The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.449
29(3), 449–462

Bootstrap estimation of long-run variance under strong dependence

Changryong Baek[a,1] · Yong Kwon[a]

[a]Department of Statistics, Sungkyunkwan University

(Received January 13, 2016; Revised February 5, 2016; Accepted February 29, 2016)

Abstract

This paper considers a long-run variance estimation using a block bootstrap method under strong dependence also known as long range dependence. We extend currently available methods in two ways. First, it extends bootstrap methods under short range dependence to long range dependence. Second, to accommodate the observation that strong dependence may come from deterministic trend plus noise models, we propose to utilize residuals obtained from the nonparametric kernel estimation with the bimodal kernel. The simulation study shows that our method works well; in addition, a data illustration is presented for practitioners.

Keywords: long-run variance, long range dependence, block bootstrap

1. 서론

정상시계열에 대한 장기적 분산(long-run variance; LRV)은 모평균의 추정값인 표본 평균의 점근적 정규성을 위한 표준화된 극한(scaling limit)으로 정의된다. 따라서 LRV는 시계열 분석의 추론에서 매우 중요한 역할을 하는 모수이다. 예를 들어 정상시계열의 모평균에 대한 추론, 자기 상관성을 갖는 회귀분석에서의 회귀 모수에 대한 검증, 단위근 검증, 변화점 감지를 위한 CUSUM 통계량 등등 그 활용 분야는 무궁무진하다.

이러한 중요성에 따라 지난 수십년 동안 LRV에 대한 일치 추정량이 활발히 연구되었다. LRV에 대한 대표적인 시간영역(time domain)에서의 추정량은 표본 자기 상관 함수들의 가중합인 heteroskedasticity and autocorrelation covariance(HAC)이 있다. 가중치를 결정하는 커널(kernel)에 따라 Newey과 West (1987)이 제안한 바틀렛(Bartlett) 커널을 이용한 바틀렛 추정량을 비롯해 Andrews (1991)이 제안한 quadratic spectral(QS) 커널을 이용한 QS 추정량 등이 있다. 이는 곧 커널 추정량의 특별한 형태로 볼 수 있으므로 띠넓이 선택(bandwidth selection)이 유한 표본의 성능을 크게 좌우하며 이와 관련한 많은 후속 연구가 진행되었다. 주파수 영역(frequency domain)에서의 추정량은 원점 근처에서의 스펙트럴 밀도함수의 일치 추정과 관련이 있으며 자세한 내용은 Brillinger (1981)을 참조하기 바란다.

붓스트랩을 이용한 LRV의 추정도 활발하게 연구가 되었다. 중요한 점은 시계열 자료의 경우 그 의존

This research was supported by the Basic Science Research Program from the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1006025).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: crbaek@skku.edu

구조(dependence structure)가 공분산 구조를 결정짓는 핵심이기 때문에 일반적인 i.i.d. 가정 하에서의 재표본(resampling)을 사용하지 못한다는 점이다. 이러한 문제는 의존 구조를 보존하기 위해서 한 개씩 자료를 뽑는 것이 아니라 시계열 자료를 길이가 ℓ인 블럭으로 뽑는 Künsch (1989)의 moving block bootstrap(MBB)에 의해 해결의 실마리를 찾게 되었다. 블럭 붓스트랩은 그 이후 겹침을 허락하지 않는 블럭 붓스트랩(nonoverlapping block boostrap; NBB), 정상성을 보존해주는 정상 블럭 붓스트랩(stationary block bootstrap; SBB), 자료를 이어 붙여 처음과 끝부분의 블럭 크기가 작아지는 단점을 해결한 순환 블럭 붓스트랩(circular block bootstrap; CBB) 등의 여러 변형된 방법으로 발전하였다. 하지만, 커널 평활법이 띠넓이에 의존하듯이 블럭 붓스트랩은 블럭의 크기가 유한 표본에서의 성능을 크게 결정하는 매우 중요한 모수다. 예를 들어 Politis와 White (2004)는 상호 의존성이 약한 단기억 시계열(short range dependence; SRD)에서 블럭의 크기를 결정하는 방법을 연구하였으며 보다 자세한 블럭 붓스트랩에 대한 논의는 Lahiri (2003)를 참고하길 바란다.

본 논문은 자료의 의존성이 매우 강한 장기간 의존 시계열(long range dependence; LRD)에 대한 붓스트랩을 이용한 LRV 추정량에 대해서 제안한다. LRD 시계열은 그 의존성이 시차가 증가함에도 불구하고 매우 천천히 감소하여 SRD 가정 하에서 개발된 방법들이 일치 추정량을 되지 못하기 때문에 수정(modification)이 필요하다. 또한, LRD 가정 하에서 유한 표본 성능을 좌우하는 커널의 띠넓이 선택, 블록 크기 결정 등의 방법들에 대한 연구는 매우 미흡한 실정이다. 따라서 본 연구는 붓스트랩 방법에서 중요한 블럭 크기 결정을 위해 수정된 교차 검증법(modified cross-validation; MCV) 혹은 $(2l+1)$-CV의 아이디어를 이용한 방법을 제안한다. 이는 곧 추세가 포함된 평균변화모형에서의 LRV 추정으로 확장 가능하며 Kim 등 (2009)에서 제안한 쌍봉형 커널(bimodal kernal)에 기반을 둔 블럭 크기 결정법을 제안한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 LRV 추정량에 대해서 엄밀한 정의를 내리고 기존의 대표적인 방법론을 설명한다. 제 3장에서는 쌍봉형 커널을 이용하여 추세함수를 추정하고 이를 이용하여 블럭 크기를 결정하는 블럭 붓스트랩 방법을 제안한다. 제 4장에서는 기존의 여러 방법론과의 비교를 통해서 본 논문에서 제안한 방법론의 유용성을 살펴보며 제 5장에서는 결론을 다루었다.

2. 장기적 분산 추정 방법

정상 시계열 $\{X_t\}_{t\in\mathbb{Z}}$는 모평균 μ가 일정하고 자기공분산 함수 $\gamma(h) = \mathrm{Cov}(X_0, X_h)$가 존재하며 시차(lag)에만 의존하는 시계열을 뜻한다. 모평균에 대한 추정값으로는 관측 자료 $X_1, \ldots, X_n$의 표본평균 $\hat{\mu} := \overline{X}_n = n^{-1}\sum_{t=1}^{n} X_t$을 생각할 수 있다. 단기억 시계열임을 추가로 가정한다면, 즉 $\sum_{h\in\mathbb{Z}} |\gamma(h)| < \infty$, 중심극한 정리에 의해서

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\left(0, s^2\right)$$

임을 보일 수 있다. 따라서 s^2은 $\sigma_n^2 := \mathrm{Var}(\sqrt{n}\overline{X}_n)$의 극한으로 이를 장기적 분산(LRV)이라고 부르고 수식으로는 다음과 같이 정의한다.

$$s^2 := \lim_{n\to\infty} \sigma_n^2 = \lim_{n\to\infty} \sum_{h=-(n-1)}^{n-1} \left(1 - \frac{|h|}{n}\right)\gamma(h) = \sum_{h\in\mathbb{Z}} \gamma(h). \tag{2.1}$$

주파수 영역에서 스펙트럴 밀도함수는 자기 공분산의 푸리에 급수로 나타낼 수 있으므로, 즉

$$f(\lambda) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h), \quad \lambda \in (-\pi, \pi], \tag{2.2}$$

LRV는 주파수 영역에서 원점에서의 스펙트럴 밀도함수의 상수배인

$$s^2 = 2\pi f(0) \tag{2.3}$$

으로 정의된다.

