응용통계연구 (2009)
22(6), 1203–1214

시계열자료 군집방법의 비교연구

홍한움[1] · 박민정[2] · 조신섭[3]

[1]서울대학교 통계학과, [2]포항수학연구소, [3]서울대학교 통계학과

(2009년 11월 접수, 2009년 11월 채택)

요 약

본 논문에서는 시계열자료의 군집분석을 위해 시간영역과 진동수영역에서의 군집 방법들을 소개하고 각 방법들의 장단점에 대해 논의하였다. KOSPI 200에 속한 15개 기업의 일별 주가자료를 이용한 비교분석 결과 비모수적인 방법인 웨이블릿을 이용한 군집분석이 가장 좋은 결과를 보였다. 비정상 시계열자료의 경우 차분 보다는 EMD를 이용하여 추세를 제거하는 방법이 스펙트럴 밀도함수를 이용한 군집분석에 더 효율적이었다.

주요용어: 시계열, 군집, 카오틱 사상, 스펙트럴분석, 웨이블릿, 경험적 모드분석.

1. 서론

군집분석(clustering analysis)이란 관측치들을 유사성(similarity)이나 거리(distance) 등과 같은 측도(measure)를 이용하여 유사한 관측치들을 몇 개의 집단(군집)으로 집단화 하는 기법을 말한다. 분류(classification)와는 다른 개념으로, 분류는 원래 속해 있는 집단에 대한 정보가 없는 새 관측치가 어느 집단에 속할 것인지를 판단하는 기법임에 비해, 군집분석은 집단 구조(group structure)에 대한 가정 없이 오직 정의된 거리를 이용하여 관측치들을 군집으로 묶는 좀 더 원초적인 기법이다. 따라서 군집분석에서 가장 중요한 것은 거리를 어떻게 정의하느냐 하는 것과 정의된 거리를 바탕으로 어떻게 군집화 할 것인가 하는 것이다. 군집화 하는 방법은 크게 계층적(hierarchical)인 방법과 비계층적(nonhierarchical)인 방법이 있는데, 군집의 수를 사전에 결정해 놓았을 때 비계층적 방법을 많이 쓰고 그렇지 않을 때 계층적 방법을 많이 쓴다.

시계열 자료의 군집분석은 시계열 자료들을 몇 개의 군집으로 집단화하는 방법이다. 시계열 자료들이 T기간 동안 관측되었다면, 개별적인 시계열 자료는 T차원에서의 관측치로 생각할 수 있다. 가장 단순히 생각할 수 있는 거리는 T차원의 두 시계열 자료의 유클리드 거리(Euclidean distance)를 구하는 것이다. 그러나 시계열 자료의 경우 T의 값이 매우 크기 때문에 이렇게 정의된 거리는 지나치게 큰 값을 가지게 되며, 유사한 시계열 자료의 경우도 한 시점에서의 값의 차이가 크면 유클리드 거리가 커져 서로 다른 군집으로 묶일 가능성이 높다. 따라서 시계열 자료의 군집분석에서 가장 중요한 것은 거리를 어떻게 효과적으로 정의할 것인가이다. 거리만 적절히 정의할 수 있다면 기존에 알려진 군집화 방법들을 이용하여 군집분석을 시행할 수 있다. 대부분의 경우 군집의 수를 사전에 알 수 없기 때문에, 시계열 자료의 군집은 계층적 방법을 이용한다.

본 연구의 박민정은 Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(Grant 2009-0094070)에 의해 지원받았음.
[3]교신저자: (151-747) 서울시 관악구 관악로 599, 서울대학교 통계학과, 교수. E-mail: sinsup@snu.ac.kr

시계열 자료의 거리를 정의하는 방법으로는 두 가지 방법을 사용한다. 첫 번째 방법은 시계열 자료를 잘 특징지을 수 있는 특성치 벡터(feature vector)를 정의해서 특성치 벡터들 사이의 거리를 이용하는 방법이다. 원 시계열 자료의 차원이 크기 때문에 특성치 벡터를 이용하는 것이므로 특성치 벡터의 차원은 작을수록 좋다. 두 번째 방법은 시계열 자료의 스펙트럴 밀도함수를 구한 후, 스펙트럴 밀도함수들 사이의 거리를 정의하는 것이다. 원 시계열 자료는 시간영역(time domain)의 자료이지만, 자료의 특성에 따라 주파수 영역(frequency domain)에서 자료를 분석하는 것이 더 효과적일 수 있다. 2절에서는 각각의 영역에서 시계열 자료를 군집하는 방법들을 소개하고 각 방법들의 장단점에 대해 알아본다. 3절에서는 KOSPI 200에 속한 기업들 중 15개 기업의 일별 주가자료를 2절에서 소개한 방법들에 의해 군집분석을 시행하고 각 방법들의 특징 및 효용성을 비교해 보았다.

2. 시계열자료 군집방법(clustering method)

2.1. 시간영역(time domain)에서의 군집

2.1.1. 상관함수(correlation function)를 이용한 군집 Piccolo (1990)는 정상성(stationary), 가역성(invertible)을 만족하는 시계열 자료에 대해 자기회귀-이동평균(ARMA)모형을 적합하고 적합한 모형의 계수를 특성치 벡터로 정의하였다.

$$\phi(B)x_t = \theta(B)\epsilon_t, \qquad \pi(B) = \theta^{-1}(B)\phi(B) = \ 1 - \pi_1 B - \pi_2 B - \cdots$$

따라서 이 특성치 벡터를 이용한 거리행렬은 다음과 같다.

$$d_\pi(x_t, y_t) = \sqrt{\sum_{j=1}^{\infty}(\pi_{j,x} - \pi_{j,y})^2}.$$

Galeano와 Pena (2000)는 정상성을 만족하는 시계열 자료의 자기상관함수(ACF)를 특성치 벡터로 사용하였다. 즉, 시계열 자료 x_t와 y_t의 ACF를 이용한 특성치 벡터를 각각 ρ_x와 ρ_y라 하면 거리 행렬은 다음과 같다.

$$d_{ACF}(x_t, y_t) = \sqrt{(\rho_x - \rho_y)^T \Omega (\rho_x - \rho_y)} \tag{2.1}$$

식 (2.1)에서 Ω는 가중행렬이고, ACF 특성치 벡터의 차원은 시계열 관측치의 개수 T보다 충분히 작아야 한다.

