The Korean Journal of Applied Statistics (2009)
22(4), 687–695

Multiple Change-Point Estimation of Air Pollution Mean Vectors

Jaehee Kim[1] · Sooyoung Cheon[2]

[1]Department of Statistics, Duksung Women's University;
[2]KU Industry-Academy Cooperation Group Team of Economics and Statistics, Korea University

(Received April 2009; accepted May 2009)

Abstract

The Bayesian multiple change-point estimation has been applied to the daily means of ozone and PM10 data in Seoul for the period 1999. We focus on the detection of multiple change-points in the ozone and PM10 bivariate vectors by evaluating the posterior probabilities and Bayesian information criterion(BIC) using the stochastic approximation Monte Carlo(SAMC) algorithm. The result gives 5 change-points of mean vectors of ozone and PM10, which are related with the seasonal characteristics.

Keywords: Bayesian change-point model, Bayesian information criterion(BIC), multivariate normal distribution, ozone, PM10, posterior, stochastic approximation Monte Carlo(SAMC), truncated Poisson.

1. Introduction

Air pollution has become local as well as regional issue of big cities, industrial centers. The primer focus of this paper is to present an application of methodology for air pollution with ozone and dust aerosols measured by PM10 fraction size (dust fractions of particles below $10\mu m$). Many of the statistical contributions focus on determining the relationship among ozone concentrations, meteorology and air pollutants. The need of innovative statistical methods for modern environmental assessment is undisputed. Its statistical applications to the environmental sciences are increasing and they play an important role in environmental monitoring and assessment. Ozone(O_3) is a ubiquitous trace gas that absorbs some of the biologically harmful ultraviolet radiation in the stratosphere. On the surface, ozone is harmful, with destructive impacts on materials, crops and health. Its levels have been high enough in certain areas to be of concern for several decades. Ozone levels are difficult to control, as it is a secondary pollutant. It results from photochemical reactions involving precursor pollutants. The precursors include a variety of volatile organic compounds, comprised mainly of non-methane hydrocarbons and nitric oxide(NO) and nitrogen dioxide(NO_2). Both non-methane hydrocarbons and nitrogen oxides are emitted from transportation and industrial processes. Volatile organic compounds are emitted from diverse sources such as automobiles,

This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(2009-03-M020-0025).
[1]Corresponding author: Professor, Department of Statistics, Duksung Women's University, 419, Ssangmun-Dong Dobong-Ku, Seoul 132-714, Korea. E-mail: jaehee@duksung.ac.kr

chemical manufactures, dry cleaners and other facilities using chemical solvents. The rates and completeness of the reactions that produce ozone, as well as its subsequent transport and deposition, are driven by meteorological conditions: the availability of sunlight, temperature and wind speed.

The 1990s was an extraordinary decade for air pollution research, regulation and control. Researchers have attempted to provide relevant evidence by using statistical approaches to separate the effect of particulate matter from the effects of other pollutants since PM10 affects the mortality. There has been a remarkably fast evolution in the extent of the evidence on particulate matter and health and in the interpretation of the findings. Large associations between the incidence of respiratory symptoms and PM10 pollution were also observed for some researches (Safadi and Pena, 2008).

For the Bayesian change analysis, Smith (1975) made a Bayesian inference about a change-point and Carlin *et al.* (1992) derived hierarchical Bayesian analysis of change-point problems. Barry and Hartigan (1993) and Crowley (1997) applied the PPM(product partition model) to the identification of multiple change-points in normal means and used a Gibbs sampling approach to obtain the product estimates. Loschi and Cruz (2005) extended the PPM to calculate probability of change points in the means and variances of normal data sequences with Gibbs sampling. Son and Kim (2005) considered Bayesian single change-point detection in a sequence of multivariate normal observations using intrinsic Bayes factor. For their Bayesian approach, the Markov sampling technique is used to calculation of the posterior probabilities. With more possible partitions, the model space becomes complex with multiple modes. For the computational algorithm of posterior probabilities of this complicated form, recent computational methods are developed and those can improve the calculation of the posterior probabilities for changing dimension. Green (1995) proposed the reversible jump Markov chain Monte Carlo(RJMCMC) algorithm that RJMCMC samplers jump between parameter subspaces of differing dimensionalities which are applicable for multiple change-point problems. SAMC has been also applied successfully to many hard computational problems, and Bayesian model selection (Liang *et al.*, 2007; Liang, 2009). Recently Cheon and Kim (2009) derived the posterior distribution of multivariate mean vectors in the multiple change-point models and used SAMC for posterior calculation. Cheon and Kim (2009) and Liang (2009) showed that SAMC performs better than RJMCMC when sample space is complex in Bayesian model selection problems.

