응용통계연구 (2009)
22(4), 735–744

장기억 과정에서 빠른 베이지안 변화점검출

김주원[1] · 조신섭[2] · 여인권[3]

[1]서울대학교 입학처, [2]서울대학교 통계학과, [3]숙명여자대학교 통계학과

(2009년 4월 접수, 2009년 7월 채택)

요 약

이 논문에서는 장기억 과정에서의 변화점을 빨리 검출하는 베이지안 추론방법에 대해 알아본다. 장기억 과정에서의 베이지안 추정은 장기억 모수값에 따라 전체 자료에 대한 부분차분을 계산해야 하기 때문에 수행시간이 많이 걸린다는 문제가 있다. 이 논문에서는 이러한 문제를 해결하고자 장기억 모수공간을 그룹화하여 순서형으로 범주화시킨 후 설명력이 가장 높은 범주의 대표값을 선택하게 하였다. 이 방법은 초기단계에서 범주의 대표값에 대해 한번씩만 부분차분을 계산하면 되기 때문에, 매번 계산해야 하는 추정하는 방법보다, 특히 시계열자료의 수가 많은 경우, 상대적으로 빠른 베인지안 추론이 가능하다. 또한 장기억 모수공간이 $(0, 0.5]$이기 때문에 모수공간을 적절하게 그룹화한다면 장기억 모수를 선택하는 것이 모수를 추정하는 것에 비해 큰 차이가 없다. 이 논문에서는 나일강 수위자료 실증분석을 통해 제안된 방법의 타당성을 확인해본다.

주요용어: 디리슈레분포, 변화점검출, 자기회귀부분누적이동평균모형.

1. 서론

시계열은 특정기간동안 시간에 따라 관측된 자료이며 대부분 자료들간에 의존성이 존재한다. 이러한 의존관계를 모형화하고 모형화된 관계식을 이용하여 미래를 예측하는 것이 시계열분석의 주요 목적 중 하나이다. 시계열분석에서는 외부상황에 따라 자료들간의 관계에 변화가 생기는 일이 종종 발생하는데, 이 변화는 그 발생시점 이후에 지속적으로 영향을 주는 경향이 있기 때문에 변화가 일어난 정확한 위치와 변화량에 대한 분석이 매우 중요하다. 이러한 이유로 시계열분석에서 변화점 검출에 대한 연구가 많이 이루어지고 있으나 대부분 ARMA 모형에서의 평균이나 분산에서의 변화를 중심으로 연구되었으며 빈도론자적 방법와 베이지안 방법 모두 비슷한 상황이다.

장기억 과정은 현재의 값이 미래의 값에 장기간 영향을 주는 특징을 가지는 시계열이다. 과학 기술이 발전함에 따라 주어진 현상에 대해 좀더 정밀한 분석을 가능케 하는 데이터가 관찰되고 있다. 시계열 분야에서는 시간상으로 세밀하게 관찰 가능하게 된 데이터라고 할 수 있다. 예를 들어, 환율시장에서 환율의 변화가 과거에는 5분 단위로 기록되었으나 현재에는 10초 단위로 기록된다거나 대기내의 오존 측정이 과거 5시간 단위로 관찰 가능하였으나 과학 기술의 발달로 10분 단위로 가능해 지고 있는 경우 등이다. 이렇게 세밀하게 측정된 자료의 일반적인 특징은, 비록 시간적으로는 얼마 안되지만, 먼 시점의 관찰값들로부터 지속적으로 영향을 받는 강한 의존 혹은 장기 의존(strong or long range dependency)이 존재한다. 이러한 특성을 가지는 시계열자료는 기술이 계속 발전하면서 더욱더 많이 생산될 것이기 때문에 이들 분야에서의 변화점 검출은 매우 중요한 관심사가 되고 있다.

[3]교신저자: (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 부교수.
Email: inkwon@sm.ac.kr

장기억 과정에서의 변화점 검출에 대한 연구는 계산과정의 복잡성 때문에 다른 시계열모형에서의 변화점 연구보다 많지 않다. Hidalgo와 Robinson (1996)와 Wright (1998)는 장기억 오차항을 수반하는 선형모형에서 회귀모수의 단일 변화점 검출에 대해 연구하였으며 Kuan과 Hsu (1998)는 장기억 과정에서 평균의 단일 변화점에 대한 연구를 하였으면 이들 연구들은 빈도론자적 관점에 의해 연구되었다. 일반적인 모형에서의 베이지안 변화점 검출에 대한 연구는 Green (1995)과 Chib (1998) 등에 의해 연구되었으나 이를 장기억 과정에 적용하는데 있어 수행시간이 많이 걸리는 문제가 발생한다. Liu와 Kao (1999), Ray와 Tsay (2002), Ko와 Vannucci (2006) 등이 장기억 과정을 기반으로 한 변화점 검출에 대해 연구하였다. 이 논문에서는 장기억 과정에서 변화점의 위치와 변화량을 빠르게 검출하는 베이지안 방법을 소개한다.

