The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.2.271
30(2), 271–280

Forecasting Korean housing price index: application of the independent component analysis

Ro Jin Pak[a,1]

[a]Dankook University, Department of Applied Statistics

(Received February 3, 2017; Revised February 28, 2017; Accepted March 6, 2017)

Abstract

Real-estate values and related economics are often the first read newspaper category. We are concerned about the opinions of experts on the forecast for real estate prices. The Box-Jenkins ARIMA model is a commonly used statistical method to predict housing prices. In this article, we tried to predict housing prices by combining independent component analysis (ICA) in multivariate data analysis and the Box-Jenkins ARIMA model. The two independent components for both the selling price index and the long-term rental price index were extracted and used to predict the future values of both indices. In conclusion, it has been shown that the actual indices and the forecast indices using ICA are more comparable to the forecasts of the ARIMA model alone.

Keywords: ARIMA, Box-Jenkins model, ICA, forecasting, real estate price

1. 서론

현재 부동산 가격 지수로 가장 많이 사용하는 것은 국민은행에서 매달 발표하는 'KB주택가격동향'자료이다. KB주택가격동향에 포함된 부동산의 매매가격(종합)지수와 전세가격(종합)지수의 관계를 설명함에 있어서 고려해야 할 사항은 이 지수들이 특정 시점의 가격 수준을 100으로 놓고 매 월 지수를 상대적으로 구하고 있다는 것이다. 2017년 1월 현재의 지수는 2015년 12월의 매매지수와 전세지수가 실체적 가격 수준에 상관없이 모두 100으로 조정된 상태에서 계산되어 있다. 우리나라의 경우 전세지수가 매매지수에 후행하면서 따라 움직이는 것으로 알려져 있고 특정 시점에서 억지로 전세지수를 매매지수에 맞추려 하다 보니 전세지수 상승세가 가파르게 나타난다. 이러한 상황을 무시하고 무작정 2015년 12월 이후의 지수를 예측한다는 것은 무리가 있다. 앞서 언급한대로 전세지수는 2015년 12월을 앞두고 매우 급하게 상승하며 따라서 그 이후를 예측하게 되면 상승기조가 계속 유지될 수밖에 없다.

그런데, 신호처리와 같은 공학 분야에서는 다양한 근원에서 발생한 소음이 혼합되어 측정된 상태에서 본래의 근원을 찾아내려는 시도가 있는데 이를 독립성분분석(independent component analysis; ICA)이라한다. 독립성분분석은 다차원 분석도구로서 통계학에서의 주성분분석, 요인분석과는 다소 다른 개념이다. 본 논문에서는 독립성분분석의 개념을 빌려서 매매지수와 전세지수의 독립성분을 구하고 이들로부터 예측을 수행하는 방법으로 지수들을 예측해보려 하였다.

[1]Department of Applied Statistics, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea. E-mail: rjpak@dankook.ac.kr

독립성분분석에 대한 자세한 응용은 Hyvärinen (1998a, 1998b)에서 확인할 수 있다. 국내에서는 주로 공학 (Cho, 2004; Hwang과 Kim, 2011; Jeon 등, 2006) 그리고 의학 (Shim 등, 2001)에서 활용된 연구들이 있으며 해외에서는 증권 관련하여 Back과 Weigend (1997)의 연구가 있고 산업생산지수와 같은 시계열 자료에 대한 연구가 최근 Garcia-Ferrer 등 (2011)에 의해 이루어졌다.

부동산 매매지수와 전세지수의 독립성분을 구하고 이들로부터 예측을 수행한 후 지수들로 역변환 하는 방법을 사용하여 2016년도 예측을 시도하였다. 결과적으로는 적어도 기존의 자료에 대하여는 독립성분을 이용한 예측이 일반적인 ARIMA 예측보다 정확함을 확인할 수 있었다. 현재 제공되고 있는 지수 데이터가 갖고 있는 구조적 문제를 극복할 수 있는 한 가지 방법이라고 사려 된다.

