The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.525
29(3), 525–537

EMD based hybrid models to forecast the KOSPI

Hyowon Kim[a] · Byeongchan Seong[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received February 26, 2016; Revised February 29, 2016; Accepted February 29, 2016)

Abstract

The paper considers a hybrid model to analyze and forecast time series data based on an empirical mode decomposition (EMD) that accommodates complex characteristics of time series such as nonstationarity and nonlinearity. We aggregate IMFs using the concept of cumulative energy to improve the interpretability of intrinsic mode functions (IMFs) from EMD. We forecast aggregated IMFs and residue with a hybrid model that combines the ARIMA model and an exponential smoothing method (ETS). The proposed method is applied to forecast KOSPI time series and is compared to traditional forecast models. Aggregated IMFs and residue provide a convenience to interpret the short, medium and long term dynamics of the KOSPI. It is also observed that the hybrid model with ARIMA and ETS is superior to traditional and other types of hybrid models.

Keywords: intrinsic mode function, exponential smoothing method, ARIMA model, nonstationary model, nonlinear model

1. 서론

전통적으로 시계열 자료 분석 및 예측을 위하여 가장 널리 사용되는 방법은 Box와 Jenkins의 ARIMA (autoregressive integrated moving-average) 모형이다. Box 등 (1993) 이후로, ARIMA 모형은 다양한 형태로 변화 및 진화하였을 뿐만 아니라 다변량 기법으로 확장되었으며 그 이론적인 배경 또한 탄탄하다. 가장 중심이 되는 개념은 정상성(stationarity)으로 볼 수 있으며 활용성이 가장 높은 모형은 선형 모형으로 볼 수 있다. 만약 관심의 시계열이 비정상적(nonstationary)이거나 비선형 모형을 따르는 경우, ARIMA 모형은 차분 또는 Box-Cox 멱변환(power transformation)과 같은 변환을 가한 시계열에 적합되는 것이 일반적이다 (Wei, 2006). 그러나, 미지의 다양한 형태의 비정상성과 비선형성이 이러한 변환을 통하여 정상화 또는 선형화된다는 보장은 없으며, 특히 변환을 통한 ARIMA 모형의 적합은 그 해석 가능성에 대하여 여러가지 한계를 가지고 있다.

1950년대부터 시작된 지수평활법 (Brown, 1959; Holt, 1957; Winters, 1960)은 최근 다양한 형태로 발전하였으며 그 실용성과 자동화된 모형 적합으로 인하여 다양한 분야에서 각광받고 있다 (Hyndman과 Khandakar, 2008). 특히 지수평활법이 잘 구현되어 있는 R 패키지인 forecast는 최근 다운로드 횟수가 폭발적으로 증가했다고 보고되고 있다 (http://robjhyndman.com/hyndsight/fpp-downloads/). 고주

This research was supported by the Chung-Ang University Research Scholarship Grants in 2014.
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 221 Heukseok-Dong, Dongjack-Gu, Seoul 06974, Korea. E-mail: bcseong@cau.ac.kr

파수 및 계절 시계열 자료 분석을 위한 지수평활법도 뛰어난 자료 적응력, 추정방법의 편리성 및 비선형 모형의 가능성 등에서 우수함을 인정받고 있으며 전력수요를 포함하는 에너지 관련 수요예측 분야, 교통량 예측 및 각종 제품의 수요 예측과 같은 연구들을 통해서 발전하고 있다 (Gould 등, 2008).

최근 시계열 자료에서 나타나는 비정상성과 비선형성과 같은 복잡성을 포용할 수 있는 다양한 방법론들이 연구되고 있다. 그 중에서도 경험적모드분해법(empirical mode decomposition; EMD)에 관한 연구들은 주목할 만한 성장을 거듭하고 있다.

Hilbert-Hwang 변환을 이용한 Huang 등 (1998)의 선구적인 연구에서 비롯되어 EMD는 비선형 또는 비정상 시계열을 고주파수에서 저주파수에 걸친 독립적인 성분들로 분해함으로써 복잡한 시계열에 대한 해석 가능성을 높여주는 분해법이다. 또한, 구현하기 쉬운 알고리즘으로 원시계열이 포함하고 있는 변동들을 자동적으로 자료에 부합되게 추출한다는 장점을 가지고 있다 (Kim 등, 2008).

최근 EMD는 시계열 예측을 위해서 기존 시계열 모형과 결합된 혼합 모형(hybrid model)의 형태로 자주 활용되고 있다. 예를 들면, Kim 등 (2008)은 EMD와 벡터 시계열 모형을 결합하여 사이버 바이러스 관련 자료를 분석 및 예측하였으며, Wei와 Chen (2012)는 지하철 승객 통행량을 예측하기 위하여 EMD와 신경망(neural networks) 모형을 결합하였다. Park과 Seong (2014)는 EMD를 이용하여 한국의 주요 거시 경제 지표를 대상으로 순환변동과 추세 성분을 추출하고 예측에 활용한다. Zhu 등 (2015)는 탄소 가격 시계열 자료를 기간별로 해석하기 위하여 EMD에 의한 IMF들의 성질을 면밀히 분석하고 계층적 군집화(hierarchical clustering)를 사용하여 그룹화하고 해석의 편의성을 제고하였다.

본 논문에서는 시계열 자료에 내포된 비정상성과 비선형성과 같은 복잡성을 효과적으로 다루기 위한 혼합 모형을 연구한다. 특히, EMD에 의한 내재모드함수(intrinsic mode function; IMF)의 해석 및 예측 편리성을 개선하기 위하여 누적에너지(cumulative energy)의 개념을 도입하여 IMF들을 통합하였으며, 통합된 IMF 및 residue의 성분들은 ARIMA 모형 및 지수평활법의 혼합 모형으로 예측한다.

