일반화 자기회귀 조건부 이분산 모형들의 베이지안 군집분석에 관한 연구: 주식자료에 응용

유희경*, 정수정**, 성경***

요약

주식시장에서의 주식 데이터들은 GARCH(p,q)모델이 될 수 있는데 본 논문에서는 GARCH(1,1) 모델이라 가정하였다. 본 논문은 각 주식데이터를 GARCH(1,1) 모델을 따르는 것으로 가정하였고 각 군집은 같은 모델과 같은 모수를 갖는 몇 개의 그룹로 나누는 모형기반 군집분석 방법을 보였다. 군집의 수를 정하는 방법으로 BIC(베이지안 정보기준)를 이용하였고 각 군집의 특성을 알기위한 모수들은 베이지안 접근방법으로 추정하였다.

The Study of Bayesian Clustering of Generalized Autoregressive Conditional Heteroscedasitcity

Hee-Kyung Yoo*, Su Jeong Jeong**, Kyung Sung***

ABSTRACT

There are many data on stocks in the stock market. They can be characterized by GARCH (p,q) models, specifically in this paper we assume that each stock data follows GARCH(1,1) model. This paper presents a model-based clustering method of stock data into several groups where each group has the same model and the same parameters. For the choosing the number of groups, we exploit the BIC (Bayesian Information Criterion). And the group parameters which present the characteristics of groups are estimated through Bayesian approach.

Key Words : Regression, Bayesian, Cluster Analysis, Clustering, Autoregression

* 강원대학교 컴퓨터공학과(✉hkyoo@kangwon.ac.kr)
** (주)시벅스테크놀러지
*** 목원대학교 컴퓨터교육과
· 제1저자(First Author) : 유희경 · 교신저자(Correspondent Author) : 유희경
· 접수일(2012년 11월 21일), 수정일(1차 : 2012년 12월 11일), 게재확정일(2012년 12월 18일)

Ⅰ. 서론

투자자들의 효과적인 거래를 위해서는 변동성(volatility)을 추정하고 예측하는 것이 투자의사결정에 있어서 상당히 중요한 역할을 한다. 변동성이라는 것은 분산을 지칭하는 것으로 종종 위험(risk)를 측정하는 수단으로 간주된다. 이러한 변동 현상이 나타나는 이유는 미래값의 분산이 현재의 상황에 의존함을 의미한다.

이러한 위험은 자기회귀 조건부 이분산(autoregressive conditional heteroscedasitcity:

ARCH) 모형을 이용하여 분석할 수 있는데 본 논문에서는 Bollerslev(1986)에 의해 일반화된 자기회귀 조건부 이분산 (generalized autoregressive conditional heteroscedasitcity: GARCH) 모형으로까지 확장하여 분석해 보았다. 조건부 이분산모형을 이용한 시계열의 군집분석은 주식시장에서 거래되는 다수의 주식종목의 수익률 패턴이 유사하거나 위험도가 비슷한 종목들로 그룹화 할 때 유용하게 이용될 수 있다.

이를 통하여 투자자들에게 투자의사결정을 하도록 도움을 줄 수 있을 것이다. 본 연구에서는 국내 시계열 자료들을 가지고 베이지안 방법을 이용한 군집분석을 연구하고자 한다.

II. 본론

2.1 베이지안 추론

2.1.1 베이지안 분포

y는 관측된 데이터, θ는 사전분포(prior distribution) 라고 할 때 베이즈 정리에 의해 θ에 대한 사후분포(posterior distribution)를 나타내면 다음과 같다.

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)} \quad (1)$$

여기서

$$p(y) = \begin{cases} \sum_{\theta} p(y|\theta)\pi(\theta) & \text{if } \theta\text{가 이산인 경우} \\ \int p(y|\theta)\pi(\theta)d\theta & \text{if } \theta\text{가 연속인 경우} \end{cases}$$

이다. 이때 (1)은 (2)와 같이 표현할 수 있다.

$$p(\theta|y) \propto p(y|\theta)\pi(\theta) \quad (2)$$

2.1.2 베이지안 추정법

복잡한 사후분포에서 표본을 추출은 MCMC (Markov Chain Monte Carlo)라고 하는 모의 실험으로 상호종속적인 확률변수를 생성하여 해결할 수 있다. MCMC는 직접적으로 사후분포에서 표본생성이 불가능하거나 생성가능해도 시간과 경비가 허락하지 않을 경우 임의의 분포에서 생성하여 그것을 사후분포의 표본으로 사용할 수 있는 방법이다.