ACF를 이용한 특성치 벡터와 유사한 방법으로 Chatfield (1979)는 정상성을 만족하는 시계열 자료의 역자기상관함수(IACF)를 특성치 벡터로 사용하여 거리를 정의하였다. 식 (2.1)에서 ACF 특성치 벡터를 IACF 특성치 벡터로 대체하면 IACF를 이용한 거리가 된다.

2.1.2. 카오틱 사상(chaotic map)을 이용한 군집 카오틱 사상을 이용한 군집(chaotic map clustering; CMC)은 물리학에서 무질서한 각 개체의 자기적 성질을 다룬 것에서 출발한다. 이 성질을 이용하여 초기에 제안된 군집분석 방법은 Blatt 등 (1996)에 의한 초상자성체군집(super paramagnetic clustering; SPC)이다. SPC에서는 유사성을 측정하기 위해 강자성을 이용한다. 강자성이랑 전하를 띄지 않는 물질들이 서로를 강하게 끌어당기는 현상이다. 강자성을 바탕으로 만든 해밀턴(Hamiltonian) 식으로 유사성 측도를 만들었는데, 이 유사성 측도를 특성치 벡터로 사용하여 군집분석을 한다.

CMC 방법은 D차원에 개체를 할당하는 점, 상관관계를 이용하는 점, 유사성 측도로서 상호정보를 구하여 군집분석에 이용한다는 점에서 SPC와 동일한 알고리즘을 가진다. CMC에서는 Manrubia와

Mikhailov (1999)가 소개한 카오틱 사상의 동시발생특성(syncronization properties)에 따라 군집이 형성된다. 비슷한 시간적 발생(time evolution)을 가지는 관측치들이 같은 군집을 형성한다. 간단한 알고리즘을 소개하면 다음과 같다.

단계1) 각 시계열사이의 상관계수를 구한다. i번째 시계열과 j번째 시계열의 상관계수를 c_{ij}라 하자.

단계2) 시계열 i와 시계열 j의 단기상호작용 J_{ij}를 다음과 같이 정의한다, Kullmann 등 (2000).

$$J_{ij} = \left[1 - \exp\left\{-\frac{n-1}{n}\left(\frac{c_{ij}}{a}\right)^n\right\}\right] I(c_{ij} > 0)$$

단, $a = 1/N \sum_{i=1}^{N} \max_j\{c_{ij}\}$로, 시계열들 사이의 상관계수들 중 가장 큰 것들의 평균이다.

단계3) 시계열의 동시발생 여부를 살펴보기 위해 다음과 같이 카오틱 사상 역학(chaotic map dynamic) $x_i(\tau)$를 구한다.

$$x_i(\tau+1) = \frac{1}{C_i}\sum_{j\neq i} J_{ij} f(x_j(\tau)), \quad \text{for } \tau = 1, 2, \ldots$$

단, $x_i \in [-1, 1]$, $C_i = \sum_{j\neq i} J_{ij}$이고, f는 로지스틱 사상(logistic map)으로 $f(x) = 1 - 2x^2$이다. 이 식은 $y_i(\tau) = f(x_i(\tau))$라 했을 때,

$$y_i(\tau+1) = f\left(\frac{1}{C_i}\sum_{j\neq i} J_{ij} y_j(\tau)\right)$$

가 되고, 이는 뉴럴 네트워크(Neural Network) 연구자들에게 친숙한 형태가 된다.

단계4) 각각의 볼츠만 엔트로피 H_i와 두 개 사이의 엔트로피 H_{ij}를 구한 후, 상호정보 I_{ij}를 다음과 같이 구한다.

$$I_{ij} = H_i + H_j - H_{ij}$$

단계5) I_{ij}를 유사성 측도로 사용하여 시계열을 군집한다.

보다 자세한 카오틱 사상 알고리즘은 Angelini 등 (2000)의 논문을 참조하라.

Basalto 등 (2005)은 CMC알고리즘을 이용하여 개별 기업의 주식을 두 개씩 쌍으로 묶어서 군집분석을 하였다. CMC방법은 시계열 자료에 대해 어떠한 모수적인 가정도 하지 않는 비모수적 군집 방법(non-parametric clustering)이므로, 주가 자료를 군집하기에 적절한 방법이라 할 수 있다. 단, 시계열 자료가 정상성을 만족해야하므로 자료를 로그 변환하여 차분하는 것이 필요하다. CMC 방법을 이용하면 시간에 따라 비슷한 형태를 가지는 주식들끼리 군집을 형성하게 된다, 김유진 (2009).

