The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.4.589
27(4), 589–603

Bayesian Detection of Multiple Change Points in a Piecewise Linear Function

Joungyoun Kim[a,1]

[a]Biostatistics and Clinical Epidemiology Center, Research Institute for Future Medicine, Samsung Medical Center

(Received April 30, 2014; Revised July 31, 2014; Accepted August 14, 2014)

Abstract

When consecutive data follows different distributions(depending on the time interval) change-point detection infers where the changes occur first and then finds further inferences for each sub-interval. In this paper, we investigate the Bayesian detection of multiple change points. Utilizing the reversible jump MCMC, we can explore parameter spaces with unknown dimensions. In particular, we consider a model where the signal is a piecewise linear function. For the Bayesian inference, we propose a new Bayesian structure and build our own MCMC algorithm. Through the simulation study and the real data analysis, we verified the performance of our method.

Keywords: Change points, Bayesian inference, MCMC, reversible jump.

1. 서론

시간 순서에 따라 얻어진 자료들의 분포가 부분구간별로 다를 때, 변화점에 대한 추론 문제가 발생한다. 이러한 변화점 검출은 순차적으로 연속된 자료들에서 분포가 달라지는 시점을 찾아내는 것을 목표로 하여, 지진학, 이미지 프로세싱, 계량경제학, 회귀분석 및 추적관찰 등의 분야에서 널리 사용되고 있다.

본 논문에서는 변화점의 개수가 불분명할 때 그 개수, 위치 및 구간별 분포를 추론하는 다중 변화점 검출 문제를 다루고자 한다. 특히, 모수의 차원이 불확정적이라는 다중 변화점 추론의 특징적 문제를 다루기 위하여, 다양한 차원의 모수탐색을 통하여 유연하고 효과적인 추론을 할 수 있는 베이지안 방법을 제시한다.

베이지안 방법을 통한 변화점 추론에 관한 연구는 Chernoff와 Zacks (1964)가 정규분포를 따르는 관측값들에서 평균의 변화에 대한 베이즈 검정을 고려한 것으로 시작하여, Yao (1984)에 의해 신호가 구분적 상수(piecewise constant) 함수인 경우의 베이지안 추정방법 등으로 이어져 왔다. 이들 연구를 토대로 이후 다중 변화점 검출을 위한 다양한 연구로 발전되었으며, Barry와 Hartigan (1993), Stephens (1994), Green (1995), Chib (1998), Fearnhead (2006) 등이 대표적이다. 특히, Green (1995)의 Reversible Jump Monte Carlo Markov chain(RJMCMC)는 Metropolis-Hasting 방법 (Hastings, 1970)을

[1]Biostatistics and Clinical Epidemiology Center, Research Institute for Future Medicine, Samsung Medical Center, Il-won ro 81, Gangnam, Seoul 135-710, Korea. E-mail: joungyoun@gmail.com

다양한 차원의 모수공간으로 확대시킴으로써 변화점의 개수가 불확실한 경우에도 효과적인 사후표본 검출을 가능하게 하였다.

일반적으로 다중 변화점 검출은 시계열 모형에서 분포가 달라지는 부분구간 추정을 목적으로 하기에, 각 부분 구간에서의 신호는 상수 평균을 따른다고 가정한다. 본 논문은 구간추정 뿐만 아니라, 각 구간내에서 시간의 흐름에 따른 신호 변화에 대해서도 추정하고자 한다. 특히 신호가 시간에 따라 구분적 선형함수(piecewise linear function)인 경우로 모형을 최소화함으로써 추정의 편리성 및 해석의 용이성을 제고하고자 한다. 제안된 방법은 시간의 부분구간과 신호의 선형 함수를 동시에 추정할 수 있다. 특히, 모든 변화점에서 신호의 연속성을 가정함으로써, 강물의 수위 변화나 주가지수 변화와 같이 변화가 특정 시점에서 불연속적으로 도약하기 보다는 점진적으로 변화하는 특징을 갖는 자료들을 분석하는데 유용한 모형을 제시하고자 한다. 아울러, 이를 위한 효과적인 사후표본 검출을 위한 MCMC 알고리즘을 개발하고자 한다.

논문의 순서는 다음과 같다. 2장에서는 구분적 선형함수 모형을 소개하고, 베이지안 추론을 위한 사전분포를 설명한다. 3장에서는 사후분포로부터의 MCMC 표본을 검출하기 위한 알고리즘을 개발한다. 4장에서는 모의 실험을 통한 알고리즘의 유용성 및 신뢰성을 평가하고, 5장에서는 제안된 방법을 나일강 최저수위 자료에 적용하여 본다.

2. 평균이 구분적 선형 모형을 따르는 회귀모형

2.1. 구분적 선형 모형

순차적 시점 $X = (x_1, \ldots, x_n)$에서 얻어진 관측값 $Y = (y_1, \ldots, y_n)$는 다음과 같은 신호와 오차의 합이라 가정하자.

$$\begin{aligned} y_i &= f(x_i) + e_i, \\ e_i &\sim N(0, \sigma^2), \\ f(x) &= \begin{cases} a_1 x + b_1, & \text{if } s_1 \le x \le s_2, \\ a_2 x + b_2, & \text{if } s_2 \le x \le s_3, \\ \quad\vdots & \quad\vdots \\ a_{k+1} x + b_{k+1}, & \text{if } s_{k+1} \le x \le s_{k+2}. \end{cases} \end{aligned}$$

위의 모형은 k개의 변화점이 존재할 때, 신호 $f(x)$를 $k+1$개의 구분적 선형함수로 나타내고 있다. $S = (s_1, \ldots, s_{k+2})$는 변화점의 개수가 k개일 때 $k+1$개의 선형함수를 정의해주는 축의 매듭점(knot)을 나타내는 모수로서, 다음을 만족한다.

$$s_1 = \min_{i=1,\ldots,n} \{x_i\} < s_2 < \cdots < s_{k+2} = \max_{i=1,\ldots,n} \{x_i\}.$$

이때, $s_2, s_3, \ldots, s_{k+1}$이 k개의 변화점에 해당한다.

본 논문에서는 $f(x)$가 도약(jump)이 없는 연속임을 가정한다. 즉, 고려되는 구분적 선형함수 $f(x)$는 $a_1(s_2) + b_1 = a_2(s_2) + b_2, \ldots, a_k(s_{k+1}) + b_k = a_{k+1}(s_{k+1}) + b_{k+1}$를 만족한다. 이러한 조건을 만족시키도록 위에서 정의된 $f(x)$를 다음과 같이 재모수화(reparameterization)한다.

$$f(x) = \frac{h_{j+1} - h_j}{s_{j+1} - s_j} x + \frac{h_j s_{j+1} - h_{j+1} s_j}{s_{j+1} - s_j}, \quad s_j \le x \le s_{j+1} (j = 1, \ldots, k+1). \tag{2.1}$$

이때 $H = (h_1, \ldots, h_{k+2})$는 각 매듭점에서의 신호 $f(x)$를 나타낸다. 즉, $h_1 = f(s_1), \ldots, h_{k+2} = f(s_{k+2})$이다. 본 연구에서는 변화점의 개수 k, 변화점의 위치 $s_2, \ldots, s_{k+1}$ 및 부분 구간에서 신호크기를 결정하는 $h_1, \ldots, h_{k+2}$를 추론하고자 한다. 다음 절에서 이 모수들의 베이지안 추정방법을 소개한다.

