Dirichlet Process Hidden Markov Multiple Change-point Model

Stanley I. M. Ko*, Terence T. L. Chong†, and Pulak Ghosh‡

Abstract. This paper proposes a new Bayesian multiple change-point model which is based on the hidden Markov approach. The Dirichlet process hidden Markov model does not require the specification of the number of change-points *a priori*. Hence our model is robust to model specification in contrast to the fully parametric Bayesian model. We propose a general Markov chain Monte Carlo algorithm which only needs to sample the states around change-points. Simulations for a normal mean-shift model with known and unknown variance demonstrate advantages of our approach. Two applications, namely the coal-mining disaster data and the real United States Gross Domestic Product growth, are provided. We detect a single change-point for both the disaster data and US GDP growth. All the change-point locations and posterior inferences of the two applications are in line with existing methods.

Keywords: Change-point, Dirichlet process, Hidden Markov model, Markov chain Monte Carlo, Nonparametric Bayesian.

1 Introduction

The earliest Bayesian change-point model is explored by Chernoff and Zacks (1964), who assume a constant probability of change at each point in time. Smith (1975) investigates the single change-point model under different assumptions of model parameters. Carlin et al. (1992) assume that the structural parameters are independent of the change-points and introduce the Markov chain Monte Carlo sampling method to derive the posterior distributions. Stephens (1994) further applies the Markov chain Monte Carlo method to the case of multiple changes. Chib (1998) allows the change-point probability to depend on the regime between two adjacent change-points. Koop and Potter (2007) propose the Poisson hierarchical prior for durations in the change-point model that allows the number of change-points to be unknown. More recent works on the Bayesian change-point model include Wang and Zivot (2000), Giordani and Kohn (2008), Pesaran et al. (2006), Maheu and Gordon (2008) and Geweke and Yu (2011).

In this paper we follow the modeling strategy of Chib (1998) which is one of the most popular Bayesian change-point models. He introduces a discrete random variable indicating the regime from which a particular observation is drawn. Specifically, let

---

*Department of Finance and Business Economics, University of Macau, Macau, StanleyKo@umac.mo

†Department of Economics and Institute of Global Economics and Finance, The Chinese University of Hong Kong, Hong Kong, and Department of International Eonomics and Trade, Nanjing University, China, chong2064@cuhk.edu.hk

‡Department of Quantitative Methods & Information Systems, Indian Institute of Management at Bangalore, India, pulak.ghosh@iimb.ernet.in

$Y_n = (y_1, y_2, \ldots, y_n)'$ be the observed time series, such that the density of y_t conditioned on $Y_{t-1} = (y_1, y_2, \ldots, y_{t-1})'$ depends on the parameter θ whose value changes at an unknown time period $1 < \tau_1 < \cdots < \tau_k < n$ and remains constant within each regime, that is,

$$
y_t \sim \begin{cases} p(y_t \mid Y_{t-1}, \theta_1) & \text{if } t \leq \tau_1, \\ p(y_t \mid Y_{t-1}, \theta_2) & \text{if } \tau_1 < t \leq \tau_2, \\ \vdots & \vdots \\ p(y_t \mid Y_{t-1}, \theta_k) & \text{if } \tau_{k-1} < t \leq \tau_k, \\ p(y_t \mid Y_{t-1}, \theta_{k+1}) & \text{if } \tau_k < t \leq n, \end{cases} \tag{1}
$$

where $\theta_i \in \mathbb{R}^l$ is an l dimension vector, $i = 1, 2, \ldots, k+1$. Note that we consider in this paper the change-point problem when the data are assumed to be generated by a parametric model where the unknown parameter θ_i changes with respect to different regimes. Let s_t be the discrete indicator variable such that

$$
y_t \mid s_t \sim p(y_t \mid Y_{t-1}, \theta_{s_t}), \tag{2}
$$

where s_t takes values in $\{1, 2, \ldots, k, k+1\}$. The indicator variable s_t is modeled as a discrete time, discrete-state Markov process with the constrained transition probability matrix

$$
P = \begin{pmatrix} p_{11} & p_{12} & 0 & \cdots & & 0 \\ 0 & p_{22} & p_{23} & \cdots & & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & p_{kk} & p_{k(k+1)} \\ 0 & 0 & \cdots & 0 & & 1 \end{pmatrix}, \tag{3}
$$

where $p_{ij} = \mathrm{pr}(s_t = j \mid s_{t-1} = i)$ is the probability of moving to regime j at time t given that the regime at time $t-1$ is i. With this parameterization, the ith change-point occurs at τ_i if $s_{\tau_i} = i$ and $s_{\tau_i+1} = i+1$.

As pointed out in Chib (1998), the above is a hidden Markov model where the transition matrix of the hidden state s_t is restricted as in (3). Hence, Chib's multiple change-point model inherits the limitation of the hidden Markov model in that the number of states has to be specified in advance. In light of this, Chib (1998) suggests to select from alternative models (e.g. one change-point vs. multiple change-points) according to the Bayes factors. In this paper, we introduce the Dirichlet process hidden Markov model (DPHMM) with left-to-right transition dynamic, without imposing restrictions on the number of hidden states. The use of the DPHMM has the following appealing features:

1. We do not have to specify the number of states *a priori*. The information provided by the observations determines the states endogenously. Hence, our method can be regarded as semiparametric.

2. Our modeling approach facilitates the sampling of states since we only need to sample the states around change-points.

We note that Kozumi and Hasegawa (2000) propose a method similar to ours in that they utilize a Dirichlet process prior for θ, but in a mixture model.

The rest of the paper is organized as follows. Section 2 provides a brief introduction of the Dirichlet process. Section 3 incorporates the Dirichlet process into the change-point model. The general Markov chain Monte Carlo sampler is discussed in Section 4. A Monte Carlo study of the normal mean-shift model is conducted in Section 5. Section 6 discusses learning of DPHMM parameters. Section 7 provides applications of our model and Section 8 concludes the paper.

2 The Dirichlet Process

Our new method employs the Dirichlet process technique which is widely used in non-parametric Bayesian models. The Dirichlet process prior is first proposed by Ferguson (1973). He derives the Dirichlet process prior as the prior on the unknown probability measure space with respect to some measurable space $(\Omega, \mathcal{F})$. Hence the Dirichlet process is a distribution over probability measures. Blackwell and MacQueen (1973) show that the Dirichlet process can be represented by the Polya urn model. Sethuraman (1994) develops the constructive sticking-breaking definition.

In the present study, we assume a Dirichlet process prior to each row of the transition matrix. The Dirichlet process is best defined here as the infinite limit of finite mixture models (Neal (1992), Neal (2000) and Beal et al. (2002)). To illustrate the idea, let us first consider the case with a finite number of states. With the left-to-right restriction to the transition dynamic, a particular state $s_{t-1} = i$ will either stay at the current state i or transit to a state $j > i$. A left-to-right Markov chain with k states will typically have the following upper triangular transition matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1k} \\ 0 & p_{22} & p_{23} & \cdots & p_{2k} \\ 0 & 0 & p_{33} & \cdots & p_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & p_{kk} \end{pmatrix}, \tag{4}$$

where the summation of each row equals 1. Note that the left-to-right Markov chain here is different from Chib's restricted band transition matrix (3). Here, the number of states k is not necessarily the number of regimes as is the case in Chib's model.

