Communications of the Korean Statistical Society Vol. 15, No. 6, 2008, pp. 925–937

이분산 시계열모형을 이용한 국내주식자료의 군집분석

박만식[1], 김나영[2], 김희영[3]

요 약

본 논문에서는 주식시장에서 거래되는 다수의 주식거래종목들을 몇 개의 그룹으로 군집화하는 주제를 연구한다. 시간에 관계없이 분산이 일정한 ARMA모형과 다르게, 주가, 환율 등의 금융시계열자료에서는 조건부 이분산성을 따르게 된다. 또한, 많은 사람들이 금융시계열자료에서 관심을 갖는 것은 바로 이 변동성이다. 그러므로, 이 연구에서는 조건부 이분산성을 모형화하기에 적합하다고 알려진 일반화 조건부 이분산성 자기회귀모형에 초점을 맞춘다. 먼저 두 개의 주식종목들 사이에 변동성(volatility)의 유사성 그리고 구조의 유사성을 재는 거리를 정의하고, 모의실험을 수행한다. 실증자료로 최근 3년 동안 관찰된 국내 11개 주가의 수익률을 변동성과 구조에 따라 군집화한다.

주요용어: 일반화 자기회귀 조건부 이분산; 무조건부 분산; 군집분석.

1. 서론

주가, 환율 등 금융시계열자료에서 수익율(= $(P_{t+1} - P_t)/P_t$, 여기서 P_t를 t시점의 지수)에서 나타나는 특징은 큰폭으로 변동하는 기간이 한동안 계속되다가 또다른 기간에는 안정 국면이 지속되는 변동성 집중(volatility clustering) 현상을 나타낸다. 변동성은 통계학의 관점에서 보면 분산에 해당되는 개념이므로 변동성 집중현상이 나타나는 이유는 미래값의 분산이 현재의 상황에 의존하는 것을 뜻한다. 즉, 조건부 분산(conditional variance)들 사이에 연관관계가 존재함을 의미한다. 이를 모형화하기 위하여 Engle (1982)은 자기회귀 조건부 이분산(autoregressive conditional heteroscedastic model: ARCH)모형을 제안하였다. 이후 Bollerslev (1986)과 Taylor (1986)은 일반화 자기회귀 조건부 이분산(generalized autoregressive conditional heteroscedastic: GARCH)모형으로 확장하였다. 정확한 변동성의 예측은 여러분야의 투자의사결정에 있어서 상당히 중요한 역할을 한다. 예를 들어 성공적인 위험관리, 자산배분결정 등에서는 신뢰할만한 변동성 예측이 성패를 좌우한다고 할 수 있다. 따라서, 투자자와 증권사의 애널리스트 등에게는 거래되는 수백개의 주식거래종목들에서 수익율의 패턴이 유사한 그리고 위험도가 비슷한 종목들로 그룹화하는 일은 의사결정에 많은 도움을 줄 수 있다.

1) (136-701) 서울시 성북구 안암동 5가1 126-1 고려대학교 의과대학 의학통계학교실 및 의과학연구원 (유전체 및 단백체 독성연구소), 연구교수.
2) (110-789) 서울시 종로구 종로2가 6 삼성증권 마케팅파트, 차장.
3) (136-701) 서울시 성북구 안암동 5가1 126-1 고려대학교 의과대학 의학통계학교실 및 의과학연구원 (유전체 및 단백체 독성연구소), 연구교수. 교신저자: starkim@korea.ac.kr

본 논문에서는 금융시계열에서 변동성을 모형화하기에 적합하다고 알려진 다수의 GARCH계열을 군집화하고자 한다. 본 논문의 구성은 다음과 같다. 먼저 2절에서는 ARCH모형, GARCH모형을 간단히 요약하고, 3절에서는 GARCH모형을 군집화한 기존의 연구들을 소개하고, 몇가지 문제점들을 짚어본다. 4절에서는 모의실험을 수행하고, 5절에서는 2005년 1월 3일부터 2007년 12월 31일까지의 총 796일 동안 관찰된 국내 11개 주가(강원랜드, 삼성엔지니어링, 웅진코웨이, 제일기획, 신한지주, 국민은행, 외환은행, 현대건설, 경남기업, 대림산업, GS건설)의 수익율을 GARCH모형으로 적합한 후에 이들을 군집화할 것이다. 결론은 6절에서 제시하고자 한다.

2. 이분산성 시계열 모형

2.1. ARCH모형

Engle (1982)이 제안한 차수가 P인 자기회귀 조건부 이분산(ARCH(P))모형은 다음과 같다.

$$X_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \gamma + \sum_{p=1}^{P} \beta_p X_{t-p}^2, \tag{2.1}$$

여기서, $\gamma \geq 0$, $\beta_p \geq 0$, $\{\epsilon_t\} \sim i.i.d.(0, 1)$이고 $\{\epsilon_t\}$와 $\{X_{t-k};\, k \geq 1\}$는 독립이다. 식 (2.1)의 확률과정이 2차적률이 유한한 강정상확률과정(strictly stationary process)이 되기 위한 필요충분조건은 $\sum_{p=1}^{P} \beta_p < 1$이다. ARCH($P$)모형에서 몇가지 특징을 요약하면 다음과 같다.

(1) $\{X_t\}$는 백색잡음(white noise: WN)이고 다음의 무조건부 분산(unconditional variance: UV)을 가진다.

$$\mathrm{Var}(X_t) = \frac{\gamma}{1 - \sum_{p=1}^{P} \beta_p}.$$

(2) $(t-1)$시점까지의 자료가 주어졌을 때 t시점의 조건부 분산(conditional variance)은

$$\mathrm{Var}(X_t \mid X_{t-k},\, k \geq 1) = \gamma + \sum_{p=1}^{P} \beta_p X_{t-p}^2 \tag{2.2}$$

으로 식 (2.1)에서의 σ_t^2에 해당한다. 즉, t시점의 조건부 분산 σ_t^2은 P개의 과거자료인 $\{X_{t-1}, \ldots, X_{t-P}\}$으로 표현된다.

