ROBUST INFERENCE IN VARYING-COEFFICIENT ADDITIVE MODELS FOR LONGITUDINAL/FUNCTIONAL DATA

Lixia Hu[1], Tao Huang[2] and Jinhong You[2]

[1] *Shanghai Lixin University of Accounting and Finance*

[2] *Shanghai University of Finance and Economics*

Supplementary Material

This part contains additional numerical studies and technical proofs of the asymptotic theories in the manuscript. In addition, we describe locally quadratic approximation (LQA) in the model identification procedure.

S1 Complementary Numerical Studies

S1.1 Example 1(Continued)

First, we fix $(n, m) = (30, 20)$, and compute three-step M-estimators under the loss functions ρ_1, ρ_2, and ρ_3, denoted as Ls, Lad, and Hub, respectively. For comparison, we also compute the spline-based oracle estimator of varying-coefficient component functions given all additive component functions in advance and, analogously, the oracle estimator of additive component functions when all varying-coefficient component functions are known. We denote the oracle estimator by the suffix -O, e.g., Ls-O denotes the oracle estimator under the quadratic loss function. Based on 500 Monte Carlo replications, Table 1 compares the average MSE (AMSE) of the three-step M-estimators with that of the oracle estimators under different error distributions.

The standard deviation is given in parentheses. From Table 1, we conclude that

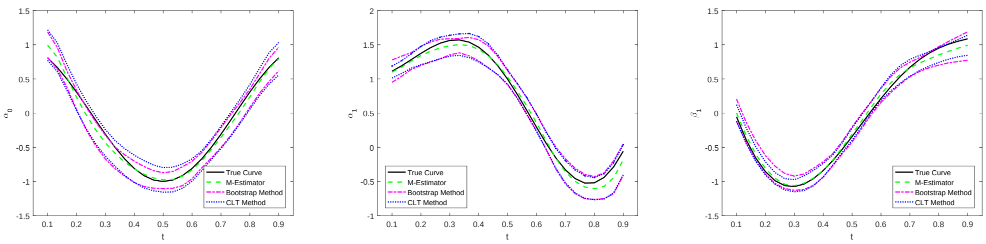
- under the normal error distribution, Huber estimators are comparable to least-squares estimators, and median estimators are slightly inferior. Moreover, the AMSE of the three-step M-estimators is similar to that of the oracle estimators, even for medium sample sizes. This embodies the oracle property of the three-step M-estimator, as if more information is known in advance.

- under non-normal error distributions, least-squares estimators exhibit worse performance than the others, especially under the $t(1)$ error distribution. The Huber and median estimators have similar performance, and their AMSEs are similar to those of the oracle estimators.

- The influence of the intra-subject covariance structure is not substantial under the different error distributions and loss functions.

Under normal error distribution $N(0, 0.2)$, Figure 1 presents the iterative least square estimator (dashed line) and 95% CLT-based CI (dotted lines) and 500 wild bootstrap sampling (dash-dotted lines). We note that the similar performance with Figure 1 (under mixed normal error distribution) in the manuscript, which indicates the rationality of our estimation method and two types CI.

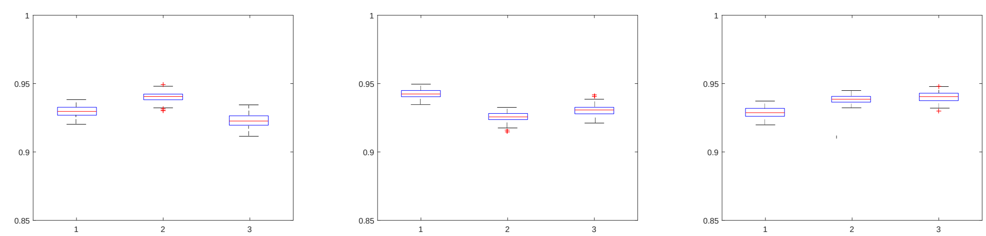
For normal error distribution and mixed normal distribution, we also investigate the average experience coverage probability (AECP) of three-step M-estimator at given 50 grid poins on the range of interested variable. We sample 200 times with dependent within-subject correlations ($\theta = 0.5$) and make 500 Monte Carlo replications in each run. Figure 2 presents the boxplot of AECP under normal error distribution for different subjects/observations combinations: 1 is for $(n, m) = (20, 30)$; 2 and 3 are for $(n, m) = (30, 20)$ and $(60, 10)$, respectively. The



(a) α_0 (b) α_1 (c) β_1

Figure 1: Three-step M-estimators under normal error distribution. Solid line: true component function, dashed line: three-step M-estimator, dotted lines: 95% CI based on asymptotic distribution, and dash-dotted lines: 95% CI based on 500 wild bootstrap resampling.

results implies that the CLT-based CI is acceptable.

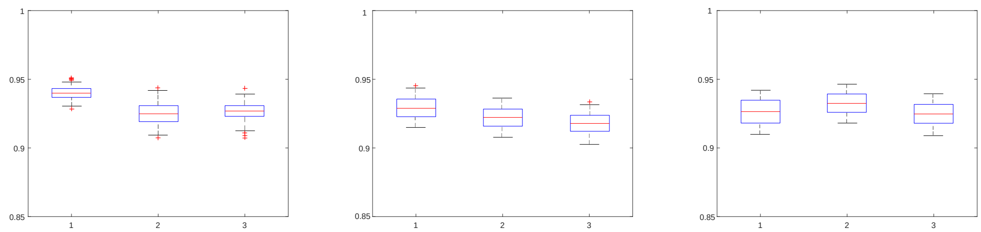


(a) AECP for α_0 (b) AECP for α_1 (c) AECP for β_1

Figure 2: Boxplots for average experience coverage probability (AECP) under normal error distribution: 1, $(n, m) = (20, 30)$; 2, $(n, m) = (30, 20)$; 3, $(n, m) = (60, 10)$.

The analogue of Figure 2 under mixed normal error distribution is given in Figure 3, from which we see that the pointwise CI is well-performed even in the presence of small proportion outliers.

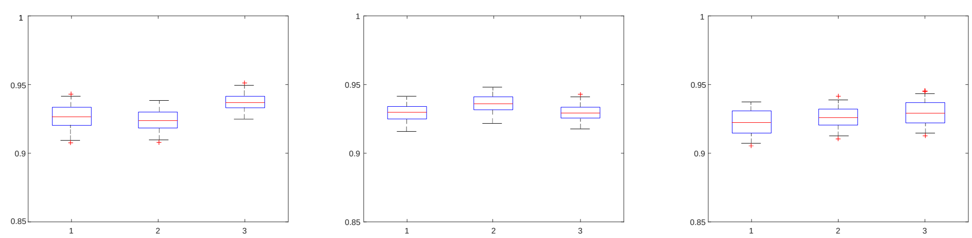
In addition, for the mixed normal error distribution, Figure 4 investigates AECPs of component functions under more general sampling plan, i.e., sparse observations for some subjects and dense observations for other subjects. Specifically speaking, we generate 30 subjects, the first r% subjects with sparse observations ($m_1 = 10$) and the last $(1 - r)$% subjects with dense observation ($m_2 = 30$). Here, we take $r = 1/2$, $1/3$ and $2/3$. The result shows that the AECPs



(a) AECP for α_0 (b) AECP for α_1 (c) AECP for β_1

Figure 3: Boxplots for average experience coverage probability (AECP) based on 200 sampling under mixed normal error distribution. 1, $(n, m) = (20, 30)$; 2, $(n, m) = (30, 20)$; 3, $(n, m) = (60, 10)$.

of component functions remain acceptable even under the mixture sampling plan and small proportion outliers.



(a) AECP for α_0 (b) AECP for α_1 (c) AECP for β_1

Figure 4: AECPs for component functions α_0, α_1 and β_1 under mixture sampling plan: the ratio of sparse ($m_1 = 10$) and dense ($m_2 = 30$) observations is 1:1 for 1, 1:2 for 2 and 2:1 for 3.

Note that he bivariate function $g(t, x) = \alpha_1(t)\beta_1(x)$ can be estimated by $\hat{g}(t, x) = \hat{\alpha}_1(t)\hat{\beta}_1(x)$, Figure 5 compares the estimated surfaces of g under different loss function with the true surface. Obviously, huber estimator is nearest to the true surface, while least squares estimator is the worst, and median estimator is in-between.

Finally, to investigate the asymptotic properties of three-step M-estimators, we take $n = 20, 40$ and $m = 20, 30$. For different combinations of (n, m) and two kinds of intra-subject covariance structure, Table 2 compares the AMSE of three-step M-estimators with different loss functions under normal error distribution and mixed normal error distribution, and Table

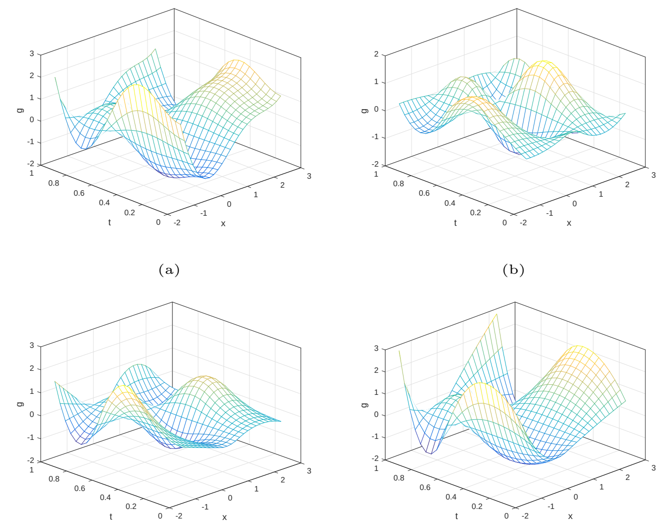


Figure 5: The estimated surfaces of $g(t,x) = \alpha_1(t)\beta_1(x)$ under different loss functions. (a) true surface; (b) least-squares-based estimation; (c) median-based estimation; (d) huber-based estimation.