본 논문의 2장에서는 기존의 시계열 모형 및 EMD 방법론에 대하여 소개하며, 3장에서 본 연구에서 제안하는 EMD로 생성된 IMF들을 통합하는 방법 및 예측 방법을 설명한다. 4장에서 제안된 방법을 일별 코스피 지수의 예측에 적용한다. 5장에서 결론을 맺는다.

2. 시계열 모형 및 분해 기법

이 장에서는 전통적으로 시계열 자료 분석 및 예측을 위해서 널리 사용되는 모형인 ARIMA 모형 및 지수평활법과, 시계열을 시간-진동수 영역에서 시계열을 분해하는 방법인 EMD 방법을 소개한다. 간단한 설명을 위하여 시계열 자료는 계절성이 없는 것으로 가정한다.

2.1. ARIMA 모형

현재의 시계열(x_t)을 자기회귀항($x_{t-1}, x_{t-2}, \ldots$) 및 오차항의 이동평균항($\varepsilon_t, \varepsilon_{t-1}, \ldots$)의 선형결합으로 표현하는 ARIMA 모형은, 단기적 예측에서 더욱 널리 사용되는 모형이다. 특히, x_t가 비정상 시계열일 경우 정상성을 위한 변환으로 차분(differencing)이 선행된다. 시계열 x_t가 ARIMA(p, d, q)를 따를 때 다음과 같이 표현할 수 있다,

$$y_t = \theta_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \tag{2.1}$$

여기서 y_t는 x_t를 d번 차분한 값으로 $y_t = (1-B)^d x_t$이며 B는 차분 연산자이다. p는 자기회귀항의 차수, q는 이동평균항의 차수를 나타낸다.

2.2. 지수평활법

지수평활법은 시계열을 수준(level), 추세(trend) 및 계절(seasonal) 성분으로 분해하여 자료를 분석 및 예측하는 방법으로서, 최근 상태공간모형(state space model)과 결합된 형태로 확장된 여러 형태를 가지고 있다 (De Livera 등, 2011). 일반적으로 가장 예측력이 높다고 알려진 지수평활 모형은 완화된(damped) 형태이다. 이 방법은 갱신방정식(update equation) 및 예측방정식(forecast equation)에 완화모수(damping parameter) ϕ를 도입하여 좀더 보수적인 예측을 한다고 알려져 있다. 계절이 없는 시계열에 대하여 완화된 지수평활 모형은 수준 ℓ_t와 추세 b_t를 위한 갱신방정식과 t-시점에서 h-시점 후 시계열 x_{t+h}를 예측하는 예측방정식으로 구성되며 다음과 같은 형태를 가진다,

$$\ell_t = \alpha x_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}), \tag{2.2}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1}, \tag{2.3}$$

$$\hat{x}_{t+h|t} = \ell_t + \left(\phi + \phi^2 + \cdots + \phi^h\right) b_t, \tag{2.4}$$

여기서 α와 β는 평활모수(smoothing parameter)이고 $0 < \phi < 1$은 완화모수이다.

2.3. EMD 방법

EMD 방법은 시계열을 서로 직교하는 IMF들로 분해하는 방법으로, IMF는 다음의 2가지 조건을 모두 만족하는 함수로 정의된다 (Kim과 Oh, 2009).

(1) 시계열 자료에 대한 극값들의 개수와 영을 통과하는 점들의 개수의 차이는 최대 한개 이하이어야 한다.

(2) 극대값들을 연결한 상위막(upper envelope)과 극소값들을 연결한 하위막(lower envelope)의 국소 평균값은 0이다.

즉, EMD 방법을 통하여 시계열 x_t는 다음과 같이 표현될 수 있다,

$$x_t = \sum_{i=1}^{m} \mathrm{imf}_t^{(i)} + r_t, \tag{2.5}$$

여기서 $\mathrm{imf}_t^{(1)}, \ldots, \mathrm{imf}_t^{(m)}$는 m개의 IMF들을 나타내며, r_t는 시계열에서 IMF들이 분해되고 남은 잔차를 나타내며 residue라고 부른다.

하나의 IMF는 시간축에 대하여 가변적인 진폭(amplitude)과 진동수(frequency)를 가질 수 있으며, 체거름 과정(shifting process)이라고 부르는 다음의 반복 과정을 통하여 생성된다.

(1) 시계열 x_t의 극대값과 극소값들을 식별한다.

(2) 모든 극대값들을 스플라인 보간법(spline interpolation)을 이용하여 연결하고 상위막 $e_t^{\max}$를 만든다.

(3) 모든 극소값들에 대하여 (2)의 과정과 동일한 방법으로 연결하여 하위막 $e_t^{\min}$를 만든다.

(4) 상위막과 하위막으로부터 평균값 $m_t = [e_t^{\max} + e_t^{\min}]/2$를 계산한다.

(5) 시계열 x_t에서 (4)의 과정에 의한 평균값을 빼고 남는 성분을 $h_t = x_t - m_t$로 정의하고, h_t가 IMF의 2가지 조건을 만족하는지 검토한다.

(6) 성분 h_t가 IMF 조건을 만족하면 $\text{imf}_t = h_t$로 두고, 성분 h_t가 IMF 조건을 만족하지 않으면 h_t를 원시계열 x_t처럼 간주하고 (1)–(5)의 과정을 다음의 정지 조건(stopping rule)이 만족될 때까지 반복한다,

$$\sum_t \left(\frac{h_t^{(i)} - h_t^{(i-1)}}{h_t^{(i-1)}} \right)^2 < \delta. \tag{2.6}$$

단, $h_t^{(i)}$와 $h_t^{(i-1)}$는 각각 현재 단계(i)와 이전 단계$(i-1)$에서 생성된 IMF의 후보값인 성분 h_t를 나타내고, δ는 미리 정해진 작은 값이다.