여기서는 MCMC 방법 중 잘 알려진 깁스 샘플러(Gibbs sampler)알고리즘에 대해 살펴보겠다.

1) 깁스 샘플러(Gibbs sampler)

깁스 샘플러는 다차원의 결합 확률분포가 복잡하여 직접 랜덤표본을 생성하기 어려운 경우 각 변수의 조건부 확률분포로부터 랜덤표본을 반복적으로 생성하면 적절한 조건 하에서 이들의 극한분포가 결합확률밀도함수가 된다는 사실에 근거하여 난수를 생성하는 방법이다. 깁스 샘플러의 알고리즘은 다음과

같다.

Step 1. 초기값 $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_p^{(0)})^T$을 선택한다.

Step 2. k를 1로 놓는다.

Step 3. 각각의 완전 조건부 분포로부터 한 번에 한 개씩 확률표본을 생성한다.

$$\theta_1^{(k)} \sim p(\theta_1 | y, \theta_2^{(k-1)}, \cdots, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \sim p(\theta_2 | y, \theta_1^{(k)}, \theta_3^{(k-1)}, \cdots, \theta_p^{(k-1)})$$

$$\vdots$$

$$\theta_p^{(k)} \sim p(\theta_p | y, \theta_2^{(k)}, \theta_3^{(k)}, \cdots, \theta_{p-1}^{(k)})$$

Step 4. k에 1을 더하고 Step 3 으로 돌아간다.

이러한 과정을 k번 반복하면 깁스 표본 $\boldsymbol{\theta}^G = (\theta_1^{(k)}, \cdots, \theta_p^{(k)})$를 얻을 수 있으며, 전체 m번 반복하여 깁스 표본 $\boldsymbol{\theta}_1^G, \boldsymbol{\theta}_2^G, \cdots, \boldsymbol{\theta}_m^G$를 얻을 수 있다.

2.2 이분산 시계열 모형

2.2.1 ARCH(p) 모형

Engle(1982)은 오차의 이분산성을 모형화하기 위해서는 자기회귀 조건부 이분산(autoregressive conditional heteroscedasitcity: ARCH) 모형을 이용한다. ARCH 모형은 오차항의 조건부 이분산이 시차를 갖는 오차항들의 제곱들의 선형 결합식으로 표현된다. ARCH(p) 모형은 다음과 같다.

$$n_t = \sigma_t e_t \qquad (3)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 n_{t-1}^2 + \cdots + \alpha_p n_{t-p}^2$$

여기서 e_t는 평균이 0이고 분산이 1인 $i.i.d.$ 확률변수이고, n_{t-i}와는 독립이다.

2.2.2 GARCH(p,q) 모형

Bollerslev(1986)는ARCH(p) 모형을 일반화 시켜 오차 항들의 제곱뿐만 아니라 시차를 갖는 조건부 이분산에 의하여 오차의 분산이 설명되어진다고 가정한 모형이 이를 일반화된 조건부 이분산(generalized autoregressive conditional heteroscedasitcity: GARCH) 모형이다.

GARCH(p,q) 모형은 다음과 같다.

$$n_t = \sigma_t e_t \qquad (4)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i n_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

여기서 e_t는 평균이 0이고 분산이 1인 $i.i.d.$ 확률변수이고, n_{t-i}와는 독립이다.

본 연구에서는 GARCH(1,1) 모형을 이용하여 군집분석을 하고자 한다. 위 모형에서 추정해야할 모수들은 $(\theta = \alpha_0, \alpha_1, \beta_1)$이다.

2.3 GARCH(1,1) 모형들의 모형기반 군집분석

본 연구의 목적은 J개의 시계열 자료를 작은 G개의 그룹으로 나누는 것이다. 그렇다면 같은 그룹에 속한 시계열 자료들은 동일한 그룹의 모수를 가진다고 할 수 있다. 그러므로 각 그룹의 모형은 각각의 그룹 모수 벡터 $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_G)$로 나타낼 수 있다($\theta_g = (\alpha_{0g}, \alpha_{1g}, \beta_g)$, $g = 1, \cdots, G$).