2.2. 주파수 영역(frequency domain)에서의 군집

2.2.1. 스펙트럴 밀도함수(spectral density function)를 이용한 군집 스펙트럴 분석(spectral analysis)은 정상성을 만족하는 시계열 자료의 ACF를 푸리에 변환(Fourier transform)을 통해 삼각 함수의 선형결합으로 나타내는 기법이다. Kakizawa 등 (1998)은 지진파 자료를 군집분석하기 위하여 스펙트럴 분석을 이용하였다. 두 시계열 자료의 스펙트럴 밀도함수를 추정하고 일반적인 밀도함수들 사이의 거리를 구하는 Kullback-Leibler(KL) 거리를 이용하여 두 시계열 자료의 거리를 정의했다. 두 밀도함수를 $p(\cdot)$와 $q(\cdot)$라고 하면 두 밀도함수 사이의 거리는 다음과 같다.

$$I(p; q) = E_p\left[\log\frac{p(x)}{q(x)}\right] \tag{2.2}$$

만일 $p(x)$와 $q(x)$가 평균 0인 다변량 정규분포의 밀도함수라면, KL 거리는 다음과 같이 쉽게 계산할 수 있다.

$$I(p;q) = \frac{1}{2}\left\{\mathrm{tr}\left(R_p R_q^{-1}\right) - \log\frac{|R_p|}{|R_q|} - mT\right\} \tag{2.3}$$

식 (2.3)에서 m은 다변량 시계열의 차원 수, T는 관측된 시계열 자료의 개수이고, R_p와 R_q는 각각 밀도함수 p와 q의 $mT \times mT$ 공분산 행렬이다.

식 (2.2)를 보면, KL거리는 거리 측도(distance measure)가 기본적으로 만족해야 하는 성질 중 대칭성(symmetry)을 만족하지 않음을 알 수 있다. 즉, $I(p;q) \neq I(q;p)$이다. 따라서 대칭성을 만족하기 위해 J-거리(J-divergence)를 다음과 같이 정의한다.

$$J(p;q) = I(p;q) + I(q;p).$$

Kazakos와 Papantoni-Kazakos (1980)는 다음 식이 성립함을 보였다.

$$2I(p;q) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\mathrm{tr}\left(fg^{-1}\right) - \log\frac{|f|}{|g|} - p\right)ds \tag{2.4}$$

식 (2.4)에서 $f(s)$와 $g(s)$는 각각 p와 q에 대응되는 스펙트럴 밀도함수이다. 따라서, 두 시계열의 J-거리는 다음과 같이 근사할 수 있다.

$$J(p;q) = \frac{1}{2}T^{-1}\sum_{s}\left[\mathrm{tr}\left(fg^{-1}\right) + \mathrm{tr}\left(gf^{-1}\right) - 2p\right].$$

대부분의 시계열 자료가 이산형(discrete) 형태로 관측되므로 시계열 자료의 스펙트럴 밀도 함수만 추정하면 위의 근사를 통해 J-거리를 쉽게 구할 수 있다.

2.2.2. 웨이블릿 특성치 벡터를 이용한 군집 주파수 영역에서 시계열 자료를 군집하는 다른 방법으로는 웨이블릿 분석이 있다. 스펙트럴 밀도함수는 푸리에 변환(Fourier transform)을 이용하여 시계열 자료를 코사인 함수의 선형 결합으로 나타내는데, 웨이블릿 분석에서도 마찬가지로 시계열을 기저(basis)들의 선형결합으로 나타낸다. 웨이블릿 분석을 이용하면 복잡한 신호(signal)를 간단한 기저 몇 개와 그 계수만으로 설명을 할 수 있다. 각 기저들은 스케일(scale)을 달리하면서 시계열을 분석하는데, 시간 스케일이 큰 창(window)를 통해 자료를 보면 시계열의 전반적인 특성을 알 수 있고, 시간 스케일이 작은 창을 통해 자료를 보면 시계열의 세부적인 특성을 알 수 있다.

Huhtala 등 (1999)은 시계열 자료의 군집을 위해 웨이블릿 분석을 이용하여 자료의 세부적인 움직임만을 비교하였다. 스케일이 작은 창을 통해 세부적인 움직임을 관찰하고 이 움직임이 비슷한 시계열끼리 군집하는 것이다. 먼저 주어진 자료를 웨이블릿 변환을 통해 웨이블릿 계수를 얻는다. 이 계수에 적절한 사후 처리(post processing)를 가하여 특성치 벡터를 얻어 시계열 자료를 군집한다. 웨이블릿 계수를 특성치 벡터로 바로 사용하지 않고 사후처리를 가하는 이유는 시계열을 사후 처리 후 다시 역변환하였을 때 위치불변성(transposition invariance), 추세불변성(trend invariance)과 척도불변성(scaling invariance)의 세 가지 성질을 만족하기 위한 것이다. 웨이블릿 변환을 통해 웨이블릿 계수를 얻었다면, 이 계수들 중 위치, 추세에 대한 정보를 나타내는 계수를 제거하고, 남은 계수들을 적절히 표준화함으로써 사후 처리를 할 수 있다.

간단한 예로 자료의 수가 8개라 가정해보자. 먼저 주어진 자료에 웨이블릿 변환을 하여 웨이블릿 계수를 얻는다. 웨이블릿 변환을 위해서는 자료의 수가 2^n개여야 한다. 2^n의 형태가 아니라면 이를 만족

할 때까지 자료의 개수를 줄여줘야 웨이블릿 변환을 할 수 있다. 먼저 피라미드 알고리즘(Pyramidal Algorithm) (Mallat, 1998)을 통해 웨이블릿 계수 벡터를 얻는다.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} \xrightarrow{\text{Transform T}} \begin{bmatrix} s_1 \\ d_1 \\ s_2 \\ d_2 \\ s_3 \\ d_3 \\ s_4 \\ d_4 \end{bmatrix} \xrightarrow{\text{Permute}} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} \xrightarrow{\text{Transform T}} \begin{bmatrix} s_1 \\ D_1 \\ s_2 \\ D_2 \\ d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} \xrightarrow{\text{Permute}} \begin{bmatrix} s_1 \\ s_2 \\ D_1 \\ D_2 \\ d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix}$$

