Communications for Statistical Applications and Methods
2013, Vol. 20, No. 6, 439–454

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.6.439

Bayesian Multiple Change-Point Estimation and Segmentation

Jaehee Kim[a], Sooyoung Cheon[1,b]

[a]Department of Statistics, Duksung Women's University
[b]Department of Informational Statistics, Korea University

Abstract

This study presents a Bayesian multiple change-point detection approach to segment and classify the observations that no longer come from an initial population after a certain time. Inferences are based on the multiple change-points in a sequence of random variables where the probability distribution changes. Bayesian multiple change-point estimation is classifies each observation into a segment. We use a truncated Poisson distribution for the number of change-points and conjugate prior for the exponential family distributions. The Bayesian method can lead the unsupervised classification of discrete, continuous variables and multivariate vectors based on latent class models; therefore, the solution for change-points corresponds to the stochastic partitions of observed data. We demonstrate segmentation with real data.

Keywords: BIC, multiple change-points, segmentation, stochastic approximation Monte Carlo.

1. Introduction

Significant attention has been focused on stochastic models for systems in which external forces cause an abrupt change in the stochastic properties at some time in the history of a system. Such models seem to derive from quality control situations where the emphasis has been to detect a change in measurable quality. An unsupervised classification of observed data may be understood as a scientific learning task where a putative hidden structure that underlies the observations remains to be discovered. Heuristic numerical methods for classification have been widely exploited for decades; however, there has been increased interest in model-based approaches as computational restrictions have become less significant. Corander *et al.* (2009) considered a Bayesian model-based method for the unsupervised classification of discrete valued vectors.

Detection of change-points and segmentation of observations are of interest in a wide range of applications. We are interested in event classifications in terms of distributions. After a Bayesian multiple change-point detection, the classification procedure can be performed with estimated change-points. Segmentation can be achieved by detecting change-points simultaneously with change-point estimation; consequently, parameters corresponding to this new segment can be estimated when a new change is detected. Finding the number of segments with these change-points is a challenging statistical problem since the dimension of the object of inference is not fixed. Segmentation or classification can be extended to a large data set. Wehrens *et al*. (2004) suggested rapid clustering for large data sets such as MRI imaging data. Segmentation procedures are used increasingly in many fields since heterogeneity should be incorporated into data analysis.

[1] Corresponding author: Assistant Professor, Department of Informational Statistics, Korea University, 2511 Sejong-ro, Sejong 339-700, Korea. E-mail: scheon@korea.ac.kr

Chernoff and Zacks (1964) considered a Bayes test to determine the mean change in normal observations. Hinkley (1970, 1972) investigated the maximum likelihood estimates of a one change-point problem. Smith (1975) considered one change-point problem in terms of distributional changes using the Gibbs sampler as a Bayesian approach for the change-point problem. Yao (1984) derived Bayes estimates in the presence of additive Gaussian noise and a signal that is a step function. Carlin *et al.* (1992) formulated the hierarchical Bayesian Markov chain model and used the Gibbs sampler for inference. Belisle *et al.* (1998) made inferences about the Bayesian hierarchical change-point model with an ensemble of sample paths for neuron spike train data. In the multiple change-point setting, Venter and Steel (1996) identified multiple abrupt change-points in a sequence of observations via hypothesis testing. Hawkins (2001) developed an approach with maximum likelihood estimates of the change-points and within-segment parameters in the exponential family.

For a Bayesian multiple change-point problem, Barry and Hartigan (1993) used the product partition model. Chib (1998) formulated the multiple change-point model in terms of a latent discrete state variable according to the Markov process with transition probabilities. Stephens (1994) discussed the use of a sampling-based technique and the Gibbs sampler that included the binomial data model. Fearnhead (2006) suggested a recursion algorithm to identify the change-points successively. Bayesian methods are attractive for change-point models since they allow for flexible relationships between parameters in various subspaces.

We isolate zones in which the distribution is homogeneous. In this context the observations can be studied with parametric models for which the parameters are supposed to be affected by abrupt changes at unknown instants. The purpose of this statistical study is to detect changes in these parameters. In the global segmentation context we require delimiting segments for which the characteristics of the signal are homogeneous within segments and are uniquely different for each segment. We assume that distributional parameters are affected by an unknown number of abrupt changes at unknown sequential points. Once the model has been specified, the problem is to estimate the location of the change-points and the parameters within segments. A Bayesian approach considered the number of breakpoints and their locations are viewed as random variables. The objective is to estimate posterior distributions with the stochastic approximation Monte Carlo algorithm(SAMC; Liang *et al.*, 2007). Once the model has been specified and the location of the breakpoints can be estimated with an appropriate method, the problem then becomes how to determine the number of segments into which the data should be partitioned. In the context of model selection, the Bayesian information criterion(BIC) provides a consistent estimator of the number of change-points. Based on posterior probabilities, the maximum a posteriori(MAP) estimator of the change-point sequence can suffer from the local maxima searching problem and SAMC can find the global maxima not trapped in the local maxima.

In this paper, we apply the Bayesian multiple change-point analysis to segmentation. Detecting change-points partitions observations into segments with different distributional parameters. In Section 2, a general multiple change-point model is defined and the Bayesian inference is reviewed for some exponential family distributions. Section 3 briefly describes the SAMC algorithm applied to the multiple change-point problem. Section 4 presents some numerical results with real data for multiple change-point estimation and segmentation. Finally, Section 5 concludes the paper with a discussion.

2. The Bayesian Multiple Change-Point Model

Change-point identification is important in data analysis since it separates observations into segments. To capture heterogeneity in the data, it is reasonable to assume that latent classes are present in the sense that each class can be characterized by a different data generating mechanism. The data are

characterized by the same distributional parameters within the same class. Interest lies in making inference about the time or position in the sequence where the change occurs and where the new segment begins. This problem can be generalized to incorporate notions of multiple changes in the system. Problems arise when different subsequences of a data series follow different statistical distributions of the same functional form, but have different parameters.

Let $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$ denote the independent observation sequence ordered in time. The set $\{1, 2, \ldots, n\}$ is partitioned into blocks such that the sequence follows the same distribution within blocks. That is, the change-points divide the partitions. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a binary vector with $x_{c_1} = x_{c_2} = \cdots = x_{c_k} = 1$ and 0 elsewhere, where $0 = c_0 < c_1 < c_2 < \cdots < c_{k+1} = n$. There are k change-points in the model where k is unknown. The multiple change-point model can be written as:

$$z_i \sim f_r(\cdot\,|\boldsymbol{\phi}_r), \quad c_{r-1} < i \leq c_r, \tag{2.1}$$

for $r = 1, 2, \ldots, k+1$ and f_r depends on the parameters $\boldsymbol{\phi}_r \in \Phi$. The parameters change at $c_1 + 1, c_2 + 1, \ldots, c_k + 1$. Each $c_1, c_2, \ldots, c_k$ is called the change-point.