3. 혼합 모형을 이용한 예측 기법

이 장에서는 2장에서 설명된 기존의 시계열 자료 분석 기법들을 조합한 혼합 모형을 소개하고 예측에 활용하는 방법을 설명한다. 특히, 본 연구에서는 EMD에 의한 IMF들을 그룹화하여 통합하고, 통합된 IMF들 및 residue의 특성에 맞게 시계열 모형을 적합하고 예측에 활용하는 방법을 제안한다.

3.1. IMF의 활용

EMD에 의하여 생성된 IMF들은, 원시계열을 다양한 주파수를 가지는 성분들로 분해한 것이다. 따라서, 각 IMF들을 이용하면 시계열의 움직임을 더 미세하게 설명할 수 있을 뿐만 아니라, 시계열 자료의 예측에도 큰 역할을 할 수 있다. 그러나, 시간-진동수 영역에서 정의되는 개별 IMF들을 예측하기는 쉽지 않으며, 기존 문헌들에서는 시간 영역에서의 기법들을 그대로 적용하고 있다.

본 연구에서는 이러한 개별 IMF들을 예측하는 어려움을 극복하기 위하여 개별 IMF를 그룹화하는 방법을 고려하였다. 편의상 IMF들은 두 개의 그룹(고주파수 그룹과 저주파수 그룹)으로 분류하여 통합하였으며, residue는 추세를 나타내는 성분이므로 하나의 독립 그룹으로 유지하였다. 즉, 고주파수 그룹 imfs_t^{high}와 저주파수 그룹 imfs_t^{low}는 다음과 같이 정의하며,

$$\text{imfs}_t^{high} = \sum_{i=1}^{m^*-1} \text{imf}_t^{(i)}, \tag{3.1}$$

$$\text{imfs}_t^{low} = \sum_{i=m^*}^{m} \text{imf}_t^{(i)}, \tag{3.2}$$

시계열 x_t는 다음과 같이 표현될 수 있다,

$$x_t = \text{imfs}_t^{high} + \text{imfs}_t^{low} + r_t. \tag{3.3}$$

m^*은 Kim 등 (2008)에서 고려된 누적에너지 개념을 사용하여 정하였으며, 누적에너지의 변화량이 가장 큰 IMF의 번호를 m^*로 결정하였다. k번째 IMF까지의 누적에너지 E_k의 정의는 다음과 같다,

$$E_k = \frac{\sum_{s=1}^{k} \sum_t w^2(t,s)}{\sum_{s=1}^{m} \sum_t w^2(t,s)}. \tag{3.4}$$

단, $w(t,s)$는 시간 t에서 s번째 IMF인 $\text{imf}_t^{(s)}$의 순간진폭(instantaneous amplitude)를 나타낸다.

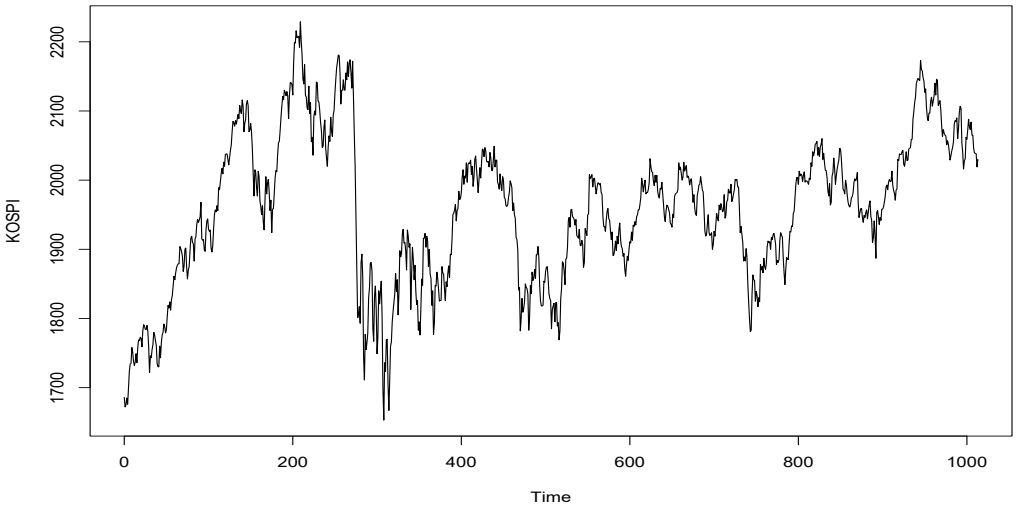

Figure 4.1. Time series plot for KOSPI from July 1, 2010 to July 31, 2015.

3.2. 혼합 모형 및 예측

기존 연구들은 EMD의 결과를 예측에 이용하기 위하여 주로 ARIMA 모형을 적용하였다. 그러나,이러한 EMD + ARIMA 혼합모형은 IMF 및 residue의 성질에 적당하지 않다고 볼 수 있다. 예를 들면, IMF들은 추세를 가질 수 없고 residue는 일반적으로 비정상적이며 고차로 적분되어 있는(integrated of high-order) 확률과정이다. 따라서, 정확한 예측을 위해서는 EMD의 성질을 잘 반영하는 모형화가 이루어져야 한다.

본 연구에서는 3.1절에서 설명한 그룹화된 IMF 및 residue에 ARIMA 모형과 지수평활법을 조합하여 적용하는 혼합 모형을 고려한다. 특히, 추세가 없다는 측면에서 정상성에 가까운 그룹화된 IMF들에 ARIMA 모형을 적용하고 고차로 적분되어 있는 residue에 지수평활법을 적용하는 혼합 모형이 하나의 예가 될 수 있다. 따라서, 혼합 모형에 의한 예측은 다음과 같이 표현될 수 있다,

$$\hat{x}_{T+h|T} = \widehat{\text{imfs}}^{high}_{T+h|T} + \widehat{\text{imfs}}^{low}_{T+h|T} + \hat{r}_{T+h|T}. \tag{3.5}$$

단, $\hat{x}_{T+h|T}$, $\widehat{\text{imfs}}^{high}_{T+h|T}$, $\widehat{\text{imfs}}^{low}_{T+h|T}$와 $\hat{r}_{T+h|T}$는 예측원점 T-시점에서 각 성분에 대한 $(T+h)$-시점의 예측치를 의미한다.

