International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

Estimation of Autoregressive Models with Multiple Structural Breaks

Ding Peng[1], Ying Han[2], Yan Zhang[3]
School of Mathematical Sciences, Guizhou Normal University, Guiyang550001, China
Email address: 212067042@qq.com

***Abstract*—*To better explain non-stationary time series data and predict changes in the data using models, this paper employs a piecewise autoregressive model. By utilizing LASSO regression, the problem of detecting change points is transformed into a variable selection problem. The number of change points and their positions are estimated, resulting in an initial grouping of the data. Subsequently, autoregressive models are used to interpret each group of data. The Innovation of This Paper Lies in the Following Aspects: Considering that the innovations of the model may follow a normal distribution or not, we employ a mixture of normal distributions to fit the unknown distribution. By utilizing the Dirichlet process to identify the unknown distribution of the innovations for each group of the model, we obtain the variances of the mixture of normal distributions and the weights of each normal distribution. Through numerical simulations, we compare our model with the LASSO regression method and the MLE method. Our model performs better.*

***Keywords*—** *LASSO Regression; Dirichlet Process Mixture Model; Autoregressive; MCMC algorithm.*

I. INTRODUCTION

The autoregressive (AR) model uses past time series values as the independent variables of the model. Since its introduction by Yule [1] in 1927, it has been widely applied in fields such as statistics, econometrics, and information science. Walker [2] further developed the AR model by proposing the p-order autoregressive model while studying atmospheric pressure at the Darwin port in India. However, as scientists have delved deeper into research, linear time series models have been found to have certain limitations, especially when fitting stock data. The closing prices of stocks can sometimes experience sudden increases or decreases. Page [3] was the first to formally introduce the change-point problem in 1954. He primarily focused on whether the distribution parameters undergo a single change, i.e., whether there is a single change point in the time series data, and proposed the cumulative sum detection method. Quandt [4] used the likelihood ratio test (LRT) to construct a test statistic for studying a simple linear regression model with one change point. For traditional methods and review literature on change-point research, please refer to [5][6][7][8]. Barry [9] developed a Bayesian approach to address change-point detection problems, utilizing a product partition model to demonstrate the effectiveness of a new product partition model. This model exhibits posterior clustering of blocks and a new independent block posterior distribution for parameters. Bai [10] proposed a likelihood ratio type test for multiple structural changes in regression models, allowing for the use of lagged dependent variables and trend regression variables, and derived the limiting distribution of the test, from which asymptotic critical values can be obtained. Zou et al. [11] proposed a nonparametric maximum likelihood method for detecting multiple change points in regression models without requiring prior knowledge of the number of change points. They used the Bayesian information criterion (BIC) to determine the number of change points and leveraged the intrinsic sequential structure of the likelihood function, employing a dynamic programming algorithm to estimate the locations of the change points.

After Tibshirani [14] introduced the LASSO method by adding a penalty term based on the L_1 norm, many scholars have applied the LASSO method to change-point detection. The idea is to transform the change-point detection problem into a variable selection problem. After analyzing the LASSO algorithm, Zou [15] derived the necessary conditions for consistent variable selection in the LASSO algorithm and suggested introducing adaptive weights, proposing the Adaptive LASSO method. This method adaptively adds weights to the LASSO penalty term in a data-driven manner. Ciuperca [16] studied the change-point problem in multivariate linear regression models using two types of Lasso methods and found that the adaptive Lasso method performs better than the least squares method in detecting change points. Ciuperca [17] also applied the adaptive Lasso method to detect change points in quantile regression models and demonstrated that this method outperforms other variable selection methods. Zhang [18] proposed a change-point linear regression method based on sparse group Lasso, which segments the available data into different regions by estimating the number and locations of change-points, and further generates sparse and interpretable models for each region. Qian and Su [19] proposed two shrinkage procedures for determining the number of structural changes in linear panel data models using the adaptive group fused Lasso. Li Aomei [20] combined the Groupwise Majorization Descent (GMD) algorithm with the adaptive group Lasso method to study the change-point problem in multivariate linear regression models, accurately estimating the number, location, and model parameters of change points, with good estimation precision. Yang Zhaoxin [21] proposed constructing a quantile Lasso statistic to estimate the change-point locations in linear regression models and obtained the convergence rate of the estimates.

After identifying the change points, this paper applies the Dirichlet mixture model to the prior of the parameters. The Dirichlet process, first proposed by Ferguson [22], does not require a pre-specified number of clusters. For each data point,

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

an auxiliary variable is assigned, and the probability of the data point belonging to each existing category as well as the probability of a new category is calculated. Then, the value of the auxiliary variable is sampled according to these probabilities, achieving the purpose of clustering. Due to the unknown parameters playing a decisive role in the number of clusters, Escobar & West [23] calculated that the posterior distribution of the parameters is a mixture of two gamma distributions, under the assumption that the prior of the parameters follows a gamma distribution. After continuous improvement by subsequent researchers, the Dirichlet process has been widely applied in many fields. Liang Hong & Ryan Martin [24] established a flexible nonparametric Bayesian model for modeling insurance losses to predict future claim amounts. Adesina [25] proposed a Dirichlet process mixture prior for generalized linear mixed models (GLMMs) and applied it to fit over-dispersed and equi-dispersed count data. In the field of time series, Dirichlet mixture models have been widely applied, and their analyses have repeatedly demonstrated the feasibility of these models for model building and estimation [26][27][28][29][30].

The subsequent content of this paper is as follows. Section 2 will briefly introduce the methods used in this paper. Section 3 will identify change points and incorporate the Dirichlet mixture model into the autoregressive model. Section 4 is about parameter estimation. Section 5 will compare the estimation methods of the Dirichlet process mixture model with the separate LASSO method and maximum likelihood estimation method through a segmented autoregressive model, demonstrating the superiority of our approach. Section 6 presents a specific case study.