Consider that f_r is a probability density parameterized by $\boldsymbol{\phi}_r$. Let $\mathbf{x}^{(k)}$ denote a configuration of $\mathbf{x}$ with k change-points. Let $\boldsymbol{\eta}^{(k)} = (\mathbf{x}^{(k)}, k, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k+1})$, and A_k be the space of models with k change-points, $\mathbf{x}^{(k)} \in A_k$ and $\mathcal{X} = \cup_{k=0}^{n} A_k$. The likelihood function of $\mathbf{Z}$ is

$$L\left(\mathbf{Z}|\boldsymbol{\eta}^{(k)}\right) = \prod_{j=c_0+1}^{c_1} f_1\left(z_j|\boldsymbol{\phi}_1\right) \times \cdots \times \prod_{j=c_k+1}^{c_{k+1}} f_{k+1}\left(z_j|\boldsymbol{\phi}_{k+1}\right).$$

We set the prior distribution for k in $\boldsymbol{\eta}^{(k)}$ as the truncated Poisson distribution,

$$P(k) = \frac{\lambda^k}{\sum_{j=0}^{n-1} \frac{\lambda^j}{j!}} \frac{(n-1-k)!}{(n-1)!}, \quad k = 0, 1, \ldots, n-1, \tag{2.2}$$

which is equivalent to assuming that A_k is subject to a truncated Poisson distribution with parameter λ and each of the $(n-1)!/[k!(n-1-k)!]$ models in A_k is a priori equally. In addition, the uniform prior is used for $\mathbf{x}^{(k)}$ given k.

The posterior density $P(\mathbf{x}^{(k)}, k|\mathbf{Z})$ is obtained by integrating over the parameter space such as

$$P\left(\mathbf{x}^{(k)}, k|\mathbf{Z}\right) \propto \int P\left(\mathbf{Z}|\mathbf{x}^{(k)}, k, \boldsymbol{\phi}\right) P\left(\boldsymbol{\phi}|\mathbf{x}^{(k)}, k\right) P\left(\mathbf{x}^{(k)}|k\right) P(k) d\boldsymbol{\phi}, \tag{2.3}$$

where $\boldsymbol{\phi}$ is the parameter vector of the distribution with its prior, $P(\mathbf{Z}|\mathbf{x}^{(k)}, k, \boldsymbol{\phi})$ is the probability density of $\mathbf{Z}$ given the value of $\mathbf{x}^{(k)}$, k and $\boldsymbol{\phi}$, $P(\boldsymbol{\phi}|\mathbf{x}^{(k)}, k)$ is the prior distribution of the parameter $\boldsymbol{\phi}$, $P(\mathbf{x}^{(k)}|k)$ is the prior distribution of configuration of the number of change-points, k, and $P(k)$ is the prior distribution of the number of change-points, k (Gelman *et al.*, 2003; Gelfand and Smith, 1990).

Kim and Cheon (2010) and Cheon and Kim (2010) provide the derivation of the full unnormalized posterior for the normal, exponential, binomial, Poisson and multivariate normal distributions for multiple change-points identified as:

(a) Normal

Let f_r be a normal density parameterized by $\boldsymbol{\phi}_r = (\mu_r, {\sigma_r}^2)$. For a uniform prior on μ_r and an

inverse-gamma prior, $IG(\gamma, \delta)$, on ${\sigma_r}^2$, which are independent of each other,

$$\log P\left(\mathbf{x}^{(k)}, k|\mathbf{Z}\right) \propto a_k + \frac{k+1}{2}\log(2\pi) - \sum_{i=1}^{k+1}\left\{\frac{1}{2}\log(c_i - c_{i-1}) - \log\Gamma\left(\frac{c_i - c_{i-1} - 1}{2} + \gamma\right)\right. \tag{2.4}$$
$$\left. + \left(\frac{c_i - c_{i-1} - 1}{2} + \gamma\right)\log\left[\delta + \frac{1}{2}\sum_{j=c_{i-1}+1}^{c_i} z_j^2 - \frac{\left(\sum_{j=c_{i-1}+1}^{c_i} z_j\right)^2}{2(c_i - c_{i-1})}\right]\right\},$$

where $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$.

(b) Exponential

Let f_r be an exponential density parameterized by $\boldsymbol{\phi}_r = \sigma_r$. For a gamma prior, $G(\gamma, \delta)$, on σ_r,

$$\log P\left(\mathbf{x}^{(k)}, k|\mathbf{Z}\right) \propto a_k + \sum_{i=1}^{k+1}\left\{\log\Gamma(c_i - c_{i-1} + \gamma) - (c_i - c_{i-1} + \gamma)\log\left(\delta + \sum_{j=c_{i-1}+1}^{c_i} z_j\right)\right\},$$

where $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$.

(c) Binomial

Let f_r be a binomial density, BIN(b, p_r), parameterized by $\boldsymbol{\phi}_r = p_r$. For a beta prior, Beta$(\alpha, \beta)$, on p_r,

$$\log P\left(\mathbf{x}^{(k)}, k|\mathbf{Z}\right) \propto a_k + \sum_{i=1}^{k+1}\left\{\sum_{j=c_{i-1}+1}^{c_i}\left[\log b! - \log(b - z_j)! - \log z_j!\right]\right. \tag{2.5}$$
$$+ \log\Gamma\left(\alpha + \sum_{j=c_{i-1}+1}^{c_i} z_j\right) + \log\Gamma\left(n(c_i - c_{i-1}) + \beta - \sum_{j=c_{i-1}+1}^{c_i} z_j\right)$$
$$\left. - \log\Gamma(\alpha + b(c_i - c_{i-1}) + \beta)\right\},$$

where $a_k = (k+1)(\log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)) + \log(n-1-k)! + k\log\lambda$.

(d) Poisson

Let f_r be a Poisson density parameterized by $\boldsymbol{\phi}_r = \phi_r$. For a gamma prior, $G(\gamma, \delta)$, on ϕ_r,

$$\log P\left(\mathbf{x}^{(k)}, k|\mathbf{Z}\right) \propto a_k + \sum_{i=1}^{k+1}\left\{\log\Gamma\left(\gamma + \sum_{j=c_{i-1}+1}^{c_i} z_j\right)\right. \tag{2.6}$$
$$\left. - \left(\sum_{j=c_{i-1}+1}^{c_i} z_j + \gamma\right)\log(c_i - c_{i-1} + \delta) - \sum_{j=c_{i-1}+1}^{c_i}\log z_j!\right\},$$

where $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$.

