응용통계연구 제 19권 1호, 2006년, pp. 121–134

# 시계열분석을 위한 주파수 공간상에서의 재표집 기법*

여인권[1] 윤화형[2] 조신섭[3]

요 약

이 논문에서는 이산코사인변환을 이용하여 시계열자료를 주파수 공간으로 변환시킨 후, 이산코사인변환 계수를 재표집하여 시계열자료에 대한 재표본을 추출하는 방법에 대해 알아본다. 기존 주파수 공간상에서의 붓스트랩 방법은 스펙트럼평균(spectral mean)에 대한 추론을 하기위해 사용되지만 제안하고자 하는 방법은 시간영역상에서의 시계열자료에 얻을 수 있다는 것이 가장 큰 차이점이다. 이 논문에서는 정상시계열의 경우, 이산코사인변환 계수의 통계적 성질을 유도하고 이 성질을 이용하여 붓스트랩하는 과정을 설명한다. 모의실험을 통해 기존에 사용되고 있는 방법과 성능을 비교하였다.

주요용어: 붓스트랩, 이산코사인변환. 점근적 독립

## 1. 서론

Efron(1979)에 의해 이론적 토대가 마련된 이후, 붓스트랩(bootstrap)은 1980년대와 90년대 통계학에서 가장 활발하게 연구되었던 분야 중 하나로 통계학의 전 분야에서 다양하게 개발되고 활용되는 방법이다. 독립적인 표본들에 대해 붓스트랩은 통계적으로 비슷한 성질을 가지는 재표본(resample)을 표본들로부터 랜덤하게 추출하여 통계 추론에 필요한 통계량의 적률이나 분포를 근사하는데 사용된다. 문제는, Singh(1981)이 지적한 것처럼, 시계열자료나 공간통계(spatial statistics)자료와 같이 자료들 간에 독립성이 만족되지 않는 상황에서 재표본을 랜덤하게 추출하면 표본들간의 연관성이 무시되기 때문에 원래의 표본과 통계적인 성질이 비슷한 재표본을 추출하기 어렵다는 것이다. 이러한 문제를 해결하기 위한 많은 연구들이 진행되고 있는데 Kunsch(1989)의 block bootstrap(이하 BB)과 Buhlmann(1997, 1998)의 sieve bootstrap(이하 SB) 등이 시계열 자료에 대한 대표적인 재표집 방법으로 사용되고 있다.

BB는 Hall(1985)의 아이디어를 Kunsch(1989)가 붓스트랩에 적용한 방법으로, 정상 시계열 자료를 여러 개의 블록으로 분할한 다음 블록을 표본처럼 랜덤하게 선택하고 선택한

* 이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2003-002-C00043)
이 논문은 2005년 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음. (KRF-2005-070-C00022)

1) (140-742) 서울시 용산구 청파동 2가, 숙명여자대학교 이과대학 수학통계학부, 조교수
E-mail: inkwon@sookmyung.ac.kr
2) (151-747) 서울시 관악구 신림동 산56-1, 서울대학교 자연과학대학 통계학과, 박사과정
E-mail: whyya@chollian.net
3) (151-747) 서울시 관악구 신림동 산56-1, 서울대학교 자연과학대학 통계학과, 교수
E-mail: sinsup@snu.ac.kr

순서대로 블록을 결합하여 재표본을 얻는 방법이다. 이 방법의 문제점은 블록과 블록 간에 연관성이 단절될 수 있으며 자료의 연관성과 블록의 개수 간에 반비례의 관계가 성립한다는 것이다. 즉, 블록의 개수가 많아지면 다양한 재표본을 얻을 수 있지만 블록에 의한 단절이 많아져 자료들 간의 관계를 충분히 설명하지 못하는 반면 블록의 개수가 작으면 자료들 간의 관계는 잘 설명되지만 다양한 형태의 재표본을 얻을 수 없다는 단점이 있다. 적절한 블록의 수를 결정하기 위한 연구가 Buhlmann and Kunsch(1999) 등에 의해 진행되어 왔지만 이 방법이 가지고 있는 근본적인 문제를 해결하지 못하고 있다.

SB 방법은 모수 또는 준모수적 모형들의 집합에서 원자료에 맞는 근사 모형을 선택하고 적합시킨 후 회귀분석에서의 붓스트랩 방법과 같이 잔차를 재표집 함으로써 재표본을 추출하는 방법이다. 이 방법은 BB 방법에서 발생했던 블록간의 단절성과 블록의 개수에 대한 문제를 해결하였다. DNA의 염기수열과 같은 범주형 시계열분석을 위한 variable length Markov chains SB 방법들도 여기에 속한다. 그러나 이 방법에서는 붓스트랩의 효율성이 선택된 모형에 영향을 받기 때문에 모형 선택이 잘못되는 경우 원자료의 특성과 상이한 재표본을 얻을 수도 있다. 일반적으로 SB방법에서는 적절한 근사 모형의 선택하기 위해 상당히 큰 차수의 AR(p)모형을 사용하고 있는데 문제는 이러한 모형으로도 적합이 잘 되지 않는 경우에 사용하기 어렵다는 것이다. 또한 분산추정이 정확하지 않기 때문에 보다 정확한 분산추정을 위해서는 Choi and Hall(2000)이 언급한 이중붓스트래핑(double boostrapping) 같은 복잡한 알고리즘을 적용해야 하는 단점이 있다.