II. Method Introduction

2.1: Introduction to LASSO

Consider the typical linear regression model, where our data consist of an n-dimensional vector Y and an $n \times p$ matrix X. As the number of variables p gradually increases, even to the point where $p \gg n$, the standard linear regression model will fail. To address this issue, Tibshirani [5] proposed the LASSO method. For the general autoregressive model:

$$y_t = \sum_{l=1}^{p} y_{t-l}\beta_l + \varepsilon_t \qquad (1)$$

The LASSO method estimates $\hat{\beta}$ as follows:

$$argmin \left\| y - \sum_{l=1}^{p} y_{t-l}\beta_l \right\|_2^2 + \lambda \sum_{l=1}^{p} |\beta_l|_1 \qquad (2)$$

Herein, λ is a regularization penalty parameter, i.e., $\lambda > 0$, and the sparsity of the solution is determined by the magnitude of λ. $\|\cdot\|_2$ denotes the l_2 norm, while $|\cdot|_1$ represents the l_1 norm. The Lasso method involves adding a constraint to the sum of the absolute values of the coefficients after the loss function reaches its minimum value, thereby bounding it. By introducing this constraint, some of the estimated coefficient values $\hat{\beta}$ become sparse, with most of their values being zero, thus achieving an optimization goal. Equation (2) can also be transformed as follows:

$$argmin \left\| y - \sum_{l=1}^{p} y_{t-l}\beta_l \right\|_2^2 \quad s.t. \sum_{l=1}^{p} |\beta_l|_1 \le q \qquad (3)$$

The LASSO method possesses favorable characteristics in optimization, with the objective function to be minimized being a convex function. Therefore, it does not suffer from the issue of multiple local minima, and the global minimum can be efficiently solved using various algorithms, such as the coordinate descent algorithm [31], Least Angle Regression and Shrinkage (LARS) [32], and so on.

2.2: Dirichlet Process Mixture Model

We first introduce the basic concept of the Dirichlet process: for any finite partition $A_1, A_2, \cdots, A_n$ in the measurable space Θ, if the following equation holds, then the distribution G is said to satisfy the Dirichlet process.

$$(G(A_1), G(A_2), \cdots, G(A_n) = Dir(\alpha H(A_1), \alpha H(A_2), \cdots, \alpha H(A_n)) \qquad (4)$$

Where α is the concentration parameter, which controls the discreteness of the distribution *G*. Specifically, the larger the value of α, the more categories there are; the smaller the value of α, the fewer categories there are. The Dirichlet process has been widely applied in nonparametric clustering.

It is difficult for readers to understand how the Dirichlet process is computed merely from its definition. Fortunately, Blackwell, D [33] derived an important formula, which can be interpreted as follows: assuming we have an infinite number of data points, and the first n data points are divided into K^* categories, the probability that the *n*+1-th data point belongs to a certain category is given by

$$p(z_{y_{n+1}} = j \mid z_{y_{1:n}}) = \begin{cases} \frac{1}{n+\alpha} \sum_{i=1}^{n} \delta(z_{y_i} = j), & j = 1, 2, \cdots, K^* \\ \frac{\alpha}{n+\alpha} H(j), & j = K^* + 1 \end{cases} \qquad (5)$$

where $\delta(\cdot)$ is the indicator function, which takes the value 1 when the equality holds and 0 otherwise. $z_{y_{n+1}}$ denotes the category to which the n+1-th data point belongs.

We also need to understand one thing: although we know that the Dirichlet process can be computed, how can we link this stochastic process with the data? This is achieved through the Dirichlet process mixture model, which is represented as follows:

$$\begin{aligned} y_i|\theta_i &\sim F(y_i|\theta_i) \\ \theta_i|G &\sim G \\ G|\alpha, H &\sim DP(\alpha, H) \end{aligned} \qquad (6)$$

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

θ_i represents the parameters of the distribution that the data y_i follows. In order to perform clustering, the distribution G is generally a discrete distribution, which is a distribution we construct. The most widespread application of the Dirichlet process is as the prior for the Dirichlet process mixture model.

III. Model Establishment and Changepoint Detection

The objective of this paper is to estimate the locations of changes $\{t_k\}$, the number of changepoints K, the coefficients $\{b_k\}$ of the autoregressive models for each segment, and the mixture normal distribution of the innovations, based on n pairs of observed data (y_t, y_t^*).

We consider the following autoregressive model:

$$y_t = \beta_t' y_t^* + \varepsilon_t \quad (7)$$

Where $y_t^* = \{u_k, y_{t-1}, y_{t-2}, \cdots y_{t-p+1}\}'$ is a $p \times 1$ dimensional column vector, u_k is the mean of each segment of the autoregressive model, β_t is a $p \times 1$ dimensional sparse coefficient vector, and ε_t is the innovation. The coefficient vector β_t in the above equation is not a fixed vector; instead, it varies with time t. Assuming that among the n data points, the value of β_t changes K times, we denote the set of all changepoints as $T = \{t_k, k = 1, 2, \cdots, K\}$, and define it as follows:

$$\beta_t = b_k, t_{k-1} \le t \le t_k - 1, k = 1, 2, \cdots, K \quad (8)$$

Where $t_0 = 1, t_{k+1} = n + 1$, and $\{b_k, k = 1, 2, \cdots, K\}$ are the values of the coefficients for each segment in equation (7). Typically, the innovation ε_t is assumed to follow the same normal distribution. However, in this paper, we assume that the innovation ε_t may follow other distributions or a mixture of several distributions, and the distribution that ε_t follows may differ for each segment. As long as the data size is sufficiently large, we can use a mixture of normal distributions to approximate the unknown distribution.

Suppose there are *n* data points, then equation (7) can be transformed into the following form:

$$Y = Y^* B' + E \quad (9)$$

Where $Y = \{y_1, y_2, \cdots, y_n\}'$ is an $n \times 1$ dimensional column vector, $B = \{\beta_1', \cdots, \beta_n'\}'$ is an $np \times 1$ dimensional vector, where each β_t can be determined from equation (8), and it is piecewise. $Y^* = diag\{(y_1^*)', (y_2^*)', \cdots, (y_n^*)'\}$ is an $n \times np$ matrix, and $E = \{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}'$ is an $n \times 1$ dimensional column vector.