위에서 변환행렬 T는 웨이블릿 필터(wavelet filter)의 종류에 따라 달라진다. 가장 많이 쓰는 웨이블릿 필터는 Daubechies-4 웨이블릿인데, 이에 해당하는 웨이블릿 변환을 행렬로 나타내면 다음과 같다.

$$\mathbf{D_4} = \begin{bmatrix} 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} & 2^{-\frac{2}{3}} \\ 0.065 & -0.241 & -0.371 & -0.548 & -0.065 & 0.241 & 0.371 & 0.548 \\ -0.354 & -0.729 & 0.046 & 0.512 & 0.171 & 0.046 & 0.137 & 0.171 \\ 0.171 & 0.046 & 0.137 & 0.171 & -0.354 & -0.729 & 0.046 & 0.512 \\ -0.837 & 0.483 & 0 & 0 & 0 & 0 & 0.129 & 0.224 \\ 0.129 & 0.224 & -0.837 & 0.483 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.129 & 0.224 & -0.837 & 0.483 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.129 & 0.224 & -0.837 & 0.483 \end{bmatrix}$$

단, $[0.129, 0.224, -0.837, 0.483]$: mother wavelet

$[0.483, 0.837, 0.224, -0.129]$: father wavelet

Daubechies-4 웨이블릿을 사용할 경우 원 시계열에 $\mathbf{D_4}$ 행렬을 곱해주면 피라미드 알고리즘을 사용한 것과 동일한 결과가 된다. 사후 처리를 위해 먼저 변환행렬 $\mathbf{D_4}$의 첫 번째 행을 곱하여 구해진 웨이블릿 계수 S_1을 제거한다. 위의 피라미드 알고리즘에서 마지막 단계의 첫 번째 성분으로, S_1은 시계열의 위치에 대한 정보를 가지고 있으므로 이를 제거함으로써 위치불변성을 만족할 수 있다. 그 다음 각 레벨의 양 끝에 있는 계수를 제거한다. 웨이블릿 계수 벡터의 k번째 성분의 레벨은 $2^{i-1} \leq k \leq 2^i$를 만족하는 i이다. 예를 들면 위의 피라미드 알고리즘에서 S_2는 레벨 1, D_1과 D_2는 레벨 2, $d_1 \sim d_4$는 레벨 3이다. 이 경우 제거하는 계수는 S_2, D_1, D_2, d_1, d_4이다.이들 계수는 추세(trend)에 관한 정보를 가지고 있으므로 이들을 제거함으로써 추세불변성을 만족할 수 있다. 마지막으로 척도불변성을 만족하기 위해 계수 벡터의 스케일을 조정해 주어야 한다. 경제 시계열의 경우 변동 폭이 기간에 비례하여 커지기 때문에 레벨 i에 해당하는 계수에 2^i을 곱해준다. 마지막으로 이렇게 얻은 벡터를 표준화(normalizing)하여 척도불변성까지 만족할 수 있고 이를 특성치 벡터로 사용하여 시계열을 군집하면 된다.

위치, 추세, 척도가 다른 시계열 자료의 경우도 시계열의 세밀한 움직임이 동일하다면 같은 군집에 속한 것으로 생각할 수 있다, 이정현 (2008). 그림 2.1(왼쪽)은 미화 1달러에 대한 남아프리카 공화국 랜드(Rand)와 스위스 프랑(Franc)의 시계열 그림을 나타내는데, 위치와 추세가 다르므로 비슷한 시계열로 생각되지 않는다. 사후처리를 거치고 난 후의 그림 2.1(오른쪽)을 보면 남아프리카 공화국 랜드와 스위스 프랑의 미국 1달러에 대한 환율이 비슷해짐을 볼 수 있다 (Huhtala 등, 1999). 따라서 시계열의 세밀한 움직임이 비슷하기 때문에 웨이블릿을 이용한 군집에서는 두 시계열을 비슷한 시계열로 생각한다.

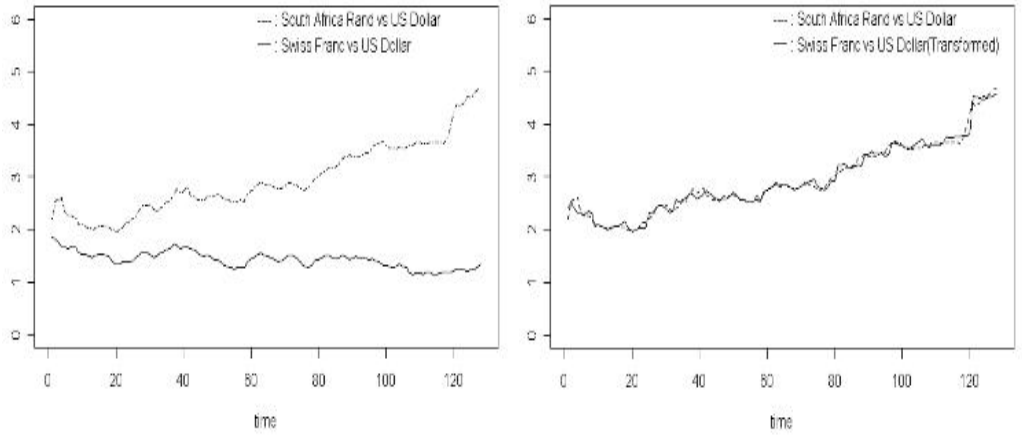

그림 2.1. (왼쪽) 미국 1달러에 대한 남아프리카 공화국 Rand와 스위스 Franc의 시계열도, 1986년 5월∼1996년 12월, 월별 자료, (오른쪽) 미국 1달러에 대한 남아프리카 공화국 Rand 와 변환한 스위스 Franc의 시계열도

사후 처리를 통해 특성치 벡터를 얻더라도 이 특성치 벡터가 원 시계열의 차원을 많이 줄여주지는 못한다. Huhtala 등 (1999)이 제시한 위의 방법대로 사후 처리를 하면, 시계열 자료의 개수가 $T = 2^n$개 일 때 특성치 벡터의 차원은 $T - 2n$이 된다. 예를 들어 128개 자료로부터 얻어진 특성치 벡터의 차원은 $128 - 2 \times 7 = 114$로 역시 매우 크다. 한 가지 대안은 PCA(principal component analysis)를 사용하는 것이다. PCA를 사용하면 114개로 설명을 해야 하는 것을 약 15개 내외로 전체 자료를 99%이상 설명 가능하다.