한편, 강한 의존성을 가지는 장기간 의존 시계열(LRD)는 주파수 영역에서

$$f(\lambda) = b_0|\lambda|^{-2d} + o\left(|\lambda|^{-2d}\right), \quad \text{as } \lambda \to 0,\ b_0 > 0,\ d \in \left(0, \frac{1}{2}\right) \tag{2.4}$$

으로 정의되며 d를 LRD 모수라고 부른다. 따라서 원점 근방에서 스펙트럴 밀도함수가 발산하므로 스펙트럴 밀도함수와 자기 공분산과의 관계인 식 (2.2)에 비추어 보면

$$\sum_{h\in\mathbb{Z}} |\gamma(z)| = +\infty$$

임을 알 수 있어 LRD 시계열에서의 표본 평균은 $\sqrt{n}$의 속도로 정규분포로 수렴하지 못한다. 약간의 추가적인 가정하에서 LRD 시계열은 또한 자기 상관함수가 시차에 따라서 멱함수형태로 감소하는 시계열임을 알 수 있다. 즉

$$\gamma(h) = \mathrm{Corr}(X_0, X_h) \sim Ch^{2d-1}, \quad C > 0$$

을 만족한다. 따라서 LRD 모수 d는 LRD 시계열의 의존성을 결정하는 핵심 모수로 d가 클수록 매우 강한 종속관계를 보이게 되며 $d = 0$ 인 경우에는 SRD 시계열이 된다. 이러한 매우 강한 종속관계 때문에 표본 평균은 SRD 시계열의 경우보다 훨씬 더 느린 속도 ($O(n^{1/2-d})$)로 정규분포로 수렴함이 잘 알려져 있으며 (Beran, 1994) LRD 시계열의 경우 LRV는

$$s^2(d) := \lim_{n\to\infty} \mathrm{Var}\left(n^{\frac{1}{2}-d}\overline{X}_n\right) = b_0 p(d) \tag{2.5}$$

으로 정의되어 원점 근방에서의 스펙트럴 밀도 함수 b_0를 LRD 모수 d에 의존하는 상수값 $p(d)$만큼 곱해준 값으로 정의된다 (Abadir 등, 2009). 상수 $p(d)$는

$$p(d) = \begin{cases} 2\dfrac{\Gamma(1-2d)\sin(\pi d)}{d(1+2d)}, & d \in \left(0, \dfrac{1}{2}\right), \\ 2\pi, & d = 0 \end{cases}$$

으로 LRD 모수 $d = 0$인 경우, 즉 SRD 시계열의 경우, LRD에서 정의한 LRV (2.5)와 SRD에서 정의한 LRV (2.1)이 일치함을 살펴볼 수 있다.

강한 종속성을 가지는 시계열에 대한 LRV의 대표적인 추정량은 표본 자기 공분산함수(sample autocovariance function; SACVF)를 이용한 HAC 추정량과 주파수 영역에서의 추정량인 memory and autocorrelation consistent(MAC) 추정량이 있다. HAC 추정량은

$$\hat{s}_H^2(d) = q^{-2d} \sum_{k=-q}^{q} K\left(\frac{h}{B}\right)\hat{\gamma}(h) \tag{2.6}$$

로 $K(\cdot)$은 커널 함수(kernel function)이며 B는 띠넓이(bandwidth)이고

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=|h|+1}^{n} \left(X_t - \bar{X}_n\right)\left(X_{t-|h|} - \bar{X}_n\right)$$

는 SACVF이다. 커널의 경우 Newey와 West (1987)가 제안한 Bartlett 커널 함수가 널리 쓰이며 다음과 같다.

$$K\left(\frac{h}{B}\right) = \begin{cases} 1 - \dfrac{h}{B+1}, & |h| \le B, \\ 0, & |h| > B. \end{cases} \tag{2.7}$$

HAC 추정량은 커널 추정량으로 평활법이 가지는 단점, 즉 띠넓이 선택이 그 성능을 좌우하는 매우 중요한 모수가 된다. Abadir 등 (2009)은 이론적으로 $q \to \infty$에 대해서 $q = O(n^{1/2})$이면 일치 추정량이 된다고 밝혔고 모의 실험 결과 $[n^{.2}]$가 일반적으로 성능이 좋음을 보여 본 논문에서는 $B = [n^{.2}]$를 사용하여 연구를 진행하였다.

한편 주파수 영역에서의 LRV 추정량으로 Robinson (2005)은 MAC 추정량을 다음과 같이

$$\hat{s}_M^2(d) := p(d)\frac{1}{m}\sum_{t=1}^{m} \lambda_t^{2d} I(\lambda_t) \tag{2.8}$$

제안하였다. $I(\lambda_t)$은 푸리에 주파수 $\lambda_t = 2\pi t/n$에 대한 피리오도그램

$$I(\lambda_t) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t e^{-ij\lambda_t}\right|^2 \tag{2.9}$$

이다. MAC 추정량의 성능 역시 튜닝 모수 m에 그 성능이 크게 좌우되며 Abadir 등 (2009)에 따르면 MSE를 가장 작게 만드는 최적의 튜닝 모수는 $O(n^{4/5})$이며 모의 실험에서는 $m = [n^{.8}]$이 가장 작은 MSE를 산출하여 본 논문에서는 이를 사용하였다.

한편, 앞서 설명한 HAC, MAC 추정량 모두 LRD 모수 d에 의존하기 때문에 이에 대한 일치 추정량이 필요하다. 본 논문에서는 피리오도그램 (2.9)에 기반한 준모수적 추정방법인 local Whittle(LW) 추정량을 사용하여 d를 추정하였다. LW 추정량은

$$\hat{d} = \underset{d\in[\Theta_1,\Theta_2]}{\operatorname{argmin}} \left\{ \log\left(\frac{1}{m^L}\sum_{t=1}^{m^L} \lambda_t^{2d} I(\lambda_t)\right) - 2d\frac{1}{m^L}\sum_{t=1}^{m^L}\log\lambda_t \right\} \tag{2.10}$$

으로 주어지며 $-1/2 < \Theta_1 < 0 < \Theta_2 < 1/2$, m^L은 LW 추정량에 쓰인 푸리에 주파수의 개수이다. 정규 분포 가정하의 최적의 $m^L = O(n^{.8})$임이 알려져 있으며 본 논문에서는 실증 분석에서 많이 쓰이는 $m^L = [n^{2/3}]$을 사용하였다.

