

# Variable Selection in Finite Mixture of Time-Varying Regression Models

Jing Liu, Wanzhou Ye

College of Science, Shanghai University, Shanghai, China

Email: wzhy@shu.edu.cn

**How to cite this paper:** Liu, J. and Ye, W.Z. (2020) Variable Selection in Finite Mixture of Time-Varying Regression Models. *Advances in Pure Mathematics*, 10, 101-113.

<https://doi.org/10.4236/apm.2020.103007>

**Received:** January 30, 2020

**Accepted:** March 3, 2020

**Published:** March 6, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## Abstract

In this paper, we research the regression problem of time series data from heterogeneous populations on the basis of the finite mixture regression model. We propose two finite mixed time-varying regression models to solve this. A regularization method for variable selection of the models is proposed, which is a mixture of the appropriate penalty functions and  $l_2$  penalty. A Block-wise minimization maximization (MM) algorithm is used for maximum penalized log quasi-likelihood estimation of these models. The procedure is illustrated by analyzing simulations and with an application to analyze the behavior of urban vehicular traffic of the city of São Paulo in the period from 14 to 18 December 2009, which shows that the proposed models outperform the FMR models.

## Keywords

Mixture Regression Models, GARCH, Block-Wise MM algorithm, LASSO, SCAD

## 1. Introduction

The problem of variable selection in FMR models has been widely discussed [1] [2] [3]. When a response variable  $y$  with a finite mixture distribution depends on covariates  $x$ , we obtain a finite mixture of regression (FMR) model. The FMR model with  $K$  components can be given as follows [3]:

$$f(y; \mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k f(y; \eta_k(\mathbf{x}), \phi_k) \quad (1)$$

where  $y$  is an independent and identically distributed (IID) response and  $x$  is a  $p \times 1$  vector of covariates.  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)^\top$  denotes the mixing proportions satisfying  $0 < \pi_k < 1$ ,  $\sum_{k=1}^K \pi_k = 1$ .  $f(y; \eta_k(\mathbf{x}), \phi_k)$  is the  $k$ th mixture

component density.  $\eta_k(\mathbf{x}) = h(\alpha_k + \mathbf{x}^T \boldsymbol{\beta}_k)$  for  $k = 1, \dots, K$ , for a given link function  $h(\cdot)$ , and a dispersion parameter  $\phi_k$ .

However, in some situations, observations were not independent. As pointed out in [2], in the analysis of the PD data, observations from each patient over time were assumed to be independent to facilitate the analysis and comparison with results from the literature. However, the validity of such assumption may be questionable. Whereupon, we consider a situation that observations were time series.

The generalised autoregressive conditional heteroskedasticity (GARCH) model is widely used in time series analysis. A mixture generalized autoregressive conditional heteroscedastic (MGARCH) model was pointed out in [4]. [5] generalized the MixN-GARCH model by relaxing the assumption of constant mixing weights. Whereupon, we combine the GARCH model and the FMR model to discuss the above problem.

There has been extensive studies about variable selection methods. A recent review of the literature regarding the variable selection problem in FMR models can be found in [6]. There are a general family of penalty functions, including the least absolute shrinkage and selection operator (LASSO), the minimax concave penalty (MCP) and the smoothly clipped absolute deviation (SCAD) in [2] and [7].

The method of the maximum penalized log-likelihood (MPL) estimation is usually the EM algorithm. [8] proposed a new algorithm (block-wise MM) for the MPL estimation of the L-MLR model. It was proved to have some desirable features such as coordinate-wise updates of parameters, monotonicity of the penalized likelihood sequence, and global convergence of the estimates to a stationary point of the penalized loglikelihood function, which are missing in the commonly used approximate-EM algorithm presented in [3].

The rest of the paper is organized as follows: in Section 2, the definition of finite mixture of time-varying regression Models and in Section 3, feature selection methods are discussed. In Section 4, the block-wise MM algorithm for its estimation and the BIC for choosing tuning parameters and components are presented, and the example of the Gaussian distribution is derived. Simulation studies on the performance of the new variable selection methods are then provided in Section 5. In Section 6, analysis of a real data set illustrates the use of the procedure. Finally, conclusions are given in Section 7.

## 2. Finite Mixture of Time-Varying Regression Models

### 2.1. Finite Mixture of Autoregression Models

Let  $\{y_t; t = 1, \dots, n\}$  be a response variable which is a time series.  $\{\mathbf{x}_t; t = 1, \dots, n\}$  is a  $p$ -dimensional vector of covariates, and each of them is a time series. For an FM-AR( $d$ ) model with  $K$  components, the conditional density function for observation  $t$  is given as follows:

$$f(y_t; \mathbf{x}_t, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k f(y_t; \eta_k(\mathbf{x}_t), \phi_k), \quad (2)$$

where

$$\eta_k(\mathbf{x}_t) = h(\alpha_k + \mathbf{x}_t^T \boldsymbol{\beta}_{k1} + \mathbf{x}_{t-1}^T \boldsymbol{\beta}_{k2} + \dots + \mathbf{x}_{t-d}^T \boldsymbol{\beta}_{kd}), \tag{3}$$

for  $k = 1, \dots, K$ , for a given link function  $h(\cdot)$ , and a dispersion parameter  $\phi_{kt}$ .

The master vector of all parameters is given by  $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, \boldsymbol{\alpha}^T, \boldsymbol{\phi}^T, \boldsymbol{\beta}^T)^T$ , with

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{11} & \dots & \boldsymbol{\beta}_{1d} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\beta}_{K1} & \dots & \boldsymbol{\beta}_{Kd} \end{pmatrix}, \tag{4}$$

where  $\boldsymbol{\beta}_{ki} = (\beta_{ki1}, \dots, \beta_{kip})^T \in \mathbb{R}^p$ ,  $i = 1, \dots, d$ . Let  $\tilde{\mathbf{x}}_t = (\mathbf{x}_t^T, \mathbf{x}_{t-1}^T, \dots, \mathbf{x}_{t-d}^T)$ , and  $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}_{k1}, \dots, \boldsymbol{\beta}_{kd})^T$ , (3) can be rewrote as  $\eta_k(\mathbf{x}_t) = h(\alpha_k + \tilde{\mathbf{x}}_t \tilde{\boldsymbol{\beta}})$ .

