Parametric Modal Regression with Autocorrelated Error Process

Tao Wang

University of Victoria

Supplementary Materials

S1 Monte Carlo Experiments

We investigate the finite sample performance of the proposed estimation procedure and the developed test using Monte Carlo simulations, where we compare the suggested estimation procedure to the traditional parametric modal regression without considering the error structure and the corresponding mean estimation. The bandwidth for the original parametric modal regression is chosen by the plug in method with the help of the asymptotically optimal bandwidth expression in Remark 4; see Ullah et al. (2021).

We simulate 200 datasets from each data generating process (DGP) with sample sizes of $n = 200, 400, 600,$ or 1000. In order to evaluate the finite sample performance, we calculate the mean, the average of the estimated standard errors (SE) as well as the mean squared errors (MSE) for each estimator and average them across 200 simulations. We also compare

SE to the sample standard deviation of the estimators (SD).

DGP 1 (fixed order) In this simulated example, we illustrate the finite sample performance of the proposed estimator with a known AR order. The data are generated from

$$Y_t = \theta_1 X_{1,t} + \theta_2 X_{2,t} + \varepsilon_t, \tag{S1.1}$$

where $\theta_1 = 1$, $\theta_2 = 2$, and X_t follows a multivariate normal distribution with two components that are independently standard normal distributed. The error term ε_t adheres to an AR process of order $d = 1$ or $d = 2$

$$\varepsilon_t = \sum_{j=1}^{d} \beta_j \varepsilon_{t-j} + \eta_t, \tag{S1.2}$$

where η_t is produced from: (1) standard normal distribution ($N(0,1)$), (2) t distribution with degrees of freedom 3 ($t(3)$), (3) mixture Laplace distribution ($0.8L_p(0,1)+0.2L_p(0,4)$), and (4) mixture normal distribution ($0.9N(0,1)+0.1N(0,5)$), respectively. We focus on the case in which $\beta_1 = 0.8$ and $\beta_2 = -0.5$ to guarantee the stationarity assumption for the model of the error terms. Since the order of the AR process is assumed to be known in this example, we directly apply the kernel-based objection function (2.8) for estimating. We set the initial values of parameters in simulations based on mean and median estimates and find that the final estimation results do not differ significantly. For simplicity, the initial values for the coefficients are thus set to be mean estimates for all cases.

Table S1: Results of Simulations—DGP 1

Sample Size	Traditional				Proposed			
	θ_1 (SE)	MSE	θ_2 (SE)	MSE	θ_1 (SE)	MSE	θ_2 (SE)	MSE
$\eta_t \sim N(0,1)$				$d=1$				
n=200	1.2490 (0.4423)	0.2169	1.9577 (0.4649)	0.2169	0.9888 (0.2036)	0.0414	2.0293 (0.3595)	0.1294
n=400	1.0689 (0.3360)	0.1199	2.0345 (0.3812)	0.1458	1.0005 (0.1594)	0.0253	1.9796 (0.2601)	0.0677
n=600	1.0687 (0.3032)	0.0962	1.9853 (0.3395)	0.1149	0.9800 (0.1287)	0.0169	2.0162 (0.2023)	0.0410
n=1000	1.0364 (0.2747)	0.0764	2.0081 (0.3246)	0.1049	0.9918 (0.1043)	0.0109	2.0089 (0.1889)	0.0356
				$d=2$				
n=200	0.9865 (0.2517)	0.0632	1.9877 (0.4013)	0.1604	0.9794 (0.1752)	0.0310	1.9864 (0.3521)	0.1236
n=400	0.9816 (0.2226)	0.0497	1.9873 (0.3394)	0.1148	0.9950 (0.1368)	0.0186	2.0100 (0.2710)	0.0732
n=600	0.9955 (0.2009)	0.0402	2.0084 (0.3146)	0.0985	0.9970 (0.1225)	0.0149	1.9907 (0.2328)	0.0540
n=1000	1.0104 (0.1951)	0.0380	1.9927 (0.2940)	0.0861	0.9933 (0.0857)	0.0073	2.0117 (0.1855)	0.0344
$\eta_t \sim t(3)$				$d=1$				
n=200	1.1694 (0.6830)	0.4928	2.0037 (0.5183)	0.2673	0.9809 (0.2497)	0.0624	1.9876 (0.4757)	0.2253
n=400	1.1013 (0.4540)	0.2154	1.9848 (0.4183)	0.1743	0.9972 (0.2035)	0.0412	2.0149 (0.3259)	0.1059
n=600	1.0140 (0.4287)	0.1831	2.0337 (0.4054)	0.1646	0.9958 (0.1582)	0.0249	1.9794 (0.2957)	0.0874
n=1000	1.0808 (0.3451)	0.1250	2.0364 (0.3744)	0.1408	0.9922 (0.1322)	0.0175	1.9943 (0.2168)	0.0468
				$d=2$				
n=200	0.9706 (0.2587)	0.0674	1.9452 (0.4536)	0.2077	1.0034 (0.2262)	0.0509	2.0037 (0.4043)	0.1627
n=400	0.9984 (0.2347)	0.0548	2.0072 (0.3833)	0.1462	1.0131 (0.1578)	0.0249	1.9983 (0.3163)	0.0996
n=600	1.0278 (0.2160)	0.0472	1.9546 (0.3713)	0.1393	0.9855 (0.1495)	0.0224	2.0036 (0.2746)	0.0750
n=1000	1.0066 (0.2076)	0.0429	1.9945 (0.3243)	0.1047	0.9934 (0.1201)	0.0144	1.9977 (0.2089)	0.0434
$\eta_t \sim 0.8L_p(0,1) + 0.2L_p(0,4)$				$d=1$				
n=200	1.2364 (0.6140)	0.4310	2.0027 (0.5595)	0.3115	1.0016 (0.2813)	0.0787	2.1007 (0.5263)	0.2857
n=400	1.0711 (0.4580)	0.2138	2.0046 (0.4734)	0.2230	0.9641 (0.2141)	0.0469	2.0496 (0.4043)	0.1651
n=600	1.0482 (0.4130)	0.1720	2.0191 (0.3953)	0.1559	0.9868 (0.2023)	0.0409	2.0145 (0.3368)	0.1131
n=1000	1.0010 (0.3487)	0.1210	2.0396 (0.3828)	0.1474	0.9917 (0.1532)	0.0234	2.0007 (0.2896)	0.0835
				$d=2$				
n=200	0.9652 (0.2773)	0.0777	1.9795 (0.4838)	0.2333	0.9983 (0.2405)	0.0575	1.9827 (0.4553)	0.2065
n=400	0.9877 (0.2626)	0.0688	1.9536 (0.4239)	0.1810	1.0026 (0.2097)	0.0437	1.9782 (0.3687)	0.1358
n=600	0.9861 (0.2486)	0.0617	2.0487 (0.3972)	0.1593	1.0137 (0.1836)	0.0337	2.0331 (0.2983)	0.0897
n=1000	1.0002 (0.2164)	0.0466	1.9999 (0.3492)	0.1213	1.0008 (0.1467)	0.0214	2.0109 (0.2714)	0.0734
$\eta_t \sim 0.9N(0,1) + 0.1N(0,5)$				$d=1$				
n=200	1.1776 (0.4697)	0.2510	1.9742 (0.4272)	0.1823	0.9677 (0.2168)	0.0478	1.9991 (0.3830)	0.1459
n=400	1.0787 (0.3205)	0.1084	1.9935 (0.3902)	0.1516	1.0228 (0.1514)	0.0233	2.0163 (0.2926)	0.0855
n=600	1.0476 (0.3105)	0.0982	2.0058 (0.3772)	0.1416	0.9831 (0.1310)	0.0174	2.0206 (0.2424)	0.0589
n=1000	1.0405 (0.2835)	0.0816	2.0132 (0.3189)	0.1014	0.9922 (0.1100)	0.0121	2.0134 (0.1721)	0.0297
				$d=2$				
n=200	1.0258 (0.2603)	0.0681	1.9578 (0.3759)	0.1424	0.9591 (0.1868)	0.0364	1.9847 (0.3472)	0.1202
n=400	0.9902 (0.2250)	0.0505	1.9653 (0.3599)	0.1301	0.9831 (0.1446)	0.0211	1.9904 (0.2521)	0.0633
n=600	1.0054 (0.2121)	0.0448	2.0312 (0.3419)	0.1173	1.0092 (0.1180)	0.0139	1.9721 (0.2276)	0.0516
n=1000	1.0003 (0.1850)	0.0341	1.9808 (0.2817)	0.0793	0.9978 (0.0978)	0.0095	1.9735 (0.1956)	0.0388

