시계열분석을 위한 주파수 공간상에서의 재표집 기법 *

여인권[1] 윤화형[2]

요 약

이 논문에서는 시계열자료을 이산코사인변환을 이용하여 주파수 공간으로 변환시킨 후 이산코사인변환 계수를 이용하여 재표본을 추출하는 방법에 대해 알아본다.

주요용어: 붓스트랩, 이산코사인변환, 이상점.

1. 서론

Efron(1979)에 의해 이론적 토대가 마련된 이후, 붓스트랩(bootstrap)은 1980년대와 90년대 통계학에서 가장 활발하게 연구되었던 분야 중 하나로써 통계학 전분야에서 다양하게 개발되고 활용되는 방법이다. 독립적인 표본들에 대해 붓스트랩은 통계적으로 비슷한 성질을 가지는 재표본(resample)을 표본들로부터 무작위로 추출하여 통계 추론에 필요한 통계량의 적률이나 분포를 근사하는데 사용된다. Singh(1981)이 지적한 것처럼 문제는 시계열자료나 공간통계(spatial statistics)자료와 같이 자료들 간에 독립성이 만족되지 않는 상황에서 무작위로 추출하는 일반적인 붓스트랩 방법을 사용하면 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 방법으로써 정상 시계열 자료를 여러 개의 블록으로 분할한 다음 일반적인 붓스트랩 방법과 같이 블록을 표본처럼 무작위로 선택하고 선택한 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 Buhlmann과 Kunsch (1999) 등에 의해 진행되어 왔지만 이 방법이 가지고 있는 근본적인 문제를 해결하지 못하고 있다.

* 이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2003-002-C00043)

1) (140-742) 서울특별시 용산구 청파동 2가 숙명여자대학교 수학통계학부 조교수
E-mail: inkwon@moak.chonbuk.ac.kr
2) (151-742)서울시 관악구 신림동 산56-1 서울대학교 통계학과 박사과정수료
E-mail: whyya@chollian.net

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. 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 비정상시계열자료나 이동평균모형과 같이 고차의 AR(p)로도 적합이 잘 되지 않는 경우에는 사용하기 어렵다는 것이다. 또한 분산추정이 정확하지 않기 때문에 보다 정확한 분산추정을 위해서는 Choi와 Hall(2000)이 언급한 이중붓스트래핑(double boostrapping) 같은 복잡한 알고리즘을 적용해야 하는 단점이 있다.

기존 연구의 대부분은 시간영역에서 재표본을 직접 추출하는 방법을 중심으로 진행되어 왔지만 제안하고자 하는 방법은 자료를 주파수 영역으로 변환시킨 자료를 이용하여 재표집을 구한다. 일반적으로 연구되고 있는 주파수 영역에서의 재표집 방법에서는 복잡한 종속성이 존재하는 시계열 관측값들이 거의 독립적인 통계량인 주기도 좌표(periogogram ordinates)로 변환될 수 있다고 가정하며 변환된 값들에 대해 독립인 표본에 적용했던 붓스트랩 방법을 사용하고 있다. Franke와 Hardle(1992)는 주기도(periodogram)과 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 Dahlhaus와 Janas(1996), Paparoditis와 Politis(1999) 등을 포함한 많은 연구자들에 의해 수정, 보완, 확장되어 연구되고 있다.

종속성이 존재하는 자료에 대한 재표집 방법의 점근적 이론들은 Lahiri(2003)을 참조하기 바란다. 이 논문에서는 위에서 언급한 형태와 전혀 다른 주파수 영역 재표집 방법을 제안하고자 한다.

2. 이산주파수변환

시계열 자료 분석은 크게 시간영역(time domain) 분석방법과 주파수영역(frequency domain) 분석방법으로 나눌 수 있다. 주파수 영역에서의 시계열 분석은 시간 영역에서 파악하기 어려운 자료의 특징을 유도할 수 있다는 장점이 있지만 대부분 복소수 형태의 값으로 표시되는 푸리에변환(Fourier transform)을 기초로 이루어지기 때문에 일반인이 사용하는데 있어 어렵고 분석 결과에 대한 해석이 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. 이 논문에서는 사용이 용이하면서 이해하기 쉽고 기능면에서 우수한 DCT를 중심으로 주파수공간 상에서의 재표집 방법에 대해 설명하고자 한다. DCT는 코사인함수 대신 지수함수를 사용하는 이산푸리에변환(discrete Fourier transform)과 밀접한 관계가 있지만 적은 수의 계수로 시계열자료에서의 주요 에너지를 나타낼 수 있는 측면에서 DFT

보다 좋은 에너지압축(energy compaction) 성질을 가지고 있다.