4. 실증분석: 코스피 예측

4.1. 자료 설명

본 논문에 사용된 자료는 한국 거래소(www.krx.co.kr)에서 제공된 일별 코스피 지수로서 관련 기간은 2010년 7월 1일부터 2015년 9월 30일까지이다. 총 시계열 자료의 길이는 1,054개이며, 모형 적합을 위한 표본내(in-sample) 기간은 2015년 7월 31일(n = 1,014)까지이다. 예측 성능 평가를 위한 표본외(out-of-sample) 기간은 2015년 8월 1일 이후이며 자료의 길이는 40개이다.

Figure 4.1은 표본내 기간의 코스피 지수에 대한 시계열 그림이다. 유럽국가 부채위기에 따른 2011년

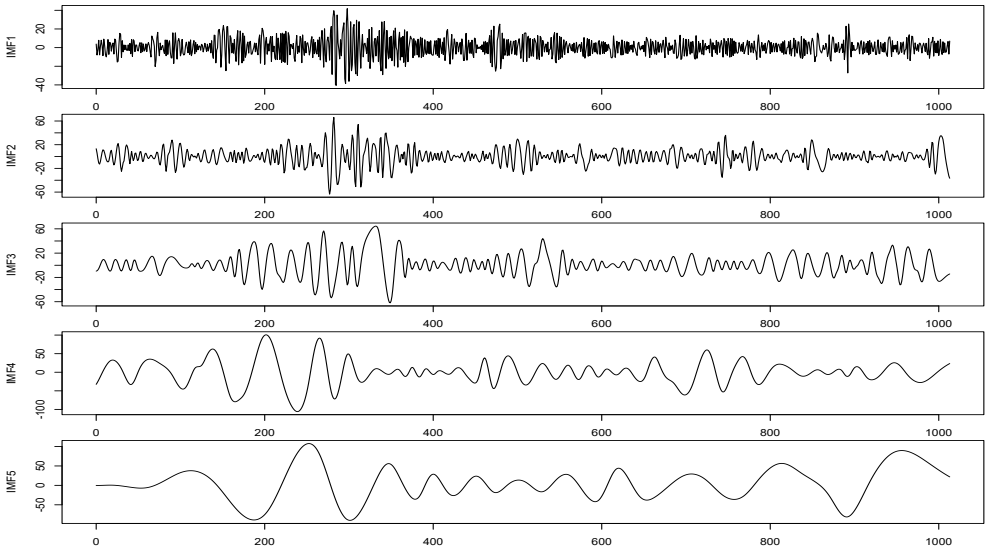

Figure 4.2. IMF1–IMF5 extracted from empirical mode decomposition (EMD).

6월 S&P의 그리스 채권 하향 조정으로 인하여, $t = 276$ 지점인 2011년 8월 초부터 지수는 급격히 떨어지고 있다. $t = 326$ 지점인 2011년 10월 중순부터는 코스피 지수가 조금씩 회복되고 있으며 변동성 또한 안정되어 가고 있는 것으로 보인다. 주기적 계절변동은 없다고 판단되므로 이후 예측 모형은 모두 비계절형 모형을 적용하기로 한다.

4.2. EMD에 의한 분해 및 IMF의 그룹화

표본내 기간의 코스피 시계열을 EMD 방법으로 분해하기 위하여 R 패키지인 EMD를 사용하였다. 패키지가 제공하는 emd( ) 함수 사용시 경계 처리를 위한 옵션으로 'boundary = wave'를 사용하였으며, 최대 IMF의 개수는 10으로 제한하였다.

Figures 4.2와 4.3은 emd( ) 함수로 생성된 7개의 IMF와 residue를 나타낸다. $\text{imf}_t^{(1)}$부터 $\text{imf}_t^{(5)}$까지는 단기적 변동성(volatility)을 반영하고 있는 것으로 고려할 수 있으며, $\text{imf}_t^{(6)}$ 이후부터는 코스피의 중장기적 움직임을 나타내는 것으로 보인다. residue는 코스피 지수의 장기적인 추세를 나타내고 있다.

Table 4.1은 각 IMF들의 해석에 합리성을 부여하기 위하여 원시계열의 분산, 각 IMF들의 분산, 원시계열에 대하여 IMF의 분산이 차지하는 비율과 IMF들의 주기(mean period)를 계산한 것이다. 주기는 전체 시계열 길이(n)를 시계열에 존재하는 극대값의 개수로 나누어 구하였다. IMF의 변동성은 $\text{imf}_t^{(5)}$와 $\text{imf}_t^{(6)}$이 가장 큰 값을 보였으며 특히 $\text{imf}_t^{(6)}$는 원시계열 분산의 59%를 차지하였다. 이러한 IMF들의 움직임은 주기와 관련시킬 때 더 의미있는 해석이 가능하다. 즉, 1년 평균 주식거래일수를 240일로 간주할 때, $\text{imf}_t^{(1)}$–$\text{imf}_t^{(3)}$는 각각 3일, 7일, 16일 정도의 주기를 가지고 있으므로 1주일 이하에서의 KOSPI 움직임을 나타내고 있다고 볼 수 있다. $\text{imf}_t^{(4)}$와 $\text{imf}_t^{(5)}$은 각각 38일과 85일의 주기를 보이고 있으므로 6개월 이하를, $\text{imf}_t^{(6)}$은 약 10개월의 주기적 움직임을 나타내고 있다. $\text{imf}_t^{(7)}$은 약 2년, residue는 그 이상의 장기적인 움직임을 나타내고 있다.