### 2.2. Finite Mixture of GARCH Models

Let  $\{y_t; t = 1, \dots, n\}$  be a response variable which is a time series. Let  $\{\mathbf{x}_t; t = 1, \dots, n\}$  is a  $p$ -dimensional vector of covariates, and each of them is a time series. For some distributions with unequal dispersion parameter  $\phi_k$ , we propose the FM-GARCH models. For an FM-GARCH  $(d, M, S)$  model with  $K$  components, the conditional density function for observation  $t$  is given as follows:

$$f(y_t; \mathbf{x}_t, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k f(y_t; \eta_k(\mathbf{x}_t), \phi_{kt}), \tag{5}$$

where  $\eta_k(\mathbf{x}_t) = h(\alpha_k + \tilde{\mathbf{x}}_t \tilde{\boldsymbol{\beta}})$  for  $k = 1, \dots, K$ , for a given link function  $h(\cdot)$ , and a conditional heteroscedastic (a dispersion parameter)

$$\phi_{kt} = \gamma_{0k} + \sum_{m=1}^M \gamma_{km} \epsilon_{k,t-m} + \sum_{s=1}^S \delta_{ks} \phi_{k,t-s}, \tag{6}$$

where  $\gamma_{0k} > 0$ ,  $\gamma_{km} \geq 0$ ,  $\delta_{ks} \geq 0$ , and  $\epsilon_{kt} = \phi_{kt} e_{kt}$ ,  $e_{kt}$  is an independent and identically distributed series with mean zero and variance unity.

The master vector of all parameters is given by  $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, \boldsymbol{\alpha}^T, \boldsymbol{\gamma}_0^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$ , with  $\boldsymbol{\gamma}_0 = (\gamma_{01}, \dots, \gamma_{0K})^T$ ,  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K)^T$ ,  $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, \dots, \gamma_{km})^T$ , and  $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_K)^T$ ,  $\boldsymbol{\delta}_k = (\delta_{k1}, \delta_{k2}, \dots, \delta_{ks})^T$ .

### 3. Feature Selection Method

Let  $\{(\mathbf{x}_t, y_t); t = 1, \dots, n\}$  be a sample of observations from the FM-AR or FM-GARCH model. The quasi-likelihood function of the parameter  $\boldsymbol{\theta}$  is given by [9]

$$L_n(\boldsymbol{\theta}) = \prod_{t=1}^n f(y_t; \mathbf{x}_t, \boldsymbol{\theta}) = \prod_{t=1}^n \left\{ \sum_{k=1}^K \pi_k f(y_t; \eta_k(\mathbf{x}_t), \phi_{kt}) \right\}. \tag{7}$$

The log quasi-likelihood function of the parameter  $\boldsymbol{\theta}$  is given by

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{t=1}^n \log \sum_{k=1}^K \pi_k f(y_t; \eta_k(\mathbf{x}_t), \phi_{kt}). \tag{8}$$

When the effect of a component of  $\mathbf{x}$  is not significant, the corresponding

ordinary maximum quasi-likelihood estimate is often close to 0, but not equal to 0. Thus this covariate is not excluded from the model. Inspired by an idea of [2], we estimate  $\theta$  by maximizing the penalized log quasi-likelihood function (MPLQ) for the model

$$\mathcal{F}_n(\theta) = \mathcal{L}_n(\theta) - \mathcal{P}_n(\theta), \tag{9}$$

with the mixture penalty (or regularization) function:

$$\mathcal{P}_{nk}(\theta) = \sum_{k=1}^K \pi_k \sum_{i=1}^d \sum_{j=1}^p p_n(\beta_{kij}; \lambda_{nk}) + \frac{1}{2} \sum_{k=1}^K \pi_k \sum_{i=1}^d \sum_{j=1}^p \nu_{nk} \beta_{kij}^2, \tag{10}$$

for some ridge tuning parameter  $\nu_{nk} \geq 0$ , and  $p_n(\beta_{kij}; \lambda_{nk})$  is a nonnegative penalty function. In the penalty function  $\mathcal{P}_n(\theta)$ , the amount of  $l_2$  penalty imposed on the componentwise regression coefficients  $\beta_{kij}$ 's are chosen proportional to  $\pi_k$ . The functions  $p_n(\beta_{kij}; \lambda_{nk})$  are designed to identify the no significant coefficients  $\beta_{kij}$ 's in the mixture components  $f(y_t; \eta_i(x_t), \phi_{kt})$ . General regularity conditions about the  $p_n(\beta_{kij}; \lambda_{nk})$  is given in [2] [3].

We estimate the new method using the following well-known penalty (or regularization) functions:

- LASSO penalty:  $p_n(\beta; \lambda_{nk}) = \lambda_{nk} |\beta|$ .
- MCP penalty:  $p'_n(\beta; \lambda_{nk}) = (\lambda_{nk} - nb_{nk} |\beta|)_+$ .
- SCAD penalty:

$$p'_n(\beta; \lambda_{nk}) = \lambda_{nk} I(n|\beta| < \lambda_{nk}) + \frac{(a_{nk} \lambda_{nk} - n|\beta|)_+}{a_{nk} - 1} I(n|\beta| > \lambda_{nk}).$$

Here,  $I$  is the indicative function. The constant  $a_{nk} \geq 2$  and  $b_{nk} \geq 0$  pointed in [2], and LASSO tuning parameter  $\lambda_{nk} \geq 0$ , which controls the amount of penalty. The asymptotic properties about these penalty functions can be analogously derived in [3] and [2]. We call the penalty function  $\mathcal{P}_{nk}(\theta)$  in (10) constructed from LASSO, MCP, SCAD jointly with the mixed  $L_2$ -norm as MIXLASSO-ML<sub>2</sub>, MIXMCP-ML<sub>2</sub>, MIXSCAD-ML<sub>2</sub> penalties.

### 4. Numerical Solutions

A new method for maximizing the penalized log-likelihood function is the block-wise Minorization Maximization (MM) algorithm inspired by [8], which is also known as block successive lower-bound maximization (BSLM) algorithm in the language of [10]. At each iteration of the method, the function is maximized with respect to a single block of variables while the rest of the blocks are held fixed. We shall now proceed to describe the general framework of the algorithm.

