Journal of Statistical Theory and Applications
Vol. **19(3)**, September 2020, *pp*. 391–396
DOI: https://doi.org/10.2991/jsta.d.200827.001; ISSN 1538-7887
https://www.atlantis-press.com/journals/jsta

Model-Based Filtering via Finite Skew Normal Mixture for Stock Data

Solmaz Yaghoubi[1], Rahman Farnoosh[2,*]

[1]*Science and Research Branch, Islamic Azad University, Tehran, Iran*
[2]*School of Mathematics, Iran University of Science and Technology, Tehran, Iran*

ARTICLE INFO

Article History

Received 09 March 2019
Accepted 28 July 2020

Keywords

Stock of banks and credit institutions
Mixture model
Clustering time series
Multivariate skew normal
GAS model

2000 Mathematics Subject Classification: 62H30.

ABSTRACT

This paper proposes a flexible finite mixture model framework using multivariate skew normal distribution for banking and credit institutions' stock data in Iran. This method clusters time series stocks data of Iranian banks and credit institutions to filter those data into four groups. The proposed model estimates matrices of time-varying parameter for skew normal distribution mixture using EM algorithm, updating the estimated parameters via generalized autoregressive score (GAS) model. Empirical studies are conducted to examine the effect of the proposed model in clustering, estimating, and updating parameters for real data from 12 sets of stocks. Our stock data were filtered in four trade clusters with best performance.

1. INTRODUCTION

In recent years, clustering algorithms for time series data have found significance because of having a good quality in different kind of applications.

It is specifically useful in stock filtering performance where big databases gathered from market stocks. These data have regularities which can be clustered automatically.

Stock of banks and credit institutions are very diverse in terms of stock trading value, trading volume, growth rate, the first price, the first opening volume, last price, close prices, and the rate of difference between high and low price. Those features help us in clustering time series data representing from market stock. Thus understanding heterogeneous features is of interest and key important in clustering the groups reliably.

Clustering of stocks provides strategies which help the trader of market stocks identify type of banks and credit institution's stock as the best candidates for buy, the best candidate for sell, as well as developing a watching list controlling for buy and sell.

In this study, we purpose a finite mixture model using Skew Normal Distribution for clustering high-dimensional time series data. The framework estimates a matrix of time-varying parameters and applies updates using score-driven approach proposed by Creal *et al.* [1], allowing us to robustly cluster the data into approximately homogenous groups.

The skew normal mixture distribution was considered in our proposed model. The main purpose of this model is to deal with data sets that may not be normal and our model is able to robustly cluster and with good performance the high-dimension data that may have an asymmetric- and/or heavy-tailed distribution.

Literature Review

Roengpitya *et al.* [2], Ayadi *et al.* [3], and Ayadi and Groen [4] illustrated cluster analysis to identify bank business models. Ahmadzade-hgoli [5] proposed The LINEX Weighted k-Means Clustering and Andre Lucas(2017) introduced a finite mixture model for multivariate normal and t distribution which updated parameters using score-driven approach. Creal *et al.* [1] and Harvey [6] introduced generalized autoregressive score (GAS) model for updating time-varying parameters while Ayadi and Groen [4] explained static clustering methods

*Corresponding author. Email: rfarnoosh@iust.ac.ir

with dynamic parameters. Catania [7] provided an example of dynamic clustering with dynamic parameters. He proposed a score-driven mixture model and used score-driven updates for all parameters that required a large number of observations.

Finally we applied our model to a multivariate panel of $N = 12$ stock data of banks and credit institutions for the period 2019/6- 2019/10, i.e., over $T = 90$ days in 18 week with $P = 8$ indicator variables for L groups of similar stock data of banks and credit institutions. We identified $L = 4$ trade model components and illustrated properties of each of group.

