Journal of the Korean Data & Information Science Society
2011, **22**(4), 727–733

한국데이터정보과학회지

확률행렬이론을 이용한 한국주식시장의 상관행렬 분석[†]

김건우[1] · 이승철[2]

[12]연세대학교 수학과

접수 2011년 3월 25일, 수정 2011년 5월 28일, 게재확정 2011년 7월 8일

요 약

주식수익률간의 상관행렬 분석을 통해 유의미한 정보를 추출 활용하는 것은 주식시장을 이해하는데 매우 중요하다. 최근 확률행렬이론을 이용 상관행렬을 분석하는 연구들이 많이 진행되어 왔는데, 본 논문에서는 단일 요인 모형을 확률행렬이론에 적용 한국주식시장에서 주식수익률간의 상관행렬에 관한 유의미한 정보를 추출하였다. 특히 단일 요인을 도입 상관행렬을 분석한 결과가 실제 데이터를 잘 설명함을 관찰하였고, 단일 요인 모형의 유용성을 확인하였다.

주요용어: 단일 요인 모형, 상관 행렬, 확률행렬이론.

1. 서론

어느 금융자산의 평균 수익률과 위험을 알고 있을 때, 그것을 이용한 최적화된 포트폴리오에 대한 연구는 매우 오래전부터 되어왔다. Markowitz (1952)의 포트폴리오 선택이론에서부터 시작되어 최근까지도 다양한 방법으로 연구 되어왔다 (변현우 등, 2010). 그만큼 포트폴리오 최적화 문제는 금융 분야의 매우 중요하고 기본적인 문제 중 하나이다. 포트폴리오 최적화 문제를 자세하게 살펴보면, 우선 N개 자산으로 이루어진 포트폴리오 P의 평균 수익률을 R_P라 하면 R_P는 다음과 같이 정의할 수 있다.

$$R_P = \sum_{i=1}^{N} w_i R_i, \ (R_i\text{는 개별자산 } i\text{의 기대수익률}, w_i\text{는 개별자산 } i\text{의 투자 비율})$$

비슷하게 포트폴리오의 위험 (risk)은 전체 분산 σ_P^2에 연관시켜 표현 할 수 있다 ($\sigma_P^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i\sigma_i\rho_{ij}\sigma_j w_j$, ρ_{ij}는 자산 i, j의 상관계수, σ_i는 자산 i의 표준편차).

여기서 최적화된 포트폴리오는 주어진 R_P에 대하여 위험 (분산)을 최소화 하는 것이다. 즉 상관행렬의 정보 분류를 통한 효율적 포트폴리오를 세우고 분석하는 것은 매우 중요한 일이다. 다시 얘기하면 포트폴리오 최적화 문제는 상관행렬 분석과 매우 밀접하게 연관되어 있다. 상관행렬은 의미 없는 무작위 요인으로 부터의 상관행렬과 무작위 요소가 제거된 의미 있는 요인의 상관행렬로 이루어져 있는데, 이 두 영역을 구분하고 분석하는 연구가 다양한 방법으로 시도되고 있다. 그 중 확률행렬이론을 통한 상관행렬 분석 방법이 최근 많이 연구되고 있다. 확률행렬이론을 이용한 시계열 금융부분의 연구는 Laloux 등 (1999), Plerou 등 (2002), Utsugi 등 (2004), Conlon 등 (2007, 2009) 등에 의하여 활발히 진행되

[†] 이 논문은 2010년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2010-0010919).
[1] 교신저자: (120-749) 서울특별시 서대문구 신촌동 134번지, 연세대학교 수학과, 대학원생.
E-mail: geonwoo@yonsei.ac.kr
[2] (120-749) 서울특별시 서대문구 신촌동 134번지, 연세대학교 수학과, 교수.

어 왔다. 우리나라 KOSPI에 대한 연구는 Shin (2006), Park과 Lee (2007), Kim 등 (2008) 등에 의해 많이 연구 되어 왔으나 확률행렬이론을 이용한 연구는 많이 되어 있지 않다. 따라서 본 논문은 한국 주식 시장의 상관행렬들이 확률행렬이론 의해 얼마나 잘 설명되는지 살펴보았다. 또한, 자본자산가격결정모형 (Capital Asset Pricing Model)에서 현실검증에 주로 사용되는 시장지수를 이용한 단일 요인 모형 (One-factor model)을 통해 시계열 자료를 생성하고 생성된 자료를 통해 대행 상관행렬 (Surrogate correlation matrix)을 구하고 대행 상관행렬이 실제 상관행렬과 어느 정도 일치 하는지 비교 분석 할 것이다. 행렬의 속성은 행렬 분해에 의한 고유값 (Eigenvalue) 분포로 잘 설명 할 수 있으며, 상관행렬의 개별 포트폴리오에서 분산 역할을 하고 있으므로 우리는 고유값 분포를 중심으로 상관행렬을 분석하도록 한다.