(e) Multivariate Normal

Let f_r be a d-dimensional multivariate normal density parameterized by $\boldsymbol{\phi}_r = (\mu_r, \Sigma_r)$ with d-dimensional mean vector $\boldsymbol{\mu}$ and $d\times d$ covariance matrix Σ. For a vague prior and an inverse-Wishart

prior $IW(\nu_0, \mathbf{\Lambda}_0)$, which are independent of each other, placed on the μ_r's and Σ_r's, respectively,

$$\log P\left(\mathbf{x}^{(k)}, k|\mathbf{Z}\right) \propto c_k - \frac{d(k+1)}{2}\log(2\pi) + \sum_{i=1}^{k+1}\log\left[\frac{2^{\frac{\nu_i d}{2}}\pi^{\frac{d(d-1)}{4}}}{(c_i - c_{i-1})^{\frac{1}{2}}}|\mathbf{\Lambda}_0 + (c_i - c_{i-1})\mathbf{S}_i|^{-\frac{\nu_i}{2}}\prod_{u=1}^{d}\Gamma\left(\frac{\nu_i + 1 - u}{2}\right)\right], \tag{2.7}$$

where

$$c_k = -(k+1)\left[\frac{\nu_0 d}{2}\log 2 + \frac{d(d-1)}{4}\log\pi + \sum_{u=1}^{d}\log\Gamma\left(\frac{\nu_0 + 1 - u}{2}\right) - \frac{\nu_0}{2}\log|\mathbf{\Lambda}_0|\right]$$

with $\nu_i = c_i - c_{i-1} + \nu_0 - 1$, $\mathbf{S}_i = 1/(c_i - c_{i-1})\sum_{j=c_{i-1}+1}^{c_i}(\mathbf{z}_j - \bar{\mathbf{z}}^i)(\mathbf{z}_j - \bar{\mathbf{z}}^i)'$, $\mathbf{z}_j = (z_{j1}, \ldots, z_{jd})'$ for $j = c_{i-1} + 1, \ldots, c_i$, and $\bar{\mathbf{z}}^i = ((\sum_{j=c_{i-1}+1}^{c_i} z_{j1})/(c_i - c_{i-1}), \ldots, (\sum_{j=c_{i-1}+1}^{c_i} z_{jd})/(c_i - c_{i-1}))'$.

We can sample from this non-normalized posterior using the SAMC technique with the partitioned sample space according to the negative posterior log-likelihood function and estimate the change-points with the greatest posterior probabilities. The BIC is often used in Bayesian model selection, discussed in Kass and Raftery (1995). The model with the highest posterior probability is the one that minimizes BIC = −2(log maximized likelihood) + log n(number of parameters). BIC is used because it penalizes more severely for the parameters and the posterior comparison is considered in change-point estimation. BIC favors simpler models and gives a rough approximation of the logarithm of the Bayes factor, which is easy to use and does not require an evaluation of the prior distributions (Raftery, 1995).

3. Multiple Change-Point Estimation and Classification

3.1. A review of the SAMC algorithm

A multiple change-point problem naturally requires the computational complexity for multi-way splitting of data. Since the number of change-points forms a variable parameter space, we use an advanced Monte Carlo algorithm, SAMC (Liang *et al*., 2007). The basic idea of SAMC is explained briefly as follows.

Suppose that we are working on inference for a posterior distribution, $f(\mathbf{x}) = (1/B)\psi(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, where B is the normalizing constant, and $\mathcal{X}$ is the sample space. In the context of Bayesian multiple change-point analysis, $\psi(\cdot)$ corresponds to the unnormalized posterior density $P(\mathbf{x}^{(k)}, k|\mathbf{Z})$ (see Section 2) and $\mathcal{X}$ corresponds to the sample space of $\mathbf{x}^{(k)}$. Suppose that the sample space has been partitioned according to the model index, *e.g.*, $E_1 = \{\mathbf{x}^{(k)} : k = 1\}$, $E_2 = \{\mathbf{x}^{(k)} : k = 2\}$, ... Without loss of generality, we consider only models with $k_{min} \le k \le k_{max}$, where k is the number of change-points, and k_{min} and k_{max} can be determined after a pilot study of the above algorithm, respectively. Outside this range, $P(\mathbf{x}^{(k)}, k|\mathbf{Z}) \approx 0$. The SAMC algorithm samples from the non-normalized negative posterior distribution, $-\log\psi(\mathbf{x})$, with the above partitioned sample space and estimates the change-points with the greatest posterior probability.

SAMC seeks to draw samples from each of the subregions with a pre-specified frequency. Let $\mathbf{x}^{(t+1)}$ denote a sample drawn from a Metropolis-Hastings (MH; Metropolis *et al.*, 1953; Hastings, 1970) kernel $K_{\theta^{(t)}}(\mathbf{x}^{(t)}, \cdot)$ with the proposal distribution $q(\mathbf{x}^{(t)}, \cdot)$ and the stationary distribution

$$f_{\theta^{(t)}}(\mathbf{x}) \propto \sum_{i=1}^{m-1}\frac{\psi(\mathbf{x})}{e^{\theta_i^{(t)}}}I(\mathbf{x} \in E_i) + \psi(\mathbf{x})I(x \in E_m), \tag{3.1}$$

where $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{m-1}^{(t)})$ is an $(m-1)$-vector in a space Θ. For convenience, we set $\theta_m^{(t)} = 0$. Here, without loss of generality, we assume that E_m is non-empty; that is, $\int_{E_m} \psi(\mathbf{x})d\mathbf{x} > 0$. In practice, E_m can be replaced by any subregion that is known to be nonempty.

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ be an m-vector with $0 < \pi_i < 1$ and $\sum_{i=1}^m \pi_i = 1$, which defines desired sampling frequencies for the subregions. Henceforth, $\boldsymbol{\pi}$ is called the desired sampling distribution. Define $H(\theta^{(t)}, \mathbf{x}^{(t+1)}) = (\boldsymbol{e}^{(t+1)} - \boldsymbol{\pi})$, where $\boldsymbol{e}^{(t+1)} = (e_1^{(t+1)}, \ldots, e_m^{(t+1)})$ and $e_i^{(t+1)} = 1$ if $\mathbf{x}^{(t+1)} \in E_i$ and 0 otherwise. Let $\{\gamma_t\}$ be a positive, non-decreasing sequence satisfying the conditions,

$$\text{(i)} \quad \sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \text{(ii)} \quad \sum_{t=0}^{\infty} \gamma_t^{\delta} < \infty, \tag{3.2}$$

for some $\delta \in (1, 2)$. In the context of stochastic approximation (Robbins and Monro, 1951), $\{\gamma_t\}_{t\geq 0}$ is called the gain factor sequence.

Let $J(\boldsymbol{x})$ denote the index of the subregion that the sample $\boldsymbol{x}$ belongs to, which takes values in $\{1, \ldots, m\}$. With the above notations, one iteration of SAMC can be described as:

The SAMC algorithm:

(a) (Sampling) Simulate a sample $\boldsymbol{x}^{(t+1)}$ by a single MH update with the target distribution as defined in (3.1).