3. 블럭 붓스트랩을 이용한 LRV 추정 및 블럭 크기 결정

따라서 LRD 시계열에서의 LRV의 추정은 정의인 수식 (2.5)에서 살펴보듯이 붓스트랩 표본을 통해 표본 평균을 구한 뒤 이들의 표본 분산에 n^{1-2d}를 곱함으로써 산출할 수 있다. 하지만 강한 종속성 때문에 한 개씩 복원 추출함으로써 붓스트랩 표본을 추출한다면 이는 자료의 의존 구조를 정확하게 반영하지 못하여 좋은 추정량이 될 수 없다. 따라서 시계열 자료에서 가장 많이 쓰이는 종속 구조를 보존해 주는 블럭 붓스트랩(block bootstrap; BB)을 이용한다면 LRV 추정량을 다음과 같이 구할 수 있다.

Step 1. 주어진 자료 $\{X_t, t = 1, \ldots, n\}$에 대해서 $Y_t := X_{t\bmod(n)}$, $t \in \mathbb{Z}$를 정의한다. $t\bmod(n)$은 t를 n으로 나눈 나머지로 $t > n$인 경우에 관측 자료를 원형의 형태로 이어 붙이는 것을 뜻한다. 반복을 나타내는 변수를 r이라 하고, 그 초기값을 0으로 놓자.

Step 2. 블럭의 시작점 i_r를 균등분포 $\{1,\ldots,n\}$에서 뽑는다.

Step 3. 주어진 시작점에 대하여 블럭 크기가 ℓ인 자료 $\{Y_{i_r+j-1}, j=1,\ldots,\ell\}$를 뽑는다. 이렇게 새로 뽑힌 표본을 $Y^*_{r\ell+j}=Y_{i_r+j-1}$, $j=1,\ldots,\ell$라 둔다.

Step 4. r을 1씩 증가시키면서 Step 2–3를 반복하여 $Y^*_1,\ldots,Y^*_n$을 얻고 붓스트랩 표본의 평균 $\overline{Y}^*_{(k)}=n^{-1}\sum_{j=1}^n Y^*_j$을 구한다.

Step 5. 붓스트랩 표본의 스케일된 극한값이 LRV이므로, 즉

$$\mathrm{Var}\left(n^{\frac{1}{2}-d}\overline{Y}^*_{(k)}\right),$$

블럭 붓스트랩을 이용한 LRV 추정량은 Step 1–4를 n_B번 붓스트랩 반복하여 얻어진 표본평균 $\overline{Y}^*_{(1)},\ldots,\overline{Y}^*_{(n_B)}$의 표본 분산에 $n^{1-2\hat{d}}$를 곱한값이며 이를

$$\hat{s}^2_B\left(\hat{d}\right):=\frac{n^{1-2\hat{d}}}{n_B-1}\sum_{k=1}^{n_B}\left(\overline{Y}^*_{(k)}-n_B^{-1}\sum_{j=1}^{n_B}\overline{Y}^*_{(j)}\right)^2$$

로 표기한다.

앞서 HAC 혹은 MAC 추정량이 띠넓이 모수 q 혹은 낮은 주파수의 개수 m에 의존하였듯이 BB를 이용한 LRV 추정량 역시 블럭의 크기를 결정하는 모수인 ℓ에 크게 의존한다. 본 연구에서는 Politis와 Romano (1995)년에 제안한 방법에 따라

$$\ell=2\cdot\min_h\left\{h\geq 1 \text{ such that } |\hat{\rho}(h)|\leq\frac{2}{\sqrt{n}}\right\} \tag{3.1}$$

을 사용하였다. 여기에서 $\hat{\rho}(h)$는 h차 표본자기상관계수이다. 위 방법은 Chu와 Marron (1991)이 제안한 $(2l+1)$-CV에 기반한 커널 띠넓이 선택과 매우 밀접한 관계가 있다. 즉, 주어진 시점을 기준으로 의존성이 거의 0이되는 시차만큼 앞 뒤로 자료를 잘라서 부분표본을 만들어 의존구조를 보존하는 방법이다. 편의상 위 방법을 'BB1'으로 지칭한다.

Remark 3.1: 자료의 겹침을 허락하지 않은 NBB의 경우에도 비슷하게 LRV 추정량을 다음과 같이 정의할 수 있다. 블럭 크기 ℓ에 대하여 첫 번째 블럭은 $\{X_1,\ldots,X_\ell\}$, 두 번째 블럭은 $\{X_{\ell+1},\ldots,X_{2\ell}\}$ 등으로 붓스트랩 부분표본을 얻을 수 있고 각 부분 표본으로부터 산출된 표본 평균을 $\overline{X}^\circ_1,\overline{X}^\circ_2,\ldots,\overline{X}^\circ_{[n/\ell]}$라고 표기하자. 그러면 NBB에 의한 LRV 추정량은

$$\frac{n^{1-2\hat{d}}}{[n/\ell]-1}\sum_{k=1}^{\left[\frac{n}{\ell}\right]}\left(\overline{X}^\circ_k-\left[\frac{n}{\ell}\right]^{-1}\sum_{j=1}^{\left[\frac{n}{\ell}\right]}\overline{X}^\circ_j\right)^2$$

으로 주어진다. Härdle 등 (2003)에 따르면 MBB의 RMSE가 작으나 NBB와 MBB의 이론적인 수렴속도는 같으며 모의 실험 결과는 거의 비슷함을 보고하였다. 본 연구에서도 NBB와 MBB의 차이는 거의 미미하여 겹침을 허락하는 붓스트랩만 모의 실험에서 다루었다. NBB의 경우도 블럭 크기는 (3.1)을 따른다.

Remark 3.2: 장기억 시계열에서의 LRV의 경우 LRD 모수 d에 의존한다. 만약 $d=0$이면 ARMA (p,q) 모형을 포함하는 단기억 시계열에서의 LRV와 일치한다. 즉 LRD에서의 LRV는 SRD를 포함하는 보다 넓은 시계열에서의 LRV이다. 따라서, SRD 시계열에 대해서 LRD 가정하에서 LRV를 추정하였을 경우, 즉 d를 추정하여 LRV를 통하여 추정하였을 때 어떠한 성능을 보이는지에 대해서 모의 실험을 통해 확인하였다.