2. 장기억 과정에서의 변화점 검출

2.1. ARFIMA 모형

확률과정 $\{X_t\}$는 k-차 자기상관함수가 $\rho(k)$인 정상과정(stationary process)이라고 하자. 이 확률과정에 대해, 다음과 같은 성질을 만족하는 상수 $\alpha \in (0,1)$와 $C_\rho > 0$가 존재하면,

$$\lim_{k\to\infty} \frac{\rho(k)}{C_\rho k^{-\alpha}} = 1$$

확률과정 $\{X_t\}$를 장기억(long-memory)을 가지는 정상과정이라고 부른다. 장기억 과정은 첨단기술의 발전으로 실시간으로 관측되어지고 있는 금융, 환경, 정보통신, 수문학(hydrology) 등의 분야에서 흔히 발생한다. 즉 기존의 자료에서는 발생하지 않았던 것이 관측시간 간격이 좁아지면서 관측값 간의 종속성이 매우 높아지는 성질을 가지게 되었다. 일반적으로, 장기억 확률과정은 자기회귀부분누적이동평균(autoregressive fractional integrated moving average: ARFIMA) 모형을 이용하여 분석하고 있다. 이 모형은 ARIMA 모형에서 차분의 값을 자연수 대신 실수을 사용한 것으로, $-0.5 < d < 0.5$에 대해 다음과 같은 구조를 가질 때 ARFIMA(p,d,q)를 따른다고 한다.

$$\begin{aligned} \phi(B)(1-B)^d(X_t-\mu) &= \psi(B)\varepsilon_t, \qquad \varepsilon_t \sim N(0,\sigma^2) \\ \phi(B) &= 1-\phi_1 B - \cdots - \phi_p B^p \\ \psi(B) &= 1-\psi_1 B - \cdots - \psi_p B^q \end{aligned}$$

부분차분모수 d는 장기간동안의 시계열 종속성을 결정하는 반면 AR 모수 p와 MA 모수 q는 일반적으로 시계열과 오차항에 대해 단기간동안의 종속성을 다양하게 모형화하는데 사용된다. 장기억 과정은 스펙트럴 밀도를 이용하여 모형화할 수 있는데 이 경우 $-0.5 < d < 0$에 대해 스펙트럴 밀도가 0이 되고 자기상관의 합이 0이 되는 문제가 발생하기 때문에 일반적인 분석에서는 ARFIMA 모형에서 d의 모수공간은 $0 \le d \le 0.5$로 정한다. 이 논문에서는 부분차분모수 d가 시간에 따라 변하는 경우, 이를 빨리 검출할 수 있는 베이지안 방법을 소개한다. 설명을 간단하게 하기 위해 이 논문에서는 X_t는 평균이 μ이고 분산이 σ^2인 ARFIMA$(0,d,0)$을 따르고 시간에 따라 d의 값만 변하는 것으로 가정한다.

가능도함수를 계산하기 위해 ARFIMA 모형을 AR형태로 표시하면 다음과 같이 쓸 수 있는데

$$\sum_{k=0}^{\infty} \pi_k(d)(X_t-\mu) = \varepsilon_t,$$

여기서 $\pi_k(d) = B^k\Gamma(k-d)/\{\Gamma(k+1)\Gamma(-d)\}$로 다음의 관계식을 이용하여 구할 수 있다.

$$(1-B)^d = \sum_{k=0}^{\infty}\binom{d}{k}(-B)^k = \sum_{k=0}^{\infty}\frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}B^k = 1 - \frac{d}{\Gamma(1-d)}\sum_{k=1}^{\infty}\frac{\Gamma(k-d)}{\Gamma(k+1)}B^k.$$

일반적으로 장기억 과정의 시계열자료는 관측값의 수가 많기 때문에 위의 정의대로 가능도 함수를 계산하면 상당한 시간이 걸린다. 이러한 문제점을 해결하기 위해 Haslett와 Raftery (1989)는 가능도함수의 근사값을 구하는 방법을 소개하였다. 그러나 MCMC에 의해 생성된 각각의 d에 대해 가능도함수를 계산해야 하기 때문에 이 근사방법은 베이지안 추론에서는 여전히 부담스러운 작업이다.

2.2. 베이지안 변화점 검출 방법

변화점에 대한 베이지안 추론은 Green (1995)과 Chib (1998) 등에 의해 연구되었다. 이들 방법은 일반적인 모형에서 사용될 수 있는 방법으로 활용범위는 높으나 장기억 과정과 같이 특수한 상황에서는 계산이 효율적이지 못하다. 장기억 과정에서의 베이지안 변화점 검출에 대한 연구는 그 중요성에 비해 계산과정의 어려움으로 Ray와 Tsay (2002)와 Ko와 Vannucci (2006)의 연구 외에는 많지 않다. Ray와 Tsay (2002)는 AR 모형에서의 McCulloch와 Tsay (1993)의 결과를 장기억 과정에 확장시켜 d에 대한 변화점 검출방법을 제안하였다. 이들 연구에서는, t시점에서의 부분차분 모수를 d_t라고 했을 때,

$$d_t = d_0 + \sum_{j=1}^{t}\delta_j\beta_j = d_{t-1} + \delta_t\beta_t$$

라고 가정하였다. 여기서 δ_t는 변화가 있을 성공확률이 $P(\delta_t = 1) = \epsilon$인 베르누이 확률변수이고 β_t는 도약의 크기에 대한 수열로 분포는 알고 있다고 가정하였다. 이들은 전체자료를 몇 개의 블록으로 나누고 블록간에 변화가 있는지와 있다면 어느 정도가 되는지를 격자망 깁스 표집기(griddy Gibbs sampler)를 이용한 MCMC 방법으로 추론하였으며 Chan과 Palma (1998)의 시간종속 칼만 필터방법을 이용하여 가능도함수를 계산하였다.