주파수 영역에서의 기존 연구에서는 복잡한 종속성이 존재하는 시계열 관측값들이 거의 독립적인 통계량인 주기도 도표(periodogram ordinates)로 변환될 수 있다고 가정하며 변환된 값들에 대해 독립인 표본에 적용했던 붓스트랩 방법을 사용하고 있다. Franke and Hardle(1992)는 주기도와 스펙트럼 밀도(spectral density)간의 관계가 승법회귀모형(multiplicative regression model)에 의해 근사될 수 있다는 근거 하에서 추정된 스펙트럼 밀도에 추정된 주파수 영역 잔차의 집합으로부터 복원 추출한 붓스트랩 오차를 곱하여 주기도의 재표본을 구하는 방법을 제안하였으며 Dahlhaus and Janas(1996)와 Paparoditis and Politis(1999) 등을 포함한 많은 연구자들에 의해 수정, 보완, 확장되어 연구되고 있다.

종속성이 존재하는 자료에 대한 재표집 방법의 점근적 이론들은 Lahiri(2003)를 참조하기 바란다. 이 논문은 시계열자료에서 자료들 간에 종속관계를 충분히 고려할 수 있으면서도 다양한 형태의 재표본을 얻을 수 있는 새로운 재표집 방법에 대한 연구를 제안하고자 한다. 제안하고자 하는 방법은 시계열자료를 이산코사인변환을 이용하여 주파수 영역으로 변환시킨 후 이산코사인변환 계수를 이용하여 재표본을 구한다. 이 과정을 효과적으로 설명하기 위해 이 논문에서는 다음과 같은 구성으로 이루어져 있다. 2절에서는 통계학에서는 생소한 이산코사인변환에 대해 간단히 알아보고 3절에서는 정상시계열의 경우 이산코사인변환 계수의 통계적 성질과 재표집 방법에 대해 소개한다. 모의실험을 통해 기존방법과 비교하여 제안된 방법의 성능이 얼마나 우수한지를 4절에서 고찰한다. 마지막 절에서는 이후 추가해야하거나 개선해야할 문제에 대해 알아본다.

## 2. 이산주파수변환

시계열자료 분석은 크게 시간영역(time domain)에서의 분석방법과 주파수영역(frequency domain)에서의 분석방법으로 나눌 수 있다. 주파수 영역에서의 시계열 분석은 시간 영역에서 파악하기 어려운 자료의 특징을 유도할 수 있다는 장점이 있지만 대부분 복소수 값으로 표시되는 푸리에변환(Fourier transform)을 기초로 이루어지기 때문에 분석 결과를 해석하는데 쉽지 않다는 단점이 있다. 멀티미디어 분야에서는 푸리에변환의 난해함을 해결하기 위해 이산코사인변환(discrete cosine transform, 이하 DCT)이나 이산웨이브렛변환(discrete wavelet transform)과 같은 새로운 툴을 개발하여 사용해 오고 있다. 이런 이산주파수변환들은 음성, 오디오, 영상 및 동영상과 같은 멀티미디어 콘텐츠를 압축하는 기술을 개선시키는데 큰 역할을 하였다. 이 논문에서는 사용이 용이하면서 이해하기 쉽고 기능면에서 우수한 DCT를 중심으로 주파수 영역상에서의 재표집 방법에 대해 설명하고자 한다. DCT는 오디오와 동영상 압축에서 가장 많이 사용되고 있는 MPEG과 영상 압축의 JPEG에서 기본이 되는 주파수변환이다. DCT는 코사인함수 대신 지수함수를 사용하는 이산푸리에변환(discrete Fourier transform, 이하 DFT)과 밀접한 관계가 있지만 적은 수의 계수로 시계열자료에서의 주요 에너지를 나타낼 수 있는 측면에서 DFT보다 좋은 에너지압축(energy compaction) 성질을 가지는 것으로 알려져 있다.

임의의 시계열 자료 $x_1, x_2, \ldots, x_n$가 있을 때, DCT의 계수는 다음과 같은 가중합으로 정의된다.

$$F_j = w_j \sum_{k=1}^{n} x_k \cos\left\{\frac{\pi}{2n}(2k-1)(j-1)\right\}, \quad j = 1, 2, \ldots, n.$$

여기서 $w_j$는 $j = 1$일 때 $1/\sqrt{n}$이고 그 외의 값에서는 $\sqrt{2}/\sqrt{n}$의 값을 가진다. 그림 1은 $n = 9$일 때, 계수 $F_j$를 계산하기 위해 사용된 9개의 가중값, $\cos\{\pi(2k-1)(j-1)/(2n)\}$, $k = 1, 2, \ldots, 9$,를 표시한 것이다. 첫 번째 계수인 $F_1$을 특별히 DC 성분(DC component)이라고 하는데, 가중치가 모두 1이 되기 때문에 $F_1$는 자료의 평균에 $\sqrt{n}$을 곱한 값이 된다. $F_2$의 경우 자료의 감소 또는 증가 추세에 대한 전반적인 특징을, $F_3$은 감소하다가 증가하는 또는 증가하다가 감소하는 형태의 특징 등을 나타내는 것을 볼 수 있다. 첨자 $j$가 작을수록 코사인함수의 주기가 길어지며 해당 계수는 시계열자료에 있어 장기간의 변화를 나타내는 저주파 대역의 특성을 나타내고 반대로 $j$가 커질수록 주기가 짧아져 단기간의 변화를 나타내는 고주파 대역의 특성을 나타낸다. 일반적인 주파수 변환과 마찬가지로 전체 또는 부분 DCT 계수를 이용하여 역으로 시계열 자료를 재구성할 수 있는데 이 때 사용되는 변환이 역이산코사인변환(inverse DCT, 이하 IDCT)이다. 역이산코사인변환은 다음과 같이 정의된다.

$$x_k = \sum_{j=1}^{n} w_j F_j \cos\left\{\frac{\pi}{2n}(2k-1)(j-1)\right\} \quad k = 1, 2, \ldots, n.$$