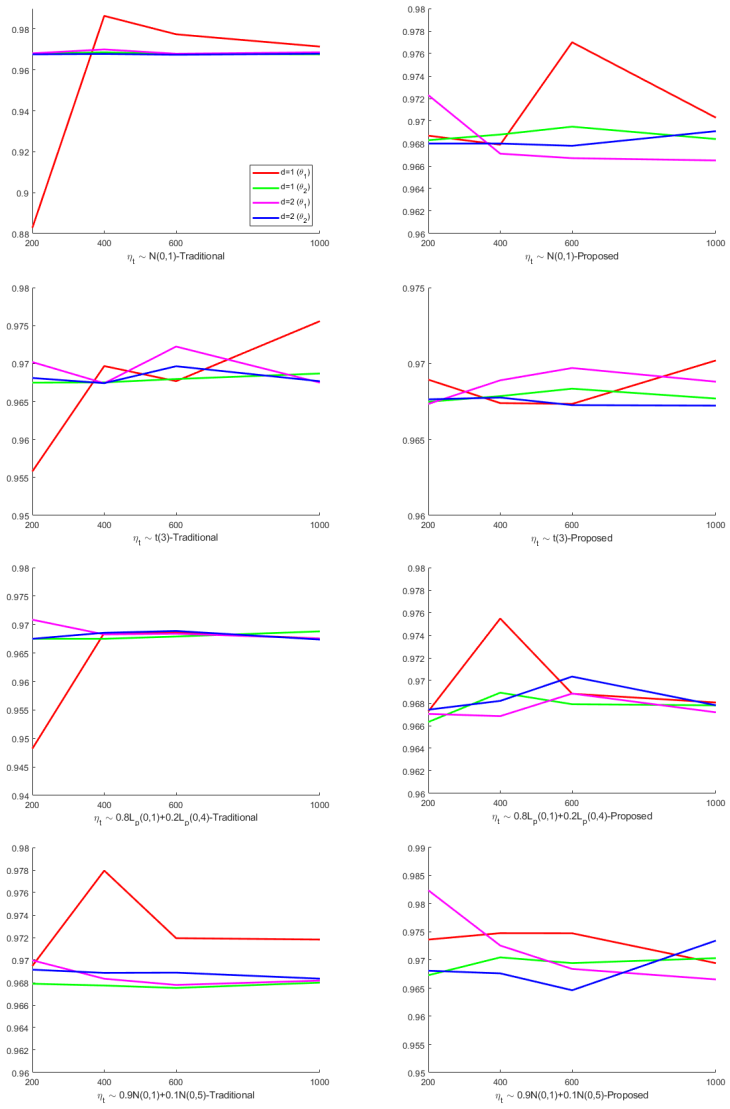

Figure S1: Ratio SE/SD of DGP 1

The simulation results are summarized in Table S1, showing that both traditional and proposed methods can reasonably estimate parameters (in terms of bias). This is expectable because the AR process has no effect on the values of the modal coefficients in this simulation case. For example, let

$\varepsilon_t = \beta_1 \varepsilon_{t-1} + \eta_t$. With ignoring the serial correlation, the modal regression line will be $Mode(Y_t \mid X_t) = X_t^T \theta + Mode(\varepsilon_t \mid X_t)$ based on the current information induced by the event at time t. Taking into account the serial correlation, we can have $Mode(Y_t \mid F_{t-1}) = X_t^T \theta + \beta_1 \varepsilon_{t-1}$ with the assumption that $Mode(\eta_t \mid F_{t-1}) = 0$ based on the information induced by all events before time t. Because mode does not have the additive property, it is difficult to guarantee that $Mode(\varepsilon_t \mid X_t)$ and $Mode(\beta_1 \varepsilon_{t-1} + \eta_t \mid X_t)$ will be the same. However, the value of θ is unaffected according to the model setting. Note that the estimates of unknown parameters become closer to the true values and the SE and MSE of each estimator in all settings decrease as the sample size n increases, which agrees with the asymptotic results in Section 3. Furthermore, the results demonstrate that the developed estimator is more efficient (with smaller SE and MSE) than the traditional modal regression estimator, especially for the AR(1) error process, where traditional modal regression estimators are slightly biased with small sample sizes. Finally, we compare SE to SD for the modal regression coefficients to check the accuracy of the estimated SE. We plot the ratio of SE over SD in Figure S1. The results indicate that the ratio is close to one and does not differ significantly between the compared two estimation methods. Thus, the proposed method performs reasonably well.

DGP 2 (order selection) In this simulation example, we present the finite sample performance of the developed estimation procedure with order selection. For each sampling scheme, the proposed estimation procedures without order selection and with SCAD penalty function are compared with regard to the efficiency improvement. In addition, we report the oracle result as a benchmark, where the true AR coefficients and order are known. The data are generated from the following parametric regression

$$Y_t = \theta_1 X_{1,t} + \theta_2 X_{2,t} + \varepsilon_t, \tag{S1.3}$$

where $\theta_1 = 2$, $\theta_2 = 3$, and X_t follows a multivariate normal distribution with two components being independently standard normal distributed. The error process ε_t is an AR process of order $d = 10$ or $d = 20$

$$\varepsilon_t = \sum_{j=1}^{d} \beta_j \varepsilon_{t-j} + \eta_t, \tag{S1.4}$$

where η_t is set to be the same as in DGP 1 (four different distributions). Our goal is to inspect whether the suggested estimation procedure with order selection can specify the model correctly and enhance estimation efficiency. We thus concentrate on the case where $\beta_1 = 0.5$ and all other β_j's are zero. According to the estimation procedure, we choose $d = 25$ as an upper bound for order in the first step estimation. For each simulation, an estimate whose absolute value is less than 10^{-4} is shrunk to 0, that is, the corresponding regression variables are removed.

Table S2: Results of Simulations—DGP 2

Sample Size	$d = 10$			$d = 20$		
	Without Selection	With Selection	Oracle	Without Selection	With Selection	Oracle
	$\eta_t \sim N(0,1)$					
n=200	0.2404	0.2037	0.1885	0.2723	0.2311	0.1993
n=400	0.1667	0.1414	0.1358	0.1542	0.1189	0.1071
n=600	0.1221	0.1049	0.1028	0.1004	0.0886	0.0788
n=1000	0.0686	0.0604	0.0594	0.0725	0.0625	0.0561
	$\eta_t \sim t(3)$					
n=200	0.2824	0.2129	0.2072	0.2907	0.2339	0.2084
n=400	0.1577	0.1327	0.1289	0.1800	0.1267	0.1138
n=600	0.1019	0.0799	0.0741	0.1509	0.1076	0.0977
n=1000	0.0676	0.0573	0.0563	0.0773	0.0616	0.0556
	$\eta_t \sim 0.8L_p(0,1) + 0.2L_p(0,4)$					
n=200	0.3696	0.3132	0.3069	0.3231	0.2838	0.2577
n=400	0.2211	0.1764	0.1692	0.2174	0.1862	0.1645
n=600	0.1445	0.1344	0.1227	0.1587	0.1334	0.1352
n=1000	0.1085	0.0977	0.0950	0.1129	0.0922	0.0883
	$\eta_t \sim 0.9N(0,1) + 0.1N(0,5)$					
n=200	0.2476	0.1921	0.1883	0.2569	0.2206	0.2058
n=400	0.1512	0.1253	0.1179	0.1571	0.1274	0.1165
n=600	0.1173	0.1018	0.1012	0.1266	0.1091	0.0947
n=1000	0.0778	0.0655	0.0610	0.0820	0.0749	0.0655

The results are summarized in Table S2, where the $MSEs$ for all parameters are reported. It can be observed that the finite sample performances of the estimation procedure with order selection are very close to those of the oracle cases, which is consistent with the theoretical results. Particularly, the proposed estimation procedure with order selection can specify AR order accurately for different error distributions and achieve a smaller MSE than the procedure without order selection. This achievement is more prominent for moderate sample size, such as $n = 200$. With a large sample size, the gain for the developed method in terms of MSE is not very big.

This is to be expected because all methods can estimate coefficients more precisely with the large sample size. Also, the difference between d=10 and d=20 is not remarkable. This implies that the proposed estimation procedure is not very sensitive to the assumption of the AR order provided that a variable selection for the AR error is conducted.

Remark S1. The developed two-step estimation procedure can be iterated to achieve better finite sample performance in practice. However, based on Monte Carlo simulation results, one iteration is enough to obtain the well finite sample performance.

