Communications for Statistical Applications and Methods
2022, Vol. 29, No. 6, 695–708

https://doi.org/10.29220/CSAM.2022.29.6.695
Print ISSN 2287-7843 / Online ISSN 2383-4757

Grouping stocks using dynamic linear models

Sihyeon Kim[a], Byeongchan Seong[1,a]

[a]Department of Applied Statistics, Chung-Ang University, Korea

Abstract

Recently, several studies have been conducted using state space model. In this study, a dynamic linear model with state space model form is applied to stock data. The monthly returns for 135 Korean stocks are fitted to a dynamic linear model, to obtain an estimate of the time-varying β-coefficient time-series. The model formula used for the return is a capital asset pricing model formula explained in economics. In particular, the transition equation of the state space model form is appropriately modified to satisfy the assumptions of the error term. k-shape clustering is performed to classify the 135 estimated β time-series into several groups. As a result of the clustering, four clusters are obtained, each consisting of approximately 30 stocks. It is found that the distribution is different for each group, so that it is well grouped to have its own characteristics. In addition, a common pattern is observed for each group, which could be interpreted appropriately.

Keywords: dynamic linear model, state space model, CAPM, k-shape clustering

1. Introduction

For investing in stocks, investors can use a lot of information when constructing their personal portfolio. This study attempts to provide useful information using a dynamic linear model (DLM). Although this model has been a well-known theory for a long time, we use a novel approach to interpret the estimated time-varying β-coefficient for stocks. A detailed description of the model fitting follows in Section 3. To fit the model, the monthly returns of each stock are expressed as a capital asset pricing model. This model attempts to explain the excess returns of individual stocks in terms of the excess returns of the overall market and systematic risk, which measures volatility. Here, excess return means individual return minus risk-free return. A specific description of the model formula is provided in Section 2. The estimated time-varying β-coefficient is then used to group stocks using k-shape clustering. The clustering method is described in detail in Section 4. This study not only estimates the regression coefficient, but also groups stocks and interprets their common characteristics to gain insight and provide useful information to investors.

2. Capital asset pricing model (CAPM)

The capital asset pricing model (CAPM) was first suggested by Markowitz (1952). It attempts to explain the equilibrium price of financial assets in the capital market based on average and variance.

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2021. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1A01073864).

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bcseong@cau.ac.kr

Published 30 November 2022 / journal homepage: http://csam.or.kr

Here, the equilibrium price means the price at which supply and demand are equal in a perfectly competitive market involving free competition. There are several assumptions that hold for CAPM to be established. Investment is based on perfect capital markets, all investors have homogeneous expectations, the model is in a single-period, risk-free assets exist, and all investors decide their portfolios by average and variance criteria. CAPM represents the expected return $E(r_p)$ of a specific portfolio p as follows (Kim, 2003):

$$E\left(r_p\right) = (1 - w)\, r_f + wE\left(r_m\right), \tag{2.1}$$

where r_m denotes the overall market return, and r_f is the risk-free rate of return obtained from a risk-free asset. If inflation is not considered, the return on risk-free assets is constant. Because the rate of return is constant, the standard deviation is zero, and the correlation coefficient with other assets is zero. We can calculate w using the fact that the correlation coefficient for risk-free returns is zero:

$$w = \frac{\sigma_p}{\sigma_m}. \tag{2.2}$$

The capital market line (CML) is the set of portfolios obtained by combining the market portfolio and risk-free assets. Combining Equations (2.1) and (2.2) yields the following CML:

$$E\left(r_p\right) = r_f + \frac{E\left(r_m\right) - r_f}{\sigma_m}\sigma_p. \tag{2.3}$$

The CML for the return on individual stocks r_i rather than the market portfolio p can be summarized as follows:

$$\begin{aligned} E\left(r_i\right) &= r_f + \frac{E\left(r_m\right) - r_f}{\sigma_m}\sigma_i \\ &= r_f + \left(E\left(r_m\right) - r_f\right)\frac{\mathrm{Cov}\left(r_i, r_m\right)}{\mathrm{Var}\left(r_m\right)}, \\ E\left(r_i\right) - r_f &= \left(E\left(r_m\right) - r_f\right)\cdot \beta_i. \end{aligned} \tag{2.4}$$

That is, the expected value of an asset's excess return has a linear relationship with the asset's systematic risk (i.e., the β-coefficient). Systematic risk cannot be reduced by diversified investments as it is a common factor in the market. Examples of systematic risk include inflation, economic cycle, and interest rate fluctuations. This study explains the relationship between the excess returns of individual assets and those of the market by establishing a model with time-varying coefficients.

3. Dynamic linear model (DLM)

3.1. DLM in state space model form

A DLM is a linear regression model in which the regression coefficients change with time (Holmes *et al.*, 2021; Montillet and Bos, 2019). The DLM allows the setting of various models by expressing them in the form of a state space model.

$$y_t = \mathbf{Z}_t\mathbf{x}_t + v_t, \tag{3.1}$$

$$\mathbf{x}_t = \mathbf{B}_t\mathbf{x}_{t-1} + \mathbf{w}_t. \tag{3.2}$$

Equation (3.1) is a measurement equation that expresses actual observed data (Holmes *et al.*, 2014). For $t = 1, \ldots, T$, there are n observations and m independent variables and y_t is an $n \times 1$ vector and can be expressed as a linear combination of $\mathbf{Z}_t\,(n \times m)$ and $\mathbf{x}_t\,(m \times 1)$. In the DLM, $\mathbf{Z}_t$ contains the independent variable, and the $\mathbf{x}_t$ vector corresponds to the regression coefficient. Equation (3.2) is a transition equation that describes how the time-evolving regression parameter $\mathbf{x}_t$ changes with time. This equation can be expressed as a linear combination of $\mathbf{x}_t$ and $\mathbf{B}_t\,(m \times m)$. v_t is the observation error. Observation errors are excluded, if necessary. If included, $\mathbf{v}$ is assumed to follow a multivariate normal distribution ($\mathbf{v} = (v_1, \ldots, v_T), \mathbf{v} \sim \text{MVN}(\mathbf{0}, \mathbf{R})$). $\mathbf{w}_t$ is the process error, which is assumed to follow a multivariate normal distribution ($\mathbf{w}_t \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$).