#### 4.1. Maximization of the Penalized Log-Likelihood Function

We follow the approach of [8] and minorize the  $\varepsilon$ -approximate of  $-\mathcal{P}_n(\theta)$  by

$$G_1(\theta; \theta^{(r)}) = -\frac{1}{2} \sum_{k=1}^K \pi_k \sum_{j=1}^d \sum_{i=1}^p p_n \left( \frac{\beta_{ijk}^2}{w_{ijk}^{(r)}}; \lambda_{ni} \right) - \frac{1}{2} \sum_{k=1}^K \pi_k \sum_{j=1}^d \sum_{i=1}^p \nu_{nk} \beta_{ijk}^2 + C_1(\theta^{(r)}), \tag{11}$$

where  $w_{ijk}^{(r)} = \sqrt{\beta_{ijk}^{2(r)} + \varepsilon^2}$ , for some  $\varepsilon > 0$ , and

$$C_1(\boldsymbol{\theta}^{(r)}) = -\frac{\varepsilon^2}{2} \sum_{k=1}^K \pi_i \sum_{j=1}^d \sum_{l=1}^p p_n^{-1}(w_{ijk}^{(r)}; \lambda_{ni}) - \frac{1}{2} \sum_{k=1}^K \pi_i \sum_{j=1}^d \sum_{l=1}^p p_n(w_{ijk}^{(r)}; \lambda_{ni}). \tag{12}$$

Moreover, minorize the log quasi-likelihood function  $\mathcal{L}_n(\boldsymbol{\theta})$  by

$$G_2(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \sum_{k=1}^K \sum_{t=1}^n \tau_{kt}^{(r)} \log \pi_i + \sum_{k=1}^K \sum_{t=1}^n \tau_{kt}^{(r)} \log f(y_i; \eta_i(x_t), \phi_{kt}^{(r)}) - \sum_{k=1}^K \sum_{t=1}^n \tau_{kt}^{(r)} \log \tau_{kt}^{(r)}, \tag{13}$$

where  $\tau_{kt}^{(r)} = \pi_i^{(r)} f(y_i; \eta_i^{(r)}(x_t), \phi_{kt}^{(r)}) / f(y_i; x_t, \boldsymbol{\theta}^{(r)})$ .

Note that  $\tau_{kt}^{(r)}$  and  $G_2(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$  are analogous to the posterior probability and the expected complete-data log-likelihood function of the expectation-maximization algorithm respectively.

The block-wise MM algorithm maximizes  $\mathcal{F}_n(\boldsymbol{\theta})$  iteratively in the following two steps:

- Block-wise Minorization-step. Conditioned on the  $r$ th iterate  $\boldsymbol{\theta}^{(r)}$ , the FM-GARCH model can be block-wise minorized in the coordinates of the parameter components  $\boldsymbol{\pi}$ ,  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\gamma}_0$ ,  $\boldsymbol{\gamma}$ ,  $\boldsymbol{\delta}$ , and  $\boldsymbol{\beta}$ , via the minorizers

$$G_{\boldsymbol{\pi}}(\boldsymbol{\pi}; \boldsymbol{\theta}) = G_2(\boldsymbol{\pi}, \boldsymbol{\alpha}^{(r)}, \boldsymbol{\gamma}_0^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\delta}^{(r)}; \boldsymbol{\theta}^{(r)}) - \mathcal{P}_n(\boldsymbol{\pi}, \boldsymbol{\beta}^{(r)}), \tag{14}$$

$$G_{\boldsymbol{\alpha}, \boldsymbol{\gamma}_0}(\boldsymbol{\alpha}, \boldsymbol{\gamma}_0; \boldsymbol{\theta}^{(r)}) = G_2(\boldsymbol{\pi}^{(r)}, \boldsymbol{\alpha}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\delta}^{(r)}; \boldsymbol{\theta}^{(r)}) - \mathcal{P}_n(\boldsymbol{\theta}^{(r)}), \tag{15}$$

$$G_{\boldsymbol{\gamma}, \boldsymbol{\delta}}(\boldsymbol{\gamma}, \boldsymbol{\delta}; \boldsymbol{\theta}^{(r)}) = G_2(\boldsymbol{\pi}^{(r)}, \boldsymbol{\alpha}^{(r)}, \boldsymbol{\gamma}_0^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}, \boldsymbol{\delta}; \boldsymbol{\theta}^{(r)}) - \mathcal{P}_n(\boldsymbol{\theta}^{(r)}), \tag{16}$$

$$G_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{\theta}^{(r)}) = G_1(\boldsymbol{\pi}^{(r)}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(r)}) + G_2(\boldsymbol{\pi}^{(r)}, \boldsymbol{\alpha}^{(r)}, \boldsymbol{\gamma}_0^{(r)}, \boldsymbol{\beta}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\delta}^{(r)}; \boldsymbol{\theta}^{(r)}), \tag{17}$$

respectively. Similar block-wise minorized can be made for FM-AR model.

- Block-wise Maximization-step. Upon finding the appropriate set of block-wise minorizers of  $\mathcal{F}_n(\boldsymbol{\theta})$ , we can maximize (14) to compute the  $(r + 1)$ th iterate block-wise update of  $\boldsymbol{\pi}$ . Solving for the appropriate root of the FOC (first-order condition) for the Lagrangian, we can compute the  $(r + 1)$ th iterate block-wise update

$$\pi_k^{(r+1)} = \frac{\sum_{t=1}^n \tau_{kt}^{(r)}}{\zeta^* + z_k}, \tag{18}$$

for each  $k$ , where  $z_k = \sum_{i=1}^d \sum_{j=1}^p p_n(\beta_{kij}; \lambda_{ni}) + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^p v_{ni} \beta_{kij}^2$ , and  $\zeta^*$  is the unique root of

$$\sum_{k=1}^K \frac{\sum_{t=1}^n \tau_{kt}^{(r)}}{\zeta^* + z_k} - 1 = 0, \tag{19}$$

in the interval  $(z^*, \infty)$ , and  $z^* = -\min_{k=1, \dots, K} \{z_k\}$ .

The block-wise updates for  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\gamma}_0$ ,  $\boldsymbol{\gamma}$ ,  $\boldsymbol{\delta}$ , and  $\boldsymbol{\beta}$  can be obtained by solving (15)-(17) via the first-order condition equal to 0.