2.2. 베이지안 방법을 이용한 구분적 선형 모형 구축

본 절에서는 식 (2.1) 모형에서 베이지안 모수 추정 방법을 알아본다. 먼저 베이즈 추정을 하기 위해, 관심 모수 $\Omega = (K, S, H)$를 다음과 같이 정의한다. K는 변화점 개수, $S = (s_1, \ldots, s_{K+2})$는 K개의 변화점이 있을 때 정의된 매듭점들, $H = (h_1, \ldots, h_{K+2})$는 각 매듭점에서의 신호 $(f(s_1), f(s_2), \ldots, f(s_{K+2}))$를 나타낸다. 관심 모수 $\Omega = (K, S, H)$에 대한 식 (2.1) 모형의 우도 함수는 다음과 같이 정의될 수 있다.

$$L(K, S, H \mid X, Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - f(x_i)\right)^2\right\}. \tag{2.2}$$

베이지안 접근을 위해 각 모수에 대한 사전분포를 다음과 같이 정의한다. 먼저, 변화점 개수 K의 사전분포함수로 다음과 같은 절단 포아송(truncated Poisson) 분포를 가정한다.

$$\pi(K) = \frac{\lambda^K / K!}{\sum_{j=0}^{K_{\max}} \lambda^j / j!}. \tag{2.3}$$

본 연구에서는 $\lambda = 1$, K의 최대값 $K_{\max} = 20$을 가정한다.

변화점 개수 $K = k$일 때, 모수 $S = (s_1, \ldots, s_{k+2})$의 사전분포는 다음과 같이 Green (1995)이 제안한 방법을 따른다. 먼저, 구간 (s_1, s_{k+2})에서 균등분포를 따르는 $2k+1$개의 순서 통계량 $U = (u_{(1)}, u_{(2)}, \ldots, u_{(2k+1)})$을 고려한다. $j \in \{2, 3, \ldots, k+1\}$에 대하여 S의 j번째 원소 s_j는 $s_j = u_{(2j-2)}$로 정의된다. 따라서 모수 $S = (s_1, \ldots, s_{k+2})$의 사전분포는 다음과 같다.

$$\pi(S \mid K = k) = \frac{(2k+1)! \left\{\prod_{j=2}^{k+2}(s_j - s_{j-1})\right\}}{(s_{k+2} - s_1)^{2k+1}}. \tag{2.4}$$

모수 $H = (h_1, \ldots, h_{k+2})$의 각 원소들은 서로 독립이고 정규분포 $N(0, \sigma_h^2)$을 따른다고 가정한다. 따라서 모수 $H = (h_1, \ldots, h_{k+2})$의 사전분포는 다음과 같다.

$$\pi(H \mid K = k) = \prod_{j=1}^{k+2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma_h^2} h_j^2\right\}. \tag{2.5}$$

위에서 구한 우도함수 식 (2.2)와 사전분포 식 (2.3)– 식 (2.5)에 의하여 사후분포는 다음과 같이 계산된다.

$$\pi(K, S, H \mid X, Y) = \frac{\pi(K)\pi(S \mid K)\pi(H \mid K) \times L(K, S, H \mid X, Y)}{\int_{K,S,H} \pi(K)\pi(S \mid K)\pi(H \mid K) \times L(K, S, H \mid X, Y) dK dS dH}. \tag{2.6}$$

사후분포 식 (2.6)의 분모에 있는 적분이 어렵기 때문에, 이 함수로부터 모수의 추정지를 직접 계산하는 것은 매우 복잡하고 어려운 문제이다. 이와 같이 복잡한 형태의 사후분포를 가질 때, 일반적으로 MCMC 방법을 이용하여 조건부 사후분포로부터 표본을 추출하여 모수를 추정할 수 있다. MCMC 사슬의 t-번째 표본을 (K^t, S^t, H^t)라고 가정하자. $(t+1)$-번째 표본 $(K^{t+1}, S^{t+1}, H^{t+1})$의 후보로 제안된

(K^*, S^*, H^*)이 받아들여질 확률은 다음과 같다.

$$R = \min\left\{\frac{\pi(K^*, S^*, H^* \mid X, Y)}{\pi(K^t, S^t, H^t \mid X, Y)} \times \frac{q(K^t, S^t, H^t \mid K^*, S^*, H^*)}{q(K^*, S^*, H^* \mid K^t, S^t, H^t)}\right\}$$
$$= \min\left\{1, \frac{\pi(K^*)\pi(S^* \mid K^*)\pi(H^* \mid K^*)}{\pi(K^t)\pi(S^t \mid K^t)\pi(H^t \mid K^t)} \times \mathrm{LR} \times \mathrm{PR}\right\}.$$

LR은 우도비(Likelihood Ratio)로 $\mathrm{LR} = L(K^*, S^*, H^* \mid X, Y)/L(K^t, S^t, H^t \mid X, Y)$로 계산된다. PR은 제안비(Proposal Ratio)로 $\mathrm{PR} = q(K^t, S^t, H^t \mid K^*, S^*, H^*)/q(K^*, S^*, H^* \mid K^t, S^t, H^t)$이다. 이때 $q(\cdot \mid K^t, S^t, H^t)$는 제안 분포(proposal distribution)로, 모수 (K, S, H)가 t시점에서 표본 (K^t, S^t, H^t)를 가질 때, 새로운 표본을 제안할 수 있는 분포로, 여러가지 방법들이 가능하다. 다음장에서는 이러한 제안분포 및 MCMC 알고리즘에 대해 좀더 자세히 알아보도록 한다.

3. MCMC 알고리즘

본 논문에서 사용되는 제안분포 $q(\cdot \mid \cdot)$는 네가지 제안분포들의 조합으로서 나머지 모수의 표본들이 고정되었다는 가정하에 표본의 일부를 갱신하는 것으로 다음과 같다. 현재 변화점의 개수 $K = k$라고 할 때, (1) $S = (s_1, \ldots, s_{k+2})$ 갱신하기 (2) $H = (h_1, \ldots, h_{k+2})$ 갱신하기 (3) 변화점 개수를 $k+1$로 증가하기 (4) 변화점 개수를 $k-1$로 감소하기 이다. 이 네가지의 제안 방법들은 각각 0.3, 0.1, 0.3, 0.3의 확률로 임의로 뽑히게 되고, 이 확률은 MCMC 표본추출의 수렴을 원활하게 하기 위해 정해졌다. 이하에서 LR은 우도비, PR은 제안비를 나타낸다.