3.2. Estimating and predicting

The process of predicting y_t involves two steps. (*i*) estimating the regression parameters at time t, and (*ii*) predicting y_t.

3.2.1. Estimating the regression parameters

A Kalman filter is used to estimate the regression parameters (Holmes *et al.*, 2021). First, define the distribution of $\mathbf{x}_{t|t} = E\left(\mathbf{x}_t | \varphi_t\right)$ to set the initial value, where φ_t means the set of information up to t:

$$\mathbf{x}_{t|t} \sim \text{MVN}\left(\boldsymbol{\pi}_t, \boldsymbol{\Lambda}_t\right). \tag{3.3}$$

With the defined distribution and Equation (3.2), the distribution of $\mathbf{x}_{t|t-1}$ is as follows:

$$\begin{aligned} E\left(\mathbf{x}_{t|t-1}\right) &= \mathbf{B}_t E\left(\mathbf{x}_{t-1|t-1}\right) \\ &= \mathbf{B}_t \boldsymbol{\pi}_{t-1}, \\ \text{Var}\left(\mathbf{x}_{t|t-1}\right) &= \mathbf{B}_t \text{Var}\left(\mathbf{x}_{t-1|t-1}\right) \mathbf{B}_t^T + \mathbf{Q} \\ &= \mathbf{B}_t \boldsymbol{\Lambda}_{t-1} \mathbf{B}_t^T + \mathbf{Q}, \\ \therefore\ \mathbf{x}_{t|t-1} &\sim \text{MVN}\left(\mathbf{B}_t \boldsymbol{\pi}_{t-1}, \mathbf{B}_t \boldsymbol{\Lambda}_{t-1} \mathbf{B}_t^T + \mathbf{Q}\right). \end{aligned} \tag{3.4}$$

3.2.2. Predicting y_t

One-step ahead prediction of y can be obtained using the regression parameter at time t estimated above, the predictor at time t, and Equation (3.1):

$$\begin{aligned} \hat{y}_{t|t-1} &= \mathbf{Z}_t E\left(\mathbf{x}_{t|t-1}\right) + v_t \text{ where } v_t \sim N(0, r), \\ E\left(\hat{y}_{t|t-1}\right) &= \mathbf{Z}_t E\left(\mathbf{x}_{t|t-1}\right) \\ &= \mathbf{Z}_t \mathbf{B}_t \boldsymbol{\pi}_{t-1}, \end{aligned} \tag{3.5}$$

$$\begin{aligned} \text{Var}\left(\hat{y}_{t|t-1}\right) &= \mathbf{Z}_t \text{Var}\left(\mathbf{x}_{t|t-1}\right) \mathbf{Z}_t^T + r \\ &= \mathbf{Z}_t \left(\mathbf{B}_t \boldsymbol{\Lambda}_{t-1} \mathbf{B}_t^T + \mathbf{Q}\right) \mathbf{Z}_t^T + r. \end{aligned} \tag{3.6}$$

3.3. Diagnostics

The DLM assumes that the error term follows a multivariate normal distribution. To check whether the assumption holds, the residual e_t is used, which is the difference between the actual observed value y_t and the predicted value $\hat{y}_{t|t-1}$. The following three visual inspections are performed.

- plot of (t, e_t) : It should be randomly scattered without any specific pattern such as changing with time.
- autocorrelation plot of e_t : As the autocorrelation function (ACF) falls within the two standard error lines, there is no autocorrelation between the residuals.
- Q-Q plot of e_t : As the residuals are distributed near the observed-expected line, the property of a normal distribution must be satisfied.

4. Clustering

Clustering is a type of unsupervised learning algorithm, that groups observations based on the distance between observations. Several clustering methods depend on how the distance between the observations is measured. In this study, k-shape clustering, which is suitable for grouping time-series data, is used (Jin and Seong, 2020).

4.1. k-shape clustering

k-shape clustering measures the distance between groups based on shape-based distance measure (SBD) (Paparrizos and Gravano, 2016). SBD calculates the distance between two time-series using cross-correlation. Cross-correlation measures the similarity of time-lagged signals. The cross-correlation CC_w with lag w of two time-series vectors $\vec{x}$ and $\vec{y}$ of length m is defined as follows ($w \in \{1, 2, \ldots, 2m-1\}$):

$$\mathrm{CC}_w(\vec{x}, \vec{y}) = R_{w-m}(\vec{x}, \vec{y}), \tag{4.1}$$

$$R_k(\vec{x}, \vec{y}) = \begin{cases} \displaystyle\sum_{l=1}^{m-k} x_{l+k} y_l, & k \geq 0, \\ R_{-k}(\vec{y}, \vec{x}), & k < 0. \end{cases}$$

The SBD of the two time-series vectors $\vec{x}$ and $\vec{y}$ are as follows:

$$\mathrm{SBD}(\vec{x}, \vec{y}) = 1 - \max_w \left(\frac{\mathrm{CC}_w(\vec{x}, \vec{y})}{\sqrt{R_0(\vec{x}, \vec{x}) \cdot R_0(\vec{y}, \vec{y})}} \right), \quad 0 \leq \mathrm{SBD}(\vec{x}, \vec{y}) \leq 2. \tag{4.2}$$

