Statistica Sinica **35** (2025), 457-478
doi:https://doi.org/10.5705/ss.202021.0405

PARAMETRIC MODAL REGRESSION WITH AUTOCORRELATED ERROR PROCESS

Tao Wang*

University of Victoria

Abstract: We propose an efficient two-step estimation procedure for a parametric modal regression with autoregressive errors. The procedure relies on estimating a parametric transformation of the dependent variable from data using a (penalized) kernel-based objective function. We establish asymptotic normality for the resulting estimator and demonstrate that it possesses oracle properties, as if the true order of autoregressive error structure were known in advance. To numerically estimate modal parameter and determine the order of error structure, two modified (penalized) modal expectation-maximization (MEM) algorithms are developed. Furthermore, we present a modal residual-based autocorrelation test and show that the statistic is asymptotically distributed as a $\mathcal{X}^2$ distribution. Monte Carlo simulations and an empirical analysis are conducted to illustrate the finite sample performance of the resultant estimator. We also discuss the extension of the results to a nonparametric modal regression model.

Key words and phrases: Autoregressive error, MEM algorithm, modal regression, oracle property, order selection, residual-based test.

1. Introduction

Modal regression has recently attracted much attention due to its robustness for skewed and heavy-tailed data, which can be treated as a complement to mean or median (quantile) regression; see Ullah, Wang and Yao (2021, 2022, 2023). The main objective of modal regression is to capture how covariates X affect the "most likely" (mode) value of a response variable Y, as denoted by

$$Mode(Y \mid X) = \underset{Y}{\operatorname{argmax}}\, f_{Y|X}(Y \mid X), \tag{1.1}$$

where $f_{Y|X}(Y \mid X)$ represents the conditional density of Y given X. The modal regression line can then be obtained by nonparametrically estimating the aforementioned conditional density function (Chen et al., 2016). However, because of the "curse of dimensionty", such a density-based estimation is difficult to implement. Similar to mean or median (quantile) regression, we can avoid nonparametrically estimating conditional density and achieve different types of modal regression models by directly imposing structural assumptions on

*Corresponding author. E-mail: taow@uvic.ca

$Mode(Y \mid X)$; see Yao and Li (2014), Chen (2018), Ullah, Wang and Yao (2021, 2022, 2023), and references therein for details.

For illustration, suppose that random samples $\{Y_t, X_t\}_{t=1}^n$ are collected in order to establish a conditional modal regression model

$$Y_t = Mode(Y_t \mid X_t) + \varepsilon_t,\ t = 1, \ldots, n, \tag{1.2}$$

where $Y_t \in \mathbb{R}$, $X_t \in \mathbb{R}^p$ (which may include lagged values of Y_t), and $\{\varepsilon_t\}_{t=1}^n$ are random errors with $Mode(\varepsilon_t \mid X_t) = 0$ almost surely. This construction of a modal regression line allows for nonuniqueness, but all of the models considered in this paper are assumed to have a global unique mode for convenience purposes. According to Ullah, Wang and Yao (2023), we can impose a linear regression structure $X^T\theta$ on $Mode(Y_t \mid X_t)$ to explain the mode relationship between response and explanatory variables, where θ is an unknown parameter vector in the parameter space $\Theta_\theta \subset \mathbb{R}^p$ and T represents the transpose of a matrix or a vector. Following that, the parameter θ can be estimated using a kernel-based objective function constructed from the density of error term ε_t (Kemp and Santos Silva, 2012; Yao and Li, 2014)

$$Q_n(\theta) = \frac{1}{nh_n} \sum_{t=1}^{n} K\left(\frac{Y_t - X_t^T\theta}{h_n}\right), \tag{1.3}$$

where $K(\cdot)$ is a kernel function with $\int_{\mathbb{R}} K(t)dt = 1$ and h_n is a non-stochastic strictly positive bandwidth dependent on n. To acquire the reliable estimator from (1.3), we need to assume that the error term ε_t, denoted as $Y_t - Mode(Y_t \mid X_t)$, is independent and identically distributed (i.i.d.).

The i.i.d. assumption, on the other hand, may be violated when data are collected sequentially in time, such as the financial return, individual income, or interest rate, which naturally imposes a correlated structure for error terms. As a result, the error terms in (1.2) will possess serial correlation and the conditional mode of ε_t on X_t is no longer zero due to the absence of mode additive property. If such correlation is not taken into account, the modal estimator of θ from (1.3)—the "most likely" effect—may be inefficient or biased (when ε_t and X_t have dependence relationship), rendering any inference based on it invalid. Ullah, Wang and Yao (2021) mentioned that incorporating the information from error autocorrelation structure can lead to a more efficient modal estimator, but they did not address the critical issues related to practical implementation. The question of how to best incorporate error correlation information to recover modal coefficient θ remains unanswered. To fill the literature gap, we in this paper assume that $\{Y_t, X_t\}_{t=1}^n$ is a sequence of strictly stationary random vectors and aim at estimating the conditional modal regression with autoregressive (AR) error ε_t by directly modeling the error process. We show that the autocorrelation function of the error process contains useful information for inferring and can be

properly used to improve the performance of modal estimators.

Serial correlation in the error terms of mean or median (quantile) regression has been investigated intensively as a main class of dynamic regression models; see Opsomer, Wang and Yang (2001), Xiao et al. (2003), Su and Ullah (2007), Wang, Li and Tsai (2007), Martins-Filho and Yao (2009), Chen, Li and Li (2015), among others. All of these studies can be viewed as extensions of the mean or median (quantile) regression literature from the typical case where error terms are i.i.d. to instances where specific parametric or nonparametric structures for error terms are allowed. In light of these research, we study parametric linear modal regression with errors represented by a stationary AR process with finite order d, where we recover modal parameters by directly incorporating the autocorrelated error process; see (2.3). However, the true AR order in the errors is rarely available in advance and is simply assumed to be an upper bound in practice. The misspecification of the lagged order will result in the reduction of estimation precision and efficiency. Although information criterion-based methods can usually identify the order, these methods are sensitive to small changes in the data and ignore stochastic errors inherited in the process of determining order (Qiu, Li and You, 2015). Furthermore, the existing information criterion in mean or median (quantile) regression cannot be utilized for modal regression with a kernel-based objective function. These motivate us to propose a modal variable selection procedure through a penalty function based on the estimated residuals to determine the AR order.

Specifically, we propose an efficient two-step estimation procedure for estimating the modal regression parameters while accounting for the AR error structure. In the first step, we select an arbitrary large value for d (upper bound) to obtain the initial estimate of θ, and in the second step, we update the estimate with order selection using penalized modal regression. Consequently, the final modal estimator of θ is based on a parametric transformation of the dependent variable, which must be estimated from data using a (penalized) kernel-based objective function. We establish asymptotic normality for the resulting estimator and demonstrate that it has the oracle properties as if the true error structure were known in advance. Following Li, Ray and Lindsay (2007) and Yao (2013), we suggest two modified (penalized) MEM algorithms to numerically estimate modal parameters. Monte Carlo simulations and an empirical analysis are conducted to illustrate the finite sample performance of the resulting estimators, where we show that accounting for autocorrelation in the errors can result in substantially more accurate and efficient modal estimates. In spite of the extensive literature on mode, there is little research on variable selection in modal regression. The developed order selection procedure can also be considered as a contribution to modal variable selection literature.

The proposed estimation methodology relies on the presence of autocorrelation in the error terms. If the modal regression model does not contain an

autocorrelated error process, the developed method may lose efficiency. As a result, it is particularly important to check for any signs of autocorrelation in modal regression. To accomplish this objective, we suggest a residual-based test for autocorrelation in modal regression models. In general, the Breusch-Godfrey LM test can be applied to the residuals of a baseline modal regression. Furno (2000), for example, recommended a LM test based on the least absolute deviation residuals. Nevertheless, Huo et al. (2017) argued that such a LM test could result in potentially large size distortions for median (quantile) regression. Particularly, the LM statistic either diverges to infinity or weakly converges to a distribution that is different from the $\mathcal{X}^2$ distribution. Given that modal regression can be regarded as a special case of quantile regression, which is obtained by maximizing the density function, such size distortions may also appear in modal regression if the LM test is used. We thus extend the results in Huo et al. (2017) to parametric modal regression to propose a modal residual-based test and show that the statistic is asymptotically distributed as a $\mathcal{X}^2$ distribution.

