Int. J. Nonlinear Anal. Appl. 16 (2025) 4, 27–39
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.23427.2540

Static finite mixture model of multivariate skew-normal to cluster time series based on the GAS

Solmaz Yaghoubi[a], Rahman Farnoosh[b,*]

[a]Science and Research Branch, Islamic Azad University, Tehran, Iran

[b]School of Mathematics, Iran University of Science and Technology, Tehran, Iran

(Communicated by Javad Damirchi)

Abstract

This paper proposes an observation-driven finite mixture model for clustering high-dimension data. A simple algorithm using static hidden variables statically clusters the data into separate model components. The model accommodates normal and skew-normal distributed mixtures with time-varying component means, covariance matrices and skewness coefficient. These parameters are estimated using the EM algorithm and updated with the Generalized Autoregressive Scale (GAS) approach. Our proposed model is preferably clustered using a skew-normal distribution rather than a normal distribution when dealing with real data that may be skewed and asymmetrical. Finally, our proposed model will be evaluated using a simulation study and the results will be discussed using a real data set.

Keywords: Clustering, Finite mixture model, Skew normal distribution, Generalized autoregressive score, Time series
2020 MSC: 91C20, 93A30, 30D45, 30G35, 91B84

1 Introduction

Recently, much research has been done in the field of clustering using finite mixture models. Finite mixture models are linear combinations of components, and each of the components has its specific characteristics. Mixture models are a common method in clustering heterogeneous statistical populations.

Methodologically, our paper falls in the category of time-series data clustering. This clustering class consists of 4 parts; static and dynamic clustering methods for time series with both types of static and dynamic parameters.

The static clustering of time series, i.e. each time series during the clustering process belongs to only one cluster with the highest probability. The parameters of each cluster can be either constant (static) or time-varying (dynamic), Peel and McLachlan [21]. Wang et al. [24] apply static time series clustering with static parameters. Fruehwirth-Schnatter and Kaufmann [12] is an example of static clustering with both static and dynamic parameters. They cluster the time series into different groups of regression models with static parameters and then extended the model to static clusters in groups of different hidden Markov models (HMMs).

In contrast, in dynamic clustering, clusters can change the position of their cluster overtime to another cluster, depending on the sample being added, the parameters of each cluster can be either static or dynamic. Creal et.al [8]

[*]Corresponding author
Email addresses: yaghoubi.solmaz@gmail.com (Solmaz Yaghoubi), rfarnoosh@iust.ac.ir (Rahman Farnoosh)

Received: May 2021 *Accepted:* October 2021

proposed dynamic clustering with static parameters, and finally Catania [6] applied dynamic clustering with dynamic parameters.

Punzo and Maruotti [22] have introduced robust Dynamic Clustering Using Gaussian Hidden-Markov Model for Multivariate longitudinal Data in the presence of noise and outliers. Maruotti et al. (2019) [19] suggested Hidden Markov and semi-Hidden Markov models with multivariate Leptokurtic normal components. these models are suitable for dynamical time series clustering of data that have noise or outliers.

The mixture normal model is one of the most common clustering methods. Unfortunately, this model tends to over-fit the data. To overcome this weakness, Lin et al. [16] introduced a skew-normal mixture model.

In this paper, an observation-driven approach to the analysis of high-dimensional data using mixture models is presented. This framework includes static hidden variables for data clustering as separate clusters in mixture models, in which each component has dynamic parameters that allow data to be more reliably divided into almost homogeneous groups.

The proposed method of this article is in the static clustering class with dynamic parameters using the score-driven approach introduced by Creal et al. [7]. Here, we compare the performance of the multivariate normal mixture (MNM) to the multivariate skew normal mixture (MSNM) when dealing with real data that may not be normal and may be skewed. We show that the MSNM model performs is better.

In the proposed model, all dynamic parameters are estimated by using maximum-likelihood method based on the EM-algorithm and updated with the Generalized Autoregressive score-driven Approach (GAS) (Creal at.al [7] and Harvey [14].

Our simulation studies show that if the time-dependent mean of clusters is far apart, or if the mean distances are such that clusters overlap, the clustering performance of our model is acceptable in both cases. This accuracy in clustering is not affected by the type of simulation data (symmetric or asymmetric) and the performance of the MSNM model is good for any type of data. This proposed algorithm also categorizes well the road traffic data for the province of Ilam in Iran.

Therefore, our innovation in this paper is clustering using a new class of multivariate skew-normal distributions, in which model parameters are estimated by the EM algorithm and updated with the Generalized Autoregressive score-driven Approach.

This article is organized in 6 sections. In Section 2 we introduce the multivariate skew-normal mixture model and estimate parameters using EM algorithm. Section 3 explains the GAS model of updating parameters through the Score-Driven approach. In Section 4 we discuss the advantages of our proposed method using a simulation study. Section 5 provides an empirical example for clustering the roads in Ilam Province, Iran. In Section 6, the results of the proposed model are discussed. Note that in this article, all calculations are performed with the R program.

2 Basic model

2.1 Finite mixture model

Let $\boldsymbol{x}_{it} \in \mathbb{R}^{P\times 1}$ be a vector of multivariate panel data from firms $i = 1, 2, \dots, N$ for time $t = 1, 2, \dots, T$ with P observed characteristics. We model $\boldsymbol{x}_{it}$ of jth component mixture by:

$$\boldsymbol{x}_{it} = \sum_{j=1}^{J} z_{ij} \cdot \nu_{ijt} \quad t = 1, 2, \dots, T, \tag{2.1}$$

where $\nu_{ijt} \sim f_j(.|\boldsymbol{\mu}_{jt}, \boldsymbol{\Sigma}_{jt}, \boldsymbol{\Lambda}_{jt})$ is multivariate skew-normal distribution, $\boldsymbol{\mu}_{jt}$, $\boldsymbol{\Sigma}_{jt}$ are mean and covariance matrix respectively and $\boldsymbol{\Lambda}_{jt}$ is represent skewness parameter of mixture component $j = 1, \dots, J$ at time t. z_{ij} are unobserved indicators for the mixture component of firm i, $z_{ij} = 1$ if firm i is in mixture component j and $z_{ij} = 0$ otherwise. We define $\boldsymbol{z_i} = (z_{i1}, z_{i2}, \dots, z_{iJ})'$ as:

$$\boldsymbol{z_i} = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{iJ} \end{bmatrix} \in \left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\},$$

We assume $\boldsymbol{z_i}$ has a multinomial distribution with $\omega_j = p(z_{ij} = 1) \in [0, 1]$ and $\sum_{j=1}^{J} \omega_j = 1$. To write the likelihood of model in (1), the observations in time t, $\boldsymbol{X}_{it} = (x_{i1}, x_{i2}, \dots, x_{it})' \in \mathbb{R}^{T\times P}, t = 1, 2, ..., T$. There are two types of parameters, i.e. static and dynamic parameters. The dynamic parameters are represented by $\boldsymbol{\mu}_{jt}$, $\boldsymbol{\Sigma}_{jt}$ and $\boldsymbol{\Lambda}_{jt}$ for all time t. While any other parameter is incorporated into the vector $\boldsymbol{\theta}_j(\boldsymbol{\Theta})$ where $\boldsymbol{\Theta}$ gathers all static parameters of the model; for example $\omega_j(\boldsymbol{\Theta})$ represents functions of $\boldsymbol{\Theta}$ and for simplicity we use the short-hand notation ω_j and use $\boldsymbol{\theta}_j$ instead of $\omega_j(\boldsymbol{\Theta})$ and $\boldsymbol{\theta}_j(\boldsymbol{\Theta})$.

