and Kohn (1994), Chib and Greenberg (1995) and Shephard (1994) in empirical work have shown that the mixing properties of the MCMC output is considerably improved when highly correlated components are grouped together and simulated as one block. Conversely, MCMC algorithms that do not exploit such groupings tend to be slow to converge. The second is the computational burden in evaluating the joint density functions $f(Y^{\tau_{k-1}+1,j}|Y_{\tau_k}, \theta_k) \times f(Y^{j+1,\tau_{k+1}-1}|Y_j, \theta_{k+1})$ for each value of j in the support $\tau_{k-1} < j < \tau_{k+1}$. This calculation must be repeated for each break point leading to individual density evaluations of the order n^m. With a long time series running into several hundreds of observations, these calculations are too burdensome to be practical for even relatively small values of m.

1.3. Outline of paper

The rest of the paper is organized as follows. In Section 2, a new parameterization of the change point model and an associated MCMC algorithm is supplied that eliminates the weaknesses of existing approaches. The MCMC implementation is shown to be straightforward and a simple consequence of the approach developed by Chib (1996) for hidden Markov models. In Section 3 the problem of model comparison is considered and approaches for estimating the likelihood function, the maximum-likelihood estimate, and the Bayes factors for comparing change-point models are provided. It is shown that the Bayes factors can be readily obtained from the method of Chib (1995) in conjunction with the proposed parameterization of the model. Section 4 contains examples of the ideas and Section 5 concludes.

2. A new parameterization

We begin this section by providing a new formulation of the change-point model that lends itself to straightforward calculations. This formulation is based on the introduction of a discrete random variable s_t in each time period, referred to as the state of the system at time t, that takes values on the integers $\{1, 2, \ldots, m+1\}$ and indicates the regime from which a particular observation y_t has been drawn. Specifically, $s_t = k$ indicates that the observation y_t is drawn from $f(y_t|Y_{t-1}, \theta_k)$.

The variable s_t is modeled as a discrete time, discrete-state Markov process with the transition probability matrix constrained so that the model is equivalent to the change-point model. To accomplish this, the transition probability matrix specifies that s_t can either stay at the current value or jump to the next higher value. This one-step ahead transition probability matrix is represented as

$$P = \begin{pmatrix} p_{11} & p_{12} & 0 & \cdots & 0 \\ 0 & p_{22} & p_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \vdots & 0 & p_{mm} & p_{m,m+1} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{3}$$

where $p_{ij} = \Pr(s_t = j|s_{t-1} = i)$ is the probability of moving to regime j at time t given that the regime at time $t-1$ is i. The probability of change thus depends on the current regime, a generalization of both Yao (1984) and Barry and Hartigan (1993) (it is easy to make the p_{ii} a function of covariates through a probit link function, if desired). It is specified that the chain begins in state 1 at $t = 1$, implying that the initial probability mass function on the states is $(1, 0, \ldots, 0)$, and the terminal state is $m + 1$. Note that there is only one unknown element in each row of P.

One way to view this parameterization is as a generalized change-point model in which the jump probabilities p_{ii} $(i \leqslant m)$ are dependent on the regime and the transitions of the state identify the change points $\Upsilon_m = (\tau_1, \ldots, \tau_m)$. The kth change occurs at τ_k if $s_{\tau_k} = k$ and $s_{\tau_k+1} = k + 1$. Note that this reparameterization automatically enforces the order constraints on the break points. Another way to view this parameterization is as a hidden Markov model (HMM) (Chib, 1996) in which the transition probabilities of the *hidden* state variable s_t are restricted in the manner described above. This view of the model forms the basis of our computational MCMC scheme.

2.1. Markov chain Monte Carlo scheme

Suppose that we have specified a prior density $\pi(\Theta, P)$ on the parameters and that data Y_n is available. In the Bayesian context, interest centers on the posterior density $\pi(\Theta, P|Y_n) \propto \pi(\Theta, P) f(Y_n|\Theta, P)$. Note that we use the notation π to denote prior and posterior density functions of (Θ, P). This posterior density is most fruitfully summarized by Markov chain Monte Carlo methods after the parameter space is augmented to include the unobserved states $S_n = (s_1, \ldots, s_n)$ (see Chib and Greenberg, 1996 for a convenient summary of these methods). In other words, we apply our Monte Carlo sampling scheme to the posterior density $\pi(S_n, \Theta, P|Y_n)$.

The general sampling method works recursively. First, the states S_n are simulated conditioned on the data Y_n and the other parameters (where s_n is set equal to $m+1$), and second, the parameters are simulated conditioned on the data and S_n. Specifically, the MCMC algorithm is implemented by simulating the following full conditional distributions:

1. Θ, $P|Y_n$, S_n, and
2. $S_n|Y_n$, Θ, P,

where the most recent values of the conditioning variables are used in all simulations. Note that the first distribution is the product of $P|S_n$ and $\Theta|Y_n$, S_n, P. The latter distribution is model specific and is therefore discussed only in the context of the examples below.

2.2. Simulation of $\{s_t\}$

Consider now the question of sampling S_n from the distribution $S_n|Y_n$, Θ, P, which amounts to the simulation of Υ_m from the joint distribution $\Upsilon_m|Y_n$, Θ, P. As mentioned earlier, sampling the change points from the latter distribution is difficult, if not impossible, however, the simulation of the states from the former distribution is relatively straightforward and requires just two passes through the data, regardless of the number of break points. Because the MCMC samples on the states are obtained by sampling the joint distribution, significant improvements in the overall mixing properties of the output (relative to the one-at-a time sampler) can be expected. In addition, this algorithm produces considerable time savings, requiring merely order n conditional density evaluations of y_t.