We now present an example of the Gaussian FM-GARCH model to specify the procedure described above, and give the following Lemma 1 about a useful minorizer for the MPL estimation of the Gaussian FM-GARCH model, which can be found in [11].

**Lemma 1** *if  $\Theta = (0, \infty)$ , then the function  $1/\sum_{i=1}^q c_i \theta_i$  satisfy that*

$$\frac{1}{\sum_{i=1}^q c_i \theta_i} \leq \sum_{i=1}^q \frac{c_i \varphi_i^2}{\left(\sum_{i=1}^q c_i \varphi_i\right)^2 \theta_i} \tag{20}$$

**Example 1** *We consider the Gaussian FM-GARCH Model,*

$$f(y_t; \mathbf{x}_t, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k N(y_t; \eta_k(\mathbf{x}_t), \sigma_{kt}^2), \tag{21}$$

where  $\eta_k(\mathbf{x}_t) = h(\alpha_k + \tilde{\mathbf{x}}_t \tilde{\boldsymbol{\beta}}_k)$ , and  $\sigma_{kt}^2 = \gamma_{0k} + \sum_{m=1}^M \gamma_{km} \varepsilon_{k,t-m}^2 + \sum_{s=1}^S \delta_{ks} \sigma_{k,t-s}^2$ .

Here,  $\varepsilon_{kt} = \sigma_{kt} e_{kt}$ , and  $e_{kt}$  is an independent and identically distributed series with mean zero and variance unity.

According to [8], and using Lemma 1, we can obtain the further minorizer of Gaussian FM-GARCH by

$$\begin{aligned} \tilde{G}_2(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) &= \sum_{k=1}^K \sum_{t=1}^n \tau_{kt}^{(r)} \log \pi_k - \frac{1}{2} \sum_{k=1}^K \sum_{t=1}^n \tau_{kt}^{(r)} \log \sigma_{kt}^2 \\ &\quad - \frac{1}{2pd} \sum_{k=1}^K \sum_{j=1}^{pd} \sum_{t=1}^n \frac{\tau_{kt}^{(r)}}{\sigma_{kt}^2} \left( y_t - \alpha_k - pdx_{ij} (\beta_{kj} - \beta_{kj}^{(r)}) - \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)} \right)^2 \\ &\quad + C_2(\boldsymbol{\theta}^{(r)}), \end{aligned} \tag{22}$$

where  $\tau_{kt}^{(r)} = \pi_k^{(r)} N(y_t; \alpha_k^{(r)} + \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)}, \sigma_{kt}^{2(r)}) / \sum_{k=1}^K \pi_k^{(r)} N(y_t; \alpha_k^{(r)} + \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)}, \sigma_{kt}^{2(r)})$ , and

$$C_2(\boldsymbol{\theta}^{(r)}) = -\frac{n}{2} \log(2\pi) - \sum_{k=1}^K \sum_{t=1}^n \tau_{kt}^{(r)} \log \tau_{kt}^{(r)}.$$

The block-wise updates of  $\boldsymbol{\pi}$  from Gaussian FM-GARCH Model come from (18), and the block-wise updates for  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\gamma}$ , and  $\boldsymbol{\delta}$ , can be obtained from (15)-(16) via the first-order condition equal to 0. By doing so, we obtain the coordinate-wise updates for  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\gamma}_0$  block

$$\alpha_k^{(r+1)} = \frac{\sum_{t=1}^n \tau_{kt}^{(r)} (y_t - \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)}) / \sigma_{kt}^{2(r)}}{\sum_{t=1}^n \tau_{kt}^{(r)} / \sigma_{kt}^{2(r)}}, \tag{23}$$

$$\gamma_{0k}^{(r+1)} = \frac{\sum_{t=1}^n \tau_{kt}^{(r)} \gamma_{0k}^{2(r)} (y_t - \alpha_k^{(r+1)} - \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)})^2 / (\sigma_{kt}^{2(r)})^2}{\sum_{t=1}^n \tau_{kt}^{(r)} \gamma_{0k}^{(r)} / \sigma_{kt}^{2(r)}}, \tag{24}$$

for each  $k$ . Moreover, the coordinate-wise updates for the  $\boldsymbol{\gamma}$  and  $\boldsymbol{\delta}$  block

$$\gamma_{km}^{(r+1)} = \frac{\sum_{t=1}^n \tau_{kt}^{(r)} \gamma_{km}^{2(r)} \varepsilon_{k,t-m}^2 (y_t - \alpha_k^{(r+1)} - \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)})^2 / (\sigma_{kt}^{2(r)})^2}{\sum_{t=1}^n \tau_{kt}^{(r)} \gamma_{km}^{(r)} \varepsilon_{k,t-m}^2 / \sigma_{kt}^{2(r)}}, \tag{25}$$

$$\delta_{ks}^{(r+1)} = \frac{\sum_{t=1}^n \tau_{kt}^{(r)} \delta_{ks}^{2(r)} \sigma_{k,t-s}^2 (y_t - \alpha_k^{(r+1)} - \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)})^2 / (\sigma_{kt}^{2(r)})^2}{\sum_{t=1}^n \tau_{kt}^{(r)} \delta_{ks}^{(r)} \sigma_{k,t-s}^2 / \sigma_{kt}^{2(r)}}, \tag{26}$$

for each  $k, m$ , and  $s$ . Finally, making the substitute (22) into (17), the coordinate-wise updates for the  $\beta$  block

$$\beta_{kj}^{(r+1)} = \frac{pd\beta_{kj}^{(r)} \sum_{t=1}^n \tau_{kt}^{(r)} x_{ij}^2 / \sigma_{kt}^{2(r)} + \sum_{t=1}^n \tau_{kt}^{(r)} x_{ij} (y_t - \alpha_k^{(r)} - \tilde{\mathbf{x}}_t^T \tilde{\boldsymbol{\beta}}_k^{(r)}) / \sigma_{kt}^{2(r)}}{G_1'(\boldsymbol{\pi}^{(r)}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(r)}) / \beta_{kj} + pd \sum_{t=1}^n \tau_{kt}^{(r)} x_{ij}^2 / \sigma_{kt}^{2(r)}}, \quad (27)$$

for each  $k$  and  $j = 1, \dots, pd$ , where  $G_1'(\boldsymbol{\pi}^{(r)}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(r)})$  is the first derivative of (11) with respect to  $\beta$ .