1. $S = (s_1, \ldots, s_{K+2})$**의 표본 제안하기.** MCMC 사슬의 t번째 표본의 변화점 개수와 매듭점을 각각 $K^t = k$, $S^t = (s_1^t, \ldots, s_{k+2}^t)$라고 하자. 집합 $\{2, 3, \ldots, k+1\}$에서 $1/k$의 확률로 한 원소 j를 추출한다. 추출된 j에 대해, 균등분포 $\mathrm{Uniform}(s_{j-1}^t, s_{j+1}^t)$로부터 임의로 u를 추출한다. j와 u가 주어졌을 때, 모수 S의 새로운 표본 $S^* = (s_1^*, \ldots, s_{k+2}^*)$은 다음과 같이 제안된다.

$$s_j^* = u,$$
$$s_l^* = s_l^t, \quad (l \neq j).$$

이때의 제안비(PR)는 1이므로, 전이확률은 다음과 같다.

$$R = \min\left\{1, \frac{\pi(S^* \mid K = k)}{\pi(S^t \mid K = k)} \times \mathrm{LR}\right\}$$
$$= \min\left\{1, \frac{(s_j^* - s_{j-1}^t)(s_{j+1}^t - s_j^*)}{(s_j^t - s_{j-1}^t)(s_{j+1}^t - s_j^t)} \times \mathrm{LR}\right\}$$
$$= \min\left\{1, \frac{(u - s_{j-1}^t)(s_{j+1}^t - u)}{(s_j^t - s_{j-1}^t)(s_{j+1}^t - s_j^t)} \times \mathrm{LR}\right\}.$$

R의 확률로, 제안된 표본 S^*를 받아들인다. 만약 받아들이면, 모수 S에 대한 $(t+1)$-번째 표본으로 $S^{t+1} = S^*$로 한다. 받아들여지지 않은 경우에는 $S^{t+1} = S^t$로 놓는다. 다른 모수들에 대한 $(t+1)$-번째 표본값으로는 확률 1로, $K^{t+1} = K^t$, $H^{t+1} = H^t$로 한다.

2. $H = (h_1, \ldots, h_{K+2})$**의 표본 제안하기.** MCMC 사슬의 t번째 표본의 변화점의 개수와 각 매듭점에서의 신호를 각각 $K^t = k$, $H^t = (h_1^t, \ldots, h_{k+2}^t)$라고 하자. 집합 $\{1, 2, \ldots, k+2\}$에서 $1/(k+2)$의 확

률로 임의의 원소 j를 추출한다. 균등분포 Uniform$(-0.5, 0.5)$에서 임의로 u를 추출한다. j와 u가 주어졌을 때, 모수 H의 새로운 표본 $H^* = (h_1^*, \ldots, h_{k+2}^*)$은 다음과 같이 제안된다.

$$\begin{aligned} h_j^* &= h_j^t + u, \\ h_l^t &= h_l^t, \quad (l \neq j). \end{aligned}$$

PR $= 1$이므로, 이때의 전이확률은 다음과 같다.

$$\begin{aligned} R &= \min\left\{1, \frac{\pi(H^* \mid K=k)}{\pi(H^t \mid K=k)} \times \mathrm{LR}\right\} \\ &= \min\left\{1, \exp\left\{\frac{1}{2\sigma^2}\left((h_j^t)^2 - (h_j^t+u)^2\right)\right\} \times \mathrm{LR}\right\}. \end{aligned}$$

R의 확률로, 제안된 H^*를 받아들인다. 만약 받아들이면, $H^{t+1} = H^*$로 놓고 그렇지 않은 경우에는 $H^{t+1} = H^t$로 놓는다. 다른 모수에 대해서는 확률 1로, $K^{t+1} = K^t$, $S^{t+1} = S^t$로 한다.

3. **변화점 개수를 증가하기.** MCMC 사슬의 t-번째 표본의 변환점의 개수, 매듭점, 신호의 크기가 각각 $K^t = k$, $H^t = (h_1^t, \ldots, h_{k+2}^t)$, $S^t = (s_1^t, \ldots, s_{k+2}^t)$라고 하자. 균등분포 Uniform$(s_1, s_{k+2})$에서 임의로 u_1를 추출한다. 뽑힌 u_1에 대해서 $s_j < u_1 < s_{j+1}$을 만족하는 j를 찾는다. u_1과 j가 주어졌을 때, 변화점의 개수가 $k+1$로 증가 된 모수 S의 새로운 표본 $S^* = (s_1^*, \ldots, s_{k+2}^*, s_{k+3}^*)$은 다음과 같이 제안된다.

$$\begin{aligned} s_l^* &= s_l^t, \quad (1 \leq l \leq j), \\ s_{j+1}^* &= u_1, \\ s_l^* &= s_{l-1}^t, \quad (j+2 \leq l \leq k+3). \end{aligned}$$

다음으로, 균등분포 Uniform$(0,1)$로부터 임의로 u_2를 추출한다. 위에서 추출된 j와 u_2가 주어졌을 때, 모수 H의 새로운 표본 $H^* = (h_1^*, \ldots, h_{k+2}^*, h_{k+3}^*)$은 다음과 같이 제안된다.

$$\begin{aligned} h_l^* &= h_l^t, \quad (1 \leq l \leq j), \\ h_{j+1}^* &= u_2 \times h_j^t + (1-u_2) \times h_{j+1}^t, \\ h_l^* &= h_{l-1}^t, \quad (j+2 \leq l \leq k+3). \end{aligned}$$

이때의 PR $= (D_{k+1}/(k+1))/B_k$로, $D_0 = 1$, $B_{K_{\max}} = 0$이고 그 외에는 $B_k = D_k = 1$이다. 따라서, 전이확률은 다음과 같이 계산된다.

$$\begin{aligned} R &= \min\left\{1, \frac{\pi(S^* \mid K=k+1)}{\pi(S^t \mid K=k)} \times \frac{\pi(H^* \mid K=k+1)}{\pi(H^t \mid K=k)} \times \frac{\pi(k+1)}{\pi(k)} \times \mathrm{PR} \times \mathrm{LR} \times J\right\} \\ &= \min\left\{1, \frac{(2k+3)(2k+2)\left(u_1 - s_j^t\right)\left(s_{j+1}^t - u_1\right)}{\left(s_{k+2}^t - s_1^t\right)^2\left(s_{j+1}^t - s_{j-1}^t\right)} \times \phi(h_{j+1}^* \mid 0, \sigma^2) \times \frac{\lambda}{k+1} \times \mathrm{PR} \times \mathrm{LR} \times J\right\}. \end{aligned}$$

$\phi(\cdot \mid \mu, \sigma^2)$은 평균 μ, 분산 σ^2을 갖는 정규분포의 확률 밀도함수를 나타낸다. 자코비안 $J = \mid h_j - h_{j+1} \mid$이다. R의 확률로, 제안된 표본을 받아들인다. 만약 받아들이면, $K^{t+1} = k+1$, $S^{t+1} = S^*$, $H^{t+1} = H^*$로 한다. 그렇지 않은 경우에는 $K^{t+1} = k$, $S^{t+1} = S^t$, $H^{t+1} = H^t$로 한다.