In addition, our study contributes to literature on statistic clustering of time series data for stocks (Roengpitya *et al.* [2], Ayadi *et al.* [3], and Ayadi and Groen [4]) by identifying stock trade model because we believe the properties of stock models are unlikely to switch their trade model over a short-term period (see, e.g., Ayadi and Groen [4]). This article is organized in 4 sections. In Section 2, we introduce the finite mixture model for skew normal distribution and estimate matrix of parameters using EM algorithm and updating parameters via GAS model through the score-driven approach. Section 3 explains an empirical study of stock data from banks and credit institutions in Iran, and a brief conclusion is presented in Section 4. Note that in this paper all of computations were run using R program.

2. INTRODUCING THE MODEL

2.1. Mixture Model

Let $\boldsymbol{y}_{it} \in \mathbb{R}^{P\times 1}$ be a multivariate panel data for the firms i = 1,2,…,N that contains p = 1,2,…,P characteristics for time t = 1,2,…,T. We show $\boldsymbol{y}_{it}$ by L-component mixture model as follows:

$$\boldsymbol{y}_{it} = \sum_{l=1}^{L} z_{il} \cdot \nu_{ilt} \quad i = 1, 2, \dots, N, t = 1, 2, \dots, T. \tag{1}$$

where z_{il} are hidden variables of the firm I. If the firm I is in the mixture component L then $z_{il} = 1$ otherwise $z_{il} = 0$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \dots, z_{il})' \in \{0, 1\}$ and $P(z_{il} = 1) = \omega_1$, where $\omega_1 + \omega_2 + \cdots + \omega_l = 1$. we define $\nu_{ilt} \sim f_l(.|\boldsymbol{\alpha}_{lt}, \boldsymbol{\beta}_{lt}, \gamma_l)$ where $\boldsymbol{\alpha}_{lt}$, $\boldsymbol{\beta}_{lt}$ and γ_l are mean, covariance matrices, and skewness parameters of skew normal distribution

It suffices to note that all observations were stacked into the matrix $\boldsymbol{Y}_{it} = (y_{i1}, y_{i2}, \dots, y_{it})' \in R^{(T\times P)}$ as parameters in each mixture component l with $\boldsymbol{\alpha}_{lt}$ and $\boldsymbol{\beta}_{lt}$ for all times t. It is important to note that we have two type of parameters: static and dynamic. Here $\boldsymbol{\Theta}$ contains all static parameters, like $\omega_l = \omega_l(\boldsymbol{\Theta})$, $\gamma_l = \gamma_l(\boldsymbol{\Theta})$ and $\boldsymbol{\theta}_l = \boldsymbol{\theta}_l(\boldsymbol{\Theta})$, for which we use the short-hand notion ω_l, γ_l, and $\boldsymbol{\theta}_l$ for simplicity. We specifically explain $\boldsymbol{\alpha}_{lt}$ and $\boldsymbol{\beta}_{lt}$ which are functions of past data only and updated using score-driven dynamics proposed by Creal *et al.* [1] while the values for γ_l are chosen to form time-invariant identity matrices ranging on the interval (−0.99, 0.99) (See Azzalini [8]).

To compute likelihood function by a standard prediction error we have

$$\log(L(\boldsymbol{\Theta})) = \sum_{i=1}^{N} \log \sum_{l=1}^{L} \omega_l f_l(\boldsymbol{Y}_{iT}; \boldsymbol{\theta}_l), \tag{2}$$

where

$$f_l(\boldsymbol{Y}_{iT}; \boldsymbol{\theta}_l) = \prod_{t=1}^{T} f_l(\boldsymbol{Y}_{it} | \boldsymbol{Y}_{i,t-1}; \boldsymbol{\theta}_{lt}), \tag{3}$$

and $f_l(\boldsymbol{Y}_{it}|\boldsymbol{Y}_{i,t-1}; \boldsymbol{\theta}_{lt})$ is the conditional distribution for the multivariate skew normal $\boldsymbol{Y} = \boldsymbol{\alpha} + \boldsymbol{\gamma\tau} + \boldsymbol{U}$ where $\boldsymbol{\tau}$ and $\boldsymbol{U}$ are independently distributed as $HN(0, \boldsymbol{I}_p)$and $N_p(0, \boldsymbol{\beta})$ respectively (see Lin [9]).