We define the following matrices:

$$A = \begin{pmatrix} I_p & 0 & \cdots & 0 \\ I_p & I_p & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ I_p & I_p & \cdots & I_p \end{pmatrix}, Y^* = \begin{pmatrix} (y_1^*)' & 0 & 0 & 0 \\ 0 & (y_2^*)' & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & (y_n^*)' \end{pmatrix} \quad (10)$$

Where I_p is the $p \times p$ identity matrix, $\mathbf{0}$ is the $p \times p$ zero matrix, and A is an $np \times np$ matrix. Therefore, the following matrix can be obtained:

$$\widetilde{Y^*} = Y^* A = \begin{pmatrix} (y_1^*)' & 0 & \cdots & 0 \\ (y_2^*)' & (y_2^*)' & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ (y_n^*)' & (y_n^*)' & \cdots & (y_n^*)' \end{pmatrix} \quad (11)$$

Where $\widetilde{Y^*}$ is an $n \times np$ matrix. If we let $\theta_1 = \beta_1 - \beta_0, \theta_n = \beta_n - \beta_{n-1}$, where $t = 1, 2, \cdots, n$ and β_0 is a $p \times 1$ zero vector, then $\Theta = \{\theta_1', \theta_2', \cdots, \theta_n'\}'$. Since β_t is a sparse vector, most of the θ_t' in Θ are zero vectors, and the majority of the elements in the non-zero θ_t' are also 0. Therefore, equation (9) can be rewritten as:

$$Y = \widetilde{Y^*}\Theta + E \quad (12)$$

To obtain the number of changepoints and the location of each changepoint, in conjunction with equation (3), the SGL estimation is transformed into solving the following optimization problem:

$$min_\theta \frac{1}{n} \left\| Y - \widetilde{Y^*}\Theta \right\|_2^2 + \gamma\lambda \sum_{t=1}^{n} \|\theta_t\|_2 + (1-\gamma)\lambda \sum_{t=1}^{n} |\theta_t|_1 \quad (13)$$

Through equation (13), we can determine where the autoregressive model begins to segment and how many segments there are. Next, we will use the Dirichlet process to estimate the coefficients of each segment of the autoregressive model, as well as the mixture normal distribution that the innovations follow.

We use the model in equation (7) for subsequent estimation. Let's make an assumption that our autoregressive model has K segments, each segment contains n_k data points, and the innovations in each segment follow a mixture of c_k normal distributions, with each normal distribution containing $N_{k,j}$ data points, where $k = 1, 2, \cdots, K, j = 1, 2, \cdots, c_k$. We expand the autoregressive model specifically, and equation (9) is rewritten as:

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

$$y_t = \begin{cases} \begin{cases} y_{1,N_{1,1}} = \beta_1' y^*_{1,N_{1,1}} + \varepsilon_{1,1} & \varepsilon_t \sim N(0,\sigma^2_{1,1}) \\ \vdots \\ y_{1,N_{1,C_1}} = \beta_1' y^*_{1,N_{1,C_1}} + \varepsilon_{1,c_1} & \varepsilon_t \sim N(0,\sigma^2_{1,c_1}) \end{cases} \\ \vdots \\ \begin{cases} y_{K,N_{K,1}} = \beta_K' y^*_{K,N_{K,1}} + \varepsilon_{K,1} & \varepsilon_t \sim N(0,\sigma^2_{K,1}) \\ \vdots \\ y_{K,N_{K,C_1}} = \beta_K' y^*_{K,N_{K,C_K}} + \varepsilon_{K,c_K} & \varepsilon_t \sim N(0,\sigma^2_{K,c_K}) \end{cases} \end{cases} \quad (14)$$

It can be observed that our assumed model is not only segmented, but also categorized within each segment according to different distribution parameters. As described in the previous section on the Dirichlet process, the prior of the parameters does not need to be specified:

$$\beta_t \sim N(0, V_t)$$
$$V_t \sim G$$
$$G \sim DP(\alpha, H) \quad (15)$$

where V_t is the true variance associated with each data point, The discrete distribution G is defined as follows by Sethuraman.J [34]:

$$G(dV) = \sum_{i=1}^{\infty} \pi_i \delta_{\sigma^2_{ki}}(dV), k = 1, 2, \cdots, K \quad (16)$$

where σ^2_{ki} are known variances sampled from the base distribution H, $\delta(\cdot)$ is the indicator function, π_i represents the probability that V_t equals a specific known σ^2_{ki}, and $\pi_i \sim GEM(\alpha)$ (GEM stands for Griffiths, Engen, and McCloskey). Specifically:

$$v_i | \alpha \sim Beta(1, \alpha)$$
$$\pi_i = v_i \prod_{ii=1}^{i-1} (1 - v_{ii}) \quad (17)$$

When $i = 1, \pi_1 = v_1$. To ensure conjugacy and facilitate the derivation of the posterior distribution, we make the following setting for the base distribution H:

$$H \equiv Inv - gamma(\psi_1, \omega_1) \quad (18)$$

If the true variance V_t associated with each data point were known, it would be straightforward to classify the data in each segment. However, the value of V_t is unknown. Therefore, for the subsequent classification work, we reintroduce the indicator variable z from equation (5), incorporating the indicator variable $Z = (z_{y_1}, z_{y_2}, \cdots, z_{y_n})$ into the model. Here, z_{y_n} represents the category to which the n-th data point belongs, such that when $V_t = \sigma^2_{kj}, j = 1, 2, \cdots, c_k$, then $z_{y_t} = kj$. The distribution of z_{y_t} is as follows:

$$z_{y_t} = \sum_{i=1}^{\infty} \pi_i \delta_i \quad (19)$$

In conjunction with equations (16) to (19), equation (15) can be rewritten as:

$$y_t = \beta_t' y_t^* + \varepsilon_t$$
$$\beta_t \sim N(0, \sigma^2_{z_{y_t}})$$
$$z_{y_t} = \sum_{i=1}^{\infty} \pi_i \delta_i$$
$$\sigma^2_{z_{y_t}} \sim Inv - gamma(\psi_1, \omega_1) \quad (20)$$

IV. The Posterior Distribution of the Model Parameters

Let $\Sigma^2 = \{\sigma^2_{kz}, k = 1, \cdots, K, z = 1, \cdots, c_k\}'$. Since the posterior distribution of the Dirichlet process mixture model does not have an analytical solution, we partition the unknown model and use Bayes' theorem to obtain the joint posterior distribution of the parameters:

$$p(\Sigma^2, B', \alpha, Z | Y)$$
$$= p(Y | \Sigma^2, B', \alpha, Z)\, p(B' | \Sigma^2, \alpha, Z)\, p(\Sigma^2 | \alpha, Z)\, p(Z | \alpha)\, p(\alpha) \quad (21)$$

Below, we will sequentially estimate the posterior distributions of the various parameters in equation (21).

4.1: Sampling of b_k

Through equation (13), we have already determined the number and locations of the changepoints in the piecewise autoregressive model. Thus, all the data between every two consecutive changepoints constitute a dataset for an autoregressive model.