오디오나 영상과 같이 근접자료들 간에 상당히 높은 양의 상관관계를 가지는 경우 일반적으로 저주파수 대역에서 큰 값을 가지며 고주파수 대역에는 주로 백색잡음(white noise)과

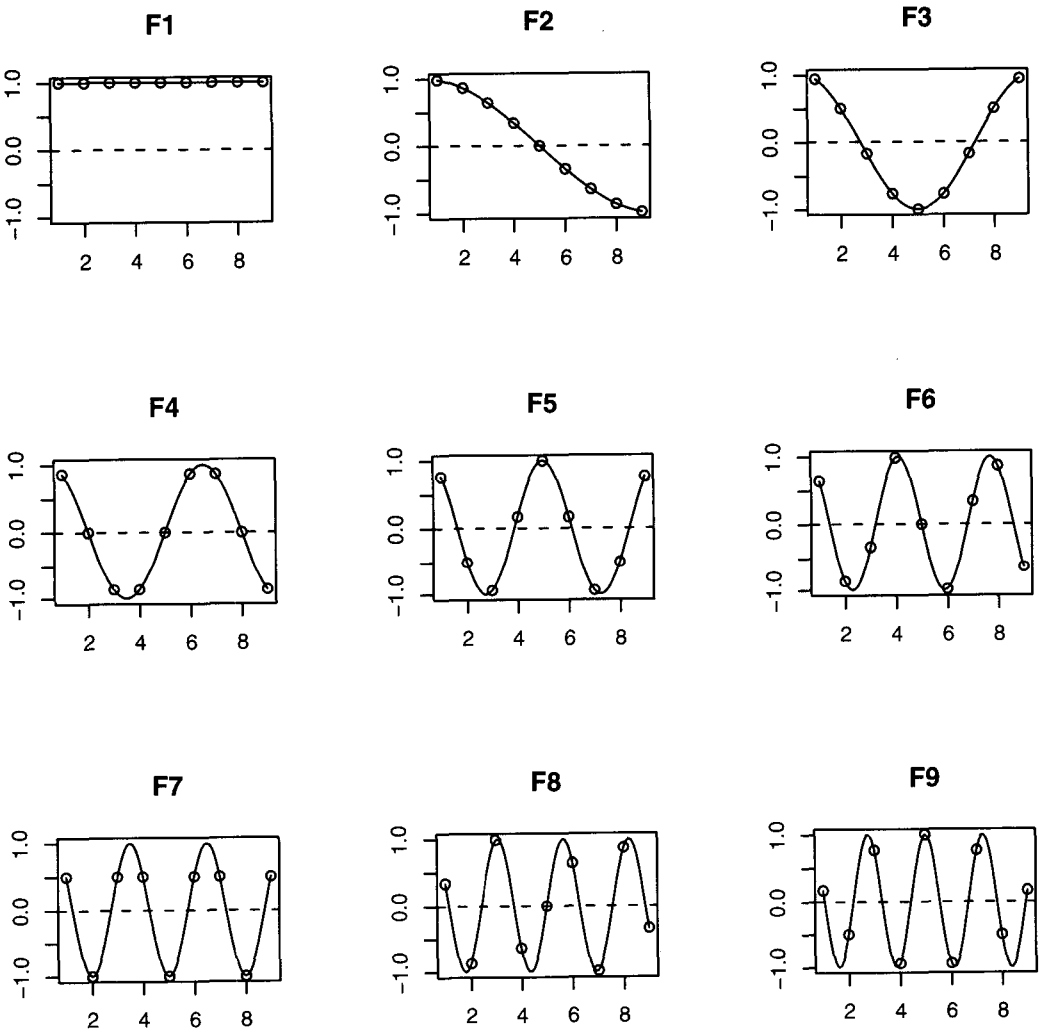


그림 2.1: 이산코사인변환의 가중치

같이 큰 의미가 없는 신호들이 표시된다. 이런 경우 저주파 대역에 대한 해석만으로도 시계열자료에 대한 특성을 충분히 설명할 수 있다. 멀티미디어에서는 고주파 대역의 계수를 제거하거나 0으로 처리하여 멀티미디어 콘텐츠를 압축한다. 주식의 일별수익률과 같은 경제자료들은 음의 상관관계를 가지는 경우가 많은데 이런 자료의 DCT 계수를 확인해 보면 고주파 대역에서도 큰 값을 가지는 경향이 있다. 이것은 자료에 따라 각 주파수 대역의 특징이 달라질 수 있기 때문에 특정 주파수 대역을 한정하여 분석하는 것을 문제가 될 수 있다는 것을 의미한다. 그림 2에는 백색잡음과 정상시계열인 AR(1)과 MA(1) 시계열과정에서 자료 50개를 100번 생성시켰을 때 각 주파수 대역에서의 DCT 계수와 (왼쪽)과 전체 DCT 계수의 히스토램(오른쪽)을 표시한 것이다.

그림 2의 윗쪽에 있는 것은 평균이 0이고 분산이 1인 정규분포를 따르는 백색잡음 과정에 대한 DCT분포를 나타낸 것인데 그림에서 보는 것과 같이 백색잡음의 경우, 각 대역에서의 DCT 계수에는 큰 차이를 보이지 않고 있으며 전체 주파수 대역에서 골고루 에너지를 가지고 있는 것을 볼 수 있다. 또한 히스토그램의 경우에도 큰 이상점이 없는 정규분포에 근사하는 것을 볼 수 있다.