As is common in our model, we do not estimate $\boldsymbol{\Theta}$ directly by numerically maximizing the log-likelihood function in (2). To overcome this problem we use EM algorithm to estimate the parameters (see Dempster *et al.* [10]). To formulate the EM algorithm for dynamic parameters we need to define the complete data $(\boldsymbol{Y}_{iT}, \boldsymbol{\tau}_{iT}, \mathbf{z}_i)$ with the likelihood function

$$\begin{aligned} \log(L_c(\boldsymbol{\Theta})) = \sum_{i=1}^{N} \sum_{l=1}^{L} z_{il} \Big\{ & \log \omega_l - \log |\boldsymbol{\beta}_{lt}| - \frac{1}{2} (\boldsymbol{Y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_l \boldsymbol{\tau}_{it})^T \boldsymbol{\beta}_{lt}^{-1} (\boldsymbol{Y}_{it} - \boldsymbol{\alpha}_{lt} \\ & - \gamma_l \boldsymbol{\tau}_{it}) - \frac{1}{2} \boldsymbol{\tau}_{it}^T \boldsymbol{\tau}_{it} \Big\}, \end{aligned} \tag{4}$$

Since $\mathbf{z_i}$ is hidden indicator we cannot perform a direct maximization and instead we maximize its conditional expectation function over $\mathbf{z_i}$ given the observed data $\boldsymbol{Y}_T = (Y_{1T}, \dots, Y_{NT})$ and some previously known values for the parameter $\boldsymbol{\Theta}^{(k-1)}$. We maximize with respect to $\boldsymbol{\Theta}$ (Lin [9]).

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k-1)}) = E[\log L_c(\boldsymbol{\Theta})|\boldsymbol{Y}_T; \boldsymbol{\Theta}^{(k-1)}] = \sum_{i=1}^{N}\sum_{l=1}^{L}\sum_{t=1}^{T}\hat{z}_{il}^{(k-1)}\{\log \omega_l - \frac{1}{2}\log |\boldsymbol{\beta}_{lt}| - \frac{1}{2}(\boldsymbol{y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_l\hat{\boldsymbol{\eta}}_{ilt}^{(k-1)})^T\boldsymbol{\beta}_{lt}^{-1}(\boldsymbol{y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_l\hat{\boldsymbol{\eta}}_{ilt}^{(k-1)}) - \frac{1}{2}tr(\boldsymbol{\beta}_{lt}^{-1}\gamma_l(\hat{\boldsymbol{\Psi}}_{ilt}^{(k-1)} - \hat{\boldsymbol{\eta}}_{ilt}^{(k-1)}\hat{\boldsymbol{\eta}}_{ilt}^{(k-1)T})\gamma_l^T)\}, \tag{5}$$

where

$$\hat{\boldsymbol{\eta}}_{ilt} = E(\boldsymbol{\tau}_{it}), \hat{\boldsymbol{\Psi}}_{ilt} = E(\boldsymbol{\tau}_{it}\boldsymbol{\tau}_{it}^T).$$

In the E-Step, the hidden indicator probabilities are updated using

$$\hat{\vartheta}_{il}^{(k)} := P[z_{il} = 1|\boldsymbol{Y}_T, \boldsymbol{\Theta}^{(k-1)}] = \frac{\omega_l^{(k-1)} f_j(\boldsymbol{y}_{iT}; \boldsymbol{\theta}_l^{(k-1)})}{\sum_{h=1}^{L}\omega_h^{(k-1)} f_l(\boldsymbol{y}_{iT}; \boldsymbol{\theta}_h^{(k-1)})}, \tag{6}$$