We denote $\{Y_{k,N_{k,j}}, k = 1, 2, \cdots, K, j = 1, 2, \cdots, c_k\}$ as the set of data for each normal distribution, and let the set $Y = \{Y_{k,N_{k,j}}\}$ represent all the data. Since β_t takes values from $\{b_k, k = 1, 2, \cdots, K\}'$, this paper needs to estimate each b_k rather than B'. Let $p(b_k)$ denote the prior distribution of b_k, and $\frac{\varepsilon_t}{\sqrt{\sigma^2_{z_{y_t}}}} \sim N(0, 1)$, then

$$p(b_k' | \Sigma^2, \alpha, Z) \propto \prod_{t=1}^{n_k} \sqrt{2\pi\sigma_k^2} \exp\left\{ - \frac{(y_t - b_k' y_t^*)^2}{2\sigma_k^2} \right\} p(b_k)$$
$$\propto \exp\left\{ - \frac{(b_k' - u_{b_k'})'(b_k' - u_{b_k'})}{2\sigma^2_{b_k'}} \right\} p(b_k) \quad (22)$$

where $z_{y_t} = k$, and it follows that:

$$\sigma^2_{b_k'} = \left(\sum_{t=1}^{n_k} \frac{y_t^* (y_t^*)'}{\sigma_k^2} \right)^{-1}, \; u_{b_k'} = \sigma^2_{b_k'} \left(\sum_{t=1}^{n_k} \frac{y_t^* y_t}{\sigma_k^2} \right) \quad (23)$$

4.2: Sampling of Z

The sampling probability of z_{y_t} is as follows:

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

$$z_{y_t}|\sigma_{kj}^2,\varepsilon_t,\alpha,j = 1,\cdots,c_k \sim \begin{cases} \dfrac{\alpha g(\varepsilon_t)\delta_{c_k+1}(dz_{y_t})}{\sum_{j=1}^{c_k} f_N(\varepsilon_t|\sigma_{kj}^2)\delta_j(dz_{y_t}) + \alpha g(\varepsilon_t)\delta_{c_k+1}(dz_{y_t})} \\ \dfrac{\sum_{j=1}^{c_k} f_N(\varepsilon_t|\sigma_{kj}^2)\delta_j(dz_{y_t})}{\sum_{j=1}^{c_k} f_N(\varepsilon_t|\sigma_{kj}^2)\delta_j(dz_{y_t}) + \alpha g(\varepsilon_t)\delta_{c_k+1}(dz_{y_t})} \end{cases} \quad (24)$$

The sampled values of z_{y_t} are $\{1,2,\cdots,K,K+1\}$. If the sampled value is within $\{1,2,\cdots,K\}$, the variance of that category will be updated. If the sampled value is $K+1$ then K is incremented by 1, increasing the number of mixture normal distributions, and the variance of the new distribution will be randomly sampled from the inverse gamma distribution.

4.3: Sampling of Σ^2

When the distribution of the innovation term is non-normal, a mixture of normal distributions can be used to approximate the unknown distribution, yielding the following mixture model:

$$\varepsilon_t \sim f(\varepsilon_t|\sigma^2) = \int F_N(\varepsilon_t|0,\sigma^2)dG(\sigma^2) \quad (25)$$

It is worth mentioning that even if ε_t does not follow a distribution with a mean of 0, as long as the time series data in equation (7) is adjusted for the mean, it will suffice. Combining the Chinese restaurant process in equation (5), we obtain:

$$\sigma_t^2|\sigma_{-t}^2,\varepsilon_t,$$

$$\alpha \sim \frac{\alpha}{n+\alpha-1} g(\varepsilon_t) G(\sigma^2|\varepsilon_t) + \frac{1}{n+\alpha-1}\sum_{i,i\neq t}\delta_{\sigma_i^2}(\sigma_k^2) \quad (26)$$

Where σ_{-t}^2 denotes the variances of all data points except for the *t*-th data point, and $g(\varepsilon_t) = \int F_N(\varepsilon_t|0,\sigma^2)dG_0(\sigma^2)$.

By the properties of conditional probability, it follows that:

$$G(\sigma^2|\varepsilon_t) \propto F_N(\varepsilon_t|0,\sigma^2)dG_0(\sigma^2) \quad (27)$$

Applying the prior information from equation (20), the posterior distribution of σ^2 can be derived as the inverse gamma distribution:

$$\begin{aligned} p(\sigma_k^2|\alpha,Z,\varepsilon_t) &\propto \prod_{t,z_{y_t}=j} F_N(\varepsilon_t|0,\sigma^2)G_0(\sigma^2) \\ &\propto \prod_{t,z_{y_t}=j} \sqrt{\sigma_k^2}\exp\left\{-\frac{\varepsilon_t^2}{2\sigma_k^2}\right\}\sigma_k^{-2(\psi_1+1)}\exp\left\{-\frac{\omega_1}{\sigma_k^2}\right\} \\ &\sim invgamma\left(\frac{N_{k,j}+2\psi_1}{2},\frac{N_{k,j}\varepsilon_t^2+2\omega_1}{2}\right) \end{aligned} \quad (28)$$

Where $N_{k,j}$ denotes the number of data points contained in the *j*-th normal distribution of the *k*th segment. If $z_{y_t} = c_k+1$, the number of mixture normal distributions in the *k*-th segment increases, and at this point, $N_{k,j}=1, j=c_k+1$. In the algorithm, a Gibbs sampler can be employed to continuously update the values of σ_k^2.

4.4: Sampling of α

The parameter α essentially controls the similarity between the distribution G generated by the Dirichlet process and the base distribution G_0. When α is larger, G is more similar to G_0, and when α is smaller, G tends to deviate from G_0. Escobar and West (1995) provided the posterior density of the parameter α under a gamma prior. Assuming a continuous prior density $p(\alpha)$ (which may depend on the sample size *n*), this implies the implied prior $p(K|n) = E[p(K|\alpha,n)]$. Using the result from Antonia [35], we have:

$$p(K|\alpha,n) = c_n(K)n!\alpha^K\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}, K=1,2,\cdots,n \quad (29)$$

Where $c_n(K) = p(K|\alpha=1,n)$. If K is known, the data are assigned to K specific groups. When both K and the weights π are known, the data Y are conditionally independent of the parameter α. Moreover, when K is known, the weights π are also conditionally independent of the parameter α. This leads to the definition of the conditional distribution:

$$p(\alpha|\pi,K,Y) = p(\alpha|K) \propto p(\alpha)p(K|\alpha) \quad (30)$$

Given that the likelihood function has been provided in Equation (29), the Gibbs sampling analysis can be extended accordingly. For a given parameter α, one can first sample all other parameters except α. Subsequently, the parameter α can be sampled using Equation (30), without requiring additional information.