The algorithm for sampling the states follows from Chib (1996), where the unrestricted Markov mixture model is analyzed by MCMC methods. The objective is to draw a sequence of values of $s_t \in \{1, 2, \dots, m+1\}$ from the mass function $p(S_n|Y_n, \Theta, P)$. Henceforth, the notation $p(\cdot)$ is used whenever one is dealing with a discrete mass function. Let

$$S_t = (s_1, \dots, s_t); \qquad S^{t+1} = (s_{t+1}, \dots, s_n),$$

denote the state history up to time t and the future from $t+1$ to n, respectively, with a similar convention for Y_t and Y^{t+1}, and write the joint density in reverse time order as

$$p(s_{n-1}|Y_n, s_n, \Theta, P) \times \cdots \times p(s_t|Y_n, S^{t+1}, \Theta, P) \times \cdots \times p(s_1|Y_n, S^2, \Theta, P). \tag{4}$$

We write the joint density in this form because only then can each of the mass functions be derived and sampled. The process is completed by sampling, in turn,

- s_{n-1} from $p(s_{n-1}|Y_n, s_n = m+1, \Theta, P)$,
- s_{n-2} from $p(s_{n-2}|Y_n, S^{n-1}, \Theta, P)$,
- $\vdots$
- s_1 from $p(s_1|Y_n, S^2, \Theta, P)$.

The last of these distributions is degenerate at $s_1 = 1$. Thus, to implement this sampling it is sufficient to consider the sampling of s_t from $p(s_t|Y_n, S^{t+1}, \Theta, P)$.

Chib (1996) showed that

$$p(s_t|Y_n, S^{t+1}, \Theta, P) \propto p(s_t|Y_t, \Theta, P)p(s_{t+1}|s_t, P), \tag{5}$$

where the normalizing constant is easily obtained since s_t takes on only two values, conditioned on the value taken by s_{t+1}. There are two ingredients in this expression – the quantity $p(s_t|Y_t, \Theta, P)$ and $p(s_{t+1}|s_t, \Theta, P)$, which is just the transition probability from the Markov chain. To obtain the mass function $p(s_t|Y_t, \Theta, P)$, $t = 1,2,\dots,n$, a recursive calculation is required. Starting with $t = 1$, the mass function $p(s_{t-1}|Y_{t-1}, \Theta, P)$ is transformed into $p(s_t|Y_t, \Theta, P)$ which in turn is transformed into $p(s_{t+1}|Y_{t+1}, \Theta, P)$, and so on. The details are as follows. Suppose $p(s_{t-1} = l|Y_{t-1}, \Theta, P)$ is available. Then, the update to the required distribution is given by

$$p(s_t = k|Y_t, \Theta, P) = \frac{p(s_t = k|Y_{t-1},\Theta,\ P) \times f(y_t|Y_{t-1}, \theta_k)}{\sum_{l=k-1}^{k} p(s_t = l|Y_{t-1},\Theta,P) \times f(y_t|Y_{t-1},\theta_l)},$$

where

$$p(s_t = k|Y_{t-1}, \Theta, P) = \sum_{l=k-1}^{k} p_{lk} \times p(s_{t-1} = l|Y_{t-1}, \Theta, P), \tag{6}$$

for $k = 1, 2, \dots, m+1$, and p_{lk} is the Markov transition probability in Eq. (3). These calculations are initialized at $t = 1$ by setting $p(s_1|Y_0, \theta)$ to be the mass distribution that is concentrated at 1. With these mass functions at hand, the states are simulated from time n (setting s_n equal to $m+1$) and working backwards according to the scheme described in Eq. (5).

It can be seen that the sample output of the states can be used to determine the distribution of the change points. Alternatively, posterior information about change points can be based on the distribution

$$\Pr(s_t|Y_n) = \int p(s_t = k|Y_{t-1}, \Theta, P)\pi(\Theta, P|Y_n)d(\Theta, P).$$

This implies that a Monte Carlo estimate of $\Pr(s_t|Y_n)$ can be found by taking an average of $p(s_t = k|Y_{t-1}, \Theta, P)$ over the MCMC iterations. By the Rao–Blackwell theorem, this estimate of $\Pr(s_t|Y_n)$ is more efficient than one based on the empirical distribution of the simulated states.

2.3. Simulation of P

The revision of the non-zero elements p_{ii} in the matrix P given the data and the value of S_n is straightforward since the full conditional distribution $P|Y_n, S_n, \Theta$ is independent of (Y_n, Θ), given S_n. Thus, the elements p_{ii} of P may be simulated from $P|S_n$ without regard to the sampling model for the data y_t (see Albert and Chib, 1993).

Suppose that the prior distribution of p_{ii}, independently of p_{jj}, $j \neq i$, is *Beta*, i.e.,

$$p_{ii} \sim Beta(a,b)$$

with density

$$\pi(p_{ii}|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_{ii}^{a-1}(1-p_{ii})^{b-1},$$

where $a \gg b$. The joint density of P is, therefore, given by

$$\pi(P) = c\prod_{i=1}^{m} p_{ii}^{(a-1)}(1-p_{ii})^{(b-1)}$$

where $c = \{(\Gamma(a+b))/(\Gamma(a)\Gamma(b))\}^m$. The parameters a and b may be specified so that they agree with the prior beliefs about the mean duration of each regime. Because the prior mean of p_{ii} is equal to $\bar{p} = a/(a+b)$, the prior density of the duration d in each regime is approximately $\pi(d) = \bar{p}^{d-1}(1-\bar{p})$ with prior mean duration of $(a+b)/b$. Let n_{ii} denote the number of one-step transitions from state i to state i in the sequence S_n. Then multiplying the prior by the likelihood function of $P|S_n$ yields

$$p_{ii}|S_n \sim Beta(a+n_{ii},\, b+1), \quad i = 1,\ldots,m, \tag{7}$$

since $n_{i,i+1} = 1$. The probability p_{ii} $(1 \leqslant i \leqslant m)$ can now be simulated from Eq. (7) by letting

$$p_{ii} = \frac{x_1}{\sum_{j=1}^{2} x_j}, \quad x_1 \sim Gamma(a+n_{ii},\, 1);\; x_2 \sim Gamma(b+1,\, 1)$$

This completes the simulation of S_n and the non-zero elements of P.

