다변량 최근접 예측 모형:
거래량을 고려한 종합주가지수의 예측

윤종훈 · 이회경

한국과학기술원 테크노경영대학원
서울특별시 동대문구 청량리2동 207-43번지

Abstract

This paper examines the mutlivariate nearest neighbor forecasting model which considers the volume traded as well as the stock price. The empirical results using the data from KOSPI indicate that the predictive power of the nearest neighbor model increases as the model becomes multivariate.

1. 서론

시계열 자료에 대한 예측의 문제는 경제학을 위시한 사회 과학의 여러 분야에서 중요한 주제로 인식되어 왔다. 그러나 재무 분야의 경우 효율적인 시장은 예측이 불가능하다는 효율적 시장가설(efficient market hypothesis)이 제기된 이후로, 주가나 환율 등 재무관련 시계열의 예측은 학계의 관심을 받지 못하였다. 실제로 다른 분야의 시계열 예측에서 주로 사용되어 왔던 기존의 선형 모형들은 모두 랜덤워크(random walk)보다 나은 예측 결과를 보이는데 실패하였다.

Scheinkman and LeBaron (1989), Hsieh (1991), Abhyankar, Copeland and Wong (1997)등의 연구는 이런 선형 모형들의 한계에 대한 해답을 제시해 주었다. 이들은 BDS검정과 상관차원 검정(correlation dimension test)등을 통하여 주가와 환율을 포함한 대부분의 재무 관련 시계열들이 혼돈된 행태(chaotic behavior) 혹은 비선형 의존성(nonlinear dependency)을 가진다는 것을 보였다. 앞의 연구들과는 별도로, Lo and Mackinlay (1988), Frankel and Froot (1990), Jegadeesh (1990), Taylor and Allen (1992) 등은 재무 관련 시계열이 공통적으로 가지고 있는 혼돈된 행태, 이분산(heteroskedasticity), 꼬리가 두터운 분포(leptokurtosis)등의 성질을 분석하여 환율과 주가의 예측 가능성에 관한 연구들을 발표하였고, 이러한 결과들은 단기의 경우 부분적으로 재무 관련 시계열의 예측이 가능하다는 결론을 제시해 주었다.

재무 관련 시계열의 예측 모형들은 1990년대 초부터 제시되기 시작하였다. 대표적인 예로는 Sugihara and May (1990), Diebold and Nason (1990), LeBaron (1992) 등의 최근접 모형(nearest neighbor), Kuan and White (1994)의 신경망 이론(neural network), Bollerslev, Chou and Kroner (1992), Hentschel (1995) 등의 자기회귀 조건부 이분산 모형(ARCH) 등을 들 수 있다.

본 논문에서는 위에서 언급한 여러 모형들 중 최근접 모형에 초점을 맞추었다. 이제까지 제시된 최근접 모형들은 대부분 미래의 주가를 예측하기 위해 과거의 주가 자료만을 사용하였으나 본 연구에서는 가격 이외에 거래량이 가지고 있는 정보를 추가로 고려하였다. 이자율이나 환율 등의 변수들이 시장 밖에서 결정되는데 반해 거래량의 경우 시장 내에서 결정되는 변수로서 주가와는 밀접한 관계를 가지는 것으로 밝혀져 있다. 거래량이 주가에 미치는 영향에 관한 연구로는 Gallant, Rossi and Tauchen (1992), Campbell, Grossman and Wang (1993), Blume, Easley and O'hara (1994) 등이 있으며, 본 연구에서는 이러한 주가와 거래량의 관계를 최근접 모형에 적용하였다. 기존의 모형들이 주가의 수익률을 비교하여 근접치(neighbor)를 찾는데 반해 새로운 모형은 주가의 수익률과 거래량의 변화량을 함께 고려하여 근접치를 찾는다. 거래량의 추가가 모형에 미치는 영향을 알아보기 위하여 보유차원(embedding dimension)과 근접치의 수, 결과치(outcome data)에 주어지는 가중치는 기존의 모형에서 제시된 방법을 따랐다.