2. 방법

2.1. 확률행렬이론

본 연구는 2003년 9월부터 2010년 12월까지 1,800일 동안 액면 병합이나 분할이 없는 KOSPI 상장종목 403개의 개별 주식을 이용하였다. 각각의 종목에 대해서는 시간에 따른 로그수익률 $R(t) = \ln P_{t+1} - \ln P_t$을 사용하였고 각각의 수익률을 다음과 같이 표준화 하였다.

$$g_i(t) = \frac{R_i(t) - \overline{R}_i}{\sigma_i} \tag{2.1}$$

여기서, $\overline{R}_i$는 개별종목 $i(i = 1 \sim 403)$의 조사기간에 대한 평균이며 σ_i는 표준편차이다. 이 때, 동시간 개별종목 i, j의 상관계수를 $g_i(t)$를 써서 표시하면

$$C_{ij} \equiv < g_i(t), g_j(t) > \tag{2.2}$$

이 되고, 상관행렬 C를 $g_i(t)$를 원소로 가지는 $N \times T$행렬 G로 표현하면 다음과 같다.

$$C = \frac{1}{T} G G^T \tag{2.3}$$

C는 대칭행렬이므로 대각화가 가능하여 다음과 같이 스펙트럴 분해를 할 수 있다.

$$C = PDP^T = \sum_{j=1}^{n} \lambda_j \overrightarrow{w_j} \overrightarrow{w_j}^T \tag{2.4}$$

여기서 $\overrightarrow{w_j}$는 포트폴리오 j의 컬럼벡터를 나타내며, 모든 포트폴리오 j에 대해서 $\overrightarrow{w_j}\overrightarrow{w_j}^T$이 일정하다면 λ_j는 포트폴리오 j의 분산역할을 한다.

경험적 상관행렬 C의 고유값 밀도 함수를 다음과 같이 정의 하도록 한다.

$$\rho_C(\lambda) = \frac{1}{N} \frac{dn(\lambda)}{d\lambda} \tag{2.5}$$

여기서 $n(\lambda)$는 λ보다 작은 C의 고유값의 수이다.

다음은 확률행렬이론에 의한 확률밀도함수이다. 상관행렬의 고유치 분포는, Sengupta 등 (1999)에 의하여 주식수 N과 시간 T가 무한하다고 가정하고 T/N가 1보다 큰 Q로 수렴한다고 가정하면 (즉 $T/N := Q$) 무작위 상관행렬의 고유치 분포는 다음과 같다.

$$\rho_C(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda}, \ \lambda_{\min}^{\max} = \sigma^2(1 + 1/Q \pm 2\sqrt{1/Q}) \tag{2.6}$$

여기서 이론적으로 모든 λ는 $\lambda_{\min}$과 $\lambda_{\max}$사이에 있어야 한다. 또한, σ^2은 G의 모든 원소 분산이다. 하지만, 우리는 (2.1)에서와 같이 표준화 했으므로 1로 생각할 수 있다. $\lambda_{\min}^{\max}$는 이론적 고유치 분포의 상한과 하한이다. 상한과 하한에 속하는 고유치는 확인할 수 없는 무작위 요인들의 속성을 표현하는 것이다. 이 범위를 벗어난 고유값은 확률행렬이론으로부터 일탈 (deviate)되어 있다고 한다. 다시 애기하면 범위 안에 있는 고유치는 확인할 수 없는 무작위 요인들의 속성을 포함하는 것이고, 반면에 범위에 벗어나는 고유값은 의미 있는 요인들의 속성을 포함하는 것으로 본다.

2.2. 단일 요인 모형

각각의 개별 종목이 Sharpe (1963)의 단일 요인으로 설명된다고 가정하자. 그 다음은 다음과 같이 단순 회귀식에 적합 후 개별종목의 로그 수익률을 종속변수로 하고 시장 로그 수익률을 설명 변수로 하여 α_i, β_i그리고 오차 ε_i를 계산한다.

$$R_i(t) = \alpha_i + \beta_i M^{market}(t) + \varepsilon_i(t) \tag{2.7}$$

여기서 $M^{market}(t)$는 일별 KOSPI 로그 수익률이며, $\varepsilon_i(t)$는 Huang 등 (2005)에서 설정한 바와 같이 독립이며 평균이 0이고, 분산 σ^2을 가정한다.