3. Bayes factor calculation

In this section we consider the comparison of alternative change-point models (e.g., a model with one change point vs. one with more than one change point). The Bayesian framework is particularly attractive in this context because these models are non-nested. In such settings, the marginal likelihood of the respective models, and Bayes factors (ratios of marginal likelihoods), are the preferred means for comparing models (Kass and Raftery, 1995; Berger and Perrichi, 1996).

The computation of the marginal likelihood using the posterior simulation output has been an area of much current activity. A method developed by Chib (1995) is particularly simple to implement. The key point is that the marginal likelihood of model $\mathcal{M}_r$

$$m(Y_n|\mathcal{M}_r) = \int f(Y_n|\mathcal{M}_r, \Theta, P)\pi(\Theta, P)|\mathcal{M}_r)\,\mathrm{d}\psi$$

may be re-expressed as

$$m(Y_n|\mathcal{M}_r) = \frac{f(Y_n|\mathcal{M}_r, \Theta^*, P^*)\pi(\Theta^*, P^*|\mathcal{M}_r)}{\pi(\Theta^*, P^*|Y_n, \mathcal{M}_r)}, \tag{8}$$

where (Θ^*, P^*) is any point in the parameter space. Note that, for convenience, the notation does not reflect the fact that the size of the parameter space, and the parameters, are model dependent. The latter expression, which has been called the basic marginal likelihood identity, follows from Bayes theorem. This expression requires the value of the likelihood function $f(Y_n|\mathcal{M}_r, \psi^*)$ at the point $\psi^* = (\Theta^*, P^*)$ along with the prior and posterior ordinates at the same point. These quantities are readily obtained from the MCMC approach discussed above. The choice of the point ψ^* is in theory completely irrelevant but in practice it is best to choose a high posterior density point such as the maximum-likelihood estimate or the posterior mean. Given estimates of the marginal likelihood for two models $\mathcal{M}_r$ and $\mathcal{M}_s$, the Bayes factor of r vs. s is defined as

$$B_{rs} = \frac{m(Y_n|\mathcal{M}_r)}{m(Y_n|\mathcal{M}_s)}.$$

Large values of B_{rs} indicate that the data support $\mathcal{M}_r$ over $\mathcal{M}_s$ (Jeffreys, 1961).

In the rest of the discussion we explain how each of the quantities required for the marginal likelihood calculation is obtained. The model index $\mathcal{M}$ is suppressed for convenience.

3.1. Likelihood function at $\psi^* = (\Theta^*, P^*)$

A simple method for computing the likelihood function is available from the proposed parameterization of the change-point model. It is based on the decomposition

$$\ln f(Y_n|\psi^*) = \sum_{t=1}^{n} \ln f(y_t|Y_{t-1}, \psi^*),$$

where

$$f(y_t|Y_{t-1},\psi^*)=\sum_{k=1}^{m} f(y_t|Y_{t-1},\psi^*,s_t=k)p(s_t=k|Y_{t-1},\psi^*) \tag{9}$$

is the one-step ahead prediction density. The quantity $f(y_t|Y_{t-1}, \psi^*, s_t = k)$ is the conditional density of y_t given the regime $s_t = k$ whereas $p(s_t = k|Y_{t-1}, \psi^*)$ is the mass in Eq. (6). The one-step ahead density (and, consequently, the joint density of the data) is thus easily obtained. It should be noted that the likelihood function is required at the selected point ψ^* for the computation of the marginal likelihood. It is not required for the MCMC simulation.

3.2. Estimation by simulation

A Monte Carlo version of the EM algorithm can be used to find the maximum likelihood estimate of the parameters (Θ, P). This estimate can be used as the point ψ^* in the calculation of the marginal likelihood.

Note that the EM algorithm entails the following steps: First, the computation of the function

$$Q(\psi,\psi^{(i)})=\int_{S_n}\ln(f(Y_n,S_n|\psi))p(S_n|Y_n,\psi^{(i)})\,\mathrm{d}S_n, \tag{10}$$

which requires integrating S_n out from the complete data joint density $f(Y_n, S_n|\psi)$ with respect to the current distribution of S_n given Y_n and the current parameter estimate $\psi^{(i)}$. The second step is the maximization of this function to obtain the revised value $\psi^{(i+1)}$.

Due to the intractability of the integral above, the first step is implemented by simulation. Because the integration is with respect to the joint distribution

$$p(S_n|Y_n,\psi^{(i)})\equiv p(s_1,s_2,\dots,s_n|Y_n,\psi^{(i)}), \tag{11}$$

the Q function may be calculated as follows (see also Wei and Tanner, 1990). Let $S_{n,j}(j \leqslant N)$ denote the N draws of S_n from $p(S_n|Y_n, \psi^{(i)})$ and let

$$\hat{Q}(\psi)=N^{-1}\sum_{j=1}^{N}\ln\{f(Y_n,S_{n,j}|\psi)\}=N^{-1}\sum_{j=1}^{N}\{\ln f(Y_n|S_{n,j},\Theta)+\ln f(S_{n,j}|P)\}. \tag{12}$$

The M-step is implemented by maximizing the $\hat{Q}$ function over ψ. This two-step process is iterated until the values $\psi^{(i)}$ stabilize (N is usually set small at the start of the iterations and large as the maximizer is approached). The quantity thus obtained is the (approximate) maximum-likelihood estimate.