First of all, let us remind that in the mixture model in (2.1) the clustering characteristic $\boldsymbol{z}_i$ is described using time-invariant cluster indicators rather than using time-varying indicators to describe $\boldsymbol{z}_{it}$. That means, in our clustering approach, the selection of clusters does not depend on time. In our empirical example we use the time-invariant cluster indicator, because the road safety strategy developed by the Road Maintenance and Transportation Organization (RMTO) is unlikely to change their model over a limited period of time such as ours.

2.2 EM estimation

Using the above definition, the likelihood function is given by:

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{N}\Big(\sum_{j=1}^{J}\omega_j f_j(\boldsymbol{X}_{iT};\boldsymbol{\theta}_j)\Big), \tag{2.2}$$

$$\log(L(\boldsymbol{\Theta})) = \sum_{i=1}^{N}\log\sum_{j=1}^{J}\omega_j f_j(\boldsymbol{X}_{iT};\boldsymbol{\theta}_j), \tag{2.3}$$

where

$$f_j(\boldsymbol{X}_{iT};\boldsymbol{\theta}_j) = \prod_{t=1}^{T} f_j(\boldsymbol{X}_{it}|\boldsymbol{X}_{i,t-1};\boldsymbol{\theta}_{jt}), \tag{2.4}$$

and $f_j(\boldsymbol{X}_{it}|\boldsymbol{X}_{i,t-1};\boldsymbol{\theta}_{jt})$ represents the conditional multivariate skew normal distribution matrix with $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\vartheta} + \boldsymbol{U}$ where $\boldsymbol{\vartheta}$ and $\boldsymbol{U}$ are independently distributed as $HN(0, \boldsymbol{I}_p)$ and $N_p(0, \boldsymbol{\Sigma})$ respectively, $\boldsymbol{\Lambda}$ is represent skewness parameter. given the past data and given the parameters for time t as collected in $\boldsymbol{\theta}_{jt}$ (see [18]). In the following denote:

$$E(\boldsymbol{\vartheta}_{it}|\boldsymbol{x}_{ij}, z_{ij} = 1) = \boldsymbol{\eta}_{ijt} \quad \text{and} \quad E(\boldsymbol{\vartheta}_{it}\boldsymbol{\vartheta}'_{it}|\boldsymbol{x}_{ij}, z_{ij} = 1) = \boldsymbol{\psi}_{ijt}. \tag{2.5}$$

To write the likelihood function, we define the complete data for firm i as: $(\boldsymbol{X}_{iT};\boldsymbol{\vartheta}_{iT}, \boldsymbol{z_i})$. If $\boldsymbol{z_i}$ is known, the complete data likelihood function is:

$$\log(L_c(\boldsymbol{\Theta})) = \sum_{i=1}^{N}\sum_{j=1}^{J} z_{ij}[\log\omega_j + \log f_j(\boldsymbol{X}_{iT};\boldsymbol{\theta}_j)]. \tag{2.6}$$

Since $\boldsymbol{z_i}$ is hidden,(2.6) cannot be maximized directly [11]. Instead we maximize the conditional expectation function (2.6) and estimate the parameters using the EM algorithm. We maximize with respect to $\boldsymbol{\Theta}$ the conditional expectation function over $\boldsymbol{z_i}$ given the observed data $\boldsymbol{X}_T = (\boldsymbol{X}_{1,T}, \ldots, \boldsymbol{X}_{N,T})$ and some initial or previously determined parameter value $\boldsymbol{\Theta}^{(k-1)}$.

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k-1)}) &= E[\log L_c(\boldsymbol{\Theta})|\boldsymbol{X}_T;\boldsymbol{\Theta}^{(k-1)}] \\ &= E[\sum_{i=1}^{N}\sum_{j=1}^{J} z_{ij}[\log\omega_j + \log f_j(\boldsymbol{X}_{iT};\boldsymbol{\theta}_j)]|\boldsymbol{X}_T;\boldsymbol{\Theta}^{(k-1)}] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J} P[z_{ij} = 1|\boldsymbol{X}_T;\boldsymbol{\Theta}^{(k-1)}][\log\omega_j + \log f_j(\boldsymbol{X}_{iT};\boldsymbol{\theta}_j)]. \end{aligned} \tag{2.7}$$

Since we defined a multivariate skew normal distribution, its density function for random vector $\boldsymbol{X}$ is:

$$f(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = 2^p\phi_p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Omega})\Phi_p(\boldsymbol{\Lambda}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})|\Delta). \tag{2.8}$$

$\boldsymbol{X}$ is a p-dimensional skew normal distribution with a $p\times 1$ location vector $\boldsymbol{\mu}$, a $p\times p$ positive definite covariance matrix $\boldsymbol{\Sigma}$, and a $p\times p$ skewness matrix $\boldsymbol{\Lambda}$ and with $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$ and $\Delta = (\boldsymbol{I}_p + \Lambda^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) = \boldsymbol{I}_p - \boldsymbol{\Lambda}^T\boldsymbol{\Omega}\boldsymbol{\Lambda}$. Moreover, $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, $\phi_p(.|\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $\Phi_p(.|\boldsymbol{\Sigma})$ denotes the probability density function of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and cumulative density function of $N_p(0, \boldsymbol{\Sigma})$, respectively. This density function is distributed as $SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$.