이 논문에서는 부분차분모수의 공간이 $(0, 0.5]$와 같이 짧은 구간으로 이루어져 있다는 것에 주목하였다. 이 공간을 m개의 구간으로 나누고 각 구간을 대표하는 m개의 값 $\mathcal{C} = \{c_1, \ldots, c_m\}$만 d의 값으로 사용한다면 베이지안 추론 초기단계에서 m개의 값에 대해서만 가능도를 한번만 계산하고 이후에는 선택된 d에 대해 이미 계산된 가능도함수값을 참조만 하면 되기 때문에 빠른 속도로 MCMC를 수행할 수 있다. 즉, 기존 연구에서는 d의 가능도함수를 $(0, 0.5]$ 상에서의 연속형 분포로 추정하였으나 이 논문에서 m개의 범주를 가지는 범주형 분포로 가정하고 사후분포를 계산하는 방법에 대해 알아본다. 변화점이 여러 개 있는 경우, 그 위치와 d의 사후분포를 효율적으로 추정하기 위해 이 논문에서는 두 단계로 이루어진 계층구조 분석방법을 제안한다. 평균 μ와 σ^2에 대한 사전분포는 각각 정규분포와 역감마분포를 가정하며 이에 대응하는 사후분포는 베이지안 분석에서 이미 많이 언급되었기 때문에 설명에서 생략한다.

2.2.1. 블록탐색(Block search) 이 탐색과정에서는 시계열자료를 다음과 같이 길이가 b인 J개의 블록으로 나눈다.

$$B_j = \{X_{(j-1)b+1}, \ldots, X_{jb}\}, \quad j = 1, \ldots, J$$

이 블록안에서는 시계열특성이 변하지 않는다고 가정한다. 즉 블록안에서는 같은 d값이 적용된다고 본다. 모수공간 $\mathcal{C}$에서 l번째 모수값을 j번째 블록에서 선택되었을 때, 그 값을 c_l^j이라고 하면 j번째 블록

표 2.1. 블록과 선택모수 간의 분할표

블록	선택된 모수			
	c_1	c_2	$\cdots$	c_m
B_1	$f_{1,1}$	$f_{1,2}$	$\cdots$	$f_{1,m}$
B_2	$f_{2,1}$	$f_{2,2}$	$\cdots$	$f_{2,m}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
B_J	$f_{J,1}$	$f_{J,2}$	$\cdots$	$f_{J,m}$

에서 모수들의 가능도함수는 다음과 같이 쓸 수 있다.

$$f(c_l^j, \mu, \sigma^2 | B_j) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^b \exp\left[-\frac{1}{2\sigma^2}\sum_{t=(j-1)b+1}^{jb}\left\{(1-B)^{c_l^j}(x_t-\mu)\right\}^2\right]$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^b \exp\left[-\frac{1}{2\sigma^2}\sum_{t=(j-1)b+1}^{jb}\left\{\sum_{k=1}^{t}\pi_k(c_l^j)(x_t-\mu)\right\}^2\right],$$

여기서 지수부분에 있는 $\sum_{k=1}^{t}\pi_k(c_l^j)(x_t-\mu) = \sum_{k=1}^{t}\pi_k(c_l^j)x_t - \mu\sum_{k=1}^{t}\pi_k(c_l^j)$가 되며 시계열 가중합 $\sum_{k=1}^{t}\pi_k(c_l^j)x_t$와 상수합 $\sum_{k=1}^{t}\pi_k(c_l^j)$을 계산하는데 μ가 영향을 주지 않는다. 즉 각 블록에서 $\mathcal{C}$의 값들에 대해 시계열 가중합을 구하여 보관하면 이후의 작업에서는 이 합들을 이용하여 가능도 함수를 구할 수 있어 계산속도도 매우 빨라진다.

$\vec{L}_j$를 j번째 블록에서 선택되어진 모수의 위치를 나타내는 벡터로 모수공간 $\mathcal{C}$에서 l번째 모수값을 선택했다면 l번째 원소의값만 1이고 나머지는 0인 m-차원 벡터라고 하자. 이 논문에서는 각 블록에 대해 설명력이 높은 d를 $\mathcal{C}$에서 선택하기 때문에 이를 모형화하기 위해 j번째 블록의 사전분포를 다음과 같이 설정한다.

$$\vec{P}_j \sim \text{Dirichlet}(\vec{\delta}_j), \quad \vec{\delta}_j = (\delta_{j1}, \ldots, \delta_{jm})$$
$$\vec{L}_j \sim \text{Multinomial}(1, \vec{P}_j), \quad \vec{P}_j = (P_{j1}, \ldots, P_{jm})^T,$$