웨이블릿 분석을 이용한 군집의 가장 큰 장점은 원 자료가 정상성을 만족하지 않더라도 의미 있는 변환이 가능하다는 것이다. 스펙트럴 분석에서 푸리에 변환을 통해 자료를 코사인 함수의 선형결합으로 나타내기 위해서는 정상성 가정이 반드시 필요하다. 정상성을 만족하지 않으면 스펙트럴 분석 결과가 아무 의미가 없다. 예를 들어 자료에 추세가 있을 경우 자료의 주기를 잡아낼 수 없으므로 스펙트럴 분석을 위해서는 자료를 차분한 후 분석을 해야 한다. 하지만 웨이블릿 분석에서는 정상성을 만족하지 않더라도 웨이블릿 분석 결과의 의미는 변하지 않는다. 사후 처리를 통해서 추세를 제거하고 스케일을 조정하기는 하지만 이는 자료의 세부적인 움직임만을 관찰하기 위한 것이지 자료가 정상성을 만족하도록 만들기 위한 것은 아니다. 그림 2.1(오른쪽)을 보아도 환율 자료의 추세가 완전하게 제거되지는 않았음을 알 수 있다. 자료가 차분(difference)되어 정보를 잃는 것을 걱정할 필요도 없다. 웨이블릿 분석에 있어서의 유일한 가정은 자료가 기저들의 선형결합으로 나타나야 한다는 것이다.

2.2.3. 경험모드 분할(EMD)과 스펙트럴 밀도함수를 이용한 군집 Huang 등 (1998)은 경험모드분할(empirical mode decomposition; EMD)을 제안하여 비정상, 비선형 시계열 자료의 Hilbert transform에 의한 순간빈도(instantaneous frequency; IF)를 구하는 것을 가능하게 하였다. 웨이블릿 분석이 시계열 자료 또는 신호를 미리 정의된 기저들의 선형 결합으로 나타내는 방법임에 비해, EMD 방법은 어떠한 가정도 없이 시계열을 경험적으로 진동(oscillation)에 근거한 내재모드함수(intrinsic mode function; IMF)들의 합으로 분할하는 방법이다.

IMF는 다음 두 가지 성질을 만족하는 함수를 말한다. 첫째, 진동을 표현할 수 있는 두 종류의 값인 극값(extrema)의 수나 함수 값이 0인 값(Zero-crossing)의 개수가 같거나 1개만 차이가 나야 하고, 둘째, 임의의 점에서 극대값(local maxima)에 의해 만들어지는 윗덮개(upper envelope)와 극소값(local

minima)에 의해 만들어지는 아래덮개(lower envlope)의 평균이 0이어야 한다. IMF의 성질을 만족하는 가장 대표적인 예는 사인 함수와 코사인 함수이나, IMF가 되기 위해서 진폭이 상수일 필요는 없다. 스펙트럴 분석이나 웨이블릿 분석과는 달리 EMD 방법은 수학적인 이론 배경이 없다는 것이 단점으로 지적될 수 있다.

$x(t)$를 분할하고자 하는 시계열이라 할 때, 첫 번째 IMF $h_1(t)$를 구하는 과정은 다음과 같다.

단계1) $y_0(t) = x(t)$

단계2) $i = 0$

단계3) $y_i(t)$ 극대값, 극소값들을 각각 B-스플라인(B-spline)으로 연결하여 윗덮개와 아래덮개를 구한다. 이를 각각 $ue_i(t)$, $le_i(t)$라 한다.

단계4) 함수 $ue_i(t)$와 $le_i(t)$ 의 평균함수를 $ae_i(t)$, 즉 $ae_i(t) = (ue_i(t) + le_i(t))/2$이라 하고, $y_{i+1}(t) := y_i(t) - ae_i(t)$를 구한다.

단계5) $ae_i(t) \approx 0$이 되거나 특정 stopping rule을 만족할 때까지 단계3)∼단계5)를 반복한다. 최종 번째 값은 IMF의 성질을 만족하게 되고 이를 첫 번째 IMF로 둔다 ($h_i(t) := y_k(t)$).

$y_0(t) = x(t) - h_1(t)$라 하고, 위와 같은 과정을 다시 거치면 두 번째 IMF $h_2(t)$를 구할 수 있다. 세 번째 IMF $h_2(t)$는 $y_0(t) = x(t) - h_1(t) - h_2(t)$라 놓고 마찬가지 방법으로 구한다. 이런 식으로 $h_j(t) \approx 0$, $r(t) \approx 0$ 혹은 $r(t)$가 단조함수가 될 때까지 IMF 함수들을 구한다. 최종적으로 시계열 $x(t)$는 다음과 같이 분할된다.

$$x(t) = \sum_{j=1}^{J} h_j(t) + r(t)$$

$r(t)$는 잔차항으로, 추세가 있는 시계열의 경우 추세를 나타내는 함수가 된다. IMF $h_j(t)$에서, j가 커질수록 시계열의 전반적인 특성을 나타내는 함수가 되고, j가 작아질수록 시계열의 세부적인 특성을 나타내는 함수가 된다. 따라서 $h_1(t)$는 오차를 나타내는 경우가 많다. 자세한 내용은 Huang 등 (1998)을 참고하라.