2. 독립성분분석

시간(t)에 따라 관측된 두 세트의 시계열 자료 $x_1(t)$와 $x_2(t)$가 주어졌다고 하자. 그런데 $x_1(t)$와 $x_2(t)$에는 내재되어 있는 어떤 신호들 $s_1(t)$와 $s_2(t)$이 존재하여 어떤 상수들 $(a_{11}, a_{12}, a_{21}, a_{22})$과 아래와 같이

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t), \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned} \tag{2.1}$$

선형관계로 엮어 있다고 하자. 만일 $s_1(t)$와 $s_2(t)$를 성공적으로 추출해 낼 수 있다면 $x_1(t)$와 $x_2(t)$를 이들 성분을 이용하여 분석할 수 있을지도 모른다는 생각을 할 수 있겠다.

위의 생각을 확장한 독립성분분석에 대하여 간단히 정리하여 보겠다. 통계적 잠재변수 모형에 기인한 독립성분분석은 Jutten과 Hérault (1991) 그리고 Comon (1994)에 의해 정의되었다. 관찰 가능한 n개의 변수 $x_1, \ldots, x_n$를 독립성분 $s_1, \ldots, s_n$으로

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n, \quad j = 1, 2, \ldots, n \tag{2.2}$$

와 같이 선형식으로 만들 수 있다면 이를 독립성분분석 모형이라고 한다. 이를 행렬 형태로 표시하면

$$\boldsymbol{x} = \boldsymbol{A}s, \tag{2.3}$$

$\boldsymbol{x} = [x_1, \ldots, x_n]^T$, $\boldsymbol{s} = [s_1, \ldots, s_n]^T$ 그리고 $\boldsymbol{A} = [(a_{ij})]_{n\times n}$로 표현할 수 있다. 만일 성분 s_j들이 독립이라는 단순한 가정만으로 $\boldsymbol{A}$를 성공적으로 추정할 수 있다면, $\boldsymbol{A}$의 역행렬을 구하여 독립성분

$$\boldsymbol{s} = \boldsymbol{A}^{-1}\boldsymbol{x} \tag{2.4}$$

을 찾아낼 수 있게 된다.

2.1. 독립성분분석의 원리

식 (2.1)을 기하학적으로 보면 x_1과 x_2가 2차원 상에서 s_1과 s_2라는 축을 회전 변환한 결과라고 할 수 있다. 그리고 $\boldsymbol{A}$의 역행렬이 존재한다면 거꾸로 s_1과 s_2는 x_1과 x_2의 회전변환의 결과라고 하겠다. s_1과 s_2는 독립성을 유지하고 싶은 상황에서 정규분포 가정이 없어도 중심극한정리에 의해 식 (2.1)을 만족하는 x_1과 x_2는 독립여부와 관계없이 근사적으로 정규성을 보장 받을 수 있다. 그런데 만일 s_1과 s_2가 정규성을 갖고 있다면 s_1과 s_2의 독립성 여부가 보장될 필요가 없다. 따라서 실제로는 x_1과 x_2가 관측된 상황에서 s_1과 s_2의 독립성을 보장받기 위해 비정규성을 극대화 시킬 필요가 있다는 것이다.

이제, $\boldsymbol{x}$가 관측되었고 가중치 벡터 $\boldsymbol{w}$가 있어서 적절한 선형결합 $y = \boldsymbol{w}^T\boldsymbol{x} = \sum_i w_i x_i$가 가능하다고 하자. 식 (2.3)에서 정의된 $\boldsymbol{A}$를 사용하여 새로운 변수 $\mathbf{z} = \boldsymbol{A}^T\boldsymbol{w}$를 정의하면 $y = \boldsymbol{w}^T\boldsymbol{A}s = \mathbf{z}^T\boldsymbol{s} = \sum_i z_i s_i$가 된다. s_i들의 확률분포에 상관없이 독립인 s_i들의 선형결합인 y는 중심극한정리에 근거하여 정규분포에 근접할 것이고 이는 우리가 바라지 않는 상황이다. 따라서 가능하다면 z_i들 중에 많은 것이 0이 되었으면 좋겠는데 그래도 적어도 단 하나만은 0이 아니 길 원한다. 그렇게 된다면 $\mathbf{z}^T\boldsymbol{s} = \boldsymbol{w}^T\boldsymbol{x}$ 자체로 부터 어떤 식으로든 s_i들 중 하나를 구하게 될 것이다. 이렇듯 $\boldsymbol{w}^T\boldsymbol{x}$의 비정규성을 극대화하는 과정에서 독립성분을 찾고자 한다.