3 is the analogue for the $0.2 \times t(1)$ and $0.5 \times t(2)$ error distributions. Note that for each given pair (n, m), the performance of the three-step M-estimator is similar to that presented in Table 1. Moreover, as the total number of observations grows, the AMSEs of three-step M-estimators decrease with normal error distribution, no matter which loss function is used. For non-normal error distribution, the estimators based upon robust loss functions ρ_2 and ρ_3 decrease, however, the least square estimators haven't significant improvements.

S1.2 Numerical Study of Model Identification Procedure

In this subsection, we will investigate the finite-sample performance of the proposed model identification procedure. A VCAM with repeated measurements is given by

$$y_{ij} = \alpha_0(t_{ij}) + \alpha_1(t_{ij})\beta_1(x_{ij1}) + 5\beta_2(x_{ij2}) + 3\alpha_3(t_{ij})x_{ij3} + w_i(t_{ij}) + e_{ij},$$

where t_{ij}, x_{ij1}, w_i, α_0, α_1, and the random noise e_{ij} are the same as in Example 1 in the manuscript, $x_{ij2} = t_{ij}^2 + \zeta_{ij}$ with ζ_{ij} independently drawn from $N(0, 0.5)$, and x_{ij3} are indepen-

Table 1: AMSE based on 500 Monte Carlo replications, with $(n, m) = (30, 20)$.

	$\theta = 0$			$\theta = 0.5$		
Est	α_0	α_1	β_1	α_0	α_1	β_1
			Normal Error			
Ls	0.0104(0.0035)	0.0153(0.0062)	0.0108(0.0052)	0.0061(0.0026)	0.0158(0.0034)	0.0095(0.0037)
Ls-O	0.0100(0.0034)	0.0122(0.0045)	0.0058(0.0024)	0.0056(0.0025)	0.0147(0.0028)	0.0080(0.0034)
Lad	0.0119(0.0047)	0.0177(0.0079)	0.0145(0.0083)	0.0084(0.0037)	0.0179(0.0049)	0.0129(0.0064)
Lad-O	0.0115(0.0045)	0.0144(0.0061)	0.0075(0.0037)	0.0074(0.0032)	0.0163(0.0042)	0.0104(0.0054)
Hub	0.0104(0.0036)	0.0152(0.0060)	0.0101(0.0046)	0.0061(0.0026)	0.0158(0.0033)	0.0090(0.0035)
Hub-O	0.0101(0.0034)	0.0123(0.0045)	0.0058(0.0024)	0.0057(0.0025)	0.0147(0.0029)	0.0078(0.0033)
			Mixed Normal Error			
Ls	0.0910(0.0593)	0.9983(1.4442)	0.5302(0.6731)	0.0803(0.0534)	0.9467(1.4342)	0.5239(0.6722)
Ls-O	0.0809(0.0591)	0.0755(0.0574)	0.0814(0.0859)	0.0755(0.0565)	0.0732(0.0547)	0.0797(0.0848)
Lad	0.0143(0.0073)	0.0231(0.0130)	0.0201(0.0174)	0.0098(0.0049)	0.0158(0.0090)	0.0250(0.0248)
Lad-O	0.0136(0.0069)	0.0162(0.0081)	0.0101(0.0060)	0.0095(0.0047)	0.0138(0.0062)	0.0115(0.0062)
Hub	0.0141(0.0072)	0.0193(0.0089)	0.0177(0.0141)	0.0081(0.0042)	0.0135(0.0071)	0.0220(0.0260)
Hub-O	0.0133(0.0067)	0.0160(0.0078)	0.0102(0.0051)	0.0080(0.0042)	0.0121(0.0062)	0.0095(0.0060)
			$0.5 \times t(2)$ Error			
Ls	0.0324(0.0316)	0.1650(0.6698)	0.1287(0.4288)	0.0316(0.0282)	0.1269(0.5516)	0.1020(0.3630)
Ls-O	0.0300(0.0312)	0.0433(0.1317)	0.0342(0.0802)	0.0306(0.0301)	0.0477(0.1453)	0.0346(0.0859)
Lad	0.0119(0.0063)	0.0181(0.0123)	0.0191(0.0159)	0.0127(0.0063)	0.0193(0.0122)	0.0171(0.0152)
Lad-O	0.0116(0.0056)	0.0144(0.0076)	0.0158(0.0083)	0.0122(0.0061)	0.0149(0.0093)	0.0135(0.0078)
Hub	0.0116(0.0055)	0.0178(0.0151)	0.0175(0.0175)	0.0129(0.0057)	0.0190(0.0122)	0.0166(0.0141)
Hub-O	0.0107(0.0049)	0.0141(0.0080)	0.0141(0.0113)	0.0122(0.0051)	0.0155(0.0088)	0.0136(0.0115)
			$0.2 \times t(1)$ Error			
Ls	30.0323(233.7587)	1.3838(1.2394)	71.6044(572.1458)	30.0769(233.8998)	1.2757(1.2426)	71.2963(570.4998)
Ls-O	12.9845(59.5655)	0.5963(0.6420)	2.9933(16.6641)	13.0272(59.8353)	0.5679(0.6370)	3.0022(16.7430)
Lad	0.0123(0.0058)	0.0220(0.0109)	0.0251(0.0188)	0.0172(0.0083)	0.0127(0.0068)	0.0183(0.0082)
Lad-O	0.0119(0.0058)	0.0175(0.0098)	0.0129(0.0085)	0.0155(0.0075)	0.0120(0.0060)	0.0152(0.0150)
Hub	0.0146(0.0068)	0.0251(0.0130)	0.0222(0.0133)	0.0170(0.0091)	0.0143(0.0072)	0.0184(0.0094)
Hub-O	0.0143(0.0069)	0.0201(0.0102)	0.0117(0.0071)	0.0155(0.0085)	0.0127(0.0065)	0.0131(0.0081)

Table 2: Comparison of AMSE under normal and mixed normal error distributions

(n,m)	Fun	$\theta=0$ Ls	$\theta=0$ Lad	$\theta=0$ Hub	$\theta=0.5$ Ls	$\theta=0.5$ Lad	$\theta=0.5$ Hub
				Normal Error			
(20, 20)	α_0	0.0173(0.0057)	0.0240(0.0094)	0.0168(0.0056)	0.0237(0.0061)	0.0345(0.0113)	0.0242(0.0063)
	α_1	0.0304(0.0117)	0.0304(0.0142)	0.0285(0.0112)	0.0218(0.0117)	0.0210(0.0117)	0.0213(0.0113)
	β_1	0.0290(0.0105)	0.0360(0.0179)	0.0248(0.0087)	0.0201(0.0137)	0.0281(0.0020)	0.0198(0.0134)
(20, 30)	α_0	0.0087(0.0033)	0.0112(0.0050)	0.0087(0.0033)	0.0111(0.0031)	0.0155(0.0055)	0.0116(0.0032)
	α_1	0.0112(0.0043)	0.0150(0.0071)	0.0117(0.0042)	0.0133(0.0049)	0.0160(0.0081)	0.0130(0.0048)
	β_1	0.0167(0.0093)	0.0248(0.0158)	0.0167(0.0081)	0.0199(0.0078)	0.0207(0.0092)	0.0184(0.0066)
(40, 20)	α_0	0.0042(0.0014)	0.0053(0.0022)	0.0042(0.0014)	0.0081(0.0028)	0.0094(0.0042)	0.0082(0.0028)
	α_1	0.0057(0.0016)	0.0089(0.0039)	0.0057(0.0016)	0.0075(0.0029)	0.0099(0.0050)	0.0074(0.0028)
	β_1	0.0124(0.0046)	0.0153(0.0067)	0.0120(0.0043)	0.0148(0.0052)	0.0169(0.0073)	0.0140(0.0045)
(40, 30)	α_0	0.0037(0.0016)	0.0047(0.0020)	0.0037(0.0015)	0.0052(0.0019)	0.0059(0.0024)	0.0045(0.0016)
	α_1	0.0034(0.0016)	0.0051(0.0029)	0.0033(0.0015)	0.0049(0.0024)	0.0062(0.0027)	0.0037(0.0014)
	β_1	0.0117(0.0030)	0.0144(0.0072)	0.0113(0.0025)	0.0124(0.0061)	0.0149(0.0058)	0.0107(0.0036)
				Mixed Normal Error			
(20, 20)	α_0	0.1578(0.1349)	0.0171(0.0069)	0.0154(0.0072)	0.1822(0.1368)	0.0414(0.0162)	0.0376(0.0142)
	α_1	0.5756(0.9397)	0.0391(0.0270)	0.0345(0.0303)	0.5923(0.9605)	0.0434(0.0346)	0.0461(0.0366)
	β_1	0.7638(1.0997)	0.0926(0.0763)	0.0930(0.0734)	0.8290(1.1297)	0.1151(0.1135)	0.1163(0.1001)
(20, 30)	α_0	0.0957(0.0840)	0.0115(0.0059)	0.0107(0.0054)	0.1080(0.1008)	0.0201(0.0079)	0.0167(0.0086)
	α_1	0.9389(1.5330)	0.0148(0.0071)	0.0156(0.0125)	0.9150(1.4540)	0.0206(0.0102)	0.0197(0.0109)
	β_1	0.6587(0.8466)	0.0533(0.0221)	0.0529(0.0030)	0.6794(0.7591)	0.0635(0.0429)	0.0582(0.0546)
(40, 20)	α_0	0.0588(0.0368)	0.0074(0.0031)	0.0072(0.0032)	0.0608(0.0358)	0.0133(0.0056)	0.0148(0.0061)
	α_1	0.5642(1.2187)	0.0129(0.0044)	0.0141(0.0070)	0.6662(1.3521)	0.0161(0.0074)	0.0161(0.0092)
	β_1	0.4627(0.8563)	0.0283(0.0182)	0.0303(0.0262)	0.5628(1.0200)	0.0363(0.0184)	0.0343(0.0234)
(40, 30)	α_0	0.0436(0.0278)	0.0056(0.0027)	0.0050(0.0022)	0.0522(0.0327)	0.0071(0.0034)	0.0099(0.0040)
	α_1	0.4408(1.1603)	0.0071(0.0035)	0.0072(0.0029)	0.4490(1.1607)	0.0077(0.0039)	0.0068(0.0039)
	β_1	0.3491(0.7644)	0.0150(0.0074)	0.0139(0.0064)	0.3604(0.7289)	0.0201(0.0098)	0.0197(0.0112)