Let $\mathbf{p}_i = (0, \ldots, p_{ii}, p_{i(i+1)}, \ldots, p_{ik})$ be the transition probabilities of the ith row of the transition matrix (4). Suppose we draw m samples $\{c_1, \ldots, c_m\}$ of s_{t+1} given $s_t = i$ with probability profile $\mathbf{p}_i$. The joint distribution of the sample is thus

$$\text{pr}(c_1, \ldots, c_m \mid \mathbf{p}_i) = \prod_{j=i}^{k} p_{ij}^{m_j}, \tag{5}$$

where m_j denotes the number of samples that take state j, $j = i, \ldots, k$. We assume a symmetric Dirichlet prior $\pi(\mathbf{p}_i \mid \beta)$ for $\mathbf{p}_i$ with positive concentration parameter β:

$$\mathbf{p}_i \mid \beta \sim \text{Dirichlet}(\beta/(k-i+1), \ldots, \beta/(k-i+1)) = \frac{\Gamma(\beta)}{\Gamma\left(\frac{\beta}{k-i+1}\right)^{k-i+1}} \prod_{j=i}^{k} p_{ij}^{\beta/(k-i+1)-1}.$$

$$(6)$$

With the Dirichlet prior, we can analytically integrate out $\mathbf{p}_i$ such that

$$\text{pr}(c_1, \ldots, c_m \mid \beta) = \int \text{pr}(c_1, \ldots, c_m \mid \mathbf{p}_i, \beta) d\pi(\mathbf{p}_i \mid \beta)$$

$$= \frac{\Gamma(\beta)}{\Gamma(m+\beta)} \prod_{j=i}^{k} \frac{\Gamma\left(m_j + \frac{\beta}{k-i+1}\right)}{\Gamma\left(\frac{\beta}{k-i+1}\right)}. \tag{7}$$

The conditional probability of a sample $c_d \in \{c_1, \ldots, c_m\}$ given all other samples is thus

$$\text{pr}(c_d = j \mid \mathbf{c}_{-d}) = \frac{m_{-d,j} + \beta/(k+i-1)}{m-1+\beta}, \tag{8}$$

where $\mathbf{c}_{-d}$ denotes the sample set with c_d deleted, and $m_{-d,j}$ is the number of samples in $\mathbf{c}_{-d}$ that take state j.

Taking the limit of equation (8) as k tends to infinity, we have

$$\text{pr}(c_d = j \mid \mathbf{c}_{-d}) = \begin{cases} \frac{m_{-d,j}}{m-1+\beta} & j \in \{i, i+1, \ldots, k\}, \\ \frac{\beta}{m-1+\beta} & \text{for all potential states .} \end{cases} \tag{9}$$

Note that the probability that c_d takes an existing state, say j, is proportional to $m_{-d,j}$, which implies that c_d is more likely to choose an already popular state. In addition, the probability that a new state (i.e. $k+1$) takes place is proportional to β. Hence, there are potentially many states available, with infinite dimension transition matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ 0 & p_{22} & p_{23} & \cdots \\ 0 & 0 & p_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{10}$$

The actual state space can be regarded as consisting of an infinite number of states, only a finite number of which are actually associated with the data. Therefore, the number of states is endogenously determined.

3   The Dirichlet Process Hidden Markov Multiple Change-point Model and the State Evolution

Let us now turn to the proposed multiple change-point model and discuss a particular state evolution. Suppose we have already generated the hidden states up to $s_t = i$. We impose the Dirichlet process as described in Section 2 to s_{t+1}. In the change-point model, the transitions that have existed so far from state i are only self transitions. With the

left-to-right restriction, we will neither see a backward transition, i.e., transition from state i to some previous states, nor a forward transition, i.e., transition to some future states. The counts of the existing transitions from state i will be used as the counts defined in equation (9). Hence, we will have

$$\text{pr}(s_{t+1} = j \mid s_t = i, s_1, \ldots, s_{t-1}) = \begin{cases} \frac{n_{ii}}{n_{ii}+\beta} & j = i, \\ \frac{\beta}{n_{ii}+\beta} & s_{t+1} \text{ takes a new state,} \end{cases} \quad (11)$$

where $n_{ii} = \sum_{t'=1}^{t-1} \delta(s_{t'}, i)\delta(s_{t'+1}, i)$ denotes the counts of transitions that have occurred so far from state i to itself.[1] Note that in equation (11), s_{t+1} depends only on the *state* that s_t takes according to the Markovian property. All other previous states merely provide the transition counts.

We introduce a self-transition prior mass α for each state. The idea here is that if s_t transits to a new state, say $s_{t+1} = i + 1$, then without previous transition records, the next state s_{t+2} conditioned on $s_{t+1} = i+1$ will further take another new state with probability 1. Hence, with α, the trivial case is avoided and we have

$$\text{pr}(s_{t+1} = j \mid s_t = i, s_1, \ldots, s_{t-1}) = \begin{cases} \frac{n_{ii}+\alpha}{n_{ii}+\beta+\alpha} & j = i, \\ \frac{\beta}{n_{ii}+\beta+\alpha} & s_{t+1} \text{ takes a new state.} \end{cases} \quad (12)$$

Therefore, the whole Markov chain is characterized by two parameters, α and β, instead of a transition probability matrix. We can see that α controls the prior tendency to linger in a state, and β controls the tendency to explore new states. Figure 1 illustrates three left-to-right Markov chains of length $n = 150$ with different α and β. Figure 1a depicts the chain that explores many new states with very short linger time. Figure 1b shows the chain with long linger time and less states. Figure 1c lies in between.

Equation (12) coincides with Chib (1998)'s model when the probability p_{ii} is integrated out. Specifically, in Chib (1998),

$$\begin{aligned} \text{pr}(s_{t+1} &= j \mid s_t = i, s_1, \ldots, s_{t-1}) \\ &= \int p(s_{t+1} = j \mid s_t = i, s_1, \ldots, s_{t-1}, p_{ii})f(p_{ii})dp_{ii} \\ &= \begin{cases} \frac{n_{ii}+a}{n_{ii}+a+b} & j = i, \\ \frac{b}{n_{ii}+a+b} & j = i+1 \end{cases} \end{aligned} \quad (13)$$

where $p_{ii} \sim \text{Beta}(a, b)$ and $f(p_{ii})$ is the corresponding density. However, our model stems from a different perspective. The derivations in the previous section and equation (12) follow the nonparametric Bayesian literature (Neal (2000)) and the infinite HMM of Beal et al. (2002). Indeed, it is known that when the Dirichlet process is truncated at a finite number of states, the process reduces to the generalized Dirichlet distribution (GDD), see Connor and Mosimann (1969) and Wong (1998). For the same reason we have (12) coincides with (13). We would like to point out that our modeling strategy facilitates the Gibbs sampler of s_t which is different from Chib (1998). We will elaborate further on the Gibbs sampler in the next section. Also, learning of α and β will be discussed in Section 6.

---

[1] The Kronecker-delta function $\delta(a, b) = 1$ if and only if $a = b$ and 0 otherwise.

(a) $\alpha = 1, \beta = 5$.   (b) $\alpha = 5, \beta = 1$.   (c) $\alpha = 5, \beta = 5$.
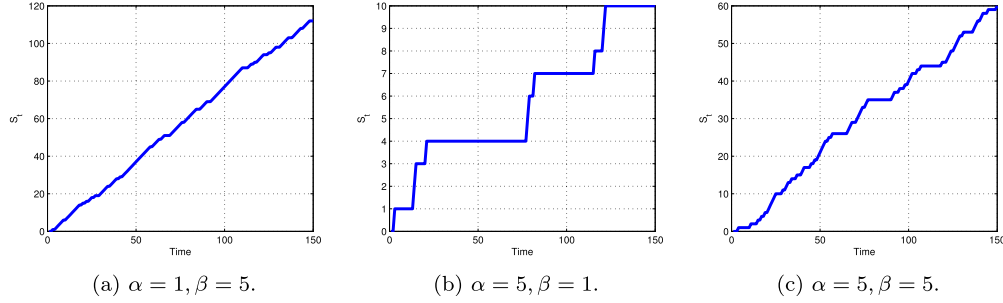
Figure 1: Left-to-right Markov chain with different α and β.

4   Markov Chain Monte Carlo Algorithm

4.1   General

Suppose we have observations $Y_n = (y_1, y_2, \ldots, y_n)'$. Given the state $S_n = (s_1, \ldots, s_n)'$ we have

$$y_t \mid s_t \sim p(y_t \mid Y_{t-1}, \theta_{s_t}), \tag{14}$$

where $\theta_{s_t} \in \mathbb{R}^l$, $Y_{t-1} = (y_1, \ldots, y_{t-1})'$. Let $\theta = (\theta_1, \ldots, \theta_k)'$ and γ denotes a hyperparameter. Recall that we impose the DPHMM to the states and a hierarchical model to the parameter, we are thus interested in sampling from the posterior $p(\theta, S_n, \gamma \mid Y_n)$ given the priors $p(\theta|\gamma)$, $p(\gamma)$ and $p(S_n)$. The general Gibbs sampler procedure is to sample the following in turn:

Step 1. $S_n \mid \theta, \gamma, Y_n$,

Step 2. $\theta \mid \gamma, S_n, Y_n$,

Step 3. $\gamma \mid \theta, S_n, Y_n$.