4. **변화점 개수를 감소하기.** MCMC 사슬의 t번째 표본의 변환점의 개수, 매듭점, 신호를 각각 $K^t = k$, $H^t = (h_1^t, \ldots, h_{k+2}^t)$, $S^t = (s_1^t, \ldots, s_{k+2}^t)$라고 하자. 집합 $\{1, 2, \ldots, k\}$의 원소 중 $1/k$의 확률로

원소 j를 추출한다. 변화점의 개수가 $k-1$로 감소된 모수 S의 새로운 표본 $S^* = (s_1^*, \ldots, s_{k+1}^*)$과 모수 H의 새로운 표본 $H^* = (h_1^*, \ldots, h_{k+1}^*)$은 다음과 같이 제안된다.

$$\begin{aligned} s_l^* &= s_l^t, \quad (1 \le l \le j), \\ s_l^* &= s_{l+1}^t, \quad (j+1 \le l \le k+1), \\ h_l^* &= h_l^t, \quad (1 \le l \le j), \\ h_l^* &= h_{l+1}^t, \quad (j+1 \le l \le k+1). \end{aligned}$$

이때의 $\mathrm{PR} = B_{k-1}/(D_k/k)$로, $D_0 = 1$, $B_{K_{\max}} = 0$이고 그 외에는 $B_k = D_k = 1$이다. 따라서 전이확률은 다음과 같다.

$$\begin{aligned} R =& \min\left\{1, \frac{\pi(S^* \mid K=k-1)}{\pi(S^t \mid K=l)} \times \frac{\pi(H^* \mid K=k-1)}{\pi(H^t \mid K=k)} \times \frac{\pi(k-1)}{\pi(k)} \times \mathrm{PR} \times \mathrm{LR} \times J\right\} \\ =& \min\left\{1, \frac{\left(s_{k+2}^t - s_1^t\right)^2 \left(s_{j+2}^t - s_j^t\right)}{(2k+1)(2k)\left(s_{j+2}^t - s_{j+1}^t\right)\left(s_{j+1}^t - s_j^t\right)} \times \sqrt{2\pi\sigma^2} \exp^{\frac{1}{2\sigma^2}\left(h_{j+1}^t\right)^2} \times \frac{k}{\lambda} \times \mathrm{PR} \times \mathrm{LR} \times J\right\}, \end{aligned}$$

이때 자코비안 $J = 1/\mid h_j - h_{j+2} \mid$이다. R의 확률로, 제안된 표본을 받아들인다. 만약 받아들이면, $(t+1)$-번째 표본으로 $K^{t+1} = k-1$, $S^{t+1} = S^*$, $H^{t+1} = H^*$로 한다. 그렇지 않은 경우에는 $K^{t+1} = k$, $S^{t+1} = S^t$, $H^{t+1} = H^t$로 한다.

4. 모의실험

제안된 방법을 검증하고자 다음과 같은 모의실험을 고려하였다. 먼저 n개의 순차적 시간 $X = (x_1, x_2, \ldots, x_n)$에서 서로 독립적으로 얻어진 관측값 $Y = (y_1, y_2, \ldots, y_n)$을 고려한다. $n = 100$이고 시간 $x_1 = 0.1, x_2 = 0.2, x_3 = 0.3, \ldots, x_{100} = 10$이라고 가정한다. 각각의 x_i에 대응되는 y_i는 다음 모형에 의해 생성된다.

$$y_i = f(x_i) + e_i, \quad e_i \sim N(0, 0.5^2),$$

이때 함수 $f(x)$에 대해 다음과 같은 두가지 경우를 고려한다.

$$f_1(x) = \begin{cases} 0.2x, & \text{if } 0 \le x \le 5, \\ 1, & \text{if } 5 \le x \le 10. \end{cases}$$

$$f_2(x) = \begin{cases} 0.5x, & \text{if } 0 \le x \le 2, \\ 0.75x - 0.5, & \text{if } 2 \le x \le 4, \\ 2.75x - 8.5, & \text{if } 4 \le x \le 6, \\ -x + 14, & \text{if } 6 \le x \le 8, \\ -3x + 30, & \text{if } 8 \le x \le 10. \end{cases}$$

첫 번째 함수 $f_1(x)$는 xy-좌표상 세개의 점 $\{(0,0), (5,1), (10,1)\}$을 지나고 점진적인 변화뒤에 상수함수가 유지된다. 두 번째 함수 $f_2(x)$는 xy-좌표상 여섯 개의 점 $\{(0,0), (2,1), (4,2.5), (6,8), (8,6), (10,0)\}$을 모두 지나고, 각 다섯 개의 구간에서 고유의 기울기를 갖는 구분적 선형함수로, 점진적인 변화와 급직적인 변화가 결합되어 있다.

각각의 함수로부터 생성된 자료에 대해, 제안된 알고리즘을 적용하여 MCMC 사후표본을 생성하였다. 각 MCMC 사슬은 길이가 600000이므로 600000개의 표본이 생성되었다. 처음 100000개의 표본은 초