Remark 3.3: 만약, 정상시계열의 평균의 추정에 관심이 있다면 점근적 정규성에 의해서

$$n^{\frac{1}{2}-d}\frac{\overline{X}-\mu}{\hat{s}(\hat{d})} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } n\to\infty$$

이므로 $100(1-\alpha)\%$ 신뢰구간은

$$\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\hat{s}(\hat{d})}{n^{\frac{1}{2}-\hat{d}}},\ \overline{X} + z_{\frac{\alpha}{2}}\frac{\hat{s}(\hat{d})}{n^{\frac{1}{2}-\hat{d}}}\right) \tag{3.2}$$

으로 주어진다.

Remark 3.4: 장기억 시계열은 평균변화(changes in mean) 모형에 SRD 오차가 더해진 모형과 매우 쉽게 혼동됨이 잘 알여져 있다. 예를 들어 Baek (2013), Baek과 Pipiras (2014) 및 인용한 논문들을 참조한다. 따라서 LRV의 추정에 있어서 SRD를 포함하는 LRD 모형에서의 LRV 추정량을 쓴다고 할지라도 정확한 추정을 기대하기는 힘들다. 따라서 평균변화모형에 대해서 적절한 추정을 한 다음 오차에 대해서 LRV를 추정하는 방법을 다음과 같이 제안한다. 먼저 평균변화모형을

$$X_t = m(t) + \epsilon_t, \quad t = 1,\ldots,n$$

이라고 하자. 본 논문에서는 일반적인 평균변화모형을 추정하기 위해서 비모수적인 커널평활을 이용한 함수 추정법을 다룬다. 커널 함수 추정법에서 가장 중요한 모수는 커널의 띠넓이(bandwidth)이다. 하지만 띠넓이의 추정이 에러항 $\{\epsilon_t\}$이 강한 종속 관계를 가질 때 매우 조심스러운 접근이 필요하며 비모수적인 방법으로 Chu와 Marron (1991)년 제안한 MCV가 널리 쓰인다. MCV는 통상적인 leave-one-out 교차 검증이 아니라 주어진 시점에 대해서 앞뒤로 l개의 관측치를 제거하여 강한 종속 관계를 제거한 leave-$(2l+1)$-out CV을 의미한다. 본 논문은 MCV를 커널 추정량으로 Kim 등 (2009)에서 제안한 쌍봉형 커널을 이용한 띠넓이 선택을 사용하여 leave-$(2l+1)$-out CV의 느린 계산 속도를 개선하였다. 띠넓이 선택은 다음과 같다

$$h_b = \underset{h\in(0,1)}{\operatorname{argmin}} \sum_{t=1}^{n} \left(m(t) - \widehat{m}_b(t)\right)^2$$

이고 평균 변화에 대한 추정은 쌍봉형 커널(bimodal kernel)을 이용한 Nadaraya-Watson 추정량

$$\widehat{m}_b(t) = \frac{1}{nh}\sum_{i=1}^{n} K^b\left(\frac{t-i}{h_b}\right)X_i, \quad K^b(x) = 630\left(4x^2-1\right)^2 x^4 1_{\{-\frac{1}{2}\le x\le\frac{1}{2}\}}$$

이다. 이렇게 선택된 띠넓이 h_b에 대해서 Nadaraya-Watson 추정량을 다시 구한 뒤 이를 통해 잔차

$$e_b(t) = X_t - \widehat{m}(t), \quad t = 1,\ldots,n, \tag{3.3}$$

$$\widehat{m}(t) = \frac{1}{nh_b}\sum_{i=1}^{n} K\left(\frac{t-i}{h_b}\right)X_i, \quad K(x) = 0.9375\left(1-x^2\right)^2 1_{\{-1\le x\le 1\}} \tag{3.4}$$

를 구한다. 마침내 붓스트랩 블럭 사이즈 ℓ_b은 $\{e_b\}$의 표본자기상관계수 $\hat{\rho}_b(h)$에 대해서

$$\ell_b = 2\cdot\min_h\left\{h\ge 1 \text{ such that } |\hat{\rho}_b(h)|\le\frac{2}{\sqrt{n}}\right\} \tag{3.5}$$

으로 계산한다. 따라서 LRV의 추정은 쌍봉형 커널에 기반한 잔차 (3.3)에서 블럭 사이즈 ℓ_b를 사용하여 붓스트렙 부분표본을 얻을 다음 이들의 평균의 표본분산을 구함으로써 추정할 수 있다. 이 방법을 편의상 'BB2'라고 명한다.

Table 4.1. Effect of bootstrap sample size n_B for FARIMA$(0, d, 0)$ models

n_B	d	0.1	0.2	0.3	0.4
200	BB1	0.0406	0.0460	0.1467	1.5890
	BB2	0.0448	0.0618	0.2491	1.8901
500	BB1	0.0248	0.0383	0.1511	1.5107
	BB2	0.0295	0.0582	0.2726	1.8861
1000	BB1	0.0287	0.0406	0.1369	1.5651
	BB2	0.0301	0.0638	0.2504	1.9179

4. 모의 실험

본 장에서는 모의 실험을 통해서 앞서 제 3장에서 제안한 LRD 가정하에서 붓스트랩을 이용하여 LRV 추정하는 방법의 성능을 기존에 제안되었던 HAC 추정량 (2.6) 및 MAC 추정량 (2.8)과 비교 연구한다. 성능비교의 측도로서는 MSE값을 사용하였다. MSE는 500번의 반복을 통하여 구하였다. 즉

$$\text{MSE} := \frac{1}{500}\sum_{i=1}^{500}\left\{\hat{s}^2(\hat{d}) - s^2(d)\right\}^2$$

을 HAC, MAC 방법 및 붓스트랩 방법인 BB1과 쌍봉형 커널을 이용하여 추세를 제거한 다음에 블럭의 크기를 결정하는 BB2 방법 4가지에 대해서 비교하였다.

먼저 붓스트랩 방법의 경우 붓스트랩 반복수 n_B를 정해야 하므로 LRD의 가장 대표적인 모형인 FARIMA$(0, d, 0)$ 모형

$$(1-B)^d X_t = \epsilon_t, \quad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$$

을 통해 가장 적절한 붓스트랩 반복수를 결정하였다. FARIMA$(0, d, 0)$ 모형의 스펙트럴 밀도함수는

$$\frac{\sigma^2}{2\pi}\left|1-e^{-i\lambda}\right|^{-2d} \sim \frac{\sigma^2}{2\pi}\lambda^{-2d}$$

이므로 FARIMA$(0, d, 0)$의 모수 d는 수식 (2.4)에 등장하는 LRD 모수와 일치한다. FARIMA$(0, d, 0)$의 참값 LRV는

$$s^2(d) = b_0 p(d) = \frac{\sigma^2}{2\pi}\frac{2\Gamma(1-2d)\sin(\pi d)}{d(1+2d)}$$

이 되며 표본수 $n = 1{,}000$, $\sigma^2 = 1$에 대해서 붓스트랩 반복횟수 $n_B = 200, 500, 1{,}000$에 대해서 MSE를 Table 4.1에 보고하였다. 일반적으로 LRD 모수 d의 값이 증가할수록 MSE가 증가하고 붓스트랩 반복수가 늘어날수록 MSE는 작아지는 경향이 있는데 대략 $n_B = 500$을 넘어서면 그 감소효과는 거의 미비하여 본 모의 실험에서는 $n_B = 500$을 채택하여 연구를 진행하였다.