We will discuss the three steps below.

4.2   Simulation of S_n

The state prior $p(S_n)$ can be easily derived from (12). Moreover, the full conditional is

$$p(S_n \mid \theta, \gamma, Y_n) \propto p(S_n)p(Y_n \mid S_n, \theta, \gamma). \tag{15}$$

Simulation of S_n from the full conditional (15) is done by the Gibbs sampler. Specifically, we draw s_t in turn for $t = 1, 2, \ldots, n$ from

$$p(s_t \mid S_{t-1}, S^{t+1}, \theta, \gamma, Y_n) \propto p(s_t \mid s_{t-1}, S_{t-2})p(s_{t+1} \mid s_t, S^{t+2})p(y_t \mid s_t, Y_{t-1}, \theta, \gamma), \tag{16}$$

where $S_{t-1} = (s_1, \ldots, s_{t-1})'$ and $S^{t+1} = (s_{t+1}, \ldots, s_n)'$. The most recent updated values of the conditioning variables are used in each iteration. Note that we write $p(s_t \mid s_{t-1}, S_{t-2})$ and $p(s_{t+1} \mid s_t, S^{t+2})$ to emphasize the Markov dynamic; the other conditioning states merely provide the counts.

With the left-to-right characteristic of the chain, we do not have to sample all s_t from $t = 1$ to T. Instead, we only need to sample the state in which a change-point takes place. To see this, let us consider a concrete example. Suppose from the last sampler, we have

s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	$\cdots$
1	1	1	2	2	3	3	3	4	5	5	$\cdots$

With left-to-right transition restriction, s_t requires a sampling from (16) if and only if s_{t-1} and s_{t+1} differ. For other cases, s_t is unchanged with probability one. Suppose we are at $t = 2$. Since s_1 and s_3 are both equal to 1, s_2 is forced to take 1. In the above chain, the first state that needs to sample from (16) is s_3, which will either take the values 1 or 2 (s_2 or s_4). If s_3 takes 2 in the sampling (i.e., joining the following regime), then the next state to sample would be s_5; otherwise (i.e., joining the preceding regime), the next state to sample is s_4 because $s_5 - s_3 \neq 0$ for $s_3 = 1$. Now suppose we are at $t = 9$. We will draw a new s_9 and s_9 will either join the regime of s_8 or the regime of s_{10}. This will look strange because a gap exists in the chain. However, our concern here is the *consecutive grouping* or *clustering* in the series. We can alternatively think that the state represented by s_9 is simply pruned away in the current sweep. Note the numbers assigned to the s_t's are nothing but indicators of regimes.[2] Therefore, we will relabel the s_t's after each sweep.

In general, suppose $s_{t-1} = i$ and $s_{t+1} = i+1$. s_t takes either i or $i+1$. Table 1 shows the corresponding probability values specified in (16). To see this, if s_t takes i, then the transition from s_{t-1} to s_t is a self-transition and that from s_t to s_{t+1} is an innovation. The corresponding probability values are in the first row of Table 1. The reasoning for $s_t = i + 1$ is similar. Note the changes of counts in different situations.

Table 1: Sampling probabilities of s_t. $n_{ii} = \sum_{t'=1}^{t-2} \delta(s_{t'}, i)\delta(s_{t'+1}, i)$ and $n_{i+1,i+1} = \sum_{t'=t+1}^{n-1} \delta(s_{t'}, i+1)\delta(s_{t'+1}, i+1)$.

	$p(s_t \mid s_{t-1} = i, S_{t-2})$	$p(s_{t+1} = i+1 \mid s_t, S^{t+2})$	$p(y_t \mid s_t, Y_{t-1}, \theta)$
$s_t = i$	$\dfrac{n_{ii} + \alpha}{n_{ii} + \beta + \alpha}$	$\dfrac{\beta}{n_{ii} + 1 + \beta + \alpha}$	$p(y_t \mid Y_{t-1}, \theta_i)$
$s_t = i+1$	$\dfrac{\beta}{n_{ii} + \beta + \alpha}$	$\dfrac{n_{i+1,i+1} + \alpha}{n_{i+1,i+1} + \beta + \alpha}$	$p(y_t \mid Y_{t-1}, \theta_{i+1})$

For the initial point s_1, if currently $s_2 - s_1 \neq 0$, then we can sample s_1 from

$$\mathrm{pr}(s_1 \mid s_2, s_3, \ldots, s_n) = \begin{cases} c \cdot \frac{\alpha}{\beta+\alpha} \cdot \frac{\beta}{\beta+\alpha} \cdot p(y_1 \mid Y_0, \theta_{s_1}) & \text{if } s_1 \text{ unchanged,} \\ c \cdot \frac{\beta}{\beta+\alpha} \cdot \frac{n_{s_2 s_2} + \alpha}{n_{s_2 s_2} + \beta + \alpha} \cdot p(y_1 \mid Y_0, \theta_{s_2}) & \text{if } s_1 = s_2, \end{cases} \quad (17)$$

---

[2]This will exclude the case of the regime with only one data point. We do not consider this situation here.

where $n_{s_2 s_2} = \sum_{t'=2}^{n-1} \delta(s_{t'}, s_2) \delta(s_{t'+1}, s_2)$ and c is the normalizing constant. For the end point s_n, if $s_n - s_{n-1} \neq 1$, we sample s_n from

$$\mathrm{pr}(s_n \mid s_{n-1}, s_{n-2}, \ldots, s_1) = \begin{cases} c \cdot \dfrac{n_{s_{n-1} s_{n-1}} + \alpha}{n_{s_{n-1} s_{n-1}} + \beta + \alpha} \cdot p(y_n \mid Y_{n-1}, \theta_{s_{n-1}}) & \text{if } s_n = s_{n-1}, \\[2ex] c \cdot \dfrac{\beta}{n_{s_{n-1} s_{n-1}} + \beta + \alpha} \cdot p(y_n \mid Y_{n-1}, \theta_{s_n}) & \text{if } s_n \text{ unchanged,} \end{cases} \tag{18}$$

where $n_{s_{n-1} s_{n-1}} = \sum_{t'=1}^{n-2} \delta(s_{t'}, s_{n-1}) \delta(s_{t'+1}, s_{n-1})$ and c is the normalizing constant.

As mentioned in Section 3, the DPHMM facilitates the Gibbs sampler of states. In sampling s_t from (16), we simultaneously use up all information of the transitions prior to t (i.e. $s_1, \ldots, s_{t-1}$) and after t (i.e. $s_{t+1}, \ldots, s_T$) which are captured in $p(s_t \mid s_{t-1}, S_{t-2})$ and $p(s_{t+1} \mid s_t, S^{t+2})$. Thus far, our algorithm only requires a record of transitions and draws at the point where the structural change takes place, whereas we have to sample all s_t in Chib (1998).

4.3   Updating θ and γ

Given S_n and Y_n, the full conditionals of θ and γ are simply

$$\begin{aligned} p(\theta_i \mid \gamma, S_n, Y_n) &\propto p(\theta_i \mid \gamma) \prod_{\{t: s_t = i\}} p(y_t \mid Y_{t-1}, \theta_i), \\ p(\gamma \mid \theta, S_n, Y_n) &\propto p(\gamma)\, p(\theta \mid \gamma, S_n, Y_n), \end{aligned} \tag{19}$$

which are model specific. In the following sections, we will study a simulated normal mean-shift model, a discrete type Poisson model, and an AR(2) model.

4.4   Initialization of States

In Section 4.2, we have discussed the simulation of the states S_n. The number of change-points is inherently estimated through the sampling of states in equation (16). Within the burn-in period, the state number will be changing around after each MCMC pass. After the burn-in period, the Markov chain converges and hence the number of states becomes stable. Theoretically, it is legitimate to set any number of change-points in the beginning and let the algorithm find out the convergent number of states. In practice, we find it is more efficient to initialize with a large number of states and let the algorithm prune away redundant states, rather than allow for the change-point number to grow from a small number. Specifically, suppose a reasonably large state number k is proposed. We initialize equidistant states, that is

$$s_t = i, \quad \text{if } \frac{(i-1) \cdot n}{k} < t \leq \frac{i \cdot n}{k}, \tag{20}$$

where $i = 1, \ldots, k$. Then the algorithm described above will work out the change-point locations and the number of states after convergence of the Markov chain.