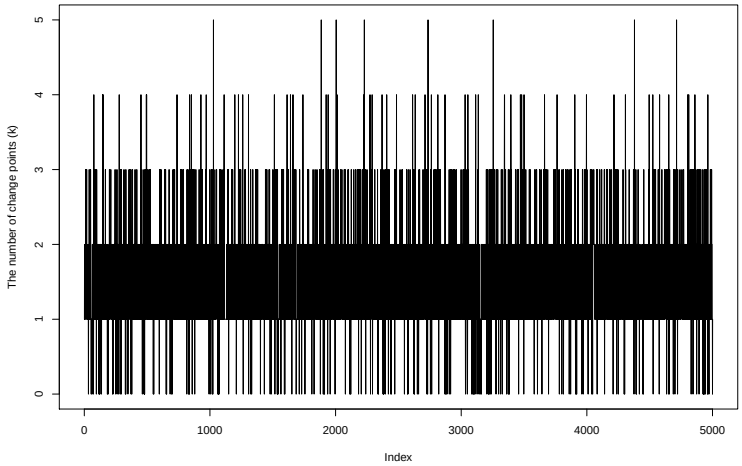

Figure 4.1. Trace plot of K over MCMC sampling order

Table 4.1. Posterior distribuition for K

K	0	1	2	3	4	5	합계
상대도수(%)	4.96	54.17	31.53	7.92	1.26	0.16	100

기값의 영향을 제거하기 위해 버려졌다. 남은 500000개의 사후표본간의 독립성을 위하여, 매 100번째 값만 추론에 사용한다. 따라서 하나의 MCMC 사슬로부터 추론에 사용되는 사후표본의 총 수는 5000개이다. MCMC사슬 생성을 위한 초기 변화점의 개수 K^0는 K의 사전분포로부터 임의로 정하였다. 초기 변화점의 개수 K^0에 대응되는 초기 $S^0 = (s_1^0, s_2^0, \ldots, s_{K^0+2}^0)$에서 $s_1^0 = x_1$, $s_{K^0+2}^0 = x_n$이다. S^0의 나머지 원소 $s_2^0, s_3^0, \ldots, s_{K^0+1}^0$의 값들은 균등분포 Uniform$(x_1, x_n)$로부터 임의로 얻은 K^0개의 순서통계량을 차례로 대응시켰다. K^0가 주어졌을 때 신호의 초기값 $H^0 = (h_1^0, h_2^0, \ldots, h_{K^0+2}^0)$의 모든 원소값들은 Y값들의 평균, 즉, $\sum_{i=1}^n y_i/n$으로 동일하게 하였다.

4.1. $f(x) = f_1(x)$인 경우

Figure 4.1은 MCMC 표본으로부터 얻어진 변화점 개수를 시간 흐름에 따라 연결한 것이다. 이 그림을 통해 MCMC가 잘 수렴하고 있음을 보여준다. Figure 4.2의 실선은 모의자료를 생성하는데 사용된 실제 함수 $f(x)$를 나타내고, 점들은 이때 생성되어진 모의자료를 나타낸다. 실선과 가장 가까운 점선은 MCMC로부터 얻어진 사후평균이다. 즉, 각각의 x_i에서의 $f(x_i)$의 추정값 $\hat{f}(x_i)$들의 사후평균을 연결한 것이다. 가장 밖의 두 점선은 95% 신뢰구간(Credible Region; C.R.)으로 각 x_i에서 얻은 $\hat{f}(x_i)$의 2.5% 분위수(quantile) 97.5% 분위수를 연결한 것이다.

Table 4.1은 MCMC 표본에서 얻어진 변화점 개수 K의 사후분포를 보여준다. 변화점 개수가 1인 MCMC 표본이 전체의 54.16%를 차지하며 최대 사후확률을 갖는데, 이는 실제 변화점의 개수와 일치한다. Figure 4.3는 변화점 개수가 1인 MCMC 부분 표본들에 대하여, 그 변화점의 위치와 각 매듭에서의 신호 (h_1, h_2, h_3)의 분포를 상자그림으로 보여준다 $K = 1$인 MCMC 표본에서 얻어진 s_2의 평균은 3.24 (표준편차 1.71)로 실제 변화점의 위치 5보다 작게 추정되었다. 하지만, 이에 해당하는 $f(x)$값인 h_2의 사후평균은 0.83 (표준편차 0.25)으로 참값에 근사하게 추정되었다.

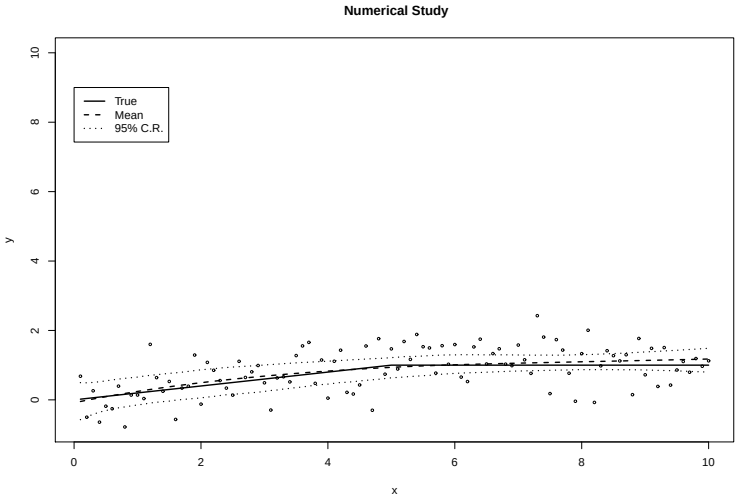

Figure 4.2. Posterior estimates of f_1

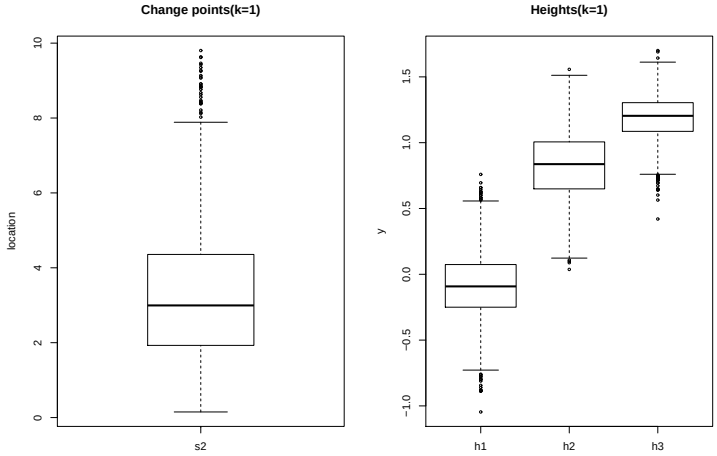

Figure 4.3. Boxplots of s_2, h_1, h_2 and h_3 from the posterior sample with $K = 1$

4.2. $f(x) = f_2(x)$인 경우

Figure 4.4은 MCMC 표본으로부터 얻어진 변화점 개수를 시간 흐름에 따라 연결한 것이다. 이 그림을 통해 MCMC가 잘 수렴하고 있음을 보여준다. Figure 4.5는 Figure 4.2와 마찬가지로 실제 함수, MCMC 표본으로부터 얻어진 사후평균 함수 및 95% CR을 보여준다. Table 4.2는 MCMC 표본에서 얻어진 변화점 개수 K의 사후분포이다. 모의실험에서 가정된 K의 참값 $K = 4$인 사후확률이 47.91%로 가장 높았다. Figure 4.6의 좌측은 $K = 4$인 MCMC 사후표본에서 변화점 (s_2, s_3, s_4, s_5)의 분포를 상자그림으로 보여준다. 우측은 $K = 4$일 때 각 매듭점에서의 신호 $(h_1, h_2, \ldots, h_6)$의 분포를 상자그림으로 보여준다. Table 4.3과 Table 4.4는 $K = 4$인 MCMC 표본에서의 변화점 및 신호에 대한 사후통계량을 보여준다.