아래에 비정규성을 극대화하는 몇 가지 방법을 간단히 소개하고 실제적으로 계산상 가장 많이 사용되는 FastICA 알고리즘을 기술하겠다.

1) 첨도(kurtosis)가 0이 되지 않도록 한다 (Delfosse와 Loubaton, 1995).
   첨도는 정규분포의 경우 0이 되며 따라서 첨도가 음이나 양이 되는 방향으로 독립성분을 찾는다.

2) 네그엔트로피(negentrophy)를 최대화하도록 한다 (Cover와 Thomas, 1991; Papoulis, 1991).
   엔트로피(H)는 확률 변수 Y가 이산형인 경우,

$$H(Y) = -\sum P(Y = y_i)\log P(Y = y_i)$$

   로 연속형인 경우, 임의의 확률 변수 벡터 $\boldsymbol{y}$에 대하여

$$H(\boldsymbol{y}) = \int f(\boldsymbol{y})\log f(\boldsymbol{y})d\boldsymbol{y}$$

   로 정의된다. 그런데, 엔트로피는 정규분포(가우시안) 확률변수의 경우 다른 확률변수들에 비해 큰 값을 갖는다. 그래서

$$J(\boldsymbol{y}) = H(\boldsymbol{y}_{\text{gaussian}}) - H(\boldsymbol{y})$$

   을 네그엔트로피라고 정의하고 $J(\boldsymbol{y})$가 0보다 커질수록 정규분포 변수가 아님을 확신하게 된다. 하지만 네그엔트로피를 정의대로 구하는 것이 사실상 어려워 다양한 근사식을 사용한다 (Hyväriene, 1998a, b; Jones와 Sibson, 1987).

3) 상호정보량(mutual information)을 최소로 한다 (Cover와 Thomas, 1991; Papoulis, 1991).
   확률 변수 $\{y_i; i = 1, \ldots, m\}$에 대하여 상호정보량을

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\boldsymbol{y})$$

   로 정의한다. 벡터 $\boldsymbol{y} = \boldsymbol{W}x$, $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$라고 하면

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\boldsymbol{x}) - \log|\det \boldsymbol{W}|$$

   가 되고 몇몇 과정을 거쳐

$$I(y_1, y_2, \ldots, y_m) = \text{constant} - \sum_i J(y_i)$$

   가 된다. 상호정보량을 최소로 하는 가중치 $\boldsymbol{W}$를 구하는데 이는 네그엔트로피를 최대화하는 것과 같다.

4) 최대우도법(maximum likelihood method) (Pham 등, 1992).

s_i에 대한 확률함수 f_i를 알고 있고 관찰이 $\boldsymbol{x}(t)$, $t = 1, \ldots, T$같이 T번 이루어진다면, 로그우도함수를 구하여

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i \left(\boldsymbol{w}_i^T \boldsymbol{x}(t) \right) + T \log |\det \boldsymbol{W}|, \quad \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$$

를 최대로 하는 $\boldsymbol{W}$를 추정한다.

2.2. FastICA 알고리즘

앞 절에서 언급한 추정 방법들은 관련된 확률함수를 모르면 사실상 사용하기 어려운 점이 있다. 그래서 실제로는 근사식을 이용하여 추정하는 방법을 사용하며 대표적인 방법인 FastICA 알고리즘에 대해 간단히 설명하고자 한다. 알고리즘의 기본 대상이 되는 $\boldsymbol{w}^T\boldsymbol{x}$에서 네그엔트로피를 최대로 만드는 $\boldsymbol{w}$를 찾아내려 한다. Hyvärinen (1998b)는 네그엔트로피가 적당한 비이차형(nonquadratic) 함수 G와 표준정규확률 변수 ν에 대하여