5 A Monte Carlo Study: the Normal Mean-Shift Model

5.1 The Model

In this section, we first study the normal mean-shift model with known variance σ^2. Suppose the normal data $Y_n = (y_1, \ldots, y_n)'$ is subject to unknown k changes in mean. We use the following hierarchical model

$$
\begin{aligned}
&y_t \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma^2) \quad \text{if } \tau_{i-1} < t \leq \tau_i, \\
&\theta_i \mid \mu, \upsilon^2 \sim \mathrm{N}(\mu, \upsilon^2), \\
&(\mu, \upsilon^2) \sim \text{Inv-Gamma}(\upsilon^2 \mid a, b),
\end{aligned}
\tag{21}
$$

where σ^2 is known and τ_i $(i = 1, \ldots, k)$ is the change-point. We set $\tau_0 = 0$ and $\tau_{k+1} = n$. Next, we apply our algorithm to the case of unknown variance. In addition to (21), we assume the Inverse-Gamma prior for σ^2

$$
\sigma^2 \sim \text{Inv-Gamma}(\sigma^2 \mid c, d).
\tag{22}
$$

All derivations of the full conditionals and the Gibbs samplers are given in the Appendix.

We simulate two normal sequences with the parameters specified in Table 2. Specifically, Model 1 is subject to one change-point occurring at $t = 50$. Model 2 is a two change-points model with breaks at $t = 50$ and $t = 100$. Both models assume variance $\sigma^2 = 3$. Two realizations with respect to Model 1 and Model 2 are shown in Figure 2. We can see the overlapping of the data ranges of different regimes and it is hard to visually identify the change-points.
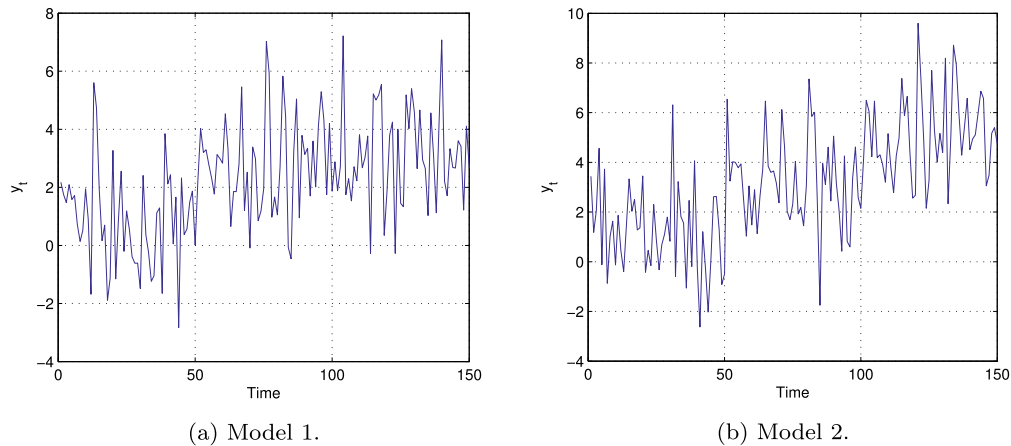

(a) Model 1.   (b) Model 2.

Figure 2: Random realizations of Model 1 and Model 2.

Table 2: Normal Mean-Shift Models 1 and 2.

|	θ_1	θ_2	θ_3	σ^2	τ_1	τ_2	k	n
Model 1 (One change-point)	1	3	-	3	50	-	1	150
Model 2 (Two change-points)	1	3	5	3	50	100	2	150

5.2 Simulation Results

To implement our algorithm, we set the inverse-Gamma hyperparameters $a = b = c = d = 1$, and the DPHMM parameters $\alpha = 3$ and $\beta = 2$. The two Gibbs samplers for the cases of known and unknown variance are conducted for 5000 sweeps with 5000 burn-in samples, respectively. The 5000 sweeps after the burn-in period are thinned with 50 draws to reduce dependence of iterations. The first column of Figure 3 shows the probabilities of regime indicator $s_t = i$ of the two models. Intersections of the lines $s_t = i$ clearly demonstrate the break locations.

To compare our proposed DPHMM to Chib's method, we also report the posterior inference of Chib's model under the true change-point number and the same model specification as in (A-3) and (A-4).[3] The posterior means and standard deviations of parameters are summarized in Table 3. First, our method performs well in all cases where the posterior distributions concentrate on the true values. The sample first-order serial correlations demonstrate good mixing of the samplers. Second, our results are comparable to those estimated from Chib's model. The Bayes factors show that in most of the cases the models with the true number of change-points are preferred to others. For example, in Model 2 where two change-points exist, the Bayes factors comparing models with $k = 1$ versus $k = 2$ are close to zero favoring the two change-point model. Likewise, the Bayes factors comparing $k = 2$ versus $k = 3$ favor the two change-point model. Hence we conclude that the model with two change-points is correctly specified with high probability. However, the Bayes factor fails to detect the correct number of changes in Model 1 with unknown variance. The values suggest a model with two change points. The posterior probabilities of states estimated by Chib's model are shown in the second column of Figure 3. In summary, the simulation results demonstrate that our algorithm works well in the normal mean-shift models and is robust to the change-point number compared to Chib's model.
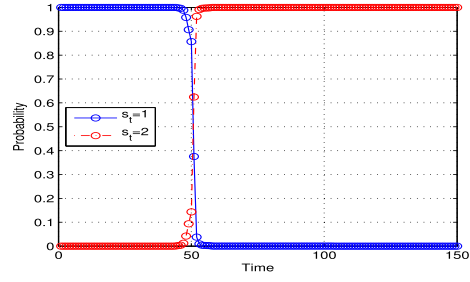
5.3 Robustness Check of Change-Point Number

In this section, we study the robustness of our algorithm in detecting the true number of change-points. Although our method does not require prespecification of the change-point number, it is still possible that our algorithm fails to estimate the correct number of change-points. Thus, we replicate the entire estimation process as in the previous section for 1000 times and record the estimated change-point number in each replication.
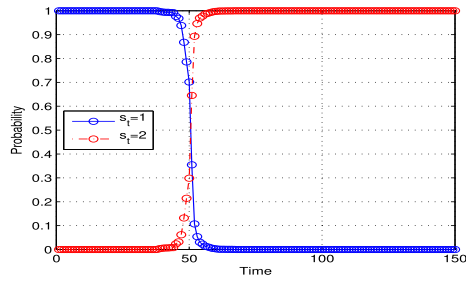
---

[3]The prior of the transition probabilities in Chib (1998)'s model is assumed to be Beta(a, b). The parameters a and b are chosen to reflect equidistant duration of each state. For example, in the case of one change-point with $n = 150$ sample size, we take $b = 0.1$ and $a = n/2 \times b = 7.5$, i.e. Beta$(7.5, 0.1)$.
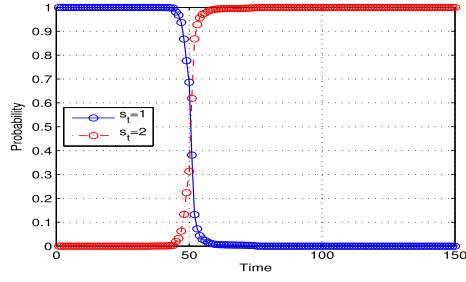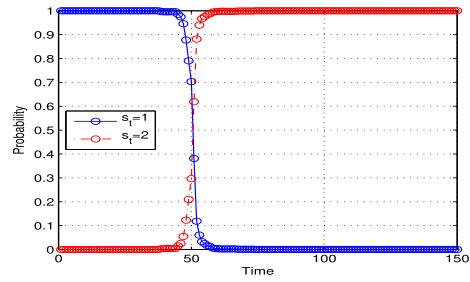
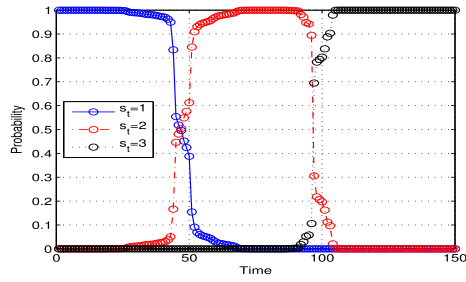DPHMM: Model 1 with known variance.   Chib: Model 1 with known variance.
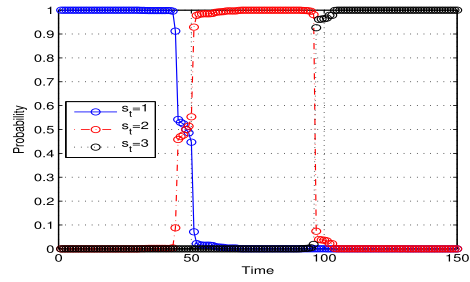
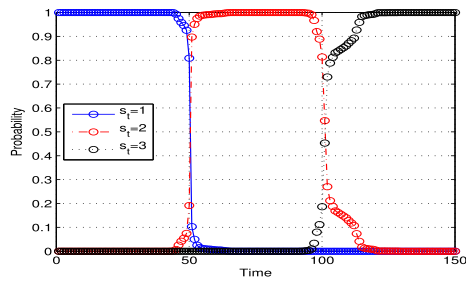DPHMM: Model 1 with unknown variance.   Chib: Model 1 with unknown variance.
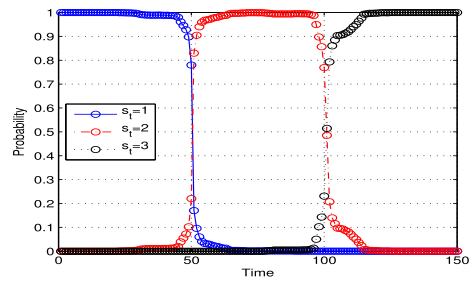
DPHMM: Model 2 with known variance.   Chib: Model 2 with known variance.