여기서 $\vec{\delta}_j$는 초월모수로 주어진 값으로 가정한다. 이 사전분포와 가능도함수를 이용하여 L_j에 대한 사후분포를 유도하면 다음과 같다.

$$\vec{P}_j^* \sim \text{Dirichlet}(\vec{\delta}_j^*), \quad \vec{L}_j \sim \text{Multinomial}(1, \vec{P}_j^*),$$

여기서 $\vec{\delta}_j^*$의 l번째 원소는 전단계에서 선택된 d가 c_j^l이면 $\delta_{jl}+1$, 아니면 δ_{jl}이 되고 $\vec{P}_j^*$의 l번째 원소는 다음과 같이 계산된다.

$$P_{jl}^* = \frac{f(c_l^j, \mu, \sigma^2|B_j)P_{jl}}{\sum_{i=1}^{m} f(c_i^j, \mu, \sigma^2|B_j)P_{ji}}$$

각각의 블록에 대해, 위와 같은 확률을 가지는 다항분포을 반복적으로 생성했을 때 j번째 블록에서 $\mathcal{C}$에서 l번째 원소를 선택한 빈도수를 $f_{j,l}$이라고 표시하면 표 2.1와 같은 분할표를 얻을 수 있다.

변화점의 유무와 위치에 대한 블록 탐색은 표 2.1에서 인접블록 간의 선택된 모수의 빈도차이를 이용한다. 빈도의 차이를 측정하기 위한 측도로 다음과 같은 측도를 생각해 볼 수 있다.

$$\psi_1(j) = \sum_{l=1}^{m}|f_{j,l} - f_{j+1,l}|^p, \quad j = 1, \ldots, m-1, \; p > 0$$

$$\psi_2(j) = \sum_{l=1}^{m} \sum_{k=l-1}^{l+1} |f_{j,l} - f_{j+1,k}|^p$$

이 측도를 기준으로 상대적으로 변화점의 발생가능성이 높은 인접 블록을 확인한 후 정확한 위치를 아래에 설명된 방법으로 찾는다.

2.2.2. 정확탐색(Exact search) 앞에서 언급한 블록탐색은 블록의 크기에 영향을 받는 경향이 있다. 블록의 크기가 큰 경우 변화점인 있는 부분과 없는 부분에서 인접블록 간의 측도에 확연한 차이가 있는 반면 정확한 위치를 파악하기 어렵고 블록의 크기가 작은 경우 장기억 모수의 효과를 너무 세분화시켜 측도의 변화가 크게 일어나는 문제가 있다. 그러므로 블록탐색이 효과적인 방법이 되기 위해서는 적정한 크기의 블록을 정하는 것이 중요하지만 시계열특성에 따라 그 크기가 다르기 때문에 일반적인 원칙을 정하는 것이 어려울 뿐만 아니라 블록탐색으로 정확한 변화점의 위치를 알아내는 것은 사실상 불가능하다. 이 논문에서는 블록탐색에 의해 얻은 개괄적 변화점 갯수와 위치를 바탕으로 정확한 변화점 위치와 변화량을 검출하는 방법에 대해 알아본다. 변화점의 갯수가 정확하게 파악되지 않는 경우, 가능한 변화점 갯수에 대해 베이즈 인자(Bayes factor)를 계산하여 비교할 수 있다.

장기억 과정에서 K개의 변화점이 있다고 할 때, 각각의 변화점 위치를 $\vec{\tau} = (\tau_1, \ldots, \tau_K)$라고 하고 같은 부분차분모수를 가지는 $K+1$개의 시계열 구간의 모수를 $\vec{c} = (c_{l_1}^1, \ldots, c_{l_{K+1}}^{K+1})$라고 하자. 이 변화점의 위치들은 앞의 블록 탐색에서 사용된 m개의 블록 중 ψ의 값이 큰 K개 위치의 인접 블록에 있다고 가정한다. 시간순서대로 표시했을 때 k번째 변화점에 해당하는 위치가 k_*번째와 k_*+1번째 블록에서 있는 경우, 이 인접블록의 위치를 $N_k = \{(k_*-1)b+1, \ldots, (k_*+1)b\}$라고 표시한다.

블록 $\{\tau_{k-1}+1, \ldots, \tau_k\}$에서 선택된 부분차분모수가 d일 때 가능도함수는

$$f(\tau_{k-1}, \tau_k, d \mid \vec{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\tau_k - \tau_{k-1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{t=\tau_{k-1}+1}^{\tau_k} \left\{(1-B)^d (x_t - \mu)\right\}^2\right]$$

이고 각각의 N_k에서 선택된 $\vec{\tau}$와 $\vec{c}$에 대해, 가능도함수는 다음과 같이 쓸 수 있다.

$$\begin{aligned} f(\vec{\tau}, \vec{c} \mid \vec{x}) &= \prod_{k=1}^{K+1} f(\tau_{k-1}, \tau_k, c_{l_k}^k \mid \vec{x}) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{K+1} \sum_{t=\tau_{k-1}+1}^{\tau_k} \left\{(1-B)^{c_{l_k}^k} (x_t - \mu)\right\}^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{K+1} \sum_{t=\tau_{k-1}+1}^{\tau_k} \left\{\sum_{j=1}^{t} \pi_j(c_{l_k}^k)(x_t - \mu)\right\}^2\right], \end{aligned}$$