Each of these steps is quite easy. Estimating the Q function requires draws on S_n, and these are obtained by the method discussed in Section 2.1. In the M-step, the maximization is usually straightforward and separates conveniently into one involving Θ in the sampling model and one involving P in the jump process. The latter estimates are

$$\hat{p}_{ii} = \frac{\sum_{j=1}^{N} n_{ii,j}}{\sum_{j=1}^{N}(n_{ii,j}+1)}, \quad i = 1, \dots, m, \tag{13}$$

where $n_{ii,j}$ is equal to the number of transitions from state i to state i in the vector $S_{n,j}$.

3.3. Marginal likelihood

The estimate of the marginal likelihood is completed by computing the value of the posterior ordinate $\pi(\psi^*|Y_n)$ at ψ^*. By definition of the posterior density

$$\pi(\Theta^*, P^*|Y_n) = \pi(\Theta^*|Y_n)\pi(P^*|Y_n, \Theta^*),$$

where

$$\pi(\Theta^*|Y_n) = \int \pi(\Theta^*|Y_n, S_n)p(S_n|Y_n)\,\mathrm{d}S_n \tag{14}$$

and

$$\pi(P^*|Y_n, \Theta^*) = \int \pi(P^*|S_n)p(S_n|Y_n, \Theta^*)\,\mathrm{d}S_n, \tag{15}$$

since $\pi(P^*|Y_n, \Theta^*, S_n) = \pi(P^*|S_n)$. The first of these ordinates may be estimated as

$$\hat{\pi}(\Theta^*|Y_n) = G^{-1}\sum_{g=1}^{G} \pi(\Theta^*|Y_n, S_{n,g}),$$

using the G draws on S_n from the run of the Markov chain Monte Carlo algorithm. The value $\pi(\Theta^*|Y_n, S_{n,g})$ may be stored at the completion of each cycle of the simulation algorithm.

The calculation of the second ordinate in Eq. (15) requires an additional simulation $\{S_{n,j}\}_{j=1}^{G}$ of S_n from $p(S_n|Y_n, \Theta^*)$. These draws are readily obtained by appending a pair of additional calls to the simulation of S_n conditioned on (Y_n, Θ^*, P) and P conditioned on (Y_n, Θ^*, S_n) within each cycle of the MCMC algorithm. Because the ordinate $\pi(P^*|S_n)$ is a product of Beta densities from the first m rows of P, the estimate of the reduced conditional ordinate in Eq. (15) is

$$\begin{aligned}\hat{\pi}(P^*|Y_n,\Theta^*) &= G^{-1}\sum_{j=1}^{G}\prod_{i=1}^{m} \pi(p_{ii}|S_{n,j}) \\ &= G^{-1}\sum_{j=1}^{G}\prod_{i=1}^{m}\left\{\frac{\Gamma(a+b+n_{ii,j}+1)}{\Gamma(a+n_{ii,j})\Gamma(b+1)}\right\} p_{ii}^{(a+n_{ii,j}-1)}(1-p_{ii})^{(b+1-1)}.\end{aligned}$$

The log of the marginal likelihood from Eq. (8) is now given by

$$\ln \hat{m}(Y_n) = \ln f(Y_n|\psi^*) + \ln \pi(\Theta^*) + \ln \pi(P^*)$$
$$- \ln \hat{\pi}(\Theta^*|Y_n) - \ln \hat{\pi}(P^*|Y_n,\Theta^*). \tag{16}$$

The calculation of Eq. (16) is illustrated in the examples.

4. Examples

4.1. Binary data with change point

Consider first a sequence of binary observations $\{y_t\}$, where $y_t \in \{0,1\}$, and suppose that $y_t \sim$ *Bernoulli*(ξ_t), where the probability ξ_t is subject to change at unknown time points. This is a canonical but non-trivial problem involving change points. To illustrate the ideas discussed above, y_t is simulated from the process

$$\xi_t = \begin{cases} \theta_1, & t \leqslant 50, \\ \theta_2, & 50 < t \leqslant 100, \\ \theta_3, & 100 < t \leqslant 150, \end{cases}$$

where $\theta_1 = 0.5$, $\theta_2 = 0.75$ and $\theta_3 = 0.25$. The data is reproduced (in cumulative sum form) in Fig. 1. To keep the discussion simple, it is assumed that θ_k $(k \leqslant 3)$ is independent of any covariates. Note that the break at $t = 50$ is not clearly distinguishable in the graph.

One important inferential question is the estimation of the jump probability matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & 0 \\ 0 & p_{22} & p_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

and the change points $\{\tau_1,\tau_2\}$. A second question is the comparison of models with a single change point ($\mathscr{M}_1$), two change points ($\mathscr{M}_2$), and three change points ($\mathscr{M}_3$). Note that $\mathscr{M}_1$ contains the parameters $(\theta_1, \theta_2, p_{11})$, whereas $\mathscr{M}_2$ and $\mathscr{M}_3$ contain the parameters $(\theta_1, \theta_2, \theta_3, p_{11}, p_{22})$ and $(\theta_1, \theta_2, \theta_3, \theta_4, p_{11}, p_{22}, p_{33})$, respectively.