Table 3: Comparison of AMSE under $0.2 \times t(1)$ and $0.5 \times t(2)$ error distributions

		$\theta = 0$			$\theta = 0.5$		
(n, m)	Fun	Ls	Lad	Hub	Ls	Lad	Hub
				$0.5 \times t(2)$ Error			
(20, 20)	α_0	0.0728(0.1858)	0.0268(0.0181)	0.0271(0.0165)	0.0760(0.1934)	0.0299(0.0191)	0.0273(0.0176)
	α_1	0.2642(0.7724)	0.0399(0.0275)	0.0483(0.0334)	0.3663(1.0033)	0.0336(0.0238)	0.0334(0.0282)
	β_1	0.2326(0.7268)	0.0760(0.0472)	0.0800(0.0454)	0.2915(0.8157)	0.0810(0.0583)	0.0735(0.0556)
(20, 30)	α_0	0.0345(0.0678)	0.0128(0.0072)	0.0112(0.0068)	0.0467(0.0529)	0.0219(0.0096)	0.0233(0.0104)
	α_1	0.1424(0.5916)	0.0215(0.0098)	0.0212(0.0090)	0.1398(0.5819)	0.0176(0.0073)	0.0180(0.0104)
	β_1	0.1156(0.2772)	0.0358(0.0270)	0.0263(0.0170)	0.1100(0.2415)	0.0340(0.0213)	0.0297(0.0170)
(40, 20)	α_0	0.0296(0.0545)	0.0120(0.0068)	0.0096(0.0053)	0.0314(0.0530)	0.0111(0.0061)	0.0109(0.0061)
	α_1	0.1897(0.7380)	0.0157(0.0081)	0.0138(0.0057)	0.1916(0.7405)	0.0129(0.0051)	0.0115(0.0047)
	β_1	0.1656(0.5006)	0.0343(0.0231)	0.0243(0.0149)	0.1691(0.4944)	0.0271(0.0248)	0.0235(0.0168)
(40, 30)	α_0	0.0293(0.0734)	0.0093(0.0047)	0.0083(0.0042)	0.0292(0.0775)	0.0100(0.0056)	0.0075(0.0037)
	α_1	0.3434(1.0457)	0.0098(0.0043)	0.0086(0.0038)	0.1912(0.7483)	0.0090(0.0045)	0.0073(0.0031)
	β_1	0.2613(0.6881)	0.0194(0.0102)	0.0182(0.0087)	0.1646(0.4902)	0.0231(0.0151)	0.0177(0.0107)
				$0.2 \times t(1)$ Error			
(20, 20)	α_0	152.8847(1296.2)	0.0247(0.0149)	0.0262(0.0140)	152.9221(1296.5)	0.0309(0.0137)	0.0345(0.0156)
	α_1	1.1684(1.0295)	0.0397(0.0199)	0.0449(0.0281)	1.3283(1.0899)	0.0349(0.0191)	0.0470(0.0312)
	β_1	429.7490(3146.3)	0.0445(0.0219)	0.0467(0.0229)	429.6629(3144.1)	0.0471(0.0319)	0.0459(0.0321)
(20, 30)	α_0	10.3490(52.0259)	0.0141(0.0072)	0.0115(0.0061)	6.4800(30.3419)	0.0157(0.0103)	0.0148(0.0068)
	α_1	1.3966(1.2809)	0.0194(0.0117)	0.0186(0.0128)	1.4232(1.3723)	0.0290(0.0140)	0.0298(0.0129)
	β_1	51.9701(370.9491)	0.0432(0.0282)	0.0381(0.0250)	51.5754(367.6787)	0.0423(0.0271)	0.0384(0.0185)
(40, 20)	α_0	44.3016(247.4062)	0.0101(0.0067)	0.0082(0.0041)	44.3580(247.7636)	0.0142(0.0066)	0.0101(0.0043)
	α_1	1.2505(1.3224)	0.0134(0.0060)	0.0170(0.0114)	1.2551(1.3329)	0.0139(0.0057)	0.0151(0.0077)
	β_1	68.2984(422.8558)	0.0306(0.0201)	0.0328(0.0262)	68.2566(422.4329)	0.0248(0.0229)	0.0274(0.0429)
(40, 30)	α_0	8.1547(29.9252)	0.0083(0.0038)	0.0062(0.0028)	8.1437(29.8633)	0.0062(0.0028)	0.0074(0.0033)
	α_1	1.3795(1.0867)	0.0097(0.0052)	0.0106(0.0103)	1.3418(1.0529)	0.0079(0.0042)	0.0099(0.0094)
	β_1	22.3450(78.9040)	0.0255(0.0147)	0.0175(0.0090)	22.3086(78.5702)	0.0163(0.0111)	0.0169(0.0116)

Table 4: Model identification in Example 2 under normal and mixed normal error distributions.

		$\theta = 0$									$\theta = 0.5$								
		AT(%)			VCT(%)			TM(%)			AT(%)			VCT(%)			TM(%)		
(n,m)	Est	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F
Normal Error																			
(20, 20)	Ls	72	0	28	94	0	6	66	0	34	69	0	31	91	0	9	60	0	40
	Lad	55	0	45	76	0	24	40	0	60	65	0	35	63	0	37	36	0	64
	Hub	60	0	40	70	0	30	30	0	70	41	0	59	84	0	16	32	0	68
(20, 30)	LS	95	0	5	94	0	6	90	0	10	98	0	2	97	0	3	96	0	4
	Lad	81	0	19	82	0	18	65	0	35	82	0	18	82	0	18	64	0	36
	Hub	86	0	14	90	0	10	77	0	23	90	0	10	87	0	13	77	0	23
(40, 20)	Ls	95	0	5	96	0	4	91	0	9	100	0	0	99	0	1	97	0	3
	Lad	83	0	17	86	0	14	73	0	27	91	0	9	90	0	10	83	0	17
	Hub	92	0	8	97	0	3	91	0	9	97	0	3	93	0	7	90	0	10
(40, 30)	LS	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
	Lad	98	0	2	99	0	1	97	0	3	95	0	5	98	0	2	93	0	7
	Hub	99	0	1	99	0	1	96	0	4	100	0	0	100	0	0	100	0	0
Mixed Normal Error																			
(20, 20)	Ls	15	0	85	13	0	87	11	0	89	9	0	91	17	0	83	6	0	94
	Lad	68	0	32	69	0	31	48	0	52	81	0	19	67	1	32	53	1	46
	Hub	68	0	32	61	16	23	52	12	36	77	0	23	48	27	25	42	17	41
(20, 30)	Ls	19	0	81	16	0	84	14	0	86	17	0	83	15	0	85	10	0	90
	Lad	75	0	25	79	0	21	59	0	41	82	0	18	76	0	24	63	0	37
	Hub	83	0	17	70	10	20	60	7	33	85	0	15	77	7	16	70	10	20
(40, 20)	Ls	11	0	89	19	0	81	9	0	91	14	0	86	20	0	80	11	8	81
	Lad	82	0	18	87	0	13	70	0	30	88	0	12	86	0	14	75	0	25
	Hub	89	0	11	76	5	19	70	3	27	94	0	6	83	5	12	78	5	17
(40, 30)	Ls	16	0	84	15	0	85	8	0	92	13	0	87	12	0	88	7	0	93
	Lad	96	0	4	94	0	6	92	0	8	98	0	2	95	0	5	90	0	10
	Hub	99	0	1	91	0	9	90	0	10	96	0	4	92	0	8	92	0	8

dent copies from $N(0, 1.5^2)$. Other component functions are given by $\alpha_3(t) = t^2 / \int_0^1 t^2 \mathrm{d}t$,

$$\beta_2(x_2) = 0.2 \sin(\pi x_2/2) - \mathrm{E}[0.2 \sin(\pi X_2/2)], \quad \text{and}$$

$$\beta_1(x_1) = 5 \sin(\pi x_1/2) + 2x_1(1 - x_1) - \mathrm{E}[5 \sin(\pi X_1/2) + 2X_1(1 - X_1)].$$

Based on 200 Monte Carlo replications, we compare the performance of the proposed model identification procedure for independent and dependent intra-subject covariance structures and three kinds of loss function used in Example 1. Table 4 lists the percentages of correct fitting (C-F), over-fitting (O-F), and under-fitting (U-F) in the identification of additive terms (AT), varying-coefficient terms (VCT), and true model (TM) for normal error and mixed normal error distributions. The counterparts for the heavy-tail error distributions of $0.2 \times t(1)$ and $0.5 \times t(2)$ are given in Table 5. From the obtained results, we notice that under normal error distribution, all of the percentage of correctly identified additive terms, varying-coefficient terms, and true models increases as the total number of observations increases, regardless of which loss function is adopted. In the case of small proportion outliers, the power of model identification increases as the number of observations nm increases if we use robust loss function ρ_2 and ρ_3. However, the least-square-based identification procedure exhibits very poor performance, and no significant improvement is obtained by increasing the total number of observations. It is expected since mean regression method is sensitive to outliers, which greatly influence the power of model identification. Also, the influence of the intra-subject covariance structure on the power of model identification is insignificant.