본격적으로 우리가 제안한 방법론에 대한 성능을 살펴보기 위해서 표본 크기 $n = 1{,}000$인 FARIMA $(1, d, 0)$ 모형을 여러 조합의 AR 계수 ρ와 모수 d에 대해서 생성하여 MSE를 구하였고 Table 4.2에 정리하였다. 먼저 FARIMA$(1, d, 0)$ 모형은

$$(1-\rho B)(1-B)^d X_t = \epsilon_t, \quad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$$

으로 FARIMA$(0, d, 0)$ 모형에 AR 구조를 더한 모형으로 FARIMA$(1, d, 0)$의 모수 d는 LRD 모수와 일치하며 LRV의 참값은 다음과 같다

$$\frac{\sigma^2}{2\pi(1-\rho)^2}\frac{2\Gamma(1-2d)\sin(\pi d)}{d(1+2d)}.$$

Table 4.2. MSE of LRV estimators for FARIMA$(1, d, 0)$ models

d	ρ	-0.7	-0.3	0	0.3	0.7
0	HAC	0.1073	0.0538	·	0.3524	79.1566
	MAC	0.0280	0.0349	·	0.5945	88.0489
	BB1	0.0188	0.0243	·	0.2302	71.1980
	BB2	0.0166	0.0263	·	0.2611	73.5576
0.1	HAC	0.0266	0.0183	0.0478	0.4402	72.4214
	MAC	0.0233	0.0303	0.0062	0.4507	69.0176
	BB1	0.0100	0.0196	0.0319	0.2087	68.8826
	BB2	0.0085	0.0175	0.0335	0.2815	71.5577
0.2	HAC	0.0064	0.0210	0.0816	0.6676	81.1261
	MAC	0.0261	0.0387	0.0119	0.3870	41.5722
	BB1	0.0060	0.0288	0.0374	0.3433	87.1400
	BB2	0.0075	0.0189	0.0551	0.5374	88.9718
0.3	HAC	0.0148	0.0651	0.2387	1.5904	112.7612
	MAC	0.0487	0.1028	0.0983	0.6521	17.9897
	BB1	0.0051	0.0343	0.1417	1.3468	133.5036
	BB2	0.0194	0.0620	0.2493	1.8563	133.5824
0.4	HAC	0.1415	0.4683	1.4415	7.1915	286.7730
	MAC	0.8768	2.6680	5.5266	14.5070	8.7998
	BB1	0.0490	0.4330	1.5694	8.6686	350.3549
	BB2	0.1774	0.6051	1.9479	9.5863	357.4557

MSE = mean squared errors, LRV = long-run variance, HAC = heteroskedasticity and autocorrelation covariance, MAC = memory and autocorrelation consistent.

먼저 $d = 0$인 경우, 즉 SRD 시계열인 AR(1) 모형의 경우 LRD 모수를 추정하여 대입한 LRV 추정량의 경우를 나타낸다. $\rho = 0$인 경우는 i.i.d. 정규분포를 따르는 확률변수로 고려하지 않았다. 상관의 정도가 음수 혹은 약할 경우 비록 LRD 모수를 추정한다 할지라도 작은 MSE값을 보여 안정적으로 추정이 잘 됨을 알 수 있다. 하지만 기존의 여러 연구들이 지적하였듯이 $\rho \approx 1$인 경우 LRV 추정량의 성능은 확연히 떨어짐을 확인할 수 있다. 방법론들을 서로 비교해 보자면 우리가 제안한 BB1 방법이 HAC 및 MAC 방법과 비교하여 작은 MSE를 보여 성능이 우수함을 확인할 수 있었으며 쌍봉형 커널을 통해 추세를 추정한 BB2의 방법도 BB1과 거의 비슷한 성능을 보여 AR(1) 모형의 추세인 평균이 0인 직선을 잘 추정하고 있음을 보여준다.

의존성을 결정하는 d가 .2보다 작은 경우, 즉 중등도의 종속 관계를 가지는 LRD 모형의 경우, BB1 및 BB2 모형이 다른 방법들과 비교하여 뒤쳐지지 않음을 알 수 있다. 특히 ρ가 음의 값인 경우는 BB를 사용한 추정량들이 좋은 성능을 보이는 것을 확인할 수 있다. 하지만, AR 계수 ρ가 매우 크거나 d가 .5에 가까워 그 상관관계가 매우 큰 경우에는 붓스트랩에 기반한 방법보다는 MAC 추정량이 다른 추정량보다 작은 MSE를 산출하였다. 이는 MAC 추정량의 경우 $p(d)$라는 상수항 부분을 모수적인 추정을 하게 되어 비모수적으로 추정하는 HAC 혹은 BB 방법에 비해서 효율적이기 때문이라 판단된다. 반대로 d가 매우 크거나 ρ가 1에 가까운 경우 HAC 혹은 BB 방법이 좋은 성능을 보이기위해서는 더 많은 표본이 필요하다.

LRV의 가장 기본적인 활용은 전체 평균 μ에 대한 추론이다. 따라서 $100(1-\alpha)\%$ 점근적 신뢰구간에 대해서 평균을 포함하는지에 대한 포함확률(coverage probability)이 LRV 추정량의 성능측도로 쓰일 수 있다. 표본 크기 $n = 1{,}000$이고 FARIMA$(1, d, 0)$를 따르는 모형에 대해서 수식 (3.2)에 기반하여 평

Table 4.3. Coverage probabilty for μ based on the t-ratio

d	ρ	-0.7	-0.3	0	0.3	0.7
0	HAC	0.934	0.910	·	0.984	1
	MAC	0.864	0.922	·	0.940	0.994
	BB1	0.820	0.894	·	0.980	1
	BB2	0.794	0.878	·	0.986	1
0.1	HAC	0.824	0.814	0.890	0.982	1
	MAC	0.820	0.868	0.916	0.968	0.998
	BB1	0.752	0.844	0.942	0.992	1
	BB2	0.732	0.828	0.924	0.986	1
0.2	HAC	0.684	0.754	0.822	0.918	1
	MAC	0.810	0.892	0.902	0.916	1
	BB1	0.714	0.862	0.900	0.962	1
	BB2	0.634	0.810	0.858	0.940	0.998
0.3	HAC	0.602	0.628	0.782	0.902	1
	MAC	0.856	0.892	0.930	0.964	1
	BB1	0.714	0.836	0.862	0.926	0.982
	BB2	0.572	0.666	0.780	0.876	0.976
0.4	HAC	0.484	0.588	0.676	0.846	0.994
	MAC	0.864	0.928	0.942	0.966	1
	BB1	0.680	0.610	0.644	0.742	0.930
	BB2	0.410	0.482	0.534	0.654	0.850

HAC = heteroskedasticity and autocorrelation covariance, MAC = memory and autocorrelation consistent.