2.3.1 자료의 구조

관측값 y_t^j는 $t(t=1, \cdots, T_j)$시점에서 관측된 $j(j=1, \cdots, J)$번째 시계열 자료를 의미한다.

2.3.2 GARCH(1,1) 모형들의 모형기반 군집분석

모형을 기반으로 군집분석하기 위해 일반적으로 혼합모형을 사용한다(Banfield and Raftery, 1993; McLachlan and Peel, 2000; Fraley and Raftery, 2002; Zhong and Chosh, 2003; Fruhwirth-Schnatter, 2006). 혼합 모형을 공식화하기 위해서는 $\boldsymbol{y} = (y^1, y^2, \cdots, y^J)$가 G개의 군집으로부터 나온 것이라 가정한다. 그러면 각 군집의 자료는 군집의 특정 모수 $\theta_g(\boldsymbol{\theta}= (\theta_1, \theta_2, \cdots, \theta_G))$를 가진 $p(y^j|\theta_g)$ 분포로부터 생성된다.

각 그룹을 따르는 자료들의 확률은 각 그룹의 크기에 따라 나타나며 $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_G)$ $(\eta_g \geq 0, \sum_{g=1}^{G} \eta_g = 1)$로서 표기되어진다.

따라서 $\boldsymbol{\eta}$를 알고 있을 때 y^j의 혼합분포는 다음과 같이 주어진다.

$$p(y^j|\boldsymbol{\eta}, \boldsymbol{\theta}) = \sum_{g=1}^{G} \eta_g p(y^j|\theta_g) \quad (5)$$

이 때의 유한한 혼합모형 우도함수는 다음과 같다.

$$p(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\theta}) = \prod_{j=1}^{J} \left(\sum_{g=1}^{G} \eta_g p(y^j|\theta_g) \right) \quad (6)$$

또 다른 혼합모형의 표현은 관측되지 않은 그룹 지시함수(group indicator)인 잠재변수 $\boldsymbol{S} = (S_1, \cdots, S_J)$를 구하는 것이다. 만약 $S_j = g$라면 j번째$(j=1, \cdots, J)$자료가 그룹 g $(g=1, \cdots, G)$를 따르는 것을 의미한다. 따라서 y^j가 그룹 g에 포함될 확률은 $P(S_j = g) = \eta_g$이다. 이때의 우도함수를 구하면 다음과 같이 나타낼 수 있다.

$$p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{\eta}, \boldsymbol{\theta}) = \prod_{g=1}^{G} \eta_g p(y^j|\theta_g) \quad (7)$$

여기서 $\boldsymbol{S} = (S_1, \cdots, S_J)$는 다항분포로부터 구할 수 있다.

1) 사전분포(prior density)

혼합모형 (7)식 의 모수 $\boldsymbol{\eta}$와 $\boldsymbol{S}$는 추정되어져야 하며, 우리가 최종적으로 추정해야 할 모수들을 $\psi = (\boldsymbol{S}, \boldsymbol{\eta}, \boldsymbol{\theta})$로 나타내자. 따라서 사전분포는 다음과 같이 구할 수 있다.

$$\varphi(\psi) = \varphi(\boldsymbol{S}|\boldsymbol{\eta})\varphi(\boldsymbol{\eta})\varphi(\boldsymbol{\theta}) \quad (8)$$

여기서

$$\varphi(\boldsymbol{S}|\boldsymbol{\eta}) = \prod_{j=1}^{J} P(S_j = g) = \prod_{g=1}^{G} \eta_g^{x_g} \quad (9)$$

$$\varphi(\boldsymbol{\theta}) = \prod_{g=1}^{G} \varphi(\theta_g) \quad (10)$$

이며, $x_g = \#(S_j = g)$ 즉, $S_j = g$인 총 개수로 정의한다.