In other words, after adjusting the scale of each time-series, SBD measures the similarity at the lag when the cross-correlation between the two time-series is the greatest. If the two time-series are equal, then the SBD is zero. Clustering determines the optimal centroid by repeating the process of calculating the centroid of each group after grouping based on the measured distance. Equation (4.3) is the centroid vector $\vec{\mu}_k^*$ of the k^{th} group P_k in k-shape clustering.

$$\vec{\mu}_k^* = \arg\max_{\vec{\mu}} \sum_{\vec{x}_i \in P_k} \left(\frac{\max_w \mathrm{CC}_w(\vec{x}_i, \vec{\mu}_k)}{\sqrt{R_0(\vec{x}_i, \vec{x}_i) \cdot R_0(\vec{\mu}_k, \vec{\mu}_k)}} \right)^2. \tag{4.3}$$

4.2. Davis-Bouldin index

In this study, the Davis-Bouldin index (DB index) is used to evaluate clustering performance (Scherl, 2010). It is necessary to define the $\mathbf{R} = (R_{ij})$ matrix to measure the similarity between clusters C_i and

C_j:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}, \tag{4.4}$$

where s_i is the average distance between the observation values of cluster C_i and the centroid; and d_{ij} denotes the distance between C_i and C_j. R_{ij} must satisfy the following conditions.

- $R_{ij} \geq 0$.
- $R_{ij} = R_{ji}$.
- if $s_i = s_j = 0$, then $R_{ij} = 0$.
- if $s_j > s_k$ and $d_{ij} = d_{ik}$, then $R_{ij} > R_{ik}$.
- if $s_j = s_k$ and $d_{ij} < d_{ik}$, then $\mathrm{v}R_{ij} < R_{ik}$.

If these are satisfied, **R** is symmetric positive semi-definite, and the diagonal element becomes zero. The DB index using **R** is as follows:

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} R_i, \tag{4.5}$$

$$R_i = \max_{j=1,\ldots,k \wedge j \neq i} R_{ij}, \tag{4.6}$$

where k denotes the number of clusters. When the DB index is the lowest, the number of clusters can be determined as optimal.

5. Model fitting

In this study, the CAPM model is applied to the DLM to explain the excess returns of stocks listed on the Korea Exchange (KRX) as market excess returns, and to interpret the time-varying β-coefficient. The CML (2.4) described in the CAPM can be rewritten as follows:

$$r_{pt} - r_{ft} = \left(r_{mt} - r_{ft}\right) \cdot \beta_t + \epsilon_t, \quad \epsilon_t \sim N\left(0, \sigma^2\right). \tag{5.1}$$

As defined in Section 2, r_{pt}, r_{ft}, and r_{mt} are the returns of individual stock, risk-free rate, and market, respectively. β_t corresponds to systematic risk.

5.1. Dataset

Among the stocks included in the Korea Composite Stock Price Index (KOSPI) 200 as of April 2022, the monthly returns of 135 stocks for which data exist since January 2009, are used as dependent variables. Considering the subprime mortgage crisis of 2007–2008, the starting point of the data is January 2009. The monthly return of the KOSPI Composite Index is used for the market return. To obtain data, the Yahoo Finance API is used. For the risk-free rate of return, we use the yield of a three-month short-term government bond (treasury bill). The FRED economic data API provides these data. Figure 1 and Figure 2 show trends in market returns and risk-free returns over time.

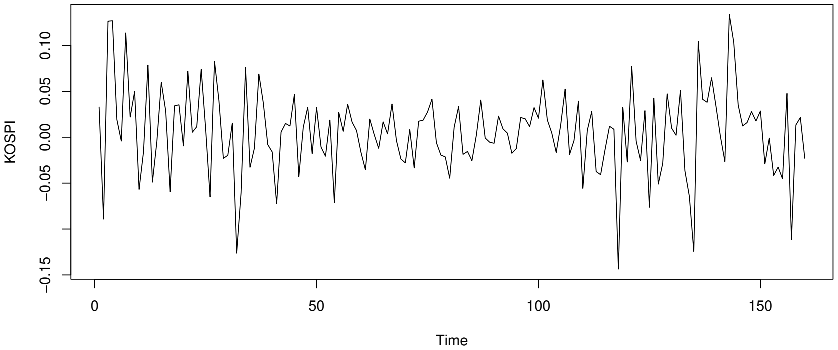

Figure 1: *Monthly return of KOSPI composite index.*

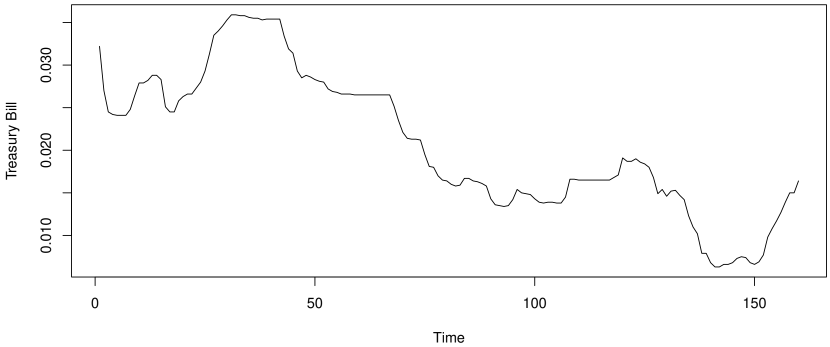

Figure 2: *Three-month rates and yields.*

5.2. Model

The state space model form of the DLM described in Section 3.1 is constructed as follows. It is based on the basic structure of Equation (5.1), but the structure has been modified to suit these data.