DGP 3 (AR test) To demonstrate the good performance of the suggested test, we generate data from

$$Y_t = \theta_1 X_{1,t} + \theta_2 X_{2,t} + \varepsilon_t, \tag{S1.5}$$

where $\theta_1 = 1$, $\theta_2 = 3$, and X_t follows a multivariate normal distribution with two components being independently standard normal distributed. For the sake of simplicity, we assume that the error process ε_t is an AR process of order $d = 1$

$$\varepsilon_t = \beta_1 \varepsilon_{t-1} + \eta_t, \tag{S1.6}$$

where η_t follows the same distribution shown in DGP 1 (four different distributions) and $\beta_1 = \{0, 0.1, 0.2, \cdots, 0.9\}$. Thus, the limiting distribution of the proposed modal residual-based test under H_0 is the $\mathcal{X}^2$ distribution with one degree of freedom. When $\beta_1 = 0$ (size performance), the specified

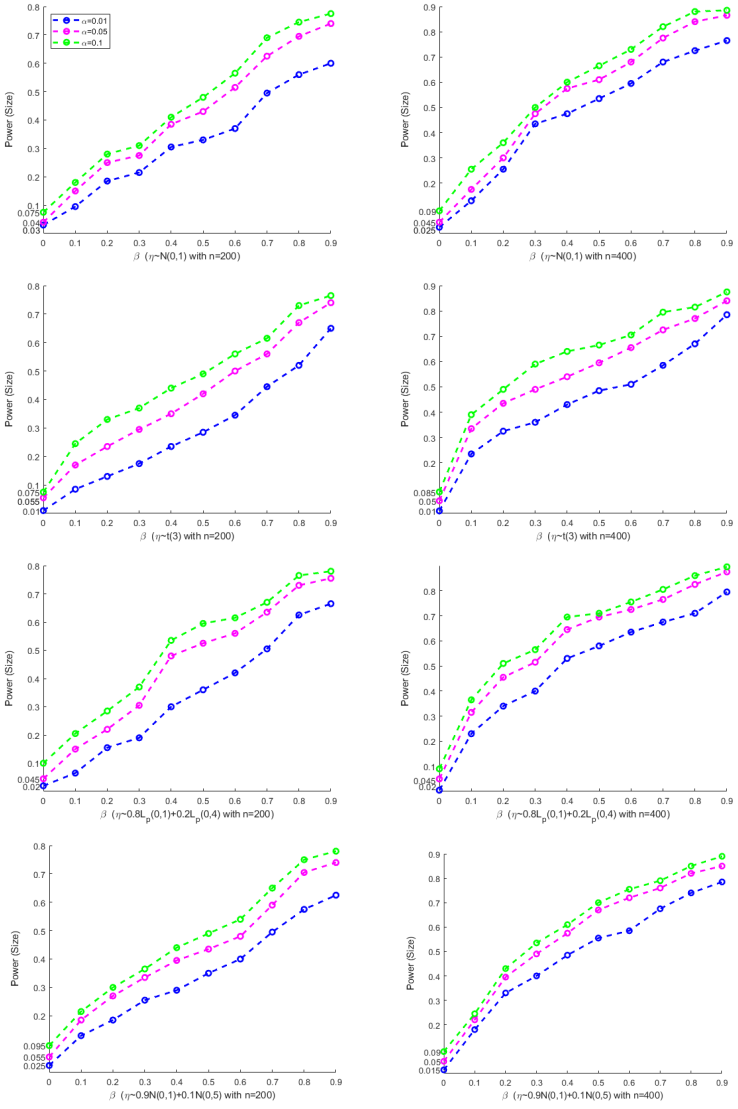

Figure S2: Power and Size of DGP 3

alternative hypothesis collapses into the null hypothesis, indicating the nonexistence of AR error structure; when β_1 is far from 0 (power performance), the existence of AR error structure is revealed. We adopt the bootstrap procedure illustrated in Remark 6 with 200 replications of the boot-

strap sampling and consider the significance levels α=0.1, 0.05 and 0.01 to investigate the size and power of the developed test. Due to the computational cost, we only report the results for n =200 or 400 with fixed lag one used in the auxiliary regression.

The simulation results for the suggested test in terms of size (Type I error) and power are reported in Figure S2, which clearly demonstrates that the proposed test performs quite well for all considered error distributions. Particularly, when the null hypothesis is true, the size of the developed test is fairly close to the significance level, regardless of the choice of error distribution; as we move away from the null hypothesis, the power increases quickly. Furthermore, the power increases as the sample size increases and the curves exhibit similar patterns under different error variables. It is important to note that there is likely a considerable small sample effect that is dominating in the test, which necessitates a large sample size in order to achieve better power results using bootstrap.

One might curious on the effect of using different values for d in the auxiliary regression. We then conduct the test with $d = 4$. The results are shown in Table S3, from which we can conclude that the developed test has the desired invariance property with respect to the value of order in terms of size. The rejection frequency does not vary much with d. Nevertheless,

with a large value d, a large sample size is suggested to be utilized to avoid the effect of a small number of degrees of freedom.

Table S3: Rejection Frequence (%) of the Test

Order	$d=1$				$d=4$			
	$N(0,1)$	$t(3)$	$0.8L_p(0,1)$ $+0.2\mathrm{L}_p(0,4)$	$0.9N(0,1)$ $+0.1\mathrm{N}(0,5)$	$N(0,1)$	$t(3)$	$0.8L_p(0,1)$ $+0.2\mathrm{L}_p(0,4)$	$0.9N(0,1)$ $+0.1\mathrm{N}(0,5)$
	$\alpha=1\%$							
n=200	3.0	1.0	2.0	2.5	2.5	2.0	1.5	3.0
n=400	2.5	1.0	0.5	1.5	1.5	1.0	1.0	2.5
	$\alpha=5\%$							
n=200	4.0	5.5	4.5	5.5	5.5	4.5	5.0	4.0
n=400	4.5	5.0	5.0	5.0	4.5	5.0	4.0	5.5
	$\alpha=10\%$							
n=200	7.5	7.5	10.0	9.5	8.5	8.0	8.5	9.0
n=400	9.0	8.5	9.0	9.0	10.5	9.0	9.5	10.0

Remark S2. (Wald-Type Test) Aside from the proposed test, it is natural to investigate the Wald test by directly examining the variability of the estimated coefficient $\tilde{\beta}$. According to the asymptotic results from Theorem 2, it is clear that under certain regularity conditions, $\|\tilde{\beta}-\beta_0\| = O_p(h_1^2 + (n_0h_1^3)^{-1})$. With $n_0h_1^7 \to 0$, it can be obtained that

$$\sqrt{n_0h_1^3}(\tilde{\beta}-\beta_0) \xrightarrow{d} \mathcal{N}(0,\boldsymbol{\Xi}),$$

where $\boldsymbol{\Xi}$ is the asymptotic variance. Following Yao and Li (2014) and Ullah et al. (2021) to consistently estimate the corresponding density derivatives, we can get the consistent estimate for the asymptotic variance matrix $\boldsymbol{\Xi}$, which is defined as $\hat{\boldsymbol{\Xi}}$. We then have

$$\sqrt{n_0h_1^3}\hat{\boldsymbol{\Xi}}^{-1}(\tilde{\beta}-\beta_0) \xrightarrow{d} \mathcal{N}(0,I_d),$$

where I_d is an identity matrix with dimension $d\times d$ and $\|\sqrt{n_0h_1^3}\hat{\boldsymbol{\Xi}}^{-1}(\tilde{\beta}-$

$\beta_0)\|^2 \xrightarrow{d} \mathcal{X}_d^2$. Based on this, we can verify the presence of an AR error structure by looking at the parameter β. The similar modal-based bootstrap methodology in Remark 6 can be utilized to implement such a test. Note that both tests do not rely on the distributional assumptions of the errors.

DGP 4 (compared with mean estimation) To illustrate the advantage of the proposed modal estimation compared to the traditional mean estimation, we generate data according to the following model

$$Y_t = 1 + \theta X_t + \varepsilon_t, \tag{S1.7}$$

where the coefficient $\theta = 1$ and the covariate X_t follows a uniform distribution $U[0, 1]$. The error term ε_t follows an AR(2) process, which is

$$\varepsilon_t = 0.8\varepsilon_{t-1} - 0.5\varepsilon_{t-2} + \eta_t. \tag{S1.8}$$

Different from the symmetric error in DGP 1, we let η_t follow a mixture normal distribution $0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ with $\mathbb{E}(\eta) = 0$ and $Mode(\eta) = 1$. Therefore, the modal and mean coefficients in (S1.8) are different by a constant. For simplicity, the initial values for the coefficients are set to be mean estimates when conducting modal estimation.

The simulation results with 200 repetitions are shown in Table S4. With autocorrelation issue and i.i.d. η_t, although the parameter estimates from mean and mode are the same (the intercept is different), it shows that the proposed modal estimators are more efficient with the finite sample size,

reflecting in smaller SE and MSE. This is expectable since modal estimation intends to capture the "most likely" value, which is also consistent with the results reported in empirical analysis in Section 5. Note that with the above model setting, the mean and modal coefficients only differ in constant. But practically, it is difficult to know whether error terms really rely on covariates. For instance, if η_t depends on covariate X_t, we should expect that modal estimate is different from mean estimate when the error distribution is skewed. Then, modal regression can be used to supplement mean regression to uncover some distinguishing features of the data.

Table S4: Results of Estimations—DGP 4

	Mean		Traditional		Proposed	
Sample Size	θ (SE)	MSE(θ)	θ (SE)	MSE(θ)	θ (SE)	MSE(θ)
n=200	0.9904 (0.2458)	0.0602	0.9646 (0.4554)	0.2076	0.9909 (0.1982)	0.0392
n=400	1.0079 (0.1771)	0.0313	1.0147 (0.3315)	0.1095	1.0141 (0.1439)	0.0208
n=600	0.9718 (0.1379)	0.0197	0.9636 (0.2981)	0.0897	0.9894 (0.1188)	0.0141
n=1000	1.0037 (0.1034)	0.0107	0.9904 (0.24611)	0.0603	1.0106 (0.0902)	0.0082

We then assess the prediction performance by comparing with the mean prediction and the prediction procedure suggested by the editor, i.e., conduct prediction by using the least squares method to estimate θ and then using the residuals to estimate the mode of η (denoted as combined prediction in Table S5). We report the coverage probabilities of prediction intervals of length 0.1σ, 0.2σ, and 0.5σ with $\sigma = \sqrt{Var(\eta)} \approx 2$, respectively.