2) 자료의 확률분포(likelihood)

관측값 y_t^j들은 다음과 같이 t분포를 따른다고 가정한다.

$$y_t^j = \epsilon_t^j \sqrt{h_t^j} \quad (11)$$
$$\epsilon_t^j \mid I_{t-1}^j \sim student(0, 1, \nu)$$
$$h_t^j = \alpha_{0g} + \alpha_{1g}(y_{t-1}^j)^2 + \beta_g h_{t-1}^j$$

이때 관측값 y_t^j들이 g그룹을 따를 경우, 모든 y_t^j에 대한 우도함수는 다음과 같이 나타낼 수 있다.

$$\prod_{j=1}^{J}\prod_{t=1}^{T_j} f(y_t^j \mid \theta_{S_j}) = \prod_{j=1}^{J} f(y^j \mid \theta_{S_j}) \quad (12)$$

3) 사후분포(posterior distribution)

모수 ψ에 대한 사후분포는 데이터의 확률분포와 ψ의 사전분포로부터 다음과 같이 나타낼 수 있다.

$$\varphi(\psi \mid \boldsymbol{y}) \propto \varphi(\boldsymbol{\eta})\prod_{g=1}^{G}\varphi(\theta_g)\prod_{j=1}^{J}\eta_{s_j} f(y^j \mid \theta_{s_j}) \quad (13)$$
$$= \varphi(\boldsymbol{\eta})\prod_{g=1}^{G}\eta_g^{x_g}\varphi(\theta_g)\prod_{j=1}^{J} f(y^j \mid \theta_{s_j})$$

여기서 $\varphi(\theta_g)$는 특정 모형에 의존한다.

2.3.3 모수추론을 위한 깁스 표집법

모수(ψ)는 직접적으로 추정하기엔 어려움이 있으므로 사후분포를 다음에 나오는 3개의 블록으로 나눈 후 이를 MCMC 방법을 이용하여 모수를 추정하기로 한다. 이는 3개의 블록들이 정상조건에 만족할 때까지 반복 시행한다.

1) $\varphi(\boldsymbol{S} \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y})$로부터 $\boldsymbol{S}$ 추출

$\boldsymbol{\eta}$, $\boldsymbol{\theta}$, $\boldsymbol{y}$가 주어졌을 때, S_j들은 서로 상호독립이다. 식 (9)와 (13)를 이용하여 다음과 같이 나타낼 수 있다.

$$\varphi(S_1, \cdots, S_J \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y}) \propto \prod_{j=1}^{J} f(y^j \mid \theta_{s_j})\varphi(S_j \mid \boldsymbol{\eta})$$
$$= \varphi(S_1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y})\varphi(S_2 \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y}) \cdots \varphi(S_J \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{y})$$

$\boldsymbol{S} = (S_1, \cdots, S_J)$는 다항과정(multinomial proccess)과 같으므로 이는 다음과 같은 확률을 가지는 이산형 분포로부터 얻을 수 있다.

$$P(S_j = g \mid \boldsymbol{\theta}, \boldsymbol{\eta}, y^j) = \frac{f(y^j \mid \theta_g)\eta_g}{\sum_{l=1}^{G} f(y^j \mid \theta_l)\eta_l}, \quad g = 1, \cdots, G \quad (15)$$
$$\propto f(y^j \mid \theta_g)\eta_g$$

2) $\varphi(\boldsymbol{\eta} \mid \boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{y})$로부터 $\boldsymbol{\eta}$ 추출

식 (13)로부터 $\boldsymbol{\eta}$를 구하기 위해서 다음과 같이 표현한다.

$$\varphi(\boldsymbol{\eta} \mid \boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{y}) = \varphi(\boldsymbol{\eta} \mid \boldsymbol{S}) \propto \varphi(\boldsymbol{\eta}) \prod_{g=1}^{G} \eta_g^{x_g} \quad (16)$$

사전분포는 모수가 $a_0 = (a_{10}, \cdots, a_{G0})'$ $(a_{i0} > 0,\ i = 1, \cdots, G)$인 디리슈레 분포(Dirichlet distribution)를 선택한다. 그 결과 $\varphi(\boldsymbol{\eta} \mid \boldsymbol{S})$는 모수가 $a_g = (a_1, \cdots, a_G)'$ $(a_g = a_{g0} + x_g,$ $g = 1, \cdots, G)$인 디리슈레 분포를 따르게 된다.