Assuming $\alpha \sim G(a,b)$, if $a \to 0$ and $b \to 0$, with $log(a)$ following a uniform distribution, Equation (30) can be expressed as a mixture of two gamma posteriors. The conditional distribution of the mixture parameters, given α, K

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

and n, is a simple beta distribution. When $\alpha > 0$, the gamma function in Equation (29) can be written as:

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\beta(\alpha+1,n)}{\alpha\Gamma(n)} \quad (31)$$

Here, $\beta(\cdot)$ denotes the commonly used beta function. For any $K = 1, 2, \cdots, n$, Equation (30) can be rewritten as:

$$p(\alpha \mid K) \propto p(\alpha)\alpha^{k-1}(\alpha+n)\int_0^1 x^{\alpha}(1-x)^{n-1}dx \quad (32)$$

This implies that $p(\alpha \mid K)$ is the marginal distribution of the joint distribution of α and the continuous variable η, such that:

$$p(\alpha \mid K,\eta) \propto p(\alpha)\alpha^{k-1}(\alpha+n)\eta^{\alpha}(1-\eta)^{n-1} \quad (33)$$

Where $\eta \in (0,1)$. Consequently, we can derive two conditional posterior distributions: $p(\alpha, \mid K,\eta)$ and $p(\eta \mid K,\alpha)$. First, under the prior distribution of $\alpha, \mathrm{G}(a,b)$.

$$p(\alpha, \mid K,\eta) \propto P(\alpha)\alpha^{K-1}(\alpha+n)\eta^{\alpha} \propto \alpha^{K+a-1}e^{-a(b-ln\eta)} + n\alpha^{K+a-2}e^{-\alpha(b-ln\eta)} \quad (34)$$

Thus, the posterior estimate of α is a mixture of two gamma distributions:

$$\alpha|K,\ \eta \sim \pi_{\eta}G(\mathrm{a}+K, b-ln\eta) + (1-\pi_{\eta})G(a+K-1, b-ln\eta) \quad (35)$$

The parameter π_{η} is defined by the following expression:

$$\pi_{\eta} = \frac{c+K-1}{n(d-ln\eta)+c+K-1} \quad (36)$$

These distributions are well-defined for all gamma priors, all η within the unit interval, and all K. The second posterior distribution is given by:

$$P(\eta \mid K,\alpha) \propto \eta^{\alpha}(1-\eta)^{n-1}, 0 < \eta < 1 \quad (37)$$

That is to say, $\eta|k,\alpha \sim Beta(\alpha+1,n)$, and the mean of this Beta distribution is $(\alpha+1)/(\alpha+n+1)$. Therefore, sampling α is carried out in two steps: First, initial values for K and α are provided, and a value for η is sampled from the Beta distribution in Equation (37). Subsequently, based on the same K value and the newly sampled η value, a new value for α is sampled from the two Gamma distributions in Equation (35).

V. Numerical Simulation

In this subsection, we compare the DPMM method used in this paper with the LASSO regression method and the MLE method. We first consider the following dataset:

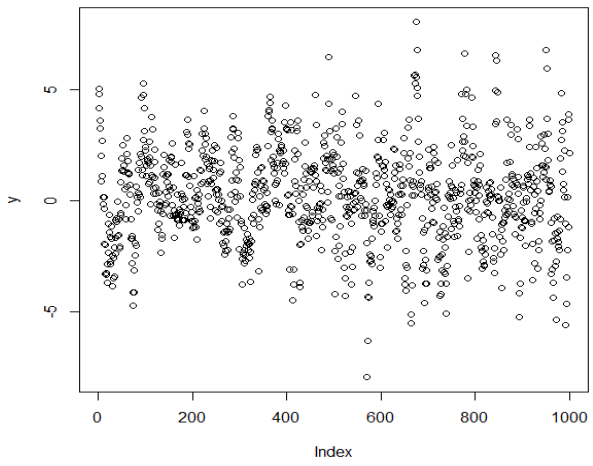

Fig. 5-1. Scatter plot of the time series

It is a two-piece autoregressive model, with the specific parameters as follows:

$$y = \begin{cases} 0.6 \times y_{t-1} + 0.2y_{t-2} + \varepsilon_1, n \le 400 \\ 0.8 \times y_{t-1} - 0.1y_{t-2} + \varepsilon_2, n > 400 \end{cases} \quad (38)$$

Where $\varepsilon_1 \sim t(df=10)$, with the mean defaulting to 0 and the standard deviation defaulting to 1. Additionally, $\varepsilon_2 \sim 0.5N(0,0.5) + 0.5N(0,2)$. The sample size n=1000, and the true location of the change point is at the 401st time point.

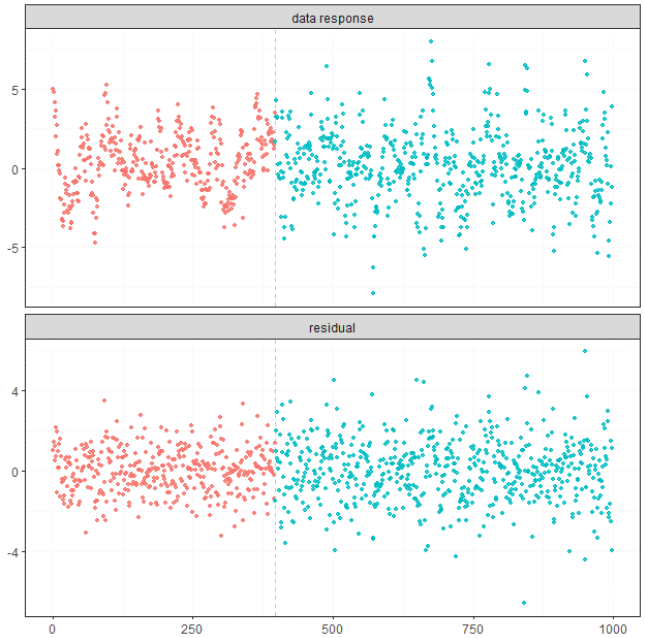

Fig. 5-2. The change point location found by the LASSO method.