$$\begin{aligned} f(\boldsymbol{X}_{it}, \boldsymbol{\vartheta}_{it}, \boldsymbol{z_i}) = &\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{T}(\frac{1}{\sqrt{2\pi\boldsymbol{\Sigma}_{jt}}} \exp\{\frac{1}{2}(\boldsymbol{X}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt}\boldsymbol{\vartheta}_{it})^T\boldsymbol{\Sigma}_{jt}^{-1} \\ &(\boldsymbol{X}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt}\boldsymbol{\vartheta}_{it})\})^{z_{ij}} \cdot \prod_{i=1}^{N}\prod_{j=1}^{J}(\frac{2}{\sqrt{2\pi}} exp\{\frac{1}{2}\boldsymbol{\vartheta}_{it}^T\boldsymbol{\vartheta}_{it}\})^{z_{ij}} \\ &\cdot \prod_{i=1}^{N}\prod_{j=1}^{J}(\omega_j)^{z_{ij}}. \end{aligned} \tag{2.9}$$

Thus, here the complete data is $(\boldsymbol{X}_{it}, \boldsymbol{\vartheta}_{jt}, \boldsymbol{z_i})$ and from (2.6), the complete-data log-likelihood function of $\boldsymbol{\Theta}$ is given by:

$$\begin{aligned} \log L_c(\boldsymbol{\Theta}) = &\sum_{i=1}^{N}\sum_{j=1}^{J} z_{ij}\{\log\omega_j - \frac{1}{2}\log|\boldsymbol{\Sigma}_{jt}| - \frac{1}{2}(\boldsymbol{X}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt}\boldsymbol{\vartheta}_{it})^T\boldsymbol{\Sigma}_{jt}^{-1} \\ &(\boldsymbol{X}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt}\boldsymbol{\vartheta}_{it}) - \frac{1}{2}\boldsymbol{\vartheta}_{it}^T\boldsymbol{\vartheta}_{it}\}. \end{aligned} \tag{2.10}$$

Formally, the Expectation-step of the EM algorithm requires Q-function.

3 Expectation-step

The conditional component indicator probability is updated using:

$$P[z_{ij} = 1|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)}] = E(z_{ij}|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)}) = \frac{\omega_j^{(k-1)} f_j(\boldsymbol{x}_{iT}; \boldsymbol{\theta}_j^{(k-1)})}{\sum_{h=1}^{J}\omega_h^{(k-1)} f_h(\boldsymbol{x}_{iT}; \boldsymbol{\theta}_h^{(k-1)})} = \hat{\tau}_{ij}^{(k)}, \tag{3.1}$$

and we have to compute:

$$\begin{aligned} E(z_{ij}, \boldsymbol{\vartheta}_{it}|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)}) &= E(z_{ij}|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)})E(\boldsymbol{\vartheta}_{it}|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)}) \\ &= \hat{\tau}_{ij}^{(k)}\hat{\boldsymbol{\eta}}_{ij}^{(k)} \qquad (3.2) \\ E(z_{ij}, \boldsymbol{\vartheta}_{it}\boldsymbol{\vartheta}_{it}^T|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)}) &= E(z_{ij}|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)})E(\boldsymbol{\vartheta}_{it}\boldsymbol{\vartheta}_{it}^T|\boldsymbol{X}_T, \boldsymbol{\Theta}^{(k-1)}) \\ &= \hat{\tau}_{ij}^{(k)}\hat{\boldsymbol{\psi}}_{ijt}^{(k)}, \qquad (3.3) \end{aligned}$$

where $\hat{\boldsymbol{\eta}_{ijt}^{(k)}}$, $\hat{\boldsymbol{\psi}_{ijt}^{(k)}}$ are $\boldsymbol{\eta}_{ijt}^{(k)}$ and $\boldsymbol{\psi}_{ijt}^{(k)}$ in(2.5) respectively. The Q-function can be written:

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k-1)}) = E[\log L_c(\boldsymbol{\Theta})|\boldsymbol{X}_T; \boldsymbol{\Theta}^{(k-1)}] = &\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{T}\hat{\tau_{ij}}^{(k-1)}\Big\{\log\omega_j - \frac{1}{2}\log|\boldsymbol{\Sigma}_{jt}| \\ &- \frac{1}{2}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt} \quad \hat{\boldsymbol{\eta}}_{ijt}^{(k-1)})^T\boldsymbol{\Sigma}_{jt}^{-1}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt}\hat{\boldsymbol{\eta}}_{ijt}^{(k-1)})\} \\ &- \frac{1}{2}tr\Big(\Sigma_{jt}^{-1}\boldsymbol{\Lambda}_{jt}(\hat{\boldsymbol{\psi}}_{ijt}^{(k-1)} - \hat{\boldsymbol{\eta}}_{ijt}^{(k-1)}\hat{\boldsymbol{\eta}}_{ijt}{}^{(k-1)^T})\boldsymbol{\Lambda}_{jt}^T\}\Big)\Big\}. \end{aligned} \tag{3.4}$$

We again note that the $\tau_{ij}^{(k)}$ does not depend on time. Once $\tau_{ij}^{(k)}$ is updated, we move to the Maximizing-step.

4 Maximizing-step

In this step we optimize the Q function for time t=T as fallows:

$$\hat{\omega}_j^{(K)} = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_{ij}^{(k-1)}. \tag{4.1}$$

$$\hat{\boldsymbol{\mu}}_{jT}^{(K)} = \left(\sum_{i=1}^{N} \hat{\tau}_{ij}^{(k-1)} \boldsymbol{x}_{iT} - \hat{\boldsymbol{\Lambda}}_{jT}^{(k-1)} \sum_{i=1}^{N} \hat{\tau}_{ij}^{(k-1)} \boldsymbol{\eta}_{ijT}^{(k-1)} \right) \Big/ \sum_{i=1}^{N} \hat{\tau}_{ij}^{(k-1)}. \tag{4.2}$$

Fix $\boldsymbol{\mu}_{jT} = \hat{\boldsymbol{\mu}}_{jT}^{(K)}$ and compute:

$$\hat{\boldsymbol{\Lambda}}_{jT}^{(K)} = \left(\sum_{i=1}^{N} \hat{\tau}_{ij}^{(k-1)} (\boldsymbol{x}_{iT} - \hat{\boldsymbol{\mu}}_{jT}^{(k)}) \hat{\boldsymbol{\eta}}_{ijT}{}^{(k-1)^T} \right) \left(\sum_{i=1}^{N}) \hat{\boldsymbol{\eta}}_{ijT}^{(k-1)} \hat{\boldsymbol{\psi}}_{ijT}^{(k-1)} \right)^{(-1)}. \tag{4.3}$$

Fix $\boldsymbol{\mu}_{jT} = \hat{\boldsymbol{\mu}}_{jT}^{(K)}$ and $\boldsymbol{\Lambda}_{jT} = \hat{\boldsymbol{\Lambda}}_{jT}^{(K)}$ and compute:

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{jT}^{(K)} = \frac{1}{\sum_{i=1}^{N} \hat{\tau}_{ij}^{(k-1)}} \Bigg\{ & \sum_{i=1}^{N} \hat{\tau}_{ij}^{(k-1)} (\boldsymbol{x}_{iT} - \hat{\boldsymbol{\mu}}_{jT}^{(k)} - \hat{\boldsymbol{\Lambda}}_{jT}^{(k)} \hat{\boldsymbol{\eta}}_{ijT}^{(k-1)}) (\boldsymbol{x}_{iT} - \hat{\boldsymbol{\mu}}_{jT}^{(k)} - \hat{\boldsymbol{\Lambda}}_{jT}^{(k)} \\ & \hat{\boldsymbol{\eta}}_{ijT}^{(k-1)})^T + \hat{\boldsymbol{\Lambda}}_{jT}^{(K)} \left(\sum_{j=1}^{N} \hat{\tau}_{ij}^{(k-1)} (\hat{\boldsymbol{\psi}}_{ijT}^{(k-1)} - \hat{\boldsymbol{\eta}}_{ijT}^{(k-1)} \hat{\boldsymbol{\eta}}_{ijT}{}^{(k-1)^T}) \right) \hat{\boldsymbol{\Lambda}}_{jT}{}^{(K)^T} \Bigg\}. \end{aligned} \tag{4.4}$$

5 Score-driven model

As explained before, we use the static clustering with dynamic parameters $\boldsymbol{\mu}_{jt}$, $\boldsymbol{\Sigma}_{jt}$ and $\boldsymbol{\Lambda}_{jt}$. In this section, we formulate score-driven parameter dynamics for $\boldsymbol{\mu}_{jt}$, $\boldsymbol{\Sigma}_{jt}$ and $\boldsymbol{\Lambda}_{jt}$.

5.1 Time-varying mean

Now we introduce a score-driven approach proposed by Creal et al. [7] to explain the updating mechanism for parameter dynamics of $\boldsymbol{\mu_{jt}}$.

$$\boldsymbol{\mu}_{jt+1} = \boldsymbol{\mu}_{jt} + \alpha_1 SC_{\boldsymbol{\mu}_{jt}}. \tag{5.1}$$

where $SC_{\boldsymbol{\mu}_{j\,t}}$ represents the scaled first derivative of the expected likelihood functions with respect to $\boldsymbol{\mu}_{j\,t}$, and $\alpha_1 = \alpha_1(\boldsymbol{\Theta})$ is a diagonal matrix that depends on the unknown parameter vector ($\boldsymbol{\Theta}$). The scale is given by:

$$\begin{aligned} \nabla_{\boldsymbol{\mu}_{jt}} &= \frac{\partial}{\partial_{\boldsymbol{\mu}_{jt}}} \left(\sum_{i=1}^{N} \sum_{j=1}^{J} \tau_{ij}^{(k)} \log\, \phi(\boldsymbol{X}_{it}; \boldsymbol{\mu}_{jt}, \boldsymbol{\Sigma}_j, \boldsymbol{\Lambda}_j) \right) \\ &= \frac{\partial}{\partial_{\boldsymbol{\mu}_{jt}}} \left(\sum_{i=1}^{N} \sum_{j=1}^{J} \tau_{ij}^{(k)} \Bigg\{ \log(\omega_j) - \frac{1}{2} \log |\boldsymbol{\Sigma_j}| - \frac{1}{2} (\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j \hat{\boldsymbol{\eta}}_{ijt}^{(k)})^T \right. \\ &\quad \times \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j \hat{\boldsymbol{\eta}}_{ijt}^{(k)}) - \frac{1}{2} tr \left(\boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Lambda}_j (\hat{\boldsymbol{\psi}}_{ijt}^{(k)} - \hat{\boldsymbol{\eta}}_{ijt}^{(k)} \hat{\boldsymbol{\eta}}_{ijt}{}^{(k)^T}) \boldsymbol{\Lambda}_j^T \right) \Bigg\} \\ \nabla_{\mu_{jt}} &= \boldsymbol{\Sigma}_j^{-1} \sum_{i=1}^{N} \tau_{ij}^{(k)} (\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j \boldsymbol{\eta}_{ijt}). \end{aligned} \tag{5.2}$$

To score our scale, we compute the inverse of an expected negative Hessian under mixture component j to obtain $\dfrac{1}{\boldsymbol{\Sigma}_j{}^{-1} \sum_{i=1}^{N} \tau_{ij}^{(k)}}$. Our scale score is in form:

$$SC_{\boldsymbol{\mu}_{jt}} = \frac{\sum_{i=1}^{N} \tau_{ij} (\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j \boldsymbol{\eta}_{ijt})}{\sum_{i=1}^{N} \tau_{ij}}. \tag{5.3}$$

The updating mechanism is:

$$\boldsymbol{\mu}_{j,t+1} = \boldsymbol{\mu}_{j,t} + \alpha_1 \frac{\sum_{i=1}^{N} \tau_{ij} (\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j \boldsymbol{\eta}_{ijt})}{\sum_{i=1}^{N} \tau_{ij}}. \tag{5.4}$$

Now we can estimate, using EM algorithm, all static parameters by starting from an initial $\Theta^{(k-1)}$ and $\mu^{(k-1)}$ then computing $\mu^{(k-1)}, t = 2, \ldots, T$ by recursion (5.4).

5.2 Time-varying covariance matrices

This section derives the scaled score updates for time-varying component covariance matrices. We have

$$\boldsymbol{\Sigma}_{jt+1} = \boldsymbol{\Sigma}_{jt} + \alpha_2 SC_{\boldsymbol{\Sigma}_{jt}}. \tag{5.5}$$

where $SC_{\Sigma_{jt}}$ is again defined as the scaled fist derivative of the expected likelihood function with respect to $\boldsymbol{\Sigma}_{jt}$, and $\alpha_2 = \alpha_2(\boldsymbol{\Theta})$. Following the equation (3.4) the scale is given by:

$$\begin{aligned}
\nabla_{\boldsymbol{\Sigma}_{j\,t}} &= \frac{\partial}{\partial_{\boldsymbol{\Sigma}_{jt}}}\left(\sum_{i=1}^{N}\sum_{j=1}^{J}\tau_{ij}^{(k)}\log\,\phi(\boldsymbol{X}_{it};\boldsymbol{\mu}_{jt},\boldsymbol{\Sigma}_{jt},\boldsymbol{\Lambda}_j)\right)\\
&= \frac{\partial}{\partial_{\boldsymbol{\Sigma}_{jt}}}\left(\sum_{i=1}^{N}\sum_{j=1}^{J}\tau_{ij}^{(k)}\left\{\log(\omega_j) - \frac{1}{2}\log|\boldsymbol{\Sigma_{jt}}| - \frac{1}{2}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\hat{\boldsymbol{\eta}}_{ijt}^{(k)})^T\right.\right.\\
&\left.\times\,\boldsymbol{\Sigma}_{jt}^{-1}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\hat{\boldsymbol{\eta}}_{ijt}^{(k)}) - \frac{1}{2}tr(\boldsymbol{\Sigma}_{jt}^{-1}\boldsymbol{\Lambda}_j(\hat{\boldsymbol{\psi}}_{ijt}^{(k)} - \hat{\boldsymbol{\eta}}_{ijt}^{(k)}\hat{\boldsymbol{\eta}}_{ijt}{}^{(k)^T})\boldsymbol{\Lambda}_j^T)\right\}\\
&= \frac{1}{2}\sum_{i=1}^{N}\tau_{ij}^{(k)}\boldsymbol{\Sigma}_{jt}^{-1}\Big((\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})^T - \boldsymbol{\Sigma}_{jt}\Big)\boldsymbol{\Sigma}_{jt}^{-1}.
\end{aligned} \tag{5.6}$$

Taking the total differential of (5.6), and then taking expectations, we obtain:

$$\begin{aligned}
&\frac{1}{2}E\Big[\sum_{i=1}^{N}\sum_{j=1}^{J}\tau_{jt}^{(k)}\Big(d\,\boldsymbol{\Sigma}_{jt}^{-1}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})^T\boldsymbol{\Sigma}_{jt}^{-1}\\
&+\boldsymbol{\Sigma}_{jt}^{-1}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})^T d\,\boldsymbol{\Sigma}_{jt} + d\,\boldsymbol{\Sigma}_{jt}^{-1}\Big)\Big]\\
&= \frac{1}{2}\sum_{i=1}^{N}\tau_{ij}^{(k)}d\,\boldsymbol{\Sigma}_{jt}^{-1} = -\sum_{i=1}^{N}\frac{1}{2}\tau_{ij}^{(K)}(\boldsymbol{\Sigma}_{jt}^{-1}\,d\boldsymbol{\Sigma}_{jt}\,\boldsymbol{\Sigma}_{jt}^{-1}).
\end{aligned} \tag{5.7}$$

we obtain $-\frac{1}{2}\sum_{i=1}^{N}\tau_{ij}^{(K)}(\boldsymbol{\Sigma}_{jt}\bigotimes\boldsymbol{\Sigma}_{jt})^{-1}vec(d\boldsymbol{\Sigma}_{jt})$, (where the operator $\bigotimes$ denotes the Kronecker product). After finding the negative inverse of this equation, we can compute the scaled score:

$$\begin{aligned}
vec(SC_{\boldsymbol{\Sigma}_{jt}}) &= (\frac{1}{2}\sum_{i=1}^{N}\tau_{jt}^{(k)})^{-1}(\boldsymbol{\Sigma}_{jt}\bigotimes\boldsymbol{\Sigma}_{jt})\cdot vec(\nabla_{\boldsymbol{\Sigma}_{jt}})\\
&= (\sum_{i=1}^{N}\tau_{jt}^{(k)})^{-1}\cdot vec(2\boldsymbol{\Sigma}_{jt}\nabla_{\boldsymbol{\Sigma}_{jt}}\boldsymbol{\Sigma}_{jt}) \Leftrightarrow\\
SC_{\boldsymbol{\Sigma}_{jt}} &= \frac{\sum_{i=1}^{N}\tau_{ij}^{(k)}\Big((\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ijt})^T - \boldsymbol{\Sigma}_{jt})}{\sum_{i=1}^{N}\tau_{ij}^{(k)}}.
\end{aligned} \tag{5.8}$$

Then we can update the $\boldsymbol{\Sigma}_{jt}{}^{(k)}$ using (5.8).

5.3 Time-varying skewness matrices

As explained above, the scaled score updates for time-varying skewness matrix is given by:

$$\boldsymbol{\Lambda}_{j\,t+1} = \boldsymbol{\Lambda}_{j\,t} + \alpha_3 SC_{\boldsymbol{\Lambda}_{j\,t}}. \tag{5.9}$$

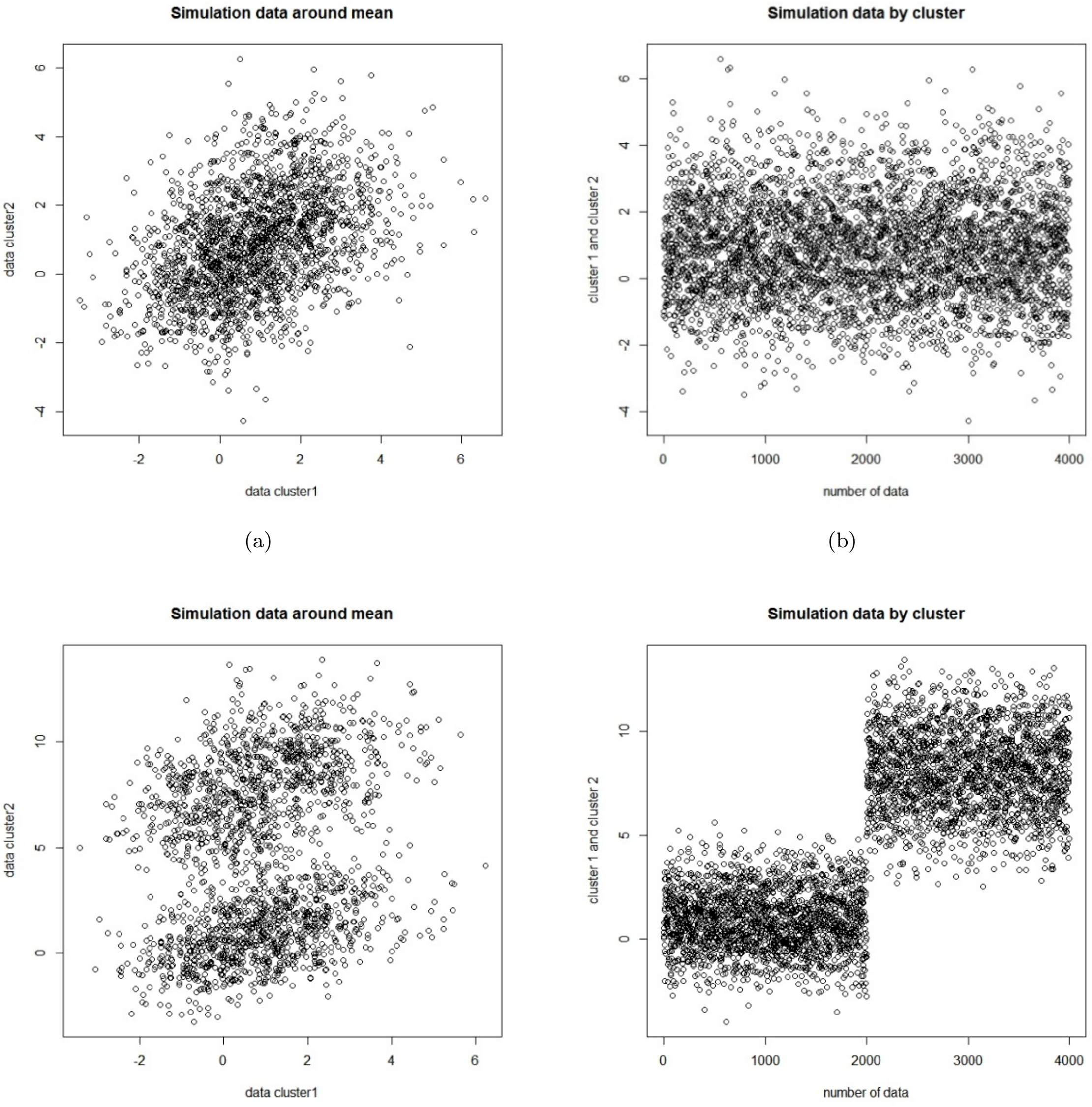
where $SC_{\boldsymbol{\Lambda}_{j\,t}}$ is the same scale as for the previous covariance matrices and mean. By repeating the calculations above, the scale is given by:

$$SC_{\boldsymbol{\Lambda}_{jt}} = \frac{\sum_{i=1}^{N}\tau_{ij}^{(k)}(\boldsymbol{x}_{it} - \boldsymbol{\mu}_{jt} - \boldsymbol{\Lambda}_{jt}\boldsymbol{\eta}_{ijt})\boldsymbol{\eta}_{ijt}}{\sum_{i=1}^{N}\tau_{ij}^{(k)}\boldsymbol{\eta}_{ijt}^2}, \tag{5.10}$$

Then we can update the parameter using (5.10). Next in the Maximizing-step (3.4) is maximized with respect α_1, α_2 and α_3. Following this process, we compute $\tau_{ij}^{(k)}$. The Expectation-step and Maximizing-step are iterated until convergence to zero is reached.

6 Simulation

In order to evaluate the performance of our proposed mixture model, we ran simulations for different types of data. This section explains the advantage of our method over clustering multivariate time series.



(a) (b)

(c) (d)

Figure 1: Simulation data in 2 types: Skew normal simulations overlapping data(a),(b) and Skew normal simulations non-overlapping data (c),(d)

We simulate data from a mixture of two dynamic bivariate skew normal densities. These densities contain mean functions in the form of sinusoid and i.i.d disturbance from bivariate normal and skew normal distribution. The covariance matrices are constant identity matrices, and in the case of the skew normal distribution the skewness coefficient are chosen to be time-invariant identity matrices (see [1]).

We chose a sample size of $N = 100$. The number of clusters is fixed at $J = 2$ while the number of time points considered is $T = 20$. We generate the data from two components around the mean that move over time. The simulation is carried out in two positions which move in overlapping and non-overlapping circles over time. In this way, we have four types of simulation data as shown in Table 1 and Figure 1.

Table 1: Type of simulation data

	Simulation	Radius	Distance
Overlapping data	Normal	2	2
	Skew normal	2	2
Non-overlapping data	Normal	2	8
	Skew normal	2	8

Since real data may not be normal and may be skewed, we use both normal and skew-normal distributions in data simulation and compare the performance of multivariate normal mixture (MNM) and multivariate skew-normal mixture (MSNM) models in the presence of both data types. The data are correctly classified by the MSNM model in both types of simulation. However, if the data are generated as skew-normal, the MNM model clusters the data incorrectly as show in Table 2 (S stands for Static, D stands for Dynamic and C stands for cluster).

Table 2: percentage of classification for each cluster

Multivariate Normal mixture model

			Overlapping data		Non-overlapping data	
simulation	M	Σ	$\%C1$	$\%C2$	$\%C1$	$\%C2$
Normal	S	S	correct specification		correct specification	
			50%	50%	50%	50%
	S	D	correct specification		correct specification	
			50%	50%	50%	50%
	D	S	correct specification		correct specification	
			50%	50%	50%	50%
	D	D	correct specification		correct specification	
			50%	50%	50%	50%
Skew Normal	S	S	correct specification		correct specification	
			65%	35%	51%	49%
	S	D	correct specification		correct specification	
			65%	35%	51%	49%
	D	S	correct specification		correct specification	
			65%	35%	51%	49%
	D	D	correct specification		correct specification	
			65%	35%	51%	49%

Since our model does time series clustering statically and only the parameters of each cluster depend on time, for a better investigation, we showed all the possible states of the parameters in Table 2. That is, the multivariate skew normal mixture (MSNM) model, which has 3 parameters of mean, variance and skewness coefficient, includes 8 different modes that the parameters can be static or dynamic. In the multivariate normal mixture (MNM) model, we have 2 parameters of mean and variance, which includes 4 different states of static and dynamic parameters. As it's showed in Table 2, in the MSNM model, clustering has a better performance than the MNM model, and by changing the parameters to static or dynamic, the performance of the MSNM model is still better than the MNM model. In this model, the clustering of data simulated with the normal distribution is completely correct and the clustering of data simulated with the skew normal distribution has a very low error. However in the MNM model, although clustering of normal simulated data is error-free, it is very error-prone compared to skew-normal data and is not reliable at all. This shows that the MSNM model that is proposed performs better than the MNM model in dealing with data that is skewed or asymmetric.

The box plot in figure 2 shows the classification error. Figure 2 demonstrates that clustering error of our model is smaller than that of the MNM model when dealing with skewed or asymmetric data.