여기서 $\tau_0 = 0$이고 $\tau_{K+1} = n$을 나타낸다. 변화점 위치가 있을 가능성이 있는 $\vec{N} = (N_1, \ldots, N_K)$와 $\vec{c}$의 선택은 블록탐색에서와 같이 독립인 다항분포를 이용하여 사전분포를 설정한다.

$$\begin{aligned} \vec{P}_k^{\tau} &\sim \text{Dirichlet}(\vec{\delta}_k^{\tau}), \quad \vec{\delta}_k^{\tau} = (\delta_{k,1}^{\tau}, \ldots, \delta_{k,2b}^{\tau}), \quad k = 1, \ldots, K, \\ \vec{L}_k^{\tau} &\sim \text{Multinomial}(1, \vec{P}_k^{\tau}), \quad \vec{P}_k^{\tau} = (P_{k,1}^{\tau}, \ldots, P_{k,2b}^{\tau}), \\ \vec{P}_j^{d} &\sim \text{Dirichlet}(\vec{\delta}_j^{d}), \quad \vec{\delta}_j^{d} = (\delta_{j,1}^{d}, \ldots, \delta_{j,m}^{d}), \quad j = 1, \ldots, K+1, \\ \vec{L}_j^{d} &\sim \text{Multinomial}(1, \vec{P}_j^{d}), \quad \vec{P}_j^{d} = (P_{j,1}^{d}, \ldots, P_{j,m}^{d}), \end{aligned}$$

사전분포와 가능도함수를 이용하여 $\vec{L}^{\tau} = (\vec{L}_1^{\tau}, \ldots, \vec{L}_K^{\tau})$, $\vec{P}^{\tau} = (\vec{P}_1^{\tau}, \ldots, \vec{P}_K^{\tau})$, $\vec{L}^{d} = (\vec{L}_1^{d}, \ldots, \vec{L}_{K+1}^{d})$ 그리고 $\vec{P}^{d} = (\vec{P}_1^{\tau}, \ldots, \vec{P}_{K+1}^{d})$의 사후분포를 유도해야 한다. 블록 탐색의 경우 고정된 위치에서 부분 차분모수를 선택했기 때문에 각각 블록에 대해 독립적으로 사후분포를 유도할 수 있었으나 정확탐색에서는 모수의 결합사후분포가 변화점의 위치에 영향을 받기 때문에 각각의 모수에 대해 단순히 주변사후분포를 곱으로 표시할 수 없다. 이 논문에서는 결합사후분포를 효율적으로 유도하기 위해 완전조건부분포(full conditional distribution)에 근거한 깁스표집기를 이용한다. 임의로 설정한 초기값 $\vec{L}^{\tau(0)}$과 $\vec{L}^{d(0)}$이라고 하면, 다음과 같은 과정을 반복적으로 실시하여 변화점 위치와 부분차분모수의 값에 대한 근사 사후분포를 구한다.

1. $k = 1, \ldots, K+1$에 대해,
   (1) $\vec{P}_k^{d(1)} \mid \vec{L}_k^{d(0)} \sim \text{Dirichlet}(\vec{\delta}_1^{d*})$, $\vec{\delta}_1^{d*} = (\delta_{1,1}^{d*}, \ldots, \delta_{1,m}^{d*})$
   (2) $\vec{L}_k^{d(1)} \mid \tau_{k-1}^{(0)}, \tau_k^{(0)}$, $\vec{P}_k^{d(1)} \sim \text{Multinomial}(1, \vec{P}_k^{d*})$, $\vec{P}_k^{d*} = (P_{k,1}^{d*}, \ldots, P_{k,m}^{d*})$

$$P_{k,j}^{d*} = \frac{P_{k,j}^{d(1)} f\left(\tau_{k-1}^{(0)}, \tau_k^{(0)}, c_j^k, \mid \vec{x}\right)}{\sum_{l=1}^{m} P_{k,l}^{d(1)} f\left(\tau_{k-1}^{(0)}, \tau_k^{(0)}, c_l^k, \mid \vec{x}\right)}$$

2. $k = 1, \ldots, K$에 대해,
   (1) $\vec{P}_k^{\tau(1)} \mid \vec{L}_k^{\tau(0)} \sim \text{Dirichlet}(\vec{\delta}_k^{\tau*})$, $\vec{\delta}_k^{\tau*} = (\delta_{k,1}^{\tau*}, \ldots, \delta_{k,2b}^{\tau*})$
   (2) $\vec{L}_k^{\tau(1)} \mid \tau_{k-1}^{(1)}, \tau_{k+1}^{(0)}$, $\vec{P}_k^{\tau(1)}, c_{l_k}^{k(1)}, c_{l_{k+1}}^{k+1(1)} \sim \text{Multinomial}(1, \vec{P}_k^{\tau*})$, $\vec{P}_k^{\tau*} = (P_{k,1}^{\tau*}, \ldots, P_{k,2b}^{\tau*})$

$$P_{k,j}^{\tau*} = \frac{P_{k,j}^{\tau(1)} f\left(\tau_{k-1}^{(1)}, (k_*-1)b+j, c_{l_k}^{k(1)} \mid \vec{x}\right) f\left((k_*-1)b+j, \tau_{k+1}^{(0)}, \tau_{k-1}^{(1)}+j, c_{l_{k+1}}^{k+1(1)} \mid \vec{x}\right)}{\sum_{i=1}^{2b} P_{k,i}^{\tau(1)} f\left(\tau_{k-1}^{(1)}, (k_*-1)b+i, c_{l_k}^{k(1)} \mid \vec{x}\right) f\left((k_*-1)b+i, \tau_{k+1}^{(0)}, \tau_{k-1}^{(1)}+j, c_{l_{k+1}}^{k+1(1)} \mid \vec{x}\right)}$$