For the recent change-point estimation with air pollutants, Jaruskova (1997) summarized the researching methods of discovering change in the behavior of meteorological series including total ozone. Borchi *et al.* (2006) considered detecting the variability of total ozone. Carslaw *et al.* (2006) estimated change-points of PM10 using CUSUM(cumulative sum) plot. However they only considered the change problem with the univariate data. The multivariate change analysis is a needed technique for environmental data.

In this paper, the Bayesian multiple change-point estimation method developed by Cheon and Kim (2009) is applied to data collected in the city of Seoul, the largest urban agglomeration of South Korea. This paper is organized as follows. Section 2 describes Bayesian multiple change-point problems for normal distribution with posterior distribution. Section 3 presents some numerical results including simulation and real data analysis with the ozone and pm10 in Seoul in 1999. The posterior distribution is calculated with the SAMC algorithm. Finally Section 4 concludes the paper with a discussion of Bayesian change-point problems.

2. Bayesian Multiple Change-Point Model of Multivariate Normal Observations

Let $Z = (z_1, \ldots, z_n)$ denote the independent observation sequence ordered in time. There exists a partition on the set $\{1, 2, \ldots, n\}$ into blocks such that the sequence follows the same distribution within blocks. A binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_{n-1})$ is introduced with $x_{c_1} = x_{c_2} = \cdots = x_{c_k} = 1$ and being 0 elsewhere, and $0 = c_0 < c_1 < \cdots < c_k < c_{k+1} = n$. There are unknown k change-points in the model.

The multiple change-point model can be written as follows.

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

for $r = 1, 2, \ldots, k+1$ and f_r depends on the parameters $\boldsymbol{\theta}_r \in \boldsymbol{\Theta}$. The parameters change at $c_1 + 1, c_2 + 1, \ldots, c_{k-1} + 1, c_k + 1$. Here, $c_1, c_2, \ldots, c_{k-1}, c_k$ are called the change-points. Let f_r be a d-dimensional multivariate normal density parameterized by $\theta_r = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with d -dimensional mean vector $\boldsymbol{\mu}$ and $d \times d$ covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{x}^{(k)}$ denote a configuration of $\mathbf{x}$ with k change-points. Let $\eta^{(k)} = (\mathbf{x}^{(k)}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_{k+1}, \boldsymbol{\Sigma}_{k+1})$ and A_k be the space of models with k change-points, $\mathbf{x}^{(k)} \in A_k$ and $\chi = \cup_{k=0}^{n} A_k$. The likelihood function of Z with d-dimensional observed vector $\mathbf{z}_j$ is

$$L\left(Z|\eta^{(k)}\right) = \prod_{j=c_0+1}^{c_1} f_1(\mathbf{z}_j|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdots \prod_{j=c_k+1}^{c_{k+1}} f_{k+1}(\mathbf{z}_j|\boldsymbol{\mu}_{k+1}, \boldsymbol{\Sigma}_{k+1}), \tag{2.2}$$

where the d-dimensional multivariate normal density is

$$f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})\right\}.$$

Therefore, the log-likelihood function of model $\eta^{(k)}$ is

$$\log L\left(Z|\eta^{(k)}\right) = -\frac{dn}{2} \log(2\pi) - \sum_{i=1}^{k+1} \left\{ \frac{c_i - c_{i-1}}{2} \log|\boldsymbol{\Sigma}_i| + \sum_{j=c_{i-1}+1}^{c_i} \frac{1}{2} (\mathbf{z}_j - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{z}_j - \boldsymbol{\mu}_i) \right\}. \tag{2.3}$$