S1.3 Comparison with existing method

Note that the two-step spline estimation method proposed by Zhang and Wang (2015) is applicable when the covariates are dependent on subjects but independent of observation time. Under this case, we compare the average MSE(AMSEs) and its standard deviation between the two types estimators based on 500 Monte Carlo replications. We consider normal error

Table 5: Model identification in Example 2 under $0.2 \times t(1)$ and $0.5 \times t(2)$ error distributions.

		$\theta = 0$									$\theta = 0.5$								
		AT(%)			VCT(%)			TM(%)			AT(%)			VCT(%)			TM(%)		
(n, m)	Est	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F	C-F	O-F	U-F
										$0.5 \times t(2)$ Error									
	Ls	19	0	81	15	0	85	11	0	89	18	0	82	15	0	85	14	0	86
(20, 20)	Lad	64	0	36	76	0	24	50	0	50	73	0	27	78	0	22	56	0	44
	Hub	65	0	35	75	7	18	55	6	39	72	0	28	77	6	17	59	4	37
	Ls	13	0	87	16	0	84	12	0	88	13	0	87	10	0	90	8	0	92
(20, 30)	Lad	74	0	26	85	0	15	62	0	38	82	0	18	80	0	20	59	0	41
	Hub	71	0	29	84	2	14	62	1	37	78	0	22	85	0	15	69	0	31
	Ls	16	0	84	12	0	88	10	0	90	15	0	85	13	0	87	9	0	91
(40, 20)	Lad	90	0	10	91	0	9	81	0	19	92	0	8	93	0	7	89	0	11
	Hub	90	0	10	90	1	9	82	1	17	90	0	10	90	0	10	82	0	18
	Ls	18	0	82	17	0	83	15	0	85	20	0	80	19	0	81	11	0	89
(40, 30)	Lad	96	0	4	92	0	8	92	0	8	97	0	3	96	0	4	94	0	6
	Hub	99	0	1	94	0	6	94	0	6	95	0	5	92	0	8	90	0	10
										$0.2 \times t(1)$ Error									
	Ls	18	0	82	20	0	80	9	0	91	18	0	82	13	0	87	9	0	91
(20, 20)	Lad	74	0	26	60	0	40	43	0	57	65	0	35	66	0	34	41	0	59
	Hub	76	0	24	67	7	26	50	7	43	69	0	31	65	10	25	42	8	50
	Ls	12	0	88	16	0	84	12	0	88	14	0	86	19	0	81	13	0	87
(20, 30)	Lad	86	0	14	71	0	29	62	0	38	79	0	21	82	0	18	67	0	33
	Hub	89	0	11	75	3	22	69	3	28	83	0	17	75	0	25	74	0	26
	Ls	15	0	85	13	0	87	13	0	87	17	0	83	15	0	85	12	0	88
(40, 20)	Lad	92	0	8	82	0	18	79	0	21	89	0	11	96	0	4	87	0	13
	Hub	96	0	4	87	1	12	85	1	14	95	0	5	88	1	11	88	0	12
	Ls	17	0	83	15	0	85	10	0	90	13	0	87	11	0	91	11	0	91
(40, 30)	Lad	94	0	6	96	0	4	90	0	10	93	0	7	100	0	0	93	0	7
	Hub	100	0	0	97	0	3	97	0	3	98	0	2	96	0	4	94	0	6

distribution $N(0, 0.2)$ and mixed normal error distribution $0.95N(0, 0.2) + 0.05N(0, 12.5^2)$, and generate $n = 40$ subjects with sparse observations ($m = 10$, 20) and dense observations ($m = 40$). Taking $p = 1$, the covariates X_i i.i.d. follow $U([-0.5, 0.5])$, the observation time points $T_{ij}(j = 1, ..., m)$ are equidistant on $[0, 1]$, and the subject-specific random trajectory w_i and univariate component functions are same with Example 1 in the manuscript. Table 6 shows that our estimators are superior to the two-step estimators of Zhang and Wang (2015) for sparse data and a small proportion outliers, while for dense data with normal error distribution, the difference between them is insignificant.

S1.4 Example 2 (Continued)

For the CD4 dataset considered in the manuscript, Figure 6 presents the estimated surfaces of bivariate functions $g_k(t, x_k) = \alpha_k(t)\beta_k(x_k)$, $k = 1, 2$.

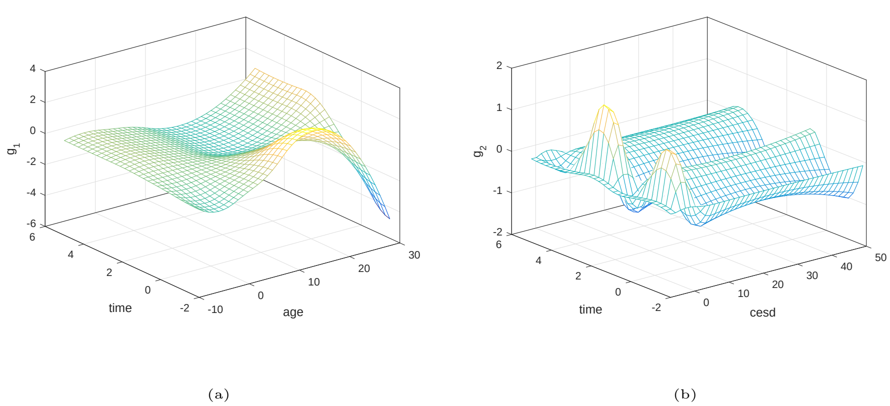


Figure 6: Estimated surfaces of CD4 dataset. (a) estimates $g_1 = \alpha_2(\text{time})\beta_2(\text{age})$; (b) estimates $g_2 = \alpha_3(\text{time})\beta_3(\text{cesd})$.

For a comparison with existing literature, we now analyze another well-known CD4 dataset from R package ‘timereg’ (Thomas (2019)). It is a subset of the Multicenter AIDS Cohort Study, which contains 1187 observations from 283 homosexual men infected with HIV during

Table 6: Comparison of AMSEs between three-step M-estimators and Zhang's estimators

Fun	$m = 10$		$m = 20$		$m = 40$	
	Ours	Zhang's	Ours	Zhang's	Ours	Zhang's
			Normal Error			
α_0	0.0149	0.0149	0.0058	0.0058	0.0027	0.0026
	(0.0053)	(0.0054)	(0.0023)	(0.0023)	(0.0008)	(0.0008)
α_1	0.0609	0.5809	0.0309	0.1580	0.0207	0.0800
	(0.0245)	(0.4871)	(0.0139)	(0.0228)	(0.0070)	(0.0347)
β_1	0.0240	0.1498	0.0103	0.1093	0.0146	0.0443
	(0.0086)	1.9450)	(0.0052)	(1.0199)	(0.0065)	(0.0135)
			Mixed Normal Error			
α_0	0.0580	0.1646	0.0334	0.0770	0.0274	0.0738
	(0.0205)	(0.0991)	(0.0049)	(0.0479)	(0.0080)	(0.0399)
α_1	0.1707	1.3124	0.1671	0.8528	0.1397	0.4184
	(0.0670)	(1.2377)	(0.0810)	(1.8620)	(0.0207)	(0.2009)
β_1	0.0354	0.1313	0.0422	0.1070	0.0118	0.1144
	(0.0203)	(1.3137)	(0.0103)	(1.3532)	(0.0042)	(0.0085)

the study period between 1984 and 1991. The time variable t_{ij} is the time (in years) of the j-th measurement of the i-th individual after HIV infection; the response Y_{ij} is the i-th individual's CD4 percent measured at time t_{ij}. Fan and Zhang (2000a); Huang, Wu and Zhou (2002, 2004) have analyzed this dataset using a VCM, which is a special case of our VCAM. They adopted three covariates: X_1 the smoking status of individual after his infection; X_2 the centered age at HIV infection, and X_3 the centered pre-infection CD4 percent. Now, we compare a VCAM and a VCM containing the three covariates. Note that X_1 is attribute variable and covariates are all time-invariant, we can write the two model as below:

$$\text{VCM}: \quad y_{ij} = \alpha_0(t_{ij}) + \alpha_1(t_{ij})x_{i,1} + \alpha_2(t_{ij})x_{i,2} + \alpha_3(t_{ij})x_{i,3} + w_{ij} + e_{ij}, \tag{S1.1}$$

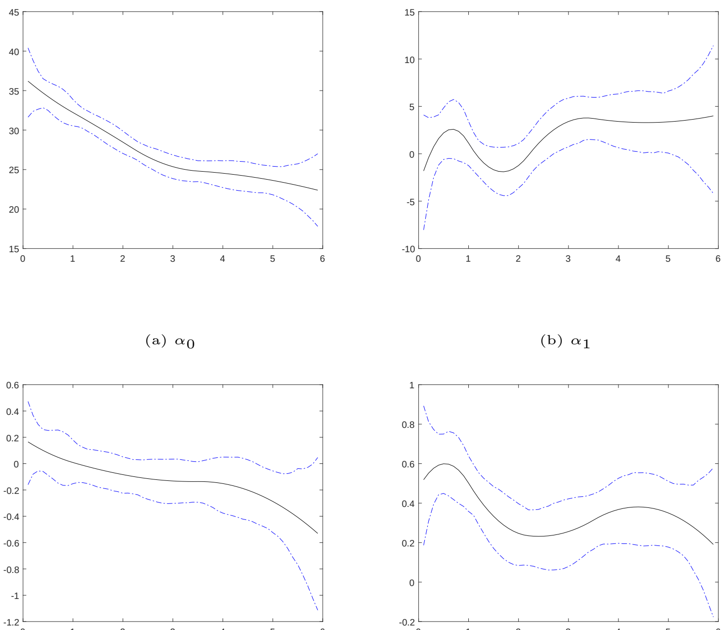
and

$$\text{VCAM}: \quad y_{ij} = \alpha_0(t_{ij}) + \alpha_1(t_{ij})x_{i,1} + \alpha_2(t_{ij})\beta_2(x_{i,2}) + \alpha_3(t_{ij})\beta_3(x_{i,3}) + w_{ij} + e_{ij} \tag{S1.2}$$

For the VCM (S1.1), Figure 7 gives the fitted curves (solid lines) of varying-coefficient functions and 95% CI (dash-dotted lines). For the VCAM (S1.2), we select the optimal number of interior knots $(\hat{\hbar}_{\mathrm{C}}, \hat{\hbar}_{\mathrm{A}}, \hat{K}_{\mathrm{C}}, \hat{K}_{\mathrm{A}}) = (2, 2, 3, 6)$ and optimal tuning parameter $(\hat{\lambda}_1, \hat{\lambda}_2) = (0.01, 0.01)$. Employing the model identification procedure, we found α_2 and α_3 are non-constant $(\|\hat{\eta}_2^{\tau} M_2\|_{L_2} = 0.2339$, $\|\hat{\eta}_3^{\tau} M_3\|_{L_2} = 0.3183)$, β_2 and β_3 are nonlinear $(\|\hat{\eta}_2 F_2\|_{L_2} = 0.2591$, $\|\hat{\eta}_3 F_3\|_{L_2} = 0.3165)$. Figure 8 presents the three-step spline estimators (solid lines) of univariate component functions in VCAM (S1.2), and 95% CI (dash-dotted lines).