(a.1) Generate $\boldsymbol{y}$ according to a proposal distribution $q(\boldsymbol{x}^{(t)}, \boldsymbol{y})$.

(a.2) Calculate the ratio

$$r = e^{\theta_{J(\boldsymbol{x}^{(t)})}^{(t)} - \theta_{J(\boldsymbol{y})}^{(t)}} \frac{\psi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{x}^{(t)})}{\psi(\boldsymbol{x}^{(t)})q(\boldsymbol{x}^{(t)}, \boldsymbol{y})}. \tag{3.3}$$

(a.3) Accept the proposal with probability $\min(1, r)$. If it is accepted, set $\boldsymbol{x}^{(t+1)} = \boldsymbol{y}$; otherwise, set $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)}$.

(b) ($\boldsymbol{\theta}$-updating) Set

$$\theta^{\left(t+\frac{1}{2}\right)} = \theta^{(t)} + \gamma_{t+1} H\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{x}^{(t+1)}\right). \tag{3.4}$$

If $\boldsymbol{\theta}^{(t+1/2)} \in \Theta$, set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t+1/2)}$; otherwise, set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t+1/2)} + c\mathbf{1}_m$, where c is a value satisfying that $\boldsymbol{\theta}^{(t+1/2)} + c\mathbf{1}_m \in \Theta$ and $\mathbf{1}_m$ denotes a constant m-vector of ones.

The self-adjusting mechanism of the SAMC algorithm is obvious. If a proposal is rejected, the weight of the subregion to which the current sample belongs will be adjusted to a larger value and the proposal of jumping out from the current subregion will be less likely rejected in the next iteration. This mechanism guarantees that the algorithm will not be trapped by local energy minima. The SAMC algorithm represents a significant advance in simulations of complex systems for which the energy landscape is rugged.

3.2. Application of SAMC to Bayesian change-point detection problems

The SAMC algorithm was implemented to the Bayesian multiple change-point model to solve a Bayesian model selection problem. For a change-point detection, the maximum *a posteriori*(MAP)

estimate of $\mathbf{x}^{(k)}$ is often a reasonable solution to the problem. The SAMC sampling method works recursively. First, the number of change-points is generated according to the proposal, and then the positions of selected change-points are generated randomly conditioning on the data Z. The sampling step of SAMC for the MAP estimate is as follows.

Let $k^{(t)}$ and $\mathbf{x}^{(k^{(t)},t)}$ denote the number of change-points and the configuration of $\mathbf{x}$ with $k^{(t)}$ change-points sampled at iteration t, respectively. The next sample can be generated from the following procedure:

(a) Generate $k^* = k-1, k, k+1$ according to the proposal matrix $Q = (Q_{k,k^*})$, where $Q_{k,k} = 1/3$ for $k_{min} \le k \le k_{max}$, $Q_{k_{min},k_{min}+1} = Q_{k_{max},k_{max}-1} = 2/3$ and $Q_{k,k+1} = Q_{k,k-1} = 1/3$ if $k_{min} < k < k_{max}$.

(b) Generate $\mathbf{x}^{(k^*,*)}$ by updating with a "death", "simultaneous" or "birth" move if $k^* = k-1, k, k+1$, respectively.

(c) Accept a new sample $(k^*, \mathbf{x}^{(k^*,*)})$ by the SAMC algorithm: *i.e*, accept a sample $(k^*, \mathbf{x}^{(k^*,*)})$ with probability

$$\min\left\{1, \frac{e^{\theta^{t,k^{(t)}}}}{e^{\theta^{t,k^*}}} \frac{f\left(\mathbf{x}^{(k^*,*)}|Z\right)}{f\left(\mathbf{x}^{(k^{(t)},t)}|Z\right)} \frac{Q\left(k^* \to k^{(t)}\right)}{Q\left(k^{(t)} \to k^*\right)}\right\}. \tag{3.5}$$

If this is accepted, then set $(k^{(t+1)}, \mathbf{x}^{(k^{(t+1)},t+1)}) = (k^*, \mathbf{x}^{(k^*,*)})$; otherwise, set $(k^{(t+1)}, \mathbf{x}^{(k^{(t+1)},t+1)}) = (k^{(t)}, \mathbf{x}^{(k^{(t)},t)})$.

The "death", "simultaneous", and "birth" moves are designed as described in Green (1995).

For effective implementation of SAMC in the change-point detection, the gain factor sequence is set to be $\tau_t = (T_0/\max(T_0, t))^{0.6}$, $t = 0, 1, 2, \ldots$, for pre-specified values of $T_0 > 1$. A practical guideline for the choice of T_0 is to examine the flatness of the histogram of the samples drawn at different subregions. A histogram is flat if the sampling frequency at each subregion is greater than 80% the average sampling frequency for all subregions. If the histogram is not flat, SAMC should be re-run with a larger value of T_0, a larger number of iterations (or both). Refer to Liang *et al*. (2007) for the choice of T_0.

4. Real Data Applications

We provide applications of our method for multiple change-point problems and automatic classification. Each segment is homogeneous in distributional parameters after the change-point detection. We performed an independence test of observations in each partition using a Bartels test (Bartels, 1982) for univariate data and a test developed by Kojadinovic and Holmes (2009) for multivariate data. The methods have been implemented in the R package *lawstat* and *copula*, respectively.

4.1. Well-log data for normal change-point model

We consider the problem to detect change-points with the well-log data in Ó Ruanaidh and Fitzgerald (1996). The data are 4050 well drilling measurements from the nuclear-magnetic responses of underground rocks. The data are used to interpret the geophysical structure of the rock surrounding the well, since the variations in mean reflect the stratification of the earth's crust. The underlying signal is roughly piecewise constant and each constant segment relates to a single rock type that has constant physical properties. The change-points in the signal occur when a new rock type is encountered;

Table 1: Ten models with the largest log-posterior values and smallest BIC in well-log data.