It is important to note that $\vartheta_{il}^{(k)}$ does not depend on time because the stock trade model is unlikely to vary in a limited time. After updating $\vartheta_{il}^{(k)}$, we move to M-Step and maximize $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k-1)})$ with respect to ω_l (See Lucas [11]).

$$\hat{\omega}_l^{(K)} = \frac{1}{N}\sum_{i=1}^{N}\hat{\vartheta}_{il}^{(k-1)}. \tag{7}$$

2.2. Updating Dynamic Parameters

Now in this section we use the score-driven approach proposed by Creal [1] to formulate dynamic parameters $\boldsymbol{\alpha}_{lt}$ and $\boldsymbol{\beta}_{lt}$.

2.2.1. Mean

As explained above, we use the score-driven approach as discussed in Lucas and Zhang [12]:

$$\boldsymbol{\alpha}_{lt+1} = \boldsymbol{\alpha}_{lt} + AU_{\boldsymbol{\alpha}_{lt}}. \tag{8}$$

where $U_{\boldsymbol{\alpha}_{lt}}$ represents the first derivation of (5) with respect to $\boldsymbol{\alpha}_{lt}$ and $A = A(\boldsymbol{\Theta})$ is a diagonal matrix of unknown parameters. By a computation similar to the one found in Lucas [13] we compute

$$U_{\boldsymbol{\alpha}_{lt}} = \frac{\sum_{i=1}^{N}\vartheta_{il}(\boldsymbol{y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_l\boldsymbol{\eta}_{ilt})}{\sum_{i=1}^{N}\vartheta_{il}}. \tag{9}$$

Then we formulate updating mechanism as follows:

$$\boldsymbol{\alpha}_{lt+1} = \boldsymbol{\alpha}_{lt} + A\frac{\sum_{i=1}^{N}\vartheta_{il}(\boldsymbol{y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_j\boldsymbol{\eta}_{ilt})}{\sum_{i=1}^{N}\vartheta_{il}}. \tag{10}$$

2.2.2. Covariance matrix

Using the same calculations and the score-driven approach, we have

$$\boldsymbol{\beta}_{lt+1} = \boldsymbol{\beta}_{lt} + BU_{\boldsymbol{\beta}_{lt}}, \tag{11}$$

As before $U_{\boldsymbol{\beta}_{lt}}$ is the first derivation of (5) with respect to $\boldsymbol{\beta}_{lt}$ and $B = B(\boldsymbol{\Theta})$ is a diagonal matrix of unknown parameters. Doing the same calculation as the one used above we have

$$U_{\boldsymbol{\beta}_{lt}} = \frac{\frac{1}{2}\sum_{i=1}^{N}\vartheta_{il}^{(k)}\left((\boldsymbol{y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_l\boldsymbol{\eta}_{ilt})^T(\boldsymbol{y}_{it} - \boldsymbol{\alpha}_{lt} - \gamma_l\boldsymbol{\eta}_{ilt}) - \boldsymbol{\beta}_{lt}\right)}{\sum_{i=1}^{N}\vartheta_{il}^{(k)}}, \tag{12}$$

Then we formulate the updating mechanism as follows:

$$\boldsymbol{\beta}_{lt+1} = \boldsymbol{\beta}_{lt} + B\frac{\frac{1}{2}\sum_{i=1}^{N}\vartheta_{il}^{(k)}\left((\boldsymbol{y}_{it}-\boldsymbol{\alpha}_{lt}-\gamma_l\boldsymbol{\eta}_{ilt})^T(\boldsymbol{y}_{it}-\boldsymbol{\alpha}_{lt}-\gamma_l\boldsymbol{\eta}_{ilt})-\boldsymbol{\beta}_{lt}\right)}{\sum_{i=1}^{N}\vartheta_{il}^{(k)}}. \tag{13}$$

After updating parameters using equations (10) and (13), we compute $\vartheta_{il}^{(k)}$ by substitution in equation (6). Next, we maximize (5) with respect to A and B for computing those values. This step is iterated until a convergence is reached.