Figure 5-2 illustrates the identification of the change point by the LASSO method. It divides the time series model into two segments, with the change point located at the 396th time point. The vertical line in the figure indicates the position of this change point, which is very close to the true change point

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

location. In Figure 5-2, the horizontal axis represents the time series of the data points, while the vertical axis represents the data values.

TABLE 5-1. The estimation of parameters for the autoregressive model

Parameters	b_{11}	b_{12}	b_{21}	b_{22}
True values	0.6	0.2	0.8	-0.1
DPMM	0.5973	0.2378	0.8015	-0.1005
LASSO	0.5950	0.2538	0.8096638	-0.1116126
MLE	0.6044	0.2416	0.8057	-0.1048

Table 5-1 presents the estimates of the coefficients for the two-piece autoregressive model obtained by the three methods. From this table, it can be observed that the DPMM provides the most accurate estimates of the autoregressive coefficients for both segments, especially for the second segment of the autoregressive model. The MLE method performs the second best, while the LASSO regression method yields the weakest results among the three.

The innovation is an important component of the autoregressive model, representing all the new information at a given time point that cannot be explained by past sequence values. It follows an unknown distribution, which may be non-normal or even a mixture distribution. We apply the Dirichlet process to the prior of the innovation, using a mixture of normal distributions to fit the unknown distribution, resulting in Table 5-2:

TABLE 5-2. The estimation of parameters for the distribution of innovation

Parameters	t-distribution	Mixture of Normal Distributions	
True values	t(df=10)	N(0,2)	N(0,0.5)
DPMM	N(0,1.0677) +N(0,2.2338)	N(0,2.1702)	N(0,0.5789)
LASSO	N(0,1.4248)	N(0,1.6093)	N(0,1.6093)
MLE	N(0, 1.1018)	N(0,1.5287)	N(0,1.5287)

In Table 5-2, it is necessary to clarify that for the first segment of the autoregressive model, the innovation follows a t-distribution with 10 degrees of freedom. The DPMM method approximates this t-distribution using a mixture of two normal distributions, with specific standard deviations and weights: 0.6302392N(0,1.0677) + 0.3697608N(0,2.2338).

For the second segment of the autoregressive model, the innovation follows a mixture of two normal distributions with equal weights of 0.5 each. The DPMM method identifies the number of components in the mixture normal distribution, with specific standard deviations and weights: 0.4972397N(0, 2.1702) + 0.5027603N(0,0.5789).

It is important to note that the Dirichlet process often identifies a number of clusters that exceeds the true number of clusters. For example, in the first segment of the autoregressive model, the maximum number of clusters identified by DPMM is 14, which occurs only three times. The corresponding weights are: (0.63023919,0.21786768,0.07667684,0.03166412,0.02159288,0.01172519,0.00424936,0.00303308,0.00107888,0.00110433,0.00051908,0.00015267,0.00008142,0.00001527). In fact, apart from the first two categories, the sample sizes of the remaining categories are very small, with only a few or a dozen samples. Therefore, this paper ignores these categories and classifies them into the second category, considering that the innovations in the second segment follow a mixture of only two normal distributions. In the second segment of the autoregressive model, the DPMM used up to 7 mixture components to fit the innovation distribution, but this occurred only once. The weights for the mixture components were (0.462809917,0.507438017,0.004958678,0.018181818,0.001652893,0.001652893,0.003305785). The Dirichlet process often identifies more clusters than the actual number of underlying components. In this case, we consider that there are effectively only two distinct components. Meanwhile, the LASSO and MLE methods assumed that the innovations followed a single normal distribution.

We process the innovations with two main objectives: to better interpret the time series data and to enable forecasting of the time series model. To this end, we conduct predictions for the first and second segments of the autoregressive model, forecasting the values for the subsequent five time points. The results are as follows:

TABLE 5-3. The estimates for the next five time points of the first segment of the autoregressive model

	the first segment of an AR model					MSE
True values	3.839809	1.883247	4.481181	3.410880	3.354103	NULL
DPMM	3.256816	3.977992	4.588088	2.272528	2.389873	1.392970
LASSO	-0.337823	-0.002889	-0.418313	-1.410194	0.151177	15.703330
MLE	1.164346	1.067213	0.993561	0.925571	0.866678	6.470308

In the five-step-ahead time series prediction for the first segment of the autoregressive model, the estimate of the first time point by DPMM is acceptable, and the estimate of the third time point is very close to the true value. However, the errors at the other three time points are all greater than 1. This is due to the randomness involved in the estimation process, yet the MSE value remains relatively small. The LASSO regression has the largest MSE value, followed by MLE, while a smaller MSE value indicates a more accurate estimation.

TABLE 5-4. The estimates for the next five time points of the second segment of the autoregressive model

	the second segment of an AR model					MSE
True values	1.817131	1.213866	1.896725	2.018733	2.114893	NULL
DPMM	2.010977	1.793854	0.750357	0.904029	2.1030925	0.586165
LASSO	3.500133	2.112954	2.979287	5.274592	6.007251	6.112773
MLE	1.372623	0.926749	0.646227	0.466911	0.3518173	1.472070

In the prediction of the second segment of the autoregressive model, the estimate of the first time series value by the DPMM method is close to the true value, and the estimate of the second time series value is also acceptable. However, the estimates of the third and fourth time series values deviate from the true values due to the characteristics of random

International Journal of Scientific Engineering and Science
Volume 9, Issue 1, pp. 133-142, 2025. ISSN (Online): 2456-7361

sampling, while the estimate of the fifth time series value is again close to the true value. The LASSO method has the largest MSE, followed by MLE, with DPMM having the smallest MSE.

Additionally, the clustering effect of the Dirichlet process is influenced by the hyperparameter α. The closer the value of α is to 0, the fewer the number of clusters. The estimated values of α for the two segments of the autoregressive model are $\alpha_1 = 0.495951$ and $\alpha_2 = 0.426070$, respectively. In practice, the calculated values of α often tend to be close to 0, and there is no so-called "true value" for α.

VI. Case Analysis

In this subsection, we apply the conclusions drawn from the numerical simulations. First, we use the LASSO regression to identify the change points and then employ the DPMM to estimate the model parameters. This approach is applied to the closing prices of QingHai Salt Lake Industry Co., Ltd. (referred to as "Salt Lake Shares," stock code 000792) from January 2, 2020, to November 10, 2023. The data consists of 2,492 time series observations, collected hourly, and is sourced from the East Money website.