Note that the fitted curves of varying-coefficient functions have similar shapes under the two models. However, except α_0 the ranges of fitted curves are different since the difference of function form of covariate effect. From the fitted curves of additive component functions in VCAM (S1.2), we can see the the linear covariates effects is not rational in the whole study period.

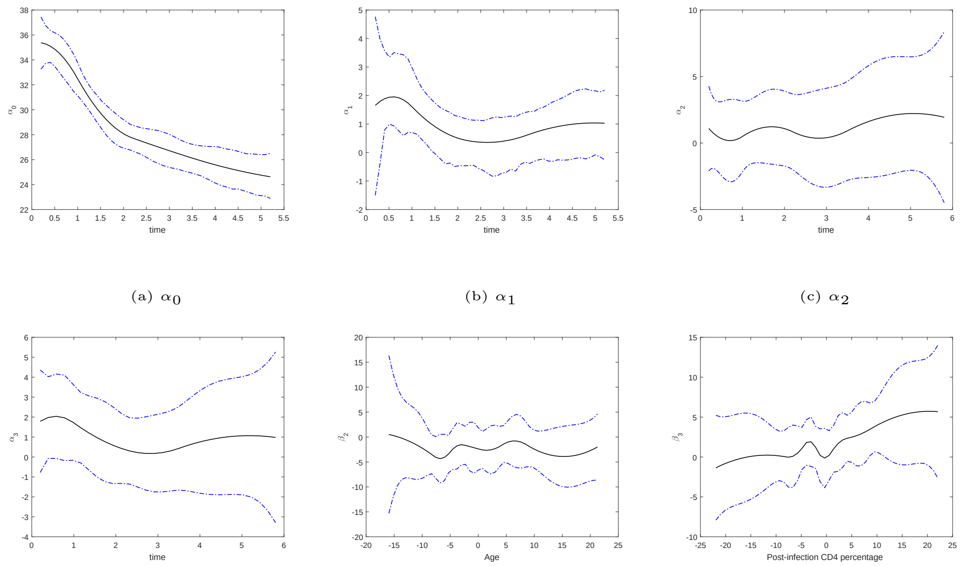
In addition, the bivariate time-varying covariates effects $g_1(t, x) = \alpha_2(t)\beta_2(x)$ and $g_2(t, x) = \alpha_3(t)\beta_3(x)$ in VCAM (S1.2), and $g_1(t, x) = \alpha_2(t)x$ and $g_2(t, x) = \alpha_3(t)x$ in VCM (S1.1) are estimated in Figure 9, which implies the VCAM (S1.2) provides more detailed information of



(a) α_0 (b) α_1

(c) α_2 (d) α_3

Figure 7: Spline estimators (solid lines) in VCM (S1.1), and 95% CI (dash-dotted lines).



(a) α_0 (b) α_1 (c) α_2

(d) α_3 (e) β_2 (f) β_3

Figure 8: Three-step spline estimators (solid lines) in VCAM (S1.2), and 95% CI (dash-dotted lines).

change points due to the nonlinear covariate effects.

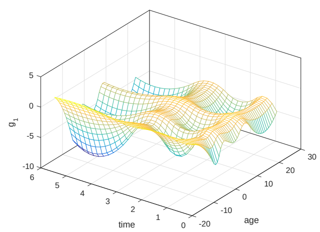
Compared with the VCM (S1.1) of Huang, Wu and Zhou (2002, 2004), the residual sum of squares (RSS) of VCAM (S1.2) decreases by 3% and the multiple determination coefficient (R^2) increases by 12%. Therefore, VCAM (S1.2) is more suitable for this real-life data.

S1.5 Example 3 (Continued)

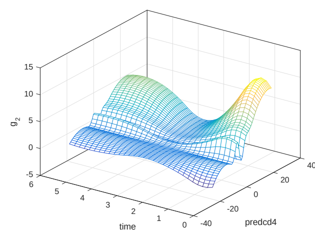
Figure 10 gives the estimated bivariate functions $g_k(t, x_k) = \alpha_k(t)\beta_k(x_k), k = 1, 2, 3$ for the ADNI data.

S1.6 Cigarette Data

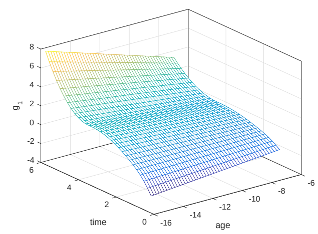
Example 3. We continue the analysis of cigarette data referred in Section 1 in the manuscript. Following Baltagi and Li (2004), we adopt covariates X_1: logarithm of the average real retail price of cigarettes; X_2: logarithm of the real disposable income per capita, and response Y:



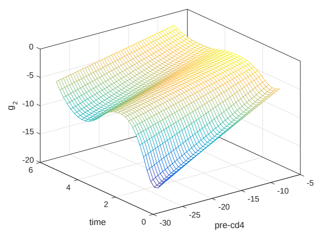
(a) g_1 in VCAM



(b) g_2 in VCAM

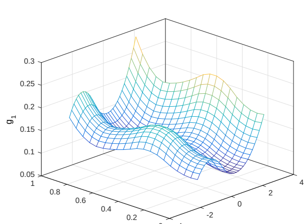


(a) g_1 in VCM

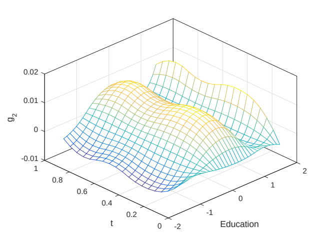


(b) g_2 in VCM

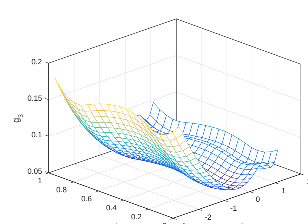
Figure 9: The estimated surfaces of time-varying covariates effects $g_1(t, x) = \alpha_2(t)\beta_2(x)$ and $g_2(t, x) = \alpha_3(t)\beta_3(x)$ in VCAM (S1.2), whilst $g_1(t, x) = \alpha_2(t)x$ and $g_2(t, x) = \alpha_3(t)x$ in VCM (S1.1).



(a)



(b)



(c)

Figure 10: Estimated bivariate surfaces for ADNI data. (a) estimates $g_1 = \alpha_1(t)\beta_1(\text{Age})$; (b) estimates $g_2 = \alpha_2(t)\beta_2(\text{Education})$; (c) estimates $g_3 = \alpha_3(t)\beta_3(\text{MMSE})$;

logarithm of the sales of cigarettes (packs of cigarettes per capita).

We use the following VCAM

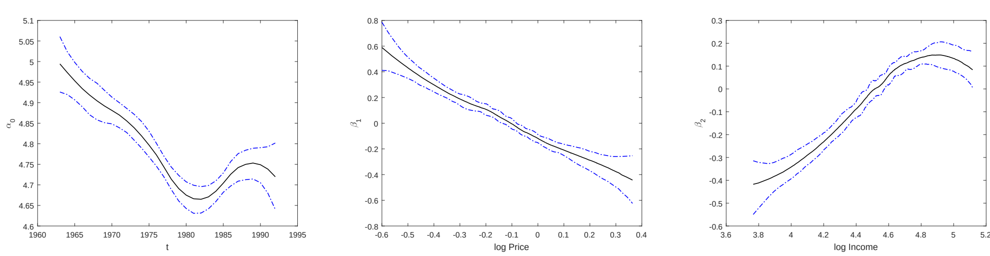
$$y_{it} = \alpha_0(t) + \sum_{k=1}^{2} \alpha_k(t)\beta(x_{itk}),\ i = 1, ..., 46;\ t = 1, ..., 30.$$

Under huber loss function given in Example 1, we obtain that the optimal knots in Step I estimation are $(\hat{\hbar}_{\mathrm{C}}, \hat{\hbar}_{\mathrm{A}}) = (2, 2)$, and smoothing tuning parameters in model identification procedure are $(\hat{\lambda}_1, \hat{\lambda}_2) = (6.31, 0.01)$, We then obtain the penalized estimators and conclude both α_1 and α_2 are time-invariant. In a word, a more parsimonious model is given by

$$y_{it} = \alpha_0(t) + \beta_1(x_{it1}) + \beta_2(x_{it2}),\ i = 1, ..., 46;\ t = 1, ..., 30. \tag{S1.3}$$

The estimated component functions are given in Figure 11, from which we see that:

- α_0 decreases before 1980, and ascends until 1990, and then decreases;
- Cigarettes consumption decreases as the Cigarettes price increases;
- Cigarettes consumption increases as the income grows until X_2 is larger than 4.8.