For the prior distributions, assume exchangeability within each model and let

$$\theta_k|\mathscr{M}_r \sim Beta(2{,}2), \qquad k \leqslant r + 1,$$

$$p_{ii}|\,\mathscr{M}_r \sim Beta(8{,}0.1), \quad i \leqslant r.$$

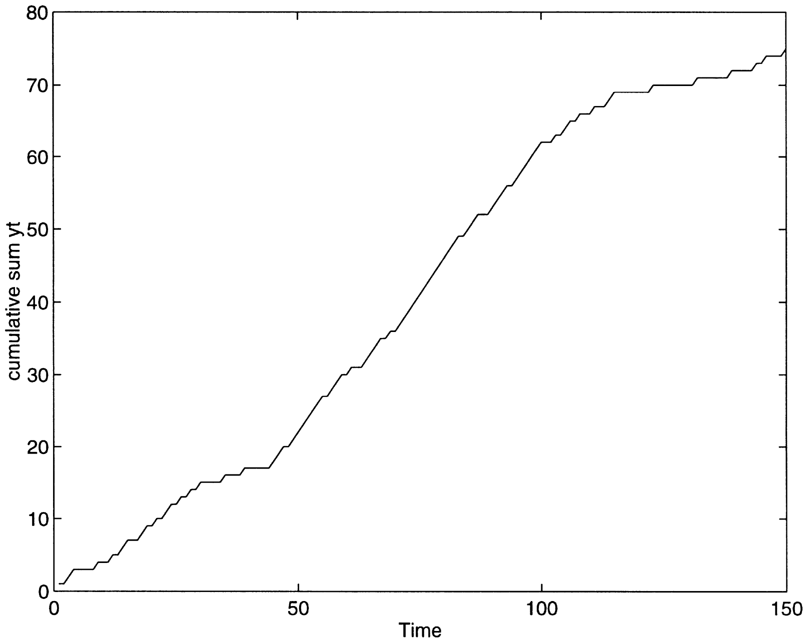

Fig. 1. Binary 0–1 data y_t in example 4.1.

The prior on p_{ii} reflects the prior beliefs that in model $\mathcal{M}_r$, the expected duration of each regime is approximately 80 periods. One could argue that this assumption is not reasonable for models $\mathcal{M}_2$ or $\mathcal{M}_3$ and a prior with shorter expected duration for these models should be specified (e.g., by letting the first parameter of the Beta distribution be 6 instead of 8). As one would expect, such a prior changes the marginal likelihood but, in this example, does not alter the rankings of the models that is reported below. In general, therefore, the key question is whether the conclusions change (not just whether the Bayes factors change) as the prior is perturbed by a reasonable amount. This must be ascertained on a case by case basis.

Under these prior distributions, $\{\theta_k\}$ in model $\mathcal{M}_r$, given the complete conditioning set, are independent with density

$$\pi(\theta_k|Y_n, S_n, P) = Beta(\theta_k|2 + U_k, 2 + N_k - U_k), \quad k \leqslant r + 1, \tag{17}$$

where $N_k = \sum_{t=1}^{n} I(s_t = k)$ is the number of observations ascribed to regime k and $U_k = \sum_{t=1}^{n} I(s_t = k)y_t$ is the sum of the y_t in regime k. This corresponds to the familiar Bayesian update for a Bernoulli probability with a single regime k. The MCMC simulation algorithm is completed by simulating $\{\theta_k\}$ from Eq. (17) and S_n and P as described in Sections 2.2 and 2.3. The MCMC sampling algorithm for all three models is conducted for $G = 6000$ iterations beyond a transient stage of 1000 iterations.

In the MCEM algorithm, the revised estimates of θ_k are obtained as $\hat{\theta}_k = U_k/N_k$ and those of P given S_n according to Eq. (13). The modal estimates were found by this algorithm at the end of a hundred EM steps. To evaluate the Q function in these iterations, N was taken to be 1 for the first 10 steps and gradually increased to 300 for the last ten steps.

Finally, in the marginal likelihood calculation the posterior ordinate at the modal estimate θ_k^* was found by averaging the beta density $\pi(\theta_k|Y_n, S_n, P)$ over the MCMC iterations, followed by a subsequent estimation of $\pi(P|S_n)$ from the reduced Gibbs run with θ_k set equal to θ_k^*. The likelihood function was estimated from Eq. (9) using

$$f(y_t|Y_{t-1}, s_t) = \theta_{s_t}^{y_t}(1 - \theta_{s_t})^{(1-y_t)}.$$

We summarize the results for the two change model in Fig. 2. The three lines in this figure correspond to the marginal probability that $s_t = k$ given the data Y_n at each time point t. It clearly indicates that the first 50 observations or so belong to the first regime, the next 50 to the second regime and the remaining to the third regime. The change points are identified very accurately by the intersections of the three lines.

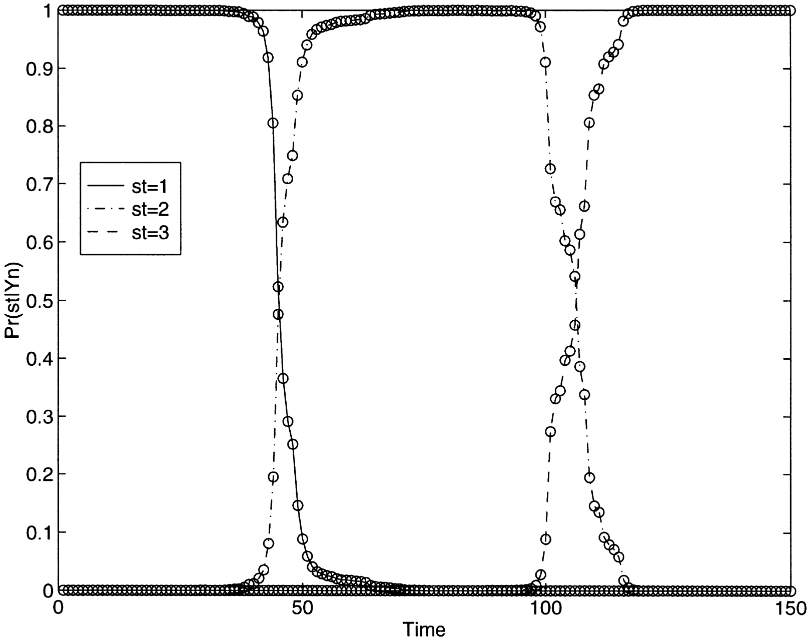

Fig. 2. Posterior probability of $s_t = k$ given binary data Y_n.