3) $\varphi(\boldsymbol{\theta} \mid \boldsymbol{S}, \boldsymbol{\eta}, \boldsymbol{y})$로부터 θ 추출

모수벡터 $\boldsymbol{\theta}$는 각각의 그룹 모수 θ_g들 사이가 독립이므로 다음과 같은 식으로 나타낼 수 있다.

$$\varphi(\boldsymbol{\theta}|\boldsymbol{S},\boldsymbol{\eta},\boldsymbol{y}) = \varphi(\boldsymbol{\theta}|\boldsymbol{S},\boldsymbol{y}) = \varphi(\theta_1|\tilde{y}^1)\cdots\varphi(\theta_G|\tilde{y}^G) \quad (17)$$

이 때, $\varphi(\theta_g|\tilde{y}^g) \propto \varphi(\theta_g)\prod_{j\in J_g} f(y^j|\theta_{S_j})$, $J_g = \{j|S_j = g\}$이며, $\tilde{y}^g$는 g 그룹에 속하는 시계열 자료들을 의미한다. $\varphi(\theta_g)$는 특정 모형에 의존하기 때문에 관심 있어 하는 모형에 대해 적절하게 명시해줘야 한다.

2.3.4 R 페키지 mclust에 의한 군집개수 및 공분산 모형 선택

mclust 패키지에서는 다양한 공분산 모형을 제공하고 있다(Fraley and Raftery, 1998). 제공된 공분산 모형은 <표 1>과 같이 나타난다.

표 1. mclust에서 제안된 공분산 모형
Table 1 Covariance model proposed in mclust

Identifier	Distribution	Volume	Shape	Orientation
EI	구형	동일함	동일함	없음
VI	구형	다름	동일함	없음
EEI	대각형	동일함	동일함	좌표축
VEI	대각형	다름	동일함	좌표축
EVI	대각형	동일함	다름	좌표축
VVI	대각형	다름	다름	좌표축
EEE	타원형	동일함	동일함	동일함
EEV	타원형	동일함	동일함	다름
VEV	타원형	다름	동일함	다름
VVV	타원형	다름	다름	다름

공분산 행렬로부터 형태(shape), 퍼짐의 정도(volume), 방위(orientation)들을 알 수 있다. 모형의 식별자(Identifier)는 모형의 기하학적 특성을 코드화한 것이다. 예를 들어 EEV 모형은 모든 군집의 부피가 동일(E:equal)하고, 모든 군집의 모양 또한 동일(E:equal)하며, 군집간에 변동(V:variable)이 있다는 것을 의미한다.

2.4. 실증분석

2.4.1 KOSPI50의 자료 설명

모형기반 군집분석을 통하여 군집모형을 GARCH(1,1)모형으로 나타내기 위해 사용된 자료는 KIS-value Plus로부터 제공된 KOSPI 50 주가지수의 일별 종가자료이다.

관측기간은 2005년 1월 1일부터 2009년 7월 31일까지 총 1137일이다. 본 연구에서는 KOSPI 50 자료 중 관측기관 동안 결측값이 있는 종목을 제외한 45개의 KOSPI 자료를 가지고 아래와 같이 계산한 연속복리 수익률을 분석에 사용하였다. R페키지 mclust 모듈은 모형기반 군집분석을 하기 위한 것으로, 자료가 결측이 있을 경우 프로그램이 실행되지 않기 때문에 결측이 있는 종목은 제외하였다.

$$R_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \quad (18)$$

여기서 R_t는 t시점의 주가수익률이고, P_t는 t시점의 종가지수를 나타낸다. 각 시계열 자료는 GARCH(1,1) 모형으로 고려하였다.

2.4.2 분석 결과

모형기반 군집분석(이하 mclust)은 데이터에 대한 베이즈 정보기준(BIC: Bayesian InformationCriterion)

을 통해 가장 최적의 모형과 군집의 수를 도출해 낼 수 있다. 이는 R페키지 mclust 모듈을 통해 분석 할 수 있으며, BIC값이 큰 경우에 모형과 군집의 수가 적절하다는 것을 뒷받침해준다. mclust 모듈을 이용하여 각각의 모형들을 통해 얻은 BIC 결과는 다음과 같다.