IMF의 해석과 예측에 편리성을 부여하기 위하여 본 연구에서는 3.1절에서 설명된 통합 IMF들인

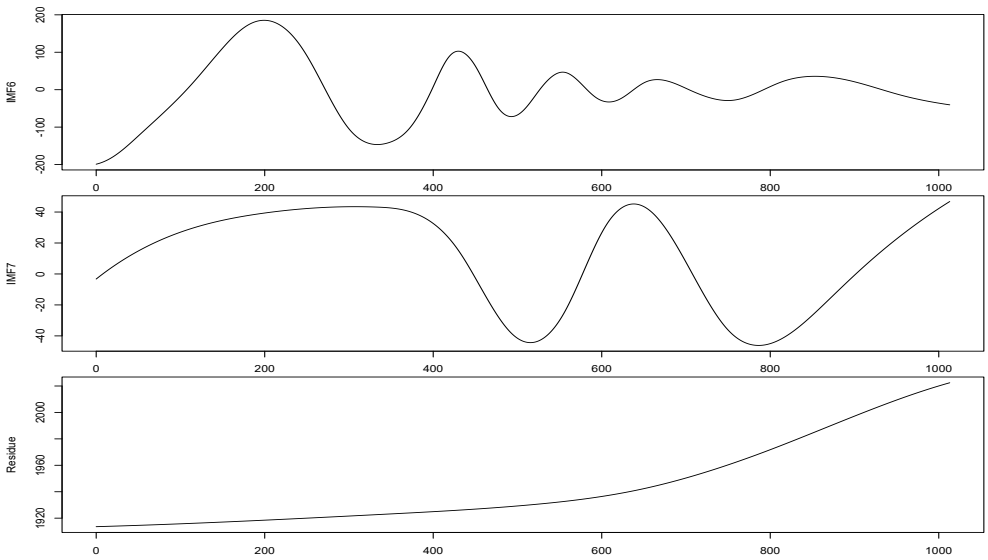

Figure 4.3. IMF6, IMF7 and residue extracted from empirical mode decomposition (EMD).

Table 4.1. Variances and mean periods of IMFs and residue

Components	Variance	Variance as % of the observed	Mean period
Observed	11358.92	1.00	NA[1]
IMF1	103.50	0.01	3.05
IMF2	181.51	0.02	6.85
IMF3	334.31	0.03	15.84
IMF4	1074.35	0.09	37.56
IMF5	1791.76	0.16	84.50
IMF6	6731.09	0.59	202.80
IMF7	929.55	0.08	507.00
Residue	1001.72	0.09	NA

[1]: NA denotes not applicable.

imfs_t^{high}와 imfs_t^{low}를 도입하였으며, 이를 위한 도구로서 Figure 4.4와 같은 누적에너지 E_k의 scree plot을 작성하였다. E_k의 변화량을 살펴볼 때, $m^* = 6$로 고려할 수 있다. 즉, $\text{imfs}_t^{high} = \sum_{i=1}^{5} \text{imf}_t^{(i)}$로 표현할 수 있는 고주파수의 통합 IMF와 $\text{imfs}_t^{low} = \sum_{i=6}^{7} \text{imf}_t^{(i)}$의 저주파수 통합 IMF로 IMF들을 그룹화할 수 있다. Table 4.2는 두 개의 통합 IMF에 대한 분산과 주기를 계산한 것이다. 분산 및 그 비율은 특이한 점이 없으나 주기에서는 흥미로운 결과를 얻을 수 있다. 고주파수 통합 IMF는 주기가 약 4일로서 1주일에 대한 움직임을 나타내는 성분으로 간주할 수 있으며 저주파수 통합 IMF는 주기가 약 203일로서 약 10개월에 대한 움직임을 나타내는 성분으로 파악된다는 점이다.

Figure 4.5는 두 개의 통합 IMF 및 residue의 시계열 그림을 원시계열 그림과 비교한 것이다. 이를 통하여 고주파수 통합 IMF는 코스피의 단기적 변동성을, 저주파수 통합 IMF는 코스피의 중기적으로 발생하는 구조적 추세 변화(structural trend breaks)를 반영하고 있으며, residue는 코스피의 장기적 증가 추세를 잘 반영하고 있다고 판단할 수 있다.

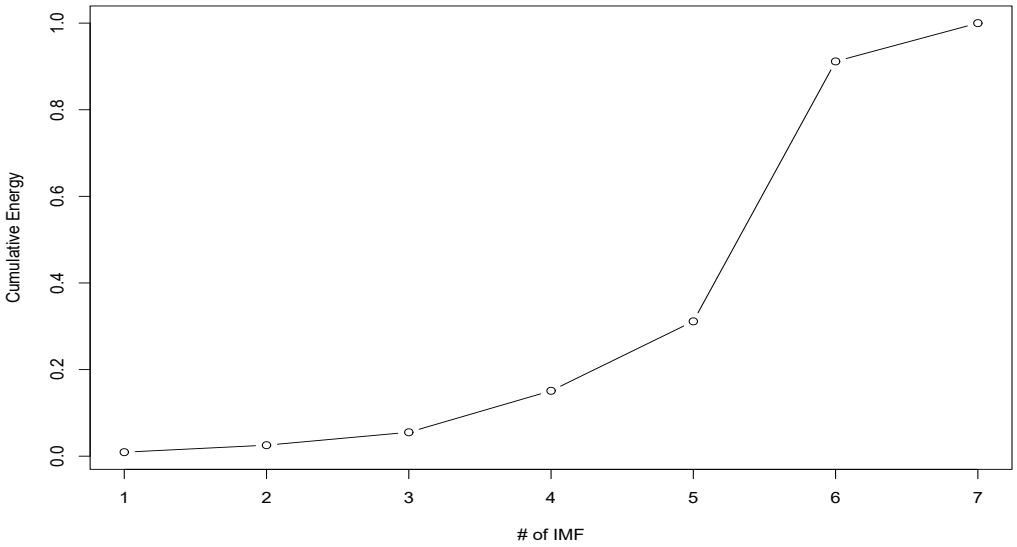

Figure 4.4. Scree plot for cumulative energy of IMFs.