(3) 추가적으로 $\{E(\epsilon_t^4)\}^{1/2} \sum_{p=1}^{P} \beta_p < 1$이라는 조건을 만족하면,

(i) $\{X_t^2\}$은 자기회귀(autoregressive: AR(P))모형이 되고, 자기상관함수는 모든 차수에서 항상 양의 값을 갖는다.

(ii) X_t의 첨도(kurtosis)는 오차항(ϵ_t)의 첨도보다 크다. 따라서, 오차항을 정규분포로 가정할 경우 X_t의 분포함수는 정규분포보다 두꺼운 꼬리를 가지게 된다.

2.2. GARCH모형

GARCH모형은 Bollerslev (1986)가 ARCH모형을 일반화시킨 것으로 ARCH구조에 조건부 분산의 시차변수(σ_{t-k}^2; $k \geq 1$)를 추가시킨 모형으로 다음의 식을 만족할 때 $\{X_t\}$는 GARCH(P, Q)모형을 따른다고 한다.

$$X_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \gamma + \sum_{p=1}^{P} \beta_p X_{t-p}^2 + \sum_{q=1}^{Q} \alpha_q \sigma_{t-q}^2, \tag{2.3}$$

여기서, $\gamma \geq 0$, $\beta_p \geq 0$, $\{\epsilon_t\} \sim i.i.d(0,1)$이고, ϵ_t와 $\{X_{t-k}, k \geq 1\}$는 독립이다. 식 (2.3)의 확률과정이 강정상성(strict stationarity)을 만족하기 위한 필요충분조건은 $\sum_{p=1}^{P} \beta_p + \sum_{q=1}^{Q} \alpha_q < 1$이다. 다음은 GARCH($P$, Q)모형에서 주목할 몇 가지 성질을 정리한 것이다.

(1) $$\{X_t\} \sim \mathrm{WN}\left(0, \frac{\gamma}{1 - \sum_{p=1}^{P} \beta_p - \sum_{q=1}^{Q} \alpha_q}\right), \tag{2.4}$$

(2) $$\mathrm{Var}(X_t \mid X_{t-k}, k > 1) = \sigma_t^2 = \frac{\gamma}{1 - \sum_{p=1}^{P} \alpha_q} + \sum_{k=1}^{\infty} d_k X_{t-k}^2, \tag{2.5}$$

여기서 d_k는 식 $\sum_{k=1}^{\infty} d_k z^k = \sum_{p=1}^{P} \beta_p z^p / (1 - \sum_{q=1}^{Q} \alpha_q z^q)$의해 결정되는 상수이다. 식 (2.5)와 ARCH(P)모형의 식 (2.2)을 비교하여 보면, GARCH(P, Q)모형은 보다 작은 갯수의 모수들을 이용하여 t시점의 조건부분산 σ_t^2에 미치는 이전 시점의 모든 자료 (X_{t-k}, $k \geq 1$)의 영향을 표현할 수 있다.

(3) 또한 $\{E(\epsilon_t^4)\}^{1/2} (\sum_{p=1}^{P} \beta_p) / (1 - \sum_{j=1}^{q} \alpha_q) < 1$의 조건을 만족하면,

(i) $\{X_t^2\}$는 자기회귀 이동평균(autoregressive moving-average: ARMA($P \vee Q$, Q)) 과정이다. 여기서, $P \vee Q \equiv \max(P, Q)$.

(ii) X_t의 첨도는 오차항(ϵ_t)의 첨도보다 크다.

위에 열거된 성질들의 이론적 증명에 대하여는 Fan와 Yao (2005)을 참고하기 바란다.

3. ARCH모형 또는 GARCH모형의 군집분석

최근 Liao (2005)는 다양한 분야에서 수행된 시계열 자료의 군집분석에 대한 기존의 연구를 정리하였는데, 그는 크게 3가지로 연구들을 분류하였다. 첫째는 원자료에 기반한 접근방법, 둘째는 자료의 특성에 기반한 접근방법 그리고 셋째는 모형에 기반한 접근방법이

다. 이분산 시계열의 군집화에 대한 연구들은 모두가 세번째 방법인 모형에 기반한 방법으로 연구되었고, 등분산 시계열(ARMA, ARIMA, vector-ARMA)자료의 군집화에 대한 연구에 비해서 매우 적다. 조건부 이분산 시계열 자료에 대한 연구로는 Otranto (2008), Caiado와 Crato (2007), Bauwens와 Rombouts (2007) 등이 있으며, 이중에서 Bauwens와 Rombouts (2007)는 베이지안 관점이므로 본 논문에서는 고려하지 않기로 한다. 이 절에서는 Otranto (2008), Caiado와 Crato (2007)의 군집방법을 소개하고자 한다.