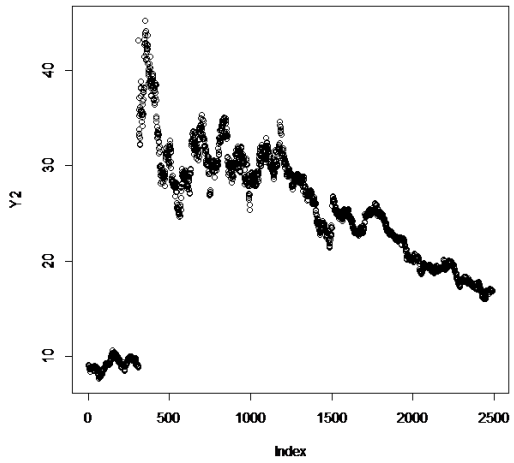

Fig. 6-1. The scatter plot of Salt Lake Shares data

Based on the raw data from Figure 6-1, an Augmented Dickey-Fuller (ADF) test was conducted. The Dickey-Fuller statistic value was -2.668, with a lag order of 13, indicating that data from the previous 13 time points were considered to assist in estimating the value of the current point. The p-value was 0.2955, which is greater than the commonly used significance levels (e.g., 0.05 or 0.01). Therefore, the time series is non-stationary.

Using LASSO regression, Figure 6-2 was obtained, which clearly identifies the change point location at 307. Subsequently, we conducted the ADF test on the two segments of time series data separately. The test results for the two segments were as follows: Dickey-Fuller = -1.4952, p-value = 0.7891 for the first segment, and Dickey-Fuller = -3.7367, p-value = 0.02215 for the second segment. Given that the p-value for the second segment is less than 0.05, it indicates the absence of a unit root, suggesting that the second segment of the time series is stationary.

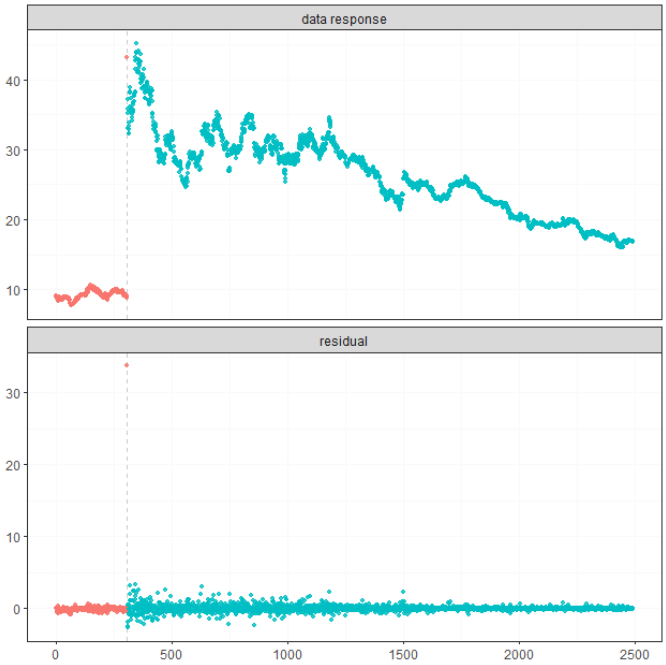

Fig. 6-2. Change point detection using LASSO regression.

In contrast, the p-value for the first segment is greater than 0.05, leading us to reject the null hypothesis and conclude that the first segment of the time series is non-stationary. To address this, we performed differencing on the first segment, resulting in a Dickey-Fuller statistic of -6.4397, with the p-value being significantly less than 0.01. This indicates that the differenced time series is stationary.

TABLE 6-1. The estimated values of the coefficients in the two autoregressive models

	b_{11}	b_{21}	b_{22}
DPMM	0. 003613054	0. 42549951	0. 07383823

In Table 6-1, based on the AIC, the optimal orders of the two autoregressive models are determined to be first-order and second-order, respectively. The estimated coefficients for the first segment of the autoregressive model are small because the original data values are close to each other, resulting in small differences after differencing. For the second segment of the autoregressive model, the time series values are strongly related to their first-order lagged values. In fact, in most autoregressive models, time points closer to the t-th moment are more useful for the model.

TABLE 6-2. The estimation of the unknown distributions of innovations in the two-segment autoregressive models

	Mixture of Normal Distributions	MSE
the first paragraph AR	N(0, 0. 1981)	0.0146692
the second paragraph AR	N(0, 0. 2327)+ N(0, 0. 8000)	1.166021

From Table 6-2, it can be observed that the unknown distribution of innovations in the first segment of the model can be approximated by a single normal distribution, while the unknown distribution of innovations in the second segment requires a mixture of two normal distributions for fitting. During the clustering process, up to 10 clusters were generated, with weights as follows: (0.5380384968,0.2873510541,0.1448212649,0.0210815765,0.0018331806,0.0018331806, 0.0009165903, 0.0009165903, 0.0009165903,0.0022914757). Considering this particular clustering result, there appear to be three normal distributions. However, across numerous iterations, the weight of the third cluster rarely exceeds 5%. Therefore, we take the average weight over 500 iterations, and the weight of the third cluster becomes negligible. Ultimately, we obtain a mixture of two normal distributions with weights: 0.5614537N(0, 0.2327) + 0.4385463N(0,0.8000).

Additionally, since the first segment of the model is derived from differenced time series values, which are relatively small, its MSE is also small. In contrast, the second segment of the model uses the original data, which contains over two thousand data points, resulting in a much larger MSE compared to the first segment. The clustering parameters generated by the two autoregressive models are $\alpha_1 = 0.2079229$ and $\alpha_2 = 0.6201116$, respectively. As previously mentioned, the closer the value of α is to 0, the fewer the number of clusters generated.

Therefore, the time series estimation model should be modified as follows:

$$y = \begin{cases} 0.0036 \times y_{t-1} + \varepsilon_1, n \le 306 \\ 0.4255 \times y_{t-1} + 0.0738 y_{t-2} + \varepsilon_2, n > 306 \end{cases} \quad (39)$$

Where $\varepsilon_1 \sim \mathrm{N}(0, 0.1981)$, $\varepsilon_2 \sim 0.5614537\mathrm{N}(0,\ 0.2327) + 0.4385463\mathrm{N}(0, 0.8000)$.