Table 4.2. Variances and mean periods of aggregated IMFs and residue

Components	Variance	Variance as % of the observed	Mean period
Observed	11358.92	1.00	NA[1]
Agg. high freq. IMFs	3228.80	0.28	3.86
Agg. low freq. IMFs	7943.44	0.70	202.80
Residue	1001.72	0.09	NA

[1]: NA denotes not applicable.

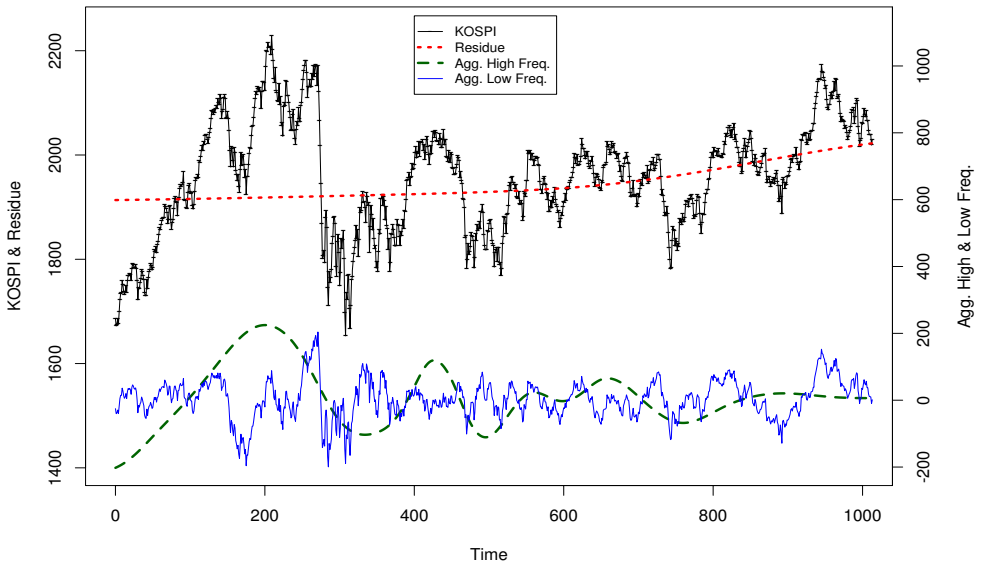

Figure 4.5. Time series plots of original series, aggregated IMFs and residue.

4.3. 혼합 모형의 예측 성능 평가

코스피 자료의 예측을 위해서 본 연구에서 고려되는 혼합 모형은 다음과 같다. 먼저, EMD에 의한 IMF를 ARIMA 또는 ETS 모형과 혼합하는 3가지를 고려하였다.

(1-1) EMD + ARIMA 모형: EMD에 의한 IMF와 residue를 ARIMA 모형화 및 예측한 후, 모든 예측치를 더하여 원시계열의 예측치를 구하는 방법.

(1-2) EMD + ETS 모형: EMD에 의한 IMF와 residue를 ETS 모형화 및 예측한 후, 모든 예측치를 더하여 원시계열의 예측치를 구하는 방법.

(1-3) EMD + ARIMA + ETS 모형: EMD에 의한 IMF와 residue를 각각 ARIMA 및 ETS 모형화하고 각각 예측한 후, 모든 예측치를 더하여 원시계열의 예측치를 구하는 방법.

다음으로는 통합 IMF들을 이용하여 예측하는 다음의 3가지 방법을 고려하였다. 단, 통합된 IMF를 사용하는 EMD 방법을 aggEMD로 표시하기로 한다.

(2-1) aggEMD + ARIMA 모형: EMD에 의한 IMF들을 고주파수와 저주파수의 두 그룹으로 그룹화한 후, 그룹화된 IMF와 residue를 ARIMA 모형화 및 예측한 후, 모든 예측치를 더하여 원시계열의 예측치를 구하는 방법.

(2-2) aggEMD + ETS 모형: EMD에 의한 IMF들을 고주파수와 저주파수의 두 그룹으로 그룹화한 후, 그룹화된 IMF와 residue를 ETS 모형화 및 예측한 후, 모든 예측치를 더하여 원시계열의 예측치를 구하는 방법.

(2-3) aggEMD + ARIMA + ETS 모형: EMD에 의한 IMF들을 고주파수와 저주파수의 두 그룹으로 그룹화한 후, 그룹화된 IMF와 residue를 각각 ARIMA 및 ETS 모형화하고 각각 예측한 후, 모든 예측치를 더하여 원시계열의 예측치를 구하는 방법.

또한, 위에서 설명된 6가지 혼합모형은 원시계열에 ARIMA 및 ETS 모형을 그대로 적용한 결과와도 비교하였다. 각 모형의 모형화 및 예측은 R의 forecast 패키지의 auto.arima( ) 및 ets( ) 함수를 이용하였으며, 미래 예측 시점에 따른 모형의 예측력을 높이기 위하여 ARIMA 및 ETS 모형의 차수 및 모수들은 고정시키지 않고 변화할 수 있도록 하였다.