No	♯ of change	Change patterns	Log-post	BIC
1	19	(26, 1034, 1070, 1210, 1220, 1420, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963)	−5659.1	11484.4
2	20	(26, 1034, 1070, 1210, 1220, 1420, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3739, 3942, 3963)	−5664.0	11502.4
3	19	(26, 1041, 1070, 1210, 1220, 1420, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963)	−5664.2	11494.5
4	20	(26, 1041, 1070, 1210, 1220, 1420, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3739, 3942, 3963)	−5669.1	11512.6
5	19	(26, 1040, 1070, 1210, 1220, 1415, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963)	−5670.3	11506.7
6	19	(26, 1040, 1070, 1210, 1220, 1415, 1436, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963)	−5671.1	11508.2
7	19	(26, 1041, 1070, 1210, 1220, 1415, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963)	−5671.3	11508.8
8	19	(26, 1040, 1070, 1210, 1220, 1415, 1436, 1525, 1684, 1866, 2046, 2408, 2470, 2532, 2591, 2771, 2780, 3942, 3963)	−5673.0	11512.1
9	20	(26, 1040, 1070, 1210, 1220, 1415, 1436, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3728, 3942, 3963)	−5679.6	11533.6
10	20	(26, 1040, 1070, 1210, 1220, 1415, 1436, 1525, 1684, 1866, 2046, 2408, 2470, 2532, 2591, 2771, 2780, 3728, 3942, 3963)	−5681.5	11537.5

consequently the detection of the change-points is important in oil-drilling (Fearnhead and Clifford, 2003). Ó Ruanaidh and Fitzgerald (1996) studied these data using MH to fit a change-point model with a fixed number of change points. We suggest to use of the normal change-point model since well-log data was assumed to follow a univariate Gaussian model in Adams and MacKay (2007).

Ó Ruanaidh and Fitzgerald (1996) detected 13 change-points with the same data. However, this is not relevant because they analyzed the data after the outliers were removed. Fearnhead and Clifford (2003) only inferred 16 change-points with an online analysis. However, they produced a conservative estimate of the number of change-points with the posterior restriction. Fearnhead (2006) found 18 more change-points than the ones of Ó Ruanaidh and Fitzgerald (1996).

We partition the sample space according to the model index with $k_{min} = 10$ and $k_{max} = 20$. We set $t_0 = 50000$, $\lambda = 15$, $\gamma = 2.0$ and $\delta = 0.00001$, for a conjugate prior on σ_i^{-2}. The uniform distribution is used as a proposal distribution, and SAMC is run for 1.0×10^7 iterations. The largest log-posterior value and smallest BIC identify the maximum posterior change-point estimates (26, 1034, 1070, 1210, 1220, 1420, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963) (Table 1), which are very similar to the results reported by Fearnhead (2006). The change-point estimates of all other examples were also obtained based on the largest log-posterior and smallest BIC. Figure 1 shows the MAP estimate as 19 change-points indicated as vertical lines with the well-log data. Table 2 shows the mean, standard deviation and p-value of an independence test of observations in each partition separated by estimated change-points that result in mostly independent observations.

4.2. Time intervals in days between successive coal mining disaster data for the exponential change-point model

The time intervals in days between successive coal mining disasters are obtained from 1875 to 1951 in which there were 109 explosions that involved the death of ten or more men. The data from Maguire *et al.* (1952) are analyzed in Cox and Lewis (1966). It is of interest to know some points where

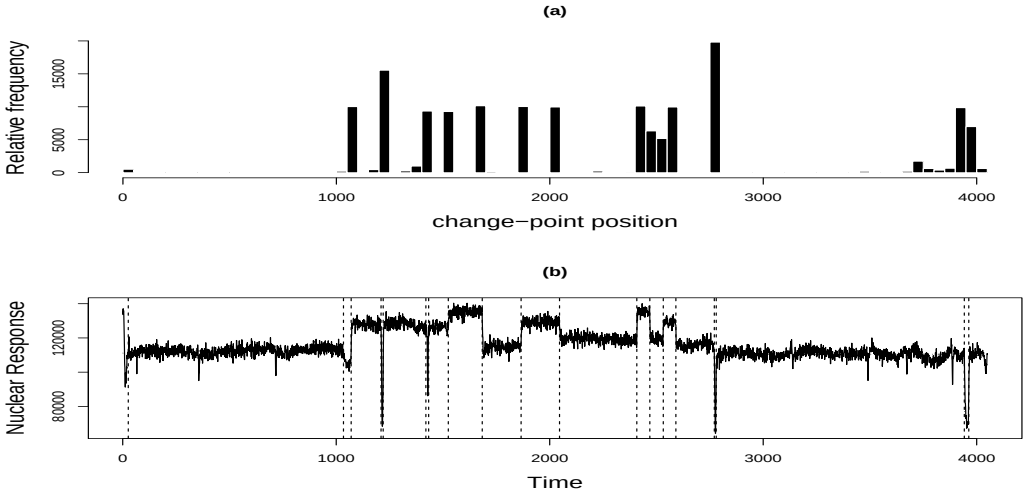

Figure 1: *The results of multiple change-point estimation in well-log data: (a) A histogram of log posterior probabilities of change-point positions; (b) A maximum posteriori estimate of the change-point positions.*

Table 2: Mean, standard deviation(SD) and p-value of an independence test of well-log in each partition separated by change-points.

Partition (♯ of data)	Mean	SD	Independence test (p-value)	Partition (♯ of data)	Mean	SD	Independence test (p-value)
1–26 (26)	111156.96	14393.56	4.9×10^{-6}	27–1034 (1008)	112384.05	2799.79	1.1×10^{-10}
1035–1070 (36)	105619.22	2647.58	2.3×10^{-4}	1071–1210 (140)	127960.16	2428.12	1.0×10^{-1}
1211–1220 (10)	87814.07	18140.37	1.6×10^{-2}	1221–1420 (200)	127574.19	2699.88	1.9×10^{-3}
1421–1433 (13)	113647.32	14032.90	7.7×10^{-3}	1434–1525 (92)	126227.92	2273.51	8.7×10^{-2}
1526–1684 (159)	134990.97	2460.49	3.9×10^{-1}	1685–1866 (182)	114869.98	2670.05	9.8×10^{-3}
1867–2046 (180)	129288.71	2483.29	2.1×10^{-1}	2047–2408 (362)	119354.03	2327.52	3.5×10^{-3}
2409–2469 (61)	135276.06	2448.93	9.7×10^{-1}	2470–2532 (63)	119679.77	2454.76	2.0×10^{-1}
2533–2591 (59)	129173.16	1910.81	2.7×10^{-1}	2592–2771 (180)	116041.76	2393.07	1.9×10^{-2}
2772–2780 (9)	82248.53	14761.79	6.9×10^{-2}	2781–3942 (1162)	110521.93	2967.77	2.2×10^{-16}
3943–3963 (21)	76930.45	9647.79	1.7×10^{-4}	3964–4050 (87)	109643.69	3589.66	2.5×10^{-7}

the disaster pattern may change. We use the exponential change-point model since the time between accidents can be modeled with an exponential distribution.

The same setting in the previous example is used except for the parameters, $\gamma = 2.0$ and $\delta = 0.002$, for a prior on σ_i. The analysis provides a change-point at 46 in the year 1897, which is the same estimate detected by Wu (2005). There seems to be variations after the change-point at 46; however, our analysis shows that there are no more mean changes. Figure 2 shows the MAP estimate in 1987. The mean and standard deviation of the first and second periods are $(117.00, 102.13)$ and $(331.44, 375.63)$, respectively. In addition, we performed an independence test in each partition which provides no significant results (p-value: 0.6889 and 0.3208).