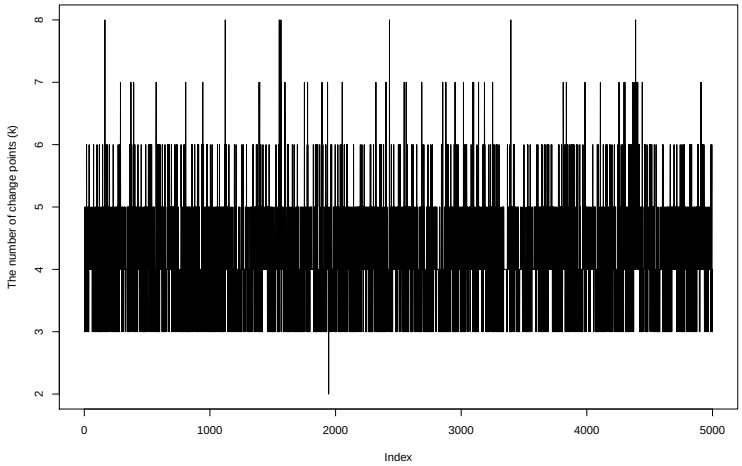

Figure 4.4. Trace plot of K over MCMC sampling order

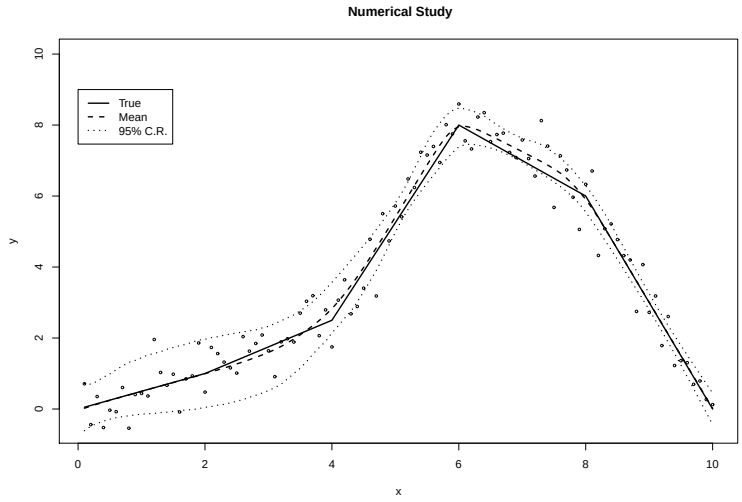

Figure 4.5. Posterior estimates of f_2

Table 4.2. Posterior distribuition for K

K	2	3	4	5	6	7	8	합계
상대도수(%)	0.02	23.42	47.91	21.58	6.06	0.88	0.14	100

5. 실제자료

변화점 추정에 자주 사용되는 자료인 나일강 수위 자료에 제안된 방법을 적용하여 분석하였다. 나일강 자료는 622년부터 1284년까지 총 633년동안 나일강의 연도별 최저수위를 측정한 663개의 관측치로 구성되어 있다. Beran과 Terrin (1996)은 장기억 과정에서 변화점의 유무를 판단할 수 있는 검정통계량

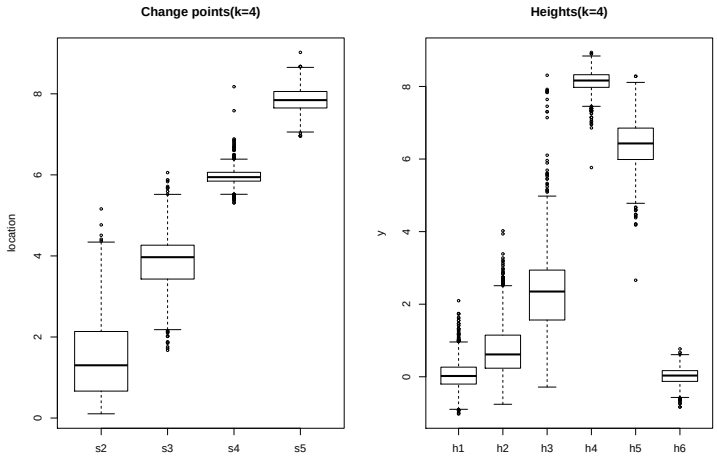

Figure 4.6. Boxplots of $s_2, \ldots, s_5, h_1, \ldots, h_6$ from the posterior sample with $K = 4$

Table 4.3. Summary of $s_2, \ldots, s_5$ from the posterior sample with $K = 4$

	s_2	s_3	s_4	s_5
평균	1.48	3.86	5.96	7.85
중앙값	1.30	3.97	5.94	7.84
표준편차	0.99	0.60	0.20	0.27

Table 4.4. Summary of $h_1, \ldots, h_6$ from the posterior sample with $K = 4$

	h_1	h_2	h_3	h_4	h_5	h_6
평균	0.04	0.75	2.30	8.14	6.41	0.02
중앙값	0.02	0.62	2.35	8.16	6.43	0.03
표준편차	0.38	0.69	1.08	0.27	0.60	0.23

Table 5.1. Posterior distribuition for K

K	5	6	7	8	9	10	11	12	합계
상대빈도(%)	1.64	18.46	45.23	25.51	7.80	1.10	0.22	0.04	100

을 제안한 뒤, 나일강 자료에 적용하여 722년을 전후로 나일강 최저수위가 유의하게 달라졌다고 보고했다. 이때 변화점의 개수는 한 개로 고정되었다. Kim 등 (2009)는 장기억 시계열 모형에서 베이지안 변화점 검출을 시도하였다. 이 방법은 임의의 변화점의 개수가 고정되었을 때, 자료가 각각의 부분 구간에서 자기회귀부분누적이동평균(autoregressive fractional integrated moving average; ARFIMA)모형을 따른다고 가정하였다. 이때 ARFIMA 모형의 모수에 대해 사전확률 분포를 가정하며 베이지안 추정을 하였다. 변화점을 하나로 가정한 경우 그 위치가 시점 672년–772년 사이이고, 변화점을 두개로 가정한 경우 그 위치가 각각 672년–772년 사이와 1022년–1122년 사이로 추정하였다. 본 논문에서 제안된 MCMC 방법을 나일강 수위 자료에 적용하고 그 결과를 이전의 연구들과 비교하고자 한다.