3.1. Caiado와 Crato (2007)의 연구

두 시계열 $\mathbf{x}_i = \{X_{i,1}, \ldots, X_{i,t}, \ldots, X_{i,T}\}$와 $\mathbf{x}_j = \{X_{j,1}, \ldots, X_{j,t}, \ldots, X_{j,T}\}$이 있고, 각각 GARCH(1, 1)모형을 따른다고 하자. 즉,

$$X_{i,t} = \sigma_{i,t}\epsilon_{i,t}, \quad \sigma^2_{i,t} = \gamma_i + \beta_{i,1}X^2_{i,t-1} + \alpha_{i,1}\sigma^2_{i,t-1}, \tag{3.1}$$

$$X_{j,t} = \sigma_{j,t}\epsilon_{j,t}, \quad \sigma^2_{j,t} = \gamma_j + \beta_{j,1}X^2_{j,t-1} + \alpha_{j,1}\sigma^2_{j,t-1}. \tag{3.2}$$

Caiado와 Crato (2007)는 조건부 이분산성을 가지는 시계열 자료의 군집화를 위하여 다음과 같은 거리를 정의하였다.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j\right)' \left(\widehat{\mathbf{V}}_i + \widehat{\mathbf{V}}_j\right)^{-1} \left(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j\right), \tag{3.3}$$

여기서, $\widehat{\boldsymbol{\theta}}_i = \left(\widehat{\beta}_{i,1}, \widehat{\alpha}_{i,1}\right)'$, $\widehat{\mathbf{V}}_i$는 i번째 시계열 자료에서 얻은 추정량들의 분산-공분산행렬(variance-covariance matrix)이다. Caiado와 Crato (2007)의 연구에서는 식 (3.3)의 거리의 성능과 효율이 실제로 GARCH(1, 1)모형에서 어느 정도인 지에 대한 모의 실험이 고려되지 않았다.

3.2. Otranto (2008)의 연구

B개의 시계열$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_B)$이 있고, $b = 1, \ldots, B$에 대하여 각각 GARCH(P_b, Q_b) 모형을 따른다고 하자. 즉,

$$X_{b,t} = \sigma_{b,t}\epsilon_{b,t}, \quad \sigma^2_{b,t} = \gamma_b + \sum_{p=1}^{P_b} \beta_{b,p}X^2_{b,t-p} + \sum_{q=1}^{Q_b} \alpha_{b,q}\sigma^2_{b,t-q}.$$

Otranto (2008)는 3단계의 계층적 검정단계(T1, T2, T3)로 이루어진 군집화 알고리즘을 제시하였다.

T1. (1) B개의 시계열자료들을 각각 모형화한 후, 무조건부 분산(UV)의 오름차순으로 정렬한다. 편의상 UV의 크기 순으로 $\mathrm{UV}(\mathbf{x}_1) < \cdots < \mathrm{UV}(\mathbf{x}_B)$라 하자.

(2) $H_{0,1}$: $\mathrm{UV}(\mathbf{x}_1) = \mathrm{UV}(\mathbf{x}_2)$인 지를 검정한다. 만약 $H_{0,1}$이 유의수준에서 기각되지 않으면, $H_{0,2}$: $\mathrm{UV}(\mathbf{x}_1) = \mathrm{UV}(\mathbf{x}_2) = \mathrm{UV}(\mathbf{x}_3)$인 지를 검정한다. 만약, $H_{0,2}$가 유의수준에서 기각되면, $\mathbb{G}_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$으로 하나의 그룹을 형성하게 된

다. 다음에는 $H_{0,3}$: UV($\mathbf{x}_3$) = UV($\mathbf{x}_4$)인 지를 검정한다. $H_{0,3}$가 기각되지 않으면, $H_{0,4}$: UV($\mathbf{x}_3$) = UV($\mathbf{x}_4$) = UV($\mathbf{x}_5$)인 지를 검정한다. $H_{0,4}$가 기각되지 않으면, $H_{0,5}$: UV($\mathbf{x}_3$) = $\cdots$ = UV($\mathbf{x}_6$)인 지 검정한다. $H_{0,5}$이 기각되면, 그룹 $\mathbb{G}_2 = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$가 형성된다. 동일한 방법으로 계속하여 검정해서 그룹 $\mathbb{G}_1, \ldots, \mathbb{G}_S$를 만든다.

T2. (1) T1에서 동일한 그룹으로 묶인 시계열자료들 내에서 시간에 의존하는 변동성(time-varying volatility: TVV)의 오름차순으로 정렬한다. Otranto (2008)은 식 (2.5)에서 $(\sum_{k=1}^{\infty} d_k^2)^{1/2}$을 TVV로 정의하고 있다.

(2) T1에서 같은 그룹으로 묶인 시계열들을 TVV가 같은 지에 따라 T1의 (2)에서와 같이 가설검정을 통하여 T1의 그룹보다 작은 소규모로 그룹화한다. 예를 들어, T1에서 $\mathbb{G}_2 = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$이고, 이들을 TVV의 오름차순으로 정렬하였을 때 TVV($\mathbf{x}_5$) < TVV($\mathbf{x}_3$) < TVV($\mathbf{x}_4$)이라 하자. $H_{0,1}$: TVV($\mathbf{x}_5$) = TVV($\mathbf{x}_3$)인 지를 검정하였더니 기각되고, $H_{0,2}$: TVV($\mathbf{x}_3$) = TVV($\mathbf{x}_4$)가 기각되지 않았다면, $\mathbb{G}_2$ 그룹은 $\mathbb{G}_{21} = \{\mathbf{x}_5\}$와 $\mathbb{G}_{22} = \{\mathbf{x}_3, \mathbf{x}_4\}$로 나뉘어진다. 이런 방법으로 T1에서의 그룹 $\mathbb{G}_1, \ldots, \mathbb{G}_S$는 T2의 과정을 통해 각각 $\{\mathbb{G}_{11}, \ldots, \mathbb{G}_{1n_1}\}, \ldots, \{\mathbb{G}_{S1}, \ldots, \mathbb{G}_{Sn_S}\}$으로 나뉘어진다.