The layout of the remainder of this paper is as follows. In Section 2, we propose an efficient two-step estimation procedure to estimate the modal regression coefficients. In Section 3, we present the asymptotic properties of the resulting modal estimators. In Section 4, we develop a modal residual-based test for autocorrelation in parametric modal regression. We report an empirical analysis in section 5 and conclude the paper in Section 6. All technical proofs and Monte Carlo simulations are presented in the supplementary file, as well as the extension to nonparametric modal regression.

2. Modal Regression with AR Errors

We begin this section by introducing the error structure of (1.2). Since most Gaussian stationary processes can be approximated by an AR process of sufficiently high order, we assume that ε_t is a stationary AR(d) series

$$\varepsilon_t = \beta_1\varepsilon_{t-1} + \cdots + \beta_d\varepsilon_{t-d} + \eta_t, \quad t = d+1, \ldots, n, \tag{2.1}$$

where $1 - \sum_{j=1}^{d} \beta_j z^j \neq 0$ for all z such that $|z| \leq 1$ on the complex plane, $\beta = (\beta_1, \ldots, \beta_d)^T$ is a $d \times 1$ vector of unknown AR coefficients, and $\{\eta_t\}_{t=d+1}^{n}$ are i.i.d. random errors with zero mode. Because the conditional modal estimators and their asymptotic properties are irrelevant to the moments of error terms, compared to mean regression, we do not impose any moment conditions on η_t, i.e., allow $AR(d)$ with $\mathbb{E}(\varepsilon_t^2) = \infty$ and the distribution of η_t to be heavy-tailed or asymmetric. Note that the distribution of the errors in practice can potentially be heteroskedastic and asymmetric simultaneously, motivating the need of the suggested modal estimation.

Remark 1. As pointed out by the editor, with i.i.d. random errors, the difference between the mode and the mean is a constant. Since all AR models can be written as a $MA(\infty)$ model, the modal regression and a zero-mean noise regression only differ by a constant term. We can then combine modal and mean regressions to achieve coefficient estimators, which has been utilized in Ullah, Wang and Yao (2021). However, compared to this combined estimation procedure, the proposed modal estimation can increase efficiency and has better prediction performance if the distribution of the error term or dependent variable is skewed; see the simulation results in the supplementary file.

The model in (1.2) can then be written as

$$Y_t = X_t^T\theta + \beta_1\varepsilon_{t-1} + \cdots + \beta_d\varepsilon_{t-d} + \eta_t \tag{2.2}$$

by incorporating the error structure information. If the values of $\{\varepsilon\}_{t=d+1}^n$ were available, (2.2) would be a valid linear modal regression equation, and the coefficients could be estimated directly using the kernel-based objective function (1.3). In practice, however, they are not available (neither directly nor indirectly), and need to be substituted with appropriate estimates.

To obtain the consistent estimate of θ, we replace ε_{t-j} with $Y_{t-j} - \theta X_{t-j}$ for $j = 1, 2, \ldots, d$ and get

$$Mode\left(Y_t \mid F_{t-1}\right) = X_t^T\theta + \sum_{j=1}^{d} \beta_j \left(Y_{t-j} - X_{t-j}^T\theta\right) \tag{2.3}$$

provided that η_t is independent of F_{t-1}, where $F_{t-1} = \sigma(\{Y_s, X_s\} : s \le t)$ is the σ-field generated by $(\{Y_s, X_s\} : s \le t)$. If the value of order d were known, the parameters can be identified and estimated by using (2.3) straightforwardly. However, we do not know the exact value of d practically. To improve estimation performance and propose a modal residual-based autocorrelation test, we instead use the estimate of θ from (2.3) with an arbitrary chosen d (e.g., upper bound chosen by ACF and PACF) to construct a preliminary consistent estimate of ε_t. We then plug it back into the AR model in (2.3) and simultaneously estimate β and select d by maximizing a penalized kernel-based objective function. Finally, we plug the consistent penalized estimate of β into (2.2) to define a new pseudo response variable, converting the AR regression problem to a parametric modal regression framework. The entire estimation procedure is built on

$$Mode\left(Y_t - \sum_{j=1}^{s} \beta_j\varepsilon_{t-j} \mid F_{t-1}\right) = X_t^T\theta, \tag{2.4}$$

where $s \le d$ is the selected order. Under some mild conditions, the final estimator of θ is shown to have similar asymptotic bias and variance as the estimator of

linear modal regression with i.i.d. observations, except for the explicit value of the density function for error terms (see Theorem 7).

2.1. Feasible estimation procedure

To efficiently account for the AR error structure, we in the **first step** estimate θ in (2.3) by maximizing the following kernel-based objective function

$$Q_{n_0}(\theta,\beta)=\frac{1}{n_0h_1}\sum_{t=d+1}^{n}K\left(\frac{Y_t-X_t^T\theta-\sum_{j=1}^{d}\beta_jY_{t-j}+\sum_{j=1}^{d}\beta_jX_{t-j}^T\theta}{h_1}\right), \quad (2.5)$$

where $n_0=n-d$ is the effective sample size and $h_1=h_1(n_0)>0$ is a scalar bandwidth sequence satisfying $h_1\to 0$ as $n_0\to\infty$. According to Yao and Li (2014) and Ullah, Wang and Yao (2021, 2022, 2023), the choice of kernel function is less important in modal estimation than the choice of bandwidth. We thus choose the Gaussian kernel in this paper for simple calculations; see the role of the Gaussian kernel in the following MEM algorithms. We use $\tilde{\theta}$ and $\tilde{\beta}$ to represent the first-step modal estimators from (2.5).

After obtaining the estimate $\tilde{\theta}$, we in the **second step** use it to construct the estimate $\hat{\varepsilon}_t$ with $\hat{\varepsilon}_t=Y_t-X_t^T\tilde{\theta}$. We then conduct a modal variable selection procedure to determine the AR order from the data by adding a penalty term into the kernel-based objective function

$$Q_{n_0}^{P}(\beta)=\frac{1}{n_0h_2}\sum_{t=d+1}^{n}K\left(\frac{\hat{\varepsilon}_t-\sum_{j=1}^{d}\beta_j\hat{\varepsilon}_{t-j}}{h_2}\right)+\sum_{j=1}^{d}p_{\lambda_j}\left(|\beta_j|\right), \quad (2.6)$$

where $h_2=h_2(n_0)>0$ is a sequence of bandwidths that depends on n_0 satisfying $h_2\to 0$ as $n_0\to\infty$, and $p_{\lambda_j}(\cdot)$ is a penalty function with the tuning parameter λ_j controlling the model complexity. In general, the larger the λ_j, the simpler the modal regression model, with fewer variables selected. The penalty function $p_{\lambda_j}(\cdot)$ and the tuning parameter λ_j are not required to be identical for all j. We denote the estimator from (2.6) as $\hat{\beta}^P$. The selected AR order $s\le d$ is the highest order whose corresponding coefficient is not zero. In Section 3, we show that the penalized modal estimator is consistent and enjoys selection consistency as well as asymptotic normality.

Remark 2. To reduce model complexity, we concentrate on penalized methodology for selecting AR order. Aside from that, we can also utilize a modified Bayesian information criterion (BIC) for finding the order, where

$$d_{opt}=\underset{d}{\operatorname{argmin}}\,BIC(d)=-Q_{n_0}(\theta,\beta)+(n_0h_1^3)^{-1}\log(n_0h_1^3)df_d$$

and df_d denotes the number of coefficients. This approach, however, disregards stochastic errors inherited in the stages of determining the AR process. It is also

computationally intensive because $BIC(d)$ depends on the estimates of θ and β, and it is challenging to compare all candidate subprocesses to choose the optimal model when the maximal order is very large.

To obtain the final modal estimator of θ, we maximize the following kernel-based objective function derived from (2.4)

$$Q_{n_0}(\theta) = \frac{1}{n_0 h_3} \sum_{t=d+1}^{n} K\left(\frac{Y_t - \sum_{j=1}^{d} \hat{\beta}_j^P \hat{\varepsilon}_{t-j} - X_t^T \theta}{h_3}\right) \tag{2.7}$$

by submitting ε_{t-j} and β_j with the corresponding estimates, where $h_3 = h_3(n_0) > 0$ is a sequence of bandwidths that depends on n satisfying $h_3 \to 0$ as $n_0 \to \infty$. The final estimator from (2.7) is denoted as $\hat{\theta}$. In Section 3, we show that under appropriate assumptions, the final modal estimator $\hat{\theta}$ is asymptotically equivalent to the infeasible estimator from (2.4).