여기서 $\vec{L}_1^{d(1)}$은 다른 구간의 부분차분모수와는 독립적으로, $\vec{L}_1^{\tau(1)}$은 다른 변화점 위치와 독립적으로 생성한다.

3. 나일강 수위자료

장기억 과정에서 변화점에 대한 예제로 자주 사용되는 자료는 622년에서 1284년 동안 나일강의 연도별 최저수위를 측정한 663개의 시계열이다. 이 자료를 분석한 기존 연구결과로 Beran과 Terrin (1996)는 722년($\hat{\tau}_1 = 100$)으로 앞의 100년 동안 $\hat{d}_1 = 0.04$였고 나머지 기간동안 $\hat{d}_2 = 0.38$인 것으로 추정되었다. Ray와 Tsay (2002)는 블록크기를 20으로 했을 때, 722년까지 $\hat{d}_1 = 0.05$였고 $\hat{d}_2 = 0.45$인 것으로 분석되었다.

제안된 블록탐색과 정확탐색에서 d는 $\mathcal{C} = \{0.025, 0.05, \ldots, 0.475\}$로 간격이 초기값을 0.025로 시작하여 0.025씩 증가하도록 하여 19가지 d를 선택할 수 있게 하였다. 블록탐색과 정확탐색에서의 사전분포의 초월모수 δ는 모두 0.01로 설정했으며 μ의 사전분포는 평균이 1150이고 분산이 8000인 정규분포, σ^{-2}은 형태모수가 0.01이고 척도모수가 0.01인, 평균이 1이고 분산이 100인, 감마분포를 따른다고 가정하였다. 블록탐색에서 블록크기는 50으로 정했고 모든 결과는 10000개의 깁스표본 중 앞의 5000개를 제외한 나머지 5000개의 표본을 이용하여 얻었다.

표 3.1은 블록크기를 50으로 블록탐색을 했을 때, 인접블록간의 ψ를 계산한 후 최대값을 각각 나누어 표준화시킨 ψ값이다. 이 표에 의하면 두 번째 블록과 세 번째 블록에서 가장 큰 차이가 있고 아홉 번째

표 3.1. 블록크기 $b = 50$일 때 블록간의 표준 ψ값

ψ	1	2	3	4	5	6	7	8	9	10	11	12
ψ_1	.157	1.000	.405	.313	.710	.375	.175	.644	.869	.379	.237	.205
ψ_2	.167	1.000	.399	.310	.721	.390	.179	.665	.878	.378	.233	.211

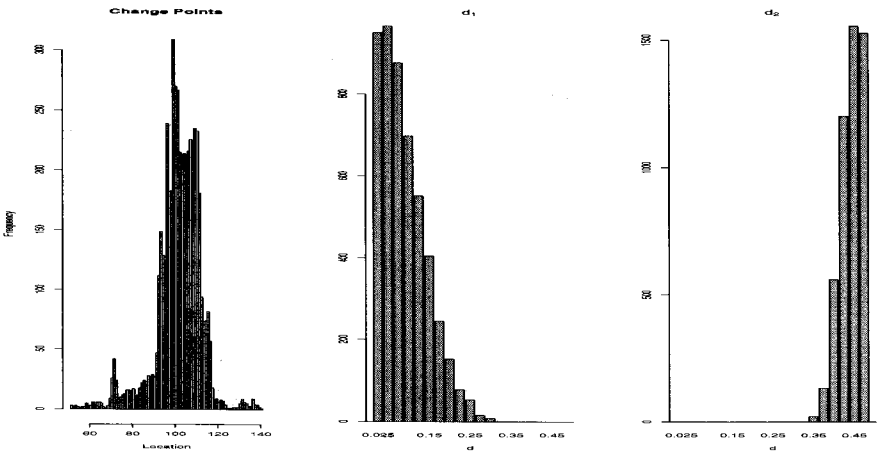

그림 3.1. 변화점이 하나인 경우 변화점 위치와 부분차분모수의 분포

와 열 번째 블록에서 그 다음으로 큰 것으로 나타났다. 이것은 변화점이 하나인 경우 그 점은 시점 50에서 150사이에 있을 가능성이 높고, 두 개인 경우 시점 50에서 150사이와 시점 400에서 500사이에 있을 가능성이 높다는 것을 의미한다.