EMD에 의한 시계열의 군집은 EMD를 이용하여 구한 IMF의 순간빈도를 이용한다. 첫 번째나 두 번째 IMF를 이용해서 시계열을 군집하면 세부적인 움직임이 비슷한 시계열끼리 군집하는 것이 되고, 세 번째, 혹은 그 뒤의 IMF를 이용해서 시계열을 군집하면 시계열의 전반적인 특성이 비슷한 시계열끼리 군집하는 것이 된다. 박쥐 신호 자료와 같이 순간빈도에 정보를 담은 신호의 경우 이러한 순간빈도를 이용하여 군집분석을 실시하는 것이 의미가 있다. 자세한 내용은 박민정 (2009)을 참고하라.

확률과정을 따르는 시계열 자료의 경우 시간에 따라 변하는 주파수의 정보를 이용하는 것은 자료의 특성상 그 의미를 찾기 힘들다. 또한 EMD 방법이 이러한 자료에 대해 filtering성향도 가지므로, Flandrin 등 (2004), 어떤 IMF를 선택하는가의 문제를 해결하기 위해서는 이론적인 연구가 선행되는 것이 필요하다. 따라서 이러한 자료의 분석을 위해서는 전체 기간 동안의 주파수의 정보 모두를 가지는 스펙트럴 밀도 함수를 다시 이용하기로 한다. 이때, Huang 등 (1998)에 언급된 marginal Hilbert spectrum을 사용하지 않는 것은 이론적인 바탕이 아직 정립되지 않았을 뿐만 아니라 Kakizawa 등 (1998)에서 언급된 스펙트럴 밀도함수의 값이 0에 가까울 경우의 문제가 marginal Hilbert spectrum에서는 더욱 심각하기 때문이다.

다만, 정상성을 만족시키기 위하여 자료를 변환하는 대신, EMD 방법을 이용하여 추세를 제거한 시계열 자료를 가지고 Kakizawa 등 (1998)의 방법을 따라 분석하도록 한다. Kakizawa 등 (1998)은 지진파 자료를 분석했는데 지진파 자료는 정상성을 따른다고 볼 수 없으나 추세가 없는 자료이므로, EMD를 이용

표 **3.1.** KOSPI 200에 속한 15개 군집대상 기업

업종	번호	기업명
금융업종	OB 1	Dawoo Securities(대우증권)
	OB 2	Samsung Securities(삼성증권)
	OB 3	Hyundai Securities(현대증권)
	OB 4	Woori Finance HLDG(우리자산운용)
	OB 5	Daishin Securities(대신증권)
건설업종	OB 6	Daelim Industrial(대림건산업)
	OB 7	Hyundai Eng&Cons(현대중공업)
	OB 8	Dongbu Construction (동부건설)
	OB 9	Doosan Construction(두산건설)
	OB 10	GS Engineering(GS건설)
에너지업종	OB 11	Samchully(삼천리)
	OB 12	Korea Gas(한국가스공사)
	OB 13	Korea Elec Power(한국전력)
	OB 14	Kyungnam Energy Co(경남에너지)
	OB 15	Kyungdong City Gas(경동도시가스)

해 추세를 제거한 후 스펙트럴 밀도함수를 이용하여 군집분석을 한다. 이는 정상성을 가정할 수 없는 지진파 자료의 스펙트럴 밀도함수를 구한 것처럼, 몇 개의 IMF 합으로 나타나는 자료 또한 스펙트럴 밀도함수를 적용할 수 있는 자료의 범위에 들어간다는 가정 하에 분석한 것이다. 또 그 결과를 2.2.1절의 정상성을 만족시키도록 자료를 변환하고 같은 방법으로 분석한 결과와 비교하여 본다. EMD에 의해 제거되는 추세의 의미에 관한 문제나 스펙트럴 밀도 함수를 이용할 수 있는 자료의 범위에 관한 문제들은 좀 더 연구가 필요한 분야가 된다.

3. 실증 분석

이 절에서는 2절에서 소개된 군집 방법들을 이용하여 실제 주가 자료들에 군집분석을 시행하여 각 방법들의 유용성을 비교해 보겠다. 분석에 사용될 경제시계열 자료는 KOSPI 200에 속하는 기업들 중 금융업, 건설업, 에너지의 3업종에서 각각 5개의 기업들을 선정하여 총 15개 개업의 2005년 1월 3일부터 2008년 12월 17일까지의 일별 주가자료를 사용했다. 경제 시계열 자료의 경우 ARMA모형이 잘 적합되지 않으므로 상관 함수를 이용한 군집은 하지 않고 나머지 비모수적인 방법들만을 사용하여 군집분석을 시행하였다. 군집분석 대상 기업은 표 3.1과 같다. 자료는 한국거래소(www.krx.co.kr)에서 구하였다. 1∼5번까지는 금융업종, 6∼10번까지는 건설업종, 11∼15번까지는 에너지업종의 기업들이다.

웨이블릿 분석을 위해서는 자료의 개수가 2^n 형태여야 하기 때문에, 각 기업 별로 분석에 사용된 자료는 1,024개의 일별 관측치이다. 같은 업종에 속한 기업의 주가 자료는 비슷한 흐름을 가질 것이다. 개별 기업의 등락폭은 달라도 같은 업종에 속해있다면 오를 때 같이 오르고 내릴 때 같이 내리는 경우가 많다. 따라서 각 군집 방법이 같은 업종의 기업들을 잘 군집하는지 살펴봄으로써 효용성을 비교해볼 수 있다.