We follow the same DGP process as above but implement the out-of-sample prediction for the additional n data points with 200 repetitions. The results reported in Table S5 indicate that the proposed modal regression provides higher coverage probabilities compared to mean regression and the combined prediction procedure suggested by the editor, which is consistent with the observations in Ullah et al. (2021). Note that if we have the skewed distribution, it is expected that modal prediction should get better results compared to mean prediction, since modal prediction is trying to capture the "most likely" results (narrow prediction interval), while mean prediction is capturing average values (widen prediction interval). Similar prediction advantage has been observed in empirical analysis in Section 5.

Table S5: Results of Predictions—DGP 4

Width	Sample	Mean Prediction	Proposed Modal Prediction	Combined Prediction
0.1σ	n=200	0.03	0.075	0.05
	n=400	0.035	0.085	0.045
	n=600	0.03	0.07	0.04
	n=1000	0.035	0.07	0.05
0.2σ	n=200	0.06	0.17	0.105
	n=400	0.055	0.155	0.1
	n=600	0.06	0.165	0.1
	n=1000	0.065	0.15	0.1
0.5σ	n=200	0.25	0.33	0.295
	n=400	0.26	0.36	0.3
	n=600	0.28	0.39	0.325
	n=1000	0.275	0.39	0.34

Remark S3. (Modal Prediction Interval) We theoretically discuss how to construct asymmetric prediction intervals for new observations based on

the modal regression. For the simplicity of explanation, we assume that the error distribution of ε is independent of X. Let $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$ be the residuals of the modal regression estimate. We use $\hat{\varepsilon}_{[i]}$ to denote the ith smallest value of the residuals. The traditionally used mean prediction interval with confidence level $1-\alpha$ for a new covariate X_{new} is $[X_{new}\hat{\theta}+\hat{\varepsilon}_{[n_1]}, X_{new}\hat{\theta}+\hat{\varepsilon}_{[n_2]}]$, where $n_1 = \lfloor n\alpha/2 \rfloor$ and $n_2 = n - n_1$. This symmetric method will be ideal if the regression error distribution is symmetric. Since the modal regression focuses on the highest conditional density region, we propose the following method for modal regression to use the information of skewed error density to construct prediction intervals. Suppose $\hat{f}(\cdot)$ is a kernel density estimate of ε based on the residuals $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$. We find the indexes $k_1 < k_2$ such that $k_2 - k_1 = n_2 - n_1 = \lceil n(1-\alpha) \rceil$ and $\hat{f}(\hat{\varepsilon}_{[k_1]}) \approx \hat{f}(\hat{\varepsilon}_{[k_2]})$ (using the iterative process). The proposed modal prediction interval for a new covariate X_{new} is then $[X_{new}\hat{\theta} + \hat{\varepsilon}_{[k_1]}, X_{new}\hat{\theta} + \hat{\varepsilon}_{[k_2]}]$.

S2 Convergence of the Penalized MEM Algorithm

We briefly discuss the convergence (a sufficiently small change in the parameters or the kernel-based objective function) of the proposed penalized MEM algorithm. From the existing results, we know that $Q_{n_0}(\cdot^{(g+1)}) \geq Q_{n_0}(\cdot^{(g)})$. However, this monotonicity may lead to unsatisfied results because the value

of the objective function could remain at the same value at any interaction. Unlike the discussion in MEM Algorithm 1, we in what follows show that if we can have a set of stationary points in the space of parameters, then $Q_{n_0}(\cdot^{(g+1)}) > Q_{n_0}(\cdot^{(g)})$. Particularly, we follow the classical EM algorithm to define a stationary point of the function $Q^P_{n_0}(\beta)$ as any point of β where the gradient vector is zero (Wu, 1983; Lim and Oh, 2014). Let $M(\beta)$ be the point-to-set map (a function from points to subsets) implicitly defined by the algorithm that goes from $\hat{\beta}^{(g)}$ to $\hat{\beta}^{(g+1)}$ for any point $\hat{\beta}^{(g)}$. We have the following result.

Lemma 1. *With an initial value $\hat{\beta}^{(0)}$, let $\hat{\beta}^{(g)} = M^g(\hat{\beta}^{(0)})$ denote the corresponding mapping. If $Q^P_{n_0}(\beta) = Q^P_{n_0}(M(\beta))$ holds only for stationary points β of $Q^P_{n_0}$ and if $\hat{\beta}^*$ is a limit point of the sequence $\{\hat{\beta}^{(g)}\}$ such that $M(\beta)$ is continuous at $\hat{\beta}^*$, then $\hat{\beta}^*$ is a stationary point of $Q^P_{n_0}(\beta)$.*

Proof. Let Θ_β denote the set of limit points of the sequence $\{\hat{\beta}^{(g)}\}$. For any $\hat{\beta}^* \in \Theta_\beta$, through the process of passing to a subsequence, we obtain $\hat{\beta}^{(g_m)} \to \hat{\beta}^*$. Since the value of $Q^P_{n_0}(\hat{\beta}^{(g_m)})$ is increasing in the iteration indicator (a hill-climbing algorithm increasing if parameter is not a stationary point), the quantity will converge to a limit as $m \to -\infty$. Thus, taking limits in the inequalities $Q^P_{n_0}(\hat{\beta}^{(g_{m+1})}) \geq Q^P_{n_0}(M(\hat{\beta}^{(g_m)})) \geq Q^P_{n_0}(\hat{\beta}^{(g_m)})$ produces $Q^P_{n_0}(\hat{\beta}^*) = Q^P_{n_0}(\lim_{m\to\infty} M(\hat{\beta}^{(g_m)}))$ with the assumption that the limit

exists. If $M(\beta)$ is continuous at $\hat{\beta}^*$, we have $Q_{n0}^P(\hat{\beta}^*) = Q_{n0}^P(M(\hat{\beta}^*))$. □

There is no general convergence theorem for the MEM algorithm because convergence depends on the starting point. Different from the Newton-Raphson algorithm, which requires calculating the inverse of the Hessian matrix and has quadratic convergence ($|\hat{\beta}^{(g+1)} - \hat{\beta}^*| \leq C|\hat{\beta}^{(g)} - \hat{\beta}^*|^2$ for $C > 0$), the convergence of $\{\hat{\beta}^{(g)}\}$ to a stationary point of $Q_{n0}^P(\beta)$ is only linear ($|\hat{\beta}^{(g+1)} - \hat{\beta}^*| \leq C|\hat{\beta}^{(g)} - \hat{\beta}^*|$).

Lemma 2. *If the penalty function $p_{\lambda_j}(\cdot)$ is differentiable, the proposed penalized MEM Algorithm 2 yields a sequence $\{\hat{\beta}^{(g)}\}$ converging to at least the unique local maximum of $Q_{n0}^P(\beta)$.*

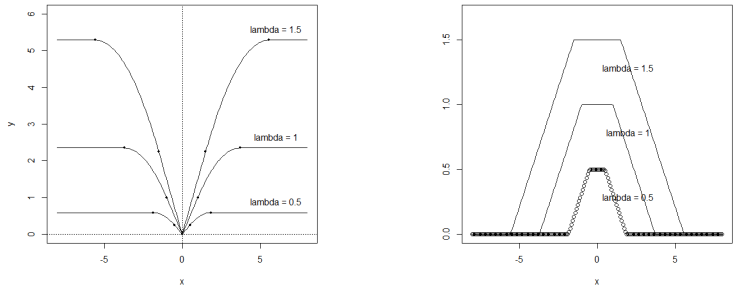

Figure S3: SCAD Function and Derivative

Since we utilize the SCAD penalty in the objective function (Figure S3), the differentiability condition is satisfied. Thus, we can conclude that the proposed penalized MEM algorithm at convergence will achieve at least the local maximum of $Q_{n0}^P(\beta)$. If we impose the existence of the global unique mode, the sequence $\{\hat{\beta}^{(g)}\}$ converges to the global maximum of $Q_{n0}^P(\beta)$.

S3 Extension to Nonparametric Regression

The developed two-step estimation procedure is heavily reliant on the assumption that the relationship between variables can be parametrically modeled. However, it is well known that for many practical econometric problems, parametric structure may be too restrictive to suffer from modeling biases or misspecification, and it is more attractive to adopt a flexible nonparametric form without any prior model structures; see Xiao et al. (2003) and Su and Ullah (2006). To the best of our knowledge, there is no research investigating nonparametric modal regression with correlated errors. Following the "letting the data speak for themselves" principle, we extend the results to nonparametric modal regression with error terms following a parametric AR process

$$Y_t = m(X_t) + \varepsilon_t, \tag{S3.9}$$

where $X_t \in \mathbb{R}$ for avoiding the "curse of dimensionality" and $m(\cdot)$ is an unknown smooth regression function.