For a Bayesian analysis, consider the prior distribution for $\mathbf{x}^{(k)}$ in $\eta^{(k)}$ as

$$\pi\left(\mathbf{x}^{(k)}\right) = \frac{\lambda^k}{\sum_{j=0}^{n-1} \frac{\lambda^j}{j!}} \frac{(n-1-k)!}{(n-1)!}, \quad k = 0, 1, \ldots, n-1. \tag{2.4}$$

For the prior setting, we set that A_k has a truncated up to $(n-1)$ Poisson distribution with parameter λ and each model in A_k is equally likely. A uniform prior and an inverse-Wishart $\mathrm{IW}(\nu_0, \Lambda_0^{-1})$ are given on $\boldsymbol{\mu}_i$'s and $\boldsymbol{\Sigma}_i$'s, respectively. These priors are independent of each other. The log-prior density is

$$\log \pi\left(\eta^{(k)}\right) = a_k - \sum_{i=1}^{k+1} \left\{ \frac{\nu_0 + d + 1}{2} \log|\boldsymbol{\Sigma}_i| + \mathrm{tr}\left(\frac{1}{2} \boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}_i^{-1}\right) \right\}, \tag{2.5}$$

where

$$a_k = -(k-1)\left[\frac{\nu_0 d}{2} \log 2 + \frac{d(d-1)}{4} \log \pi + \sum_{u=1}^{d} \log \Gamma\left(\frac{\nu_0 + 1 - u}{2}\right) - \frac{\nu_0}{2} \log|\boldsymbol{\Lambda}_0|\right] + \log(n-1-k)! + k \log \lambda.$$

Here ν_0, Λ_0 and λ are fixed hyperparameters. The log posterior of $\eta^{(k)}$ (up to an additive constant) can be obtained by adding (2.3) and (2.5). The combined log probability is as follows.

$$\begin{aligned}
&\log P\left(Z|\eta^{(k)}\right) P\left(\boldsymbol{\Sigma}|\mathbf{x}^{(k)}\right) \pi\left(\mathbf{x}^{(k)}\right) \\
&= a_k - \sum_{i=1}^{k+1} \left\{ \frac{c_i - c_{i-1} + \nu_0 + d + 1}{2} \log|\boldsymbol{\Sigma}_i| + \operatorname{tr}\left(\left\{ \frac{\boldsymbol{\Lambda}_0}{2} + \frac{(c_i - c_{i-1})}{2} \mathbf{S}_i \right\} \boldsymbol{\Sigma}_i^{-1} \right) \right. \\
&\quad \left. + \frac{(c_i - c_{i-1})}{2} (\boldsymbol{\mu}_i - \bar{\mathbf{z}}^i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \bar{\mathbf{z}}^i) \right\}
\end{aligned}$$

where $\mathbf{z}_j = (z_{j1}, \ldots, z_{jd})'$, for $j = c_{i-1} + 1, \ldots, c_i$, $\mathbf{1}_{(c_i - c_{i-1}) \times 1} = (1, \ldots, 1)'$,

$$\mathbf{z}^i = \begin{pmatrix} z_{c_{i-1}+1,1} & \cdots & z_{c_{i-1}+1,d} \\ \vdots & \ddots & \vdots \\ z_{c_i,1} & \cdots & z_{c_i,d} \end{pmatrix}, \quad \bar{\mathbf{z}}^i = \left(\frac{1}{c_i - c_{i-1}} \sum_{j=c_{i-1}+1}^{c_i} z_{j1}, \ldots, \frac{1}{c_i - c_{i-1}} \sum_{j=c_{i-1}+1}^{c_i} z_{jd} \right)',$$

$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})', \quad \bar{\mathbf{z}}^i = \left(\frac{1}{c_i - c_{i-1}} \sum_{j=c_{i-1}+1}^{c_i} z_{j1}, \ldots, \frac{1}{c_i - c_{i-1}} \sum_{j=c_{i-1}+1}^{c_i} z_{jd} \right)',$$