Remark 3 (Variable and Order Selection). In the absence of prior knowledge, a large number of variables may be included in model (2.2) to reduce potential model bias, but could result in less predictive power and greater interpretation difficulty. In this case, we can apply the penalized objective function $Q_{n_0}(\theta, \beta) + \sum_{k=1}^{p} p_{\lambda_k}(|\theta_k|) + \sum_{j=1}^{d} p_{\mu_j}(|\beta_j|)$ to simultaneously select the significant explanatory variables and determine the order of autocorrelation with the properly chosen penalty functions. A nature approach to obtain estimates is to utilize an iterate procedure by maximizing the above objective function with respect to θ and β, separately. We leave the specifics of such an investigation to another research.

2.2. Practical algorithms

2.2.1. MEM algorithm

There exist both θ and β in (2.5), implying that we need to maximize the objective function with respect to (θ, β) iteratively. To be more specific, for a given value $\tilde{\beta}_j$, $j = 1, 2, \ldots, d$, we maximize the following kernel-based objective function to obtain the estimate of θ

$$\frac{1}{n_0 h_1} \sum_{t=d+1}^{n} K\left(\frac{Y_t - X_t^T \theta - \sum_{j=1}^{d} \tilde{\beta}_j Y_{t-j} + \sum_{j=1}^{d} \tilde{\beta}_j X_{t-j}^T \theta}{h_1}\right). \tag{2.8}$$

Then, we maximize the kernel-based objective function as outlined below to achieve a new estimate of β given the estimate $\tilde{\theta}$

$$\frac{1}{n_0 h_1} \sum_{t=d+1}^{n} K\left(\frac{Y_t - X_t^T \tilde{\theta} - \sum_{j=1}^{d} \beta_j Y_{t-j} + \sum_{j=1}^{d} \beta_j X_{t-j}^T \tilde{\theta}}{h_1}\right). \tag{2.9}$$

The above two functions are maximized iteratively until convergence. The choice of bandwidths will be introduced later, whereas the initial values can be obtained by running a mean or median (quantile) regression.

Algorithm 1 MEM Algorithm.

E-Step. Calculate the weight $\pi(t \mid \theta^{(g)})$ with the preliminary estimate of the modal parameter as

$$\pi\left(t \mid \theta^{(g)}\right) = \frac{K\left(\tilde{Y}_t - X_t^T\theta^{(g)} + \sum_{j=1}^d \tilde{\beta}_j X_{t-j}^T\theta^{(g)}/h_1\right)}{\sum_{t=d+1}^n K\left(\tilde{Y}_t - X_t^T\theta^{(g)} + \sum_{j=1}^d \tilde{\beta}_j X_{t-j}^T\theta^{(g)}/h_1\right)}.$$

M-Step. Update $\theta^{(g+1)}$ with the weight calculated in E-Step

$$\begin{aligned}\theta^{(g+1)} &= \underset{\theta}{\operatorname{argmax}} \sum_{t=d+1}^n \left\{\pi\left(t \mid \theta^{(g)}\right) \log \frac{1}{h_1} K\left(\frac{\tilde{Y}_t - X_t^T\theta + \sum_{j=1}^d \tilde{\beta}_j X_{t-j}^T\theta}{h_1}\right)\right\} \\ &= (X^{*T} W_X X^*)^{-1} X^{*T} W_X Y^*,\end{aligned}$$

where g is the iteration indicator, $\tilde{Y}_t = Y_t - \sum_{j=1}^d \tilde{\beta}_j Y_{t-j}$, $X_t^* = X_t^T - \sum_{j=1}^d \tilde{\beta}_j X_{t-j}^T$, $X^* = (X_{d+1}^*, \ldots, X_n^*)^T$, $Y^* = (\tilde{Y}_{d+1}, \ldots, \tilde{Y}_n)^T$, and W_X is an $(n-d) \times (n-d)$ diagonal matrix consisting of diagonal elements $\{\pi\left(t \mid \theta^{(g)}\right)\}_{t=d+1}^n$.

Nevertheless, because there is no explicit expression for the estimator in modal regression, obtaining a modal estimator by maximizing the kernel-based objective function is difficult. To numerically estimate the proposed models, we develop a modified MEM Algorithm 1 by virtue of Gaussian kernel based on Li, Ray and Lindsay (2007) and Yao (2013), which can provide an explicit expression for the modal estimator in M-Step (log-maximization). Due to space constraints, we only present the algorithm for (2.8), while other kernel-based objective functions can be solved using the same procedures.

We iterate E-Step and M-Step until the total error of the estimate approaches the preassigned constraint. In practice, a tolerance ϵ is set as 10^{-5} and the algorithm is iterated until $\|\tilde{\theta}^{(g+1)} - \tilde{\theta}^{(g)}\| < \epsilon = 10^{-5}$, where $\|\cdot\|$ denotes the Euclidean norm, defined as $\|A\| = \{tr(AA^T)\}^{1/2}$. Consistent with the result in Yao and Li (2014), the proposed MEM algorithm has the ascending property, which means that at each iteration $Q_{n_0}(\theta^{(g+1)}, \tilde{\beta}) \geq Q_{n_0}(\theta^{(g)}, \tilde{\beta})$ and the equality holds if and only if $\theta^{(g+1)} = \theta^{(g)}$, ensuring the convergence of MEM algorithm. In general, the MEM algorithm leads to optimization problems suffering from the local maximum with small bandwidths. To address this issue, we can try different starting points of parameters (i.e., mean, median, or quantile estimates) on each occasion to obtain a stable estimate. If the $Q_{n_0}(\cdot)$ function is assumed to be unimodal, the initial values for the algorithm will not produce much effect on the results of estimation. Accordingly, the algorithm will not be trapped at a local maximum.

2.2.2. Penalized MEM algorithm

There are numerous penalty functions available in the literature, including LASSO, adaptive LASSO, ridge, elastic net, among others (Fan and Lv, 2010). In this paper, we choose the smoothly clipped absolute deviation (SCAD) penalty because of its unbiasedness for a true coefficient, sparsity to reduce model complexity, and continuity to avoid unnecessary variation. The first derivative of $p_\lambda(|\beta_j|)$ for the SCAD penalty is defined as

$$p_\lambda^{(1)}(|\beta|) = \lambda\left\{I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\right\} \tag{2.10}$$

for $\beta > 0$, where $(t)_+ = tI(t > 0)$ with $I(\cdot)$ being the indicator function and $a = 3.7$ suggested by Fan and Li (2001) from a Bayesian point of view. Notice that the SCAD penalty is a quadratic spline with knots at $\pm\lambda$ and $\pm a\lambda$. With the proper choice of tuning parameter, we can shrinkage some coefficients to zero with probability converging to one, providing the theoretical support for AR order choice.

The maximization of the SCAD penalized objective function is not easy because it is irregular at the origin and lacks a second derivative at some points. To circumvent this difficulty, we take the local quadratic approximation for the SCAD penalty function suggested by Fan and Li (2001). Suppose we can obtain an estimate $\beta_j^{(g)}$ in the gth step that is close to the true parameter β_j. If $|\beta_j^{(g)}|$ is close to 0, then set $\hat{\beta}_j^P = 0$. Otherwise, the SCAD penalty can be locally approximated by a quadratic function as

$$\left\{p_{\lambda_j}(|\beta_j|)\right\}^{(1)} = p_{\lambda_j}^{(1)}(|\beta_j|) \cdot \operatorname{sgn}(\beta_j) \approx \frac{p_{\lambda_j}^{(1)}\left(|\beta_j^{(g)}|\right)}{|\beta_j^{(g)}|}\beta_j, \tag{2.11}$$

which is equivalent to

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_{j0}|) + \frac{1}{2}\left\{\frac{p_{\lambda_j}^{(1)}\left(|\beta_j^{(g)}|\right)}{|\beta_j^{(g)}|}\right\}\left(\beta_j^2 - \beta_j^{(g)2}\right). \tag{2.12}$$

We then propose a penalized MEM algorithm for (2.6). Starting from an initial estimate, we iterate the E-Step and M-Step until some convergence criterion is met (we explain more about convergence property in the supplementary file). In contrast to Algorithm 1, we now need to select the tuning parameter λ_j. Since the magnitude of λ_j is proportional to the standard error of the estimate of β_j, we follow Fan and Li (2004) to set $\lambda_j = \lambda SE(\hat{\beta}_j)$, where λ is a scalar variable and $SE(\hat{\beta}_j)$ is the standard error from (2.6) that can be acquired by a modal Bootstrap procedure; see Ullah, Wang and Yao (2021). After that, the original d-dimensional optimization is reduced to a one dimensional problem. We select

λ based on the BIC-type criterion relying on the kernel-based objective function. The simulation results in the supplementary file suggest that the correct order of AR error terms could be identified by setting some of coefficients to zero if d is chosen large.