$$J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2 \tag{2.5}$$

가됨을 보였다. $J(y) = J(\boldsymbol{w}^T\boldsymbol{x})$를 $\boldsymbol{w}$에 대해 최대화한다는 것은 $E\{G(\boldsymbol{w}^T\boldsymbol{x})\}$를 $\boldsymbol{w}$에 대해 최대화한다는 것과 같다. 다만 $\boldsymbol{w}$가 가중치로서의 역할을 하기위해 $||\boldsymbol{w}||^2 = 1$로 제한하기로 하였다. 그래서 라그랑주 방법론에 따라 $E\{G(\boldsymbol{w}^T\boldsymbol{x})\}$-$\lambda(||\boldsymbol{w}||^2 - 1)$를 $\boldsymbol{w}$에 대해 미분하여 0으로 놓은 $E\{\boldsymbol{x}g(\boldsymbol{w}^T\boldsymbol{x})\}$-$\lambda\boldsymbol{w} = 0$를 뉴튼-랩슨의 방법을 이용하여 수치 해석적으로 아래의 방정식의 근 $\boldsymbol{w}^+$를 구한다. 여기서, g는 G의 도함수를 의미한다. 즉,

$$\boldsymbol{w}^+ = \boldsymbol{w} - \frac{E\left\{\boldsymbol{x}g(\boldsymbol{w}^T\boldsymbol{x})\right\} - \lambda\boldsymbol{w}}{E\left\{g'(\boldsymbol{w}^T\boldsymbol{x})\right\} - \lambda}$$

이다. 함수 G는 $G(\nu) = -\exp(\nu^2/2)$와 $G(\nu) = (1/\alpha)\log\cosh(\alpha\nu)$, $1 \le \alpha \le 2$가 적당하다고 되어있다. 실제적인 계산은 R에서 제공하는 fastICA 패키지를 사용하였는데 R에서는 $\alpha = 1$로 하는 logcosh함수가 디폴트형태로 주어진다.

위의 과정을 Hyvärinen (1999a), Hyvärinen과 Oja (1997)는 다음과 같이 알고리즘화 하고 FastICA라고 불렀다.

1. $\boldsymbol{w}$의 초기치를 (무작위로) 정한다.
2. 출력값 $\boldsymbol{w}^+ = E\{\boldsymbol{x}g(\boldsymbol{w}^T\boldsymbol{x})\} - E\{g'(\boldsymbol{w}^T\boldsymbol{x})\}\boldsymbol{w}$로 한다.
3. 입력값 $\boldsymbol{w} = \boldsymbol{w}^T/||\boldsymbol{w}||$를 표준화 한다.
4. 출력값이 원하는 수준에서 수렴할 때까지 단계 2, 3을 반복한다.

3. 자료 분석

국민은행에서 제공하는 전국 주택의 매매 및 전세가격 지수 자료(1989년 1월부터 2015년 12월까지)를 사용하여 2016년도 일 년간의 가격 지수의 예측을 시도하여 보았다. 국민은행 자료는 전국 아파트/주택을 대상으로 층화 2단 집락 확률비례추출법을 통해 추출된 표본들의 가격 지수를 매달 1번씩 조사한 내용을 담고 있다. 가격 지수는 표본들의 특성에 따른 가중치를 이용한 가격의 가중평균을 구하여 2015년

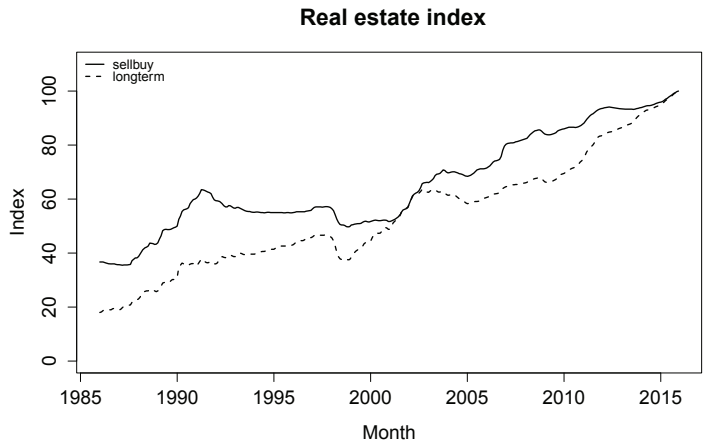

Figure 3.1. Indices for real estate.