Table 1
Maximized log-likelihood and log marginal likelihood for binary data

	$\mathscr{M}_1$ (One change point)	$\mathscr{M}_2$ (Two change points)	$\mathscr{M}_3$ (Three change points)
$\ln f(Y_n\|\psi^*)$	$-$ 97.548	$-$ 92.912	$-$ 92.326
$\ln m(Y_n)$	$-$ 103.665	$-$ 101.872	$-$ 102.402

We report in Table 1 the evidence in the data for each of the three models. The results imply that the Bayes factor B_{21} is approximately six, while B_{23} is approximately 1.7. These Bayes factor provide (moderate) evidence in favor of the two change-point model in relation to the two competing models.

4.2. Poisson data with change point

As another illustration of the methods in the paper consider the much analyzed data on the number of coal-mining disasters by year in Britain over the period 1851–1962 (Jarrett, 1979). Let the count y_t in year t be modeled via a hierarchical Poisson model

$$f(y_t|\xi_t) = \xi_t^{y_t}\mathrm{e}^{-\xi_t}/y_t! \quad (t \leqslant 112)$$

and consider determining the evidence in favor of three models $\mathscr{M}_1 - \mathscr{M}_3$. Under $\mathscr{M}_1$, the no change point case, $\xi_t = \lambda$, $\lambda \sim Gamma(2,1)$. Under $\mathscr{M}_2$,ξ_t is subject to a single break:

$$\xi_t = \begin{cases} \lambda_1 & \text{for } t \leqslant \tau_1, \\ \lambda_2 & \text{for } \tau_1 + 1 \leqslant t \leqslant 112 \end{cases}$$

with

$$\lambda_1,\, \lambda_2 \sim Gamma(2,1).$$

Finally, under model $\mathscr{M}_3$, ξ_t is subject to two breaks with

$$\lambda_1,\, \lambda_2,\, \lambda_3 \sim Gamma(3,1).$$

The data on y_t is reproduced in Fig. 3. In this context it is of interest to fit all three models and to determine the Bayes factor B_{12} for 1 vs. 2, B_{13} for 1 vs. 3 and B_{23} for 2 vs. 3.

Carlin et al. (1992), in their analysis of these data, use a hierarchical Poisson Bayesian model and fit the one change-point model and find that the posterior mass function on τ_1 is concentrated on the observations 36–46 with a mode at y_{41}. The three largest spikes are at $t = 39, 40$, and 41, corresponding to a change sometime between 1889 and 1892. These results are also easily reproduced from

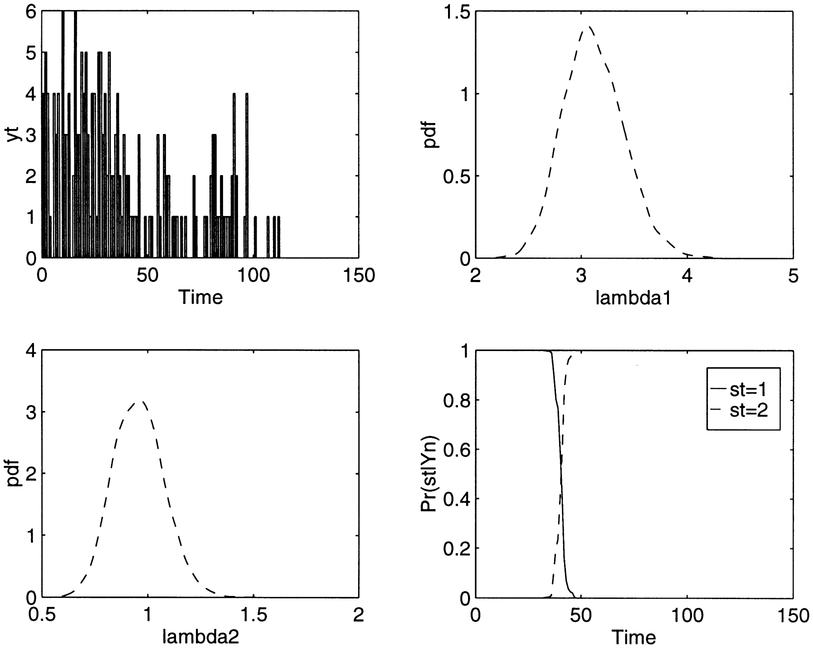

Fig. 3. Data on y_t, posterior marginal densities of λ_1 and λ_2 and $\Pr(s_t = k|Y_n)$.

our approach. It is not difficult to check that $\{(\lambda_1, \lambda_2)|(S_n, Y_n, P)\}$ factors into independent terms

$$\lambda_k|Y_n, S_n, P \sim Gamma(2 + U_k, 1 + N_k), \quad k \leqslant 2,$$

where $U_k = \sum_{t=1}^{n} I(s_t = k)y_t$ and $N_k = \sum_{t=1}^{n} I(s_t = k)$. As before, the MCMC algorithm is initialized arbitrarily (at 0.9 for p_{11} and 2 for λ_1 and λ_2) and implemented for $G = 6000$ iterations beyond a transient stage of a thousand steps. The prior on p_{11} is $Beta(8, 0.1)$. The MCEM algorithm, which yields the modal estimates λ_1^*, λ_2^* and P^*, is also implemented in the fashion described above, i.e., the Q function is estimated from a single realization of S_n in the first ten steps and $N = 300$ realizations in the final 10 steps. Finally, the marginal likelihood is computed in a manner analogous to that described for the Bernoulli model.