총 8개의 경쟁모형들에 대한 예측 성능을 비교하기 위하여 표본외 기간의 자료($p = 40$)를 이용하여 h-일 미래 예측에 대한 다음과 같은 평균 예측오차 제곱 평균 를 계산하였다.

$$\mathrm{RMSE}(h) = \sqrt{\sum_{t=n}^{n+p-1} \frac{(x_{t+h} - \hat{x}_{t+h}(t))^2}{p}}, \quad 1 \le h \le 20. \tag{4.1}$$

단, $\hat{x}_{t+h}(t)$는 t시점에서 x_{t+h}의 예측치를 의미한다. 평균 예측오차 절대값(mean absolute forecast error; MAE)을 사용할 수도 있으나 모든 경우에 있어서 RMSE에 의한 결과와 동일한 결과를 보여주었으므로 생략하였다)

Table 4.3 및 Figure 4.6에서 볼 수 있는 것처럼, 거의 모든 시점의 예측에서 aggEMD + ARIMA + ETS 모형이 가장 우수한 성능을 보여주었다; 단, 3–5시점 예측은 제외. 대체로 그룹화된 IMF를 이용한 혼합 모형들이 우수한 예측력을 가지고 있다. aggEMD + ARIMA 모형의 경우, aggEMD + ARIMA + ETS 모형과 유사한 예측력을 보였다. EMD에 의한 IMF를 통합없이 그대로 이용한 혼합 모형의 경우, 지수평활법(ETS)을 사용하는 경우를 제외하고 단기(1–5시점) 예측에서 예측력이 아주 좋지 않았다. 그러나, 예측 시점이 멀어질수록 예측력은 점점 좋아졌다.

Table 4.3. Comparison of out-of-sample RMSE for lead-times from 1 to 20 days

h-step ahead	EMD-based hybrid models							
	ARIMA	ETS	M1-1	M1-2	M1-3	M2-1	M2-2	M2-3
1	22.66	22.55	54.15	28.92	54.14	21.97	22.61	**21.97**
2	38.61	38.30	72.83	44.47	72.81	31.79	38.45	**31.79**
3	50.22	49.64	78.03	55.09	78.01	**40.47**	49.91	40.48
4	59.57	58.68	73.13	62.76	73.08	**53.71**	59.06	53.72
5	66.27	64.97	70.33	68.34	70.26	**62.52**	65.52	62.53
6	70.23	68.43	71.92	71.43	71.84	64.63	69.20	**64.62**
7	74.01	71.68	73.02	74.33	72.90	62.85	72.67	**62.82**
8	76.13	73.36	73.89	75.67	73.74	60.68	74.54	**60.61**
9	79.17	76.08	78.42	79.92	78.23	65.44	77.42	**65.35**
10	83.97	80.56	84.09	86.24	83.86	68.71	82.06	**68.59**
11	90.23	86.56	89.67	92.55	89.41	69.90	88.15	**69.73**
12	96.97	93.14	92.36	97.70	92.05	70.86	94.72	**70.64**
13	100.81	96.95	95.82	100.11	95.48	75.37	98.50	**75.14**
14	101.56	97.71	98.84	101.27	98.48	82.83	99.23	**82.61**
15	100.53	96.94	99.51	100.69	99.14	89.61	98.36	**89.40**
16	95.78	92.34	91.32	96.67	90.91	81.71	93.70	**81.48**
17	92.40	89.18	82.91	93.31	82.47	70.84	90.40	**70.59**
18	92.70	89.49	78.43	93.68	77.94	68.35	90.66	**68.07**
19	92.15	88.85	73.23	92.87	72.67	62.85	90.00	**62.52**
20	92.01	88.53	74.24	92.61	73.64	66.19	89.72	**65.84**
Average	78.80	76.20	80.31	80.43	80.05	63.57	77.24	**63.42**

RMSE = root mean squared forecast error, EMD = empirical mode decomposition, ARIMA = autoregressive integrated moving-average, ETS = exponential smoothing method, M1-1 = EMD + ARIMA, M1-2 = EMD + ETS, M1-3 = EMD + ARIMA + ETS, M2-1 = aggEMD + ARIMA, M2-2 = aggEMD + ETS, M2-3 = aggEMD + ARIMA + ETS.
The underlined bold numbers denote the minimum RMSEs.

aggEMD + ARIMA + ETS 모형이 다른 혼합 모형들에 비하여 우수한 예측력을 나타내고 있는 것은, 두 가지 원인에 의한 것으로 추정된다. 첫째, 그룹화된 IMF가 그룹화되지 않은 IMF들의 복잡성을 감소시켰다고 볼 수 있으며, 둘째, residue를 예측할 때는 ARIMA 모형보다 ETS 모형이 더 유리하다는 점이다.

5. 결론

빅데이터 시대가 도래함에 따라 시계열 자료의 형태 또한 복잡해지고 있으며 전통적인 시계열 모형은 여러가지 제약점을 보이고 있다. 특히, 시계열 자료에 흔히 존재하는 비정상성과 비선형성을 해결하기 위한 모형의 필요성은 시급하다고 하겠다. 또한 최근에는 에너지 수요 예측 및 교통량 예측 분야에서 고주파수 시계열 자료 분석을 위한 모형도 그 중요성이 날로 높아지고 있다. 본 논문에서는 이러한 맥락에서 복잡한 시계열 자료의 해석의 편리성과 예측 가능성을 높이기 위하여 EMD를 활용한 혼합 모형을 연구하였다. EMD에 의한 IMF의 해석을 더욱 편리하게 만들기 위하여 IMF들을 누적에너지 개념을 사용하여 그룹화하였고, 이것을 이용한 ARIMA 및 지수평활법의 혼합 모형은 예측의 성능을 더욱 높여 주었다. 향후, IMF 그룹화와 혼합 모형의 정교성을 좀더 높일 수 있다면 다양한 분야의 복잡한 시계열 자