12월을 100으로 하여 상대적으로 계산된다. 따라서 2015년 12월의 매매(가격)지수와 전세(가격)지수는 모두 100이 되고 그것은 매매가와 전세가가 같다는 의미는 아닌 것이다. 결국 구조적으로 매매지수와 전세지수는 매우 상관관계가 높을 수밖에 없다. 이러한 사실을 간과하고 지수들을 이용하여 시계열 예측 모형을 추정한다면 여러모로 왜곡된 결과를 초래할 수밖에 없다. Figure 3.1의 지수 그래프에서 보듯이 두 개의 지수는 지속적인 상승추세를 보이며 2015년 12월은 일치하게 된다. 무엇보다 전세 지수가 최종적으로 100이 되기 위해 기울기가 급격하게 될 수밖에 없다.

3.1. 박스-젠킨스의 모델

먼저, 매매지수(y_t)와 전세지수(x_t)에 대한 박스-젠킨스 모형을 추정하고 예측을 수행하여 보았다. 지면상 자세히 서술할 수 없었으나 실제로 다양한 차수(order)의 모형이 가능하였으며 아래에는 나름 합리적이라고 판단된 모형을 제시하였다.

- 매매지수: ARIMA$(3,1,3)(0,1,1)_{12}$

$$\left(1-1.66B+1.45B^2-0.64B^3\right)\left(1-B^{12}\right)(1-B)\,y_t = \left(1-0.90B+0.81B^2-0.22B^3\right)\left(1-0.91B^{12}\right)\epsilon_t.$$

- 전세지수: ARIMA$(3,1,3)(0,1,1)_{12}$

$$\left(1-1.80B-1.74B^2-0.77B^3\right)\left(1-B^{12}\right)(1-B)\,y_t = \left(1-1.09B-0.99B^2-0.18B^3\right)\left(1-0.93B^{12}\right)\epsilon_t.$$

3.2. ICA의 활용

ICA를 이용하여 아래와 같은 절차를 따라 예측을 수행하였다.

1. 원 시계열을 필요하다면 차분과 변환을 통해 정상시계열로 만든다.
2. 단계 1에서 구한 시계열에서 독립성분 추정치를 FastICA 알고리즘으로 구한다.
3. 단계 2에서 구한 독립성분 추정치들에 박스-젠킨스의 ARIMA모형을 추정한다.
4. 단계 3에서 추정한 모형으로 예측을 수행한다.

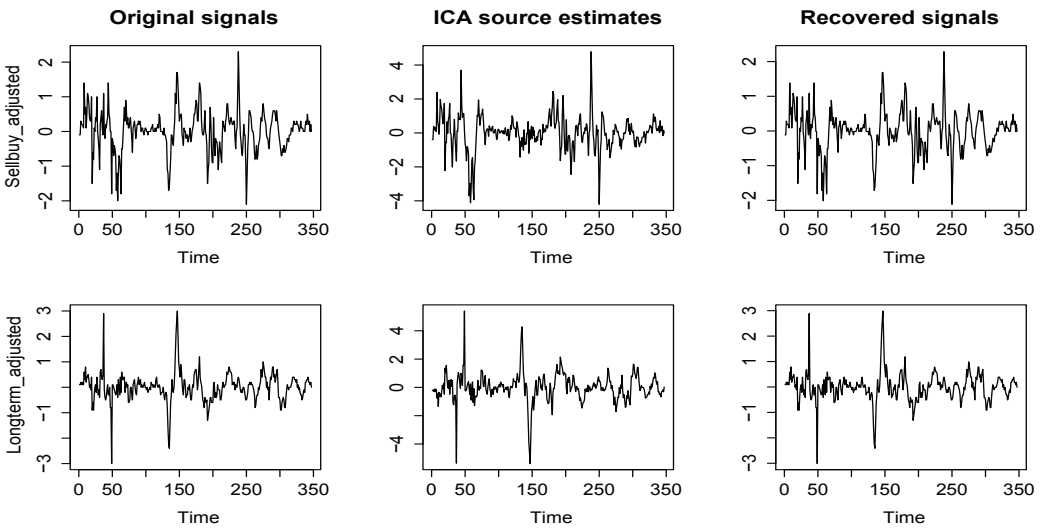

Figure 3.2. Differenced series, ICA processed estimate and recovered series. ICA = independent component analysis.