본 논문의 진행은 다음과 같다. 2절에서는 최근접 모형에 대한 이론적 고찰과 기존의 최근접 모형, 거래량을 함께 고려한 다변량 최근접 모형을 제시하고, 3절에서는 한국종합주가지수(KOSPI)의 예측을 통해 기존의 모형과 새로운 모형을 비교하였다. 마지막으로, 4절에서는 결론과 최근접 모형의 한계점 및 앞으로의 연구 방향을 제시하였다.

2. 이론적 고찰과 모형의 설계

시계열의 분석에서 근접치의 개념은 Stone (1977)의 연구에서 도입되었다. 그는 시계열을 일정한 기간의 단위로 나누어 벡터의 형태로 변환시킨 후, 비슷한 구조를 가지는 두 벡터를 근접치로 정의하였다. 시계열을 벡터로 변환하는 단위를 보유차원이라고 하며 $\{x_1, x_2, \cdots, x_{t-1}, x_t\}$라는 시계열을 가정할 때 일반적으로 t시점에서 변환된 벡터는

$$x_t = \left(x_{t-(E-1)}, x_{t-(E-2)}, \cdots, x_{t-1}, x_t\right) \quad (1)$$

가 된다. 벡터간의 거리를 측정하는 기준으로는 주로 기하학적 거리(euclidean distance)[1]가 사용되나 Casdagli (1992), Jaditz and Sayers (1998) 등의 연구에서는 최상기준(supreme norm)[2]도 사용된다.

근접치를 이용한 시계열의 예측은 크게 두 방

향으로 발전하였다. 하나는 근접치를 기존의 회귀분석 모형에 이용한 Cleveland (1979), Barkoulas, Baum and Onochie (1997), Jaditz and Sayers (1998) 등의 국부 가중 회귀분석(locally weighted regression)이며, 다른 하나는 근접치의 결과치를 예측에 이용한 Mizrach (1992), Linden, Satchell and Yoon (1993), Fernandez-Rodriguez and Sosvilla-Rivero (1998) 등의 최근접 모형이다. 본 논문에서는 분석의 대상을 최근접 모형에 한정하기로 한다.

2.1 최근접 모형

최근접 모형의 일반적인 형태는 $t \in [1, \cdots, n]$에서 다음과 같이 정의된다:

$$x_t = \sum_{i=1}^{n-1} w_{it} I[d(\boldsymbol{x}_i, \boldsymbol{x}_{t-1}) < \eta] y_i + \varepsilon_i \tag{2}$$

단, 여기에서 $I[\cdot]$는 근접치를 찾는 함수이고 $d(\cdot)$는 벡터간의 거리를 구하는 함수이며 η는 임의의 상수이다. y_i는 x_i의 결과치이며 w_{it}는 y_i에 주어지는 가중치이다.

최근접 모형을 이용하여 재무 관련 자료에 대해 예측을 시행한 연구들은 대부분 보유차원을 사전적으로 정한 뒤 분석을 행하였다. Mizrach (1996)는 보유차원이 1과 4의 경우에 대하여 예측을 시행하여 기존의 선형 모형과 비교하였고 Fernandez-Rodriguez and Sosvilla-Rivero (1998)는 보유차원을 6으로 고정하여 분석을 시행하였다. 그러나 기존의 연구들은 보유차원을 사전적으로 정함에 있어 이론적인 뒷받침을 제시하지 못하였다. 보유차원을 정하는 방법은 최근접 모형의 적용에 있어서 앞으로 해결해야 할 과제이다.