표 2. 각 모형에 따른 BIC
Table 2. Each model according to the BIC

Group	EI	VI	EEI	VEI	EVI	VVI
1	226864.2	226864.2	-848513.2	-848513.2	-848513.2	-848513.2
2	227429.7	236189.8	-846310.3	236020.2	-813283.0	-846726.5
3	227520.5	237850.0	-844674.0	236657.8	-760161.7	-846941.0
4	226906.3	245352.9	-845041.8	243256.9	-725056.5	NA
5	225109.6	243148.9	-843989.5	232114.5	-667241.1	-2917445.7
6	223463.2	241408.5	-842381.2	230253.7	-616154.0	NA
7	221865.3	241055.2	-842220.1	228525.4	-570214.6	NA
8	220293.0	*247553.3*	-839765.9	226711.4	-528199.7	NA
9	218600.2	244884.2	-838188.7	224358.6	-466562.3	NA

<표 2>를 통해 VI모형(군집의 부피가 서로 다르지만, 군집의 모양은 동일한 모형)에서 BIC값이 247553.3으로 가장 크게 나타났으며, 8개 군집을 가지는 것이 적절하다는 결과를 얻을 수 있다. 따라서 8개의 군집으로 분류한 결과 [표 3]과 같이 나타났다.

군집1은 S-Oil, 현대차, 신세계, POSCO, 삼성전자, 현대모비스, KT&G, KT, 강원랜드, 한국전력, 삼성화재, SK텔레콤, 한국가스공사와 같은 기업들이 하나의 군집으로 이루어졌고, 군집2는 기아차, 하이닉스, 삼성물산, 삼성증권, OCI, 대한항공, LG, SK, KCC, 현대제철, 대우증권, 대우인터내셔널, LG디스플레이, LG전자, LG화학, GS와 같은 기업으로 이루어졌다. 군집3은 주업종이 건설인 기업으로 현대건설, GS걸설, 현대산업, 대우건설로 이루어졌으며, 군집6은 주업종이 은행·금융인 기업으로 신한지주, 우리금융, 외환은행, 기업은행으로 이루어졌다. 군집8은 주업종이 기계인 기업으로 두산인프라코어, 삼성중공업, 현대중공업, 두산중공업, 대우조선해양으로 이루어졌다. 군집4는 현대상선, 군집5는 SK네트웍스, 군집7은 NHN으로 각각 하나의 기업으로 군집이 이루어졌다.

표 3. 군집 분류표
Table 3. Clustering classification table

군집1	S-Oil(7), 현대차(13), 신세계(16), POSCO(17), 삼성전자(18), 현대모비스(21), KT&G(26), KT(29), 강원랜드(30), 한국전력(36), 삼성화재(37), SK텔레콤(39), 한국가스공사(44)
군집2	기아차(1), 하이닉스(2), 삼성물산(4), 삼성증권(5), OCI(6), 대한항공(10), LG(11), SK(12), KCC(14), 현대제철(15), 대우증권(20), 대우인터내셔널(24), LG디스플레이(27), LG전자(28), LG화학(33), GS(34)
군집3	현대건설(3), GS건설(19), 현대산업(22), 대우건설(23)
군집4	현대상선(8)
군집5	SK네트웍스(9)
군집6	신한지주(25), 우리금융(35), 외환은행(40), 기업은행(42)
군집7	NHN(31)
군집8	두산인프라코어(32), 삼성중공업(38), 현대중공업(41), 두산중공업(43), 대우조선해양(45)

<표 3>에서 확인한 8개의 군집을 통해 각각의 군집에 대한 확률을 구하여 본 결과는 <표 4>와 같다. 이를 통해 어떠한 군집에 더 많은 자료들이 포함되어 있는지 알 수 있다. 군집1의 확률은 약 η_1=0.2889, 군집2는 η_2=0.3556, 군집3과 군집6은 η_3=η_6=0.0889, 군집4, 군집5, 군집7은 η_4=η_5=η_7=0.0222, 군집8은 η_8=0.1111임을 알 수 있다.

표 4. 8개 군집에 따른 군집확률
Table 4. Eight clustering in the cluster probability

군집1	0.2889	군집5	0.0222
군집2	0.3556	군집6	0.0889
군집3	0.0889	군집7	0.0222
군집4	0.0222	군집8	0.1111

군집2가 가장 큰 확률을 가지고 있으므로 군집2에 많은 시계열 자료들이 포함되어 있으며, 군집4 군집5, 군집7은 가장 작은 확률을 가지고 있으므로 가장 적은 시계열 자료들이 포함되어 있다.