5. 단계 4에서 예측한 값에 FastICA 알고리즘을 역으로 적용하고 역차분과 역변환을 통해 원 시계열에 대한 예측치로 삼는다.

실제적인 계산은 R에서 제공하는 FastICA 패키지를 사용하였다. FastICA는 가중치 행렬 $\boldsymbol{W}$를 구하되 먼저 데이터 행렬($\boldsymbol{X}$)를 사전백색화(pre-whitening)행렬 $\boldsymbol{K}$를 이용하여 일종의 표준화를 수행하고 성분 행렬($\boldsymbol{S}$)과의 관계 $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{K}\boldsymbol{W}$가 성립되도록 수행한다. 즉, $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{K}\boldsymbol{W} \Leftrightarrow \boldsymbol{X} = \boldsymbol{S}(\boldsymbol{K}\boldsymbol{W})^{-1}$를 만족하는 $\boldsymbol{W}$를 구해준다.

1. 먼저, 매매지수와 전세지수에서 추세와 계절성을 제거한 뒤 ICA를 수행하고 ICA가 합당한 지 복원한 결과를 Figure 3.2의 왼쪽부터 오른쪽으로 추출된 독립성분과 독립성분을 이용하여 다시 역변환하여 복원된 시계열 그림을 연속하여 그려 보았다. 원 시계열과 복원된 시계열이 동일함을 확인할 수 있다. 이 과정에서 사용된 $\boldsymbol{K}$와 $\boldsymbol{W}$는 다음과 같다.

$$\boldsymbol{K} = \begin{pmatrix} -0.970 & -2.147 \\ -0.924 & 2.259 \end{pmatrix}, \qquad \boldsymbol{W} = \begin{pmatrix} -0.517 & 0.855 \\ -0.855 & -0.517 \end{pmatrix}.$$

2. FastICA를 통해 추출된 독립성분 $s_1(t)$과 $s_2(t)$에 대하여 아래와 같이 각각 ARMA$(3,0,3)$과 ARMA$(3,0,2)$를 적합할 수 있었다. Tables 3.1과 3.2에 관련된 통계수치들을 적어 놓았다.

$$s_1(t) = \frac{1 + 0.7829B - 0.4044B^2 + 0.2034B^3}{1 - 1.5240B + 0.9901B^2 - 0.3944B^3}\epsilon(t),$$
$$s_2(t) = \frac{1 - 0.3684B^1 - 0.9999B^2}{1 - 0.3987B + 0.6803B^2 - 0.6934B^3}\epsilon(t).$$

3. 2에서 구한 독립성분들 $s_1(t)$와 $s_2(t)$의 시계열 모형으로 부터 2016년도 1월부터 12월까지의 예측을 수행하고 $\boldsymbol{X} = \boldsymbol{S}(\boldsymbol{K}\boldsymbol{W})^{-1}$를 이용하여 역변환한 후 역차분한 결과를 Figure 3.3에 그려 보았다. ICA에 의한 결과는 2016년도 전반기에는 실제치와 매우 유사하게 나타나고 있다. 앞서 일반적인 ARIMA에 의한 예측결과는 실제치를 상회하고 있다.

Table 3.1. Statistics for $s_1(t)$

Parameter	ar1	ar2	ar3	ma1	ma2	ma3
Estimate	1.5240	−0.9901	0.3944	−0.7829	0.4044	−0.2034
t-test	32.1550	−18.1341	14.1697	−13.0433	6.7236	−2.0646
p-value	<.0001	<.0001	<.0001	<.0001	<.0001	0.0389
Ljung-Box			$\chi^2 = 26.241$	p-value = 0.158		

Table 3.2. Statistics for $s_2(t)$

Parameter	ar1	ar2	ar3	ma1	ma2
Estimate	0.3987	−0.6803	0.6934	0.3684	0.9999
t-test	9.7618	−23.1092	17.6584	40.6517	66.2786
p-value	<.0001	<.0001	<.0001	<.0001	<.0001
Ljung-Box			$\chi^2 = 24.814$	p-value = 0.208	