근접치의 수 역시 보유차원과 마찬가지로 사전적으로 결정되는 부분이다. 이제까지의 연구에 의하면 근접치의 수가 증가할수록 예측오차가 작아지는 경향을 보이나, 그 수가 너무 많아지면 오히려 오차가 증가한다는 것이 밝혀져 있다. 결과치에 적용되는 가중치는 연구에 따라 조금씩 다르지만, 거리가 커질수록 가중치가 줄어든다는 가정은 모든 연구에 대해 일치한다. 즉, 예측하고자 하는 시점의 벡터와 더 비슷한 구조를 가지는 벡터의 결과치에게 더 많은 가중치를 주는 것인데, Mulhern and Caprara (1994)는 식 (3)의 방법을, Linden, Satchell and Yoon (1993)은 식 (4)의 방법을 사용하여 각 결과치에 가중치를 주었다.

$$w_i = \frac{d_i^p}{\sum_{j=1}^{q} d_j^p} \tag{3}$$

$$w_i = \frac{e^{-d_i}}{\sum_{j=1}^{q} e^{-d_j}} \tag{4}$$

1) $d(\boldsymbol{x}_i, \boldsymbol{x}_t) = \sqrt{\sum_{k=0}^{E-1} (x_{i-k} - x_{t-k})^2}$

2.2 다변량 최근접 모형

기존의 최근접 모형과 본 연구에서 제시하는 모형의 가장 큰 차이점은 주가를 예측함에 있어 거래량이 가지는 정보를 이용한다는 것이다. 기존의 모형은 단지 주가의 벡터만을 비교하여 근접치를 찾았지만 새로운 모형은 거래량의 벡터도 함께 비교하여 근접치를 찾는다. 즉, 주가와 거래량이 모두 비슷한 구조를 가지는 벡터를 근접치로 찾는다. 기존의 연구에서는 두 벡터의 거리가

$$d(\boldsymbol{x}_i, \boldsymbol{x}_t) = \sqrt{\sum_{k=0}^{E-1} (x_{i-k} - x_{t-k})^2} \tag{5}$$

로 주어졌으나 거래량을 함께 고려하는 모형에서는 식 (6)과 같이 거래량을 포함하여 시계열을 정의하고 두 벡터의 거리를 식 (7)과 같이 정의한다. 단, 각 시계열의 분산이 다르므로 이를 보정하기 위하여 각각의 시계열을 정규화(normalization)한 뒤 예측에 적용한다.

$$\boldsymbol{x}_t = \begin{pmatrix} r_{t-(E-1)} & \cdots & r_{t-1} & r_t \\ v_{t-(E-1)} & \cdots & v_{t-1} & v_t \end{pmatrix} \tag{6}$$

$$d(\boldsymbol{x}_i, \boldsymbol{x}_t) = \alpha \sqrt{\sum_{k=0}^{E-1} (r_{i-k} - r_{t-k})^2} + (1-\alpha) \sqrt{\sum_{k=0}^{E-1} (v_{i-k} - v_{t-k})^2} \tag{7}$$

기존의 최근접 모형에 비해 위에서 제시된 모형이 가지는 장점은 더 설명력이 높은 근접치를 찾을 수 있다는 것이다. 본 연구에서는 식 (7)의 α를 변화시키며 예측을 시행하여 거래량이 가지는 정보가 주가의 예측에 미치는 영향의 크기를 분석하고자 한다.

3. 예측결과 및 분석

본 연구에서는 한국증권경제연구원에서 제공하는 한국의 종합주가지수의 일별 종가와 거래량을 사용하여 실증분석을 시행하였다. 자료의 수는 1980년 1월 4일부터 1995년 12월 27일까지의 4688개이며 1995년 1월 3일부터 1995년 12월 27일까지의 292일의 종가에 대하여 예측치를 구하여 실제의 자료와 비교하였다. 정상성(stationarity)과 관련된 문제들을 피하기 위해서 모든 시계열은 로그차분(log difference)하여 사용하였다.