[그림 1]은 45개 종목의 수익률을 8개의 그룹으로 나누어 그린 시계열 그림이며, 이는 각 군집의 수익률 특성을 대표한다고 볼 수 있다. 군집4, 군집5와 군집7은 각 그룹에 해당되는 자료가 하나뿐이므로 이는 곧 각각의 군집을 대표한다.

그림 1. 45개 종목의 수익률에 대한 군집분석 결과
Figure1. The result of cluster analysis on the yield of the 45 events

각 군집들의 모수를 MCMC방법으로 샘플 5000개를 추정하였고, 추정값을 찾아가는 과정에서 초반에 추정된 값들은 후반에 추정된 추정값들과 차이가 나기 때문에 추정값의 신뢰성을 높이기 위해서 그 중 1000개를 소진기간(burn in period)으로 제거하였다.

프로그램 [R]에서는 bayesGARCH 패키지를 통해 분석 할 수 있으며 그 결과 <표 5>와 같은 식으로 나타낼 수 있다.

표 5. 8개의 군집에 따른 모형
Table 5. Eight clusters according to the model

군집	모형
군집1	$y_t^1 = \sqrt{h_t^1}\,\epsilon_t^1, \quad \epsilon_t^1 \mid I_{t-1}^1 \sim student(0,1,2.0055)$ $h_t^1 = 0.0227 + 0.5469\,(y_{t-1}^1)^2 + 0.0903\,h_{t-1}^1$ $(0.0042)\ (0.2897)\qquad (0.0872)$
군집2	$y_t^2 = \sqrt{h_t^2}\,\epsilon_t^2, \quad \epsilon_t^2 \mid I_{t-1}^2 \sim student(0,1,2.0702)$ $h_t^2 = 0.0040 + 0.7716\,(y_{t-1}^2)^2 + 0.1101\,h_{t-1}^2$ $(0.0015)\ (0.1067)\qquad (0.0777)$
군집3	$y_t^3 = \sqrt{h_t^3}\,\epsilon_t^3, \quad \epsilon_t^3 \mid I_{t-1}^3 \sim student(0,1,2.0630)$ $h_t^3 = 0.0101 + 0.7715\,(y_{t-1}^3)^2 + 0.0704\,h_{t-1}^3$ $(0.0024)\ (0.1262)\qquad (0.0511)$
군집4	$y_t^4 = \sqrt{h_t^4}\,\epsilon_t^4, \quad \epsilon_t^4 \mid I_{t-1}^4 \sim student(0,1,4.7370)$ $h_t^4 = 0.0006 + 0.2022\,(y_{t-1}^4)^2 + 0.6898\,h_{t-1}^4$ $(0.0012)\ (0.1460)\qquad (0.1568)$
군집5	$y_t^5 = \sqrt{h_t^5}\,\epsilon_t^5, \quad \epsilon_t^5 \mid I_{t-1}^5 \sim student(0,1,2.0886)$ $h_t^5 = 0.0116 + 0.9189\,(y_{t-1}^5)^2 + 0.0163\,h_{t-1}^5$ $(0.0007)\ (0.0590)\qquad (0.0137)$
군집6	$y_t^6 = \sqrt{h_t^6}\,\epsilon_t^6, \quad \epsilon_t^6 \mid I_{t-1}^6 \sim student(0,1,2.0320)$ $h_t^6 = 0.0110 + 0.8830\,(y_{t-1}^6)^2 + 0.0310\,h_{t-1}^6$ $(0.0005)\ (0.0744)\qquad (0.0260)$
군집7	$y_t^7 = \sqrt{h_t^7}\,\epsilon_t^7, \quad \epsilon_t^7 \mid I_{t-1}^7 \sim student(0,1,5.4955)$ $h_t^7 = 0.0006 + 0.2304\,(y_{t-1}^7)^2 + 0.4258\,h_{t-1}^7$ $(0.0008)\ (0.1244)\qquad (0.1222)$
군집8	$y_t^8 = \sqrt{h_t^8}\,\epsilon_t^8, \quad \epsilon_t^8 \mid I_{t-1}^8 \sim student(0,1,2.1055)$ $h_t^8 = 0.0061 + 0.7313\,(y_{t-1}^8)^2 + 0.1370\,h_{t-1}^8$ $(0.0026)\ (0.1529)\qquad (0.0999)$

()안의 값은 표준오차 값이며, 각 군집에 해당되는 추정된 모수의 분포는 그 결과를 생략했다.