그림 3.1(a)와 (b)는 각각 카오틱 사상과 스펙트럴 밀도함수를 이용한 군집 결과이다. 이 두 가지 방법은 자료가 정상성을 만족한다는 가정 하에 군집이 이루어지기 때문에 주가를 이용하지 않고 로그수익률을 사용한다. $P_i(t)$를 i번째 기업의 t번째 일의 마감 가격이라 했을 때, 군집에 사용한 i번째 기업의 로그 수익률 $y_i(t)$는 다음과 같다.

$$y_i(t) = \ln r_i(t) = \ln \frac{P_i(t)}{P_i(t-1)} = \ln P_i(t) - \ln P_i(t-1)$$

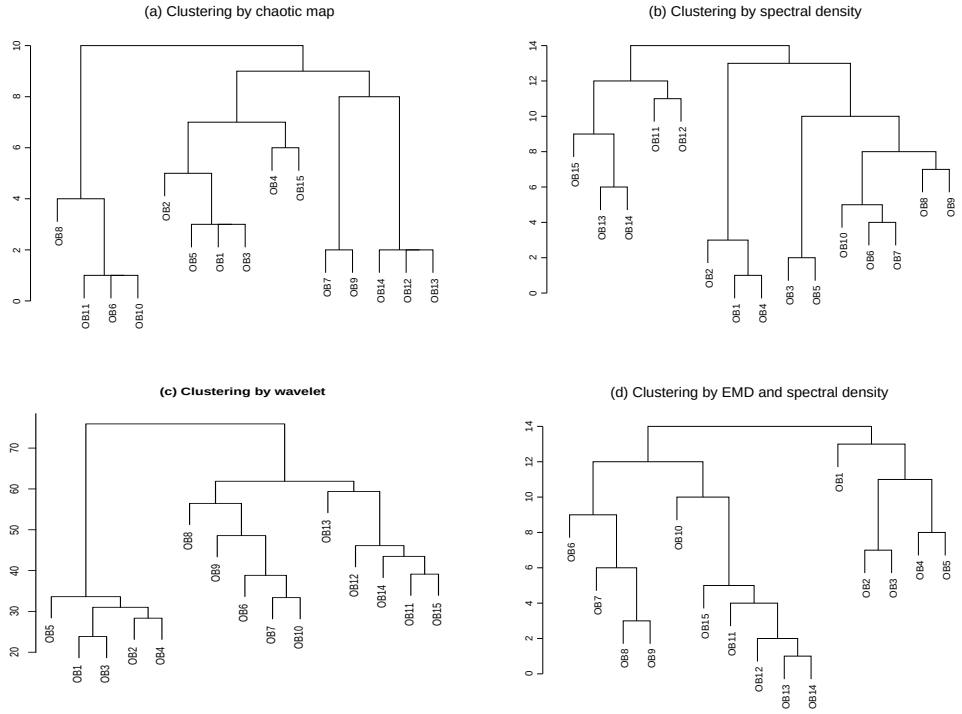

그림 3.1. 주가자료 군집분석

카오틱 사상을 이용한 군집에서는 금융업종에 속한 기업들이 하나의 군집으로 잘 묶여 있고 건설업종과 에너지 업종의 기업들은 서로 섞여 있다. 카오틱 사상을 이용한 군집은 자료들 간의 상관관계를 이용하여 동시발생 특성에 따라 군집을 형성한다. 따라서 전체적으로 거시적인 움직임이 비슷한 자료들끼리 같은 군집에 묶이기 때문에, 금융업종이 다른 업종에 비해 동종 업종에 속한 기업들 간의 주가 움직임이 비슷하다고 생각할 수 있다.

스펙트럴 밀도함수를 이용한 군집의 경우에는 현대증권(OB3)과 대신증권(OB5)이 건설업종과 함께 군집을 이룬 것을 제외하고는 동일한 업종들의 군집이 비교적 잘 형성되어 있다. 특히 금융업종과 건설업종에 속한 기업들이 강하게 묶여있음을 볼 수 있다. 스펙트럴 밀도함수를 이용한 군집은 주기가 비슷한 자료들끼리 군집을 형성하기 때문에 금융업종과 건설업종에 속한 기업들의 주가는 비슷한 주기를 가진다고 볼 수 있다. 이는 이 두 업종의 기업들은 전체적인 경기 변동 상황에 따라 민감하게 반응하기 때문인 것으로 생각된다.

그림 3.1(c)는 웨이블릿을 이용한 군집 결과이다. 웨이블릿을 이용한 군집은 자료를 차분할 필요가 없으므로 로그 수익률을 사용하지 않고 주가를 사용하여 군집분석을 실시하였다. 모든 기업이 동종 업종끼리 군집을 잘 형성하고 있음을 볼 수 있다. 특히 금융업종들이 서로 강하게 묶여있는 것이 눈에 띈다. 금융업종에 속한 기업들, 특히 증권업종의 기업들은 전체적인 경기 변동 상황에 예민하게 반응할 수밖에 없으므로 주가의 세밀한 움직임이 타 업종들보다 더욱 서로 긴밀하게 연관되어 있을 것이라 기대할 수 있는데, 군집 결과에서 금융 업종의 기업들이 타 업종의 기업들보다 가까운 거리에서 묶여 있는 것으로 이를 확인할 수 있다. 전체적인 군집 결과가 좋은 것으로 보아 경제 시계열을 군집하는데 있어서 자료의

세밀한 움직임만을 이용하여 군집하는 것이 타당하다고 생각된다.