The results on fitting $\mathscr{M}_2$ are reproduced in Fig. 3. This figure gives the posterior marginal distributions of λ_1 and λ_2 and $\Pr(s_t = k|Y_n)$. The posterior means of λ_1 and λ_2 are found to be 3.119 and 0.957 with posterior standard deviations of 0.286 and 0.120, respectively.

The break in the process is identified as occurring at around $t = 41$. We also derive the posterior mass function on τ_1 by recording the time of the transition from $s_t = 1$ to $s_t = 2$ in the MCMC simulations. The frequency distribution of these transition points is given in Fig. 4. This mass function is virtually indistinguishable from the one computed in Carlin et al. (1992) by alternative means.

Next, we consider the efficacy of the proposed parameterization in fitting the two change-point model and contrast it with the results from the one-at-a time approach using the direct $\{\tau_1,\tau_2\}$ parameterization (which in this case can be implemented). The prior on λ_k is *Gamma*(3,1), $k \leqslant 3$ and that on p_{11} and p_{22} is *Beta*(5,0.1). The MCMC algorithm is run for 10,000 iterations in each case. The results for the marginal posterior distributions of τ_1 and τ_2 from the one-at-a time algorithm and the new algorithm are reported in Figs. 5 and 6, respectively.

Interestingly, the results in Fig. 5 are more diffuse and less plausible (both approaches produce the same posterior distributions for the remaining parameters of the model).

Finally, we report on the evidence for models $\mathcal{M}_1$ to $\mathcal{M}_3$. From Fig. 6 we see that the posterior distributions of the change points are concentrated around the same region. A formal calculation summarized in Table 2 confirms the lack of support for two change points. The log marginal likelihood for $\mathcal{M}_1$

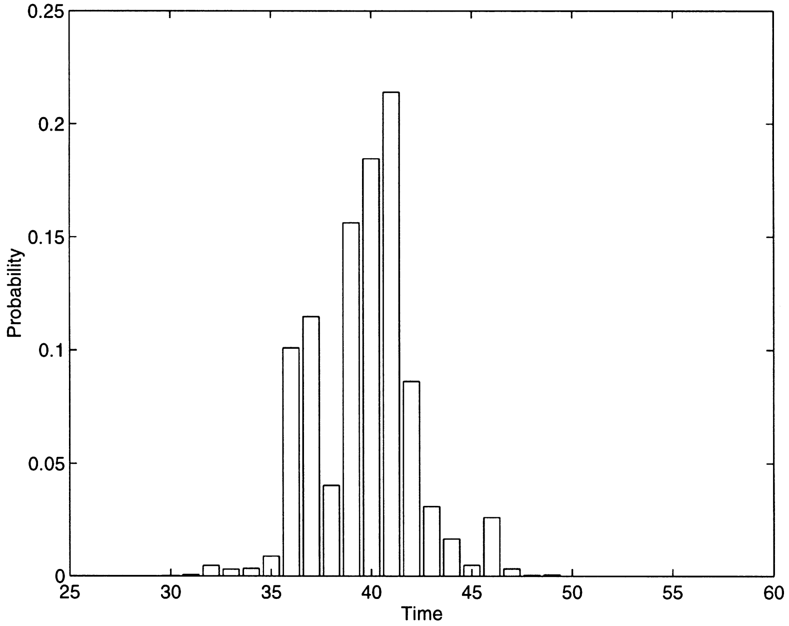

Fig. 4. Posterior probability mass function of τ_1.

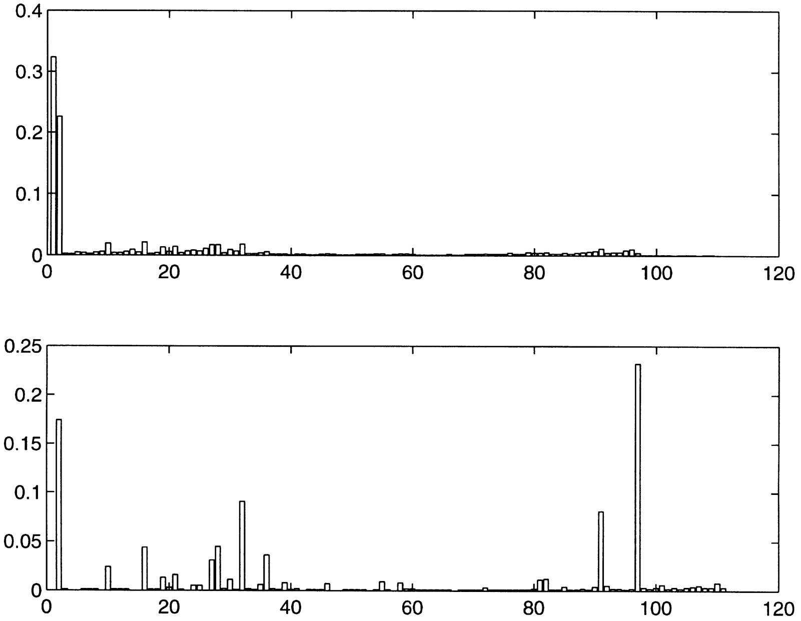

Fig. 5. Posterior probability function of τ_1 (top panel) and τ_2: One-at-a time algorithm.