먼저 $M^{market}(t)$의 분산 σ_M^2을 추정한 뒤, 1차 회귀식으로 구해진 α_i, β_i와 시장 로그 수익률을 이용하여 오차 $R_i(t) - \alpha_i - \beta_i M^{market}(t) = \varepsilon_i(t)$를 계산하고, 이것을 통하여 전체 오차의 분산 σ_ε^2을 추정한다. 그러면 각각의 자산 402개의 α_i, β_i와 분산 σ_M^2, σ_ε^2, 총 806개의 모수가 추정된다. 그 다음, 시장의 로그 수익률 과 오차항의 분포가 정규 분포를 따른다고 가정하고, 앞서 구한 분산에 대한 모수를 이용, 평균 0과 분산 σ_M^2인 시장 수익률을 난수 생성기를 통해 조사 기간만큼의 독립적인 데이터 $\widehat{M}(t)$를 생성하고 오차항 역시 같은 방법으로 평균 0과 분산 σ_ε^2인 난수 $\widehat{\varepsilon}(t)$를 생성한다. 이렇게 생성된 난수와 각 자산별 α_i, β_i를 이용하여 실제 자료와 같은 추정 수익률 $\widehat{R_i}(t)$를 산출한다. 추정된 개별 자산 수익률로 대행 상관행렬을 구하고 그 상관행렬에 대해 2.1절에서 설명한 확률행렬이론을 적용 해본다. 마지막으로 실제 데이터를 이용한 결과와 대행 상관행렬의 고유값 분포를 비교 분석 하였다.

3. 결과

3.1. 실제 데이터 고유값 분석

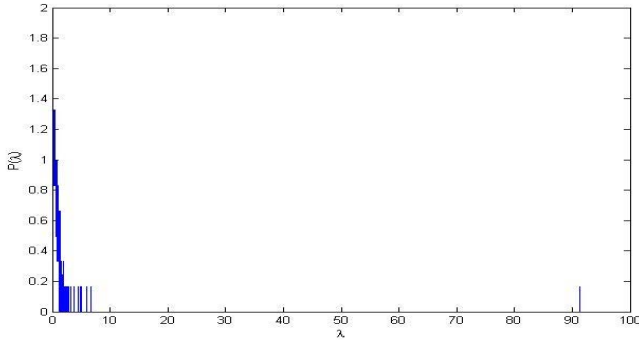

그림 3.1 실제 시장 데이터의 표준화된 상관행렬 고유값 분포

그림 3.1은 실제 데이터를 이용한 고유값들을 식 (2.5)를 이용해 표현한 결과이다. 그림에서 보듯이 시장 요인으로 볼 수 있는 상대적으로 매우 큰 고유값이 발견되었다. 그 차이가 다른 고유값과 매우 크므로 가장 큰 고유값만이 의미있는 단일요인으로 생각한다. 그 이외의 고유값에 대해서는 설명 할 수 없는 부분으로 생각하고 의미 있는 요인 (가장 큰 고유값)을 제거하면 나머지 부분의 상관행렬은 순수한 무작위라는 가정에 기초하여 그것의 분포가 확률행렬이론의 이론적 식 (2.6)과 어느 정도 일치하는지 확인하였다. 실제 자료 상관행렬의 가장 큰 고유값은 91.26이며 전체 고유 값의 합 402(=N)의 약 23%를 차지한다. 이와 같은 사실에 기초하여 표준화 한 자료의 $\sigma^2 = 1$로부터, 가장 큰 고유값 부분을 제거한 $\sigma^2 = (1 - \lambda_{\max})/N = 0.77$을 이용하여 살펴보았다.

나머지 고유값의 밀도와 확률행렬이론과의 적합결과는 그림 3.2에 나타내었다. 그 결과, 가장 큰 고유값을 제거한 $\sigma^2 = 0.77$을 이용하는 경우에는 고유값의 13%가 식 (2.6)의 이론적 상한에 벗어나 있는 것을 알 수 있었다. 이러한 결과는 고유값이 이론적 상한에 초과되어지는 부분이 많은 것으로 판단되어, Laloux 등 (1999)에서 보여진바와 같이 주어진 고유값의 분포의 범위를 가장 적합도가 좋은 σ^2를 0.01단위로 찾아보았고, 그 결과 $\sigma^2 = 0.85$를 이용하는 경우 이론적 상한에 남아있는 고유값 중 7%가 벗어나 무작위 고유값 분포의 93%가 식 (2.6)의 범위에 들어가 있는 것을 확인하였다.