3. EMPIRICAL STUDY

3.1. Data

In this section we use an empirical example to examine the ability of our proposed model. The sample studied here contains $N = 12$ stocks of banks and credit institutions for the period 2019/6-2019/10. This covers $T = 90$ day. We accept that drivers in stocks trade model can be characterized by four dimensions as shown in Figure 1. The best candidate for buy, the best candidate for sell, the watching list controlling for buy and sell.

We select a set of $P = 8$ features from these four categories. We consider stocks' trading value, trading volume, growth rate, the first price, the first opening volume, last price, close prices, and the rate of difference between high and low price.

3.2. Model Selection

In this section, the number of clusters for our empirical analysis was selected using some of well-known criteria, i.e., Akaike information criterion (AIC), Bayesian information criterion (BIC), Davies-Bouldin index (DBI), and silhouette index (SI). The purpose of this criterion is to evaluate the structure of clusters created by clustering algorithms. Many criteria have been introduced to evaluate the accuracy of the clustering results.

These indices try to measure the similarity of members within the cluster and the similarity between the clusters. Therefore, the appropriate method is the one that results in the highest level of similarity within a cluster or the greatest differentiation between clusters.

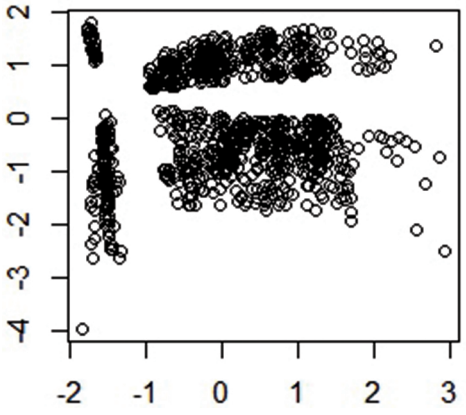

Figure 1 | Real stock data of banks and credit institutions in Iran

Table 1 | Information criteria.

Index	DBI	SI	AIC	BIC
$L = 2$	0.5769	0.5296	19.6562	**1901.23**
$L = 3$	0.5615	0.4552	**13.6587**	2851.70
$L = 4$	**0.5831**	**0.5570**	**17.5505**	**1802.14**
$L = 5$	**0.6683**	**0.6292**	21.4242	4752.62

Table 1 presents likelihood-based (AIC, BIC) and distance-based (DBI, SI) information criteria indices for different values of L = 2,...,5. The minimum value (AIC, BIC) and maximum value (DBI, SI) of components suggested L = 4.

As a likelihood-based model was utilized here, we used standard-likelihood-based criteria, including AIC and BIC, to determine the number of clusters (Hurvich and Tsai [14] and Bai and Ng [15]). The smaller are the values obtained for these two criteria, the more accurate will be the number of clusters. The silhouette index (SI, see de Amorim and Hennig [16] and Davies–Bouldin index (DBI, see Davies and Bouldin [17] criteria express the greatest similarity within a cluster, and larger values found for these two criteria indicate a better choice in terms of selecting the number of clusters. The results are presented in Table 1.