$$\begin{aligned} r_{pt} - r_{ft} &= \alpha_t + \beta_t \left(r_{mt} - r_{ft} \right) + \gamma_t, \\ \alpha_t &= \alpha_{t-1} + v_{\alpha,t}, \quad v_{\alpha,t} \sim N\left(0, \sigma_\alpha^2\right), \\ \beta_t &= \beta_{t-1} + v_{\beta,t}, \quad v_{\beta,t} \sim N\left(0, \sigma_\beta^2\right), \\ \gamma_t &= \phi_0 + \phi_1 \gamma_{t-1} + v_{\gamma,t}, \quad v_{\gamma,t} \sim N\left(0, \sigma_\gamma^2\right). \end{aligned} \tag{5.2}$$

The intercept term α_t is added in a manner similar to the linear regression model, but α_t and β_t corresponding to the regression coefficients are set to follow a random walk. Therefore, the two regression coefficients are estimated differently over time (Kim, 2003). The observation error term of DLM follows a multivariate normal distribution. However, fitting these data based on this assumption causes problems with the lag 1 ACF of the residuals for almost all stocks. As an example, the residual diagnostic results of NAVER and LG Electronics are shown in Figure 3. It can be observed that the lag 1 ACF is large in the negative direction. To resolve the autocorrelated residuals, the observation

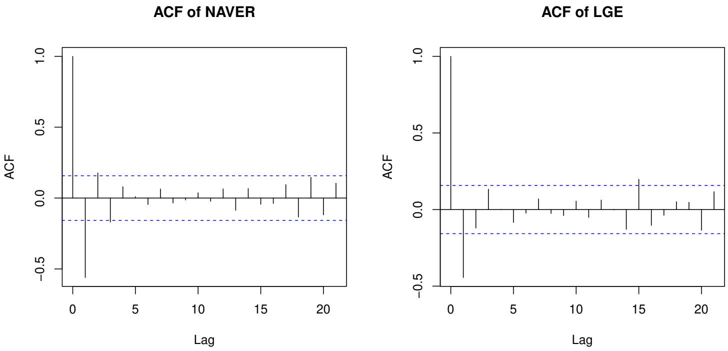

Figure 3: *Residual diagnostics for NAVER and LG Electronics (before modifying the model).*

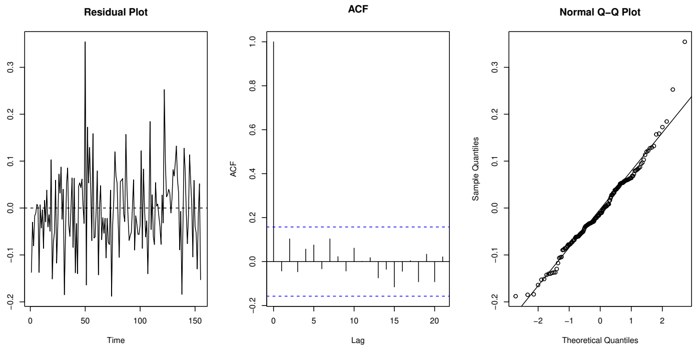

Figure 4: *Residual diagnostic for NAVER (final model).*

error term is removed, and the γ_t state is added. According to the corresponding transition equation, γ_t has an AR(1) structure as shown below:

$$\gamma_t = \phi_0 + \phi_1\gamma_{t-1} + v_{\gamma,t}, \quad v_{\gamma,t} \sim N\left(0, \sigma^2_{\gamma}\right). \tag{5.3}$$

Consequently, the lag 1 ACF problem is solved by applying the AR(1) structure to the error term of the observation equation. Equation (5.2) can be changed to the following matrix form.

$$\mathbf{y}_t = \mathbf{Z}_t\mathbf{x}_t, \tag{5.4}$$

$$\mathbf{x}_t = \mathbf{U}_t + \mathbf{B}_t\mathbf{x}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \text{MVN}\left(0, \mathbf{Q}\right). \tag{5.5}$$

$$\left(r_{pt} - r_{ft}\right) = \begin{pmatrix} 1 & \left(r_{rm} - r_{ft}\right) & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}_t .$$

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}_t = \begin{pmatrix} 0 \\ 0 \\ \phi_0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \phi_1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}_{t-1} + \begin{pmatrix} v_\alpha \\ v_\beta \\ v_\gamma \end{pmatrix}_t, \quad \begin{pmatrix} v_\alpha \\ v_\beta \\ v_\gamma \end{pmatrix}_t \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & 0 & 0 \\ 0 & \sigma^2_\beta & 0 \\ 0 & 0 & \sigma^2_\gamma \end{pmatrix}\right).$$

This is fitted using the MARSS() function in R (Holmes *et al.*, 2020). It is necessary to set several arguments in the function. First, the maximum number of iterations is set to be 3,000 to achieve