Note that (15)-(17) from Gaussian FM-GARCH Model are concave in the alternative parameterization  $\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\delta}$ , and  $\boldsymbol{\beta}$ , thus (23)-(27) globally maximize (15)-(17) over the parameter space.

### 4.2. Selection of Thresholding Parameters and Components

To implement the methods described in Sections 3 and 4.1, we need to select the size of the tuning parameters  $\lambda_{nk}$  and  $\nu_{nk}$ , the constant  $a_{nk}$  and  $b_{nk}$ , for  $k = 1, \dots, K$ , and components  $K$ . The current theory provides some guidance on the order of  $\boldsymbol{\lambda}$  in [3] and [8] by using generalized cross validation (GCV) and Bayesian Information Criterion (BIC), to ensure the sparsity property. Following the example of [8], we develop a suitable BIC criterion for the FM-AR and FM-GARCH models. Let  $\boldsymbol{\Psi} = (\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{a}, \boldsymbol{b}, K)$ , and they are chosen one at a time by minimizing

$$BIC_{\Psi} = -2\mathcal{L}_n(\boldsymbol{\theta}) + (\tilde{p} + \tilde{q} - 1)\log n, \quad (28)$$

where  $\tilde{p}$  is the dimensionality of  $\boldsymbol{\beta}$  (i.e. the total number of non-zero regression coefficients in these model), and  $\tilde{q}$  equal to  $3K$  (FM-AR models) or  $5K$  (for FM-GARCH models).

The Block-wise MM algorithm is iterated until some convergence criterion is met. In this article, we choose to use the absolute convergence criterion, where  $TOL > 0$  is a small tolerance constant from [8]. Based on the discussion above, we summarise our algorithm in 1.

## 5. Simulated Data Analysis

In this section, we evaluate the performance of the proposed method and algorithm via simulations. We consider the Gaussian FM-AR models and Gaussian FM-GARCH models. Following [2] and [8], we used the correctly estimated zero coefficients (S1), correctly estimated non-zero coefficients (S2) and the mean estimate over all falsely identified non-zero predictors ( $M_{NZ}$ ). The selection of thresholding parameters and components are solving by using Simulated Annealing (SA) algorithm. All simulations were evaluated with varying values of dimension  $p$  with 100 repetitions done for each.

### 5.1. Simulated Data Analysis of Gaussian FM-AR

The first simulations are based on the Gaussian FM-AR (2) model. Assuming that  $K$  is known, the model for the simulation was a  $K = 2$  and  $d = 2$  model of

$$\pi N(\alpha_1 + \mathbf{x}_t^T \boldsymbol{\beta}_{11} + \mathbf{x}_{t-1}^T \boldsymbol{\beta}_{12}, \sigma_1^2) + (1 - \pi) N(\alpha_2 + \mathbf{x}_t^T \boldsymbol{\beta}_{21} + \mathbf{x}_{t-1}^T \boldsymbol{\beta}_{22}, \sigma_2^2), \quad (29)$$

---

**Algorithm 1**

---

**Require:**

Initialize  $\boldsymbol{\lambda}^{(0)}, \mathbf{v}^{(0)}, \mathbf{a}^{(0)}, \mathbf{b}^{(0)}, K, \text{TOL}$ ;  
 Initialize  $\boldsymbol{\pi}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\gamma}_0^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\beta}^{(0)}$ ;

**Ensure:**

Estimated value of the optimal solution  $\boldsymbol{\pi}^{(*)}, \boldsymbol{\alpha}^{(*)}, \boldsymbol{\gamma}_0^{(*)}, \boldsymbol{\gamma}^{(*)}, \boldsymbol{\delta}^{(*)}, \boldsymbol{\beta}^{(*)}$  with appropriate  $\boldsymbol{\lambda}^{(*)}, \mathbf{v}^{(*)}, \mathbf{a}^{(*)}, \mathbf{b}^{(*)}$ , and given  $K$ ;

```

1: while  $BIC_{\Psi}^{(t+1)} < BIC_{\Psi}^{(t)}$  do
2:   Update of  $\boldsymbol{\lambda}^{(t)}, \mathbf{v}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}$ .
3:   while  $|\mathcal{F}_n^{(r+1)}(\boldsymbol{\theta}) - \mathcal{F}_n^{(r)}(\boldsymbol{\theta})| < \text{TOL}$  do
4:     Using Brent's algorithm in [8] with reverse communication to find unique root of (19).
5:     Update  $\boldsymbol{\pi}^{(r+1)}$  by (18).
6:     Update  $\boldsymbol{\alpha}^{(r+1)}$  and  $\boldsymbol{\gamma}_0^{(r+1)}$  by solving (15).
7:     Update  $\boldsymbol{\gamma}^{(r+1)}$  and  $\boldsymbol{\delta}^{(r+1)}$  by solving (16).
8:     Update  $\boldsymbol{\beta}^{(r+1)}$  by solving (17).
9:     Update  $\mathcal{F}_n^{(r+1)}(\boldsymbol{\theta})$  by (9).
10:   end while
11:   Update  $BIC_{\Psi}^{(t+1)}$  by (28).
12: end while
    
```

---

where  $n = 300$ ,  $p = 10, 20, 100$ ,  $\pi = 0.3$ ,  $\alpha_1 = 1$ ,  $\alpha_2 = 5$ ,  $\sigma_1 = 1$ , and  $\sigma_2 = 1$ . Columns of  $\mathbf{x}$  are drawn from a multivariate normal, with mean 0, variance 1, and two correlation structures:  $\rho_{ij} = \text{cor}(x_i, x_j) = 0.5^{|i-j|}$ . The regression coefficients are

$$\boldsymbol{\beta}_{11} = (1, 0, 0, 3, 0, \dots, 0)^T, \boldsymbol{\beta}_{12} = (-3, 0, -1, 0, 2, \dots, 0)^T;$$

$$\boldsymbol{\beta}_{21} = (-1, 2, 0, 0, 3, \dots, 0)^T, \boldsymbol{\beta}_{22} = (0, 0, 3, 0, -2, \dots, 0)^T.$$

**Table 1** reports the results. We can see that when the dimension  $p = 100$ , the S2 in com1 of  $X_{t-1}$  from MIXSCAD-ML<sub>2</sub> is 100, however, the S2 in com1 of  $X_{t-1}$  from MIXLASSO-ML<sub>2</sub> (S2 = 70.7) and MIXMCP-ML<sub>2</sub> (S2 = 51.3) model are small, which indicates that MIXSCAD-ML<sub>2</sub> ensures that non-zero coefficients can be correctly identified and some non-zero coefficients in the MIXLASSO-ML<sub>2</sub> and MIXMCP-ML<sub>2</sub> model are not estimated. The mean estimate over all falsely identified non-zero predictors ( $M_{NZ}$ ) of  $\boldsymbol{\beta}$  from MIXSCAD-ML<sub>2</sub> are between 0.001 and 0.01.