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

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

여기서 w_j는 $j = 1$일 때 $1/\sqrt{n}$이고 그 외의 값에서는 $\sqrt{2}/\sqrt{n}$의 값을 가진다. 첫 번째 계수인 F_1을 특별히 DC 성분(DC component)이라고 하는데 가중치가 모두 1이 되기 때문에 자료의 평균에 $\sqrt{n}$을 곱한 값이 된다. F_2의 경우 자료의 감소 또는 증가 추세에 대한 전반적인 특징을, F_3은 감소하다가 증가하는 또는 증가하다가 감소하는 형태의 특징 등을 나타내는 것을 볼 수 있다. 일반적으로 j가 작을수록 코사인함수의 주기가 길어지며 해당 계수는 시계열자료에 있어 장기간의 특징을 나타내는 저주파 대역의 특성을 나타내고 반대로 j가 커질수록 주기가 짧아져 단기간의 특징을 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.

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

오디오나 영상과 같이 근접자료들 간에 상당히 높은 양의 상관관계를 가지기 경우 일반적으로 저주파수 대역에서 큰 값을 가지며 고주파수 대역에는 주로 백색잡음(white noise)과 같이 큰 의미가 없는 신호들이 표시된다. 이런 경우 저주파 대역에 대한 해석만으로도 시계열자료에 대한 특성을 충분히 설명할 수 있다. 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 2.1은 주요 시계열과정에서 시계열자료 50개를 100번 생성시켰을 때 각 주파수 대역에서의 DCT 계수 절대값의 평균(왼쪽)과 전체 DCT 계수의 히스토램(오른쪽)을 표시한 것이다.

이들 그림에서 볼 수 있는 것과 같이 각각의 시계열과정에 따라 DCT 계수 분포의 특징이 다르다는 것을 알 수 있다. 평균이 0이고 분산이 1인 정규분포를 따르는 백색잡음을 ε_t라고 할 때 그림 2.1의 (a)와 (b)는 $X_t = \varepsilon_t$에 대한 결과로 전체 주파수 대역에서 골고루 에너지를 가지고 있는 것을 볼 수 있으며 전체 DCT 계수의 분포도 큰 이상점이 없는 정규분포의 형태를 가지고 있다. 그림 2.1의 (c)와 (d)는 모형식이 $X_t = 0.5X_{t-1} + \varepsilon_t$인 AR(1) 과정에 대한 것으로 저주파수 대역의 절대값들이 고주파수 대역의 값들에 비해 상대적으로 큰 것을 볼 수 있다. 또한 (b)와 비교하여 양쪽 꼬리부분이 길게 뻗어 있는 것을 볼 수 있는데 이것은 저주파수 대역의 값이 시계열의 특성을 나타낸다고 볼 수 있다. 그림 2.1의 (e)와 (f)는 $X_t = \varepsilon_t - 0.5\varepsilon_{t-1}$인 MA(1) 과정의 DCT 계수 분포를 나타내는데 (c)와 반대로 고주

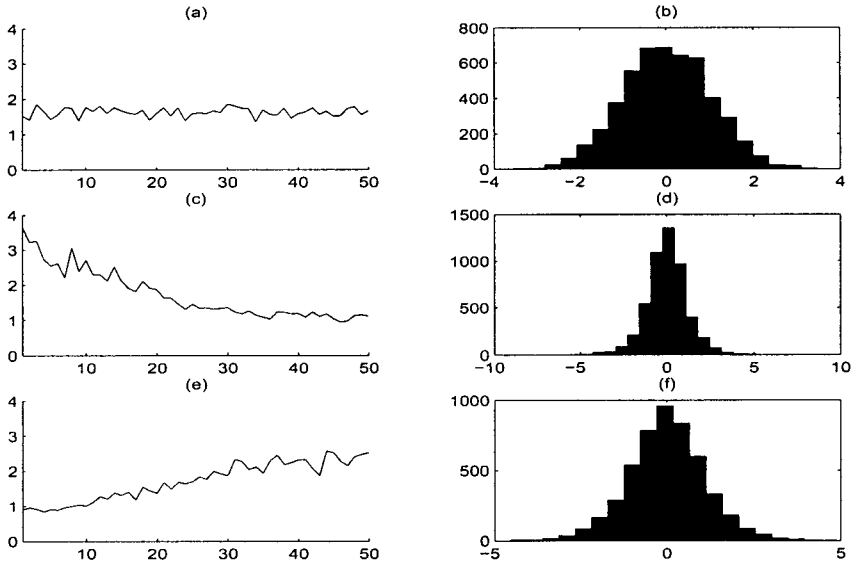

그림 2.1: 정상 시계열 이산코사인변환 계수의 분포

파수 대역의 절대값이 저주파수 대역의 값보다 큰 것을 볼 수 있으며 이들 고주파수 대역의 값이 (f)에서 양쪽 꼬리부분으로 분포해 있는 것을 확인할 수 있다. 그러므로 DCT를 이용한 주파수 공간상에서의 재표집방법에서는 압축에서와 같이 저주파만을 고려하는 것이 아니라 전체 대역을 종합적으로 고려해야 할 필요가 있다.