T3. T2에서의 하위 그룹들 내에서의 시계열 자료의 구조(structure)가 같은 지를 검정하여 최종그룹으로 묶는다. 여기서, 구조가 같다는 것은 두 모형에서의 모수가 일치한다는 것을 의미한다. 예를 들어 $\mathbf{x}_1$과 $\mathbf{x}_2$가 같은 구조인 지를 검정하는 것은, 모든 $p = 1, \ldots, \max(P_1, P_2)$와 $q = 1, \ldots, \max(Q_1, Q_2)$에 대하여,

$$H_0 : \gamma_1 = \gamma_2, \ \{\beta_{1,p} = \beta_{2,p}\}, \ \{\alpha_{1,q} = \alpha_{2,q}\}$$

임을 뜻한다.

Otranto (2008)는 위에서 기술한 세 가지의 가설검정과정(T1-T3)에서 모두 Wald검정통계량을 사용하였고, 그가 제안한 통계량의 판별력를 평가하기 위하여 모의실험을 수행하였다.

4. 모의실험

4절에서는 GARCH모형의 두 시계열자료 간에 무조건부 분산에 대한 유사성 거리와 구조에 대한 유사성 거리를 소개하고 군집의 판별력을 평가하기 위하여 모의실험을 수행하고자 한다. 두 시계열자료, $\mathbf{x}_i$와 $\mathbf{x}_j$가 식 (3.1)과 (3.2)와 같이 GARCH(1,1)모형을 따른다고 가정하자.

4.1. 무조건부 분산의 유사성거리

Otranto (2008)에서 사용한 Wald통계량을 구성하기 위한 행렬들($\mathbf{A}$, $\boldsymbol{\delta}_{ij}$, $\mathbf{G}_{ij}$, $\boldsymbol{\Lambda}_{ij}$)을 다음과 같이 정의하고자 한다.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$\boldsymbol{\delta}'_{ij} = \begin{pmatrix} \dfrac{\gamma_i}{1-\beta_{i,1}-\alpha_{i,1}} & \dfrac{\gamma_j}{1-\beta_{j,1}-\alpha_{j,1}} \end{pmatrix} \equiv \begin{pmatrix} \delta_{(ij),1} & \delta_{(ij),2} \end{pmatrix},$$

$$\mathbf{G}_{ij} = \begin{pmatrix} \dfrac{\partial \delta_{(ij),1}}{\partial \gamma_i} & \dfrac{\partial \delta_{(ij),1}}{\partial \beta_{i,1}} & \dfrac{\partial \delta_{(ij),1}}{\partial \alpha_{i,1}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \dfrac{\partial \delta_{(ij),2}}{\partial \gamma_j} & \dfrac{\partial \delta_{(ij),2}}{\partial \beta_{j,1}} & \dfrac{\partial \delta_{(ij),2}}{\partial \alpha_{j,1}} \end{pmatrix}$$

그리고

$$\boldsymbol{\Lambda}_{ij} = \begin{pmatrix} \mathrm{Cov}\begin{pmatrix} \gamma_i \\ \beta_{i,1} \\ \alpha_{i,1} \end{pmatrix} & \mathbf{O} \\ \mathbf{O} & \mathrm{Cov}\begin{pmatrix} \gamma_j \\ \beta_{j,1} \\ \alpha_{j,1} \end{pmatrix} \end{pmatrix},$$

여기서, $\mathbf{O}$는 크기가 3×3인 영행렬이다.

만약 두 시계열 자료, $\mathbf{x}_i$, $\mathbf{x}_j$의 무조건부 분산이 같으면$\left(\delta_{(ij),1} = \delta_{(ij),2}\right)$ 다음의 통계량 W_1^{ij}은

$$W_1^{ij} = \left(\mathbf{A}\widehat{\boldsymbol{\delta}}_{ij}\right)' \left[\left(\mathbf{A}\widehat{\mathbf{G}}_{ij}\right) \widehat{\boldsymbol{\Lambda}}_{ij} \left(\mathbf{A}\widehat{\mathbf{G}}_{ij}\right)'\right]^{-1} \left(\mathbf{A}\widehat{\boldsymbol{\delta}}_{ij}\right) \tag{4.1}$$

이며 자유도가 1인 카이제곱분포를 근사적으로 따르게 된다. 따라서, 본 논문에서는 무조건부 분산의 유사성을 측정하는 거리를

$$d_1\left(\mathbf{x}_i, \mathbf{x}_j\right) = 1 - W_1^{ij}\text{로부터의 유의확률} \tag{4.2}$$

로 제안한다. 식 (4.1)의 통계량 W_1^{ij}는 $H_0 : \mathrm{UV}(\mathbf{x}_i) = \mathrm{UV}(\mathbf{x}_j)$를 검정하기 위한 Wald통계량이다. 3.2절의 Otranto (2008)의 첫 번째 가설검정(T1)과의 차이점은 1) 식 (4.1)에서 $\boldsymbol{\delta}_{ij}$와 Otranto (2008)에서의 $\boldsymbol{\delta}_{ij}$는 서로 다르게 정의되며, 2) 전통적인 군집분석 방법에서와 같이 식 (4.2)는 개체들 간의 거리를 쌍별로(pairwise) 먼저 정의한 후 군집분석을 수행하게 된다.