Algorithm 2 Penalized MEM for Order Selection.

Selection of λ_j. Set $\lambda_j = \lambda SE(\tilde{\beta}_j)$. Utilize a modified BIC to select λ

$$\min_{\lambda} BIC(\lambda) = -\frac{1}{n_0 h_2} \sum_{t=d+1}^{n} K\left(\frac{\hat{\varepsilon}_t - \sum_{j=1}^{d} \hat{\beta}_j^P \hat{\varepsilon}_{t-j}}{h_2}\right) + \frac{\log(n_0 h_2^3)}{n_0 h_2^3} edf_\lambda,$$

where edf_λ represents the effective degrees of freedom (as measured by the number of nonzero coefficients of $\hat{\beta}^P$ and $n_0 h_2^3$ indicates the effective sample size (consistent with the modal convergence rate).

E-Step. Calculate the weight $\pi(t \mid \beta^{(g)})$ with the preliminary estimate of the modal parameter as

$$\pi\left(t \mid \beta^{P(g)}\right) = \frac{K\left(\hat{\varepsilon}_t - \sum_{j=1}^{d} \hat{\varepsilon}_{t-j} \beta_j^{P(g)} / h_2\right)}{\sum_{t=d+1}^{n} K\left(\hat{\varepsilon}_t - \sum_{j=1}^{d} \hat{\varepsilon}_{t-j} \beta_j^{P(g)} / h_2\right)}.$$

M-Step. Update $\beta^{P(g+1)}$ with the weight calculated in E-Step

$$\begin{aligned}\beta^{P(g+1)} = \underset{\beta}{\operatorname{argmax}} \sum_{t=d+1}^{n} \Bigg[& \pi\left(t \mid \beta^{P(g)}\right) \log\left\{\frac{1}{h_2} K\left(\frac{\hat{\varepsilon}_t - \sum_{j=1}^{d} \hat{\varepsilon}_{t-j} \beta_j}{h_2}\right)\right\} \\ & - \frac{n_0}{2} \sum_{j=1}^{d} \left\{\frac{p_{\lambda_j}^{(1)}(|\beta_j^{P(g)}|)}{|\beta_j^{P(g)}|}\right\} \beta_j^2 \Bigg] = \{\hat{e}^T W_e \hat{e} + n_0 \Sigma_\lambda(\beta^{P(g)})\}^{-1} \hat{e}^T W_e \hat{\varepsilon},\end{aligned}$$

where $\hat{e} = (\hat{\varepsilon}_{-1}, \ldots, \hat{\varepsilon}_{-d})$, $\hat{\varepsilon}_{-j} = (\hat{\varepsilon}_{d+1-j}, \ldots, \hat{\varepsilon}_{n-j})$, $\hat{\varepsilon} = (\hat{\varepsilon}_{d+1}, \ldots, \hat{\varepsilon}_n)^T$, W_e is an $(n-d) \times (n-d)$ diagonal matrix with diagonal elements $\{\pi(t \mid \beta^{P(g)})\}_{t=d+1}^{n}$, and $\Sigma_\lambda(\beta^{(g)}) = diag\{p_{\lambda_1}^{(1)}(|\beta_1^{P(g)}|)/|\beta_1^{P(g)}|, \ldots, p_{\lambda_d}^{(1)}(|\beta_d^{P(g)}|)/|\beta_d^{P(g)}|\}$ for nonvanished $\beta^{P(g)}$.

3. Asymptotic Properties

Since the AR order does not necessarily increase with sample size (Wang, Li and Tsai, 2007), we do not assume an increasing d (i.e., being independent of n) when investigating the theoretical properties of the proposed estimators. To facilitate the asymptotic analysis, we make the following assumptions.

C1. Parameter Space: The true values of parameters θ_0 and β_0 are in the interior of the known compact parameter space $\Theta_\theta \times \Theta_\beta$, which is a subset of Euclidean space $\mathbb{R}^p \times \mathbb{R}^d$.

C2. Stationary: The strictly stationary process $\{(X_t^T, \varepsilon_t)^T\}$ is strong mixing with mixing coefficients $\alpha(j)$ that satisfy $\sum_{j=1}^{\infty} j^2 \alpha(j)^{\delta/(1+\delta)} < \infty$ for some

$\delta > 0$, where $\alpha(n) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_n} |P(A\cap B) - P(A)P(B)|$ with $\mathcal{F}^0_{-\infty}$ being a σ-field generated by $\{(X_t^T, \varepsilon_t) : t \le 0\}$ and $\mathcal{F}^\infty_n$ being a σ-filed generated by $\{(X_t^T, \varepsilon_t) : t \ge n\}$.

C3. Kernel Function: The kernel function $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a nonnegatively symmetric density function. In addition, it is Lipschitz continuous on $\mathbb{R}$ and $\int_{\mathbb{R}} t^2K^2(t)dt < \infty$.

C4. Density Function: The density function of η, denoted by $g_\eta(\cdot) : \mathbb{R} \to \mathbb{R}$, is bounded away from zero and infinity and continuous at η for all η. Also, $g_\eta(\cdot)$ is assumed to have the fourth continuous derivative and the global unique mode zero, i.e., $g_\eta(\cdot) < g_\eta(0)$ for all $\eta \neq 0$.

C5. Moment: The covariates X_t are strictly stationary and ergodic with $\mathbb{E}\|X\|^{2s} < \infty$ for some $s > 2$. The matrices J_{β_0}, J_β, and J_θ defined in the following theorems are positive definite.

Although a little bit lengthy, these assumptions are actually quite mild; see Kemp and Santos Silva (2012), Yao and Li (2014), and Ullah, Wang and Yao (2021, 2022, 2023). C1 is common and can be easily satisfied in practice. When the estimators take values in the parameter space that is bounded and closed, calculating modal estimators is more useful since all mean estimators are biased at extreme boundary points. Under the mixing condition imposed in C2, the dependence among $\{(X_t^T, \varepsilon_t)^T\}$ will diminish as the distance between indices increases and is thus asymptotically ignorable. Similar to Ullah, Wang and Yao (2021), we can show that the α-mixing condition will make estimator behave in the same way as the independence case, which is typical in nonparametric problems. We do not impose a bounded support condition for the kernel function $K(\cdot)$ in C3. As argued by a large number of research (Ullah, Wang and Yao, 2023), it is not indispensable for the kernel function to have a bounded support as long as its tails are thin. For example, the Gaussian kernel, which is the default kernel utilized in this paper, is allowed. C4 is employed to ensure the existence of the global unique mode, which is the same as that in Kemp and Santos Silva (2012) and Ullah, Wang and Yao (2021, 2022, 2023). Such a condition can be released to capture the multimodal estimators by using different initial estimates in the MEM algorithm. C5 is necessary when deriving asymptotic distributions for estimators. All conditions related to bandwidths are illustrated in the following theorems.

We primarily show the asymptotic properties of the initial estimator $\tilde{\theta}$, while the results for $\tilde{\beta}$ can be obtained accordingly (i.e., $\|\tilde{\beta} - \beta_0\| = O_p\{(n_0h_1^3)^{-1/2} + h_1^2\}$). In what follows, we let $g_\eta^{(c)}(\cdot)$ denote the cth derivative of $g_\eta(\cdot)$ with $\|g_\eta^{(c)}(\cdot)\|_\infty$ bounded from above.

Theorem 1. *Under the regularity Conditions* C1.–C5. *and the restriction* $\|\tilde{\beta} - \beta_0\|/h_1^2 \to 0$, *with probability approaching one, as* $n_0 \to \infty$, $h_1 \to 0$, *and* $n_0 h_1^5 \to \infty$, *there exists a consistent maximizer* $\tilde{\theta}$ *of* (2.5) *such that* $\|\tilde{\theta} - \theta_0\| = O_p\{(n_0 h_1^3)^{-1/2} + h_1^2\}$.

Theorem 2. *With* $n_0 h_1^7 = O(1)$, *under the same conditions as Theorem* 1, *the estimator satisfying the consistency result in Theorem* 1 *has the following asymptotic result*

$$\sqrt{n_0 h_1^3}\left(\tilde{\theta} - \theta_0 - \frac{h_1^2}{2}\frac{g_\eta^{(3)}(0)}{g_\eta^{(2)}(0)} J_{\beta_0}^{-1} M_{\beta_0}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_\beta^{-1}\right).$$

Furthermore, under the assumption that $n_0 h_1^7 \to 0$, *we have*

$$\sqrt{n h_2^3}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_\beta^{-1}\right),$$

where $\xrightarrow{d}$ *denotes convergence in distribution,* $J_{\beta_0} = \mathbb{E}(X_{\beta_0} X_{\beta_0}^T)$, $M_{\beta_0} = \mathbb{E}(X_{\beta_0})$, $X_{\beta_0}^T = (X_{\beta_0,d+1}, \ldots, X_{\beta_0,n})^T$, *and* $X_{\beta_0,t} = X_t^T - \sum_{j=1}^d \beta_{0j} X_{t-j}^T$.