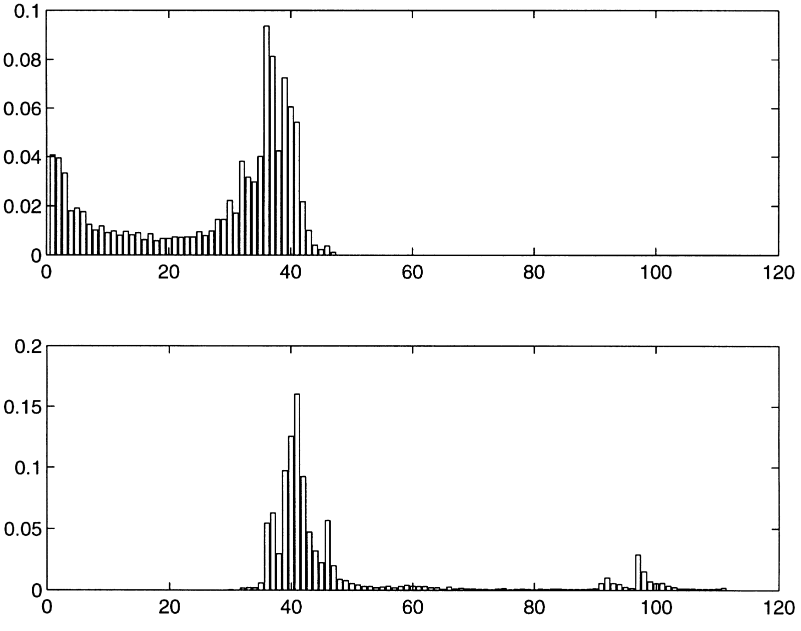

Fig. 6. Posterior probability function of τ_1 (top panel) and τ_2: New algorithm.

Table 2
Maximized log-likelihood and log marginal likelihood for coal mining disaster data

	$\mathscr{M}_1$ (No change point)	$\mathscr{M}_2$ (One change points)	$\mathscr{M}_3$ (Two change points)
$\ln f(Y_n \mid \psi^*)$	-203.858	-172.181	-171.450
$\ln m(Y_n)$	-206.365	-179.684	-180.836

is -206.365 and that of $\mathscr{M}_2$ is -172.181, implying decisive evidence in favor of $\mathscr{M}_2$.

The log marginal likelihood of the two change-point model $\mathscr{M}_3$ is -180.836, slightly lower than that of $\mathscr{M}_2$. Thus, we are able to conclude that these data do not support a second change point.

5. Concluding remarks

This paper has proposed a new approach for the analysis of multiple change-point models. The centerpiece of this approach is a formulation of the change-point model in terms of an unobserved discrete state variable that indicates the regime from which a particular observation has been drawn. This state variable is specified to evolve according to a discrete-time, discrete-state Markov process with the transition probabilities constrained so that the state variable can either stay at the current value or jump to the next higher value. This parameterization exactly reproduces the change-point model. In addition, it was shown that the MCMC simulation of this model is straightforward and improves on existing approaches in terms of computing time and speed of convergence.

The paper also provides a means for comparing alternative change-point models through the calculation of Bayes factors. This calculation which was hitherto not possible, in general, due to the intractability of the likelihood function is based on the computation of the marginal likelihood of each competing model from the output of the MCMC simulation. These calculations were illustrated in the context of models for binary and count data.

It is important to mention that the approach proposed here should prove useful in the development of new approaches for the classical analysis of the change-point model. Besides providing a simple approach for the computation of the likelihood function and maximum-likelihood estimates, it should allow for the construction of new tests due to the connection with hidden Markov models that is developed in this paper.

Finally, the approach leads to a new analysis for the class of epidemic change-point models. In a version of this model considered by Yao (1993), an epidemic state is followed by a return to a normal state. This model becomes a special case of the above framework if one restricts the state variable to take three values such that the parameter values in the first and last state (corresponding to the normal state) are identical. The MCMC analysis of the model then proceeds with little modification.

Acknowledgements

The comments of the editor, associate editor and two annonymous referees are gratefully acknowledged.

References

Albert, J., Chib, S., 1993. Bayes inference via Gibbs Sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business and Economic Statistics 11, 1–15.

Barry, D., Hartigan, J., 1993. A Bayesian analysis for change point problems. Journal of the American Statistical Association 88, 309–319.

Berger, J., Pericchi, L., 1993. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91, 109–122.

Carlin, B., Gelfand, A., Smith, A.F.M., 1992. Hierarchical Bayesian analysis of change-point problems. Applied Statistics 41, 389–405.

Carter, C.K., Kohn, R., 1994. On Gibbs Sampling for state space models. Biometrika 81, 541–553.

Chernoff, H., Zacks, S., 1964. Estimating the current mean of a Normal distribution which is subject to changes in time. Annals of Mathematical Statistics 35, 999–1018.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

Chib, S., 1996. Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics 75, 79–98.

Chib, S., Greenberg, E., 1995. Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. Journal of Econometrics 68, 339–360.

Chib, S., Greenberg, E., 1996. Markov Chain Monte Carlo simulation methods in econometrics. Econometric Theory 12, 409–431.

Inclan, C., 1993. Detection of multiple changes of variance using posterior odds. Journal of Business and Economic Statistics 11, 289–300.

Jarrett, R.G., 1979. A note on the interval between coal mining disasters. Biometrika 66, 191–193.

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

Kass, R., Raftery, A., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Liu, J.S., Wong, W.H., Kong, A., 1994. Covariance structure of the Gibbs Sampler with applications to the comparisons of estimators and data augmentation schemes. Biometrika 81, 27–40.

Shephard, N., 1994. Partial non-Gaussian state space. Biometrika 81, 115–131.

Stephens, D.A., 1994. Bayesian retrospective multiple-changepoint identification. Applied Statistics 43, 159–178.

Wei, G.C.G., Tanner, M.A., 1990. A Monte Carlo implementation of the EM Algorithm and the Poor Man's data augmentation algorithm. Journal of the American Statistical Association 85, 699–704.

Yao, Y.C., 1984. Estimation of a noisy discrete-time step function: Bayes and Empirical Bayes approaches. Annals of Statistics 12, 1434–1447.

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