Consistent with the parametric case, we remove the constraints imposed on the mean and variance of the model errors and are interested in estimating the modal regression line $m(x)$ given $X_t = x$. The errors follow the same AR structure as in (2.4). We concentrate on the AR process with little loss of generality because both moving average (MA) and ARMA processes can

be well approximated by an AR process provided that the latent roots of the MA polynomial lie inside the unit circle (Brockwell and Davis, 1991). Following the same procedures proposed in the paper, we can estimate the nonparametric model based on an additive partially linear modal regression and show that the resultant estimator has similar asymptotic properties to the estimator of nonparametric modal regression with i.i.d. observations under some mild conditions. So far as we know, there is no research investigating additive partially linear modal regression in the literature. Since B-splines can be used to describe complex, non-linear relationships between response and explanatory variables, we novelly introduce a B-splines based procedure for estimating modal regression.

In conjunction with (2.4), we can have

$$Mode(Y_t \mid F_{t-1}) = m(X_t) + \sum_{j=1}^{d} \beta_j (Y_{t-j} - m(X_{t-j})). \tag{S3.10}$$

To model the unknown function $m(\cdot)$, we implement the B-splines technique. Let $s_1, \cdots, s_M$ be the M interior knots with $s_0 < s_1 < \cdots < s_M < s_{M+1}$ and $s_{-(q-1)}, \cdots, s_{-1}$ and $s_{M+2}, \cdots, s_{M+q}$ be the $2(q-1)$ additional boundary knots such that $s_{-(q-1)} = \cdots = s_{-1} = s_0$ and $s_{M+1} = s_{M+2} = \cdots = s_{M+q}$, where q is the order of B-splines (for quadratic B-splines, $q = 3$). We denote a set of basis functions as $\{B_{i,q}(.)\}_{i=-(q-1)}^{M}$ and approximate $m(X_t) \approx \sum_{i=-(q-1)}^{M} a_i B_i(X_t)$, where

$$B_{i,q}(s) = \frac{s - s_i}{s_{i+q-1} - s_i} B_{i,q-1}(s) + \frac{s_{i+q} - s}{s_{i+q} - s_{i+1}} B_{i+1,q-1}(s), \tag{S3.11}$$

$i = -(q-1), \cdots, M, B_{i,1}(s) = 1$ if $s \in [s_j, s_{j+1}]$, and $B_{i,1}(s) = 0$ otherwise. We then have the following regression model

$$Y_t \approx \sum_{j=1}^{d} \beta_j Y_{t-j} + \sum_{i=-(q-1)}^{M} a_i \left[B_{i,q}(X_t) - \sum_{j=1}^{d} \beta_j B_{i,q}(X_{t-j}) \right], \tag{S3.12}$$

where the parameters are estimated by maximizing the following kernel-based objective function

$$\frac{1}{nh_B} \sum_{t=1}^{n} K \left(\frac{Y_t - \sum_{j=1}^{d} \beta_j Y_{t-j} - \sum_{i=-(q-1)}^{M} a_i \left[B_{i,q}(X_t) - \sum_{j=1}^{d} \beta_j B_{i,q}(X_{t-j}) \right]}{h_B} \right) \tag{S3.13}$$

with respect to β_j and a_i. It is noted that in B-spline smoothing, the interior knots are typically placed on a grid of equally spaced empirical quantiles. Although (S3.13) can provide unbiased estimators, we may have efficiency loss due to the unknown order of AR error terms.

Denote the preliminary estimates from (S3.13) as $\tilde{\beta}_j$ and $\tilde{a}_i$. We can construct the corresponding estimate $\hat{\varepsilon}_t$ as

$$\hat{\varepsilon}_t = Y_t - \sum_{i=-(q-1)}^{M} \tilde{a}_i B_{i,q}(X_t). \tag{S3.14}$$

Then, we can use a penalized kernel-based function associated with bandwidth h_C to reestimate β_j, where the corresponding estimate is defined as $\hat{\beta}_j^P$. In the last step, we recalculate a_i by maximizing

$$\frac{1}{nh_D}\sum_{t=1}^{n} K\left(\frac{Y_t^* - \sum_{i=-(q-1)}^{M} a_i B_{i,q}(X_t)}{h_D}\right), \tag{S3.15}$$

where $Y_t^* = Y_t - \sum_{j=1}^{d} \hat{\beta}_j^P \hat{\varepsilon}_{t-j}$ and $h_D = h_D(n) \to 0$ as $n \to \infty$ is a bandwidth. The resulting regression has a pseudo-residual term that is uncorrelated. The final estimate of $m(\cdot)$ is $\hat{m}(\cdot) = \sum_{i=-(q-1)}^{M} \hat{a}_i B_{i,q}(X_t)$, in which $\hat{a}_i$ is the estimate from (S3.15). The asymptotic theorems can be derived directly following the procedures in this paper (i.e., using a smaller bandwidth in the previous stage to control the bias in the preliminary step of the estimation), where we can show that the resultant estimator has the oracle property as if the true variables were known in advance, and enjoys the nice asymptotic properties parallel to the independent error case. Note that the last step estimation can be considered as a post model selection estimation, which is to reduce the estimation bias resulting from simultaneous shrinkage towards both significant and non-significant AR coefficients.

Remark S4. For the linear mean regression with serial correlation, we can apply the standard generalized least squares to incorporate the error autocorrelation function to improve the efficiency of mean estimators. However, such a method may not be suitable for modal regression unless we impose a zero conditional mode value on error term and assume the additivity property for mode satisfied.

S4 Technical Proofs

S4.1 Proof of Theorem 1

Recall that

$$\begin{aligned}&\frac{1}{n_0h_1}\sum_{t=d+1}^{n}K\left(\frac{Y_t-X_t^T\theta-\sum_{j=1}^{d}\tilde{\beta}_jY_{t-j}+\sum_{j=1}^{d}\tilde{\beta}_jX_{t-j}^T\theta}{h_1}\right)\\=&\frac{1}{n_0h_1}\sum_{t=d+1}^{n}K\left(\frac{Y_t-X_t^T\theta-\sum_{j=1}^{d}\beta_{0j}Y_{t-j}+\sum_{j=1}^{d}\beta_{0j}X_{t-j}^T\theta+error_t}{h_1}\right),\end{aligned}\tag{S4.16}$$

where $error_t=\sum_{j=1}^{d}\beta_{0j}Y_{t-j}-\sum_{j=1}^{d}\tilde{\beta}_jY_{t-j}+\sum_{j=1}^{d}\tilde{\beta}_jX_{t-j}^T\theta-\sum_{j=1}^{d}\beta_{0j}X_{t-j}^T\theta$.

Define $\delta_n=h_1^2+\sqrt{(n_0h_1^3)^{-1}}$. Then, it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\mu\|=c}Q_{n_0}\left(\theta_0+\delta_n\mu\right)<Q_{n_0}\left(\theta_0\right)\right\}\geq 1-\eta,\tag{S4.17}$$

where θ_0 is the true parameter and $\|\cdot\|$ represents the Euclidean distance. The above equation implies that with probability tending to one, there is a local maximum in the ball $\{\theta_0+\delta_n\mu:\|\mu\|\leq c\}$. Using the Taylor expansion, it follows that

$$\begin{aligned}&Q_{n_0}\left(\theta_0+\delta_n\mu\right)-Q_{n_0}\left(\theta_0\right)\\=&\frac{1}{n_0h_1}\sum_{t=d+1}^{n}K\left(\frac{\eta_t-\delta_n\mu^T(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j})+error_t}{h_1}\right)\\&-\frac{1}{n_0h_1}\sum_{t=d+1}^{n}K\left(\frac{\eta_t+error_t}{h_1}\right)\\=&\frac{1}{n_0h_1}\sum_{t=d+1}^{n}\left[-K^{(1)}\left(\frac{\eta_t+error_t}{h_1}\right)\left(\frac{\delta_n\mu^T(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j})}{h_1}\right)\right.\end{aligned}$$

$$+\frac{1}{2}K^{(2)}\left(\frac{\eta_t+error_t}{h_1}\right)\left(\frac{\delta_n\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1}\right)^2$$
$$-\frac{1}{6}K^{(3)}\left(\frac{\eta_t^*}{h_1}\right)\left(\frac{\delta_n\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1}\right)^3\Bigg] \tag{S4.18}$$
$$=I_1+I_2+I_3,$$

where η_t^* is between $\eta_t+error_t$ and $\eta_t+error_t-\delta_n\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})$. Based on the result $T_n=\mathbb{E}(T_n)+O_p(\sqrt{\mathrm{Var}(T_n)})$, we consider each part of the above Taylor expansion.