$\boldsymbol{\Sigma}_i^{-1}$, for $i = 1, \ldots, k+1$

and

$$\mathbf{S}_i = \frac{1}{c_i - c_{i-1}} \sum_{j=c_{i-1}+1}^{c_i} (\mathbf{z}_i - \bar{\mathbf{z}}^i)(\mathbf{z}_i - \bar{\mathbf{z}}^i)'.$$

Integrating out $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k+1}$ and $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{k+1}$ from the full posterior distribution shown in Cheon and Kim (2009), it is obtained as

$$\begin{aligned}
\log \pi(\mathbf{x}^{(k)}|Z) \;=\; & b_k + \sum_{i=1}^{k+1} \left\{ \frac{\nu_i d}{2} \log 2 + \frac{d(d-1)}{4} \log \pi \right. \\
& \left. + \sum_{u=1}^{d} \log \Gamma\left(\frac{\nu_i + 1 - u}{2} \right) - \frac{\nu_i}{2} \log|\boldsymbol{\Lambda}_0 + (c_i - c_{i-1}) \mathbf{S}_i| \right\}
\end{aligned}$$

where $\nu_i = c_i - c_{i-1} + \nu_0 - 1$.

We can choose the model that minimizes Bayesian Information Criterion(BIC) of competing models for a data set. The model with the highest posterior probability may be the one that minimizes BIC $= -2$(log maximized likelihood) + (number of parameters).

3. Numerical Results

3.1. A simulated example with bivariate normal observations

In this Section, we first explain the SAMC (Liang *et al.*, 2007) algorithm briefly, and then test the ability of SAMC to find the global maximum in multiple change-point problems.

Let $f(\mathbf{x}) = c\psi(\mathbf{x})$, for $\mathbf{x} \in \chi$, denote the target probability density/mass function we are working with, where χ is the sample space and c is an unknown constant. Let $E_1, E_2, \ldots, E_m$ denote a

Table 3.1. Relative sampling frequencies of the subregions for the simulated example.

Subregion	frequency(%)	Subregion	frequency(%)	Subregion	frequency(%)
$(-\infty, -672)$	106.06	$-672, -670$	100.03	$-670, -668$	98.25
$[-668, -666)$	96.88	$-666, -664$	98.43	$-664, -662$	97.41
$[-662, \infty)$	102.94				

partition of χ according to the negative log-likelihood function $H(\mathbf{x}) = -\log \psi(\mathbf{x})$, and let $w_i = \int_{E_i} \psi(\mathbf{x}) d\mathbf{x}$ for $i = 1, \ldots, m$. SAMC seeks to sample from the trial distribution

$$f_w(\mathbf{x}) = \sum_{i=1}^{m} \frac{\pi_i \psi(\mathbf{x})}{w_i} I(\mathbf{x} \in E_i),$$

where π_i's are pre-specified constants such that $\pi_i > 0$ for all i and $\sum_{i=1}^{m} \pi_i = 1$. In Liang *et al.* (2007), $\pi = (\pi_1, \ldots, \pi_m)$ is called the desired sampling distribution of the subregions. It is easy to see that if $w_1, w_2, \ldots, w_m$ can be well estimated, sampling from $f_w(\mathbf{x})$ will result in a "random walk" in the space of subregions (by regarding each subregion as a "point"). Each subregion is sampled with a frequency proportional to π_i. Hence, the local-trap problem can be overcome essentially, provided that the sample space is partitioned appropriately. The success of SAMC depends crucially on the estimation of this w_i.