분석의 편의를 위하여 두 모형 모두에 대해서 보유차원은 3, 근접치의 수는 10으로 사전적으로 제한하였다. 결과치에 주어지는 가중치는 식 (3)을 사용하였고 p는 Mulhern and Caprara (1994)의 연구에서와 같이 -2로 정하였다. 예측에 대한 오차는 MAE(mean absolute error)[3]와 RMSE(root mean squared error)[4]를 구하였으며 통계적인 유의성을 비교하기 위하여 Granger and Newbold (1986)가

2) $d(\boldsymbol{x}_i, \boldsymbol{x}_t) = \max_{k=0}^{E-1} |x_{i-k} - x_{t-k}|$

3) $\frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$

도출하고 Mizrach (1995)가 개량한 통계량을 구하였다.

Granger and Newbold (1986)에 의하면, 비교하고자 하는 두 예측 모형을 통해 구한 예측오차의 시계열을 각각 e_{1t}, e_{2t}라고 하고 (e_{1t}, e_{2t})가 평균이 0이고 분산이 각각 σ_1^2, σ_2^2인 이변량 정규분포(bivariate normal)을 따른다고 가정하면 $U=(e_1+e_2)$와 $V=(e_1-e_2)$ 역시 이변량 정규분포를 따르게 되고 U와 V의 상관계수는 $\sigma_1^2-\sigma_2^2$가 된다. 두 모형의 예측력의 비교는 U와 V의 상관계수가 0이라는 귀무가설 하에 표본분산의 차이가 근사적으로 정규분포를 따른다는 것을 이용하여 확인할 수 있다.

Mizrach (1995)는 Granger and Newbold (1986)의 가정을 완화하여 예측오차가 정규분포(normal distribution)를 따르지 않거나 예측오차에 편의(bias)나 상관관계(correlation), 이분산이 존재하여도 검정할 수 있는 식 (8)와 같은 통계량을 도출해 내었다. 식 (8)의 통계량 역시 U와 V의 상관계수가 0이라는 귀무가설 하에서 근사적으로 정규분포를 따른다.

$$\sqrt{n}\frac{1/n\sum_{j=1}^{n}U_jV_j}{\left[\sum_{l=-k}^{k}\left(1-\left[|l|/(k+1)\right]\right)s'_{UVUV}(l)\right]^{1/2}}\overset{asy.}{\sim}N(0,1) \qquad (8)$$

$$s'_{UVUV}(l)=1/n\sum_{j=l+1}^{n}U_jV_jU_{j-l}V_{j-l} \quad for \quad j\geq 0$$

단,
$$=1/n\sum_{j=-l+1}^{n}U_{j+l}V_{j+l}U_jV_j \quad for \quad j<0$$

$$k=k(n),\ with\ \lim_{n\to\infty}\frac{k(n)}{\sqrt{n}}=0$$

<표1> 최근접 모형의 예측오차와 검정 통계량

α	MAE	RMSE	Z-value
1	.00833982	.0110000	-
.99	.00829815	.0108769	2.23271648
.98	.00829313	.0108201	1.96465379
.97	.00823435	.0107221	2.35142655
.96	.00822028	.0106833	2.27600820
.95	.00826267	.0106985	1.92614247
.90	.00829676	.0106703	1.53130739
.85	.00824981	.0106524	1.35350910
.80	.00827359	.0106443	1.24424073
.75	.00830426	.0106501	1.13389711
.70	.00831241	.0107251	0.85119247
.65	.00835220	.0107222	0.82824418
.60	.00828324	.0106770	0.98184030
.55	.00823929	.0106402	1.10154191
.50	.00816866	.0105012	1.46884485
.45	.00817181	.0105359	1.28518878
.40	.00824213	.0105376	1.32747434
.35	.00825421	.0105788	1.19181646
.30	.00817346	.0105232	1.34000838
.25	.00812964	.0104734	1.43110472
.20	.00818479	.0105686	1.11995580
.15	.00817405	.0105317	1.20820366
.10	.00828603	.0106691	0.79498134
.05	.00832824	.0106642	0.87831507
0	.00870980	.0119350	-0.5143563