(i) For the first part, which is $I_1=\frac{1}{n_0h_1}\sum_{t=d+1}^n\Big[-K^{(1)}\left(\frac{\eta_t+error_t}{h_1}\right)\left(\frac{\delta_n\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1}\right)\Big]$, by Taylor expansion, we can rewrite it as

$$\begin{aligned}\mathbb{E}(I_1)=&\frac{-\delta_n}{h_1}\mathbb{E}\left(K^{(1)}\left(\frac{\eta_t+error_t}{h_1}\right)\left(\frac{\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1}\right)\right)\\
=&\frac{-\delta_n}{h_1}\mathbb{E}\Bigg(K^{(1)}\left(\frac{\eta_t}{h_1}\right)\frac{\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1}\\
&+K^{(2)}\left(\frac{\eta_t}{h_1}\right)\frac{error_t\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1^2}\\
&+\frac{1}{2}K^{(3)}\left(\frac{\eta_t^{**}}{h_1}\right)\frac{(error_t)^2\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})}{h_1^3}\Bigg)\\
=&I_{11}+I_{12}+I_{13},\end{aligned} \tag{S4.19}$$

where η_t^{**} is between η_t and $\eta_t+error_t$. As the order of η_t^{**} is the same as that of η_t, when we do the calculation associated with I_{13}, we instead use η_t directly. By some direct calculations for each part, we can get

$$I_{11} = \frac{-\delta_n}{h_1}\mathbb{E}\left(K^{(1)}\left(\frac{\eta_t}{h_1}\right)\frac{\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j})}{h_1}\right) = O_p(\delta_n c h_1^2). \tag{S4.20}$$

$$\begin{aligned} I_{12} =& \frac{-\delta_n}{h_1}\mathbb{E}\left(K^{(2)}\left(\frac{\eta_t}{h_1}\right)\frac{\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j})}{h_1}\frac{error_t}{h_1}\right) \\ =& \frac{-\delta_n}{h_1}\iint K^{(2)}\left(\frac{\eta}{h_1}\right)\frac{\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j})}{h_1} g_\eta(\eta)\frac{error_t}{h_1} d\eta dF(X) \\ =& \frac{-\delta_n}{h_1}\iint K(\tau)(\tau^2-1)\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j}) g_\eta(\tau h_1)\frac{error_t}{h_1} d\tau dF(X) \\ =& O_p(\delta_n c h_1^2). \end{aligned} \tag{S4.21}$$

With the condition that $\|\tilde{\beta} - \beta_0\|/h_1^2 \to 0$, it can be seen that I_{11} dominates I_{12} and I_{13}. Meanwhile, we obtain

$$\frac{\delta_n^2}{h_1^2}\mathbb{E}\left(K^{(1)}\left(\frac{\eta_t}{h_1}\right)\frac{\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j})}{h_1}\right)^2 = O_p(\delta_n^2 c^2 (h_1^3)^{-1}). \tag{S4.22}$$

These equations show that $I_1 = O_p(\delta_n c h_1^2) + O_p(\sqrt{\delta_n^2 c^2 (n_0 h_1^3)^{-1}}) = O_p(\delta_n^2 c)$.

(ii) For the second part, which is $I_2 = \frac{1}{n_0 h_1}\sum_{t=d+1}^n \left(\frac{1}{2}K^{(2)}\left(\frac{\eta_t + error_t}{h_1}\right)\left(\frac{\delta_n \mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j})}{h_1}\right)^2\right)$, we can rewrite it as

$$\begin{aligned} \mathbb{E}(I_2) =& \frac{\delta_n^2}{2h_1}\mathbb{E}\left(K^{(2)}\left(\frac{\eta_t + error_t}{h_1}\right)\frac{(\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j}))^2}{h_1^2}\right) \\ =& \frac{\delta_n^2}{2h_1}\mathbb{E}\Bigg(K^{(2)}\left(\frac{\eta_t}{h_1}\right)\frac{(\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j}))^2}{h_1^2} \\ & + K^{(3)}\left(\frac{\eta_t}{h_1}\right)\frac{error_t(\mu^T(X_t - \sum_{j=1}^d \beta_{0j}X_{t-j}))^2}{h_1^3} \end{aligned}$$

$$+\frac{1}{2}K^{(4)}\left(\frac{\eta_t^{**}}{h_1}\right)\frac{error_t^2(\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}))^2}{h_1^4}\Bigg) = I_{21}+I_{22}+I_{23}, \tag{S4.23}$$

where η_t^{**} is between η_t and $\eta_t + error_t$. Notice that as the order of η_t^{**} is the same as that of η_t, when we do the calculation associated with I_{23}, we instead use η_t directly. By some calculations for each part, we can get

$$I_{21} = \frac{\delta_n^2}{2h_1}\mathbb{E}\left(K^{(2)}\left(\frac{\eta_t}{h_1}\right)\frac{(\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}))^2}{h_1^2}\right) = O_p((\delta_n c)^2). \tag{S4.24}$$

$$I_{22} = \frac{\delta_n^2}{2h_1}\mathbb{E}\left(K^{(3)}\left(\frac{\eta_t}{h_1}\right)\frac{error_t(\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}))^2}{h_1^3}\right) = o_p((\delta_n c)^2). \tag{S4.25}$$

Meanwhile, we can prove that $I_{23} = o_p((\delta_n c)^2)$ as well. Following the same steps in (i), we obtain the following result

$$\frac{\delta_n^4}{4h_1^2}\mathbb{E}\left(K^{(2)}\left(\frac{\eta_t}{h_1}\right)\frac{(\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}))^2}{h_1^2}\right)^2 = O_p((\delta_n c)^4(h_1^5)^{-1}). \tag{S4.26}$$

With the condition $n_0h_1^5 \to \infty$ held, the above equations indicate that the second part will dominate the first part when we choose c big enough.

(iii) Following the same way, we can calculate the third part. As the order of η_t^* is the same as the order of η_t, by direct calculation, we have

$$\frac{\delta_n^3}{6h_1}\mathbb{E}\left(K^{(3)}\left(\frac{\eta_t}{h_1}\right)\frac{(\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}))^3}{h_1^3}\right) = O_p(\delta_n^3). \tag{S4.27}$$

$$\frac{\delta_n^6}{36h_1^2}\mathbb{E}\left(K^{(3)}\left(\frac{\eta_t}{h_1}\right)\frac{\left(\mu^T(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})\right)^3}{h_1^3}\right)^2=O_p(\delta_n^6(h_1^7)^{-1}). \tag{S4.28}$$

These indicate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, $P\left\{\sup_{\|\boldsymbol{\mu}\|=c}Q_{n_0}\left(\theta_0+\delta_n\mu\right)<Q_{n_0}\left(\theta_0\right)\right\}\geq 1-\eta$ holds naturally.

□

S4.2 Proof of Theorem 2

At first, the estimator $\tilde{\theta}$ must satisfy

$$\begin{aligned}-\frac{1}{n_0h_1^2}\sum_{t=d+1}^n K^{(1)}\left(\frac{\eta_t-(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})^T(\tilde{\theta}-\theta_0)+error_t^{\tilde{\theta}}}{h_1}\right)\\ \left(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}\right)=0,\end{aligned} \tag{S4.29}$$

where $error_t^{\tilde{\theta}}=\sum_{j=1}^d\beta_{0j}Y_{t-j}-\sum_{j=1}^d\tilde{\beta}_jY_{t-j}+\sum_{j=1}^d\tilde{\beta}_jX_{t-j}^T\tilde{\theta}-\sum_{j=1}^d\beta_{0j}X_{t-j}^T\tilde{\theta}$.

By taking Taylor expansion, we can obtain

$$\begin{aligned}&-\frac{1}{n_0h_1^2}\sum_{t=d+1}^n K^{(1)}\left(\frac{\eta_t}{h_1}\right)\left(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}\right)\\ &+\frac{1}{n_0h_1^3}\sum_{t=d+1}^n K^{(2)}\left(\frac{\eta_t}{h_1}\right)\left(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j}\right)\\ &(error_t^{\tilde{\theta}}-(X_t-\sum_{j=1}^d\beta_{0j}X_{t-j})^T(\tilde{\theta}-\theta_0))\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{n_0 h_1^2} \sum_{t=d+1}^{n} K^{(1)}\left(\frac{\eta_t}{h_1}\right)\left(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j}\right) \\
&+ \frac{1}{n_0 h_1^3} \sum_{t=d+1}^{n} K^{(2)}\left(\frac{\eta_t}{h_1}\right)\left(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j}\right) \\
&(error_t^{\tilde{\theta}} - (X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})^T(\tilde{\theta} - \theta_0)) \qquad \text{(S4.30)} \\
&- \frac{1}{n_0 h_1^4} \sum_{t=d+1}^{n} K^{(3)}\left(\frac{\tilde{\eta}_t^*}{h_1}\right)\left(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j}\right) \\
&(error_t^{\tilde{\theta}} - (X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})^T(\tilde{\theta} - \theta_0))^2 = 0,
\end{aligned}
$$

where $\tilde{\eta}_t^*$ is between η_t and $\eta_t + error_t^{\tilde{\theta}} - (X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})^T(\tilde{\theta} - \theta_0)$. Assuming $\|\tilde{\beta} - \beta_0\|/h_1^2 \to 0$, from Theorem 1, we know $\|\tilde{\theta} - \theta_0\| = O_p(\delta_n)$, which indicates that $|error_t^{\tilde{\theta}} - (X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})^T(\tilde{\theta} - \theta_0)| = O_p(\|\tilde{\theta} - \theta_0\|) = O_p(\delta_n)$. We can see that the third part, which is associated with $(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})(error_t^{\tilde{\theta}} - (X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})^T(\tilde{\theta} - \theta_0))^2$, is dominated by the second part, which is associated with $(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})(error_t^{\tilde{\theta}} - (X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j})^T(\tilde{\theta} - \theta_0))$. We then mainly focus on the first two parts of the left side of the above equation.