Now we show the ability of SAMC that the new Bayesian multiple change-point model via SAMC can be well applied to multiple change-point problems. A total of 200 observations were generated independently according to

$$z_1, \ldots, z_{50} \sim N\left(\begin{pmatrix} -5 \\ -5 \end{pmatrix}, \begin{pmatrix} 1 & -0.1 \\ -0.1 & 1 \end{pmatrix}\right), \quad z_{51}, \ldots, z_{90} \sim N\left(\begin{pmatrix} -1 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}\right),$$

$$z_{91}, \ldots, z_{140} \sim N\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}\right), \quad z_{141}, \ldots, z_{200} \sim N\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & -0.1 \\ -0.1 & 1 \end{pmatrix}\right).$$

The sequence of the simulated data is shown in the two-dimensional plot in Figure 3.1. We assume that there are no more than 99 change-points in the observation sequence, and the change-points only occur after even observations; *i.e.*, $c_i \in \{2, 4, \ldots, 198\}$. In this simulation, we partitioned the phase space into $E_1, \ldots, E_7$ with an equal bandwidth of 2.0; that is, we set $E_1 = \{\mathbf{x} \in \chi : H(\mathbf{x}) < -672\}, E_2 = \{\mathbf{x} \in \chi : -672 \leq H(\mathbf{x}) < -670\}, \ldots, E_7 = \{\mathbf{x} \in \chi : H(\mathbf{x}) \geq -662\}$. We also set $\lambda = 3, \nu_0 = 2$ and $\Lambda_0^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, which corresponds to a conjugate prior on $\mathbf{\Sigma}_i$. For a proposal distribution, the uniform distribution was used.

SAMC was run for 1.0×10^5 iterations. The overall acceptance probability was 0.1512. Table 3.1 shows the relative sampling frequencies obtained in a run, where the relative sampling frequency of subregion i is defined as $N_i/\bar{N} \times 100\%$, and N_i and $\bar{N}$ denote the sampling frequency of subregion i and the average sampling frequency of the subregions, respectively. The approximate equality of sampling frequencies at each subregion indicates that the SAMC samplers converge for this example. Table 3.2 lists the 10 models with the largest log-posterior values. Note that the true change-point model was at the first rank both in log-posterior values among all sampled models and based on the BIC criterion. Other results of the run are shown in Figure 3.2. The underlined model is true and is ranked 1 in log-posterior values among all models sampled by SAMC. The second column shows the differences of the log-posterior values of the models from the true model. BIC means Bayesian Information Criterion, a popular criterion for model selection.

Table 3.2. The 10 models with the largest log-posterior values sampled by SAMC.

No	Log − posterior*	# change-points	Change patterns	BIC
1	0.0000	3	(50, 90, 140)	−1324.896
2	−0.4198	4	(50, 88, 90, 140)	−1318.758
3	−0.9091	5	(50, 90, 112, 114, 140)	−1312.481
4	−1.3186	6	(50, 88, 90, 112, 114, 140)	−1312.481
5	−1.4819	5	(50, 90, 140, 158, 164)	−1311.336
6	−1.7974	5	(50, 90, 140, 194, 198)	−1310.705
7	−1.8248	4	(50, 90, 140, 158)	−1315.948
8	−1.8451	4	(2, 50, 90, 140)	−1315.908
9	−1.8902	4	(50, 90, 140, 158)	−1315.817
10	−1.8914	6	(50, 88, 90, 140, 158, 164)	−1305.218

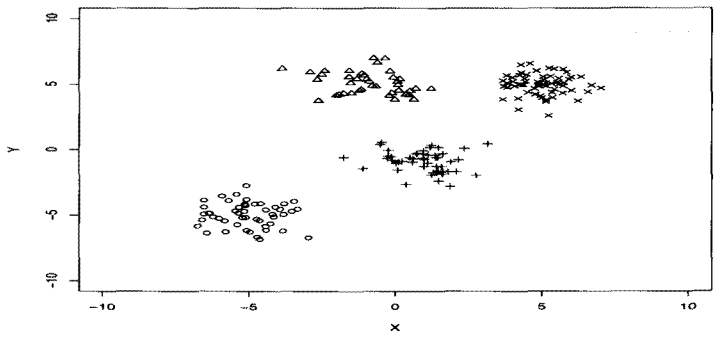

Figure 3.1. A sequence plot of simulated data from two-dimensional normal distribution

Figure 3.2(a) is the posterior histogram of the identified change-point positions. It shows that most of the models with high posterior probabilities include 3 change-points. This figure shows that the three most likely change-points are around 50, 90 and 140. Note that there is more uncertainty around second and third clusters of the histogram bars in Figure 3.2(a). This is consistent with the results shown in Table 3.1; *i.e.*, the models with the change-points at 50, around 102(88–114), and around 152(140–164) have very close log-posterior values. Figure 3.2(b) and (c) show the maximum posteriori estimate of the change patterns of the data in each variable. The maximum posteriori estimate of the change-points is (50, 90, 140), which is the same as the true values.