DPHMM: Model 2 with unknown variance.   Chib: Model 2 with unknown variance.

Figure 3: Posterior probability of $s_t = i$: DPHMM vs. Chib's model.

Table 3: Posterior estimates of Normal Mean-Shift Models 1 and 2. Mean and SD denote, respectively, posterior mean and posterior standard deviation.

	Model 1					
	Known variance			Unknown variance		
	Mean	SD	True value	Mean	SD	True value
			DPHMM			
θ_1	0.9123	0.2519	1.0000	0.9154	0.2432	1.0000
θ_2	2.9375	0.1812	3.0000	2.9354	0.1723	3.0000
σ^2				2.8244	0.3343	3.0000
			Chib's Model with $k = 1$			
θ_1	0.8980	0.1889	1.0000	0.9521	0.2441	1.0000
θ_2	2.9520	0.1324	3.0000	2.9331	0.1619	3.0000
σ^2				2.7933	0.4121	3.0000
			Bayes Factor Analysis			
$k = 1$ vs. $k = 2$	1.730			0.548		
$k = 1$ vs. $k = 3$	1.374			1.010		
$k = 2$ vs. $k = 3$	0.794			1.850		
	Model 2					
	Known variance			Unknown variance		
	Mean	SD	True value	Mean	SD	True value
			DPHMM			
θ_1	1.1746	0.2777	1.0000	1.2770	0.2584	1.0000
θ_2	2.9758	0.2874	3.0000	3.2176	0.2672	3.0000
θ_3	5.3344	0.2459	5.0000	5.1108	0.2682	5.0000
σ^2				3.1102	0.3726	3.0000
			Chib's Model with $k = 2$			
θ_1	1.1770	0.2052	1.0000	1.3680	0.2643	1.0000
θ_2	2.9480	0.1916	3.0000	3.1920	0.2730	3.0000
θ_3	5.3320	0.1823	5.0000	5.0310	0.2461	5.0000
σ^2				3.0780	0.6794	3.0000
			Bayes Factor Analysis			
$k = 1$ vs. $k = 2$	0.000			0.0193		
$k = 1$ vs. $k = 3$	0.000			0.0583		
$k = 2$ vs. $k = 3$	3.090			3.0332		

Specifically, in each of the 1000 replications, we iterate the Gibbs sampler for 5000 times and the change-point number of the last sample is recorded. Therefore, we obtain 1000 collections of change-point numbers. Table 4 reports the frequencies of the detected change-point numbers. We see that with high frequency (over 99%) our method detects one change-point ($k = 1$) in Model 1 in both cases of known and unknown variance. In Model 2, our results show that over 90% of the 1000 replications detect two change-points ($k = 2$), and over 99% detect at least one change-point. The figures demonstrate

that our algorithm correctly detects the change-point number with high probability in different cases.

Table 4: Frequencies of estimated change-point numbers.

	Known variance			Unknown variance		
	$k = 0$	$k = 1$	$k = 2$	$k = 0$	$k = 1$	$k = 2$
Model 1 (One change-point)	0.3%	99.7%	0.0%	0.5%	99.5%	0.0%
Model 2 (Two change-points)	0.0%	6.5%	93.5%	0.2%	8.7%	91.1%

6   Learning α and β

In the previous section, we set the DPHMM parameters $\alpha = 3$ and $\beta = 2$ in estimating the simulated models. In order to learn about α and β, we propose to use vague Gamma priors, see Beal et al. (2002). Note that with the number of states specified in each MCMC sweep, the DPHMM reduces to the generalized Dirichlet distribution (GDD), see Connor and Mosimann (1969) and Wong (1998). Hence the posterior is

$$p(\alpha, \beta \mid S_n) \propto \text{Gamma}(a_\alpha, b_\alpha)\text{Gamma}(a_\beta, b_\beta) \prod_{i=1}^{k+1} \frac{\beta\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(n_{ii} + \alpha)}{\Gamma(n_{ii} + 1 + \alpha + \beta)}, \quad (23)$$

where $n_{ii} = \sum_{t=1}^{T-1} \delta(s_t, i)\delta(s_{t+1}, i)$ denotes the counts of self transitions. We set $a_\alpha = b_\alpha = a_\beta = b_\beta = 1$ here and in the subsequent sections. Below we consider two alternative approaches for sampling: the first based on maximum-a-posteriori (MAP) estimation and a second approach using a random walk sampler.

6.1   The Maximum-a-Posteriori

We first solve for the maximum-a-posteriori (MAP) estimates for α and β which are obtained as the solutions to the following gradients using the Newton-Raphson method,

$$\frac{\partial \ln p(\alpha, \beta \mid S_n)}{\partial \alpha} = \frac{a_\alpha - 1}{\alpha} - b_\alpha$$
$$+ \sum_{i=1}^{k+1} [\psi(\alpha + \beta) + \psi(n_{ii} + \alpha) - \psi(\alpha) - \psi(n_{ii} + 1 + \alpha + \beta)] = 0 \quad (24)$$
$$\frac{\partial \ln p(\alpha, \beta \mid S_n)}{\partial \beta} = \frac{a_\beta - 1}{\beta} - b_\beta + \sum_{i=1}^{k+1} \left[\frac{1}{\beta} + \psi(\alpha + \beta) - \psi(n_{ii} + 1 + \alpha + \beta)\right] = 0,$$

where $\psi(\cdot)$ is the digamma function defined as $\psi(x) = d\ln\Gamma(x)/dx$.

We implement our algorithm in the previous section together with the MAP update of α and β in each sweep. The DPHMM with MAP update correctly detects the true

Table 5: MAP of α and β in Normal Mean-Shift Models 1 and 2. Average MAP values of α and β are reported with standard deviations within parentheses. For other parameters, the values are posterior means and posterior standard deviations.

	Model 1		Model 2	
	known variance	*unknown variance*	*known variance*	*unknown variance*
α	0.6353 (0.0007)	0.6353 (0.0007)	0.9453 (0.0008)	0.9451 (0.0005)
β	0.1937 (0.0005)	0.1937 (0.0005)	0.2364 (0.0001)	0.2361 (0.0005)
θ_1	0.9145 (0.2560)	0.9049 (0.2430)	1.1843 (0.2847)	1.2675 (0.2585)
θ_2	2.9326 (0.1745)	2.9385 (0.1731)	2.9859 (0.2892)	3.2222 (0.2716)
θ_3			5.3349 (0.2468)	5.1026 (0.2661)
σ^2		2.8207 (0.3340)		3.0908 (0.3752)

number of change-points in all cases. Table 5 shows the MAP solutions for α and β, and the posterior estimates of all parameters in each model. We can see that the average MAP values of α and β are 0.6353 and 0.1937 respectively in Model 1 with known and unknown variance. The results are slightly different in Model 2 such that average MAP values are 0.9451 and 0.2360 respectively. We also report the sample standard errors which show evidence of stability of the MAP values after the burn-in period. In all cases, α is greater than β indicating that the algorithm tends to linger in existing states rather than exploring a new one. Besides, all parameter estimates are in line with the results in the previous section when α and β are prespecified.