Considering $-\frac{1}{n_0 h_1^2} \sum_{t=d+1}^{n} K^{(1)}\left(\frac{\eta_t}{h_1}\right)(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j}) + \frac{1}{n_0 h_1^3} \sum_{t=d+1}^{n} K^{(2)}\left(\frac{\eta_t}{h_1}\right)(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j}) error_t^{\tilde{\theta}}$, by direct calculations, we can obtain

$$
\mathbb{E}\Bigg(- \frac{1}{n_0 h_1^2} \sum_{t=d+1}^{n} K^{(1)}\left(\frac{\eta_t}{h_1}\right)\left(X_t - \sum_{j=1}^{d} \beta_{0j} X_{t-j}\right)
$$

$$
\begin{aligned}
&+\frac{1}{n_0h_1^3}\sum_{t=d+1}^{n}K^{(2)}\left(\frac{\eta_t}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)error_t^{\tilde{\theta}}\Bigg)\\
=&-\frac{1}{h_1^2}\iint K^{(1)}\left(\frac{\eta}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)g_\eta(\eta)d\eta dF(X)\\
&+\frac{1}{h_1^3}\iint K^{(2)}\left(\frac{\eta}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)g_\eta(\eta)error_t^{\tilde{\theta}}d\eta dF(X)\\
=&\frac{1}{h_1}\iint K(\tau)\,\tau\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)g_\eta(\tau h_1)d\tau dF(X)\\
&-\frac{1}{h_1^2}\iint K(\tau)\,(\tau^2-1)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)g_\eta(\tau h_1)error_t^{\tilde{\theta}}d\tau dF(X)\\
=&\frac{h_1^2}{2}\mathbb{E}\left(\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)g_\eta^{(3)}(0)\right)\{1+o_p(1)\}.
\end{aligned}
\tag{S4.31}
$$

Considering $\frac{1}{n_0h_1^3}\sum_{t=d+1}^{n}K^{(2)}\left(\frac{\eta_t}{h_1}\right)(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j})(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j})^T$, by direct calculations, we have

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{n_0h_1^3}\sum_{t=d+1}^{n}K^{(2)}\left(\frac{\eta_t}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)^T\right)\\
&=\mathbb{E}\left(\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)^T g_\eta^{(2)}(0)\right).
\end{aligned}
\tag{S4.32}
$$

Based on the above two equations, we can achieve

$$
\tilde{\theta}-\theta_0=\frac{h_1^2}{2}\left(\mathbb{E}\left(\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)^T g_\eta^{(2)}(0)\right)\right)^{-1}
$$

$$\mathbb{E}\left(\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)g_\eta^{(3)}(0)\right)\{1+o_p(1)\}. \tag{S4.33}$$

Meanwhile, with the condition $\|\tilde{\beta}-\beta_0\|/h_1^2\to 0$ held, we can obtain

$$\begin{aligned}
&\mathrm{Var}\Big(-\frac{1}{n_0h_1^2}\sum_{t=d+1}^{n}K^{(1)}\left(\frac{\eta_t}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)\\
&+\frac{1}{n_0h_1^3}\sum_{t=d+1}^{n}K^{(2)}\left(\frac{\eta_t}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)error_t^{\tilde{\theta}}\Big)\\
=&\frac{1}{n_0h_1^4}\iint K^{(1)2}\left(\frac{\eta}{h_1}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)^T g_\eta(\eta)\\
&d\eta dF(X)(1+o_p(1))\\
=&\frac{\int\tau^2K^2(\tau)d\tau}{n_0h_1^3}\mathbb{E}\left(\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)\left(X_t-\sum_{j=1}^{d}\beta_{0j}X_{t-j}\right)^T g_\eta(0)\right)\\
&(1+o_p(1)).
\end{aligned} \tag{S4.34}$$

For the remaining part, we can follow the same idea in Yao and Li (2014) and Ullah et al. (2021, 2022, 2023) to easily obtain the results.

□

S4.3 Proof of Theorem 3

Following the steps to prove Theorem 1, we define $\delta_n=h_2^2+\sqrt{(n_0h_2^3)^{-1}}+a_n$. Then, it is sufficient to show that for any given η, there exists a large number

constant c such that

$$P\left\{\sup_{\|\mu\|=c} Q_{n_0}^P(\beta_0+\delta_n\mu) < Q_{n_0}^P(\beta_0)\right\} \geq 1-\eta. \tag{S4.35}$$

Using $p_{\lambda_j}(0)=0$ and Taylor expansion, it follows that

$$\begin{aligned}
&Q_{n_0}^P(\beta_0+\delta_n\mu) - Q_{n_0}^P(\beta_0) - \sum_{j=1}^{d}[p_{\lambda_j}(|\beta_{0j}+\delta_n\mu_j|) - p_{\lambda_j}(|\beta_{0j}|)] \\
=&\delta_n Q_{n_0}^{P(1)}(\beta_0)^T\mu + \frac{1}{2}\delta_n^2\mu^T Q_{n_0}^{P(2)}(\beta_0)^T\mu + \frac{1}{6}\delta_n^3\mu^T Q_{n_0}^{P(3)}(\beta_0^*)^T\mu^T\mu \\
&- \sum_{j=1}^{s}\left[\delta_n p_{\lambda_j}^{(1)}(|\beta_{0j}|)\operatorname{sgn}(\beta_{0j})\mu_j + \delta_n^2 p_{\lambda_j}^{(2)}(|\beta_{0j}|)\mu_j^2\{1+o_p(1)\}\right] \\
=&M_1+M_2+M_3+M_4,
\end{aligned} \tag{S4.36}$$

where $\|\beta_0^*-\beta_0\| \leq c\delta_n$. From the Proof of Theorem 1, we know

$$M_1 = O_p(\delta_n^2 c), M_2 = O_p(\delta_n^2 c^2), \text{ and } M_3 = O_p(\delta_n^3). \tag{S4.37}$$

By choosing bigger enough c, M_2 could domain M_1 and M_3 with probability $1-\eta$. Note that M_4 is bounded by

$$\sqrt{s}\delta_n \max\left\{p_{\lambda_j}^{(1)}(|\beta_{j0}|) : \beta_{0j}\neq 0\right\}\|\mu\| + \delta_n^2 \max\left\{p_{\lambda_j}^{(2)}(|\beta_{j0}|) : \beta_{0j}\neq 0\right\}\|\mu\|^2, \tag{S4.38}$$

which is also dominated by M_2 as $\max\left\{p_{\lambda_j}^{(2)}(|\beta_{0j}|) : \beta_{0j}\neq 0\right\} \to 0$. Because $Q_{n_0}^{P(2)}(\boldsymbol{\beta}_0) < 0$, we have $Q_{n_0}^P(\beta_0+\delta_n\mu) < Q_{n_0}^P(\beta_0)$ with probability $1-\eta$ for $\eta > 0$ by choosing a sufficiently large c.

□

S4.4 Proof of Theorem 4

By the property of SCAD penalty function, as $\lambda_{max} \to 0$, it can be shown that $a_n = 0$ for large n_0. Then, according to Theorem 3, it is sufficient to show that for any β^P that satisfies $\left\|\beta^P - \beta_0\right\| = O_p(\delta_n)$ and for some small $\epsilon = c\delta_n$ in which $\delta_n = h_2^2 + \sqrt{(n_0 h_2^3)^{-1}}$, when $n_0 \to \infty$, with probability tending to one, we have

$$\frac{\partial Q_{n_0}^P(\beta)}{\partial \beta_j^P} < 0, \quad \text{for } 0 < \beta_j^P < \epsilon,\ j = s+1, \cdots, d, \tag{S4.39}$$

$$\frac{\partial Q_{n_0}^P(\beta)}{\partial \beta_j^P} > 0, \quad \text{for } -\epsilon < \beta_j^P < 0,\ j = s+1, \cdots, d, \tag{S4.40}$$

which indicates that the maximizer of $Q_{n_0}^P(\beta)$ gets at $\beta_j^P = 0$.