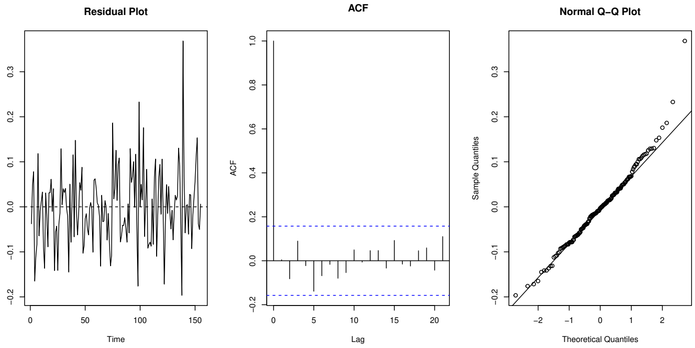

Figure 5: *Residual diagnostic for LG Electronics (final model).*

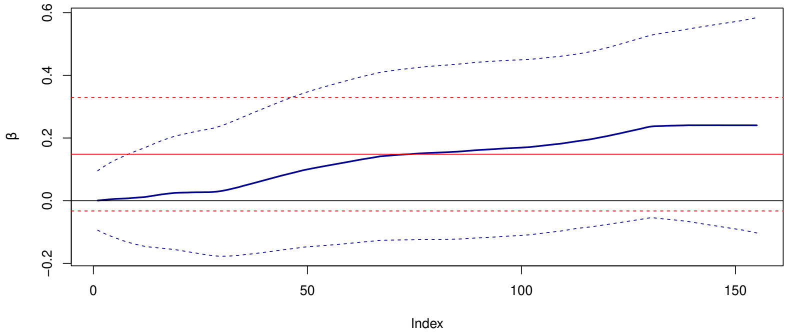

Figure 6: *The estimate of time-varying β for KT&G.*

convergence through a sufficient number of steps. Additionally, the initial values of the time-varying coefficients are set as the first element of $\mathbf{Z}$ (Holmes *et al.*, 2014).

$$\begin{aligned}\alpha_0 &= Z_{11} = 1,\\ \beta_0 &= Z_{12} = 0.0007,\\ \gamma_0 &= Z_{13} = 1.\end{aligned}$$

As a result of the fitting, an estimate of the time-varying β-coefficient can be obtained. It is judged that it does not converge until the first five months, so the results are interpreted only after six months. Prior to the interpretation, it is necessary to review whether these fitted models violate these assumptions. The residual diagnostic results for the same stocks as in Figure 3 are shown in Figure 4 and Figure 5. As these figures show, these are white noises because they satisfy the independence and normality of the error. It is confirmed that none of the other 133 stocks violate the assumptions. Figure 6 and Figure 7 show examples of the estimated β_t for KT&G and CJ CGV, respectively. The blue solid line corresponds to the DLM estimates, and the blue dotted line represents the 95% confidence interval for the DLM estimates. The OLS estimates are also included for comparison. The red solid line represents the OLS estimates, and the red dotted line corresponds to the 95% confidence interval. It is shown that the standard error of the DLM coefficients is larger than that of the OLS. Additionally, there are times when the DLM estimate is not included in the OLS interval. In other words, it can be concluded that estimates do not constantly follow the volatility of the entire market, but they have an unstable section. In the case of KT&G, no time points deviated from the OLS confidence interval

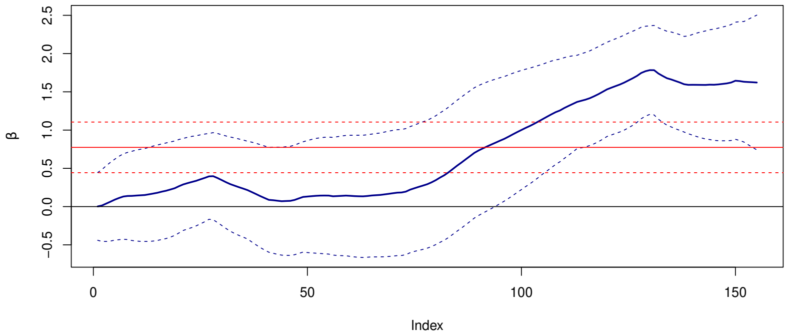

Figure 7: *The estimate of time-varying β for CJ CGV.*

Table 1: Time frequency for instability

Stock Name	Ticker	$\#\left\{\beta_t^{\text{DLM}} \mid \beta_t^{\text{DLM}} \notin \left(\beta_{\text{lwr}}^{\text{OLS}}, \beta_{\text{upr}}^{\text{OLS}}\right)\right\}$
CJ CGV	079160.KS	134
Hyundai Green Food	005440.KS	133
Lotte Chemical Corporation	011170.KS	124
Hanssem	009240.KS	124
SNT Motiv	064960.KS	112
Daewoong	003090.KS	112
GS Holding Corp.	078930.KS	110
LG Chem.	051910.KS	110
Samsung Securities	016360.KS	109
Sebang Global Battery	004490.KS	109
⋮	⋮	⋮
Celltrion	068270.KS	19
KT Corporation	030200.KS	18
Ncsoft Corporation	036570.KS	18
Hyundai Doosan Infracore	042670.KS	18
HANDSOME Corp.	020000.KS	17
Hyundai Mobis	012330.KS	17
KG Dongbu Steel	016380.KS	14
Posco International Corporation	047050.KS	14
Dongwon Systems Corporation	014820.KS	8
KT&G Corporation	033780.KS	0

(Figure 6). That is, stock generally follows the volatility across the market. Conversely, 134 time points of the CJ CGV deviate from the OLS confidence interval and can be interpreted as unstable movements, independent of market volatility (Figure 7). Results for the instability of the other stocks are described in Table 1.