Theorem 2 shows that the asymptotic properties of the first-step estimators are the same as those of the Yule-Walker estimators based on modal estimation for the AR model, implying that $\tilde{\theta}$ is as efficient as if the true regression parameter β were known in advance. In contrast to mean estimation, the modal estimator has an asymptotic bias term associated with bandwidth h_1 as a result of mode estimation and the use of local data. To control the bias in estimation and satisfy the condition $\|\tilde{\beta} - \beta_0\|/h_1^2 \to 0$, the norm of the estimator of β should be of a smaller order than h_1^2, which can be achieved through undersmoothing.

Remark 4 (Optimal Bandwidth). The asymptotic bias of $\tilde{\theta}$, according to Theorem 2, is $2^{-1}h_1^2 g_\eta^{(3)}(0)\{g_\eta^{(2)}(0)\}^{-1} J_{\beta_0}^{-1} M_{\beta_0}$, whereas the asymptotic variance is $g_\eta(0) g_\eta^{(2)}(0)^{-2}\int t^2 K^2(t)dt J_\beta^{-1}$. The asymptotically optimal bandwidth h_1 can be obtained by minimizing the asymptotic weighted mean squared errors, i.e., $\mathbb{E}\{(\tilde{\theta} - \theta_0)^T W_{\beta_0}(\tilde{\theta} - \theta_0)\} \approx g_\eta^{(3)}(0)^2\{g_\eta^{(2)}(0)\}^{-2} M_{\beta_0}^T J_{\beta_0}^{-1} W_{\beta_0} J_{\beta_0}^{-1} M_{\beta_0}^T h_1^4/4 + (n_0 h_1^3)^{-1} tr(J_\beta^{-1}) g_\eta(0) g_\eta^{(2)}(0)^{-2}\int t^2 K^2(t)dt$, where $tr(\cdot)$ denotes the trace and W_β represents a weight matrix. Accordingly, the asymptotically optimal bandwidth is

$$\hat{h}_1 = \left[\frac{3tr(J_\beta^{-1}) g_\eta(0) g_\eta^{(2)}(0)^{-2}\int t^2 K^2(t)dt}{g_\eta^{(3)}(0)^2\{g_\eta^{(2)}(0)\}^{-2} M_{\beta_0}^T J_{\beta_0}^{-1} W_{\beta_0} J_{\beta_0}^{-1} M_{\beta_0}^T}\right]^{1/7} n_0^{-1/7}.$$

If we let $W_{\beta_0} = J_{\beta_0}$, which is proportional to the inverse of the asymptotic variance of $\tilde{\theta}$, we can have

$$\hat{h}_1 = \left[\frac{3dg_\eta(0)g_\eta^{(2)}(0)^{-2} \int t^2 K^2(t)dt}{g_\eta^{(3)}(0)^2\{g_\eta^{(2)}(0)\}^{-2} M_{\beta_0}^T J_{\beta_0}^{-1} M_{\beta_0}^T} \right]^{1/7} n_0^{-1/7}.$$

As a result, the asymptotically optimal bandwidth value in modal regression is larger than that in nonparametric mean regression with order $n_0^{-1/5}$.

To investigate the asymptotic properties of the shrinkage modal estimator, we decompose the AR regression coefficient vector β_0 into $\beta_0 = (\beta_{0'}^T, \beta_{0''}^T)^T \in \mathbb{R}^d$ without loss of generality, where $\beta_{0'} = (\beta_{01}, \ldots, \beta_{0s})^T \in \mathbb{R}^s$ consists of all nonzero components of β_0 and $\beta_{0''} = (\beta_{0s+1}, \ldots, \beta_{0d})^T \in \mathbb{R}^{d-s}$ includes all zero components of β_0. Define

$$a_n = \max_{1 \le j \le d} \left\{ |p_{\lambda_j}^{(1)}(|\beta_{0j}|)| : \beta_{0j} \neq 0 \right\}, \; b_n = \max_{1 \le j \le d} \left\{ |p_{\lambda_j}^{(2)}(|\beta_{0j}|)| : \beta_{0j} \neq 0 \right\},$$
$$\Psi_\lambda = \left(p_{\lambda_1}^{(1)}(|\beta_{01}|), \ldots, p_{\lambda_s}^{(1)}(|\beta_{0s}|) \right)^T, \; \Phi_\lambda = diag\left\{ p_{\lambda_1}^{(2)}(|\beta_{01}|), \ldots, p_{\lambda_s}^{(2)}(|\beta_{0s}|) \right\},$$

where $p_{\lambda_j}^{(2)}(.)$ indicates the second derivative of penalty. We can establish the following theoretical properties about the consistency and sparsity of the penalized modal estimator of the AR model.

Theorem 3 (Consistency). *Under the conditions in Theorem* 2*, with probability approaching one, as* $b_n \to 0$ *with* $n_0 \to \infty$*, there exists a consistent maximizer* $\hat{\beta}^P$ *of* (2.6) *such that* $\|\hat{\beta}^P - \beta_0\| = O_p\{(n_0 h_2^3)^{-1/2} + h_2^2 + a_n\}$.

Theorem 4 (Sparsity). *Under the same conditions in Theorem* 3*, let* $\delta_n = h_2^2 + (n_0 h_2^3)^{-1/2}$ *and* $\lambda_{min} = \min_j\{\lambda_j\}$*, if* $\lambda_{max} = \max_j\{\lambda_j\} \to 0$*,* $\delta_n^{-1}\lambda_{min} \to \infty$ *when* $n \to \infty$*, and* $\liminf_{n\to 0} \liminf_{\beta_j \to 0^+} p_{\lambda_j}^{(1)}(|\beta_j|)/\lambda_j > 0$ *for all* j*, then the penalized modal estimator can correctly identify all zero elements; that is* $P(\hat{\beta}_{0''}^P = 0) \to 1$.

Theorem 3 demonstrates the existence of the penalized modal estimator $\hat{\beta}^P$ that converges to the true parameter at the rate $O_p\{(n_0 h_2^3)^{-1/2} + h_2^2 + a_n\}$. In other words, by choosing an approximate regularization parameter λ_j, there exists a $\sqrt{n_0 h_2^3}$-consistent penalized modal estimator. It also indicates that the difference between the SCAD penalty estimate and the true parameter is asymptotically negligible when λ_j is small enough such that $a_n = O_p(h_2^2)$. Theorem 4 states that the proposed penalized modal regression is consistent in order selection; that is, by selecting an appropriate regularization parameter λ_j, the penalized modal estimation procedure estimates a zero coefficient exactly as zero with probability tending to one.

We establish the asymptotic distribution of the modal estimator for nonzero coefficients under suitable conditions in Theorem 5, which demonstrates that $\hat{\beta}_{0'}^P$ has oracle properties, i.e., performs as well as if we knew the submodel. In what follows, we define $e = (\varepsilon_{-1}, \ldots, \varepsilon_{-d})^T$ and $\varepsilon_{-j} = (\varepsilon_{d+1-j}, \ldots, \varepsilon_{n-j})$ for $j = 1, 2, \ldots, d$.

Theorem 5 (Asymptotic Normality). *With $n_0 h_2^7 = O(1)$ and $n_0 h_2^3 \Psi_\lambda^2 = O(1)$, under the same conditions in Theorem 4, the estimator satisfying the consistency result in Theorem 3 has the following asymptotic result*

$$\sqrt{n_0 h_2^3}(J_{(1)} + \Phi_\lambda)\left[\hat{\beta}_{0'}^P - \beta_{0'} + (J_{(1)} + \Phi_\lambda)^{-1}\left\{\Psi_\lambda - \frac{h_2^2}{2}\frac{g_\eta^{(3)}(0)}{g_\eta^{(2)}(0)} M_{(1)}\right\}\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_{(1)}^{-1}\right).$$

In addition, if $\sqrt{n_0 h_2^3}\Psi_\lambda = o_p(1)$ and $\Phi_\lambda = o_p(1)$, we can obtain

$$\sqrt{n_0 h_2^3} J_{(1)}\left\{\hat{\beta}_{0'}^P - \beta_{0'} - \frac{h_2^2}{2}\frac{g_\eta^{(3)}(0)}{g_\eta^{(2)}(0)} J_{(1)}^{-1} M_{(1)}\right\} \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_{(1)}^{-1}\right).$$

Furthermore, if $n_0 h_2^7 \to 0$, we have

$$\sqrt{n_0 h_2^3} J_{(1)}\left(\hat{\beta}_{0'}^P - \beta_{0'}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_{(1)}^{-1}\right),$$

where $J_{(1)}$ and $M_{(1)}$ are the $s \times s$ submatrices of $J_\beta = \mathbb{E}(ee^T)$ and $M_\beta = \mathbb{E}(e)$ corresponding to the nonzero components $\beta_{0'}$.