3. 주파수 공간상에서의 재표집

앞 절에서 본 것과 같이 DCT 계수를 이용한 주파수 영역에서의 분석에서는 주파수 계수의 절대값이 상대적으로 큰 주파수 대역이 시계열 자료의 특성을 나타낸다고 본다. 이 논문에서 제안하고자 하는 재표집 방법에서는 절대값이 큰 특징 주파수 대역의 값은 그대로 두고 작은 값을 가지는 표집 주파수 대역의 값들을 이용하여 재표집하는 것이다. 이러한 작업을 하기 위해서는 먼저 특징 주파수 대역과 표집 주파수 대역을 어떻게 분류할 것인지에 대한 연구와 작은 값으로 분류된 대역에서 어떻게 재표집할 것인지에 대한 연구가 필요하다.

3.1. 주파수 대역의 분류

주파수 대역을 분리하는 방법에 대해서는 여러가지로 생각해 볼 수 있다. 그림 2.1의 (b)에서 보는 것과 같이 백색잡음과정의 경우 DCT 계수 분포는 다른 과정들의 DCT와 비해 정규분포에 더 근사하고 있다. 이와 반해 (d)와 (f)의 경우 첨도가 상당히 큰 것으로 나

타난는데 이것은 분포의 꼬리부분이 두터운 것을 의미한다. 또한 특징이 있는 과정의 경우 그 특징과 관련이 있는 주파수 대역의 값이 상대적으로 커서 해당 주파수 값이 마치 이상점과 같은 역활을 하는 것으로 볼 수 있다. 이러한 점을 고려하면 어떤 주파수 대역의 값이 이상점인지 아닌지를 판정함으로써 특징 주파수 대역과 표집 주파수 대역을 분류할 수도 있다.

이상점을 판정하기 위한 많은 기준들이 Barnett과 Lewis (1993)의 책에 소개되어 있는데 이들 기준들은 대부분 자료를 표준화하기 위해 표준편차를 사용하고 있어 기준 자체가 이상점에 영향을 받고 있다. 이 논문에서는 이상점에 로버스트한 다음과 같은 통계량을 이용하여 이상점을 정의하고자 한다.

$$M_i = \left| \frac{2F_i}{F_{84\%} - F_{16\%}} \right|.$$

여기서 $F_{a\%}$는 $a\%$ 백분위수를 표시한 것으로 S_F를 $F_1, \ldots, F_n$들의 표본표준편차라고 할 때 정규분포의 경우 $F_{84\%} - F_{16\%} \simeq 2S_F$가 되기 때문에 $M_i \simeq |F_i/S_F|$가 된다. 그러므로 M_i가 일반적으로 이상점을 판정할 때 사용되는 3 또는 4 보다 큰 값인 경우 이상점으로 판정한다.

위와 같은 분석을 하지 않고 단순히 DCT 계수의 절대값이 큰 상위 몇 %를 특징 주파수 대역에 해당하는 것으로 판정하고 나머지 부분을 표집주파수로 정하는 방법도 생각해 볼 수 있다. 다른 방법으로는 DCT 계수를 두 그룹으로 군집분석을 하는 것을 생각할 수 있다.

3.2. DCT 계수의 재표집

일반적인 재표집 방법과 마찬가지로 DCT 계수에 대한 재표집 방법은 크게 모수적 방법과 비모수적 방법으로 나눌 수 있다. 모수적 방법에서 DCT 계수들의 분포를 가정하고 작은 값으로 분류된 DCT 계수를 이용하여 이 분포의 모수를 추정한 후 추정된 분포에서 난수를 발생시켜 재표본을 추출한다. 참고로 그림 에서 보는 것과 같이 백색잡음과정의 DCT 계수의 분포가 정규분포와 큰 차이가 없는 것으로 볼 수 있다. 비모수적 방법에서는 작은 값으로 분류된 DCT 계수를 랜덤하게 복원추출하여 재표본을 구한다.