4.2. 구조의 유사성 거리

두 시계열자료 간에 구조에 대한 유사성 거리, d_2는 아래와 같이 정의된다.

$$d_2\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\widehat{\boldsymbol{\theta}}_i^* - \widehat{\boldsymbol{\theta}}_j^*\right)' \left(\widehat{\mathbf{V}}_i^* + \widehat{\mathbf{V}}_j^*\right)^{-1} \left(\widehat{\boldsymbol{\theta}}_i^* - \widehat{\boldsymbol{\theta}}_j^*\right), \tag{4.3}$$

표 4.1: 모의실험에서 고려한 두 가지 시나리오

True Cluster	Series	Scenario 1.				Scenario 2.			
		γ	β	α	UV	γ	β	α	UV
I	1	0.10	0.50	0.00	0.20	0.10	0.50	0.00	0.20
	2	0.10	0.45	0.05	0.20	0.10	0.48	0.02	0.20
	3	0.10	0.40	0.10	0.20	0.10	0.46	0.04	0.20
	4	0.10	0.35	0.15	0.20	0.10	0.44	0.06	0.20
	5	0.10	0.30	0.20	0.20	0.10	0.42	0.08	0.20
	6	0.10	0.25	0.25	0.20	0.10	0.40	0.10	0.20
II	1	0.10	0.80	0.00	0.50	0.10	0.80	0.00	0.50
	2	0.10	0.75	0.05	0.50	0.10	0.78	0.02	0.50
	3	0.10	0.70	0.10	0.50	0.10	0.76	0.04	0.50
	4	0.10	0.65	0.15	0.50	0.10	0.74	0.06	0.50
	5	0.10	0.60	0.20	0.50	0.10	0.72	0.08	0.50
	6	0.10	0.55	0.25	0.50	0.10	0.70	0.10	0.50

여기서, $\widehat{\boldsymbol{\theta}}_i^* \equiv (\widehat{\gamma}_i, \widehat{\beta}_{i,1}, \widehat{\alpha}_{i,1})'$이고 $\widehat{\mathbf{V}}_i^*$는 i번째 시계열 자료에서 얻은 추정량들의 분산-공분산행렬이다. 식 (4.3)과 3.1절의 Caiado와 Crato (2007)의 식 (3.3)과의 차이점은 모수 γ_i, γ_j의 포함여부이다. 또한, 3.1절에서 언급한 바와 같이 Caiado와 Crato (2007)의 논문은 모의실험이 고려되어 있지 않다.

4.3. 모의실험결과

앞에서 언급하였듯이, 임의의 두 시계열 자료들 간에 무조건부 분산에 대한 거리(d_1, 식 (4.2) 참조)와 구조에 대한 거리(d_2, 식 (4.3) 참조)에 기반해서 두 가지 계층적 군집분석 알고리즘(hierarchical clustering algorithm)을 이용한 결과를 설명하고자 한다; 완전연결 알고리즘(complete linkage algorithm)과 평균연결 알고리즘(average linkage algorithm). 다수의 GARCH(1,1)을 따르는 모의자료들이 d_1과 d_2에 따라 얼마나 올바르게 군집화되는 지를 알아보기 위하여 다음의 두 가지 시나리오를 고려하였다. 표 4.1에서 알 수 있듯이, 전체 12개의 GARCH(1,1)모형을 따르는 자료를 임의생성하였다. 각 6개씩의 자료들은 무조건부 분산이 일정하도록 ARCH 모수(β)와 GARCH 모수(α)를 적절히 조정하였다. 즉, 두 시나리오에서 첫 번째 그룹(I)에 속하게 되는 자료들은 무조건부 분산이 0.20이고 두 번째 그룹(II)은 0.50의 무조건부 분산을 갖도록 하였다. 두 시나리오들 간의 차이점은 고정된 모수, γ를 제외한 나머지 두 모수들을 통해 알 수 있듯이 두 번째 시나리오가 첫 번째 시나리오에 비해 그룹 내의 자료들이 보다 더 동질적(homogeneous)이라고 할 수 있다. 따라서 동질성(homogeneity)이 강한 자료들로부터 구조의 동일성을 평가할 수 있는 T3의 성능을 알아 볼 수 있다. 각 자료의 표본크기(T)는 500, 1000, 5000을 고려하였고 전체 반복수는 100으로 고정하였으며 두 거리들에 의해 미리 선언된 두 군집들로 얼마나 잘 그룹화가 되는 지를 경험적 정분류율을 통해 알아보았다. 표 4.2는 시나리오와 표본크기에 따라 전체 반복수 중에서 미리 선언된 두 군집들(I, II)로 제대로 그룹화가

표 4.2: 모의실험 하에서의 경험적 정분류율(%)

Scenario	size	Complete linkage		Average linkage	
		T1(d_1)	T3(d_2)	T1(d_1)	T3(d_2)
1.	500	45	27	41	28
	1000	89	38	87	45
	5000	100	38	100	87
2.	500	41	44	40	45
	1000	72	74	73	76
	5000	100	100	100	100

된 횟수가 얼마나 되는 지를 의미하는 경험적 정분류율을 제시하고 있다. 우선 두 계층적 군집분석 알고리즘에 상관없이 거의 동일한 결과를 보이고 있다. 즉, 시나리오에 상관없이 대체로 표본크기가 증가할수록 올바르게 군집화하는 확률이 증가함을 알 수 있다. 다만, 첫 번째 시나리오의 표본크기 5000인 경우에 완전연결방법에 의한 T3의 결과가 표본크기 1000인 경우의 결과와 변화가 없고 평균연결방법에 의한 결과와는 다소 차이가 있다. 그리고 무조건부 분산의 거리를 이용한 T1은 어떠한 군집분석 알고리즘을 사용하더라도 완벽하게 그룹화함을 알 수 있다. 두 번째 시나리오 하에서는 구조의 동질성 역시 표본크기가 증가할수록 올바르게 그룹화함을 알 수 있고 표본크기 5000인 경우 완벽한 군집화가 이루어지고 있다.