If the lags considered are not equally significant, the preceding asymptotic result instantly alleviates the constraint on the magnitude of the order d, because increasing d will no longer impose proportionally greater burden on the estimating efficiency. Consequently, the developed SCAD penalty procedure can be utilized to determine the complexity of the AR process. This result provides underlying support for choosing an arbitrary upper bound in the first step estimation.

The following asymptotic theorems are for the final modal estimator of θ based on the transformation of the dependent variable.

Theorem 6. *Under the conditions in Theorem 5 and the restriction $h_2/h_3 \to 0$, with probability approaching one, as $n_0 \to \infty$, $h_3 \to 0$, and $n_0 h_3^5 \to \infty$, there exists a consistent maximizer $\hat{\theta}$ of (2.7) such that $\|\hat{\theta} - \theta_0\| = O_p\{(n_0 h_3^3)^{-1/2} + h_3^2\}$.*

Theorem 7. *With $n_0 h_3^7 = O(1)$, under the same conditions as Theorem 6, the estimator satisfying the consistency result in Theorem 6 has the following asymptotic result*

$$\sqrt{n_0 h_3^3}\left\{\hat{\theta} - \theta_0 - \frac{h_3^2}{2}\frac{g_\eta^{(3)}(0)}{g_\eta^{(2)}(0)} J_\theta^{-1} M_\theta\right\} \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_\theta^{-1}\right).$$

Furthermore, under the assumption that $n_0 h_3^7 \to 0$, we have

$$\sqrt{n_0 h_3^3}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2 K^2(t)dt J_\theta^{-1}\right),$$

where $J_\theta = \mathbb{E}(XX^T)$, $M_\theta = \mathbb{E}(X)$, *and* $X^T = (X_{d+1}^T, \ldots, X_n^T)^T$.

By undersmoothing the previous estimator ($h_2/h_3 \to 0$), the bias term from preliminary estimation will be smaller than the leading bias term. As a consequence, we have a sort of oracle property, in which the modal estimator $\hat{\theta}$ is asymptotically equivalent to the estimator where the true values of $\{\varepsilon_{t-j}\}_{j=1}^d$ were known. Note that the asymptotic bias and variance are comparable to those from parametric modal regression with i.i.d. observations. The main difference is that under i.i.d. errors, $g(\cdot)$ is the density function of ε_t evaluated at 0, whereas in the current work, $g(\cdot)$ is the density of η_t. Compared to the initial estimator $\tilde{\theta}$, we do not need to account for uncertainty in lag terms, resulting in a potential increase in efficiency. Following the same procedure as in Remark 4, we can show the asymptotically optimal bandwidth $h_3 = O(n_0^{-1/7})$. With undersmoothing $\lim_{n_0\to\infty} n_0 h_3^7 = 0$, the estimator can be asymptotically centered at the true value.

Remark 5. The limiting distributions given in the preceding theorems cannot be used directly for inference or constructing confidence intervals because of the presence of many unknown terms. Although we can apply nonparametric estimation to achieve the corresponding density estimates, we have to introduce additional tuning parameters. The alternative method we can utilize is bootstrap resampling on the basis of mode, facilitating statistical inference about the parameter of interest; see Ullah, Wang and Yao (2021).

Remark 6 (Bandwidth Selection). The bandwidth in modal regression not only plays an essential role in the trade-off between reducing bias and variance, but also affects the target objective (either modal or mean estimate). In addition, when undersmoothing is used, choosing bandwidths is difficult because it does not allow for data-driven selection, and the traditional cross-validation method based on mean squared errors cannot be applied directly to modal regression. In such a case, we can work with the undersmoothing assumption on bandwidths following Ullah, Wang and Yao (2023) to apply the grid search method to select a number of potential bandwidths for h_1, h_2, and h_3. Specifically, we compute the mean regression residual first and then select 50 values of bandwidth between 50MAD and $0.5\text{MAD}n_0^{-\lambda_{h_j}}(\lambda_{h_1} = 0.16, \lambda_{h_2} = 0.15, \lambda_{h_3} = 0.143)$, where MAD is the median value of the absolute deviation of the mean regression residual from the corresponding median value. Based on simulation experience, the selected bandwidths are appropriate for the model developed in this paper. For empirical analysis, we simply set the bandwidth to $1.6\text{MAD}n_0^{-\lambda_{h_j}}$.

4. Modal Autocorrelation Test

When applying modal regression on data with serial correlation, it is particularly important to check for any signs of autocorrelation in order to make valid inferences and estimate more efficiently. However, no formal investigation into this important issue has been conducted thus far. By extending the results in Huo et al. (2017), we propose a residual-based test for autocorrelation in modal regression models. Notice that the error terms $\{\varepsilon_t\}_{t=1}^n$ are allowed to exhibit autocorrelation unless $\beta_j = 0$ for all $j = 1, \ldots, d$. Therefore, the primary null hypothesis that we wish to test for is given by

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0, \tag{4.1}$$

whereas the alternative hypothesis is $H_1 : \beta_j \neq 0$ for some $1 \leq j \leq d$. The developed autocorrelation test is based on the modal error $\varepsilon_t = Y_t - Mode\ (Y_t \mid X_t)$ in (1.2) and the assumption that the mode version of the orthogonality condition $\mathbb{E}\{Q_n(Y_t, X_t) \mid X_t\} = 0$ almost surely for every t, where $Q_n(Y_t, X_t)$ is the corresponding kernel-based objective function. With the available modal residuals $\{\hat{\varepsilon}_t\}_{t=1}^n$ from (1.3), the auxiliary regression of $\hat{\varepsilon}_t$ on X_t and $\{\varepsilon_{t-l}\}_{l=1}^d$ by linear mean regression can be carried out

$$\hat{\varepsilon}_t = X_t^T\gamma + \beta_1\hat{\varepsilon}_{t-1} + \cdots + \beta_d\hat{\varepsilon}_{t-d} + v_t, \tag{4.2}$$

in which $\gamma \in \mathbb{R}^p$ is the parameter and v_t is the error term in the auxiliary regression. We then have the following test statistic

$$\text{Mode}_T = \frac{\sum_{t=d+1}^{\text{n}} \tilde{v}_t^2 - \sum_{t=d+1}^{\text{n}} \hat{v}_t^2}{\sum_{t=d+1}^{\text{n}} \hat{v}_t^2/(n-d-p)}, \tag{4.3}$$

where $\hat{v}_t$ is the residual from the unrestricted auxiliary regression and $\tilde{v}_t$ is the residual from the restricted auxiliary regression with the null hypothesis imposed. The suggested test can be treated as the usual F test for the null hypothesis. When H_0 is true, we can show that the proposed statistic is asymptotically distributed as the $\mathcal{X}^2$ distribution with d degrees of freedom ($\mathcal{X}_d^2$). The order d is unimportant when testing the null hypothesis since the suggested test does not have size distortions (see simulation results in the supplementary file). Once the null hypothesis is rejected, we can utilize the developed penalty methodology to select an appropriate model

Theorem 8. *Under the conditions in Theorem* 2*, with the restriction that* ε_t *is homoskedastic with a constant variance, we have* $\text{Mode}_T \xrightarrow{d} \mathcal{X}_d^2$ *as* $n \to \infty$*, provided that the null hypothesis is correct.*

If the errors are heteroskedastic, we can use the typical robust variance-covariance estimator and achieve the same result. To prove the preceding

theorem, we rewrite the $Mode_T$ statistic in a form of the Wald statistic

$$Mode_T = \frac{(\sqrt{n-d}(0_{d,p}, I_d)\hat{\gamma}_m)((0_{d,p}, I_d)M^{-1}(0_{d,p}, I_d))^{-1}(\sqrt{n-d}(0_{d,p}, I_d)\hat{\gamma}_m)}{s^2}, \tag{4.4}$$

where I_d is the $d \times d$ identity matrix, $M = (n-d)^{-1}\sum_{t=d+1}^{n} Z_t Z_t^T$, $Z_t = (X_t^T, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-d})^T$, $\gamma_m = (\gamma^T, \beta_1, \ldots, \beta_d)^T$, and $s^2 = \sum_{t=d+1}^{n} \hat{v}_t^2/(n-d-p)$. Combined with the result that the estimator $\hat{\gamma}_m$ is asymptotically normally distributed, we can straightforwardly show that $Mode_T$ follows a $\mathcal{X}^2$ distribution. If $Mode_T > \mathcal{X}^2_{d,\alpha}$, the null hypothesis H_0 is rejected at significance level α, where $\mathcal{X}^2_{d,\alpha}$ is the $100(1-\alpha)\%$ quantile of the $\mathcal{X}^2_d$ distribution.