6.2   The Metropolis-Hastings Sampler

We also consider a Metropolis-Hastings (M-H) sampler for the posterior (23). The candidate-generating density is assumed to be the random walk process with positive support

$$f(\alpha'|\alpha) \propto \phi(\alpha' - \alpha), \quad \alpha' > 0, \tag{25}$$

where α is the value of the previous draw, $\phi(\cdot)$ is the standard normal density function. The acceptance ratio given β is thus

$$A(\alpha, \alpha') = \frac{p(\alpha', \beta \mid S_n)\,\Phi(\alpha)}{p(\alpha, \beta \mid S_n)\,\Phi(\alpha')}, \tag{26}$$

where $\Phi(\cdot)$ is the standard normal distribution function. The same M-H sampler is also applied to β given the updated α. We incorporate the M-H sampler of α and β in the Gibbs sampler in Section 5. The posterior estimators are shown in Table 6. The posterior means and standard deviations of parameter θ_i and σ^2 are similar to those obtained in the previous analyses. The posterior mean of α is greater than the posterior mean of β in all models. This affirms the conclusion in the MAP results that the algorithm tends to linger in existing states rather than exploring a new one. Both the MAP and M-H methods correctly estimate the number of change-points in each simulation.

Table 6: M-H sampler of α and β in Normal Mean-Shift Models 1 and 2. Posterior means and posterior standard deviations within parentheses.

	Model 1		Model 2	
	known variance	*unknown variance*	*known variance*	*unknown variance*
α	1.8073 (1.3345)	1.8145 (1.3646)	2.2007 (1.5179)	2.1482 (1.4731)
β	0.3446 (0.2168)	0.3455 (0.2176)	0.3667 (0.2087)	0.3622 (0.2086)
θ_1	1.1182 (0.2521)	1.0839 (0.2399)	1.1256 (0.2558)	1.0892 (0.2457)
θ_2	2.9447 (0.1724)	3.0939 (0.1701)	2.8277 (0.2644)	3.0341 (0.2449)
θ_3			5.0138 (0.2461)	5.1172 (0.2454)
σ^2		2.8190 (0.3298)		2.8793 (0.3455)

6.3 Comparison between MAP and M-H

We can see that the estimates of α and β from the two approaches are quite different as shown in Tables 5 and 6. The MAP as a point estimator may not reflect the variations of α and β, whereas the M-H is a typical Bayesian method which can be incorporated into the MCMC sampler of other parameters in question. Moreover, the MAP approach may be limited when the posterior happens to be multi-modal. Therefore, the M-H method is preferred in practice and the following empirical studies are conducted with the M-H sampler.

7 Empirical Applications

7.1 Poisson Data with Change-Point

We first apply our Dirichlet process multiple change-point model to the much analyzed data set on the number of coal-mining disasters by year in Britain over the period 1951–1962 (Jarrett (1979), Carlin et al. (1992) and Chib (1998)).

Let the disaster count y be modeled by a Poisson distribution

$$f(y \mid \lambda) = \lambda^y e^{-\lambda}/y!. \tag{27}$$

The observation sequence $Y_n = (y_1, y_2, \ldots, y_{112})'$ is subject to some unknown change-points. We plot the data y_t in Figure 4. Chib (1998) estimates the models with one change-point ($k = 1$) and with two change-points ($k = 2$), respectively. He assumes the parameter λ following the prior Gamma(2, 1) in the one-change-point case and the prior Gamma(3, 1) in the other. Hence, given the regime indicators S_n, the corresponding parameter λ_i in regime i has the following posteriors with respect to the two priors:

$$\text{Posterior 1:} \quad \lambda_i \mid S_n, Y_n \sim \text{Gamma}(\lambda_i \mid 2 + U_i, 1 + N_i), \tag{28}$$

and

$$\text{Posterior 2:} \quad \lambda_i \mid S_n, Y_n \sim \text{Gamma}(\lambda_i \mid 3 + U_i, 1 + N_i), \tag{29}$$
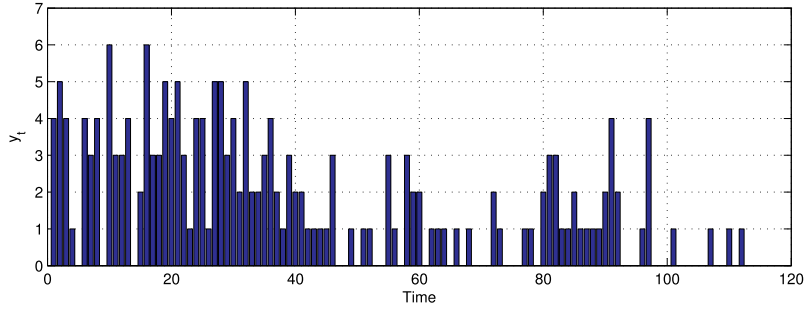
Figure 4: Data on coal mining disaster count y_t.

(a) Under prior 1.

(b) Under prior 2.
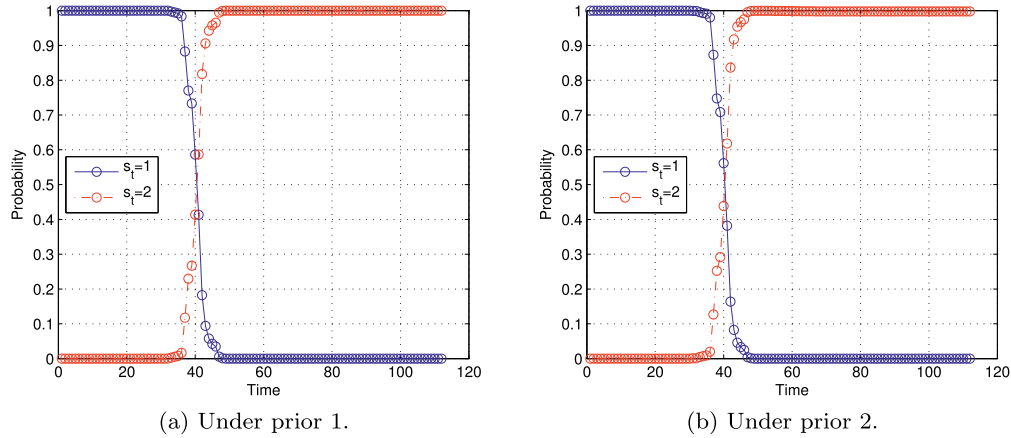
Figure 5: Posterior probability of $s_t = i$.

where $U_i = \sum_{t=1}^{112} \delta(s_t, i) y_t$ and $N_i = \sum_{t=1}^{112} \delta(s_t, i)$. We perform our algorithm with the following Gibbs steps:

Step 1. Sample $S_n \mid \lambda, Y_n$ as in (16) and obtain k,

Step 2. Sample $\lambda_i \mid S_n, Y_n$ as in (28) or (29),

Step 3. Update α and β with the Metropolis-Hastings Sampler as in Section 6.2.

The above Gibbs sampler is conducted for 5000 sweeps with 1000 burn-in samples. To reduce the sampler dependency, the 5000 sweeps are thinned by 50 draws. The sampler estimates one change-point in the data. Figure 5 shows the posterior probabilities of the regime indicator $s_t = i$ at each time point t. The intersections of the two lines $s_t = 1$ and $s_t = 2$ show that the break location exits at around $t = 40$. Figure 6 provides the
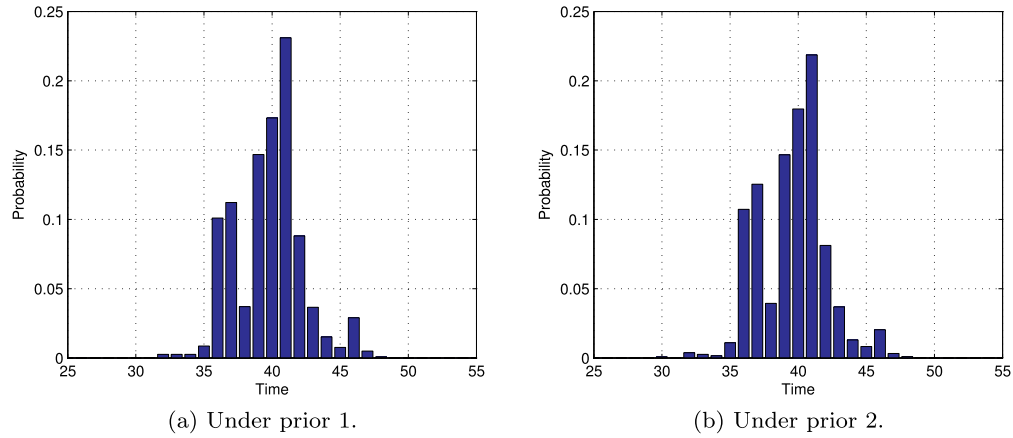
(a) Under prior 1.   (b) Under prior 2.