중간에 있는 그림은 AR(1) 모형을 따르는 확률과정으로 구조식은 $X_t = \phi X_{t-1} + \varepsilon_t$ 형태를 가지는데 모의실험에서는 $\varepsilon_t$를 평균이 0이고 분산이 1인 정규백색잡음을 따른다고 가정하였으며 $\phi$는 0.5로 지정하였다. 모수 $\phi$가 양수이면 시계열들 간에 양의 상관관계가 존재하여 시계열 에너지의 많은 부분이 저주파수 대역에 집중되는 것을 볼 수 있다. 즉 저주파수 대역의 계수 값은 커지는 반면 고주파수 대역의 값은 상대적으로 작다.

아래에 있는 그림은 모형 $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$에 대한 DCT 계수의 분포를 나타낸다. 여기서 $\theta$는 0.5가 사용되었다. 위의 AR(1)과 반대로 음의 1차 자기상관관계가 되기 때문에 고주파수 대역의 값이 커지고 저주파수 대역의 계수 값은 상대적으로 작아지는 것을 볼 수 있다. 히스토그램의 경우에도 자기회귀모형의 것과 큰 차이를 보이지 않는 것으로 나타났다. 그림 2의 히스토그램을 보고 판단하건데 정상 시계열의 경우 DCT 계수의 분포는 정규분포에서 크게 벗어나지 않는 것으로 생각된다. 자기상관관계는 주파수 대역에서의 DCT 계수 값의 크기에 영향을 주는 것으로 생각된다.

## 3. 정상 시계열에 대한 주파수 영역상에서의 재표집

시계열 자료가 정상성을 만족하는 경우에 DCT 계수는 다음과 같은 성질을 만족한다.

정리 3.1 시계열 $\{X_t\}_{t=1}^n$가 평균이 $E(X_1) = \mu$이고 자기공분산이 $cov(X_i, X_j) = \gamma_{|i-j|}$인 정상시계열(stationary time series)이면, DCT 계수는 다음과 같은 통계적 성질을 가진다.

1 $F_j$들은 점근적으로 독립(asymptotically independent)인 변수이다.

2 $E(F_1) = \mu/\sqrt{n}$이고, $j = 2, \ldots, n$에 대해서, $E(F_j) = 0$이다.

3 $var(F_1) = \sigma_1^2 = \gamma_0 + 2\sum_{i=1}^{n-1}\{(1 - i/n)\gamma_i\}$이고, $j = 2, \ldots, n$에 대해서,

$$var(F_j) = \sigma_j^2 = \gamma_0 + \frac{4}{n}\sum_{t_1<t_2}\left\{\cos(\omega_{jt_1})\cos(\omega_{jt_2})\gamma_{|t_1-t_2|}\right\} \tag{3.1}$$