5.3. Clustering result

The time-series of the estimated time-varying β-coefficients are applied to k-shape clustering and grouped according to their characteristics (Sard*á*-Espinosa, 2017). As mentioned in Section 4.2, the optimal number of clusters is chosen, which can best classify 135 time-series based on the DB index. After setting the range of the number of clusters to a minimum of three and a maximum of ten, and checking each DB index, the optimal number of clusters is four, in which case the DB index is the lowest (Figure 8). As described in Section 4.1, $k = 4$ is set, and β_t time-series clustering is performed

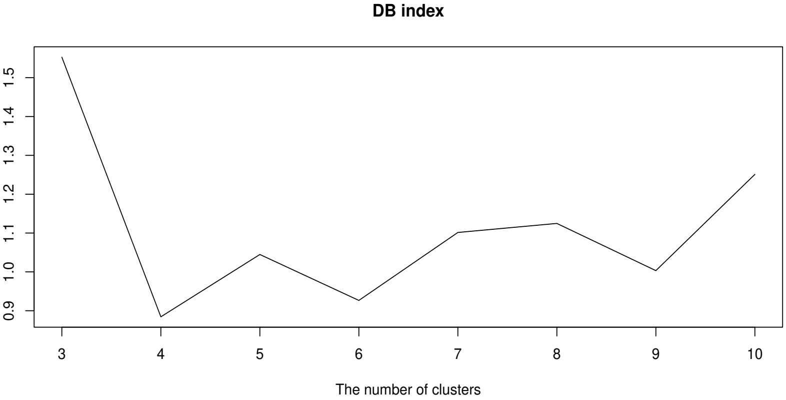

Figure 8: *DB index for determining the number of clusters.*

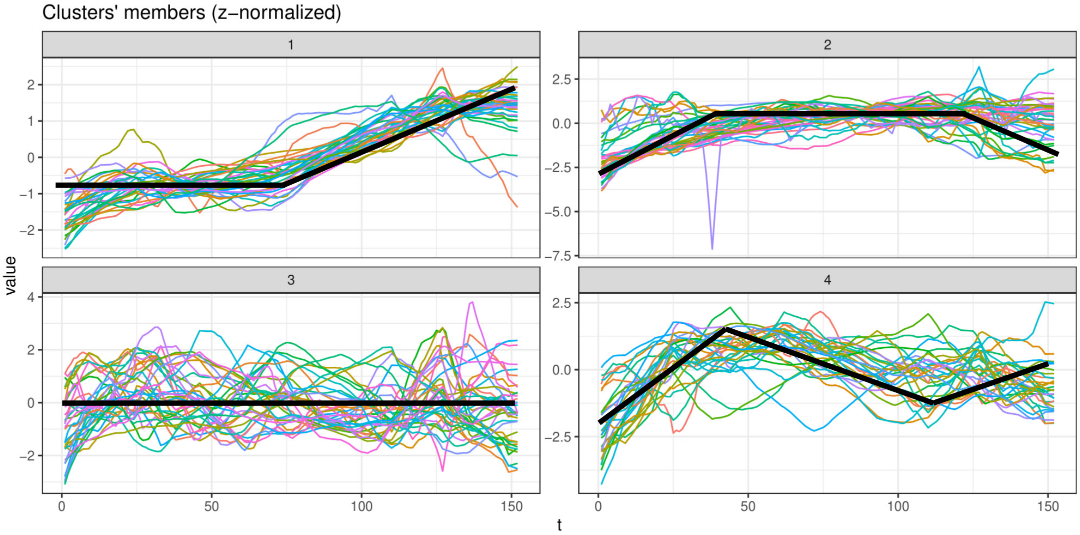

Figure 9: *Clustering result and the overall trend of each cluster.*

for 135 stocks. For the clustering performance, each time-series is standardized. Figure 9 shows the β_t time-series for each cluster. The overall trend of each cluster is represented by a thick black line. Section 6.2 will interpret this. Table 2 lists the stocks included in each cluster. In each cluster, 33, 38, 35, and 29 stocks are included, respectively.

It may be interesting to use the second best as 6 clusters. When we use it, the 6 clusters have 31, 32, 9, 20, 24 and 19 stocks in each cluster. The result can be found in Figure 10. When comparing the two results of 4 and 6 clusters, the union of the 3rd and 5th groups in the 6 clusters is similar to the 3rd group in the 4 clusters. Other than that, it is found that there no big difference.

As can be seen from Equation (5.2), α_t is also a time-varying regression coefficient. Therefore, α_t can also be used for clustering. However, it was found that for most stocks, α_t was estimated as a constant. As we focus on time varying effects, we do not use them for clustering.