Figure 6: Posterior probability mass function of change-point location τ_i.

distribution of the transition points τ_i. Interestingly, our model produces exactly the same figure as the one in Chib (1998). The change-point is identified as occurring at around $t = 41$.

The corresponding posterior means of the parameters λ_1 and λ_2 are 3.1006 and 0.9387 with posterior standard deviations, 0.2833 and 0.1168, respectively, under the prior Gamma$(2, 1)$. The posterior means of α and β are 1.8101 and 0.3697 with standard deviations 1.3577 and 0.2464. When using the prior Gamma$(3, 1)$, we have the posterior means of λ_1 and λ_2 equal to3.1308 and 0.9567 with posterior standard deviations 0.2877 and 0.1218, respectively. The posterior means of α and β are 1.8375 and 0.3715 with standard deviations 1.3456 and 0.2360, respectively under prior 2. All our results closely match those of the literature and we show a certain robustness of our model under different prior assumptions.

In order to check the robustness of the estimation of the number of change-points, k, we conduct 1000 replications of the above estimation process and collect 1000 change-point numbers. When the first prior is assumed, 77.23% of the 1000 replications detect one change-point. We find a similar result for prior 2. Hence, we conclude that without assuming the number of change-points *a priori*, our algorithm detects the same change-point number as in the model developed by Chib (1998) with high probability.

7.2   Real Output

We also apply our algorithm to estimate structural changes in real Gross Domestic Product growth. The data and model are drawn from Maheu and Gordon (2008) (see also Geweke and Yu (2011)). Let $y_t = 100[\log(q_t/q_{t-1}) - \log(p_t/p_{t-1})]$, where q_t is quarterly US GDP seasonally adjusted and p_t is the GDP price index. The data range from the second quarter of 1947 to the third quarter of 2003, for a total of 226 observations (see Figure 7). We model the data with a Bayesian AR(2) model with structural change.
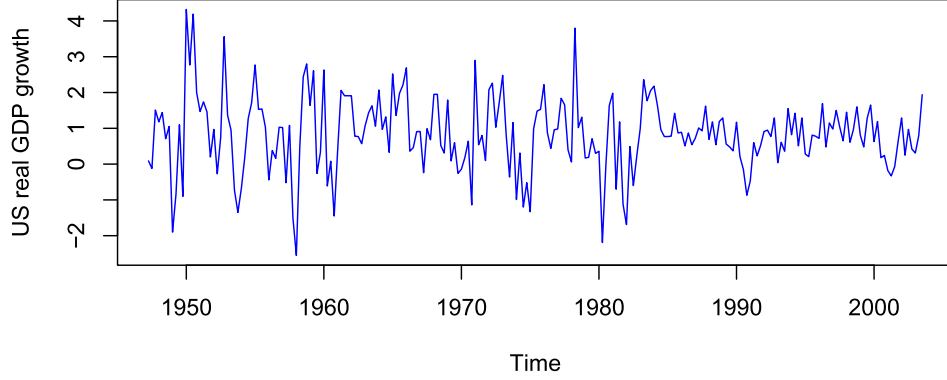
Figure 7: US real GDP growth from the second quarter of 1947 to the third quarter of 2003.

The frequentist autoregressive structural change-model can be found in Chong (2001). Suppose the data are subject to k change-points and follow

$$y_t = \beta_{0,s_t} + \beta_{1,s_t} y_{t-1} + \beta_{2,s_t} y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma_{s_t}^2), \quad s_t = 1, 2, \ldots, k+1. \quad (30)$$

We assume the following hierarchical priors to $\beta_{0,i}, \beta_{1,i}$ and $\beta_{2,i}$:

$$\beta_i = (\beta_{0,i}, \beta_{1,i}, \beta_{2,i})' \sim \mathrm{N}(\mu, V), \quad i = 1, \ldots, k+1, \quad (31)$$

where $\mu = (\mu_0, \mu_1, \mu_2)'$ and $V = \mathrm{Diag}(v_0^2, v_1^2, v_2^2)$, such that

$$p(\mu_j, v_j^2) \propto \mathrm{Inv\text{-}Gamma}(v_j^2 \mid a, b), \quad j = 0, 1, 2. \quad (32)$$

We assume the noninformative prior for σ_i^2 such that

$$p(\sigma_i^2) \propto 1/\sigma_i^2, \quad i = 1, \ldots, k+1. \quad (33)$$

Conditional on σ_i^2, the sampling of β_i, μ and V is similar to Section 5. For σ_i^2, we can draw from the following full conditional:

$$\sigma_i^2 \mid \beta_i, S_n, Y_n \sim \mathrm{Inv\text{-}}\chi^2 \left(\sigma_i^2 \mid \tau_i - \tau_{i-1}, \frac{\omega_i^2}{\tau_i - \tau_{i-1}} \right), \quad i = 1, \ldots, k+1, \quad (34)$$

where $\omega_i^2 = \sum_{\tau_{i-1} < t \le \tau_i} (y_t - \beta_{0,s_t} - \beta_{1,s_t} y_{t-1} - \beta_{2,s_t} y_{t-2})^2$.

As in the previous applications, we set the inverse-Gamma hyperparameters $a = b = 1$. The M-H update of α and β follows the discussion in Section 6.2. The Gibbs sampler is conducted for 5000 sweeps with 1000 burn-in samples. The 5000 sweeps are thinned by 50 draws. The posterior probabilities of the regime indicator s_t in Figure 8a suggest that the structural break exists between the years 1980 and 1990. Figure 8b

(a) Posterior probability of $s_t = i$.

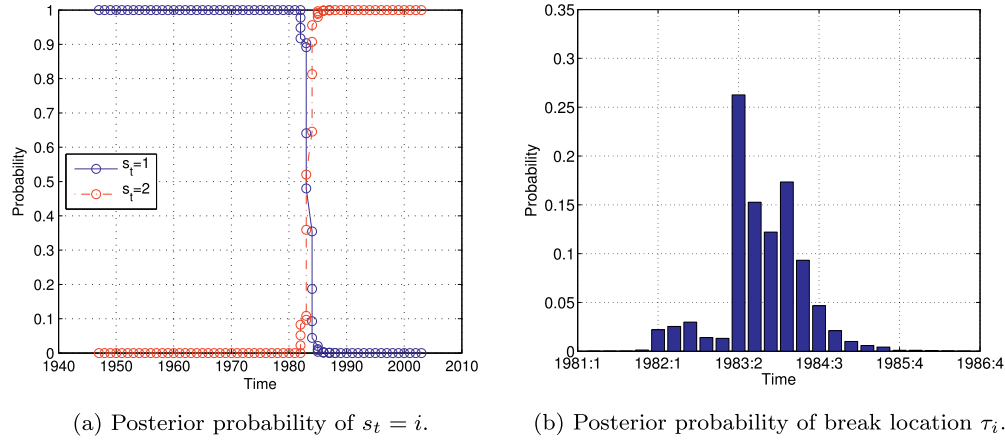(b) Posterior probability of break location τ_i.

Figure 8: US real GDP growth structural change model.

Table 7: US real GDP growth structural change model with one change-point. Posterior means and posterior standard deviations within parentheses. The results applying Chib's model are drawn from Maheu and Gordon (2008).