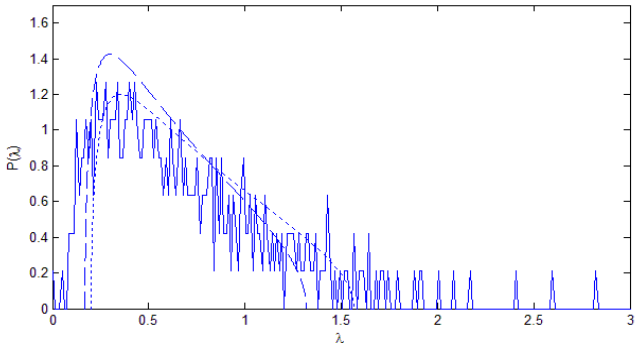

그림 3.2 실선은 자산 N=402의 실제자료 상관행렬의 고유값 밀도.
긴 점선은 $\sigma^2 = 0.77$에 대한 식 (2.6)에 의한 이론적 결과.
짧은 점선은 가장 좋은 적합도를 보인 $\sigma^2 = 0.85$에 대한 식 (2.6)에 의한 이론적 결과.

3.2. 단일 요인 모형 분석

2.2절에서 설명되어진 방법을 통한 단일 요인 모형에 의한 대행 상관행렬 고유값을 이용하여 3.1절에서의 결과를 다시 구현해 보았다. 그 결과는 그림 3.3에 나타내었다.

그림 3.3 역시 그림 3.1과 같이 매우 큰 고유값이 발견되었으며, 가장 큰 고유값 만이 의미있는 요인이라 생각하여, 가장 큰 고유값을 제거한 부분이 무작위 부분을 설명하는 식 (2.6)에 얼마나 잘 적합하는지를 살펴보았다.

실험 결과 단일 요인 모형을 이용한 대행 상관행렬의 고유값의 ($\sigma^2 = (1 - \lambda_{\max})/N$=0.82를 이용) 상위 8%의 고유값이 이론적 상한에 벗어나 있음을 확인 할 수 있다. 이 결과는 또한 그림 3.4를 통해 나타내었다. 이것은 단일 요인 모형의 무작위 부분이 확률행렬이론으로 92%가 설명되고 있음을 나타내는 것이며, 이 결과는 3.1절에서 살펴본 실제자료의 무작위 고유값 분포가 확률행렬이론에 의해 매우 비슷하게 설명되고 있음을 알 수 있다.

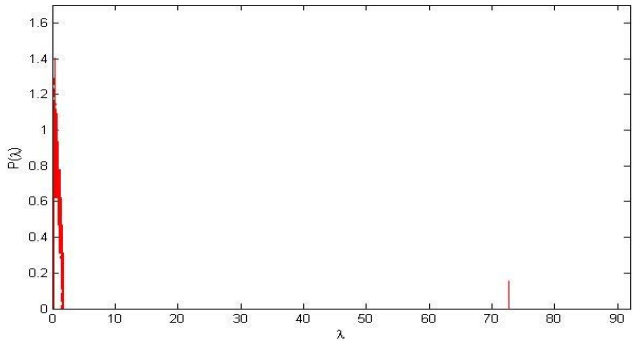

그림 3.3 단일 요인 모형 데이터의 표준화된 상관행렬 고유값 분포

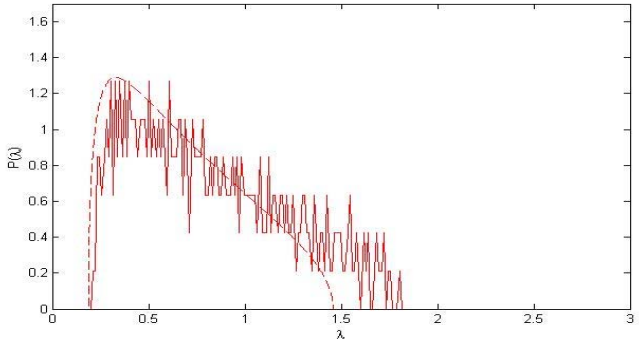

그림 3.4 실선은 단일요인모형 상관행렬의 고유값 밀도.
긴 점선은 $\sigma^2 = 0.82$에 대한 식 (2.6)에 의한 이론적 결과.