3.2. Example with air pollution data (Ozone and pm10) of Seoul in 1999

Meteorological data and air pollutants were measured hourly from January 1, 1999 to December 31, 1999 at 27 sites monitored by Korean Ministry of Environment. The hourly 24 averages on 27 sites in Seoul were computed. Meteorological data were obtained from Korea Meteorological Administration. Those data consist of temperature, precipitation, relative humidity, wind speed, wind direction, ozone and pm10. The standardized values were used in the statistical analysis. Figure 3.3 shows the plot of bivariate vectors, ozone and pm10, of Seoul according to date in 1999. For a Bayesian analysis, the hyperparameters should be decided from the previous information. We assume that there are no more than 181 change-points in the observation sequence, and the change-points only occur after even observations; *i.e.*, $c_i \in \{2, 4, \ldots, 362\}$. We partitioned the phase

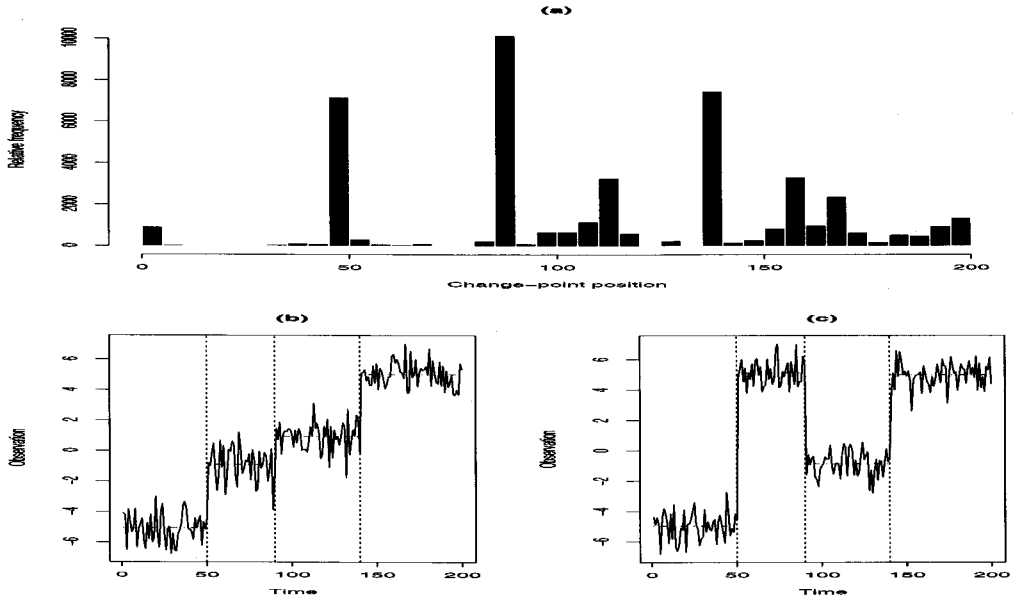

Figure 3.2. The simulation results for the change-point example: (a) The posterior histogram of change-point positions; (b) Maximum posteriori estimate of the change-point positions: the vertical (dotted) lines indicated the change-point positions, and the horizontal (dashed) lines indicate the mean value of observations separated by change-point positions; (c) same with (b) except for observations of the second variable.