	Chib's Model		DPHMM	
	$s_t = 1$	$s_t = 2$	$s_t = 1$	$s_t = 2$
β_{0,s_t}	0.5642 (0.1228)	0.4434 (0.1162)	0.5499 (0.1303)	0.3894 (0.1169)
β_{1,s_t}	0.2716 (0.0734)	0.2792 (0.1052)	0.2812 (0.0837)	0.2796 (0.1173)
β_{2,s_t}	0.0800 (0.0739)	0.1588 (0.1010)	0.0913 (0.0855)	0.2253 (0.1124)
$\sigma^2_{s_t}$	1.3331 (0.1542)	0.3362 (0.0516)	1.4089 (0.1722)	0.2672 (0.0460)

further shows the change-point at the second quarter of 1983, which is close to the results in Maheu and Gordon (2008).[4] The posterior estimates are summarized in Table 7. Finally, the posterior means of α and β are 1.7749 and 0.3045 with standard deviations 1.3422 and 0.1939 respectively. All of our results are consistent with Chib's estimates.

Finally, we replicate 1000 times the whole estimation process and check the robustness of the detected change-point number. The result suggests that nearly 100% of the replications detect one break point.

8   Concluding Remarks

In this paper, we have proposed a new Bayesian multiple change-point model, that is the Dirichlet process hidden Markov model. Our model is semiparametric in the sense that the number of states is not built-in to the model but endogenously determined. As a result, our model avoids the model misspecification problem. We have proposed an MCMC sampler which only needs to sample the states around change-points. We

---

[4]See Figure 4 in Maheu and Gordon (2008).

have also proposed the MAP and M-H updates of hyperparameters in the DPHMM process. We have presented three specific models, namely, the discrete Poisson model, the continuous normal model, and the AR(2) model with structural changes. Results from the simulations and empirical applications showed that our Dirichlet process hidden Markov multiple change-point model detected the true change-point numbers and locations with high accuracy.

Appendix

In the appendix, we give the derivations of the full conditionals and the Gibbs samplers in Section 5. For the case of known variance, we first rewrite the hierarchical model (21) as the joint distribution

$$p(Y_n, \theta, \mu, \upsilon^2 \mid S_n, \sigma^2) \propto \prod_{i=1}^{k+1} \mathrm{N}(\tilde{y}_i \mid \theta_i, \sigma_i^2) \prod_{i=1}^{k+1} \mathrm{N}(\theta_i \mid \mu, \upsilon^2) p(\mu, \upsilon^2), \qquad \text{(A-1)}$$

where $p(\mu, \upsilon^2)$ corresponds to Inv-Gamma$(\upsilon^2 \mid a, b)$, $\theta = (\theta_1, \ldots, \theta_{k+1})'$ and

$$\tilde{y}_i = \frac{\sum_{\tau_{i-1} < t \le \tau_i} y_t}{\tau_i - \tau_{i-1}}, \quad \text{and} \quad \sigma_i^2 = \frac{\sigma^2}{\tau_i - \tau_{i-1}}. \qquad \text{(A-2)}$$

From (A-1) and (A-2), we have the following full conditionals:

$$p(\theta_i \mid \mu, \upsilon^2, \sigma^2, S_n, Y_n) \propto \prod_{i=1}^{k+1} \mathrm{N}(\tilde{y}_i \mid \theta_i, \sigma_i^2) \mathrm{N}(\theta_i \mid \mu, \upsilon^2)$$

$$\propto \mathrm{N}\left(\theta_i \mid \frac{\tilde{y}_i/\sigma_i^2 + \mu/\upsilon^2}{1/\sigma_i^2 + 1/\upsilon^2}, \frac{1}{1/\sigma_i^2 + 1/\upsilon^2}\right),$$

$$p(\mu \mid \theta, \upsilon^2, S_n, Y_n) \propto \prod_{i=1}^{k+1} \mathrm{N}(\theta_i \mid \mu, \upsilon^2) p(\mu, \upsilon^2) \qquad \text{(A-3)}$$

$$\propto \mathrm{N}(\mu \mid \bar{\theta}, \upsilon^2/(k+1)),$$

$$p(\upsilon^2 \mid \theta, \mu, S_n, Y_n) \propto (\upsilon^2)^{-(k+1)/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^{k+1} (\theta_i - \mu)^2/\upsilon^2\right\} p(\mu, \upsilon^2)$$

$$\propto \text{Inv-Gamma}\left(\upsilon^2 \mid a + \frac{k+1}{2}, b + \frac{1}{2} \sum_{i=1}^{k+1} (\theta_i - \mu)^2\right),$$

where $\bar{\theta} = \sum_{i=1}^{k+1} \theta_i/(k+1)$. Therefore, we can perform the following Gibbs sampler:

Step 1. Sample $S_n \mid \theta, \mu, \upsilon, Y_n$ as in (16) and obtain k,

Step 2. Sample $\theta, \mu, \upsilon \mid S_n, Y_n$ as in (A-3).

For the case of unknown variance, the full conditional with respect to (22) is

$$\sigma^2 \mid \theta, S_n, Y_n \sim \text{Inv-Gamma}\left(\sigma^2 \middle| c + \frac{n}{2}, d + \frac{1}{2} \sum_{t=1}^{n} (y_t - \theta_{s_t})^2\right). \qquad \text{(A-4)}$$

Conditional on σ^2, we apply the same estimation strategy discussed above. The Gibbs sampler is thus

Step 1. Sample $S_n \mid \theta, \mu, \upsilon, \sigma^2, Y_n$ as in (16) and obtain k,

Step 2. Sample $\theta, \mu, \upsilon, \sigma^2 \mid S_n, Y_n$ as in (A-3) and (A-4).

References

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). "The Infinite Hidden Markov Model." In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems*, 577–584. MIT Press. 277, 279, 287

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Polya Urn Schemes." *The Annals of Statistics*, 1(2): 353–355. 277

Carlin, P. B., Gelfand, A. E., and Smith, A. F. M. (1992). "Hierarchical Bayesian Analysis of Changepoint Problems." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2): 389–405. 275, 289

Chernoff, H. and Zacks, S. (1964). "Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time." *The Annals of Mathematical Statistics*, 35(3): pp. 999–1018. 275

Chib, S. (1998). "Estimation and comparison of multiple change-point models." *Journal of Econometrics*, 86(2): 221 – 241. 275, 276, 279, 282, 284, 289, 291

Chong, T. T.-L. (2001). "Structural Change in AR(1) Models." *Econometric Theory*, 17(1): 87–155. 292

Connor, R. J. and Mosimann, J. E. (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution." *Journal of the American Statistical Association*, 64(325): pp. 194–206. 279, 287

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230. 277

Geweke, J. and Yu, J. (2011). "Inference and prediction in a multiple-structural-break model." *Journal of Econometrics*, 163(2): 172–185. 275, 291

Giordani, P. and Kohn, R. (2008). "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models." *Journal of Business and Economic Statistics*, 26(1): 66–77. 275

Jarrett, R. G. (1979). "A Note on the Intervals Between Coal-Mining Disasters." *Biometrika*, 66(1): 191–193. 289

Koop, G. and Potter, S. M. (2007). "Estimation and Forecasting in Models with Multiple Breaks." *The Review of Economic Studies*, 74(3): pp. 763–789. 275

Kozumi, H. and Hasegawa, H. (2000). "A Bayesian analysis of structural changes with an application to the displacement effect." *The Manchester School*, 68(4): 476–490. 277

Maheu, J. M. and Gordon, S. (2008). "Learning, forecasting and structural breaks." *Journal of Applied Econometrics*, 23(5): 553–583. 275, 291, 293

Neal, R. M. (1992). "The Infinite Hidden Markov Model." In Smith, C. R., Erickson, G. J., and Neudorfer, P. O. (eds.), *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, 197–211. Kluwer Academic Publishers. 277

— (2000). "Markov Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. 277, 279

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). "Forecasting Time Series Subject to Multiple Structural Breaks." *Review of Economic Studies*, 73(4): 1057–1084. 275

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4(2): 639–650. 277

Smith, A. F. M. (1975). "A Bayesian approach to inference about a change-point in a sequence of random variables." *Biometrika*, 62: 407–416. 275

Stephens, D. A. (1994). "Bayesian Retrospective Multiple-Changepoint Identification." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1): 159–178. 275

Wang, J. and Zivot, E. (2000). "A Bayesian Time Series Model of Multiple Structural Changes in Level, Trend, and Variance." *Journal of Business & Economic Statistics*, 18(3): 374–386. 275

Wong, T. (1998). "Generalized Dirichlet distribution in Bayesian analysis." *Applied Mathematics and Computation*, 97(2-3): 165–181. 279, 287