VII. Summary

This paper considers a piecewise autoregressive model with multiple change points. First, the LASSO method is employed to identify the number and locations of the change points, transforming the originally non-stationary autoregressive model into a piecewise stationary autoregressive model. Subsequently, the DPMM is introduced into the autoregressive model, with the DP serving as the prior for the parameters of the unknown distribution of innovations. A mixture of normal distributions is used to approximate this unknown distribution, thereby enhancing the flexibility of parameter estimation in the model. Through numerical simulations and comparisons with LASSO regression and MLE, it is demonstrated that the DPMM outperforms these two methods in both parameter estimation and time series prediction. This advantage is particularly pronounced when the unknown distribution is a mixture of normal distributions.

References

[1]. Yule G U. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers [J]. Philosophical Transactions of the Royal Society of London Series A, 1927, 226:267-298.
[2]. Walker G T. On periodicity in series of related terms [J]. Proceedings of the Royal Society of London Series A, 1931, 131(818):518-532.
[3]. Page, E. S. "Continuous Inspection Schemes." *Biometrika* 41, no. 1/2 (1954): 100–115.
[4]. Quandt R E. The estimation of the parameters of a linear regression system obeying two separate regimes[J]. Journal of the American Statistical Association,1958, Vol.53, No.284:873-880.
[5]. Chen X R. Introduction of the statistical analysis of change points. Journal of Applied Statistics and Managment, 1991, 10(1): 55–58.
[6]. Wang L M. Research advances in statistical analysis of change point. Statistical Research, 2003, 20(1): 50–51.
[7]. Brodsky E, Darkhovsky B S. Nonparametric Methods in Change Point Problems. Berlin: Springer Science & Business Media, 2013.
[8]. Aue A , Horvath L . Structural breaks in time series. Journal of Time Series Analysis, 2013, 34(1): 1–16.
[9]. Barry D, Hartigan J A. A Bayesian analysis for change point problems. Journal of the American Statistical Association, 1993, 88(421): 309–319.
[10]. Bai J. Likelihood ratio tests for multiple structural changes. Journal of Econometrics, 1999, 91: 299–323.
[11]. Zou C L, Yin G S, Feng L, et al. Nonparametric maximum likelihood approach to multiple change-point problems. The Annals of Statistics, 2014, 42(3): 970–1002.
[12]. Zhou Y H. Ni Z X. Zhu P F. Study of change points in Shanghai Composite Index based on least absolute deviation criterion. Chinese Journal of Management Science, 2015, 23(10): 38–45.
[13]. Chen Z, Xu Q, Li H. Inference for multiple change points in heavy-tailed time series via rank likelihood ratio scan statistics. Economics Letters, 2019, 179(7): 53–56.
[14]. R.Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B, 58:267–288, 1996.
[15]. Zou, H., Hastie, T., Tibshirani, R. Sparse principal component analysis[J]. Journal of Computational and Graphical Statistics, 2006, 15(2): 265-286.
[16]. Ciuperca G. Model selection by LASSO methods in a change-point model[J]. Statistical Papers,2012, Vol.55, No.2: 349-374.
[17]. Ciuperca G. Adaptive Lasso model selection in a multiphase quantile regression[J]. Statistics,2016, Vol.50, No.5:1100-1131.
[18]. Zhang B, Jun G and Li-Feng L. Multiple change-points estimation in linear regression models via sparse group Lasso[J]. IEEE Transaction on Signal Processing. 2015, Vol.63, No.9:2209-2224.
[19]. Qian J and L. Su. Shrinkage estimation of common breaks in panel models via adaptive group fused lasso[J]. Journal of Econometrics, 2016, Vol.191, No.1:86-109.
[20]. Liao, A. M. (2018). Multiple Change-Points Estimation in Linear Regression Models via Adaptive Group Lasso Algorithm. Xiamen: Xiamen University.
[21]. Yang, Z. X., Wei, Y. S., & Jia, W. Y. (2020). Convergence Speed of Change-Point Location Estimation in Linear Regression Models. *Journal of Huaihua Normal University (Natural Science Edition)*, 41(1), 12-17.
[22]. Ferguson T S. A Bayesian analysis of some nonparametric problems[J]. The annals of statistics, 1973: 209-230.
[23]. Escobar M D, West M. Bayesian density estimation and inference using mixtures[J]. Journal of the american statistical association, 1995, 90(430): 577-588.
[24]. Hong L, Martin R. A flexible Bayesian nonparametric model for predicting future insurance claims[J]. North American Actuarial Journal, 2017, 21(2): 228-241.
[25]. Adesina O S, Agunbiade D A, Oguntunde P E. Flexible Bayesian Dirichlet mixtures of generalized linear mixed models for count data[J]. Scientific African, 2021, 13: e00963.
[26]. Zhang, Y. X., Meng, S. W., & Tian, M. Z. (2021). A Semi-Parametric Bayesian Hierarchical Quantile Regression Model and Its Application in Insurance Company Cost Analysis. *Journal of Applied Statistics and Management*, 40(3), 381-394.
[27]. Kalli M, Griffin J E. Bayesian nonparametric vector autoregressive models[J]. Journal of econometrics, 2018, 203(2): 267-282.
[28]. Liu, W. Y. (2015). Estimation of Semi-Parametric Double Autoregressive Models Based on Dirichlet Mixture Processes. Dissertation, Xiamen University.
[29]. Lau J W, So M K P. A Monte Carlo Markov chain algorithm for a class of mixture time series models[J]. Statistics and computing, 2011, 21(1): 69-81.

[30]. Nakatsuma T. A Markov-chain sampling algorithm for GARCH models[J]. Studies in Nonlinear Dynamics & Econometrics, 1998, 3(2).
[31]. Osborne, M. R., Presnell, B., Turlach, B. A. On the lasso and its dual[J]. Journal of Computational and Graphical Statistics, 2000, 9(2): 319-337.
[32]. Friedman, J., Hastie, T., Tibshirani, R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1.
[33]. D Blackwell, and JB Macqueen. "Ferguson Distributions Via Polya Urn Schemes", Annals of Statistics 1.2 (1973): 353-355.
[34]. Sethuraman, Jayaram . "A Constructive Definition of the Dirichlet Prior." Statistica Sinica 4.2(1994):639-650.
[35]. Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the american statistical association, 90(430), 577-588.
[36]. Antoniak C E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems[J]. The annals of statistics, 1974: 1152-1174.