Table 2: Result of clustering

Cluster	Stocks
# 1 (33 stocks)	CJ Cheiljedang Corporation(097950.KS), GS Holdings Corp.(078930.KS), KB Financial Group(105560.KS), LG Corp. (003550.KS), LG Uplus Corp.(032640.KS), LG Electronics(066570.KS), LX International Corp.(001120.KS), S-Oil Corporation(010950.KS), SNT Motiv(064960.KS), Korea Zinc Company(010130.KS), Kumho Petrochemical(011780.KS), Industrial Bank of Korea(024110.KS), Nexen Tire Corporation(002350.KS), Daewoong(003090.KS), Meritz Securities(008560.KS), Mirae Asset Securities(006800.KS), Samsung Card(029780.KS), Samsung Electro-Mechanics(009150.KS), Samyang Holdings Corporation(000070.KS), AMOREPACIFIC Group(002790.KS), Asiana Airlines(020560.KS), SL Corporation(005850.KS), S-1 Corporation(012750.KS), Young Poong Corporation(000670.KS), Ottogi Corporation(007310.KS), Taekwang Industrial(003240.KS), Hanmi Science(008930.KS), Hanon Systems(018880.KS), Hanwha Solutions Corporation(009830.KS), Hyundai Doosan Infracore(042670.KS), Hyundai Mobis(012330.KS), Hyundai Marine & Fire Insurance(001450.KS), TKG Huchems(069260.KS).
# 2 (38 stocks)	CJ Logistics Corporation(000120.KS), DB HiTek(000990.KS), HMM(011200.KS), GS Engineering & Construction Corporation(006360.KS), KCC Corporation(002380.KS), KG Dongbu Steel(016380.KS), LG H&H(051900.KS), LG Innotek(011070.KS), Samsung Card(029780.KS), POSCO Holdings(005490.KS), SK Telecom(017670.KS), Green Cross Holdings Corporation(005250.KS), Nongshim(004370.KS), LOTTE Corporation(004990.KS), Lotte Chemical Corporation(011170.KS), Samsung Engineering(028050.KS), Sebang Global Battery(004490.KS), Shinhan Financial Group(055550.KS), Amorepacific Corporation(090430.KS), ORION Holdings Corp.(001800.KS), Cheil Worldwide(030000.KS), Kakao Corp.(035720.KS), COWAY(021240.KS), Posco International Corporation(047050.KS), Posco Chemical(003670.KS), Poongsan Corporation(103140.KS), Hankook & Company(000240.KS), Korea Investment Holdings(071050.KS), Hanssem(009240.KS), Hanall Biopharma(009420.KS), Korea Shipbuilding & Offshore Engineering(009540.KS), Hanwha Corporation(000880.KS), Hanwha Aerospace(012450.KS), Hyundai Green Food(005440.KS), Hyundai Mipo Dockyard(010620.KS), Hyundai Engineering & Construction(000720.KS), Hotel Shilla(008770.KS), Foosung(093370.KS).
# 3 (35 stocks)	CJ CGV(079160.KS), KT&G Corporation(033780.KS), LG Display(034220.KS), LS Corp.(006260.KS), LS ELECTRIC(010120.KS), NAVER Corporation(035420.KS), NH Investment & Securities(005940.KS), OCI Company(010060.KS), SK Networks(001740.KS), SK Discovery(006120.KS), SK hynix(000660.KS), Daesang Corporation(001680.KS), Daewoong pharmaceutical(069620.KS), Korean Air Lines(003490.KS), Dong Suh(026960.KS), Dongwon Systems Corporation(014820.KS), Doosan Corporation(000150.KS), Doosan Enerbility(034020.KS), LOTTE Fine Chemical(004000.KS), Lotte Chilsung Beverage(005300.KS), Samsung Electronics(005930.KS), Samsung Securities(016360.KS), Samsung Heavy Industries(010140.KS), Kiwoom Securities(039490.KS), Pan Ocean(028670.KS), Korea Gas Corporation(036460.KS), HANDSOME Corp.(020000.KS), Hansol Chemical(014680.KS), KEPCO Plant Service & Engineering(051600.KS), Hyundai Glovis(086280.KS), Hyundai Elevator(017800.KS), Hyundai Department Store(069960.KS), Hyundai Motor Company(005380.KS), SK Innovation(096770.KS), Hyosung Corporation(004800.KS).
# 4 (29 stocks)	CJ Corporation(001040.KS), DB Insurance(005830.KS), DL Holdings(000210.KS), LG Chem(051910.KS), KT Corporation(030200.KS), Kangwon Land(035250.KS), Kumho Tire(073240.KS), Kia Corporation(000270.KS), Green Cross Corporation(006280.KS), Korea Petrochemical(006650.KS), Daewoo Engineering & Construction(047040.KS), Dongkuk Steel Mill(001230.KS), Dongwon Industries(006040.KS), Lotte Shopping(023530.KS), Boryung Corporation(003850.KS), Bukwang Pharmaceutical(003000.KS), Samsung SDI(006400.KS), Samsung Fire & Marine(000810.KS), Celltrion(068270.KS), SHINSEGAE(004170.KS), Shin Poong Pharm.(019170.KS), IS DongSeo(010780.KS), SsangYong C&E(003410.KS), Yungjin Pharm.(003520.KS), Ncsoft Corporation(036570.KS), Yuhan Corporation(000100.KS), Hana Financial Group(086790.KS), Hyundai Steel Company(004020.KS), Korea Electric Power Corporation(015760.KS).

6. Result

6.1. Equal-variance test

Before interpreting the characteristics of each cluster, a procedure to check whether each cluster is well grouped should be conducted. The Bartlett test of equal variances is performed to test whether

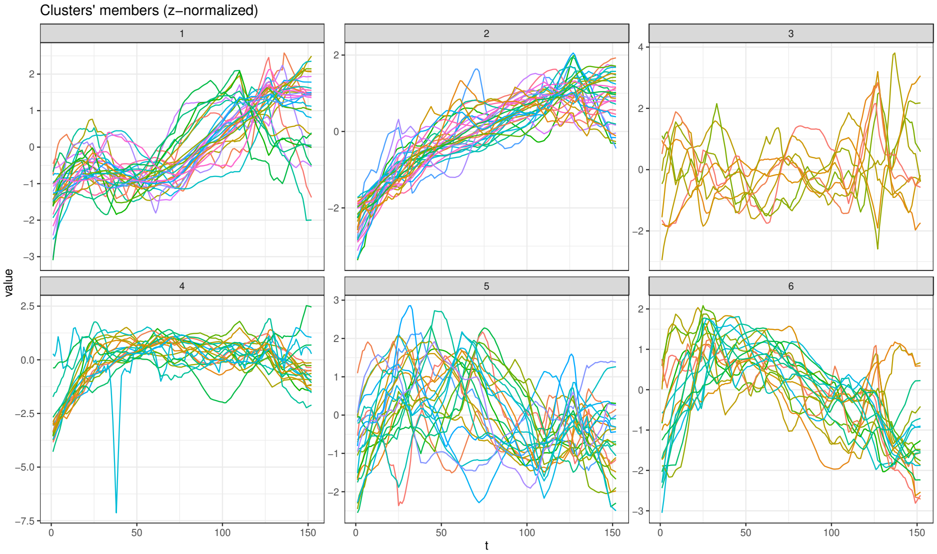

Figure 10: *Clustering result for* $k = 6$.

the distribution of each cluster is different (Bartlett, 1937). In the following hypothesis, σ_i^2 denotes the variance of the i^{th} cluster ($i = 1, 2, 3, 4$).