4.3. Weekly log price relative data for the binomial change-point model

The data in Hsu (1979) is the weekly 161 log price relative(LPR) of the Dow Jones Industrial Average for the period from July 1, 1971 to August 2, 1974. Let Y_i be the LPR in week i following a $N(\mu, \sigma^2)$ distribution. The value of μ should be practically zero under modest "efficient market" assumptions.

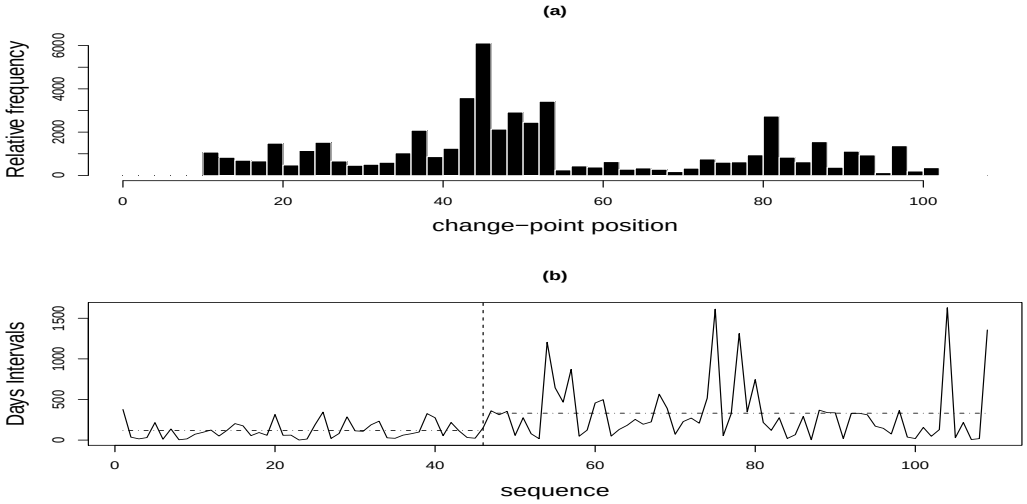

Figure 2: *Multiple change-point estimation results with data on the time intervals in days between successive coal mining disasters: (a) A histogram of log posterior probabilities of change-point positions; (b) MAP estimate of the change-point positions.*

However, there can be doubts about the constancy of σ. Visually the sequence seems to be more variable in the later than in the earlier period. As a different formulation, Hawkins (2001) created a Bernoulli variable which is set to 1 if the absolute value of the LPR is greater than or equal to 0.0133, the median absolute LPR, and 0 if it is less than 0.0133. This transformation may isolate subsequences of higher than average variability. We use the binomial change-point model to detect the LPR change.

Hsu (1979) found that the pattern of weekly log price between 1971 and 1974 is bullish up to early 1973 and bearish from early 1973 and later because of the Arab oil embargo and a steady climb in US prime interest rates occurred during the later part (1973–1974) of this period that may have impacted stock prices. He also found the second portion of the series (roughly from mid-March 1973) had a larger variation than the first segment.

We used the same settings as in the previous examples when applying SAMC to the same data, except for the partitioned sample space with $\alpha = 1.0$, and $\beta = 1.0$ for a prior on p_i. Figure 3 indicates that a change-point in the time intervals may take place at position 89; *i.e.*, the shift occurred during the week March 19–23, 1973 (the third week of March, 1973) which is the same estimate reported by Hsu (1979). The mean and standard deviation in each segment separated by one of the identified change-point positions are (0.3371, 0.7361) and (0.4754, 0.4438), respectively. The independence test in each partition results in no significant results (p-value: 0.0511 and 0.5385).

4.4. British coal-mine accident data for the Poisson change-point model

We apply our method to the British coal-mine accident data in Maguire *et al.* (1952). Jarrett (1979) extended the analysis with the same prior setting as in Carlin *et al.* (1992). The data are 112 annual counts from 1851–1962. Raftery and Akman (1986) examined the collection of times between occurrence as a continuous process and Carlin *et al.* (1992) studied the occurrence count data. We consider the Poisson change-point model for this accident count data.

For a Bayesian analysis, we set $\lambda = 1$, $\gamma = 0.5$ and $\delta = 0.9$ as a conjugate gamma prior for θ_i. Figure 4(b) shows a change-point at position 41 with the data, where the shift occurred in 1891 and

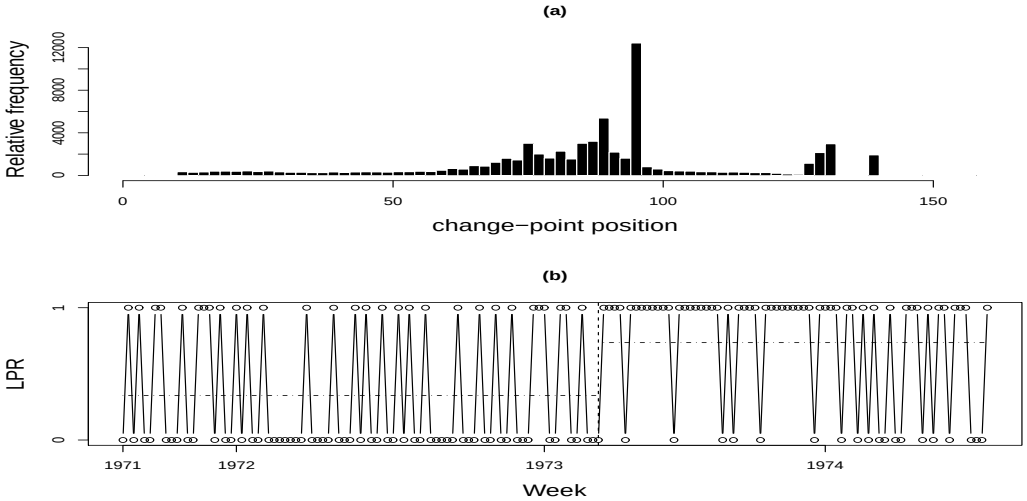

Figure 3: *Multiple change-point estimation results with weekly log price relative binomial data: (a) A histogram of log posterior probabilities of change-point positions; (b) MAP estimate of the change-point positions.*