6.1 Estimation parameters

We use ML estimation by updating score function for our proposed model in a set of independent random samples $x_1, x_2, \ldots, x_n$. The ML are obtained as:

$$\hat{\Theta} = \text{argmax}_{\theta} l(\theta|\boldsymbol{X}). \tag{6.1}$$

Multivariate Skew normal mixture model

	Overlapping data					Non-overlapping data	
simulation	M	Σ	Λ	$\%C1$	$\%C2$	$\%C1$	$\%C2$
Normal	S	S	S	correct specification		correct specification	
				50%	50%	50%	50%
	S	D	S	correct specification		correct specification	
				50%	50%	50%	50%
	D	S	S	correct specification		correct specification	
				50%	50%	50%	50%
	S	S	D	correct specification		correct specification	
				50%	50%	50%	50%
	D	D	S	correct specification		correct specification	
				50%	50%	50%	50%
	D	S	D	correct specification		correct specification	
				50%	50%	50%	50%
	S	D	D	correct specification		correct specification	
				50%	50%	50%	50%
	D	D	D	correct specification		correct specification	
				50%	50%	50%	50%
Skew Normal	S	S	S	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	S	D	S	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	D	S	S	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	S	S	D	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	D	D	S	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	D	S	D	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	S	D	D	correct specification		correct specification	
				50.009%	49.991%	50%	50%
	D	D	D	correct specification		correct specification	
				50.009%	49.991%	50%	50%

We adopt the EM algorithm to determine $\hat{\Theta}$.The ML estimation converges after a number of iterations. The EM algorithm is iterated until $|\text{Ł}\boldsymbol{\Theta}^{(k)} - \text{Ł}\boldsymbol{\Theta}^{(k-1)}|$ converges to zero. Figure 3 illustrates the convergence of estimated parameters.

7 Empirical example

Traffic data often contains hidden relationships which can be a certain cause of various types of traffic accidents. Safety can be enhanced on roads by revealing these hidden relationships through data analysis and clustering the roads accordingly.

In this section, we examine clustering models for 18 roads located in Ilam Province, Iran. The data collected for these roads cover the time period 2017-2018. This implies $T = 20$ for weekly reports. We assume that differences in road's models can be characterized along 3 components which is shown in Figure 4 and Table 3. We selected a set of 6 indicators for these 3 categories. These indicators include:

1-Traffic rate for light vehicles.

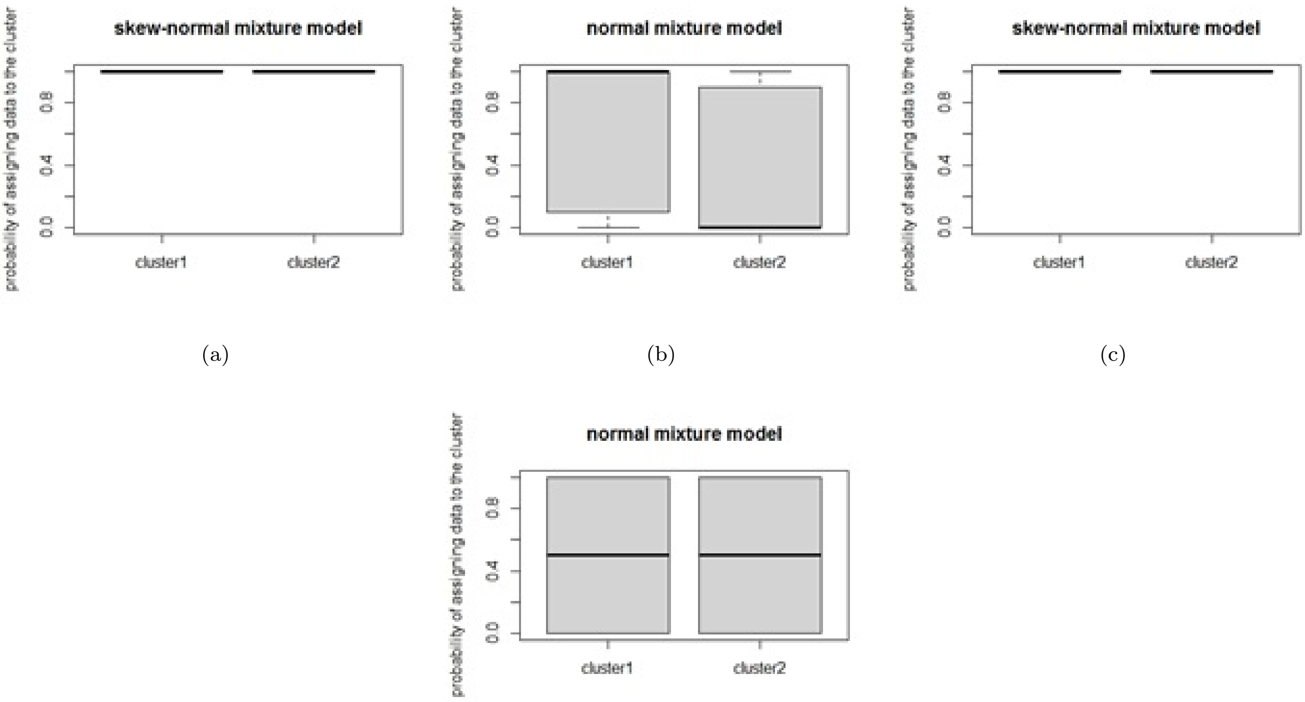
2-Traffic rate for heavy vehicles.

3-Average Speed rate.

4-Illegal overtaking rate.

5-Speed limit violation.

6-Safe distance rule violation.



(a)

(b)

(c)

(d)

Figure 2: Box plot showing clustering error for two types of data: skew normal simulations for overlapping data (a) and (b), skew normal simulations for non-overlapping data (c) and (d).

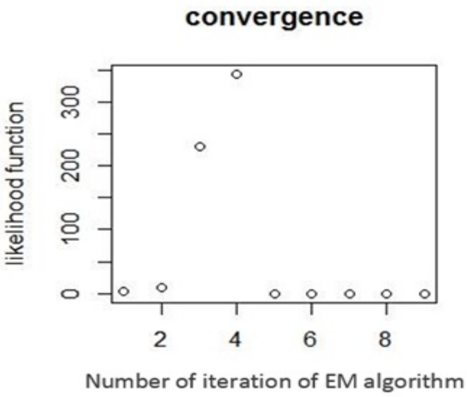


Figure 3: convergence of parameters estimation by using EM algorithm

7.1 Model selection

In this section, we decide about the number of clusters by using the following famous criteria: AIC, BIC, DBI and SI. As our framework is likelihood-based, we can apply standard information criteria such as AICc and BIC for model selection [15] and [4].

The Davies-Bouldin index (DBI, see [9]) also utilizes the Mahalanobis distance. The silhouette index (SI, see [10]) measures how well each observation fits into its respective cluster. The purpose of an internal criteria index is to evaluate the structure of clusters by clustering algorithm.