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2, \\ H_1 &: \text{not } H_0. \end{aligned}$$

The Bartlett test is tested four times using four different variables. Four variables representing the characteristics of each cluster were selected. The variables used in the test are as follows:

- variance of r_p: It can be checked whether the volatility of returns for each stock is classified differently depending on the cluster. The variance in r_p is calculated as follows:

$$\hat{\sigma}_{r_p}^2 = \frac{1}{T-1} \sum_{t=1}^{T} \left(r_{pt} - \bar{r}_p \right)^2, \tag{6.1}$$

  where $T = 160$, which is the length of the time-series.

- market capitalization: It can be checked whether the market capitalization (aggregate market value) in April 2022 differs for each cluster.

- historical return: It can be checked whether the rate of return from January 2009 (t_0) to April 2022 (t_1), which corresponds to the data collection period, differs for each cluster. y_t is the adjusted closing price at the time of t.

$$R_H = \log\left(y_{t_1}\right) - \log\left(y_{t_0}\right). \tag{6.2}$$

- future return : After the data collection period, it can be checked whether the monthly return in May 2022 differs for each cluster. In Equation (6.3), y_{t_0} is the adjusted closing price on May 1,

Table 3: Result of Bartlett test

Variables	Test statistic	p-value
$\hat{\sigma}^2_{r_p}$	8.8282	0.03166
market capitalization	242.61	<0.0001
R_H	35.69	<0.0001
R_F	8.3195	0.03985

2022, and y_{t_1} is the adjusted closing price on May 31, 2022.

$$R_F = \log(y_{t_1}) - \log(y_{t_0}). \tag{6.3}$$

As a result of the test, the null hypothesis is rejected for all four variables under the significance level of 0.05. In other words, the distributions of the four variables for each cluster are not the same. The test results are listed in Table 3. Therefore, each cluster is considered to have different characteristics, which are analyzed in Section 6.2.

6.2. Interpretation

The characteristics of each cluster are summarized below.

- Cluster 1: In the past, β_t did not change significantly over time and remained relatively stable. However, from March 2015, β_t shows an increasing trend with a steep slope, unlike the return of the market. In other words, there is a significant departure from market movement in the positive direction.
- Cluster 2: From 2009 to 2011, β_t gradually increased, but stabilized in 2012. From 2019, a decreasing trend was observed over time. There are times that deviate from the market movement for a certain period, but the volatility is not great.
- Cluster 3: No particular pattern was observed throughout the data period. This can be explained by the fact that movement does not differ significantly from market movement. In fact, it includes stocks with high market capitalization and large trading volumes (e.g., NAVER, SK Hynix, Samsung Electronics). These are heavy stocks that can affect the movement of the entire market, so they can move similarly to the market returns.
- Cluster 4: Compared with Cluster 2, β_t increased steeply until 2011. Most of them declined again until 2018, but recently showed an upward trend. Among the individual solid lines, stocks show an upward pattern from 2020 to 2021. It includes most pharmaceutical-related stocks such as Green Cross Corporation, Boryung Corporation, Bukwang Pharmaceutical, Celltrion, Shin Poong Pharm., Yungjin Pharm., and Yuhan Corporation. It has been shown that after COVID-19, their return increase, due to issues with diagnostic kits and treatments.

7. Conclusion

The paper attempts a new approach to explain the characteristics of stocks with a time-varying β-coefficient by adopting the well-known theories, CAPM and DLM. Instead of using the simple structure of the DLM formula, this study devised a form that added an AR(1) structure to create an optimal model for the data. By grouping the estimated β-coefficient time-series into four groups, stocks with similar characteristics can be grouped. Thus, it is possible to explain the distinctive features of each cluster. Investors can then utilize grouping information when composing their individual portfolios.

References

Bartlett MS (1937). Properties of sufficiency and statistical tests, *Proceedings of the Royal Society of London, Series A-Mathematical and Physical Sciences*, **160**, 268–282.

Holmes EE, Scheuerell MD, and Ward EJ (2020). *Applied Time Series Analysis for Fisheries and Environmental Data*, Seattle: Northwest Fisheries Science Center.

Holmes EE, Ward EJ, and Scheuerell MD (2014). *Analysis of Multivariate Time-series Using the MARSS package*, NOAA Fisheries, Northwest Fisheries Science Center.

Holmes EE, Ward EJ, Scheuerell MD, and Wills K (2021). *MARSS Package Manual*, Available from: https://cran. r-project. org/web/packages/MARSS/MARSS.pdf

Jin J and Seong B (2020). Analysis of the differences in living population changes and regional responses by COVID-19 outbreak in Seoul, *The Korean Journal of Applied Statistics*, **33**, 697–712.

Kim M and Chang K (2003). *Financial Econometrics* (2nd ed), Kyung Moon Sa, Seoul.

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

Montillet JP and Bos MS (2019). *Geodetic Time Series Analysis in Earth Sciences*, Springer Cham, Switzerland .

Paparrizos J and Gravano L (2016). *K*-shape: Efficient and accurate clustering of time series, *ACM SIGMOD Record*, **45**, 69–76.

Sard*á*-Espinosa A (2017). Comparing time-series clustering algorithms in R using the dtwclust package, *R Package Vignette*, **41**, 1–46.

Scherl M (2010). *Benchmarking of Cluster Indices*, Doctoral disseartation, Institut f*ü*r Statistik.

Received July 02, 2022; Revised August 16, 2022; Accepted August 17, 2022