III. 결론

실제로 우리가 접하는 금융 시계열 자료들의 변동성은 GARCH(p,q) 모형으로 모형화 할 수 있고, 대개 GARCH(1,1) 모형으로 모형화할 수 있다. 여기서는 GARCH(1,1) 모형들의 모형기반 군집분석을 실시하였다.

본 연구에서 사용한 자료는 KOSPI50 주가지수의 일별 종가자료에서 해당 기간에 결측이 있는 종목을 제외한 45개의 종목을 이용하였다. 45개의 종목의 주가 수익률 자료를 가지고 모형기반 군집방법을 이용하여 군집분석을 하였다. 각각의 종목들은 GARCH(1,1) 모형을 따르는 것으로 가정하였다. BIC 값이 큰 경우의 모형과 군집의 수를 선택하면, VI모형(군집의 부피가 서로 다르지만, 군집의 모양은 동일한 모형)이면서 군집의 수가 8개인 것이 적절하다는 결과를 얻을 수 있었다. 8개로 군집을 나눈 자료로 MCMC 방법을 이용하여 각각의 군집에 대한 모수를 추정하였다. 각 군집의 모수들은 군집에 속하는 자료들을 대표하고 있다.

참고문헌

[1] Ardia, D. (2008) ; Financial Risk Management with Bayesian Estimation of GARCH Model : theory and applications, Springer

[2] Asai, M. (2006) ; "Comparison of MCMC methods for Estimating GARCH models", *Journal of the Japan Statistical Society*, Vol. 36, No. 2, pp. 199-212.

[3] Banfield, J., and Raftery, A. (1993) ; "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics,* Vol. 49, No. 3, pp. 803-821.

[4] Bauwens, L., and Lubrano, M. (1998) ; "Bayesian Inference on GARCH Models Using the Gibbs Sampler", *The Econometrics Journal,* Vol. 1, No. 1, pp. 23-46.

[5] Bauwens, L., and Rombouts, J.V.K. (2007) ; "Bayesian clustering of many GARCH models", *Econometric Reviews,* 26, pp. 365-386.

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

[7] Chris, F., and Raftery, A. (2002) ; "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association,* Vol. 97, pp. 611-631.

[8] Fruhwirth-Schnatter, S. (2006) ; Finite mixture and Markov switching models, Springer Series in Statistics, Springer

[9] Fruhwirth-Schnatter, S., and Kaufmann, S. (2005) ; "Model-based Clustering of multiple time-series", Working Paper, Johannes Kepler Universitat Linz

[10] McLachlan, G., and Peel, D. (2000) ; Finite Mixture Models, Wiley Series in Probability and Statistics

[11] Nakatsuma, T. (2000) ; "Bayesian analysis of ARMA-GARCH models: a Markov chain sampling approach", *Journal of Econometrics*, Vol. 95, No. 1, pp. 57-69.

[12] Otranto, E. (2008) ; "Clustering Heteroskedastic Time Series by Model-Based Procedures", *Computational Statistics & Data Analysis,* Vol. 52, No.10, pp. 4685-4698.

[13] Pamminger, C. (2008) ; Bayesian clustering of categorical time series : an approach using finite mixtures of markov chain models, VDM Verlag Dr. Mueller

저자소개

유희경(Hee-Kyung Yoo)

동국대학교 대학원 이학박사

강원대학교 컴퓨터공학과 교수

※관심분야 : 컴퓨터보안, 컴퓨터시뮬레이션, 데이터마이닝

정수정(Su-Jeong Jeong)

동국대학교 대학원 이학석사

㈜시벅스테크놀러지
※관심분야 : 데이터 마이닝, 다변량분석, 베이지안 이론

성 경(Kyung Sung)

2003년 2월 : 한남대학교
컴퓨터공학과 (공학박사)
1994 ~ 2004년 동해대학교
컴퓨터공학과 교수
2004년~현재 목원대학교
컴퓨터교육과 교수

※ 관심분야 : 정보보호 및 정보관리, 컴퓨터네트워크, 신경회로망, 컴퓨터교육