(a) α_0 (b) β_1 (c) β_2

Figure 11: Reduced model (S1.3) for cigarette data set. Solid line: three-step estimator, and dash-dot lines: 95% CI.

S2 Proofs

We start with the properties of B-spline basis. Let $\{b_1, ..., b_L\}$ be standardized version of B-spline basis defined on $[a, b]$. Then, it holds that:

(1) $b_l(x) \geq 0$ for any $x \in [a, b]$, $1 \leq l \leq L$;

(2) $\sum_{l=1}^{L} b_l(x) = 1$ for any $x \in [a, b]$;

(3) for any vector $\alpha = (\alpha_1, ..., \alpha_L)^\tau$,

$$|\alpha|^2 \preceq \int_0^1 \Big\{ \sum_{l=1}^{L} \alpha_l b_l(x) \Big\}^2 \mathrm{d}x \preceq |\alpha|^2. \tag{S2.1}$$

S2.1 A Proposition

To prove the main results, we start with the following proposition, which gives the convergence rates of the initial estimators of varying-coefficient component functions. Let $\underline{h} = \hbar_{\mathrm{A}} \wedge \hbar_{\mathrm{C}}$ and $\bar{h} = \hbar_{\mathrm{A}} \vee \hbar_{\mathrm{C}}$ be the minimum and maximum of $\hbar_{\mathrm{A}}$ and $\hbar_{\mathrm{C}}$, respectively, and $\bar{N}_{\mathrm{H}} = \big(\frac{1}{n} \sum_{i=1}^{n} n_i^{-1}\big)^{-1}$ is the harmonic average of sequence $\{n_i\}$.

Proposition 1. *Under Assumptions A1–A5, M1-M2 or N1-N2, if* $\bar{h}^4 = o(n\bar{N}_{\mathrm{H}})$, $\bar{h}^{2r}/n \to C_1$, $\bar{h}^{2r+2}/(n\bar{N}_{\mathrm{H}}) \to C_2$, *and* $\bar{h}^2/\bar{N}_{\mathrm{H}} \to C_3$ *as* $n \to \infty$, *where* $0 \leq C_1 < \infty$, $0 \leq C_2, C_3 \leq \infty$. *Then, we obtain the convergence rates*

$$\|\hat{\alpha}_{0,\mathrm{I}} - \alpha_0\|_{L_2}^2 = O_p\Big(\underline{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big), \quad \textit{and}$$

$$\|\hat{\alpha}_{k,\mathrm{I}}(t|t_{k0}, x_{k0}) - \alpha_k(t)\|_{L_2}^2 = O_p\Big(\underline{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big)$$

in the L_2 norm sense, and

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big[\hat{\alpha}_{0,\mathrm{I}}(t_{ij}) - \alpha_0(t_{ij})\big]^2 = O_p\Big(\underline{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big), \quad \textit{and}$$

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big[\hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0}) - \alpha_k(t_{ij})\big]^2 = O_p\Big(\underline{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big)$$

in the MSE sense.

Remark 1. From Proposition 1, we see that if $\bar{N}_{\rm H}/n^{\frac{1}{r}} \to 0$, and $\bar{h} \asymp \big(n\bar{N}_{\rm H}\big)^{\frac{1}{2r+2}}$, then $n^{-1} = o\big(\bar{h}/\big(n\bar{N}_{\rm H}\big)\big)$ and the asymptotic variance obtains the rate of bivariate nonparametric function. When $\bar{N}_{\rm H}/n^{\frac{1}{r}} \to C$ and $\bar{h} \asymp n^{\frac{1}{2r}}$ yield $n^{-1} \asymp \bar{h}/\big(n\bar{N}_{\rm H}\big)$; and $\bar{N}_{\rm H}/n^{\frac{1}{r}} \to \infty$ and $\bar{h} = o(n^{\frac{1}{2r}})$ imply $\bar{h}/\big(n\bar{N}_{\rm H}\big) = o(n^{-1})$, that is, the asymptotic variance has a parametric rate. Thus, we can split data as sparse or dense according to whether $\bar{N}_{\rm H}/n^{\frac{1}{r}} \to 0$ or $\bar{N}_{\rm H}/n^{\frac{1}{r}} \to C$, where $0 < C \leq \infty$. The result is slightly different from Remark 5 in the manuscript since we now estimate bivariate nonparametric function and require larger sample size.

According to Corollary 6.21 and Theorem 12.7 of Schumaker (1981) and Assumption (A3), there exists positive constants $D_0, ..., D_p$, such that $\alpha_0(t) = \tilde{\alpha}_0(t) + R_0(t) = \tilde{\gamma}_0^\tau \mathbf{b}_{\rm C}(t) + R_0(t)$ and $g_k(t, x_k) = \tilde{g}_k(t, x_k) + R_k(t, x_k) = \tilde{\gamma}_k^\tau \mathcal{T}_k(t, x_k) + R_k(t, x_k)$ satisfy

$$\sup_{t\in[0,1]} |R_0(t)| \leq D_0 \hbar_{\rm C}^{-r} \quad \text{and} \quad \sup_{(t,x_k)} |R_k(t, x_k)| \leq D_k(\hbar_{\rm C}^{-r} + \hbar_{\rm A}^{-r}), \quad \text{for } k = 1, ..., p. \tag{S2.2}$$

Denote $R_{ij} = R_0(t_{ij}) + \sum_{k=1}^p R_k(t_{ij}, x_{ijk})$, then (S2.2) implies

$$\max_{i,j} |R_{ij}| \leq D_1^* \hbar_{\rm C}^{-r} + D_2^* \hbar_{\rm A}^{-r} \leq D^* \underline{h}^{-r}, \tag{S2.3}$$

where $D^* = D_1^* \vee D_2^*$.

Let $\pi(t, \mathbf{x}) = \{\mathbf{b}_{\rm C}^\tau(t), \mathcal{T}_1^\tau(t, x_1), ..., \mathcal{T}_p^\tau(t, x_p)\}^\tau$, $\pi_{ij} = \pi(t_{ij}, \mathbf{x}_{ij})$ and $\tilde{\pi}_{ij} = S_n^{-1}\pi_{ij}$, where $S_n^2 = \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \pi_{ij}\pi_{ij}^\tau$. Denote $\boldsymbol{\zeta} = S_n(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})$, $\hat{\boldsymbol{\zeta}} = S_n(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})$. Then, the minimizing problem (2.3) can be rewritten as

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{\gamma}} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \rho(y_{ij} - \pi_{ij}^\tau \boldsymbol{\gamma}) &= \operatorname*{argmin}_{\boldsymbol{\zeta}} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \rho(\varepsilon_{ij} - \tilde{\pi}_{ij}^\tau \boldsymbol{\zeta} + R_{ij}) \\ &= \operatorname*{argmin}_{\boldsymbol{\zeta}} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\rho(\varepsilon_{ij} - \tilde{\pi}_{ij}^\tau \boldsymbol{\zeta} + R_{ij}) - \rho(\varepsilon_{ij} + R_{ij})\right].\end{aligned}$$

Denote $\Gamma_n(\boldsymbol{\zeta})$ be the objective function above, $\Phi_n(\boldsymbol{\zeta}) = {\rm E}[\Gamma_n(\boldsymbol{\zeta})|\mathcal{J}]$, and $\Delta_n(\boldsymbol{\zeta}) = \Gamma_n(\boldsymbol{\zeta}) - \Phi_n(\boldsymbol{\zeta}) + \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\pi}_{ij}^\tau \boldsymbol{\zeta} \phi(\varepsilon_{ij})$. Then

$$\Gamma_n(\boldsymbol{\zeta}) = \Phi_n(\boldsymbol{\zeta}) - \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\pi}_{ij}^\tau \boldsymbol{\zeta} \phi(\varepsilon_{ij}) + \Delta_n(\boldsymbol{\zeta}). \tag{S2.4}$$

The following lemmas are useful to prove Proposition 1.

Lemma 1. *Under Assumption A1 and A2, there exists positive constants* L_1 *and* L_2*, it holds that* $L_1 I \le S_n^2/n \le L_2 I$*, except on an event whose probability tends to zero, where* I *is* l*-order identity matrix, with* $l = J_{\mathrm{C}} + pJ_{\mathrm{C}}J_{\mathrm{A}}$*,* $J_{\mathrm{C}} = q + \hbar_{\mathrm{C}}$*, and* $J_{\mathrm{A}} = q + \hbar_{\mathrm{A}} - 1$*.*

Proof. Let $\mathcal{G} = \{g(t,\mathbf{x}) = \boldsymbol{\gamma}^\tau \pi(t,\mathbf{x}), \boldsymbol{\gamma} = (\gamma_0^\tau, ..., \gamma_p^\tau)^\tau \in \mathcal{R}^l\}$ be a family of functions defined on $t \in [0,1]$, and $\mathbf{x} = (x_1, ..., x_p)^\tau \in \prod_{k=1}^p [a_k, b_k]$. For any $g^{(1)}, g^{(2)} \in \mathcal{G}$, define the theoretical inner product and empirical inner product are $\langle g^{(1)}, g^{(2)} \rangle = \mathrm{E}[g^{(1)}(T,\mathbf{X}) g^{(2)}(T,\mathbf{X})]$ and

$$\langle g^{(1)}, g^{(2)} \rangle_n = \frac{1}{n}\sum_{i=1}^n \frac{1}{n_i}\sum_{j=1}^{n_i} g^{(1)}(t_{ij}, \mathbf{x}_{ij}) g^{(2)}(t_{ij}, \mathbf{x}_{ij}).$$

The induced theoretical norm and empirical norm are denoted as $\|g\|$ and $\|g\|_n$, respectively.