5. 사례분석

4절에서의 모의실험을 근거로 이 연구에서 사용하고 있는 군집화 접근방법들의 성능은 충분하다고 판단되어진다. 따라서, 실제 이분산 시계열자료들을 이용하여 무조건부 분산의 동질성 및 구조 자체의 동질성을 측정할 수 있는 거리들로 군집분석을 수행하고자 한다. 모의실험에서와 같이 두 가지 군집분석 알고리즘을 사용하였으나 결과가 유사하여 완전연결 알고리즘의 결과를 설명하고자 한다. 실제자료는 2005년 1월 3일부터 2007년 12월 31일까지의 총 796일 동안 관찰된 국내 11개 주가(강원랜드, 삼성엔지니어링, 웅진코웨이, 제일기획, 신한지주, 국민은행, 외환은행, 현대건설, 경남기업, 대림산업, GS건설)의 수익율이다. 군집분석에 이용할 ARCH 또는 GARCH 모수의 추정단계 이전에 관측자료의 7일 전까지의 값들을 이용하여 자기회귀(AR(7))모형을 고려하였다. 따라서 각 시계열 자료의 모형은 AR(7) - GARCH(2,2)로 고려하였으며 군집분석에 관련된 모수은 γ, β_1, β_2, α_1, α_2와 이들의 분산공분산행렬이다.

그림 5.1은 군집분석에 사용된 각 시계열 자료의 수익률의 시도표이다. 단순히 시도표를 통해 조건부 이분산의 존재 유무 혹은 정도에 대한 정보를 도출하기는 쉽지 않은 일일 것이다. 하지만, 탐색적인 관점에서 본다면, 즉, 분산이라는 잘 알려진 변이의 척도를 시도표에 나타난 자료의 흐름에 비추어 본다면, 강원랜드, 제일기획, 신한지주, 국민은행 등이 다른 기업들에 비해 상대적으로 작은 변동성을 보이고 있다. 그리고 전체 11개 기업들 중에서 경남기업이 가장 큰 변동성을 보이고 있음을 알 수 있다. 그림 5.1에는 각 시계열

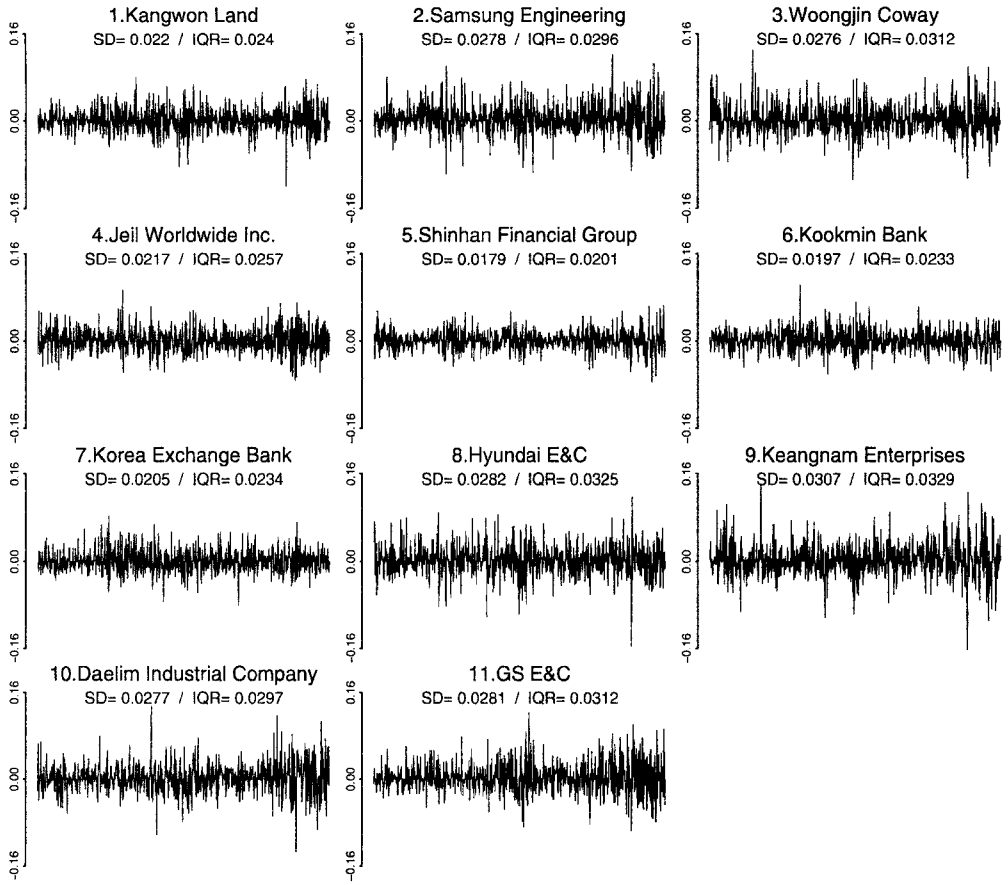


그림 5.1: 국내 11개 기업의 수익률에 대한 시도표.

자료의 변동성에 대한 통계량들을 제시하고 있는 데, 바로 표본표준편차(sample standard deviation: SD)와 사분위범위(interquartile range: IQR)이다. 이 값들로 동질적인 무조건부 분산을 가지는 기업들의 그룹화도 탐색적인 관점에서 가능하리라 판단된다. 하지만, 앞에서 언급한 구조의 동질성에 대한 군집분석의 결과는 그림 5.1의 시도표로는 판가름하기 쉽지 않다.