Similar to the Proof of Theorem 3, as $Q_{n_0}^{P(1)}(\beta_0) = O_p(\delta_n)$ and $\|\beta^P - \beta_0\| = O_p(\delta_n)$, we can obtain

$$\begin{aligned}
&n_0 Q_{n_0}^{P(1)}(\beta) - n_0 p_{\lambda_j}^{(1)}\left(|\beta_j^P|\right)\operatorname{sgn}\beta_j^P \\
=&n_0 Q_{n_0}^{P(1)}(\beta_0) + n_0 Q_{n_0}^{P(2)}(\beta_0)(\beta_{0j} - \beta_j^P) + \frac{n_0}{2} Q_{n_0}^{P(3)}(\beta_0^*)(\beta_{0j} - \beta_j^P)^2 \\
&- n_0 p_{\lambda_j}^{(1)}\left(|\beta_j^P|\right)\operatorname{sgn}\beta_j^P \\
=&- n_0\lambda_j\left\{\lambda_j^{-1} p_{\lambda_j}^{(1)}\left(|\beta_j^P|\right)\operatorname{sgn}\beta_j^P + O_p\left(\delta_n/\lambda_j\right)\right\},
\end{aligned} \tag{S4.41}$$

where β_0^* is between β and β_0. As $\delta_n^{-1}\lambda_j \geq \delta_n^{-1}\lambda_{min} \to \infty$ when $n_0 \to \infty$ and $\liminf_{n_0\to 0}\liminf_{\beta_j^P \to 0} p_{\lambda_j}^{(1)}(|\beta_j^P|)/\lambda_j > 0$, the sign of the derivation is completely determined by that of β_j^P. This completes the proof.

□

S4.5 Proof of Theorem 5

From the Proof of Theorem 4, we can know that for $j = 1, \cdots, s$, we have

$$\begin{aligned}
&Q_{n_0}^{P(1)}(\hat{\beta}_{0'}^P) - p_{\lambda_j}^{(1)}\left(|\hat{\beta}_{0'j}^P|\right)\operatorname{sgn}\hat{\beta}_{0'j}^P \\
=&Q_{n_0}^{P(1)}(\beta_{0'}) + Q_{n_0}^{P(2)}(\beta_{0'})(\beta_{0'j} - \hat{\beta}_{0'j}^P) + \frac{1}{2}Q_{n_0}^{P(3)}(\beta_0^*)(\beta_{0'j} - \hat{\beta}_{0'j}^P)^2 \\
&- \left\{p_{\lambda_j}^{(1)}\left(|\beta_{0'j}|\right)\operatorname{sgn}\beta_{0'j} + (p_{\lambda_j}^{(2)}\left(|\beta_{0'j}|\right) + o_p(1))(\hat{\beta}_{0'j}^P - \beta_{0'j})\right\}.
\end{aligned} \tag{S4.42}$$

Combining the equations, we have

$$\begin{aligned}
&Q_{n_0}^{P(1)}(\beta_{0'}) + Q_{n_0}^{P(2)}(\beta_{0'})(\beta_{0'j} - \hat{\beta}_{0'j}^P) + \frac{1}{2}Q_{n_0}^{P(3)}(\beta_0^*)(\beta_{0'j} - \hat{\beta}_{0'j}^P)^2 \\
&- \left\{\Psi_\lambda + (\Phi_\lambda + o_p(1))(\hat{\beta}_{0'j}^P - \beta_{0'j})\right\} = 0.
\end{aligned} \tag{S4.43}$$

From Theorem 4, following by Slutskys theorem and the central limit theorem, we know

$$\frac{h_2^2}{2}M_{(1)} - J_{(1)}(\hat{\beta}_{0'}^P - \beta_{0'}) - \left\{\Psi_\lambda + (\Phi_\lambda + o_p(1))(\hat{\beta}_{0'}^P - \beta_{0'})\right\} = 0, \tag{S4.44}$$

and the asymptotic distribution

$$\begin{aligned}
&\sqrt{nh_2^3}(J_{(1)} + \Phi_\lambda)\left(\hat{\beta}_{0'}^P - \beta_{0'} + (J_{(1)} + \Phi_\lambda)^{-1}\left(\Psi_\lambda - \frac{h_2^2}{2}\frac{g_\eta^{(3)}(0)}{g_\eta^{(2)}(0)}M_{(1)}\right)\right) \\
&\qquad \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\eta(0)}{g_\eta^{(2)}(0)^2}\int t^2K^2(t)dt J_{(1)}^{-1}\right),
\end{aligned} \tag{S4.45}$$

where $J_{(1)}$ and $M_{(1)}$ are the submatrices of J_β and M_β.

□

S4.6 Proof of Theorem 6

Recall that

$$Q_{n_0}(\theta) = \frac{1}{n_0 h_3} \sum_{t=d+1}^{n} K\left(\frac{Y_t - \hat{e}_t^T\hat{\beta} - X_t^T\theta}{h_3}\right)$$
$$= \frac{1}{n_0 h_3} \sum_{t=d+1}^{n} K\left(\frac{Y_t - e^T\beta - X_t^T\theta + e^T\beta - \hat{e}_t^T\beta + \hat{e}_t^T\beta - \hat{e}_t^T\hat{\beta}}{h_3}\right), \tag{S4.46}$$

where $e^T\beta - \hat{e}_t^T\beta + \hat{e}_t^T\beta - \hat{e}_t^T\hat{\beta} = O_p\left((n_0 h_3^3)^{-1/2} + h_3^2\right)$ when $\beta = \beta_0$. Define $\delta_n = h_3^2 + \sqrt{(n_0 h_3^3)^{-1}}$. Then, it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\mu\|=c} Q_{n_0}\left(\theta_0 + \delta_n\mu\right) < Q_{n_0}\left(\theta_0\right)\right\} \geq 1 - \eta. \tag{S4.47}$$

Following the same steps as the Proof of Theorem 1, with assumption $h_2/h_3 \to 0$, we can show $\|\hat{\theta} - \theta_0\| \leq \delta_n$.

□

S4.7 Proof of Theorem 7

Recall that if $\hat{\theta}$ is the optimal estimator, it will satisfy the following equation

$$Q_{n_0}(\theta) = -\frac{1}{n_0 h_3^2} \sum_{t=d+1}^{n} K\left(\frac{Y_t - \hat{e}_t^T\hat{\beta} - X_t^T\hat{\theta}}{h_3}\right) X_t = -\frac{1}{n_0 h_3^2}$$
$$\sum_{t=d+1}^{n} K\left(\frac{Y_t - e^T\beta - X_t^T\theta + X_t^T\theta - X_t^T\hat{\theta} + e^T\beta - \hat{e}_t^T\beta + \hat{e}_t^T\beta - \hat{e}_t^T\hat{\beta}}{h_3}\right) = 0 \tag{S4.48}$$

when $\theta = \theta_0$ and $\beta = \beta_0$. By taking Taylor expansion, we can obtain

$$
\begin{aligned}
&-\frac{1}{n_0 h_3^2}\sum_{t=d+1}^{n} K^{(1)}\left(\frac{\eta_t}{h_3}\right) X_t + \frac{1}{n_0 h_3^3}\sum_{t=d+1}^{n} K^{(2)}\left(\frac{\eta_t}{h_3}\right) X_t \\
&(X_t^T\theta - X_t^T\hat{\theta} + e^T\beta - \hat{e}_t^T\beta + \hat{e}_t^T\beta - \hat{e}_t^T\hat{\beta}) \\
&-\frac{1}{n_0 h_3^4}\sum_{t=d+1}^{n} K^{(3)}\left(\frac{\tilde{\eta}_t^*}{h_3}\right) X_t (X_t^T\theta - X_t^T\hat{\theta} + e^T\beta - \hat{e}_t^T\beta + \hat{e}_t^T\beta - \hat{e}_t^T\hat{\beta})^2 = 0,
\end{aligned}
\tag{S4.49}
$$

where $\tilde{\eta}_t^*$ is between η_t and $\eta_t + X_t^T\theta - X_t^T\hat{\theta} + e^T\beta - \hat{e}_t^T\beta + \hat{e}_t^T\beta - \hat{e}_t^T\hat{\beta}$.

It can be shown that the third term on the left-hand side of the above equation is dominated by the second term. With the assumption $h_2/h_3 \to 0$, we can then follow the same proof steps as those of Proof of Theorem 2 to achieve the results.

□

Bibliography

Brockwell, P. J. and Davis, R. A. (1991). Time Series: Theory and Methods. *New York: Springer-Verlag.*

Lim, Y. and Oh, H. (2014). Variable Selection in Quantile Regression When the Models Have Autoregressive Errors. *Journal of the Korean Statistical Society*, 43, 513-530.

Su, L. and Ullah, A. (2006). More Efficient Estimation in Nonparamet-

ric Regression with Nonparametric Autocorrelated Errors. *Econometric Theory*, 22, 98-126.

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*, 185 (3), 1424-1453.

Ullah, A., Wang, T., and Yao, W. (2023). Semiparametric Partially Linear Varying Coefficient Modal Regression. *Journal of Econometrics*, 235 (2), 1001-1026.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *Annals of Statistics*, 11 (1), 95-103.

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. and Li, L. (2014). A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics*, 41, 656-671.