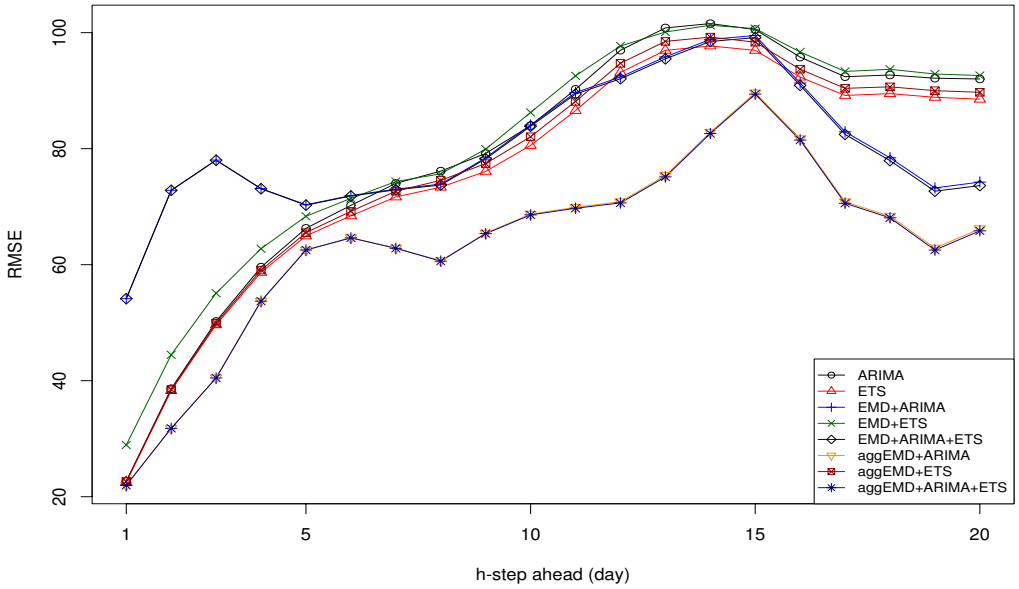

Figure 4.6. RMSE comparison in out-of-sample forecasts (RMSE = root mean squared forecast error, ARIMA = autoregressive integrated moving-average, ETS = exponential smoothing method, EMD = empirical mode decomposition).

료의 분석 및 예측에 두루 사용할 수 있는 잠재성 있는 모형으로 평가한다.

References

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1993). *Time Series Analysis: Forecasting and Control*, Prentice Hall, New Jersey.

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*, McGraw-Hill, New York.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing, *Journal of the American Statistical Association*, **106**, 1513–1527.

Gould, P. G., Koehler, A. B., Ord, J. K., Snyder, R. D., Hyndman, R. J., and Vahid-Araghi, F. (2008). Forecasting time series with multiple seasonal patterns, *European Journal of Operational Research*, **191**, 207–222.

Holt, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving average, *Office of Naval Research*, Research Memorandum, **52**, Carnegie Institute of Technology.

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 analysis, In *Proceeding of the Royal Society London A*, **454**, 903–995.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R, *Journal of Statistical Software*, **26**, 1–22.

Kim, D. and Oh, H.-S. (2009). A multi-resolution approach to non-stationary financial time series using the Hilbert-Huang transform, *The Korean Journal of Applied Statistics*, **22**, 499–513.

Kim, D., Paek, S.-H., and Oh, H.-S. (2008). A Hilbert-Huang transform approach for predicting cyber-attacks, *Journal of the Korean Statistical Society*, **27**, 277–283.

Park, M. and Seong, B. (2014). Comparison of EMD and HP filter for cycle extraction, *The Korean Journal of Applied Statistics*, **27**, 431–444.

Wei, W. W. (2006). *Time Series Analysis*, 2nd ed., Addison-Wesley, Redwood City, California.

Wei, Y. and Chen, M.-C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks, *Transportation Research Part C*, **21**, 148–162.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages, *Management Science*, **6**, 324–342.

Zhu, B., Wang, P., Chevallier, J., and Wei, Y. (2015). Carbon price analysis using empirical mode decomposition, *Computational Economics*, **45**, 195–206.

코스피 예측을 위한 EMD를 이용한 혼합 모형

김효원[a] · 성병찬[a,1]

[a]중앙대학교 응용통계학과

(2016년 2월 26일 접수, 2016년 2월 29일 수정, 2016년 2월 29일 채택)

요 약

본 연구에서는 시계열 자료의 비정상성과 비선형성과 같은 복잡성을 효과적으로 포용할 수 있는 경험적모드분해법(empirical mode decomposition; EMD)을 토대로 시계열 자료의 분석 및 예측을 위한 혼합(hybrid) 모형을 연구한다. EMD에 의하여 생성되는 내재모드함수(intrinsic mode function; IMF)는 해석 및 예측의 편리성을 개선하기 위하여 누적에너지의 개념을 사용하여 그룹화하였으며, 그룹화된 IMF 및 residue의 성분들은 그 성질에 따라서 ARIMA 모형 및 지수평활법과 결합된 혼합 모형으로 예측된다. 제안된 방법은 일별 코스피 지수의 예측을 위해서 적용하였다. 다양한 형태의 혼합 모형을 사용하여 코스피 지수를 예측하였으며 전통적인 예측 방법과 비교하였다. 분석 결과, 그룹화된 성분들은 코스피 지수의 움직임을 단기적, 중기적, 장기적으로 해석하는데 편리함을 주었으며, 그룹화된 IMF 및 residue를 각각 ARIMA 모형과 지수평활법으로 조합한 혼합 모형이 우수한 예측력을 보여주었다.

주요용어: 내재적모드함수, 지수평활법, 자기회귀적분이동평균 모형, 비정상 모형, 비선형 모형

이 논문은 2014년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (06974) 서울 동작구 흑석동 221, 중앙대학교 응용통계학과. E-mail: bcseong@cau.ac.kr