그림 3.1은 변화점이 하나 있다고 가정한 경우의 정확분석결과로 얻어진 변화점의 위치와 부분차분모수의 분포를 보여준다. 변화점의 위치는 $t = 100$이 309번 선택되어 최빈값을 나타났으며 $t = 95$에서 105사이에서 전체의 43%이상이 선택되어 변화점이 시점 100을 중심으로 변화점이 모여 있는 형태를 가지는 것으로 나타났다. 첫 번째 구간에서 선택된 d는 0.025인 경우 949번, 0.05인 경우 966번, 0.075인 경우 876번으로 0.05 근처에서의 확률이 높은 것으로 나타났으며 두 번째 부분차분모수는 0.0425에서 1203, 0.45에서 1556, 0.475에서 1528번 선택되어 0.45근처에서 높은 확률을 가지는 것으로 분석되었다.

그림 3.2은 변화점이 두 개인 경우 분석결과에서 얻어진 변화점의 위치와 부분차분모수의 분포를 보여준다. 변화점이 하나인 경우와 같이 첫 번째 변화점은 시점 100 근처에서 높은 확률을 가지고 첫 번째 구간의 부분차분모수 또한 0.05근처에서 높은 확률을 가지는 반면 두 번째 변화점은 확연하게 높은 확률을 가지는 부분이 없으며 두 번째 구간의 부분차분모수와 세 번째 구간의 부분차분모수간의 분포의 차이가 첫 번째와 같이 크지 않은 것으로 볼 수 있다. 이것을 종합하건데, 베이즈요인을 계산하지 않아도, 변화점이 두 개인 경우보다 하나인 경우가 설명력이 높다고 할 수 있다.

그림 3.3은 변화점이 하나인 경우와 두 개인 경우 μ와 σ^2의 추이를 보여주고 있다. 변화점이 하나인 경우 μ의 평균은 1149, 중앙값은 1149, 표준편차는 10.5인 것으로 나타났으며 σ^2의 평균은 4739, 중앙값은 4732, 표준편차는 259.0인 것으로 나타났다. 변화점이 두 개인 경우, μ의 평균은 1149, 중앙값은 1149, 표준편차는 10.4, σ^2의 평균은 4741, 중앙값은 4732, 표준편차는 259.9로 변화점이 한 개인 경우와 차이가 거의 없는 것으로 분석되었다.