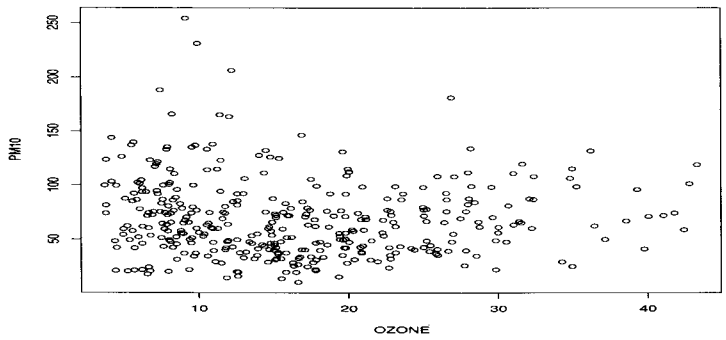

Figure 3.3. A Plot of ozone and pm10, in Seoul, 1999

space with an equal energy bandwidth of 2.0 and used the exactly same setting as that used for above simulated data, except that we set $\lambda = 5, \nu_0 = 2$ and $\Lambda_0^{-1} = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 32 \end{pmatrix}$, which corresponds to a inverse Wishart prior on $\mathbf{\Sigma}_i$. The run of SAMC consists of 5.0×10^5. The overall acceptance probability was 0.1421.

The best log-likelihood value found by SAMC is shown in Table 3.3. Figure 3.4(a) shows that most of the models with high posterior probabilities include five change-points; *i.e.*, around 24, 28, 92, 202 and 270. Figure 3.4(a) and Table 3.3 indicate that the models with the third one change-point being around 95(92–98), and the fourth one change-point being around 200(198–202) have very close log-posterior values. We used BIC for model selection. Figure 3.4(b) shows the Minimum

Table 3.3. The 10 models with the largest log-posterior values sampled by SAMC.

No	BIC	#change-points	Change patterns	Log-posterior
1	**−3006.701**	5	(24, 28, 94, 202, 270)	**1521.050**
2	−3004.030	5	(24, 28, 98, 202, 270)	1519.715
3	−3000.375	6	(24, 28, 94, 198, 200, 270)	1520.837
4	−3000.281	6	(24, 28, 94, 202, 270, 272)	1520.790
5	−3000.181	6	(24, 28, 92, 198, 200, 270)	1520.740
6	−2998.033	6	(24, 28, 94, 198, 202, 270)	1519.666
7	−2997.959	6	(24, 26, 28, 94, 202, 270)	1519.629
8	−2997.839	6	(24, 28, 92, 198, 202, 270)	1519.569
9	−2993.959	7	(24, 28, 94, 198, 200, 270,272)	1520.579
10	−2993.766	7	(24, 28, 92, 198, 200, 270,272)	1520.483