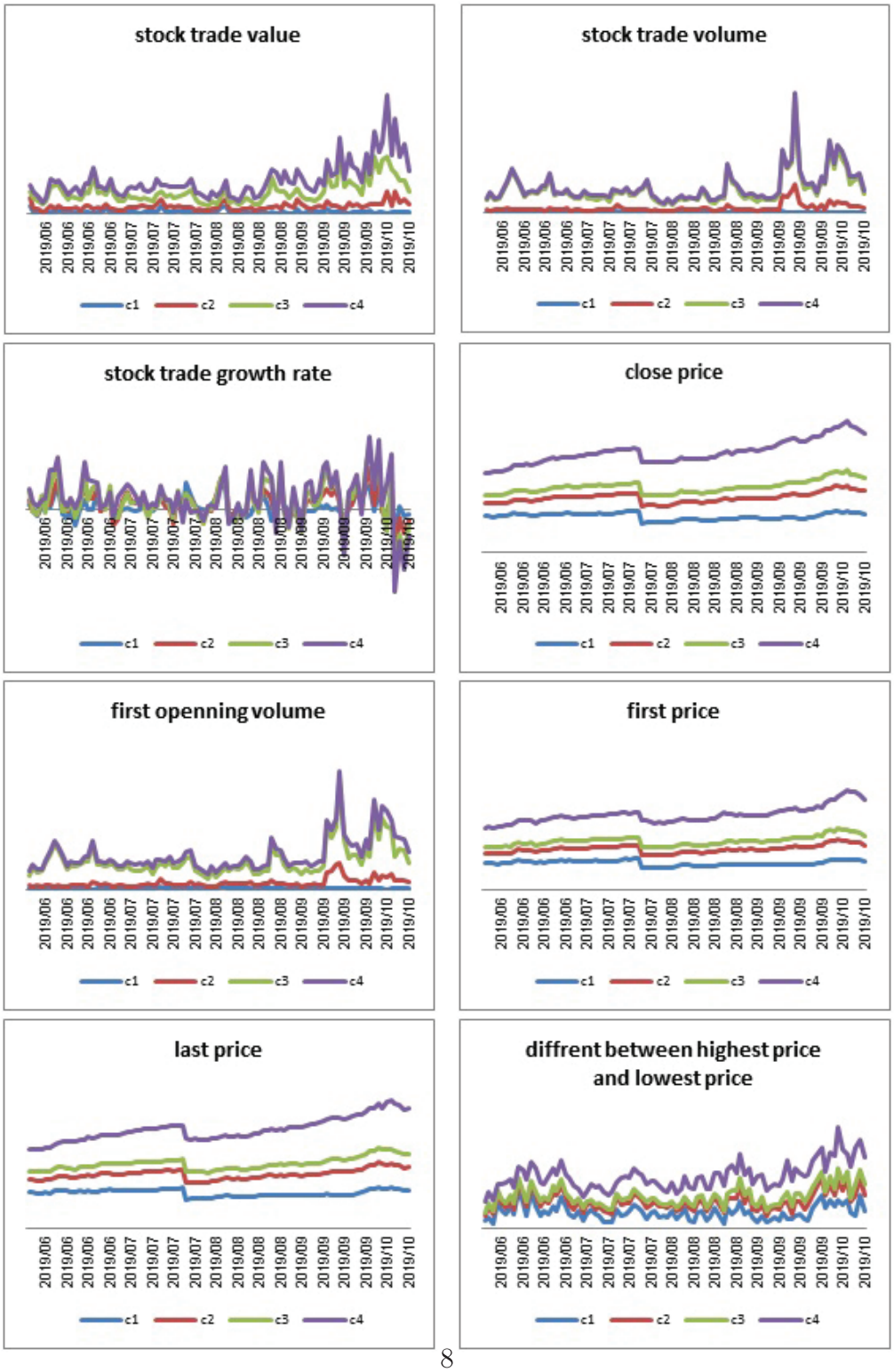

Figure 2 Time-varying component medians. C1: The best candidate for sell, C2: Watch list controlling for sell, C3: Watch list controlling for buy, C4: The best candidate for buy.

3.2.1. Discussion of stock' trade model

In this section, $L = 4$ different component densities are applied to different business models. We label a trade model on each cluster as illustrated in Figure 2 which plots the stock trade model for each feature characterization.