For any $g = \boldsymbol{\gamma}^\tau \pi(t,\mathbf{x}) \in \mathcal{G}$, according to Assumption (A1), (A2) and Lemma 1 of Stone (1985), we have

$$\begin{aligned}
\|g\|^2 &= \mathrm{E}\big[\mathrm{E}\big[\{\gamma_0^\tau \mathbf{b}_{\mathrm{C}}(T) + \sum_{k=1}^p \gamma_k^\tau \mathcal{T}_k(T, X_k)\}^2 | T\big]\big] \\
&= \int_0^1 \mathrm{E}\big[\big\{\gamma_0^\tau \mathbf{b}_{\mathrm{C}}(T) + \sum_{k=1}^p \gamma_k^\tau \mathcal{T}_k(T, X_k)\big\}^2 \big| T = t\big] f_T(t) \mathrm{d}t \\
&\ge d_1 \int_0^1 \big\{(\gamma_0^\tau \mathbf{b}_{\mathrm{C}}(t))^2 + \sum_{k=1}^p \mathrm{E}[\{\gamma_k^\tau \mathcal{T}_k(t, X_k)\}^2 | T = t]\big\} f_T(t) \mathrm{d}t,
\end{aligned}$$

where d_1 is some positive constant.

On the other hand, there exists a positive constant d_2 ($> d_1$) such that

$$\|g\|^2 \le d_2 \int_0^1 \big\{(\gamma_0^\tau \mathbf{b}_{\mathrm{C}}(t))^2 + \sum_{k=1}^p \mathrm{E}[\{\gamma_k^\tau \mathcal{T}_k(t, X_k)\}^2 | T = t]\big\} f_T(t) \mathrm{d}t.$$

By Assumption (A2) and (S2.1), we obtain that

$$\begin{aligned}
&\int_0^1 \mathrm{E}[\{\gamma_k^\tau \mathcal{T}_k(t, X_k)\}^2 | T = t] f_T(t) \mathrm{d}t \\
&= \int_0^1 \int_0^1 \Big[\sum_{l_1=1}^{J_\mathrm{C}} \sum_{l_2=1}^{J_\mathrm{A}} b_{kl_1}(t) B_{kl_2}(x_k) \gamma_{k,l_1 l_2} \Big]^2 f_{X_k|T}(x_k|t) f_T(t) \mathrm{d}x_k \mathrm{d}t \\
&\asymp \sum_{l_2=1}^{J_\mathrm{A}} \sum_{l_2=1}^{J_\mathrm{C}} \gamma_{k,l_1 l_2}^2,
\end{aligned}$$

which yields $\|g\|^2 \asymp |\gamma|^2$.

Along the line of Lemma A.2 of Huang, Wu and Zhou (2004), we can show $\|g\|_n^2 \asymp \|g\|^2$ for any $g \in \mathcal{G}$. Therefore, $\gamma^\tau S_n^2 / n \gamma = \|g\|_n^2 \asymp \|g\|^2 \asymp |\gamma|^2$. □

Similar to the procedure of Lemma 3.2 of He and Shi (1994), we have the next lemma.

Lemma 2. *Suppose that Assumptions A1–A5 and M1-M2 hold. Then for any L satisfying $1 \le L \le \bar{h}^{\frac{\eta}{10}}$ for some $0 < \eta < \frac{r-\frac{1}{2}}{2r+1}$, it holds that $\sup_{\|\boldsymbol{\zeta}\| \le L} |\kappa^{-1} \Delta_n(\kappa^{1/2} \boldsymbol{\zeta})| = o_p(1)$, where $\kappa = \bar{h}^2 / \bar{N}_\mathrm{H} + 1 - \bar{N}_\mathrm{H}^{-1}$.*

Lemma 3. *Under the assumptions of Proposition 1,*

$$P\left(\inf_{\|\boldsymbol{\zeta}\| \ge L\sqrt{\kappa}} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \rho\left(\varepsilon_{ij} - \tilde{\pi}_{ij}^\tau \boldsymbol{\zeta} + R_{ij}\right) > \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \rho\left(\varepsilon_{ij} + R_{ij}\right) \right) \to 1, \tag{S2.5}$$

Proof. We will show (S2.5) for convex loss function and non-convex loss function, respectively. Assume the convex loss function $\rho(\cdot)$ satisfies conditions A5, M1 and M2. Notice that

$$\max_{i,j} \kappa \|\tilde{\pi}_{ij}^\tau \boldsymbol{\zeta}\|^2 \le \max_{i,j} \kappa \tilde{\pi}_{ij}^\tau \tilde{\pi}_{ij} \|\boldsymbol{\zeta}\|^2 = O_p\big(\kappa \bar{h}^2 \|\boldsymbol{\zeta}\|^2 / n\big) = o(1),$$

which implies $\max_{i,j}(|R_{ij}| + \kappa^{1/2} |\tilde{\pi}(t_{ij})^\tau \boldsymbol{\zeta}|) = o_p(1)$. In combination with Lemma 2 and (S2.4), we can show (S2.5) along the lines of Theorem 1 of Tang and Cheng (2008).

For the non-convex loss function $\rho(\cdot)$, we assume that conditions A5, N1 and N2 hold.

Notice that

$$\kappa^{-1}\Gamma_n(\sqrt{\kappa}\boldsymbol{\zeta}) = \kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{R_{ij}}^{R_{ij}-\sqrt{\kappa}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}}\phi(\varepsilon_{ij}+u)\mathrm{d}u$$
$$=\kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{R_{ij}}^{R_{ij}-\sqrt{\kappa}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}}\left\{\phi(\varepsilon_{ij})+\phi'(\varepsilon_{ij})u+\left[\phi(\varepsilon_{ij}+u)-\phi(\varepsilon_{ij})-\phi'(\varepsilon_{ij})u\right]\right\}\mathrm{d}u$$
$$=: I_1+I_2+I_3,$$

where

$$I_1 = \kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{R_{ij}}^{R_{ij}-\sqrt{\kappa}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}}\phi(\varepsilon_{ij})\mathrm{d}u,$$
$$I_2 = \kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{R_{ij}}^{R_{ij}-\sqrt{\kappa}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}}\phi'(\varepsilon_{ij})u\mathrm{d}u, \quad \text{and}$$
$$I_3 = \kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{R_{ij}}^{R_{ij}-\sqrt{\kappa}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}}\left[\phi(\varepsilon_{ij}+u)-\phi(\varepsilon_{ij})-\phi'(\varepsilon_{ij})u\right]\mathrm{d}u.$$

Similar to the proof of Theorem 1 of Tang and Cheng (2008), we can show

$$|I_1| = \kappa^{-1/2}\left|\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi(\varepsilon_{ij})\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}\right| = O_p(\|\zeta\|/\sqrt{\kappa}) \tag{S2.6}$$

by Assumption A4. After direct computations, we obtain that

$$I_2 = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi'(\varepsilon_{ij})(\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta})^2 - \kappa^{-1/2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi'(\varepsilon_{ij})R_{ij}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta} =: I_{21}+I_{22}.$$

Note that $\sum_{i=1}^{n_i}\frac{1}{n_i}\sum_{j=1}^{n_i}(\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta})^2 = \|\boldsymbol{\zeta}\|^2$, we have

$$I_{21} = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi_{\varepsilon}(t_{ij})(1+o_p(1))(\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta})^2 > \frac{c}{2}\|\boldsymbol{\zeta}\|^2 \tag{S2.7}$$

and

$$|I_{22}| = \kappa^{-1/2}\left|\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi_{\varepsilon}(t_{ij})(1+o_p(1))R_{ij}\tilde{\pi}_{ij}^{\tau}\boldsymbol{\zeta}\right| = O_p(\sqrt{n}\kappa^{-1/2}\underline{h}^{-r}\|\boldsymbol{\zeta}\|) \tag{S2.8}$$

from Assumption A5 and (S2.3). According to Assumption N2, we derive that

$$\kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{R_{ij}}^{R_{ij}-\sqrt{\kappa}\tilde{\pi}_{ij}^{\tau}\zeta}\mathrm{E}\Big[\sup_{u}|\phi(\varepsilon_{ij}+u)-\phi(\varepsilon_{ij})-\phi'(\varepsilon_{ij})u||t_{ij}=t\Big]\mathrm{d}u$$
$$=\kappa^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(-\sqrt{\kappa}\tilde{\pi}^{\tau}\boldsymbol{\zeta}\right)o\left(|R_{ij}|+|\sqrt{\kappa}\tilde{\pi}^{\tau}\boldsymbol{\zeta}|\right),$$

which implies that $|I_3| = o\big(\sqrt{n}\kappa^{-1/2}\underline{h}^{-r}\,\|\zeta\|\,\big) + o\big(\,\|\zeta\|^2\,\big)$.

In combination with (S2.6), (S2.7) and (S2.8), we have that

$$P\left(\inf_{\|\boldsymbol{\zeta}\|\geq L}\kappa^{-1}\Gamma_n\left(\sqrt{\kappa}\boldsymbol{\zeta}\right) > \frac{c}{2}L^2 > 0\right) \to 1,$$

which yields (S2.5). □

Proof of Proposition 1.

Proof. From Lemma 3, we have $\|\hat{\boldsymbol{\zeta}}\| = O_p(\kappa^{1/2})$, which implies

$$\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|^2 \asymp (\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})^\tau S_n^2(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})/n = \|\,\hat{\boldsymbol{\zeta}}\,\|^2\,/n = O_p(\kappa/n) = o_p(1)$$

from Lemma 1. Employing approximation theories of spline functions and (S2.1), we have

$$\begin{aligned}
\|\hat{g}_k - g_k\|_{L_2}^2 &= \int_0^1\int_{a_k}^{b_k}\big[\hat{g}_k(t,x_k) - g_k(t,x_k)\big]^2\mathrm{d}t\mathrm{d}x_k \\
&\preceq \int_0^1\int_{a_k}^{b_k}\big[\hat{g}_k(t,x_k) - \tilde{g}_k(t,x_k)\big]^2\mathrm{d}t\mathrm{d}x_k + (D^*)^2\underline{h}^{-2r} \\
&= \int_0^1\int_{a_k}^{b_k}[(\hat{\gamma}_k - \tilde{\gamma}_k)^\tau\mathcal{T}_k(t,x_k)]^2\mathrm{d}t\mathrm{d}x_k + (D^*)^2\underline{h}^{-2r} \\
&\preceq \|\hat{\gamma}_k - \tilde{\gamma}_k\|^2 + (D^*)^2\underline{h}^{-2r} \\
&= O_p(\kappa/n + \underline{h}^{-2r}),
\end{aligned}$$

which implies $\|\hat{g}_k(t|x_k) - g_k(t|x_k)\|_{L_2}^2 := \int_0^1[\hat{g}_k(t,x_k) - g_k(t,x_k)]^2\mathrm{d}t = O_p(\kappa/n + \underline{h}^{-2r})$.