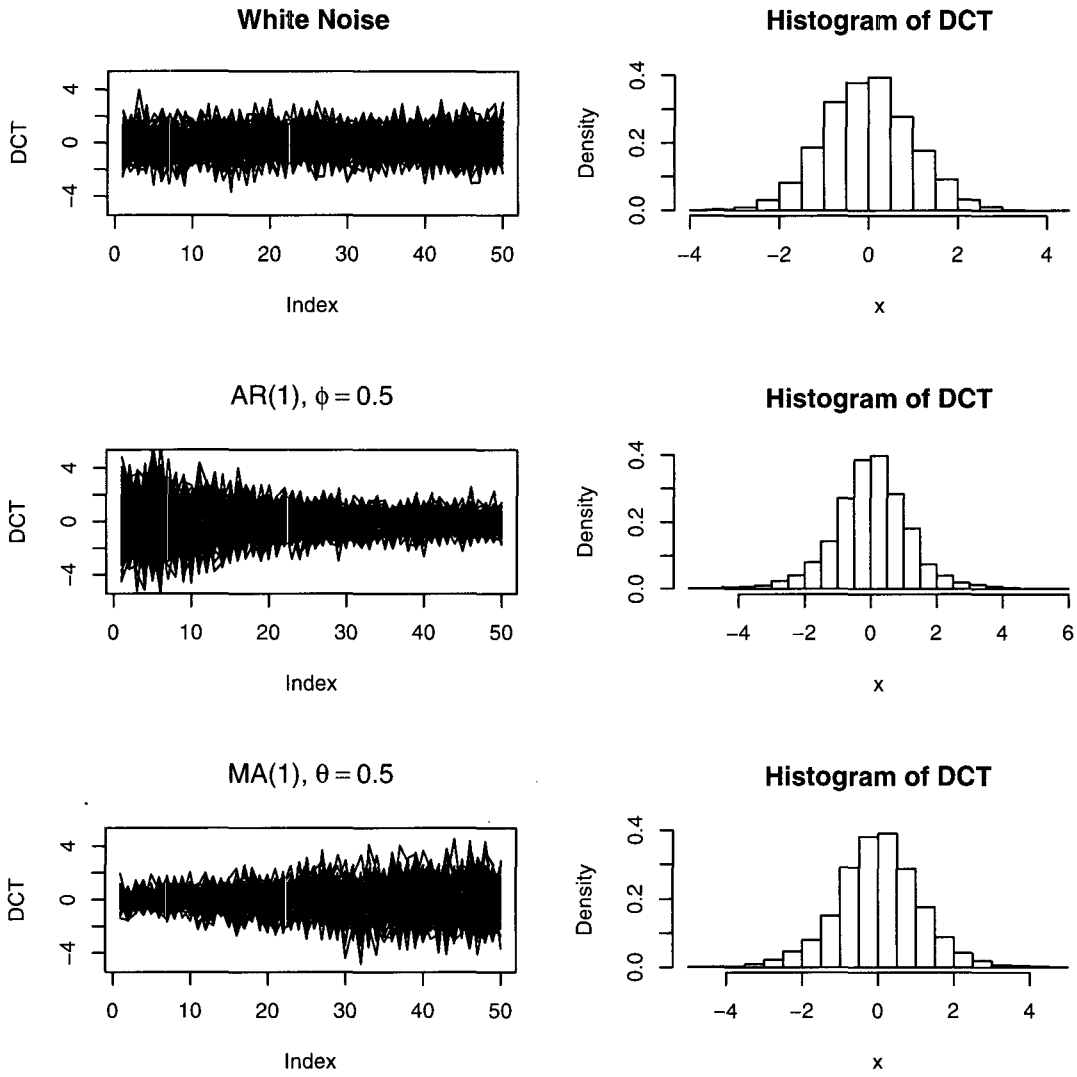


그림 2.2: 백색잡음, AR(1), MA(1)에 대한 DCT 계수와 DCT의 히스토그램

가 된다. 여기서, $\omega_{jt} = \pi(2t-1)(j-1)/(2n)$를 나타낸다.

4 만약 $\{X_t\}_{t=1}^n$가 가우시안시계열(Gaussian time series)이면, $F_j$들은 다변량정규분포를 따른다.

정리 3.1이 의미하는 것은 먼저 시계열의 크기가 어느 정도 커지면 DCT 계수들은 서로 영향을 거의 주지 않으며 DC 성분을 제외한 DCT 계수의 평균은 0으로 모두 같게 된다. 또한 분산은 3에서 언급한 것과 같이 되기 때문에 DCT 계수를 각각의 표준편차로 나눈 수정된 DCT 계수 $F_j/\sigma_j$는 분산이 모두 1이 되고 가우시안시계열인 경우 $F_j/\sigma_j, j = 2, \ldots, n$,는 점근적으로 독립이고 평균이 0이고 분산이 1인 동일한 정규분포를 따른다고 할 수 있다. 이것은 시계열자료를 DCT를 이용하여 주파수영역으로 표시한 경우 DCT 계수를 약간 수정하면 일반적으로 사용되고 있는 붓스트랩 상황과 비슷하게 될 수 있다는 것으로 의미한다. 즉 수정된 DCT 계수를 이용하여 재표집하는 것이 의미가 있다는 것을 나타낸다.

붓스트랩 방법에는 크게 모수적 방법과 비모수적 방법이 있다. 모수적 방법은 자료가 어떤 특정분포를 따른다는 가정 하에서 분포의 특성을 결정하는 모수를 표본을 이용하여 추정하고 추정된 분포로부터 재표본으로 얻는다. 비모수적 방법은 특정분포를 가정하지 않고 있는 표본으로부터 재표본을 복원추출한다. 비모수적인 방법은 표본이 많지 않는 경우에 같은 표본이 여러 번 추출될 수 있다는 단점이 있지만 분포에 대한 가정이 없기 때문에 모든 경우에 사용가능하다. 반면 모수적인 방법은 자료의 분포가 정규분포와 같이 일반적으로 많이 사용하고 있는 분포가 아닌 경우에는 사용하기 어렵다는 단점이 있지만 다양한 재표본을 얻을 수 있다. 정상시계열에서의 DCT 계수는 앞 절에서 보인 것과 같이 정규분포에 가까운 형태를 가지고 있으므로 비모수적 방법과 모수적 방법을 모두 사용할 수 있다.

- 비모수적인 재표집 방법

  단계 1. 표본을 이용하여 $\gamma_i, i = 0, \ldots, n-1$의 추정값 $\hat{\gamma}_i$를 구한다.

  단계 2. 식 (3.1)의 $\gamma_i$에 $\hat{\gamma}_i$를 대입하여 $\sigma_j, j = 2, \ldots, n$,의 추정값 $\hat{\sigma}_j$를 계산한다.

  단계 3. DCT 계수 $F_1, \ldots, F_n$를 계산한다.

  단계 4.수정된 DCT 계수 $\hat{F}_j = F_j/\hat{\sigma}_j$를 계산한다.

  단계 5. 수정된 DCT 계수를 중심 이동시켜 표준환된 DCT계수 $\tilde{F}_j = \hat{F}_j - \bar{\hat{F}}$를 구한다. 여기서 $\bar{\hat{F}} = (n-1)^{-1}\sum_{j=2}^n \hat{F}_j$를 의미한다.

  단계 6. 표준화된 $\left\{\tilde{F}_2, \tilde{F}_3, \ldots, \tilde{F}_n\right\}$ 에서 $(n-1)$개의 재표본 $\left\{\tilde{F}_2^*, \tilde{F}_3^*, \ldots, \tilde{F}_n^*\right\}$을 복원추출하여 얻는다.

  단계 7. 붓스트랩 DCT 계수 $F_j^* = \tilde{F}_j^* \cdot \hat{\sigma}_j$를 계산한다.

  단계 8. $F_1, F_2^*, \ldots, F_n^*$에 IDCT를 적용하여 붓스트랩 시계열 $X_1^*, X_2^*, \ldots, X_n^*$를 구한다.

- 모수적인 재표집방법

단계 1. 비모수적인 방법의 단계1과 단계2를 수행한다.

단계 2. 표준정규분포로부터 랜덤하게 $(n-1)$개의 재표본 $\left\{\tilde{F}_2^*, \tilde{F}_3^*, \ldots, \tilde{F}_n^*\right\}$을 얻는다.

단계 3. 비모수적인 방법의 단계 7과 8를 수행한다.