균에 대한 95% 점근적 신뢰구간을 구한다음에 참값인 $\mu = 0$을 몇번 포함하는지 500번의 반복을 통해서 구한 경험적 포함확률을 구하여 Table 4.3에 보고하였다. 그 결과 붓스트랩에 기반한 방법인 BB1의 경우 LRD 모수 d가 클 경우 일반적으로 HAC 추정량보다 비슷하거나 대체로 더 좋은 포함확률을 보임을 살펴볼 수 있다. 추세를 추정하는 BB2의 경우 BB1과 비교하여 성능이 약간 떨어지지만, 참값을 사용하는 HAC보다는 더 좋은 결과를 줌을 알 수 있다. 가장 높은 정확성을 보여주는 경우는 MAC 추정량으로 ρ나 d가 증가하여 매우 강한 종속관계를 보일수록 타 방법과 비교하여 좋은 포함확률을 기록하였다. 이 역시 앞선 모의 실험에서처럼 MAC 추정량이 상수항 $p(d)$에 대해서는 모수적인 추정을 하고 스펙트럴 밀도함수에 대해서는 비모수적인 추정을 하는 준모수적인 추정(semiparametric estimation) 방법이기 때문에 모의실험에서 쓰인 모수적인 LRD 모형인 FARIMA$(1, d, 0)$에서 비모수적인 HAC 및 붓스트랩 방법과 비교하여 좋은 성능을 보인것으로 본다.

마지막으로 우리가 고려한 모의 실험은 다음과 같다. 이미 Remark 3.4에서 지적하였듯이 LRD 모형은 평균변화모형에 SRD 에러가 더해진 모형과 유한표본에서 매우 쉽게 혼동이 된다. 이를 극복하기 위해서 쌍봉형 커널을 이용하여 추세를 추정하고 붓스트랩의 블럭 크기를 결정하는 BB2방법을 제안하였다. 그 성능을 검증하기 위해서 Mills (2007)에서 등장하는 평균변화모형

$$\begin{aligned} m(t) &= -14.24 + 0.00531t(14.77 - 0.00149t)S_{1t} + (-5.72 + 0.00378t)S_{2t}, \\ &S_{1t} = [1 + \exp\{0.003(t - 1450)\}]^{-1}, \quad S_{2t} = [1 + \exp\{-0.004(t - 456)\}]^{-1}, \\ &t = 1, \ldots, 2000 \end{aligned} \tag{4.1}$$

에 AR(1) 모형의 에러를 추가한 모형에서의 LRV 추정을 비교하였다. AR(1) 모형을 생성하는데 필요

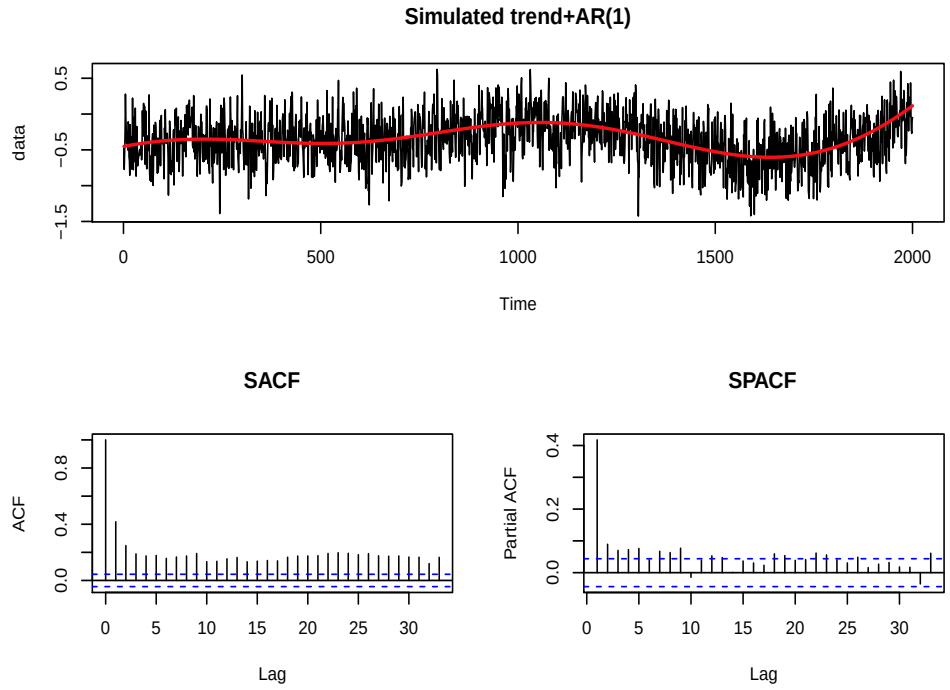

Figure 4.1. Realization of trend plus AR(1) model.

Table 4.4. MSE of LRV estimators for trend plus AR(1) model

ρ	-0.7	-0.3	0.3	0.7
HAC	.00021	.0008	.0134	.639
MAC	.0457	.0071	.0022	.663
BB1	.00001	.0003	.0022	.526
HAC2	.0083	.0086	.0122	.502
MAC2	.0064	.0072	.0027	.676
BB2	.0041	.0026	.0033	.412

MSE = mean squared errors, LRV = long-run variance, HAC = heteroskedasticity and autocorrelation covariance, MAC = memory and autocorrelation consistent.

한 이노베이션의 분산은 .3이다. Figure 4.1은 이렇게 생성된 자료의 시계열도와 표본자기상관, 부분자기상관함수 그림이다. 시차가 30이 넘어도 종속성이 살아남아 LRD 모형이 쉽게 혼동이 됨을 알 수 있다.

이러한 자료에 대해서 LRV를 추정하기 위해서 먼저 평균변화 추세를 추정하지 않고서 LRV를 추정하는 방법 세 가지와 쌍봉형 커널을 이용하여 추세를 제거한 뒤 LRV를 추정한 방법의 MSE를 Table 4.4에 정리하였다. 추세 제거 유무에 대한 혼동을 피하기 위해서 추세를 제거한 뒤 잔차에 대해서 LRV를 추정한 방법을 HAC2, MAC2 그리고 BB2라고 표시하였다. AR(1)의 계수 ρ가 음수여서 양의 종속관계를 가지지 못하는 경우에는 추세제거의 유무가 LRV의 추정에 큰 영향을 주지 않는다. 하지만 ρ값이 증가하여 LRD 시계열인 것처럼 보이기 시작하면 추세를 제거한 방법들이 더 작은 MSE를 주었으며 BB2의 방법이 HAC 혹은 MAC 방법과 비교하여 작은 MSE를 기록하여 좋은 성능을 보여주었다.