### 5.2. Simulated Data Analysis of Gaussian FM-GARCH

The second simulations are based on the Gaussian FM-GARCH(2,1,1) model. Also assuming that  $K$  is known, the model for the simulation was a  $K = 2$ ,  $d = 2$ ,  $M = 1$  and  $S = 1$  model of

$$\pi N(\alpha_1 + \mathbf{x}_t^T \boldsymbol{\beta}_{11} + \mathbf{x}_{t-1}^T \boldsymbol{\beta}_{12}, \sigma_{1,t}^2) + (1 - \pi) N(\alpha_2 + \mathbf{x}_t^T \boldsymbol{\beta}_{21} + \mathbf{x}_{t-1}^T \boldsymbol{\beta}_{22}, \sigma_{2,t}^2), \quad (30)$$

$$\sigma_{kt}^2 = \gamma_{0k} + \gamma_{k1} \epsilon_{k,t-1}^2 + \delta_{k1} \sigma_{k,t-1}^2, \quad (31)$$

for  $k = 1, 2$ , where  $n = 300$ ,  $p = 10, 20$ ,  $\pi = 0.3$ ,  $\alpha_1 = 2$  and  $\alpha_2 = 5$ ,  $\gamma_{01} = 1$  and  $\gamma_{02} = 1$ ,  $\gamma_{11} = 0.5$  and  $\gamma_{21} = 0.2$ ,  $\delta_{11} = 0.4$  and  $\delta_{21} = 0.6$ .  $\epsilon_{kt} = \sigma_{kt} e_{kt}$ ,  $e_{kt}$  is an independent and identically distributed series with mean zero and variance unity. Columns of  $\mathbf{x}$  are drawn from a multivariate normal, with mean 0, variance 1, and two correlation structures:

$\rho_{ij} = \text{cor}(x_i, x_j) = 0.5^{|i-j|}$ . The regression coefficients are



**Table 1.** Summary of MIXLASSO-ML<sub>2</sub>, MIXMCP-ML<sub>2</sub> and MIXSCAD-ML<sub>2</sub>-penalized FM-AR (2) model with BIC method form the simulated scenario. Average correctly estimated zero coefficients (specificity; S<sub>1</sub>), average correctly estimated non-zero coefficients (sensitivity; S<sub>2</sub>), and the mean  $\beta$  estimate over all incorrectly estimated non-zero coefficients ( $M_{NZ}$ ) are also reported.

Method	$K * d * p$	Com	$X_t$			$X_{t-1}$		
			S <sub>1</sub> (%)	S <sub>2</sub> (%)	$M_{NZ}$	S <sub>1</sub> (%)	S <sub>2</sub> (%)	$M_{NZ}$
MIXSCAD-ML <sub>2</sub>	2*2*10	com1	86.0	99.5	0.097	90.0	99.7	-0.012
			91.2	99.5	0.067	91.6	99.7	-0.003
			81.7	100.0	0.016	82.6	100.0	0.009
	2*2*100	com2	94.3	99.3	0.020	95.5	100.0	-0.093
			94.2	99.3	0.013	96.1	100.0	-0.018
			90.7	100.0	-0.015	90.5	100.0	0.008
MIXMCP-ML <sub>2</sub>	2*2*10	com1	80.1	100.0	0.040	87.6	100.0	0.005
			91.9	100.0	0.100	92.8	100.0	0.027
			98.1	81.0	0.304	98.1	51.3	0.205
	2*2*100	com2	93.0	100.0	0.041	96.5	100.0	-0.015
			96.8	100.0	0.055	98.4	100.0	0.084
			97.4	100.0	0.076	97.2	100.0	0.037
MIXLASSO-ML <sub>2</sub>	2*2*10	com1	76.1	100.0	0.089	76.3	99.7	-0.019
			81.6	100.0	0.066	81.4	100.0	-0.011
			80.5	76.0	0.053	81.1	70.7	0.041
	2*2*100	com2	85.1	100.0	0.015	88.3	100.0	-0.001
			91.2	87.3	0.001	90.8	100.0	-0.015
			79.1	99.3	0.048	87.1	100.0	-0.039

$$\beta_{11} = (1.5, 0, 2.5, 0, 0, \dots, 0)^T, \beta_{12} = (-3.5, 0, -1, 0, 2, \dots, 0)^T;$$

$$\beta_{21} = (-1, 2, 0, 0, 3, \dots, 0)^T, \beta_{22} = (0, 0, 3, 0, -2, \dots, 0)^T.$$

From **Table 2**, we can see that in all simulations, the value of S1 in com1 and com2 of  $X_t$  and  $X_{t-1}$  from MIXSCAD-ML<sub>2</sub> are the biggest, which indicates that MIXSCAD-ML<sub>2</sub> perform better than MIXLASSO-ML<sub>2</sub> and MIXMCP-ML<sub>2</sub> in correctly estimated zero coefficients. The mean estimate over all falsely identified non-zero predictors ( $M_{NZ}$ ) of  $\beta$  from MIXSCAD-ML<sub>2</sub> is smaller than which from MIXLASSO-ML<sub>2</sub> and MIXMCP-ML<sub>2</sub>.