주의할 것은 DC 성분은 시계열 평균에 절대적으로 영향을 주고 있고 다른 DCT 계수 성분과 통계적 성질이 차이가 있기 때문에 DCT 계수와 관련된 작업에서는 제외시켰다. 모수적 방법는 재표본 $\left\{\tilde{F}_2^*, \tilde{F}_3^*, \ldots, \tilde{F}_n^*\right\}$을 얻는 중간 과정만 제외하고 비모수적 방법와 같기 때문에 여기에서는 비모수적 방법을 중심으로 설명하고자 한다. 단계 1에서는 자기공분산을 시계열자료를 이용하여 추정는데 추정값으로는 일반적으로

$$\hat{\gamma}_j = \frac{\sum_{i=1}^{n-j}(X_i - \bar{X})(X_{i+j} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

가 사용된다. 이렇게 추정된 자기공분산은 DCT 계수의 분산을 추정하는데 사용되는데 이러한 작업을 단계 2에서 수행한다. 단계 3에서는 DCT를 이용하여 시계열 자료를 주파수 영역상으로 변환시켜 DCT 계수를 계산한다. 앞에서 언급한 것과 같이 DCT 계수의 분산이 다르기 때문에 단계 2에서 추정한 표준편차를 이용하여 DCT 계수를 분산이 같도록 수정한다. 비모수적 방법에서 단계 5는 수정한 DCT 계수의 평균을 계산하고 다시 DCT 계수에서 이 값을 빼 평균이 0이 되도록 만든다. 이것은 모형근거 붓스트랩에서 잔차의 합이 0이 되지 않기 때문에 잔차의 합이 0이 되도록 중심화(centered)시키는 것과 일맥상통한다. 단계 6은 표준화된 DCT 계수들로부터 재표본을 복원추출하고 모수적 방법에서는 추정된 정규분포로부터 확률표본을 추출한다. 단계 7에서는 단계 6에서 추출된 재표본에 단계 2에서 추정한 표준편차를 곱한다. 이렇게 하면 본래 자료에서 변환된 DCT 계수와 통계적으로 비슷한 성질의 가지는 DCT 계수를 얻을 수 있다. 단계 8에서는 단계 7에서 구한 DCT 계수와 DC 성분을 IDCT를 이용하여 변환시켜 시계열자료를 구한다.

## 4. 모의실험

제안된 방법의 성능을 알아보기 위해서 모의실험에서는 시계열자료에 대한 재표집 방법으로 많이 사용되고 있는 BB와 SB를 제안된 방법과 비교하였다. 시계열자료의 크기로는 $n = 64$와 125를 고려하였다. SB에서는 차수가 10인 AR모형을 사용하였으며, BB에서는 블록의 길이가 $l = n^{1/3}$로 고정한 비겹침블록(nonoverlapping block)방법이 사용되었다. 자료는 다음과 같은 모형에서 발생시켰다.

- AR(1): $X_t = 0.8X_{t-1} + \varepsilon_t$.
- MA(1): $X_t = \varepsilon_t - 0.8\varepsilon_{t-1}$.
- ARMA(1,1): $X_t = 0.3X_{t-1} + \varepsilon_t + 0.2\varepsilon_{t-1}$.
- AR(2): $X_t = 0.7X_{t-1} - 0.2X_{t-2} + \varepsilon_t$.

- MA(2): $X_t = \varepsilon_t - 0.7\varepsilon_{t-1} + 0.2\varepsilon_{t-2}$.
- Threshold AR(1): $X_t = (0.9X_{t-1} + \varepsilon_t)\, I_{(X_{t-1} \le -2.5)} + (0.8X_{t-1} + \varepsilon_t)\, I_{(X_{t-1} > -2.5)}$.

이들 모형은 Nordgaard(1992), Buhlmann and Kunsch(1999), Pararoditis and Politis(1999), 그리고 Lahiri(2003)에 의해 예제로 언급되었던 모형이다.

성능비교를 위한 통계량으로 이 논문에서는 Nordgaard(1992)에 의해 언급되었던 1차 표본자기상관 $\hat{\rho}_1 = \sum_{t=1}^{n-1}(X_t - \bar{X})(X_{t+1} - \bar{X}) / \sum_{t=1}^{n}(X_t - \bar{X})^2$의 분포의 5-, 20-, 50-, 80-, 그리고 95-백분위수를 고려하였다. 표본크기$n = 64$와 125에 대해 이들 백분위수의 참값을 모르기 때문에 10000번의 모의실험을 통해 이들 참값에 대한 근사값을 구하였다. 아래의 표들은 500번 붓스트랩 반복을 200번 모의실험하여 얻은 결과들이다. 표에서 NDCT와 PDCT는 각각 이 논문에서 제안한 비모수적 방법과 모수적 방법을 의미하고 TAR은 Threshold AR을 의미한다.

표본크기 $n = 64$인 경우 모의실험 결과를 다음과 같이 정리할 수 있다.

- AR(1), AR(2), 그리고 ARMA(1,1)에서는 분포의 왼쪽은 NDCT이 우수한 것으로 나타났으며 오른쪽은 PDCT가 우수한 것으로 나타났다. 실제 모형과 추정모형의 차수 차이가 컸던 SB의 경우 과대적합에 의해 성능이 오히려 떨어지는 것으로 나타났다.
- MA(1)와 MA(2)에서는 SB의 성능이 우수한 것으로 나타났는데 이것은 MA 모형이 AR($\infty$)로 표시되기 때문에 차수가 큰 AR 모형이 근사모형으로 적합하기 때문인 것으로 생각된다. 또한 95백분위수를 제외한 NDCT의 결과가 SB와 큰 차이가 없는 것으로 나타났다.
- TAR의 경우, 왼쪽의 분포에 대해서는 NDCT가 우수한 것으로 나타났으며 다른 방법들은 실제 값과 차이가 있는 것으로 나타났다.
- 전반적으로 BB의 성능이 떨어지는 것으로 나타났다.