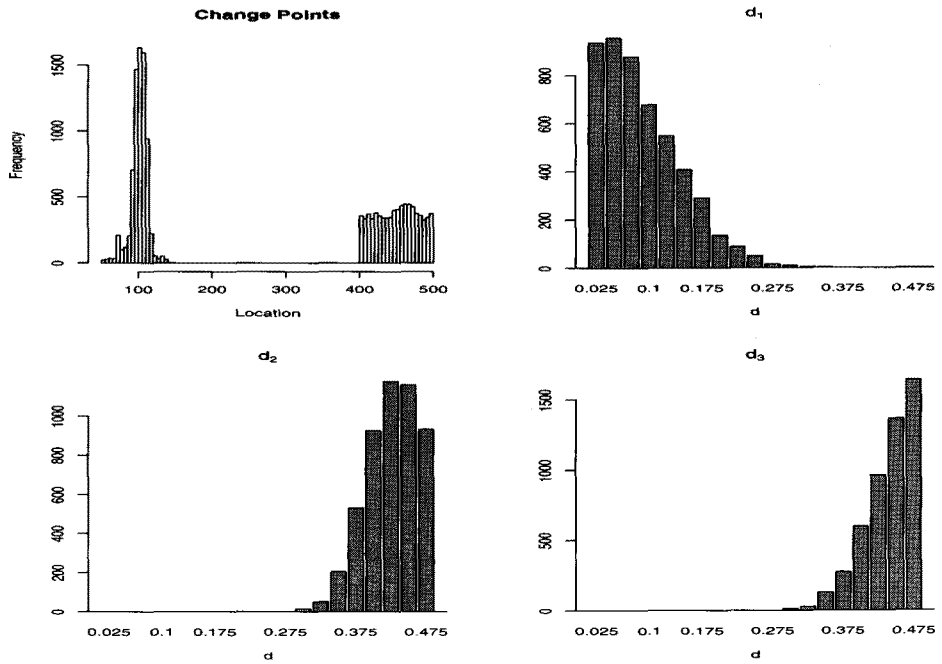

그림 3.2. 변화점이 둘인 경우 변화점 위치와 부분차분모수의 분포

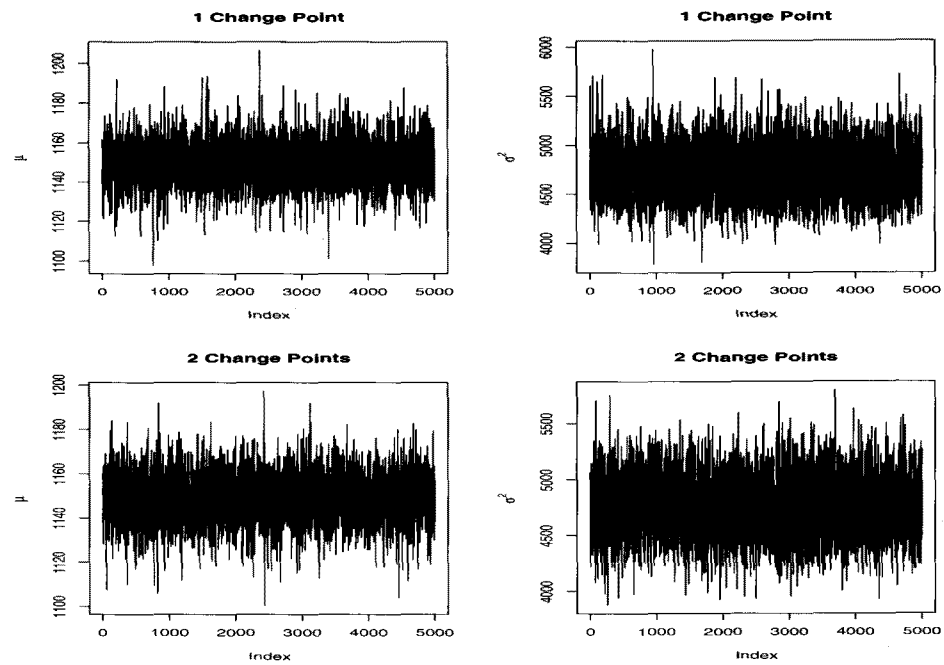

그림 3.3. μ와 σ^2의 추이분석

4. 결론

장기억 과정은 시계열자료간에 종속성이 장기간 존재하는 특징을 가지는 시계열로 과학기술의 발전에 따라 이러한 시계열 자료는 더욱더 많이 생산되고 이러한 자료에서의 변화점 검출은 매우 중요한 관심사가 되고 있다. 장기억 과정분석에서 많이 사용되고 있는 ARFIMA 모형을 이용하여 자료를 분석하기 위해서는 상당량의 연산과정이 필요하고 MCMC와 같은 방법을 이용하는 베이지안 분석에서는 이 연산과정이 부담스럽고 변화점이 있는 경우에는 더욱 많은 연산을 필요로 한다.

이 논문에서는 장기억 과정에서 변화점의 위치와 이에 따른 부분차분모수를 빠르게 검출하는 베이지안 방법을 소개하였다. 부분차분모수의 공간이 $(0, 0.5]$와 같이 짧은 구간으로 이루어져 있다는 것에 주목하고 부분차분모수를 추정하는 대신 이 공간을 몇 개의 그룹으로 나누고 각 구간을 대표하는 값을 이 중에서 선택하도록 하였다. 이렇게 하면 베이지안 추론 초기단계에서 선택가능한 값에 대해서만 가능도를 한번만 계산하고 이후에는 선택된 값에 대한 가능도함수값을 참조만 하면 되기 때문에 빠른 속도로 MCMC를 수행할 수 있다. 나일강 최저수위 자료분석에서 기존 연구결과와 유사하면서도 빠르면서도 더 확대된 분석이 가능하다는 것을 보였다.

참고문헌

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

Chan, N. H. and Palma, W. (1998). State space modeling of long-memory processes, *The Annals of Statistics*, **26**, 719–739.

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

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

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource, *Applied Statistics*, **38**, 1–50.

Hidalgo, J. and Robinson, P. M. (1996). Testing for structural change in a long-memory environment, *Journal of Econometrics*, **70**, 159–174.

Ko, K. and Vannucci, M. (2006). Bayesian wavelet-based methods for the detection of multiple changes of the long memory parameter, *IEEE Transactions on Signal Processing*, **54**, 4461–4470.

Kuan, C. M. and Hsu, C. C. (1998). Change-point estimation of fractionally integrated processes, *Journal of Time Series Analysis*, **19**, 693–708.

Liu, S. I. and Kao, M. H. (1999). Bayesian analysis for multiple changes of the long memory parameter, Technical Report, Graduate Institute of Statistics, National Central University, Chung-Li, Taiwan.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series, *Journal of the American Statistical Association*, **88**, 968–978.

Ray, B. K. and Tsay, R. S. (2002). Bayesian methods for change-point detection in long-range dependent processes, *Journal of Time Series Analysis*, **23**, 687–705.

Wright, J. H. (1998). Testing for a structural break at unknown data with long-memory disturbances, *Journal of Time Series Analysis*, **19**, 369–376.

A Fast Bayesian Detection of Change Points in Long-Memory Processes

Joo Won Kim[1] · Sinsup Cho[2] · In-Kwon Yeo[3]

[1]Office of Admissions, Seoul National University; [2]Department of Statistics, Seoul National University; [3]Department of Statistics, Sookmyoung Women's University

(Received April 2009; accepted July 2009)

Abstract

In this paper, we introduce a fast approach for Bayesian detection of change points in long-memory processes. Since a heavy computation is needed to evaluate the likelihood function of long-memory processes, a method for simplifying the computational process is required to efficiently implement a Bayesian inference. Instead of estimating the parameter, we consider selecting a element from the set of possible parameters obtained by categorizing the parameter space. This approach simplifies the detection algorithm and reduces the computational time to detect change points. Since the parameter space is (0, 0.5], there is no big difference between the result of parameter estimation and selection under a proper fractionation of the parameter space. The analysis of Nile river data showed the validation of the proposed method.

Keywords: ARFIMA models, change point detection, Dirichlet distribution.

[3]Corresponding author: Associate Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: inkwon@sm.ac.kr