### 6. Real Data Analysis

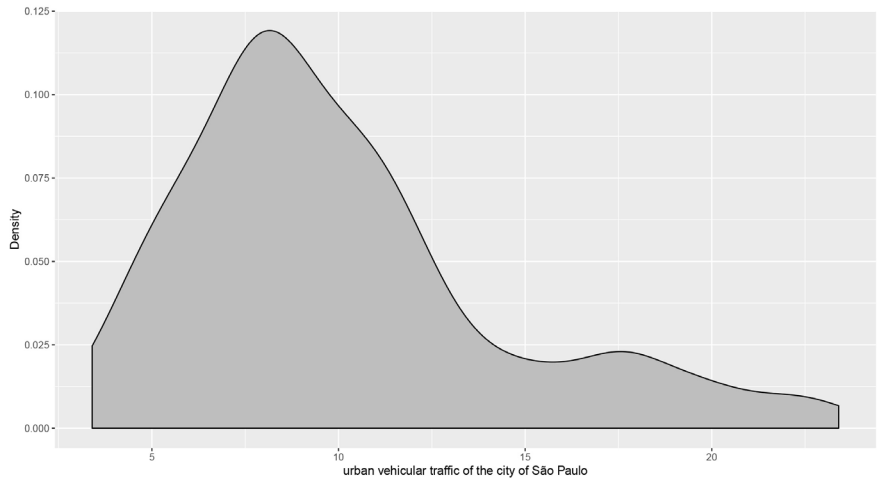
In this section, we evaluate the performance of the proposed method and algorithm via the analysis of the behavior of urban vehicular traffic of the city of São Paulo. This data set were collected notable occurrences of traffic in the metropolitan region of São Paulo in the period from 14 to 18 December 2009. This was acquired from the website <http://archive.ics.uci.edu/ml/datasets.php>. Registered from 7:00 to 20:00 every 30 minutes. It contains 135 observations and 18

**Table 2.** Summary of MIXLASSO-ML<sub>2</sub>, MIXMCP-ML<sub>2</sub> and MIXSCAD-ML<sub>2</sub>-penalized FM-GARCH(1, 1) model with BIC method from the simulated scenario. Average correctly estimated zero coefficients (specificity;  $S_1$ ), average correctly estimated non-zero coefficients (sensitivity;  $S_2$ ), and the mean  $\beta$  estimate over all incorrectly estimated non-zero coefficients ( $M_{NZ}$ ) are also reported.

Method	$K * d * p$	Com	$X_t$			$X_{t-1}$		
			$S_1$ (%)	$S_2$ (%)	$M_{NZ}$	$S_1$ (%)	$S_2$ (%)	$M_{NZ}$
MIXSCAD-ML <sub>2</sub>	2*2*10	com1	88.8	89.5	0.408	92.4	84.0	-0.048
			89.9	84.5	0.432	91.5	79.0	0.168
	com2	94.9	96.3	0.051	97.0	98.0	-0.139	
		96.3	92.0	0.076	95.7	95.0	0.008	
MIXMCP-ML <sub>2</sub>	2*2*10	com1	80.8	94.0	0.417	87.4	81.3	0.115
			85.8	78.5	0.540	87.3	68.0	0.031
	com2	89.4	95.7	0.158	94.0	99.0	0.138	
		93.4	91.0	0.269	95.6	95.5	0.118	
MIXLASSO-ML <sub>2</sub>	2*2*10	com1	73.9	84.5	0.426	79.9	76.0	-0.015
			81.3	66.5	0.579	83.5	56.7	-0.117
	com2	76.7	96.0	0.080	83.6	99.5	0.018	
		88.2	75.0	0.111	93.4	90.5	-0.126	

variables as well as one response variable. Covariate acronyms are hour (HO), immobilized bus (IB), broken truck (BT), vehicle excess (VE), accident victim (AV), running over (RO), fire vehicles (FV), occurrence involving freight (OIF), incident involving dangerous freight (IIDF), lack of electricity (LOE), fire (FI), point of flooding (POF), manifestations (MA), defect in the network of trolley-buses (DNT), tree on the road (TRR), semaphore off (SO), intermittent Semaphore (IS) and the response is slowness in traffic percent. Consider the effect of date on the behavior of traffic, we add a new variable that is day (DA). **Figure 1** shows the heterogeneity of the data set, and the FM-AR or FM-GARCH model is applicable.

The levels of the covariates attributes from FMR, FM-AR (2) and FM-GARCH (2,1,1) with  $K = 2$  models are given in **Table 3**. From **Table 4**, we can see that the MIXSCAD-ML<sub>2</sub> penalized FM-GARCH (2,1,1) with  $K = 2$  model had the lowest BIC (622.9) across all analyses, the FM-AR (2) with  $K = 2$  model being ranked second (BIC = 677.3), which is lower than the BIC (682.3) of FMR model. The predicted slowness in traffic percent from the FM-GARCH  $K = 2$  model had a MSE of 1.93 and a regression  $\tilde{R}^2$  of 0.90. The predicted slowness in traffic percent from the FM-AR (2)  $K = 2$  model had a MSE of 2.09 and a regression  $\tilde{R}^2$  of 0.89. The predicted slowness in traffic percent from the FMR  $K = 2$  model had a MSE of 2.41 and a regression  $\tilde{R}^2$  of 0.87. These results suggest that the FM-GARCH (2,1,1) model had the smallest MSE and explained the largest proportion of variance for the slowness in traffic percent data. The results of the predicted response from these models are presented in **Figure 2**.



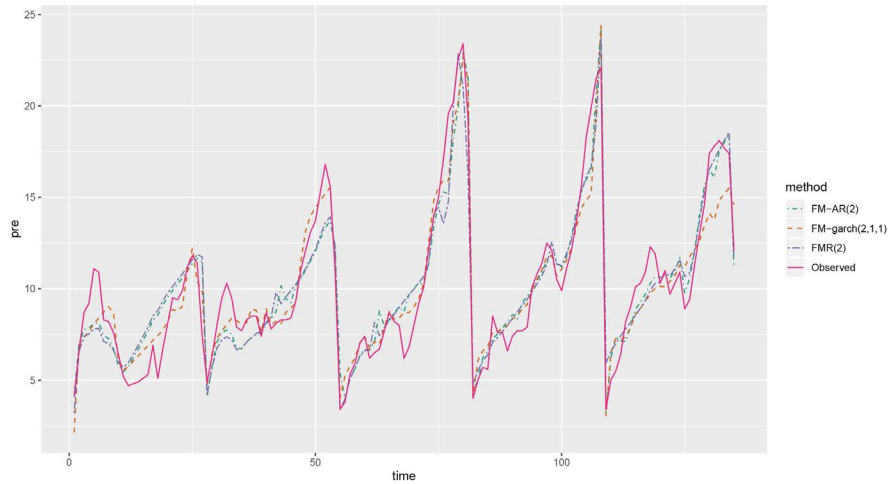
**Figure 1.** Density of slowness in traffic percent in the metropolitan region of São Paulo in the period from 14 to 18 December 2009.