Remark 7 (Bootstrap Implementation). Although the asymptotic level of $Mode_T$ is available, it may not perform well in practice if the sample size is insufficient. The parametric wild bootstrap approach built on mean can then be used to evaluate the p-value. First, we generate the wild bootstrap residuals $\{\tilde{v}_t^*\}_{t=1}^{n-d}$ from the mean-centered parametric residuals $\{\hat{v}_t^*\}_{t=1}^{n-d}$, where $\hat{v}_t^* = \hat{v}_t - Mean(\hat{v}_t)$, and define the bootstrap sample $\hat{\varepsilon}_t^* = X_t^T\hat{\gamma} + \sum_{j=1}^{d}\hat{\beta}_{m,j}\hat{\varepsilon}_{t-j} + \tilde{v}_t^*$, in which $\hat{\gamma}$ and $\hat{\beta}_{m,j}$ are the corresponding mean estimates from (4.2). Based on the bootstrap sample $\{\hat{\varepsilon}_t^*, X_t\}$, we can calculate the bootstrap test statistic $Mode_T^*$ and reject the null hypothesis H_0 when $Mode_T$ is greater than the upper α point of the conditional distribution of $Mode_T^*$ given $\{\hat{\varepsilon}_t, X_t\}$. The p-value of the test is then evaluated using the relative frequency of the event $\{Mode_T^* \geq Mode_T\}$ in the replications of the bootstrap sampling.

5. Real Data Application

We now illustrate the proposed method through an application to analyze spirit consumption data in the United Kingdom from 1870 to 1938, which can be found in Fuller (1996). The dataset contains 69 daily observations of the annual per capita consumption of spirits (Y_t), per capital income ($X_{1,t}$), and the price of spirits ($X_{2,t}$). In this illustration, we fit the data with the following parametric modal regression model

$$Y_t = \text{constant} + \theta_1 X_{1,t} + \theta_2 X_{2,t} + \theta_3 X_{3,t} + \theta_4 X_{4,t} + \varepsilon_t, \tag{5.1}$$

where variables $X_{3,t} = t/100$ and $X_{4,t} = (t-35)^2/10000$ for $t = 1, \ldots, 69$, and the error terms $\{\varepsilon_t\}_{t=1}^{69}$ are assumed to be a stationary process.

We first use the mean regression to obtain the initial estimate of θ by ignoring the AR structure, yielding the following equation

$$\mathbb{E}(Y_t \mid X_t) = \underset{(0.2707)}{2.1209} + \underset{(0.1323)}{0.6975}\, X_{1,t} - \underset{(0.0529)}{0.6322}\, X_{2,t} - \underset{(0.0837)}{0.9555}\, X_{3,t} - \underset{(0.1539)}{1.1525}\, X_{4,t}, \tag{5.2}$$

where the numbers in brackets represent standard errors. We can then calculate

the estimated mean residuals. The autocorrelation plot of the residuals in Figure 1 clearly shows that the independence assumption for residuals is questionable and there exists a periodic structure in residuals. The partial-autocorrelation plot suggests that an AR(d) with $d \leq 10$ may fit the errors well. The blue lines in Figure 1 indicate the confidence intervals.

The autocorrelation check presented above is based on the mean. To further demonstrate the existence of autocorrelation, we run the parametric modal regression and plot the corresponding autocorrelation and partial-autocorrelation functions in Figure 1. The results follow a pattern similar to mean regression. Nevertheless, the modal estimates are different from the mean estimates (although not much), which is expected given that the distribution of Y is not symmetric. The standard errors for modal coefficients are calculated through the bootstrap procedure, which are generally smaller than those in mean regression (Figure 2).

$$Mode(Y_t \mid X_t) = \underset{(0.2094)}{2.4735} + \underset{(0.0898)}{0.5873}\, X_{1,t} - \underset{(0.0671)}{0.7177}\, X_{2,t} - \underset{(0.0804)}{0.7939}\, X_{3,t} - \underset{(0.1235)}{1.3232}\, X_{4,t}. \tag{5.3}$$

Since the order d is unimportant when testing the null hypothesis (see Section 4 and simulation results in the supplementary file), we apply the proposed test to validate the autocorrelation structure with the AR(1) error process. The relative frequency of the event $\{Mode_T^* \geq Mode_T\}$ we obtain is 0.025, which strongly suggests that the null hypothesis of no autocorrelation should be rejected. According to the partial-autocorrelation plot in Figure 1, we then assume an AR(10) model on errors and apply the penalized modal regression with SCAD penalty to select order and estimate modal coefficients. The estimation results are shown as follows (Figure 2)

$$\begin{aligned} Mode(Y_t \mid F_{t-1}) = {} & \underset{(0.0384)}{1.9592} + \underset{(0.0182)}{0.8302}\, X_{1,t} - \underset{(0.0096)}{0.6792}\, X_{2,t} - \underset{(0.0127)}{0.9409}\, X_{3,t} \\ & - \underset{(0.0350)}{1.2215}\, X_{4,t} + \underset{(0.0243)}{0.6558}\, \varepsilon_{t-1} - \underset{(0.0257)}{0.2379}\, \varepsilon_{t-10}, \end{aligned} \tag{5.4}$$

where standard errors are calculated using bootstrap procedure.

After inspection, we confirm that the estimates satisfy the stationarity condition. In comparison to the traditional modal regression results in (5.3), the "most likely" effect of per capital income on annual per capital consumption of spirits is larger, while the effect of price of spirits is smaller, demonstrating that ignoring the AR error structure may result in not only inefficient but also inconsistent estimators (heteroskedasticity). Furthermore, after taking the information in the error structure into consideration, the modal estimators become more efficient. We plot the autocorrelation and partial autocorrelation functions in Figure 1 for (5.4). The new residuals do not have any significant pattern and appear to be a white process.

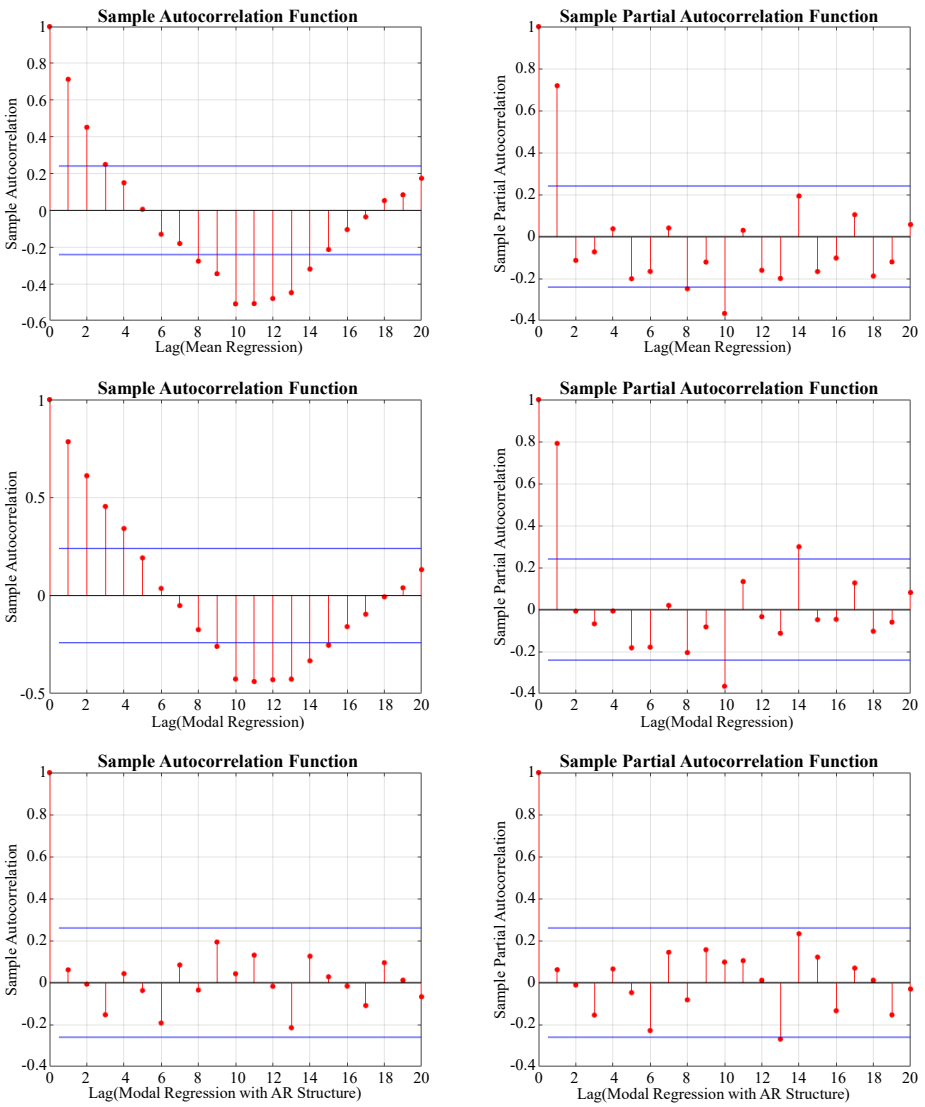

Figure 1. Correlogram of Residuals.