표본크기 $n = 128$의 경우 TAR모형에서 PDCT에서 우수한 결과를 얻었다는 것을 제외하고 모든 모형에서 NDCT가 우수한 성능을 가지고 있는 것으로 나타났다.

표 4.1: 표본크기가 $n = 64$일 때, 각 백분위수의 근사참값과 붓스트랩 추정값

| 모형 | 방법 | 5% Mean(SD) | 20% Mean(SD) | 50% Mean(SD) | 80% Mean(SD) | 95% Mean(SD) |
|---|---|---|---|---|---|---|
| AR(1) | Exact | .567 | .657 | .742 | .806 | .854 |
| | BB | .362(.105) | .448(.094) | .527(.085) | .598(.077) | .658(.071) |
| | SB | .526(.124) | .613(.111) | .691(.098) | .758(.084) | .808(.072) |
| | NDCT | .575(.116) | .658(.107) | .732(.097) | .789(.086) | .828(.075) |
| | PDCT | .519(.125) | .630(.112) | .724(.097) | .797(.081) | .847(.067) |
| MA(1) | Exact | -.613 | -.553 | -.480 | -.404 | -.320 |
| | BB | -.515(.076) | -.445(.078) | -.366(.080) | -.283(.081) | -.201(.084) |
| | SB | -.604(.085) | -.542(.088) | -.473(.092) | -.401(.097) | -.329(.102) |
| | NDCT | -.605(.088) | -.548(.089) | -.483(.089) | -.415(.092) | -.347(.094) |
| | PDCT | -.638(.083) | -.565(.086) | -.480(.089) | -.391(.092) | -.300(.096) |
| ARMA(1,1) | Exact | .239 | .333 | .424 | .507 | .580 |
| | BB | .097(.108) | .192(.104) | .285(.100) | .371(.097) | .445(.094) |
| | SB | .205(.128) | .292(.120) | .378(.113) | .459(.106) | .530(.101) |
| | NDCT | .237(.117) | .316(.115) | .395(.113) | .468(.110) | .531(.107) |
| | PDCT | .176(.120) | .284(.117) | .391(.112) | .489(.107) | .574(.101) |
| AR(2) | Exact | .394 | .477 | .554 | .622 | .678 |
| | BB | .209(.093) | .300(.088) | .388(.083) | .466(.079) | .534(.075) |
| | SB | .356(.107) | .436(.100) | .515(.092) | .585(.086) | .645(.080) |
| | NDCT | .393(.102) | .467(.098) | .537(.093) | .600(.089) | .651(.086) |
| | PDCT | .335(.106) | .436(.099) | .531(.093) | .615(.087) | .684(.080) |
| MA(2) | Exact | -.547 | -.476 | -.398 | -.314 | -.228 |
| | BB | -.469(.089) | -.397(.091) | -.315(.091) | -.229(.092) | -.142(.094) |
| | SB | -.532(.099) | -.462(.102) | -.385(.105) | -.303(.108) | -.224(.111) |
| | NDCT | -.527(.099) | -.466(.100) | -.398(.101) | -.325(.101) | -.251(.100) |
| | PDCT | -.570(.093) | -.487(.098) | -.393(.100) | -.293(.102) | -.193(.101) |
| TAR | Exact | .578 | .680 | .768 | .837 | .887 |
| | BB | .384(.119) | .468(.103) | .545(.089) | .615(.080) | .675(.073) |
| | SB | .533(.136) | .622(.124) | .703(.110) | .768(.097) | .817(.086) |
| | NDCT | .586(.128) | .674(.122) | .746(.113) | .801(.100) | .838(.088) |
| | PDCT | .531(.139) | .643(.129) | .738(.113) | .807(.095) | .856(.078) |

표 4.2: 본크기가 $n = 125$일 때, 각 백분위수의 근사참값과 붓스트랩 추정값

| 모형 | 방법 | 5% Mean(SD) | 20% Mean(SD) | 50% Mean(SD) | 80% Mean(SD) | 95% Mean(SD) |
|---|---|---|---|---|---|---|
| AR(1) | Exact | .655 | .717 | .772 | .816 | .850 |
| | BB | .488(.065) | .548(.058) | .603(.052) | .652(.048) | .694(.044) |
| | SB | .623(.071) | .683(.065) | .738(.058) | .785(.052) | .823(.047) |
| | NDCT | .672(.069) | .729(.065) | .782(.062) | .827(.058) | .860(.053) |
| | PDCT | .635(.075) | .710(.069) | .777(.062) | .833(.055) | .873(.047) |
| MA(1) | Exact | -.582 | -.537 | -.484 | -.427 | -.373 |
| | BB | -.498(.061) | -.448(.063) | -.392(.065) | -.333(.068) | -.274(.071) |
| | SB | -.577(.061) | -.529(.063) | -.476(.065) | -.421(.068) | -.366(.070) |
| | NDCT | -.584(.063) | -.538(.063) | -.487(.063) | -.435(.065) | -.384(.066) |
| | PDCT | -.612(.060) | -.552(.061) | -.485(.063) | -.416(.065) | -.347(.068) |
| ARMA(1,1) | Exact | .307 | .375 | .440 | .499 | .556 |
| | BB | .206(.068) | .273(.065) | .340(.064) | .401(.062) | .456(.061) |
| | SB | .290(.082) | .355(.077) | .419(.073) | .480(.069) | .534(.066) |
| | NDCT | .322(.078) | .379(.076) | .436(.074) | .491(.073) | .541(.073) |
| | PDCT | .276(.080) | .354(.077) | .434(.074) | .509(.072) | .576(.068) |
| AR(2) | Exact | .461 | .514 | .568 | .617 | .659 |
| | BB | .325(.069) | .387(.067) | .448(.065) | .504(.063) | .552(.063) |
| | SB | .448(.074) | .502(.069) | .555(.064) | .604(.061) | .647(.058) |
| | NDCT | .476(.072) | .525(.069) | .574(.067) | .620(.064) | .660(.063) |
| | PDCT | .438(.076) | .505(.071) | .571(.067) | .632(.063) | .685(.059) |
| MA(2) | Exact | -.508 | -.457 | -.399 | -.338 | -.280 |
| | BB | -.436(.059) | -.381(.061) | -.320(.063) | -.257(.065) | -.194(.068) |
| | SB | -.501(.063) | -.447(.065) | -.387(.067) | -.326(.069) | -.266(.071) |
| | NDCT | -.502(.070) | -.453(.069) | -.399(.069) | -.343(.069) | -.289(.071) |
| | PDCT | -.537(.067) | -.470(.067) | -.396(.068) | -.320(.070) | -.245(.073) |
| TAR | Exact | .674 | .743 | .804 | .853 | .892 |
| | BB | .510(.076) | .568(.068) | .623(.060) | .672(.054) | .713(.050) |
| | SB | .658(.080) | .715(.074) | .767(.067) | .810(.061) | .844(.055) |
| | NDCT | .708(.080) | .763(.078) | .813(.074) | .853(.066) | .882(.058) |
| | PDCT | .671(.086) | .745(.081) | .808(.074) | .857(.064) | .892(.053) |