마지막으로 EMD와 스펙트럴 밀도함수를 이용한 분석결과는 그림 3.1(d)와 같다. GS건설(OB 10)을 제외하고는 각 업종별로 군집이 잘 이루어진 것으로 보인다. 이것은 EMD 결과 나타나는 추세를 나타내는 잔차항과 큰 움직임을 나타내는 몇 개의 IMF들을 제외하고 처음 5개의 IMF들의 합으로 이루어진 자료를 사용하였으므로 세밀한 움직임이 비슷한 금융업종의 기업들이 더욱 강하게 묶여있음을 볼 수 있다. 잔차항을 제거할 경우 IMF를 몇 개까지 사용하는가는 결과에 큰 영향을 미치지 않았다.

4. 결론

본 논문에서는 시계열 자료를 군집하는 다양한 방법들에 대해 알아보았다. 모수적인 방법으로는 ARMA모형을 적합하여 군집하는 방법이 있고, 비모수적인 방법으로는 카오틱 사상을 이용한 방법, 스펙트럴 밀도 함수를 이용한 방법, 웨이블릿 특성치 벡터를 이용한 방법, 경험모드분할(EMD)을 이용한 방법이 있다. 카오틱 사상을 이용한 방법과 스펙트럴 밀도 함수를 이용한 방법은 자료의 정상성 가정이 필요하므로, 비정상 시계열의 경우 자료를 먼저 차분한 뒤 군집분석을 시행해야 한다. 웨이블릿 특성치 벡터를 이용한 방법과 경험 모드 분할을 이용한 방법은 정상성 가정이 필요 없으므로 자료의 차분으로 인한 정보의 손실을 걱정할 필요가 없다.

비모수적 방법들을 이용하여 실제 주가자료를 군집해본 결과 큰 차이는 없으나 웨이블릿을 이용한 군집이 가장 좋은 결과를 보였음을 확인 할 수 있다. 또한, 자료를 차분하여 스펙트럴 분석을 시행한 것 보다는 EMD를 이용해 추세를 제거하여 스펙트럴 군집을 한 것이 더 좋은 결과를 보였다. 차분으로 인한 정보의 손실이 시계열을 군집하는데 있어 영향을 미침을 확인할 수 있다. EMD 방법의 경우 수학적인 이론 배경이 없다는 것이 가장 큰 약점인데, EMD를 이용해 분리된 추세와 IMF 함수들의 의미에 대한 연구가 진행된다면 좀 더 효과적인 군집을 시행할 수 있을 것이다.

참고문헌

김유진 (2009). Comparison study of time series methods, Master thesis, Seoul National University.

박민정 (2009). Time-Frequency analysis by multiscale methods with applications to bat signals, Ph.D thesis, Seoul National University.

0|정현 (2008). Clustering analysis of financial time series using wavelet analysis, Master thesis, Seoul National University.

Angelini, L., De Carlo, F., Marangi, C., Pellicoro, M. and Stramaglia, S. (2000). Clustering data by inhomogeneous chaotic map lattices, *Physical Review Letters*, **85**, 554–557.

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P. and Pascazio, S. (2005). Clustering stock market companies via chaotic map synchronization, *Physica A: Statistical Mechanics and its Applications*, **345**, 196–206.

Blatt, M., Domany, E. and Wiseman, S. (1996). Super-paramagnetic clustering of data, *Physical Review Letters*, **76**, 3251–3254.

Chatfield, C. (1979). Inverse autocorrelations, *Journal of Royal Statistical Society Series A*, **142**, 363–377.

Flandrin, F., Rilling, G. and Goncalves, P. (2004). Empirical mode decomposition as a filter bank, *IEEE Signal Processing Letters*, **11**, 112–114.

Galeano, P. and Pena, D. (2000). Multivariate analysis in vector time series, *Resenhas*, **4**, 383–404.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H. (1998). The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series, *Proceedings of the Royal Society London A*, **454**, 903–995.

Huhtala, Y., Karkkainen, J. and Toivonen, H. (1999). Mining for similarities in aligned time series using wavelets, *Data Mining and Knowledge Discovery: Theory, Tools, and Technology, Proceedings of SPIEE*, **3695**, 150–160.

Kakizawa, Y., Shumway, R. H. and Taniguchi, M. (1998). Discrimination and clustering for mutivariate time series, *Journal of the American Statistical Association*, **93**, 320–340.

Kazakos, D. and Papantoni-Kazakos, P. (1980). Spectral distance measures between Gaussian processes, *IEEE Trans. Automatatic Control*, **vAC-25**, 950–959.

Kullmann, L., Kertesz, J. and Mantegna, R. N. (2000). Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions, *Physica A*, **287**, 412–419.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*, Academic Press, San Diego.

Manrubia, S. C. and Mikhailov, A. S. (1999). Mutual synchronization and clustering in randomly coupled chaotic dynamical networks, *Physical Review E*, **60**, 1579–1589.

Piccolo, D. (1990). A distance measure for classifying ARIMA models, *Journal of Time Series Analysis*, **11**, 153–164.

Comparison Study of Time Series Clustering Methods

Hanwoom Hong[1] · Minjeong Park[2] · Sinsup Cho[3]

[1]Department of Statistics, Seoul National University; [2]Pohang Mathematics Institute
[3]Department of Statistics, Seoul National University

(Received November 2009; accepted November 2009)

Abstract

In this paper we introduce the time series clustering methods in the time and frequency domains and discuss the merits or demerits of each method. We analyze 15 daily stock prices of KOSPI 200, and the nonparametric method using the wavelet shows the best clustering results. For the clustering of nonstationary time series using the spectral density, the EMD method remove the trend more effectively than the differencing.

Keywords: Time series, clustering, chaotic map, spectral analysis, wavelet, EMD.

Minjeong Park is partially supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(Grant 2009-0094070).

[3]Corresponding author: Professor, Department of Statistics, Seoul National University, 599, Gwanak-ro Gwanak-gu, Seoul 151-747, Korea. E-mail: sinsup@snu.ac.kr