제안된 방법에서는 특징주파수와 표집주파수를 나누는 과정에서 발생할 수 있는 오분류의, 특히 특징주파수에 속하는데 표집주파수로 분류될, 가능성을 고려하고 재표집의 효율을 높이기 위해 다음과 같은 단계를 추가한다. 변수 $\boldsymbol{S} = (S_1, S_2, \ldots, S_t)$는 $\boldsymbol{F}$에서 표집주파수 대역의 속하는 t개의 DCT 계수를 차례로 뽑아 표시한 것이고 L_i는 S_i의 주파수 대역의 위치를 나타낸다고 하자. 변수 R_i는 $\boldsymbol{S}$를 크기순서대로 나열 했을 때 $\boldsymbol{S}$ 중 S_i의 순위라고 하자. 모수적 또는 비모수적 방법을 이용하여 $\boldsymbol{S}$로 부터 재표본 $S_1^*, S_2^*, \ldots, S_t^*$를 추출한다. 주파수 대역 L_i 위치에 순서통계량 $S^*_{(R_i)}$를 대입하여 새로운 주파수 계수 $\boldsymbol{F}^* = (F_1^*, F_2^*, \ldots, F_n^*)$를 구한다. 이렇게 구해진 $\boldsymbol{F}^*$를 IDCT를 이용하여 시간영역 상에서의 재표본을 구한다.

예제 3.1: 시계열자료 8개를 이용하여 DCT 계수를 구한 결과가

$$\boldsymbol{F} = (2.3, -1.4, 0.6, 1.1, -0.2, 0.4, -3.1, 0.9)$$

라고 하고 특징주파수와 표집주파수를 나누는 기준은 절대값이 2이라고 하자. 그러면 6개의 주파수값이 표집주파수에 속하고 $\boldsymbol{S} = (-1.4, 0.6, 1.1, -0.2, 0.4, 0.9)$, $\boldsymbol{L} = (2, 3, 4, 5, 6, 8)$, $\boldsymbol{R} = (1, 4, 6, 2, 3, 5)$가 된다. 비모수적 방법으로 재표집한 결과가 (-1.4, -0.2, -0.2,0.9,0.9,1.1)이라고 하면 $F^*_{L_1} = F^*_2 = S_{(R_1)} = S^*_{(1)} = -1.4$를 대입한다. 결론적으로 재표집 주파수는

$$\boldsymbol{F}^* = (2.3, -1.4, 0.9, 1.1, -0.2, -0.2, 0.9)$$

가 되고 이 값들을 IDCT를 이용하여 변환시키면 시간 공간상에서의 재표본을 얻을 수 있다.

참고문헌

Barnett, V. and Lewis, T. (1993), Outliers in Statistical Data, 3 rd Ed., John Wiley & Sons, New York.

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

Buhlmann, P. (1998), Sieve bootstrap for smoothing in non-stationary time series. Ann. Statist., 26, 48-83.

Buhlmann, P. and Kunsch, .H. R. (1999), Block length selection in the bootstrap for time series, Comput. Statist. Data Anal., 31, 295-310.

Choi, E. and Hall, P. (2000), Bootstrap confidence regions computed from autoregressions of arbitrary order, J. Roy. Statist. Soc. (Ser. B), 62, 461-477.

Dahlhaus, R. and Janas, D. (1996), A frequency domain bootstrap for ratio statistics in time series analysis, Ann. Stat. 24, 1934-1963.

Diaconis, P. and Ylvisaker, D (1985), Quantifying prior opinion. In: Bayesian Statistics 2, Eds. J.M. Bernardo et al., North-Holland, Amsterdam, 133-156.

Efron, B (1979), Bootstrap methods: another look at the jackknife (with discussion), Ann. Statist., 7, 1-26.

Franke, J. and Hardle, W. (1992), On bootstrapping kernel spectral estimates, Ann. Stat., 20, 121-145.

Hall, P. (1985), Resampling a coverage pattern, Stoch. Process. Appl., 20, 231-246.

John, J. A. and Draper, N. R. (1980), An alternative family of transformations, Appl. Statist. 29, 190-197.

Kreiss, J. P. and Paparoditis, E. (2003), Autoregressive-aided periodogram bootstrap for timeseries, Ann. Statis. 31, 1923-1955.

Kunsch, H. R. (1989), The jackknife and the bootstrap for general stationary observations, Ann. Statist., 17, 1217-1241.

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

Paparoditis, E. and Politis, D. N. (1999), The local bootstrap for periodogram statistics, Journal of time series analysis 20, 193-222.

Yeo, I. K. and Johnson, R. A. (2000), A new family of power transformations to improve normality or symmetry, Biometrika, 87, 954-959.