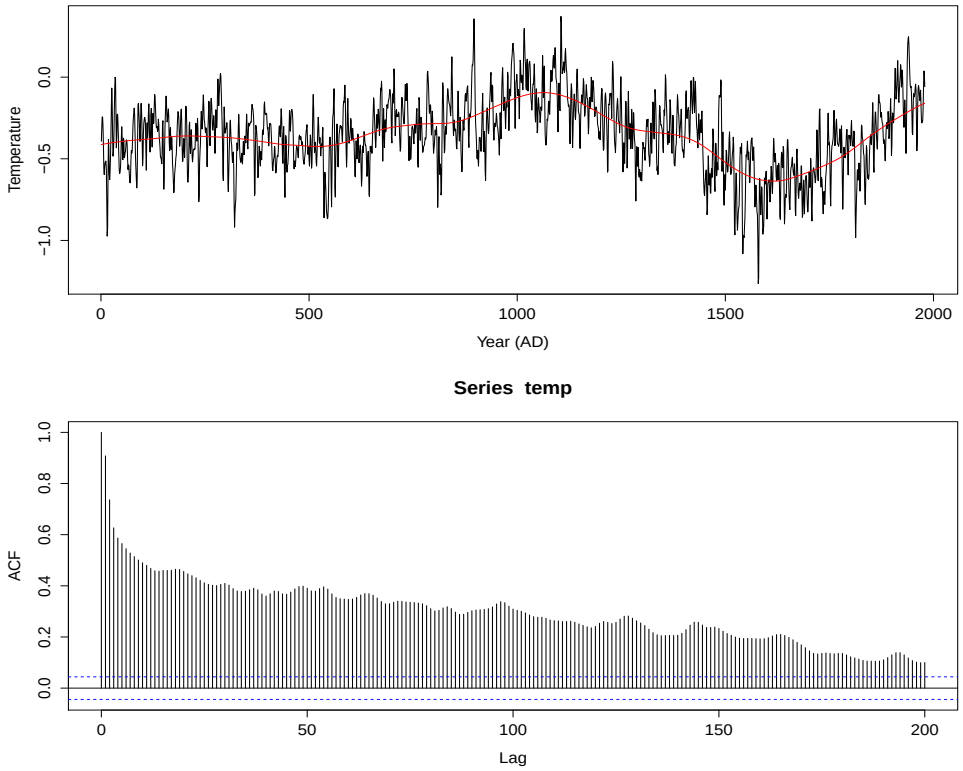

Figure 5.1. Rconstructed Northern Hemesphere temperatures and sample autocorrelation function.

5. 실증자료 분석

본 논문이 제안한 방법론의 실생활 응용으로 Mills (2007)에서 사용된 AD 1년부터 1979년까지 재구성된 연간 북반구 기온 데이터를 사용하여 실증자료 분석을 시행하였다. 재구성된 자료는 Moberg 등 (2005)이 구현한 것으로 실측 자료가 존재하지 않은 먼 과거의 온도는 나이테와 같이 잘 드러나는 정보와 깊은 호수나 바다 퇴적물에서 얻을 수 있는 잘 드러나지 않는 정보를 결합하여 서기 1년부터 기온 자료를 재구성해서 구하였다. 기상학의 전통에 따라 1961년부터 1990년까지의 평균 온도와의 차이(anomaly)로 주어지며 섭씨온도를 사용한다.

Figure 5.1은 자료의 시계열도와 표본자기상관회귀함수 그림을 나타낸다. 시차가 200에서도 매우 강한 종속성을 보이고 있음이 명확하고 비모수 함수 추정에서 이러한 강한 종속성을 제거하기 위해서 쌍봉형 커널을 사용하여 띠넓이를 선택하고 Nadaraya-Watson 추정량 (3.4)을 시계열도에 첨부하였다. 그림에서 살펴보듯 매우 뚜렷한 추세가 관측됨을 알 수 있으며 1700년도 이후에는 꾸준히 증가되어 왔음을 관찰할 수 있다. 따라서 이러한 강한 종속성이 추세함수 때문인지 아니면 LRD 시계열의 특성 때문인지 많은 논란이 지난 수십년동안 진행되었다. 온도 변화 자료 분석의 궁극적인 목적은 미래의 온도에 대해서 추정을 하는 것에 있다. 또 이러한 추정을 토대로 지구 온난화가 실존하는지, 실존한다면 어떤 속도로 진행되고 있는지가 과학적 핵심 문제이다.

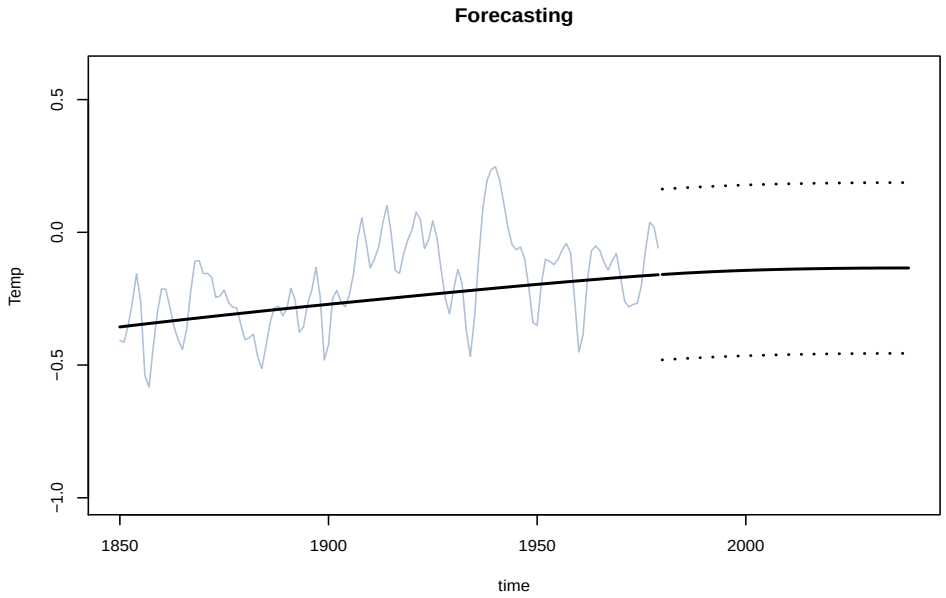

Figure 5.2. Forecasting of NH temperature from 1980 to 2040.