GARCH(2,2)의 초기모형을 근간으로 Akaike의 정보기준(Akaike information criterion: AIC)과 모수추정값의 유의확률 하에서의 최적모형을 각 시계열 자료에 적합한 결과가 표 5.1에 제시되어 있다. 추정된 5개의 모수 추정값들로부터 다음의 식을 이용하여 무조건부 분산(UV) 값을 구하게 된다.

$$\mathrm{Var}(X_t) = \frac{\gamma}{1 - \sum_{p=1}^{P} \beta_p - \sum_{q=1}^{Q} \alpha_q}.$$

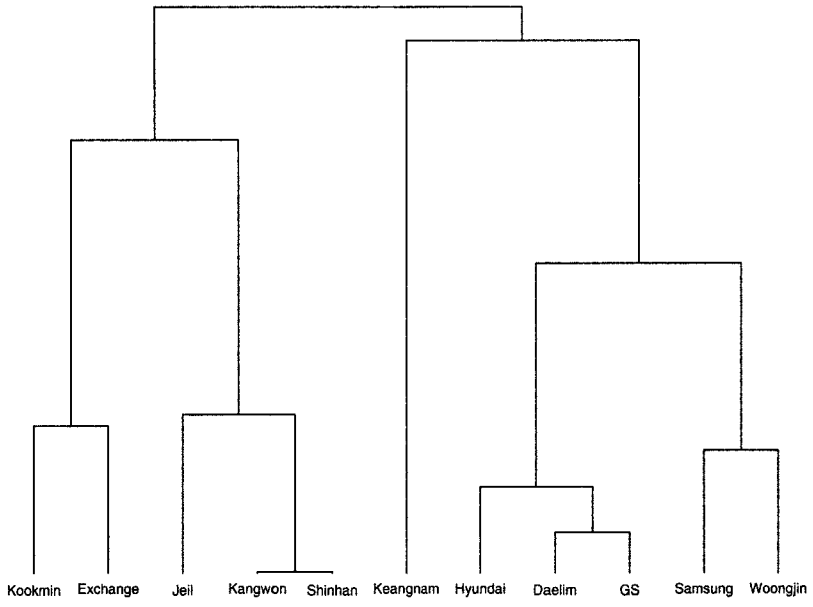
표 5.1: 시계열자료의 최적모형 및 모수추정값과 유의확률(†: $a^b \equiv a \times 10^{-b}$)

Series	$\gamma^\dagger$	ARCH parameters		GARCH parameters		UV†
		β_1	β_2	α_1	α_2	
강원랜드	2.83^5(0.013)	0.033(0.003)	0.097(0.001)		0.813(0.001)	4.96^4
삼성엔지니어링	7.34^5(0.003)	0.110(0.001)		0.794(0.001)		7.65^4
웅진코웨이	6.65^5(0.001)	0.029(0.006)	0.068(0.001)		0.814(0.001)	7.47^4
제일기획	1.72^5(0.022)		0.059(0.001)	0.904(0.001)		4.65^4
신한금융지주	5.50^6(0.100)	0.087(0.001)	0.036(0.039)		0.856(0.001)	2.62^4
국민은행	3.68^5(0.014)	0.137(0.001)			0.768(0.001)	3.87^4
외환은행	1.01^5(0.010)	0.029(0.007)	0.946(0.001)			4.04^4
현대건설	2.94^4(0.001)	0.168(0.001)	0.462(0.001)			7.95^4
경남기업	8.10^5(0.001)	0.087(0.001)	0.825(0.001)			9.20^4
대림산업	1.19^5(0.030)	0.057(0.001)	0.929(0.001)			8.50^4
GS건설	7.42^6(0.114)		0.055(0.001)	0.937(0.001)		9.28^4

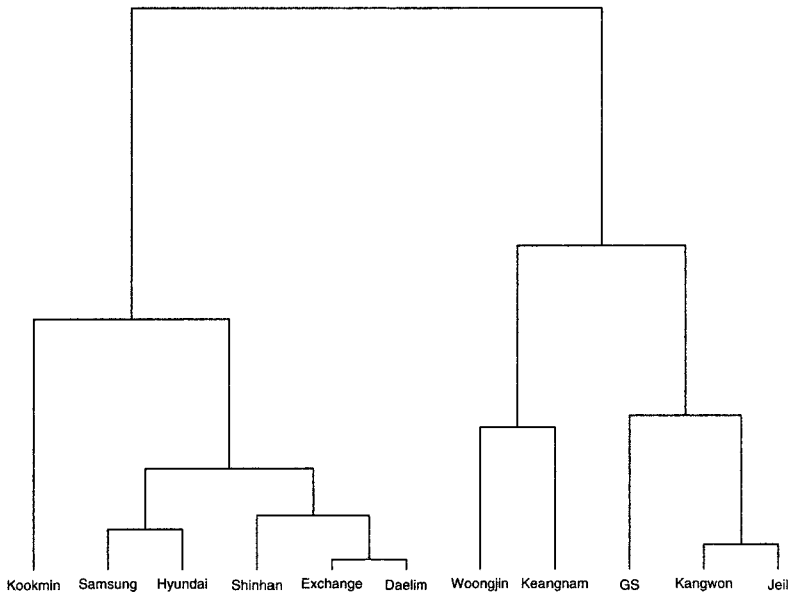
각 시계열자료의 적합된 최적모형으로부터 무조건부 분산의 값을 도출하고 이를 비교하면, 시도표로부터 얻은 자료의 변동성에 대한 시각적인 탐색 결과와 상당히 유사하다는 것을 알 수 있다. 즉, 무조건부 분산이 0.0005보다 작은 그룹과 큰 그룹으로 구분되고 변동성이 상대적으로 작은 그룹에는 강원랜드, 제일기획, 신한금융지주, 국민은행 그리고 외환은행이 포함되고, 변동성이 상당히 큰 기업들은 경남기업, 대림기업 그리고 GS건설 등으로 나타났다. 따라서, 무조건부 분산의 군집분석 결과는 대략 두 개 정도의 그룹화로 나타날 것으로 예상할 수 있다.