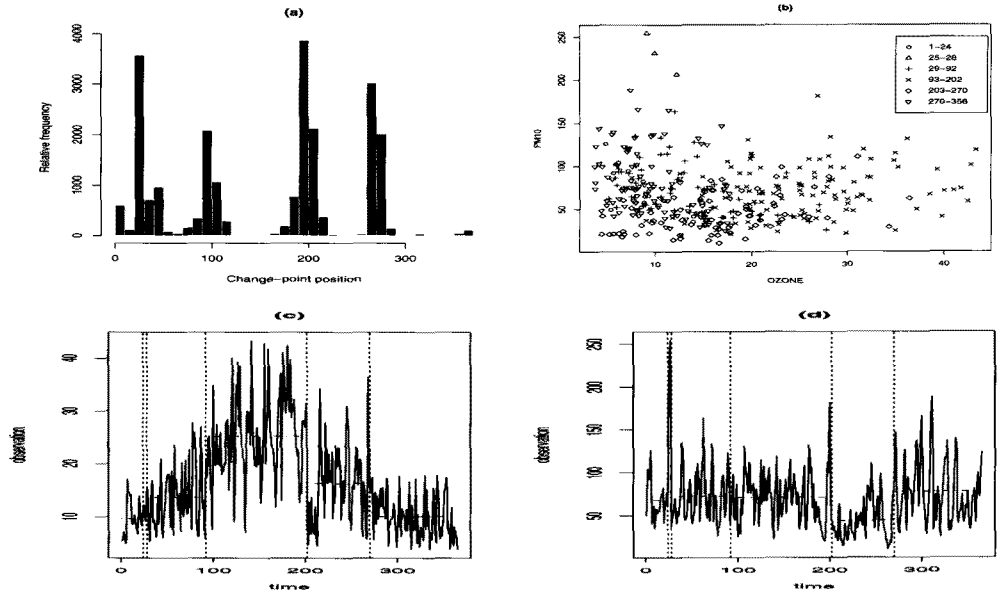

Figure 3.4. multiple Change-point estimation of ozone and pm10 (a) The posterior histogram of change-point positions sampled by SAMC; (b) A plot of bivariate, ozone and pm10, with the minimum BIC estimate of the change-point positions (c) A plot of the minimum BIC estimate of the change-point positions (the vertical (dotted) lines indicate the change-point positions, and the mean values of observations separated by change-point positions identified by the minimum BIC model) in ozone. (d) same with (c) except for pm10.

BIC estimate of the change-points is (24, 28, 92, 202, 270); *i.e.*, (Jan. 24, Jan. 28, Apr. 02, Jul. 21, Sep. 27). A detailed data exploration shows that change-points usually occurred the middle of each season. Table 3.4 implies that the changing pattern during winter may be due to the rapid increase in PM10.

4. Concluding Remark

In this paper, we have applied the Bayesian multiple change-point model with multivariate air pollution vectors. As a computational tool for the posterior calculation, SAMC was adopted because it efficiently estimates the optimum values without sticking to a local minimum or maximum.

Table 3.4. The mean and standard deviation(s.d.) values of observations separated by change-point positions of the minimum BIC model.

Partition	Period	Ozone		PM10	
		mean	s.d.	mean	s.d.
1	Jan.01–Jan.24	9.6592	3.4526	68.1146	22.7505
2	Jan.25–Jan.28	11.6408	2.7637	182.1950	99.0340
3	Jan.29–Apr.02	13.6492	5.8023	72.4048	29.4460
4	Apr.03–Jul.21	25.2095	7.8108	71.1220	29.3365
5	Jul.22–Sep.27	16.2660	7.1248	44.4502	23.4765
6	Sep.28–Dec.31	10.0605	4.0924	77.7125	36.9649

Simulation results and real data results support suitable identification of multiple change-points. Especially the multiple change-points in ozone and pm10 reflect the characteristic of the climate of Seoul. As a final remark, for the data in the example the multiple change-point problem of time series model can be considered and further research is expected.

References

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

Borchi, F., Naveau, P., Keckhut, P. and Hauchecorne, A. (2006). Detecting variability changes in Arctic total ozone column, *Journal of Atmoshperic and Solar-Terrestrail Physics*, **68**, 1383–1395.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of change point problems, *Journal of the Royal Statistical Society. Series C(Applied Statistics)*, **41**, 389–405.

Carslaw, D., Ropkins, K. and Bell, M. C. (2006). Change-point detection of gaseous and particulate traffic-related pollutants at a roadside location, *Environmental Science & Technology*, **40**, 6912–2918.

Cheon, S. and Kim, J. (2009). Multiple Change-point detection of multivariate mean vectors with Bayesian approach, *Computational Statistics & Data Analysis*, in revision.

Crowley, E. M. (1997). Product partition models for normal means, *Journal of the American Statistical Association*, **92**, 192–198.

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

Jaruskova, D. (1997). Some problems with application of change-point detection methods to evnironmental data, *Environmetrics*, **8**, 469–483.

Liang, F. (2009). Improving SAMC using smoothing methods: Theory and applications to Bayesian model selection problems, *The Annals of Statistics*, in press.

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

Loschi, R. H. and Cruz, F. R. B. (2005). Extension to the product partition model: Computing the probability of a change, *Computational Statistics & Data Analysis*, **48**, 255–268.

Safadi, T. and Pena, D. (2008). Bayesian analysis of dynamic factor models: An application to air pollution and mortality in Sao Paulo, Brazil, *Environmetrics*, **19**, 582–601.

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

Son, Y. S. and Kim, S. W. (2005). Bayesian single change point detection in a sequence of multivariate normal observations, *Statistics*, **39**, 373–387.