<표1>은 α를 변화시키며 실시한 예측에 대한 오차와 검정 통계량이다. α가 1인 행은 기존의 모형으로 예측한 결과이며 나머지 행은 거래량이 가지고 있는 정보가 예측에 반영된 결과이다. 각 행의 Z값은 기존의 모형에서 구한 예측오차와 α를 변화시키며 구한 예측오차를 식 (8)에 적용하여 구한 통계량이다. Z값이 양인 경우는 거래량을 고려한 모형으로 예측한 예측오차의 표본분산이 기존의 모형을 통한 예측의 표본분산보다 작은 경우이다.

대부분의 경우에서 거래량을 고려한 모형이 기존의 모형보다 오차가 작게 나왔다. 특히 근접치를 구하는 거리 함수에서 거래량이 차지하는 비율이 5%이하인 경우에는 Z값이 5% 유의수준에서 단측검정의 기각역인 1.645보다도 크게 나왔다. 이는 거래량을 고려한 모형의 예측력이 기존의 모형에 비해 통계적으로도 유의하다는 것을 의미한다.

4. 결론 및 연구 방향

본 연구에서는 1980년에서 1995년까지의 한국의 종합주가지수의 자료를 이용하여 거래량을 고려한 최근접 모형과 기존의 최근접 모형으로 예측을 시행하고 그 결과를 비교하였다.

주가와 거래량을 함께 고려한 모형의 예측오차가 기존의 모형에서 구한 예측오차보다 작았으며, 통계적으로도 유의한 것으로 나타났다.

본 연구에서 분석한 한국의 주식시장의 경우, 본 모형이 적용된 1980년부터 1995년까지는 가격에 제한폭이 존재하였다.[5] 일단 가격이 상한가나 하한가에 도달하면 거래가 중지되므로 정보가 가격과 거래량에 제대로 반영이 되지 않는 경우가 많았다. 그러므로 이러한 제약이 거의 없는 외국의 자료의 경우, 거래량이 가격에 미치는 영향을 더 정확히 분석할 수 있을 것으로 생각된다. 또한, 환율이나 이자율 등 주가 이외의 재무관련 시계열에 대해서도 본 모형의 적용이 가능하므로 이에 대한 연구 역시 필요할 것으로 보인다.

최근접 모형의 적용에서 상관차원과 근접치의 수 등 예측에 필요한 모수들은 분석자에 의해 사전적으로 정해진다. 그러나, 실증분석에서 예측오차는 이런 모수들의 변화에 상당히 민감하다. 그러므로 예측을 시행하기 전에 시계열의 행태를 분석하여 이런 모수들을 합리적으로 추정하는 이론적인 접근이 요구된다.

또한, 시계열 자료에 들어있는 여러 종류의 잡음을 제거할 수 있는 방법의 제시와 거래량이 가지고 있는 정보를 모형에 적용하는 방법의 개선 역시

4) $\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}}$

5) 1980년부터 1995년 3월까지는 전날 종가의 크기에 따라 범위를 나누고 각 범위별로 가격의 제한폭을 금액으로 명시하였으며, 1995년 4월부터 12월까지는 주가의 범위에 무관하게 전일 종가의 6%를 가격의 제한폭으로 규정하였다. 1998년 9월 현재 가격의 제한폭은 전일 종가의 12%이다.

필요하다.

5. 참고문헌

Abhyankar, A., L. S. Copeland and W. Wong (1997), "Uncovering Nonlinear Structure in Real-Time Stock-Market Indexes: The S&P 500, the DAX, the Nikkei 225 and the FTSE-100," *Journal of Business and Economic Statistics*, 15, 1-14.