먼저 h 시차 온도 변화에 대한 예측을 단측커널(one-sided kernel)을 이용하여 다음과 같이 계산하였다.

$$\widehat{m}_f(n+h) = \frac{1}{nh_b}\sum_{i=1}^{n+h-1} K\left(\frac{n+h-i}{h_b}\right) X_i, \quad K(x) = 1.875\left(1-x^2\right)^2 1_{\{-1\le x\le 0\}}$$

$$X_{n+t} = \widehat{m}_f(n+t), \quad t = 1,\ldots,h-1.$$

그리고 이에 대한 $100(1-\alpha)\%$ 예측구간(prediction interval)은 점근적으로

$$\left(\widehat{m}_f(n+h) - z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2 + \frac{\hat{s}^2(\hat{d})}{n^{1-2\hat{d}}}},\ \widehat{m}_f(n+h) + z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2 + \frac{\hat{s}^2(\hat{d})}{n^{1-2\hat{d}}}}\right)$$

으로 근사시킬 수 있으며 $\hat{\sigma}^2$은 잔차의 제곱합의 평균으로 추정이 가능하다. 본 논문에서 고려한 LRV의 추정값은 $\hat{d} = .224$에 대해서 $\hat{s}_H^2(\hat{d}) = 0.0315$, $\hat{s}_M^2(\hat{d}) = 0.0309$, $\hat{s}_B^2(\hat{d}) = 0.0348$로 주어지며 $\hat{\sigma}^2 = .026$으로 주어져 LRV 추정값의 차이가 적으며 블럭 붓스트랩 LRV를 이용한 95% 예측구간을 계산하여 Figure 5.2에 나타냈다. 먼저 $\hat{d}$값이 .2에 가까워 추세를 제거한 온도 변화는 추세와 LRD 모두 영향을 받고 있으며, 향후 60년간 평균 온도는 완만한 증가추세를 보이고 있음을 확인할 수 있다.

6. 결론

본 논문은 블럭 붓스트랩을 이용하여 장기적 분산을 추정하는 방법에 대해서 논의하였다. 기존의 SRD 시계열에 대한 아이디어를 확장하여 LRD 시계열에서의 적용에 대해서 논의하였다. 또한 LRD 시계열이 평균편화에 SRD 에러가 더해진 모형과 매우 쉽게 잘 혼동됨에 따라 쌍봉형 커널을 이용하여 비모수적으로 추세를 추정하고 잔차를 이용하여 LRV를 추정하는 방법을 제안하였다. 모의 실험 결과 우리가

제안한 붓스트랩에 기반한 방법이 대체로 LRV를 잘 추정하였으나 종속관계가 매우 클 경우에는 MAC 추정량보다 성능이 좋지는 못하였다. 이는 MAC 추정량이 준모수적 방법의 특성을 가지고 있기 때문이라 판단된다. 따라서 실증 자료 분석에서는 하나의 LRV 추정량에만 의존하지 말고 여러 추정량을 동시에 살펴보면서 특히 MAC 추정량과 차이가 많이 나올 경우에는 더 세심한 주의를 기울여야 할 것이다. 또한 LRV 추정량의 일상생활 응용 자료로 서기 1년에서부터 1979년까지 재구성된 북반구의 평균 온도 자료를 분석하여 향후 60년에 대한 95% 예측구간을 제공하였다. 그 결과 온도 변화는 평균변화와 LRD 성질 모두를 가지는 것으로 나타나 Mills (2007)가 주장한 것과는 달리 두 모형 모두를 동시에 고려해야하는 자료임으로 밝혀져 후속 연구가 필요할 것으로 본다.

References

Abadir, K. M., Distaso, W., and Giraitis, L. (2009). Two estimators of the long-run variance: beyond short memory, *Journal of Econometrics*, **150**, 56–70.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, **59**, 817–858.

Baek, C. (2013). Time series modelling of air quality in Korea: long range dependence or changes in mean?, *The Korean Journal of Applied Statistics*, **26**, 987–998.

Baek, C. and Pipiras, V. (2014). On distinguishing multiple changes in mean and long-range dependence using local Whittle estimation, *Electronic Journal of Statistics*, **8**, 931–964.

Beran, J. (1994). *Statistics for Long-Memory Processes*, **61**, CRC press, New York.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*, SIAM.

Chu, C. K. and Marron, J. S. (1991). Comparison of two bandwidth selections with dependent errors, *The Annuals of Statistics*, **19**, 1906–1918.

Härdle, W., Horowitz, J., and Kreiss J.-P. (2003). Bootstrap methods for time series, *International Statistical Review*, **71**, 435–459.

Kim, T. Y., Park, B. U., Moon, M. S., and Kim, C. (2009). Using bimodal kernel for inference in nonparametric regression with correlated errors, *Journal of Multivariate Analysis*, **100**, 1478–1497.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations, *Annals of Statistics*, **17**, 1217–1241.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*, Springer.

Mills, T. C. (2007). Time series modelling of two millennia of northern hemisphere temperatures: long memory or shifting trends?, *Journal of the Royal Statistical Society, Series A*, **170**, 83–94.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlen, W. (2005). Highly variable northern hemisphere temperatures reconstructed from low- and high-resolution proxy data, *Nature*, **433**, 613–617.

Newey, W. K. and West, K. (1987). A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703–708.

Politis, N. D. and Romano, J. P. (1995). Bias-corrected nonparametric spectral estimation, *Journal of Time Series Analysis*, **16**, 67–103.

Politis, N. D. and White, H. (2004). Automatic block-length selection for the dependent bootstrap, *Economic Reviews*, **23**, 53–70.

Robinson, P. M. (2005). Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction, *Econometric Theory*, **21**, 171–180.

장기간 의존 시계열에서 붓스트랩을 이용한 장기적 분산 추정

백창룡[a,1] · 권용[a]

[a]성균관대학교 통계학과

(2016년 1월 13일 접수, 2016년 2월 5일 수정, 2016년 2월 29일 채택)

요 약

본 논문은 시계열 분석의 추론에서 매우 중요한 역할을 하는 장기적 분산에 대해서 붓스트랩을 이용한 추정을 다룬다. 본 논문은 기존의 방법을 두가지 측면에서 확장한다. 첫째, 단기억 시계열에서의 장기적 분산 추정을 확장하여 자료의 의존성이 매우 강한 장기간 의존 시계열에서 붓스트랩을 이용한 장기적 분산의 추정에 대해서 논의한다. 또한 장기간 의존 시계열이 평균변화모형과 매우 쉽게 잘 혼동됨이 잘 알려져 있기에 이를 해결하기 위해서 쌍봉형 커널을 이용한 추세 추정 및 붓스트랩의 블럭을 결정하는 방법을 제안한다. 모의 실험결과 제안한 방법이 매우 유의하였으며 북반구 평균 온도 변화 자료 분석으로 실증 자료 예제도 아울러 제시하였다.

주요용어: 장기적 분산, 장기간 의존 시계열, 블럭 붓스트랩

이 논문은 2014년도 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (NRF-2014R1A1A1006025).

[1]교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: crbaek@skku.edu