4. 결 론

본 논문에서는 주식 시장의 실제 데이터의 상관행렬과 단일 요인 모형을 이용한 상관행렬의 속성을 확률행렬이론을 통하여 비교 분석하였다. 그 결과 확률행렬이론에 의해 상관행렬의 고유값 분포의 무작위 부분의 상당부분 (실제 자료의 87%, 단일 요인 모형의 92%)을 설명하고 있음을 알 수 있었다. 또한 대행 상관행렬로서 단일 요인 모형 상관행렬의 고유값 분포 역시 실제 자료와 매우 비슷한 분포를 가지고 있는 것을 확인하였으며, 단일 요인 모형에서 가장 큰 고유값 제거 분산인 0.82와 실제자료의 가장 큰 고유값 제거 분포에서 가장 적합한 분산인 0.85가 비슷하다는 결과는 단일 요인 모형의 유용성을 설명한다. 이 결과의 내용들을 바탕으로 주식시장의 상관행렬의 고유값을 확률행렬이론에 의해 분해하고 이것을 재무분야 이론에 반영한다면 주식간의 관계 이해를 통한 포트폴리오 최적화 문제에 유용할 것이라고 생각된다.

참고문헌

변현우, 송치우, 한성권, 이태규, 오경주 (2009). 변동성 지수기반 유전자 알고리즘을 활용한 계층구조 포트폴리오 최적화에 관한 연구. <한국데이터정보과학회지>, **20,** 1049-1060.

Conlon, T., Ruskin, H. J. and Crane, M. (2007). Random matrix theory and fund of funds portfolio optimisation. *Physica A*, **382**, 565-576.

Conlon, T., Ruskin, H. J., and Crane, M. (2009). Cross-correlation dynamics in financial time series. *Physica A*, **388**, 705-714.

Huang, H. T. and Cheng, W. H. (2005). Test of the CAPM under structural changes. *International Economic Journal*, **19,** 523-541.

Kim, K. K., Cho, M. H. and Park, E. S. (2008). Forecasting the volatility of KOSPI 200 using data mining. *Journal of the Korean Data & Information Science Society,* **19**, 1305-1325.

Laloux, L., Cizeau, P., Bouchaud, J. P. and Potters, M. (1999). Noise dressing of financial correlation matrices. *Physical Review Letters*, **83**, 1467-1470.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance,* **7,** 77-91.

Park, S. Y. and Lee, S. Y. (2007). Modelling KOSPI200 data based on GARCH(1,1) parameter change test. *Journal of the Korean Data & Information Science Society,* **18**, 11-16.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N. and Stanley, H. E. (2000). A random matrix theory approach to financial cross -correlations. *Physica A*, **287**, 374-382.

Sengupta, A. M. and Mitra, P. P. (1999). Distribution of singular values for some random matrices. *Physical Review E,* **60,** 3389-3392.

Sharpe, W. (1963). A simplified model for portfolio analysis. *Management Science,* **9**, 277-293.

Shin, Y. K. (2006). An empirical study on stock trading value of each investor type in the korean stock market. *Journal of the Korean Data & Information Science Society,* **17**, 1099-1106.

Utsugi, A., Ino, K. and Oshikawa, M. (2004). Random matrix theory analysis of cross correlation in financial markets. *Physical Review E,* **70,** 1-11.

Journal of the Korean Data & Information Science Society
2011, **22**(4), 727–733

한국데이터정보과학회지

A Random Matrix Theory approach to correlation matrix in Korea Stock Market[†]

Geon Woo Kim[1] · Sung Chul Lee[2]

[12]Department of Mathematics, Yonsei University

Received 25 March 2011, revised 28 May 2011, accepted 8 July 2011

Abstract

To understand the stock market structure it is very important to extract meaningful information by analyzing the correlation matrix between stock returns. Recently there has been many studies on the correlation matrix using the Random Matrix Theory. In this paper we adopt this random matrix methodology to a single-factor model and we obtain meaningful information on the correlation matrix. In particular we observe the analysis of the correlation matrix using the single-factor model explains the real market data and as a result we confirm the usefulness of the single-factor model.

Keywords: Correlation matrix, Random Matrix Theory, single-factor model.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0010919).
[1] Corresponding author: Graduate student, Department of Mathematics, Yonsei University, Seoul 120-749, Korea. E-mail: geonwoo@yonsei.ac.kr
[2] Professor, Department of Mathematics, Yonsei University, Seoul 120-749, Korea.