MCMC 생성 방법은 모의실험과 같은 방법으로 하였다. Figure 5.1은 각 MCMC 표본에서 얻어진 변화점 개수로 MCMC가 잘 수렴하고 있음을 보여준다. Figure 5.2는 MCMC에 의해 추정된 시간변화에 따른 나일강 수위 변화이다. 점은 관측값을 나타내고, 실선은 MCMC로부터 얻어진 각각의 x_i에서의

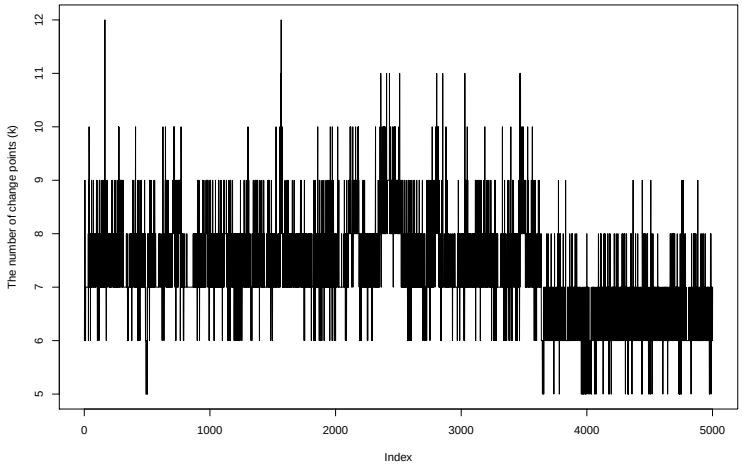

Figure 5.1. Trace plot of K over MCMC sampling order

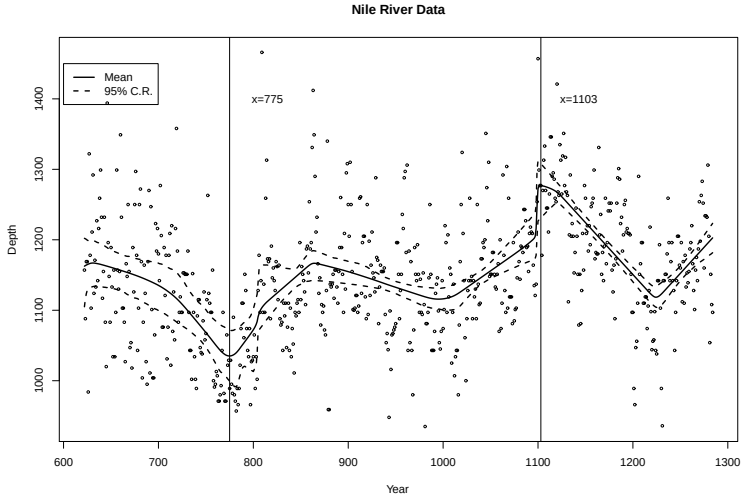

Figure 5.2. Posterior estimates of Nile river depth change

$f(x_i)$ 사후평균을 나타낸다. 점선은 95% 신뢰구간(95% C.R.)으로 각 x_i에서 얻어진 $\widehat{f(x_i)}$ 추정값들의 2.5% 분위수와 97.5% 분위수를 연결한 것이다. 제안된 방법에 의한 분석 결과, 나일강의 수위는 622년부터 최저수위가 점차적으로 줄어 775년에 최저값을 갖는다. 이후 최저수위가 점차적으로 증가 및 감소를 반복하다, 1103년에 최고값을 갖는다. 1103년 이후 최저수위는 다시 이전의 증가속도와 비슷한 속도로 감소하는 경향을 보인다. 최저 수위를 갖는 시점과 최고 수위를 갖는 시점의 추정값은 김주원 등 (2009)에서 변화점 개수를 2개로 가정했을 때 추정된 변화점 위치와 유사하다. Table 5.1은 MCMC 표본에서 얻어진 변화점 개수 K의 사후분포로서, $K = 7$일 사후확률이 45.23%로 가장 높았다. Figure 5.3의 좌측은 사후표본들 중에서, $K = 7$일 때의 각각의 변화점 $s_2, s_3, \ldots, s_7, s_8$의 분포를 상자그림으로 보여준다. 우측은 $K = 7$일 때 각각의 높이 $h_1, h_2, \ldots, h_8, h_9$에 대한 상자그림이다. Table 5.2와

Table 5.2. Summary of $s_2, \ldots, s_8$ from the posterior sample with $K = 7$

	s_2	s_3	s_4	s_5	s_6	s_7	s_8
평균	676.75	765.82	833.22	970.69	1079.55	1104.87	1224.28
중앙값	677.22	774.43	859.28	992.88	1097.40	1100.10	1224.20
표준편차	37.99	32.76	37.95	53.74	34.44	8.43	3.43

Table 5.3. Summary of $h_1, \ldots, h_9$ from the posterior sample with $K = 7$

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9
평균	1159.35	1155.80	1049.69	1143.10	1126.28	1172.15	1291.47	1114.66	1203.02
중앙값	1163.12	1157.35	1032.82	1169.05	1116.41	1185.72	1295.61	1114.40	1203.67
표준편차	25.05	30.46	50.07	57.85	25.79	34.86	15.43	6.97	11.14

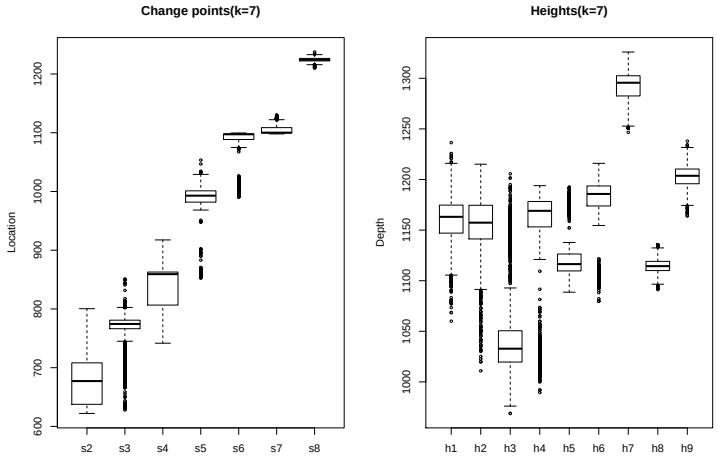

Figure 5.3. Boxplots of $s_2, \ldots, s_8, h_1, \ldots, h_9$ from the posterior sample with $K = 7$

Table 5.3은 Figure 5.3에 대응되는 수치결과이다. $K = 7$일 때, 최저 수위를 갖는 시점과 최고 수위를 갖는 시점의 사후평균은 $\widehat{s_3} = 765.82$, $\widehat{s_7} = 1104.87$이고, 이 추정값들은 Kim 등 (2009)에서 변화점이 두개일 때의 변화시점과 일치한다.