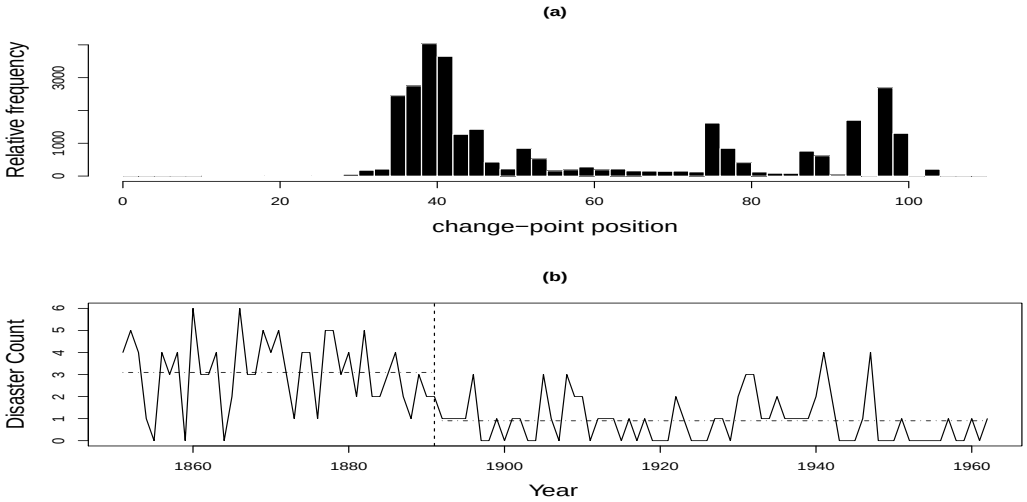

Figure 4: *Multiple change-point estimation results with British coal-mine accident data: (a) A histogram of log posterior probabilities of change-point positions; (b) MAP estimate of the change-point positions.*

is the same estimate reported by Carlin *et al.* (1992). The three largest spikes occurred at positions 39, 40 and 41; consequently, the change probably occurred sometime between late 1889 and early 1892. Raftery and Akman (1986) found a change-point, 1890. Lee (1998) used Bayesian analysis and found that the change-point was at 1890, the same change-point as Siegmund (1988) estimated with the likelihood ratio approach. The means and standard deviations of the two identified segments are (3.0976, 0.9014) and (1.5938, 1.0303) respectively. The independence test in each partition results in no significant results (p-value: 0.6377 and 0.7132).

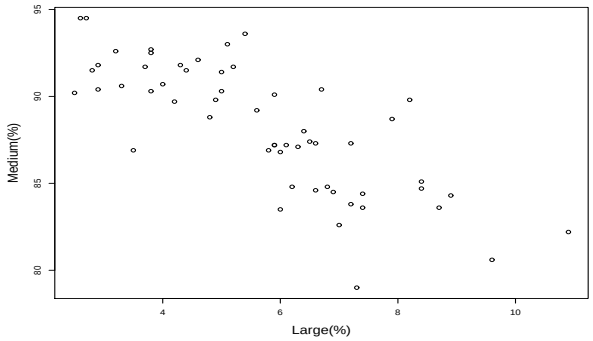

Figure 5: *A plot of large and medium particles data from a European grit producing plant*

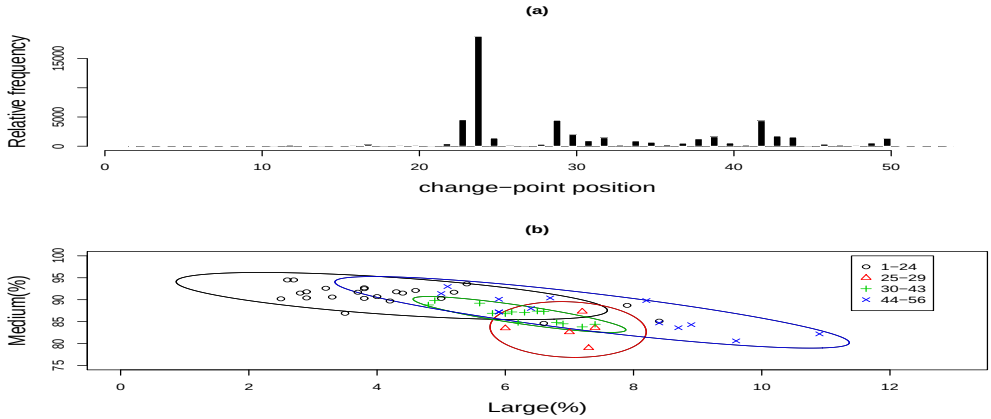

Figure 6: *Multiple change-point estimation results with data from a European grit producing plant: (a) A posterior histogram of change-point positions constructed using the posterior samples by SAMC; (b) A plot of large and medium particles, with the maximum log-posterior estimate of the change-point positions.*

4.5. European plant data for the bivariate normal change-point model

The data are 56 individual bivariate observations from a European plant that produces grit (or gravel) giving the percent of the particles (by weight) that are large and medium in size. Holmes and Mergen (1993) studied the particle size distribution for this data. Sullivan and Woodall (2000) used a preliminary control chart to detect change in the mean vector or the covariance matrix (or in both). Figure 5 shows the sequential design points determined and sorted by weight. It is of interest to know where the multivariate observation changes in order to distinguish particle sizes. We use the multiple change-point model for multivariate normal distributions. We used the exact same settings as in the previous example, except the prior

$$\Lambda_0^{-1} = \begin{pmatrix} 4.0 & -5.0 \\ -5.0 & 13.0 \end{pmatrix}.$$

The highest posterior log-likelihood value occurs at (24, 29, 43), which is the same as that reported by Sullivan and Woodall (2000). Figure 6(a) shows that most models with high posterior probabilities include four clustering of change-points around 24, 29 and 43. Figure 6(b) is a plot of large (%) and

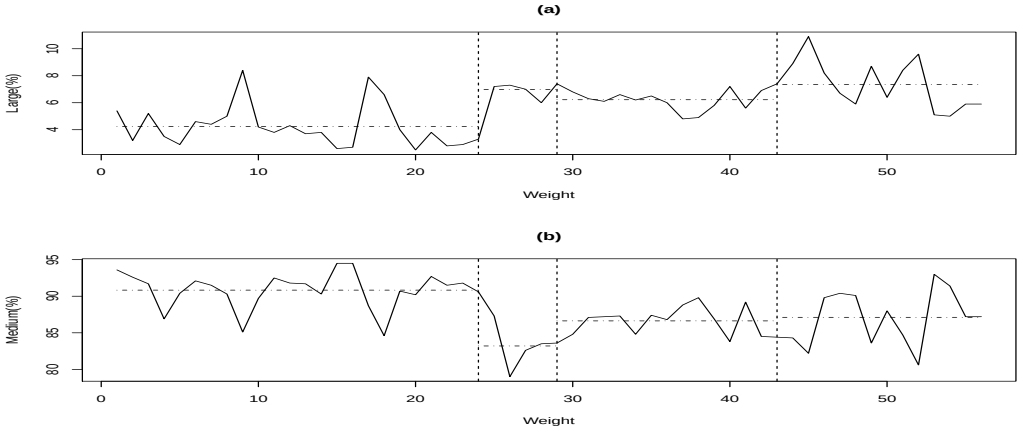

Figure 7: *The results for a European grit producing plant example: (a) A plot of the maximum log-posterior estimate of the change-point positions (vertical lines) and means (horizontal lines) of large particles in each subregion. (b) is the same with (a) except for medium particles.*

Table 3: The mean and covariance estimates in each subregion separated by estimated change-points in data from a grit producing plant.