Barkoulas, J. T., C. F. Baum and J. Onochie (1997), "A Nonparametric Investigation of the 90-day T-bill rate," *Review of Financial Economics*, 6, 187-198.

Blume, L., D. Easley and M. O'hara (1994), "Market Statistics and Technical Analysis: The Role of Volume," *Journal of Finance*, 49, 153-181.

Bollerslev, T., R. Y. Chou and K. F. Kroner (1992), "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics*, 52, 5-59.

Campbell, J. Y., S. J. Grossman and J. Wang (1993), "Trading Volume and Serial Correlation in Stock Returns," *Quarterly Journal of Economics*, 108, 905-939.

Casdagli, M. (1992), Chaos and Deterministic versus Stochastic Nonlinear Modeling, *Journal of Royal Statistical Society*, 54, 303-328.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829-836.

Diebold, F. X. and J. A. Nason (1990), "Nonparametric Exchange Rate Prediction?" *Journal of International Economics*, 28, 315-332.

Fernandez-Rodriguez, F. and S. Sosvilla-Rivero (1998), "Testing Nonlinear Forecastability in Time Series: Theory and Evidence from the EMS," *Economics Letters*, 59, 49-63.

Frankel, J. A. and K. A. Froot (1990), "Chartists, Fundamentals and Trading in the Foreign Exchange Market," *American Economic Review Papers and Proceedings*, 80, 181-185.

Gallant, A. R., P. E. Rossi and G. Tauchen (1992), "Stock Prices and Volume," *Review of Financial Studies*, 5, 199-242.

Granger, C. W. J. and P. Newbold (1986), *Forecasting in Business and Economic Time Series*, Academic Press.

Hentschel, L. (1995), "All in the Family: Nesting Symmetric and Asymmetric GARCH Models," *Journal of Financial Economics*, 39, 71-104.

Hsieh, D. A. (1991), "Chaos and Nonlinear Dynamics: Application to Financial Markets," *Journal of Finance*, 46, 1839-1877.

Jaditz, T. and C. L. Sayers (1998), "Out-of-Sample Forecast Performance as a Test for Nonlinearity in Time Series," *Journal of Business and Economic Statistics*, 16, 110-117.

Jegadeesh, N. (1990), "Evidence of Predictable Behavior of Security Returns," *Journal of Finance*, 45, 881-898.

Kuan, C. M. and H. White (1994), "Artificial Neural Networks: An Econometric Perspective," *Econometric Reviews*, 13, 1-91.

LeBaron, B. (1992), "Forecast Improvements Using a Volatility Index," *Journal of Applied Econometrics*, 7, S137-S149.

Linden, N., S. Satchell and Y. Yoon (1993), "Predicting British Financial Indices: An Approach Based on Chaos Theory," *Structual Change and Economic Dynamics*, 4, 145-162.

Lo, A. W. and A. C. Mackinlay (1988), "Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, 1, 41-66.

Mizrach, B. (1992), "Multivariate Nearest Neighbor Forecasts of EMS Exchange Rates," *Journal of Applied Econometrics*, 7, S151-S163.

Mizrach, B. (1995), "Forecast Comparisons in L2," Working Paper 95-24, Rutgers University.

Mizrach, B. (1996), "The Information in the Term Structure: A Nonparametric Investigation," *Journal of Forecasting*, 15, 137-153.

Mulhern, F. J. and R. J. Caprara (1994), A Nearest Neighbor Model for Forecasting Market Response, *International Journal of Forecasting*, 10, 191-207.

Scheinkman, J. A. and B. LeBaron (1989), "Nonlinear Dynamics and Stock Returns," *Journal of Business*, 62, 311-338.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *Annals of Statistics*, 5, 595-645.

Sugihara, G. and R. M. May (1990), "Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series," *Nature*, 344, 734-741.

Taylor, M. P. and H. Allen (1992), "The Use of Technical Analysis in the Foreign Exchange Market," *Journal of International Money and Finance*, 11, 304-314.