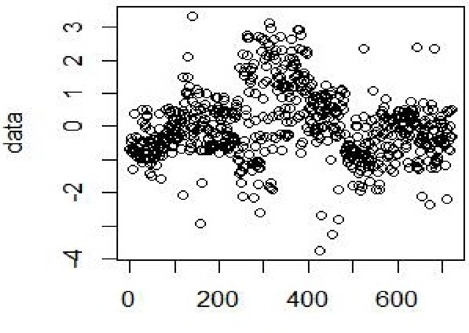


Figure 4: Real data of Ilam province roads for two individual indicators

Table 3: Information criteria

Index	DBI	SI	AIC	BIC
j=2	0.5870	**0.7476**	85.51	1220.99
j=3	**0.6962**	**0.7029**	**17.51**	**183.66**
j=4	**0.7183**	0.6701	**22.49**	**244.79**
j=5	0.6723	0.6366	26.49	305.81

As stated above, the purpose of clustering is to create groups with the highest internal resemblance (DBI, SI) and the least similarity between the categories (AIC, BIC). As a result, in the clustering internal evaluation approach, the combination of distance or similarity functions for clusters is used. The results are presented in Table 3.

Finally the category of road was:

	[Group1]	[Group2]	[Group3]
[road1]	0	0	1
[road2]	0	0	1
[road3]	1	0	0
[road4]	1	0	0
[road5]	1	0	0
[road6]	1	0	0
[road7]	1	0	0
[road8]	1	0	0
[road9]	0	1	0
[road10]	0	1	0
[road11]	1	0	0
[road12]	1	0	0
[road13]	0	0	1
[road14]	0	0	1
[road15]	0	0	1
[road16]	1	0	0
[road17]	0	0	1
[road18]	0	0	1

As seen in Figure 5, following real data clustering, the box plot showed a smaller error for our proposed model.

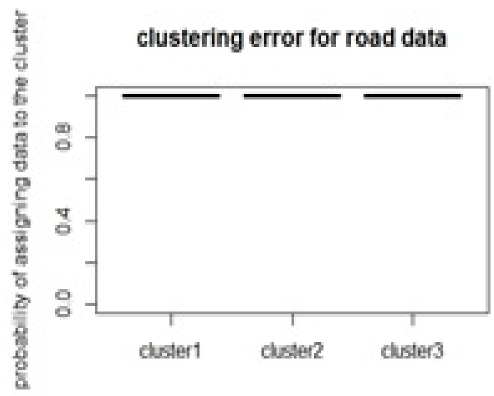


Figure 5: Box plot showing clustering error for Real data of Ilam province roads

8 Conclusion

Our proposed finite mixture model by using multivariate skew-normal distribution and score-driven approach is a novel method for static clustering time series by dynamic parameters. The results show that when dealing with real data that may not be normal and maybe skewed or asymmetric, the multivariate skew-normal mixture model (MSNM) produces better clustering results. We also show that our proposed model performance is acceptable in both types of the simulation that may be data overlapped or non-overlapped rather than a multivariate normal mixture model (MNM). We have proved our assertion through a simulation study and an empirical example provided in the study of a sample of roads taken for this purpose.

References

[1] A. Azzalini, *A class of distributions which includes the normal ones*, Scand. J. Statist. **12** (1985), 171–178.

[2] A. Azzalini and A.D. Valle, *The multivariate skew-normal distribution*, Biometrika **83** (1996), no. 2, 715–726.

[3] C.C. Aggarwal and C.K. Reddy, *Data Clustering*, Algorithms and Applications, Chapman and Hall/CRC, 2014.

[4] J. Bai and S. Ng, *Determining the number of factors in approximate factor models*, Econometrica **70** (2002), no. 1, 191–221.

[5] M.A. Benjamin, R.A. Rigby, and D.M. Stasinopoulos, *Generalized autoregressive moving average models*, J. Amer. Statist. Assoc. **98** (2003), no. 461, 214–223.

[6] L. Catania, *Dynamic adaptive mixture models*, University of Rome Tor Vergata, arXiv preprint arXiv:1603.01308, 2016.

[7] D. Creal, S. Koopman, and A. Lucas, *Generalized autoregressive score models with applications*, J. Appl. Economet. **28**, (2013) no. 5, 777–795.

[8] D. Creal, B. Schwaab, S.J. Koopman, and A. Lucas, *An observation driven mixed measurement dynamic factor model with application to credit risk*, Rev. Econ. Statist. **96** (2014), no. 5, 898–915.

[9] D.L. Davies and D.W. Bouldin, *A cluster separation measure*, IEEE Trans. Pattern Anal. Machine Intell. **2** (1979), 224–227.

[10] R.C. De Amorim and C. Hennig, *Recovering the number of clusters in data sets with noise features using feature rescaling factors*, Inf. Sci. **324** (2015), no. 10, 126–145.

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Statist. Soc.: Ser. B **39** (1977), no. 1, 1–22.

[12] S. Fruehwirth-Schnatter and S. Kaufmann, *Model-based clustering of multiple time series*, J. Bus. Econ. Statist. **26** (2008), 78–89.

[13] A. Hajrajabi and M. Maleki, *Nonlinear semiparametric autoregressive model with finite mixtures of scale mixtures of skew normal innovations*, J. Appl. Statist. **46** (2019), no. 11, 2010–2029.

[14] A.C. Harvey, *Dynamic Models for Volatility and Heavy Tails, with Applications to Financial and Economic Time Series*, Cambridge University Press, 2013.

[15] C.M. Hurvich and C.-L. Tsai, *Regression and time series model selection in small samples*, Biometrika **76** (1989), 297–307.

[16] T.I. Lin, J.C. Lee, and W.J. Hsieh, *Robust mixture modeling using the skew t distribution*, Statist. Comput. **17** (2007), 81–92.

[17] T.I. Lin, J.C. Lee, and S.Y. Yen, *Finite mixture modelling using the skew normal distribution*, Statist. Sin. **17** (2007), 909–927.

[18] T.I. Lin, *Maximum likelihood estimation for multivariate skew normal mixture models*, J. Multivar. Anal. **100** (2009), no. 2, 257–265.

[19] A. Maruotti, A. Punzo, and L. Bagnato, *Hidden Markov and semi-Markov models with multivariate leptokurtic-normal components for robust modeling of daily returns series*, J. Financ. Economet. **17** (2019), no. 1, 91–117.

[20] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2000.

[21] D. Peel and G.J. McLachlan, *Robust mixture modelling using the t distribution*, Statist. Comput. **10** (2000), 339–348.

[22] A. Punzo and A. Maruotti, *Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model*, J. Comput. Graph. Statist. **25** (2016), no. 4, 1097–1116.

[23] N. Shephard, *Generalized linear Autoregressions*, Nuffield College, University of Oxford, 1995.

[24] Y. Wang, R.S. Tsay, J. Ledolter, and K.M. Shrestha, *Forecasting simultaneously highdimensional time series: A robust model-based clustering approach*, J. Forecast. **32** (2013), no. 8, 673–684.