**Table 3.** Summary of FMR, FM-AR and FM-GARCH model with BIC method and MIXLASSO-ML<sub>2</sub> penalty.

Covariates	FMR		FM-AR				FM-GARCH			
	<i>com1</i>	<i>com2</i>	<i>com1</i>		<i>com2</i>		<i>com1</i>		<i>com2</i>	
			$x_t$	$x_{t-1}$	$x_t$	$x_{t-1}$	$x_t$	$x_{t-1}$	$x_t$	$x_{t-1}$
Intercept	7.32	-2.31	7.56	-	-1.89	-	1.39	-	6.24	-
$\pi$	0.37	0.63	0.34	-	0.66	-	0.47	-	0.53	-
DA	-	1.47	-	-	-	1.54	-	-	0.99	-
HO	0.13	0.52	0.11	0.36	-	0.13	0.29	0.39	-	-0.03
IB	-	-	-	-	-	-	-	-	-	-
BT	-	-	-	-	-	-	-	-	-	-
VE	-	-	-	-	-	-	-	-	-	-
AV	-	-	-	-	-	-	-	-	-	-
RO	-	-	-	-	-	-	-	-	-	-
FV	-	-	-	-	-	-	-	-	-	-
OIF	-	-	-	-	-	-	-	-	-	-
IIDF	-	-	-	-	-	-	-	-	-	-
LOE	-	1.75	-	-	-	1.88	-	-	-	1.80
FI	-	-	-	-	-	-	-	-	-	-
POF	-	0.61	-	1.25	-	-	-	1.41	-	-
MA	-	-	-	-	-	-	-	-	-	-
DNT	-	-	-	-0.91	-	-	-	-0.71	-	-
TRR	-	-	-	-	-	-	-	-	-	-
SO	-	-	-	-	-	-	-	-	-	-
IS	-	-	-	-	-	-	-	-	-	-

**Table 4.** Summary of the values of BIC, MSE, and adjusted regression (predicted response on observed response)  $\tilde{R}^2$  from FMR, FM-AR (2) and FM-GARCH (2,1,1) models.

model	$K$	BIC	MSE	$\tilde{R}^2$
FM-GARCH (2,1,1)	2	622.90	1.93	0.90
FM-AR (2)	2	677.32	2.09	0.8
FMR	2	682.36	2.41	0.87



**Figure 2.** Summary of predicted and observed slowness in traffic percent in the metropolitan region of São Paulo in the period from 14 to 18 December 2009.

### 7. Discussion

In this article, we discussed that the modeling of response variable which is time series and with a finite mixture distribution depends on covariates, and the variable selection problem of them. We propose the FM-AR models and FM-GARCH models for modeling data that arise from a heterogeneous population which is time series, and propose a new regularization method (MIXLASSO- $ML_2$ , MIXMCP- $ML_2$ , MIXSCAD- $ML_2$ ) for the variable selection in these model, which composed of the mixture of the  $l_1$  penalty and  $l_2$  penalty proportional to mixing proportions. In addition, we estimate the maximum log quasi-likelihood estimate for the new penalized FM-AR and FM-GARCH model, and derive a general expression for the block-wise minimized maximization (MM) algorithm with better features. The simulation results of Gaussian FM-AR and Gaussian FM-GARCH models and an actual data set illustrate the capability of the methodology and algorithm, and MIXSCAD- $ML_2$  is always superior to other penalty methods.

### Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

- [1] McLachlan, G.J. and Peel, D. (2000) Finite Mixture Models. Wiley, New York. <https://doi.org/10.1002/0471721182>
- [2] Abbas, K. and Shili, L. (2013) Regularization in Finite Mixture of Regression Models with Diverging Number of Parameters. *Biometrics*, **69**, 201-235. <https://doi.org/10.1111/biom.12020>
- [3] Abbas, K. and Chen, J.H. (2007) Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association*, **102**, 1025-1038. <https://doi.org/10.1198/016214507000000590>
- [4] Zhang, Z., Li, W.K. and Yuen, K.C. (2005) On a Mixture GARCH Time-Series Model. *Journal of Time Series Analysis*, **27**, 577-597. <https://doi.org/10.1111/j.1467-9892.2006.00467.x>
- [5] Haas, M., Krause, J., Paoletta, M.S. and Steude, S.C. (2013) Time-Varying Mixture GARCH Models and Asymmetric Volatility. *North American Journal of Economics and Finance*, **26**, 602-623. <https://doi.org/10.1016/j.najef.2013.02.024>
- [6] Abbas, K. (2011) An Overview of the New Feature Selection Methods in Finite Mixture of Regression Models. *Journal of the Iranian Statistical Society*, **10**, 201-235.
- [7] Fan, J.Q. and Li, R.Z. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. <https://doi.org/10.1198/016214501753382273>
- [8] Lloyd-Jones, L.R., Nguyen, H.D. and McLachlan, G.J. (2018) A Globally Convergent Algorithm for Lasso-Penalized Mixture of Linear Regression Models. *Computational Statistics and Data Analysis*, **119**, 19-38. <https://doi.org/10.1016/j.csda.2017.09.003>
- [9] Hwang, S.Y., *et al.* (2014) Quasilikelihood and Quasi-Maximum Likelihood for GARCH-Type Processes: Estimating Function Approach. *Journal of the Korean Statistical Society*, **1**, 5. <https://doi.org/10.1016/j.jkss.2014.01.005>
- [10] Razaviyayn, M., Hong, M. and Luo, Z.-Q. (2013) A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization. *SIAM Journal on Optimization*, **23**, 1126-1153. <https://doi.org/10.1137/120891009>
- [11] Lange, K. (2013) Optimization. Springer, New York. <https://doi.org/10.1007/978-1-4614-5838-8>