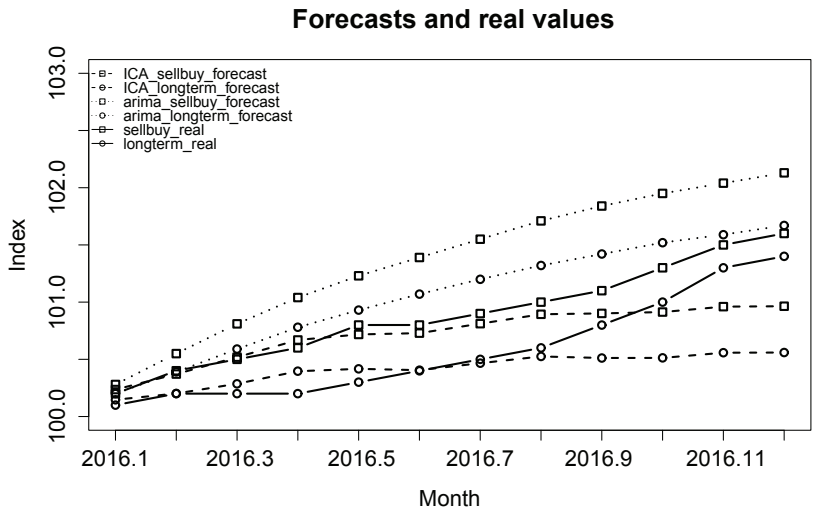

Figure 3.3. Forecasts of indices for real estate.

추가적으로 매매지수($y(t)$)와 전세지수($x(t)$)간의 전이함수모형(transfer function model)을 적합시켜 Table 3.3에 정리하여 보았다. 매매지수는 전세지수를 3개월 선행하는 것으로 보이며 구체적인 모형은

$$\left(1-B^{12}\right)(1-B)y(t) = \frac{-0.0885}{1-0.8889B}B^3\left(1-B^{12}\right)(1-B)x(t) + \frac{1-1.9064B+1.8382B^2-0.8389B^3}{(1-1.2156B+1.0625B^2-0.2374B^3)(1-0.9055B^{12})}\epsilon(t)$$

와 같이 구해졌다.

위의 전이함수모델을 이용하여 3개월 치(2016년1월–3월)를 예측하니 {100.37, 100.86, 101.86}으로 실제값 {100.1, 100.2, 100.2}와 앞서 수행한 다른 예측보다도 매우 높은 값이 나왔다. 이는 매매지수와 전세지수를 2015년 12월에 100으로 억지로 맞추는 과정에서 3개월 후행하는 전세지수가 매매지수에 비해 기울기가 급격히 상승하고 결국 예측할 때 상승기조가 유지되면서 실제값보다 큰 예측값이 계산되었다고 보인다.

Table 3.3. Statistics for a transfer function model of real estate indices

Parameter	Estimate	t-test	p-value
ma1	1.2156	14.68	<.0001
ma2	−1.0625	−12.01	<.0001
ma3	0.2374	3.23	0.0014
sma1	0.9055	34.80	<.0001
ar1	1.9064	31.35	<.0001
ar2	−1.8382	−24.20	<.0001
ar3	0.8389	16.11	<.0001
num1	−0.0885	−2.16	0.0317
den1	0.8889	12.65	<.0001
Ljung-Box		$\chi^2 = 26.000$	p-value = 0.3008

결국 부동산 지수의 구조적 특성상 일반적인 박스-젠킨스의 모형을 직접 적합하기는 무리가 있다고 하겠다. 독립성분분석과 같은 사전 변환을 통한 예측이 유효해 보인다.

4. 맺음말

현재 널리 사용되는 부동산 가격지수는 특정시점에 100이라는 수치를 갖도록 조정되어 데이터 구조상 일반적인 ARIMA모형을 적용하기에 문제가 있다. 독립성분분석이라는 기법과 ARIMA를 동시에 활용하여 예측을 하는 경우 바람직한 결과를 얻을 수 있음을 보였다. 그런데 후반기에는 ARIMA 예측치가 실제치에 더 가까운 모습을 보이는 데 본 논문이 특정한 데이터를 분석한 결과라는 점 ICA를 활용한 방법의 우수성을 이론적으로 밝히지 못한 한계점을 갖고 있다. 본 논문에서 사용한 자료의 인과적 특성상 보다 정교한 예측을 위해 벡터자기회귀이동평균모형(VARMA) 또는 벡터오차수정모형(VECM)이 더 유용할 수도 있겠다. 통계적으로 널리 사용되는 ARIMA와 공학적으로 유용한 ICA를 융합했다는 것이 부동산 지수 예측에 있어서는 의미가 있다고 하겠다.