## 5. 결론

붓스트랩은 지난 20여 년간 시계열과 공간통계 분야를 제외한 거의 모든 통계학 분야에서 많이 연구되고 활용된 강력한 통계적 방법이다. 비록 시계열 자료에 대한 여러 가지 재표집 기법들이 제안되었지만 사용하는데 문제점이 있는 것으로 알려져 있다. 이 논문에서 제안하는 DCT에 의한 주파수 영역상의 재표집 기법은 기존의 방법에서 발생했던 문제를 많이 해결할 수 있을 것으로 생각된다. 모의실험에서 본 것처럼 BB의 경우 성능이 좋지 않은 것으로 나타났으며 SB의 경우 실제모형에 영향을 많이 받는 것으로 나타났다. 제안된 방법의 경우 특정 모형에 영향를 상대적으로 덜 받고 성능 또한 나쁘지 않는 것으로 나타났다. 문제는 비정상 시계열자료에 대해 사용할 수 있는가 인데 많은 실험결과 비정상성의 특성이 이상점 형태의 DCT 계수로 표시되는 것을 확인하였다. 이 성질을 이용하면, 비정상 시계열자료에 대한 재표집방법을 얻을 것으로 생각되며 이 문제는 다음 연구과제로 남겨두고자 한다.

## 참고문헌


Buhlmann, P. (1997). Sieve bootstrap for time series, *Bernoulli* **3**, 123-148.

Buhlmann, P. (1998). Sieve bootstrap for smoothing in non-stationary time series. *The Annals of Statistics* **26**, 48-83.

Buhlmann, P. and Kunsch, H. R. (1999). Block length selection in the bootstrap for time series, *Computational Statistics and Data Analysis* **31**, 295-310.

Choi, E. and Hall, P. (2000). Bootstrap confidence regions computed from autoregressions of arbitrary order, *Journal of the Royal Statistical Society, Series B* **62**, 461-477.

Dahlhaus, R. and Janas, D. (1996). A frequency domain bootstrap for ratio statistics in time series analysis, *The Annals of Statistics* **24**, 1934-1963.

Efron, B (1979). Bootstrap methods: another look at the jackknife (with discussion), *The Annals of Statistics* **7**, 1-26.

Franke, J. and Hardle, W. (1992). On bootstrapping kernel spectral estimates, *The Annals of Statistics* **20**, 121-145.

Hall, P. (1985). Resampling a coverage pattern, *Stochastic Processes and Their Applications* **20**, 231-246.

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

Lahiri, S. N. (2003). *Resampling methods for dependent data*, Springer, New York.

Nordgaard, A. (1992). Resampling a stochastic process using a bootstrap approach, *Bootstrapping and Related Techniques*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, p.376.

Paparoditis, E. and Politis, D. N. (1999). The local bootstrap for periodogram statistics, *Journal of Time Series Analysis* **20**, 193-222.

Singh, K. (1981). On the asymptotic accuracy of the Efron's boostrap, *The Annals of Statsitics* **9**. 1187-1195.

[ 2005년 9월 접수, 2005년 10월 채택 ]

# Resampling Methods on Frequency Domains for Time Series*

In-Kwon Yeo[1] Wha-hyung Yoon[2] Sinsup Cho[3]

ABSTRACT

This paper presents the resampling method for time series data in the frequency domain obtained by using discrete cosine transforms(DCT). The advantage of the proposed method is to generate bootstrap samples in time domain comparing with existing bootstrapping method. When time series are stationary, statistical properties of DCT coefficients are investigated and provide the verification of the proposed procedure.

*Keywords:* Bootstrap, Discrete cosine transform, Asymptotically independent

* This work was in part supported by the Korea Research Foundations, Korea, under grant KRF-2003-002-C00043.
This work was supported by the Korea Research Foundation Grant funded by Korea. Government(MOEHRD, Basic Research Promotion Fund)(KRF-2005-070-C00022).

1) Assistant Processor, Division of Mathematics and Statistics, Sookmyung Women's University. Chungpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Korea.
E-mail: inkwon@sookmyung.ac.kr

2) Graduate student, Department of Statistics, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-742, Korea.
E-mail: whyya@chollian.net

3) Professor, Department of Statistics, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-742, Korea.
E-mail: sinsup@snu.ac.kr