For comparison, we further report the results of mean estimation when autocorrelation information is taken into account. By applying the penalized mean regression with SCAD penalty, we obtain the following result

$$\begin{aligned}\mathbb{E}(Y_t \mid F_{t-1}) = &\underset{(0.2551)}{2.0546} + \underset{(0.1224)}{0.6378}\, X_{1,t} - \underset{(0.0516)}{0.6880}\, X_{2,t} - \underset{(0.0753)}{0.9523}\, X_{3,t} \\ &- \underset{(0.1449)}{0.9826}\, X_{4,t} + \underset{(0.1218)}{0.4795}\, \varepsilon_{t-1} - \underset{(0.1177)}{0.2568}\, \varepsilon_{t-8}.\end{aligned} \tag{5.5}$$

It is interesting to observe that the mean estimation results differ from the modal estimation results. Especially, mean regression selects the AR model with lags

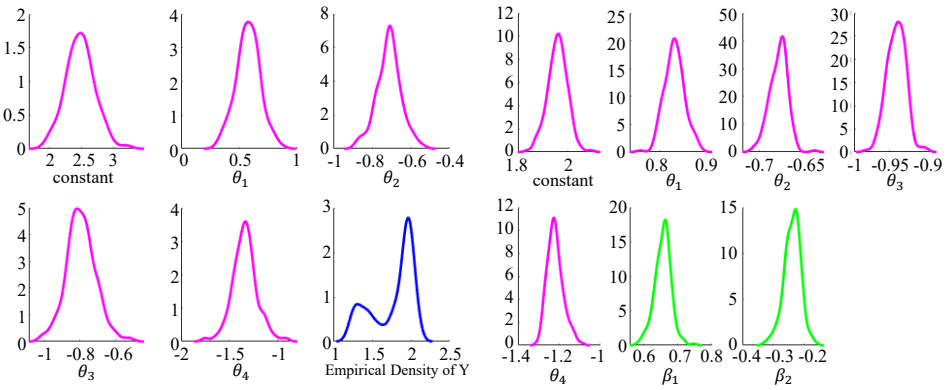

Figure 2. Bootstrap Results and Empirical Density.

1 and 8, and produces estimates with larger standard errors. In addition, the magnitudes of mean coefficients of $X_{1,t}$ and $X_{4,t}$ are smaller than those obtained from modal estimation. All of these suggest that modal estimation can provide some additional data information that mean estimation may ignore. Moreover, to compare the prediction ability, we utilize both mean and modal regressions with AR errors to predict the last five data points (out-of-sample prediction). The mean absolute prediction errors we obtain are 0.2853 (mean) and 0.1926 (mode), respectively. Therefore, modal regression also has better prediction performance, which is consistent with the simulation results in the supplementary file.

6. Concluding Remarks

As one of the center measures, the mode preserves some important features of the underlying distribution function and provides a reliable estimate of location. Built on mode value, we in this paper propose an efficient estimation procedure for parametric linear modal regression with AR errors by applying the kernel-based objective functions. We utilize a penalized objective function to select the order of the AR process and construct a computationally simple residual-based test for detecting autocorrelation in modal regression models. We investigate the asymptotic properties of the resultant modal estimators under some mild conditions. Two modal algorithms are introduced to arithmetically estimate models. The numerical results show that the developed method is superior to parametric modal regression without considering AR error structure and can effectively improve estimation and prediction accuracy in moderate-sized samples compared to mean regression. We also discuss the extension of the estimation procedure to nonparametrically established modal regression models.

We in this paper concentrate on the strictly stationary case. In practice, this assumption may be difficult to justify since time series are frequently observed with trends. We can combine the proposed estimation procedure with the

technique that removes the deterministic trend, or we can consider a locally stationary time series model. In addition, the dimension of covariates in this paper is fixed. It would be appealing to extend the results to high dimensional case, where the dimension of covariates depends on sample size, i.e., $d = O(n^{\alpha})$, $\alpha > 1$. Nevertheless, with growing d, sparseness generally refers to the proportion of zero parameters, and the initial modal estimator is not consistent. Also, with the $d > n$ setting, it is necessary to choose $\lambda > \log(n_0 h^3)$ to obtain model selection consistency with BIC-type criterion. We leave all of these interesting research for the future.

Supplementary Material

The online supplementary file contains all simulation results and technical proofs, the extension to nonparametric modal regression with autocorrelated error process, and the convergence of the penalized MEM algorithm.

Acknowledgments

The author is deeply grateful to the Co-Editor Rong Chen, Associate Editor, and two anonymous referees for their constructive comments, leading to the substantial improvement of the article. He would also like to thank Yanqin Fan, Michael Jansson, Aman Ullah, Weixin Yao, and the participants at the 2022 Canadian Economics Association Conference for their helpful comments.

References

Chen, Y. C. (2018). *Modal Regression Using Kernel Density Estimation: A Review.* Wiley Interdisciplinary Reviewers: Computational Statistics.

Chen, Y. C., Genovese, C. R., Tibshirani, R. J. and Wasserman, L. (2016). Nonparametric modal regression. *Annals of Statistics* **44**, 489–514.

Chen, Z., Li, R. and Li, Y. (2015). Varying coefficient models for data with auto-correlated error process. *Statistica Sinica* **25**, 709–723.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle Properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710–723.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Fuller, A. (1996). *Introduction to Statistical Time Series.* 2nd Edition. Wiley, New York.

Furno, M. (2000). LM tests in the presence of non-normal error distributions. *Econometric Theory* **16**, 249–261.

Huo, L., Kim, T., Kim, Y. and Lee, D. J. (2017). A residual-based test for autocorrelation in quantile regression models. *Journal of Statistical Computation and Simulation* **87**, 1305–1322.

Kemp, G. C. R. and Santos Silva, J. M. C. (2012). Regression towards the mode. *Journal of Econometrics* **170**, 92–101.

Li, J., Ray, S. and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* **8**, 1687–1723.

Martins-Filho, C. and Yao, F. (2009). Nonparametric regression estimation with general parametric error covariance. *Journal of Multivariate Analysis* **100**, 309–333.

Opsomer, J., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

Qiu, J., Li, D. and You, J. (2015). SCAD-penalized regression for varying-coefficient models with autoregressive errors. *Journal of Multivariate Analysis* **137**, 100–118.

Su, L. and Ullah, A. (2007). More efficient estimation of nonparametric panel data models with random effects. *Economics Letter* **96**, 375–380.

Ullah, A., Wang, T. and Yao, W. (2021). Modal regression for fixed effects panel data. *Empirical Economics* **60**, 261–308.

Ullah, A., Wang, T. and Yao, W. (2022). Nonlinear modal regression for dependent data with application for predicting COVID-19. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **185**, 1424–1453.

Ullah, A., Wang, T. and Yao, W. (2023). Semiparametric partially linear varying coefficient modal regression. *Journal of Econometrics* **235**, 1001–1026.

Wang, H., Li, G. and Tsai, C. L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69**, 63–78.

Xiao, Z., Linton, O. B., Carroll, R. J. and Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association* **98**, 980–992.

Yao, W. (2013). A note on EM algorithm for mixture models. *Statistics and Probability Letters* **83**, 519–526.

Yao, W. and Li, L. (2014). A new regression model: Modal linear regression. *Scandinavian Journal of Statistics* **41**, 656–671.

(Received November 2021; accepted May 2023)