References

Back, A. and Weigend, A. (1997). A first application of independent component analysis to extracting structure from stock returns, *International Journal of Neural Systems*, **8**, 473–484.

Cho, Y. (2004). Independent component analysis for clustering analysis components by using kurtosis, *The KIPS Transactions: Part B*, **4**, 429–436.

Comon, P. (1994). Independent component analysis - a new concept?, *Signal Processing*, **36**, 287–314.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*, John Wiley & Sons, Hoboken.

Delfosse, N. and Loubaton, P. (1995). Adaptive blind separation of independent sources: a deflation approach, *Signal Processing*, **45**, 59–83.

Garcia-Ferrer, A., Gonzalez-Prieto, E., and Pena, D. (2011). *Exploring ICA for time series decomposition*, In *Working paper 11-16*, Statistics and Econometrics Series 11, Universidad Carlos III de Madrid: Getafe.

Hyvärinen, A. (1998a). Independent component analysis in the presence of gaussian noise by maximizing joint likelihood, *Neurocomputing*, **22**, 49–67.

Hyvärinen, A. (1998b). *New Approximation of Differential Entropy for Independent Component Analysis and Projection Pursuit, in Advances in Neural Information Processing Systems*, MIT Press, Cambridge.

Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis, *Neural Computation*, **9**, 1483–1492.

Hwang, J. S. and Kim, J. H. (2011). A study on the estimation of modal parameters using independent component analysis method, *Journal of the Architectural Institute of Korea Structure & Construction*,

27, 27–35.

Jeon, C. H., Lee, H. S., Park, H. S., and Hong, J. H. (2006). Estimation of pure component fractions in a mixture using independent component analysis. In *Proceedings of the Korean Operations and Management Science Society Conference*, 753–757.

Jones, M. and Sibson, R. (1987). What is projection pursuit?, *Journal of the Royal Statistical Society Series 4*, **150**, 1–36.

Jutten, C. and Hérault, J. (1991). Blind separation of source, part I: an adaptive algorithm based on neuromimetic architecture, *Signal Processing*, **24**, 1–10.

Papoulis, A. (1991). *Probability, Random Variables and Stochastic Processes* (3rd Ed), McGraw-Hill, New York.

Pham, D. T., Garrat, P., and Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. In *Proceedings of EUSIPCO*, 771–774.

Shim, Y. S., Choi, S. H., and Lee, I. K. (2001). Eyeball movements removal in EEG by independent component analysis, *Korean Journal of Clinical Neurophysiology*, **3**, 26–30.

부동산 매매지수와 전세지수 예측: 독립성분분석을 활용한 분석

박노진[a,1]

[a]단국대학교 응용통계학과

요 약

우리나라 뉴스에서 매일 빠지지 않는 내용은 아마도 부동산 경제에 관한 것이라고 생각된다. 많은 사람들은 부동산 가격의 변동에 관한 전문가들의 예측에 관심을 갖고 있다. 매매가격 혹은 전세가격을 예측하기위해 일반적으로 많이 사용되는 방법은 박스-젠킨스에 기반을 둔 자기회귀이동평균모형이다. 본 논문에서는 자기회귀모형과 다변량 자료분석에서 사용하는 독립성분분석을 결합하여 예측하는 방법을 시도하여 보았다. 매매가격과 전세가격을 두 개의 독립성분으로 재설정하고 독립성분들을 이용하여 예측한 후 역변환을 통해 매매가격과 전세가격을 예측하는 방법을 시도하였다. 그 결과 일반적인 자기회귀이동평균모형을 사용할 때 보다 독립성분을 활용한 예측이 실제 지수에 더 유사한 값들을 얻을 수 있음을 보였다.

주요용어: 독립성분분석, 박스-젠킨스 모형, 부동산 가격지수, 예측, 자기회귀모형

[1](16890) 경기도 용인시 수지구 죽전로 152, 단국대학교 응용통계학과. E-mail: rjpak@dankook.ac.kr