먼저 무조건부 분산의 동질성(혹은 유사성)에 대한 군집분석에 대해 설명하겠다. 식 (4.1)에 근거한 식 (4.2)의 거리를 이용하여 완전연결 알고리즘에 근거해 계층적 군집분석을 수행한 덴드로그램이 그림 5.2(a)에 제시되어 있다. 군집화의 전반적인 형태는 그림 5.1과 표 4.1에서 도출된 결과와 상당히 일치함을 알 수 있다. 즉, 무조건부 분산의 값이 작은 기업들(강원랜드, 신한금융지주, 제일기획, 국민은행, 한국외환은행)과 큰 기업들(현대건설, 대림산업, GS건설, 삼성엔지니어링, 웅진코웨이, 경남기업)로 그룹화가 되었음을 알 수 있다. 보다 구체적으로 설명하면, 강원랜드와 신한금융지주가 제일 유사한 무조건부 분산을 가지고 있고 대림산업과 GS건설이 그 다음의 순으로 나타났으며 경남기업은 다른 기업들과 무조건부 분산의 유사성이 제일 낮은 것으로 나타났다.

표 5.2(b)는 구조의 유사성 거리를 이용한 군집분석 덴드로그램이다. 우선 대림산업과 한국외환은행이 가장 유사한 일반화 조건부 이분산 구조를 가지고 있고 강원랜드와 제일기획이 그 다음의 순으로 나타났다. 그림 5.2(a)에 나타난 군집분석 결과와 비교할 때 서로 다른 그룹에 속했던 기업들이 구조의 유사성에 의한 군집분석에서는 매우 유사한 것으로 나타났다. 대림산업과 한국외환은행이 대표적인 경우인데, 실제로 표 5.1에서 알 수 있듯이 AIC에 의한 최적모형이 ARCH(2)모형으로 동일하고 각 모수의 추정값들 역시 상당히 유사함을 알 수 있다. Otranto (2008)의 연구와는 달리 축차적이 아닌, 단순히 두 기업 간의 거리(pairwise distance)만을 고려하고(가설검정을 수행하지 않고) 군집분석을 수행하였으므로 T3과 T1의 관련성(그룹 세분화)이 이 연구에서는 고려되지 않았다.



(a) 무조건부 분산의 유사성 거리(d_1)



(b) 구조의 유사성 거리(d_2)

그림 5.2: 완전연결 알고리즘을 이용한 군집분석 덴드로그램

6. 결론

이 연구에서는 조건부 이분산 구조를 가지는 시계열 자료의 군집분석에 대해 알아보았다. 등분산 구조 하에서의 시계열 군집화는 비교적 많은 연구가 이루어졌고 최근 Caiado와 Crato (2007) 그리고 Otranto (2008)는 무조건부 분산의 유사성 및 구조의 유사성을 측도화하여 일반화 자기회귀 조건부 이분산(GARCH)모형을 따르는 시계열 자료의 군집분석을 시도하였다. 본 논문에서는 위 연구들을 기반으로 모의실험과 더불어 국내 기업들의 최근 3년 일일 주가자료를 이용하여 실제 군집분석을 수행하였다.

위 연구들은 GARCH모형 만을 고려하였으나 추후 시계열 자체의 인과관계를 함께 포함하는 ARMA-GARCH모형에 대한 연구가 필요하리라 판단된다.

참고문헌

Bauwens, L. and Rombouts, J. V. K. (2007). Bayesian clustering of many GARCH models, *Econometric Reviews,* **26**, 365–386.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics,* **31**, 307–327.

Caiado, J. and Crato, N. (2007). A GARCH-based method for clustering of financial time series: International stock markets evidence, *Munch Personal RePEc Archive,*

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica,* **50**, 987–1008.

Fan, J. and Yao, Q. (2005). *Nonlinear Time Series: Nonparametric and Parametric methods,* Springer-Verlag, New York.

Liao, T. W. (2005). Clustering of time series data—a survey, *Pattern Recognition,* **33**, 1857–1874.

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures, *Computational Statistics & Data Analysis,* **52**, 4685–4698.

Taylor, S. J. (1986). *Modelling Financial Time Series,* John Wiley & Sons, New York.

[2008년 9월 접수, 2008년 9월 채택]

Clustering Korean Stock Return Data Based on GARCH Model

Man Sik Park[1], Na Young Kim[2], Hee-Young Kim[3]

Abstract

In this study, we considered the clustering analysis for stock return traded in the stock market. Most of financial time-series data, for instance, stock price and exchange rate have conditional heterogeneous variability depending on time, and, hence, are not properly applied to the autoregressive moving-average(ARMA) model with assumption of constant variance. Moreover, the variability is front and center for stock investors as well as academic researchers. So, this paper focuses on the generalized autoregressive conditional heteroscedastic(GARCH) model which is known as a solution for capturing the conditional variance(or volatility). We define the metrics for similarity of unconditional volatility and for homogeneity of model structure, and, then, evaluate the performances of the metrics. In real application, we do clustering analysis in terms of volatility and structure with stock return of the 11 Korean companies measured for the latest three years.

Keywords: Generalized autoregressive conditional heteroscedasticity; conditional variance; clustering analysis.

1) Research professor, Department of Biostatistics & Department of Preventive Medicine, Medical Research Center for Environmental Toxico-Genomics and Proteomics, College of Medicine, Korea University, 126-1 Anam-Dong, Sungbuk-Gu, Seoul 136-705, Korea.

2) Deputy general manager, Marketing Department, Private Banker Business Div. Samsung Securities Co., LTD., 6 Jongno 2-Ga, Jongno-Gu, Seoul, 110-789, Korea.

3) Research professor, Department of Biostatistics & Department of Preventive Medicine, Medical Research Center for Environmental Toxico-Genomics and Proteomics, College of Medicine, Korea University, 126-1 Anam-Dong, Sungbuk-Gu, Seoul 136-705, Korea.
Correspondence: starkim@korea.ac.kr