6. 결론

변화점 검출은 순차적으로 누적된 자료들의 분포 추론에서 중요한 역할을 한다. 특히 장기간에 걸친 대용량 자료들이 더욱 많이 생산됨에 따라 이러한 자료에서 변화점 검출은 매우 중요한 관심사가 되고 있다. 변화점 개수의 변화에 따른 모수 공간의 차원변화라는 복잡한 상황에서의 효율적 추론을 위해서 다양한 베이지안 방법들이 연구되어 왔다. 이 논문에서는 평균이 구분적 선형 함수이고, 관측값들이 서로 독립인 모형에서 변화점을 효율적으로 검출하는 베이지안 방법을 소개하였다. 기존에 많이 고려되어 온 부분적 상수 함수와 달리, 신호가 시간의 흐름에 따라 점진적으로 변화하는 경우에 적합한 모형을 고려하기 위해, 연속성 가정까지 만족시키는 베이지안 모형을 개발하였다. Reversible Jump MCMC를 바탕으로 개발된 알고리즘은 다양한 차원의 모수 공간을 효율적으로 탐색할 수 있었으며, 모의실험을 통해 제안된 베이지안 모형 및 방법의 정확성 및 효율성을 확인하였다. 또한 나일강 최저수위 자료분석에서

기존 연구결과와 유사하면서도 점진적 수위변화 분석이 가능하다는 것을 보였다. 나일강 자료 분석 결과에서 기존의 연구보다 많은 개수의 변화점이 검출되었다. 추후 연구에서 변화점 개수에 대한 사전분포 조절, 또는 각 변화점들의 유무에 대한 베이지안 검정을 추가함으로써 변화점 개수 추론이 개선될 수 있을 것으로 기대한다.

부록

본 절에서는 식 (2.1)에서 정의된 함수가 각 마디$(s_j,\ j=2,\ldots,K+1)$에서 연속임을 증명한다.

구간 $[s_j, s_{j+1}]$에서의 식 (2.1)은 다음과 같다.

$$\begin{aligned} f(x) &= a_j x + b_j \\ &= \frac{h_{j+1}-h_j}{s_{j+1}-s_j}x + \frac{h_j s_{j+1} - h_{j+1}s_j}{s_{j+1}-s_j}. \end{aligned}$$

식 (2.1)의 연속성을 증명하는 것은 $a_j = (h_{j+1}-h_j)/(s_{j+1}-s_j)$일 때, $a_j s_{j+1}+b_j = a_{j+1}s_{j+1}+b_{j+1}$을 만족하는 b_j가 다음과 같음을 보이는 것으로 충분하다.

$$b_j = \frac{h_j s_{j+1} - h_{j+1}s_j}{s_{j+1}-s_j}.$$

1. 먼저, $j=1$일 때, b_1을 구해본다.

$$\begin{aligned} h_1 &= \frac{h_2-h_1}{s_2-s_1}s_1 + b_1, \\ b_1 &= h_1 - \frac{h_2-h_1}{s_2-s_1}s_1 \\ &= \frac{h_1(s_2-s_1)-(h_2-h_1)s_1}{s_2-s_1} \\ &= \frac{h_1 s_2 - h_2 s_1}{s_2-s_1}. \end{aligned}$$

2. 다음으로 $j=2$일 때, b_2를 구해본다.

$$\frac{h_2-h_1}{s_2-s_1}s_2 + b_1 = \frac{h_3-h_2}{s_3-s_2}s_2 + b_2,$$

$$\begin{aligned} b_2 &= \frac{h_2-h_1}{s_2-s_1}s_2 + b_1 - \frac{h_3-h_2}{s_3-s_2}s_2 \\ &= \frac{h_2-h_1}{s_2-s_1}s_2 + \frac{h_1 s_2 - h_2 s_1}{s_2-s_1} - \frac{h_3-h_2}{s_3-s_2}s_2 \\ &= h_2 - \frac{h_3-h_2}{s_3-s_2}s_2 \\ &= \frac{h_2 s_3 - h_3 s_2}{s_3-s_2}. \end{aligned}$$

3. $j=x$일 때,

$$b_x = \frac{h_x s_{x+1} - h_{x+1}s_x}{s_{x+1}-s_x}$$

를 만족한다고 가정하자. $j = x+1$일 때의 b_{x+1}을 구해본다.

$$\frac{h_{x+2}-h_{x+1}}{s_{x+2}-s_{x+1}}s_{x+1} + b_{x+1} = \frac{h_{x+1}-h_x}{s_{x+1}-s_x}s_{x+1} + b_x,$$

$$\begin{aligned} b_{x+1} &= \frac{h_{x+1}-h_x}{s_{x+1}-s_x}s_{x+1} + b_x - \frac{h_{x+2}-h_{x+1}}{s_{x+2}-s_{x+1}}s_{x+1} \\ &= \frac{h_{x+1}-h_x}{s_{x+1}-s_x}s_{x+1} + \frac{h_x s_{x+1} - h_{x+1}s_x}{s_{x+1}-s_x} - \frac{h_{x+2}-h_{x+1}}{s_{x+2}-s_{x+1}}s_{x+1} \\ &= h_{x+1} - \frac{h_{x+2}-h_{x+1}}{s_{x+2}-s_{x+1}}s_{x+1} \\ &= \frac{h_{x+1}s_{x+2} - h_{x+2}s_{x+1}}{s_{x+2}-s_{x+1}}. \end{aligned}$$

4. 따라서 모든 $j \geq 1$에 대하여

$$b_j = \frac{h_j s_{j+1} - h_{j+1}s_j}{s_{j+1}-s_j}$$

임을 증명하였다.

References

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

Beran, J. and Terrin, N. (1996). Testing for a change of the long-memory parameter, *Biometrika*, **83**, 627–638.

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**, 949–1417.

Chib, S. (1998). Estimation and comparison of multiple change-point models, *Journal of Econometrics*, **86**, 211–241.

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

Green, P. J. (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.

Kim, J. W., Cho, S. and Yeo, I. K. (2009). A fast Bayesian detection of change points long-memory processes, *The Korean journal of applied statistics*, **22**, 735–744.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **43**, 159–178.

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches, *The Annals of Statistics*, **12**, 1151–1596.

구분적 선형함수에서의 베이지안 변화점 추출

김정연[a,1]

[a]삼성 서울병원, 의생명 정보센터

(2014년 4월 30일 접수, 2014년 7월 31일 수정, 2014년 8월 14일 채택)

요 약

본 연구는 시간의 순서에 따라 순차적으로 발생한 신호 자료에 있어서, 변화점 검출을 위한 베이지안 방법을 개발하고자 한다. 특히, Reversible Jump MCMC를 이용하여, 차원이 정해지지 않은 모수 공간을 탐색할 수 있는 효율적인 베이지안 추론 모형을 개발한다. 신호가 각 구간에서 선형함수인 경우에 대한 모형과 이해가 용이한 모형을 제안하고, 추정을 위해 고유의 MCMC알고리즘을 개발하였다. 제안된 방법을 모의실험 자료에 적용함으로써 그 정확성 및 효율성을 검증하였고, 실제 자료에도 적용하여 보았다.

주요용어: 변화점, 베이지안 추론, MCMC, Reversible Jump.

[1](135-710) 서울시 강남구 일원로 81, 삼성 서울병원, 미래의학 연구원, 의생명 정보센터.
E-mail: joungyoun.kim@gmail.com