Partition (♯ of data)	Sorted weight interval	Mean $\begin{pmatrix} \text{Large} \\ \text{Medium} \end{pmatrix}$	Covariance matrix $\begin{pmatrix} \text{Var(L)} & \text{Cov(L, M)} \\ \text{Cov(M, L)} & \text{Var(M)} \end{pmatrix}$	Bivariate normal test (p-value)	Independence test (p-value)
1 (24)	1–24	4.2292 90.8333	2.4552 −2.3810 −2.3810 6.2632	0.0208	0.2812
2 (5)	25–29	6.9800 83.2000	0.3220 −0.1450 −0.1450 8.7650	0.7058	0.9545
3 (14)	30–43	6.2214 86.6286	0.5972 −1.2653 −1.2653 3.6360	0.2605	0.0050
4 (13)	44–56	7.3538 87.1154	3.4944 −6.0276 −6.0276 14.4514	0.8323	0.0005

medium (%), distinguished by the estimated change-points, (24, 29, 43); *i.e.*, the size and orientation of the ellipses are considerably different and visually indicate the dissimilarity between the sample covariance matrices. Figure 7 shows the change-point locations of weights as large and medium, respectively. A modest shift in the sample mean vectors is evident in the plot.

Table 3 provides the mean vector, the covariance matrix in each segmentation, p-value from a Shapiro-Wilks test for normality and p-value of an independence test of observations in each partition separated by estimated change-points. Each segment is characterized by gravel size as large and medium according to weight. The shift at 24 is roughly due to a shift in the mean vector and the covariance matrix; however, the segmentation at 29 and 43 occur due to the change in the covariance matrix.

5. Conclusion

One feature of the change-point problem is the complicated nature of the distribution theory for inferential statistics. The Bayesian multiple change-point estimation in this paper may simplify data analysis by providing automatic segmentation. The Bayesian simultaneous change-point estimation precludes the need for a complicated likelihood function with all of the conditional parameter esti-

mates. We have discussed Bayesian multiple change-point models for the exponential family distributions developed by Kim and Cheon (2010) and Cheon and Kim (2010), and applied those models to the observation classification. We illustrated the application of posterior distributions to several data sets such as the well-log data, weekly log price relative data, British coal-mining disaster accident data, time interval data, and grit production plant data. After providing each change-point, parameter estimation can be performed to yield information about each partition divided by each change-point. Hence, our method is simple to understand and is easily applied to change-point estimation and segmentation for sequences of independent random variables or random vectors from the provided exponential family distributions.

References

Adams, R. and Mackay, D. (2007). *Bayesian Online Changepoint Detection*, University of Cambridge Technical Report.

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

Bartels, R. (1982). The rank version of von Neumann's Ratio test for randomness, *Journal of the American Statistical Association*, **77**, 40–46.

Belisle, P., Joseph, L., MacGibbon, B., Wolfson, D. and du Berger, R. (1998). Change-point analysis of neuron spike train data, *Biometrics*, **54**, 113–123.

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

Cheon, S. and Kim, J. (2010). Multiple change-point detection of multivariate mean vectors with Bayesian approach, *Computational Statistics & Data Analysis*, **54**, 406–415.

Chernoff, H. and 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. (1998). Estimation and comparison of multiple change-point models, *Journal of Econometrics*, **86**, 221–241.

Corander, J., Gyllenberg, M. and Koski, T. (2009). Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy, *Advanced Data Analysis and Classification*, **3**, 3–24.

Cox, D. and Lewis, P. (1966). Methuen's Monographs on Applied Probability and Statistics, In *The Statistical Analysis of Series of Events*, John Wiley, London.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems, *Statistics and Computing*, **16**, 203–213.

Fearnhead, P. and Clifford, P. (2003). Online inference for hidden Markov models, *Journal of the Royal Statistical Society Series B*, **65**, 887–899.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis*, Chapman and Hall, New York.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97–109.

Hawkins, D. (2001). Finding multiple change-point models to data, *Computational Statistics & Data Analysis*, **37**, 323–341.

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

Hinkley, D. (1972). Time-ordered classification, *Biometrika*, **59**, 509–523.

Holmes, D. and Mergen, A. (1993). Improving the performance of the T^2 control chart, *Quality Engineering*, **5**, 619–625.

Hsu, D. (1979). Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis, *Journal of the American Statistical Association*, **74**, 31–40.

Jarrett, R. (1979). A note on the intervals between coal-mining disasters, *Biometrika*, **66**, 191–193.

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

Kim, J. and Cheon, S. (2010). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo, *Computational Statistics*, **25**, 215–239.

Kojadinovic, I. and Holmes, M. (2009). Tests of independence among continuous random vectors based on Cramer-von Mises functionals of the empirical copula process, *Journal of Multivariate Analysis*, **100**, 1137–1154.

Lee, C.-B. (1998). Bayesian analysis of a change-point in exponential families with applications, *Computational Statistics & Data Analysis*, **27**, 195–208.

Liang, F., Liu, R. and Carroll, R. (2007). Stochastic approximation in Monte Carlo computation, *Journal of the American Statistical Association*, **102**, 305–320.

Maguire, B., Pearson, E. and Wynn, A. (1952). The time intervals between industrial accidents, *Biometrika*, **38**, 168–180.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1091.

Ó Ruanaidh, J. and Fitzgerald, W. (1996). *Bayesian Online Changepoint Detection*, Springer, New York.

Raftery, A. (1995). Bayesian model selection in social research, In Marsden V (Ed) *Sociological methodology*, **25**, 111–163.

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

Robbins, H. and Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics*, **22**, 400–407.

Siegmund D. (1988). Confidence sets in change point problem, *International Statistical Review*, **56**, 31–48.

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

Stephens, D. (1994). Bayesian retrospective multiple-change-point identification, *Applied Statistics*, **43**, 159–178.

Sullivan, J. and Woodall, W. (2000). Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations, *IIE Transactions*, **32**, 537–549.

Venter, J. and Steel, S. (1996). Finding multiple abrupt change-points, *Computational Statistics & Data Analysis*, **22**, 481–504.

Wehrens, R., Buydens, L. M. C., Fraley, C. and Raftery, A. E. (2004). Model-based clustering for image segmentation and large datasets via sampling, *Journal of Classification*, **21**, 231–253.

Wu, Y. (2005). Inference for change point and post change means after a CUSUM test, *Lecture notes in Statistics*, **180**.

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

Received July 12, 2013; Revised September 22, 2013; Accepted September 30, 2013