(C1) The best candidate for sell (8.33 of firms; e.g., Middle East Bank)

(C2) Watch list controlling for sell (41.66 of firms; e.g., Saderat Bank, Parsian Bank, Sina Bank, Karafarin Bank, Melal Credit Institution)

(C3) Watch list controlling for buy (16.67 of firms; e.g., Tejarat Bank, Pasargad Bank)

(C4) The best candidate for buy (33.34 of firms; e.g., Melat Bank, Eghtesad Novin Bank, Dey Bank, Post Bank)

The best candidate for sell (blue line): These stocks belong to banks and credit institutions that have the lowest trading volume, value of trade, and daily growth rate over a 90-day period. These stocks are the best choice for selling.

Watch list controlling for sell (red line): This cluster shows the stocks ranked as the second lowest in terms of volume, trading value, and daily growth rate over the same period. These stocks are best candidate on the watch list for sale.

Watch list controlling for buy (green line): These stocks belong to a category that ranks the second highest in terms of volume, trading value, and daily growth rate over the same period. These stocks are best placed on the watch list for purchase.

The best candidate for buy (Purple line): These banks and credit institutions have the highest trading volume, value of transactions, and daily growth rate over same time. These stocks are the best choice for buying.

4. CONCLUSION

We proposed a novel finite mixture model for studying stock data, constructing time-varying component parameters matrices, and providing a skew normal distribution mixture. The advantage of using this model over other models is its performance in robust clustering when dealing with any type of data. In an empirical example, we clustered 12 sets of stocks for Iranian banks and credit institutions into four trade model components. The result indicated clusters that recommend selling or buying and controlling for selling and buying.

ACKNOWLEDGMENTS

The authors acknowledge that this article is not in the "conflict of interest" and "author involvement" of others. There is also no "budget statement" for this article. We also appreciate from Referee and associate editor who led to a number of improvements.

REFERENCES

1. D. Creal, S. Koopman, A. Lucas, J. Appl. Econom. 28 (2013), 777–795.
2. R. Roengpitya, N. Tarashev, K. Tsatsaronis, Bank Business Models, BIS Quarterly Review, The bank for International settlement, 2014, pp. 55–65.
3. R. Ayadi, E. Arbak, W.P. de Groen, Business Models in European Banking: A Preand Post-Crisis Screening, CEPS Discussion Paper, Centre for European Policy Studies, 2014, pp. 1–104.
4. R. Ayadi, W.P.D. Groen, Bank Business Models Monitor Europe, CEPS Working Paper, The International Research Centre on Cooperative Finance, 2015, pp. 0–122.
5. N. Ahmadzadehgoli, A. Mohammadpour, M.H. Behzadi, J. Stat. Theory Appl. 18 (2019), 147–154.
6. A.C. Harvey, Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series. Econometric Society Monograph, Cambridge University Press, Cambridge. 2013.
7. L. Catania, Dynamic Adaptive Mixture Models, University of Rome Tor Vergata. Unpublished Working Paper, 2016. arXiv:1603.01308 [stat.ME].
8. A. Azzalini, Scand. J. Stat. 12 (1985), 171–178.
9. T.I. Lin, J. Multivar. Anal. 100 (2009), 257–265.
10. A.P. Dempster, N.M. Laird, D.B. Rubin, J. R. Stat. Soc. B. 39 (1977), 1–38.
11. A. Lucas, J. Schaumberg, B. Schwaab, Bank business models at zero interest rates, in Tinbergen Institute Discussion Paper, Taylor and Francis, TI 2016-066/IV, 2016. https://doi.org/10.2139/ssrn.2831922
12. A. Lucas, X. Zhang, Int. J. Forecast. 32 (2016), 293302.
13. A. Lucas, J. Schaumberg, B. Schwaab,Bank Business Model Satzero Interest Rates, Tinbergen Institute Discussion Paper, TI2016-066/IV, 2017.
14. C.M. Hurvich, C.-L. Tsai, Biometrika. 76 (1989), 297307.
15. J. Bai, S. Ng, Econometrica. 70 (2002), 191–221.
16. R.C. De Amorim, C. Hennig, Inf. Sci. 324 (2015), 126145.
17. D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979.