Then, Cauthy-Schwartz inequality means

$$\big|\,\|\hat{g}_k(t|x_k)\|_{L_1} - \|g_k(t|x_k)\|_{L_1}\,\big| \leq \|\hat{g}_k(t|x_k) - g_k(t|x_k)\|_{L_2} = O_p\big(\sqrt{\kappa/n} + \underline{h}^{-r}\big), \qquad \text{(S2.9)}$$

where $\|f\|_{L_1} = \int f(x)\mathrm{d}x$. Using the identification condition $\|\alpha_k\|_{L_1} = 1$, we have

$$\begin{aligned}
&\|\hat{\alpha}_{k,\mathrm{I}}(t|t_{k0}, x_{k0}) - \alpha_k(t)\|_{L_2}^2 \\
&= \int_0^1 \left(\frac{\hat{\xi}_k(t|t_{k0}, x_{k0})}{\left\|\hat{\xi}_k(t|t_{k0}, x_{k0})\right\|_{L_1}} - \frac{\xi_k(t|t_{k0})}{\|\xi_k(t|t_{k0})\|_{L_1}} \right)^2 \mathrm{d}t = \int_0^1 \left(\frac{\hat{g}_k(t, x_{k0})}{\|\hat{g}_k(t|x_{k0})\|_{L_1}} - \frac{g_k(t, x_{k0})}{\|g_k(t|x_{k0})\|_{L_1}} \right)^2 \mathrm{d}t \\
&\leq 2\int_0^1 \frac{1}{\|\hat{g}_k(t|x_{k0})\|_{L_1}^2} \left(\hat{g}_k(t, x_{k0}) - g_k(t, x_{k0})\right)^2 \mathrm{d}t \\
&\quad + 2\int_0^1 g_k^2(t, x_{k0}) \left(\frac{1}{\|\hat{g}_k(t|x_{k0})\|_{L_1}} - \frac{1}{\|g_k(t|x_{k0})\|_{L_1}} \right)^2 \mathrm{d}t \\
&= O_p(\kappa/n + \underline{h}^{-2r}) = O_p\big(\bar{h}^2/(n\bar{N}_{\mathrm{H}}) + 1/n + \underline{h}^{-2r}\big).
\end{aligned}$$

The derivation of $\int_0^1 [\hat{\alpha}_{0,\mathrm{I}}(t) - \alpha_0(t)]^2 \mathrm{d}t = O_p(\kappa/n + \underline{h}^{-2r})$ is straightforward, and omitted the details. Finally, we show the convergence rate in the mean squared error sense. Similar to Lemma 1, we can show the largest eigenvalues of

$$\frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} \pi_{ij}\pi_{ij}^\tau \qquad \text{and} \qquad \frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} \mathcal{T}_k(t_{ij}, x_{k0})\mathcal{T}_k^\tau(t_{ij}, x_{k0})$$

are bounded, which to yield

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} \left[\hat{g}_k(x_{ij}, x_{ijk}) - g_k(t_{ij}, x_{ijk})\right]^2 &= O_p(\kappa/n + \underline{h}^{-2r}), \\
\frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} \left[\hat{g}_k(t_{ij}, x_{k0}) - g_k(t_{ij}, x_{k0})\right]^2 &= O_p(\kappa/n + \underline{h}^{-2r}).
\end{aligned}$$

Again from (S2.9), we obtain that

$$\frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} \left[\hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0}) - \alpha_k(t_{ij}|t_{k0}, x_{k0})\right]^2 = O_p(\kappa/n + \underline{h}^{-2r}),$$

which completes the proof. □

S2.2 Proof of Theorem 1

Under Assumption (A3), there exists positive constants $c_{k,\mathrm{A}}$ $(k = 1, ..., p)$ such that

$$\begin{cases}
\beta_k(x_k) = \tilde{\beta}_k(x_k) + r_{k,\mathrm{A}}(x_k) = \tilde{\theta}_k^\tau \mathbf{B}_{k,\mathrm{A}}(x_k) + r_{k,\mathrm{A}}(x_k), \\
\sup_{x_k \in [a_k, b_k]} |r_{k,\mathrm{A}}(x_k)| \leq c_{k,\mathrm{A}} K_{\mathrm{A}}^{-r}.
\end{cases}$$

Let $r_{ij,\mathrm{A}} = \sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0}) r_{k,\mathrm{A}}(x_{ijk})$, then

$$\max_{i,j} |r_{ij,\mathrm{A}}| = O_p(K_{\mathrm{A}}^{-r}) \tag{S2.10}$$

from Proposition 1 and Assumption (A3).

Let $\hat{\Psi}_{ij} = \left\{\hat{\psi}_1^{\tau}(x_{ij1}), ..., \hat{\psi}_p^{\tau}(x_{ijp})\right\}^{\tau}$ and $\tilde{\Psi}_{ij} = S_{n,\mathrm{A}}^{-1}\hat{\Psi}_{ij}$, where $S_{n,\mathrm{A}}^2 = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\Psi}_{ij}\hat{\Psi}_{ij}^{\tau}$ and $\hat{\psi}_k(x_{ijk}) = \hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0})\mathbf{B}_{k,\mathrm{A}}(x_{ijk})$. Denote $\boldsymbol{\vartheta} = S_{n,\mathrm{A}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\vartheta}} = S_{n,\mathrm{A}}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})$, and $\Delta_{ij} = \hat{\alpha}_{0,\mathrm{I}}(t_{ij}) - \alpha_0(t_{ij}) + \sum_{k=1}^{p}\left[\hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0}) - \alpha_k(t_{ij})\right]\beta_k(x_{ijk})$. Then, the minimizing problem (2.5) can be rewritten as

$$\operatorname*{argmin}_{\boldsymbol{\vartheta}} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\rho(\varepsilon_{ij} - \tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta} + r_{ij,\mathrm{A}} - \Delta_{ij}) - \rho(\varepsilon_{ij} + r_{ij,\mathrm{A}} - \Delta_{ij})\right].$$

Denote $\Gamma_{n,\mathrm{A}}(\boldsymbol{\vartheta})$ be the objective function above,

$$\Delta_{n,\mathrm{A}}(\boldsymbol{\vartheta}) = \Gamma_{n,\mathrm{A}}(\boldsymbol{\vartheta}) - \Phi_{n,\mathrm{A}}(\boldsymbol{\vartheta}) + \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta}\phi(\varepsilon_{ij}).$$

Then,

$$\Gamma_{n,\mathrm{A}}(\boldsymbol{\vartheta}) = \Phi_{n,\mathrm{A}}(\boldsymbol{\vartheta}) - \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta}\phi(\varepsilon_{ij}) + \Delta_n(\boldsymbol{\vartheta}). \tag{S2.11}$$

Lemma 4. *Under Assumption A2 and A3, except on an event whose probability tends to zero, the eigenvalues of $S_{n,\mathrm{A}}^2/n$ has positive lower bound $L_{1,\mathrm{A}}$ and upper bound $L_{2,\mathrm{A}}$.*

Proof. Let $\mathcal{G}_{\mathrm{A}_t} = \{g(\mathbf{x}|t) = \sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(t|t_{k0}, x_{k0})\theta_k^{\tau}\mathbf{B}_{k,\mathrm{A}}(x_k)\}$ be a family of functions defined on $\prod_{k=1}^{p}[a_k, b_k]$ for any given $t \in [0, 1]$. For any $g_1, g_2 \in \mathcal{G}_{\mathrm{A}_t}$, i.e.,

$$g_1(\mathbf{x}|t) = \sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(t|t_{k0}, x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(x_k)\theta_k^{(1)}, \quad g_2(\mathbf{x}|t) = \sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(t|t_{k0}, x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(x_k)\theta_k^{(2)},$$

define theoretical inner product

$$\langle g_1, g_2\rangle_{\mathrm{A}} = \mathrm{E}\Big[\Big\{\sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(T|t_{k0}, x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(X_k)\theta_k^{(1)}\Big\}\Big\{\sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(T|t_{k0}, x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(X_k)\theta_k^{(2)}\Big\}\Big|T = t\Big],$$

and empirical inner product

$$\langle g_1, g_2\rangle_{n,\mathrm{A}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\Big\{\sum_{k=1}^{p}\hat{\alpha}_{k,\mathrm{I}}(t|t_{k0},x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(x_{ijk})\theta_k^{(1)}\Big\}\Big\{\sum_{k=1}^{p}\hat{\alpha}_{k,\mathrm{I}}(t|t_{k0},x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(x_{ijk})\theta_k^{(2)}\Big\}.$$

Denote the induced theoretical norm and empirical norm as $\|g\|_{\mathrm{A}}$ and $\|g\|_{n,\mathrm{A}}$, respectively. Then, for any $g\in\mathcal{G}_{\mathrm{A}_t}$,

$$\begin{aligned}\|g\|_{\mathrm{A}}^2 &= \mathrm{E}\big[\big\{\sum_{k=1}^{p}\hat{\alpha}_{k,\mathrm{I}}(T|t_{k0},x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(X_k)\theta_k\big\}^2\big|T=t\big]\\ &= \int_{a_k}^{b_k}\big[\sum_{k=1}^{p}\hat{\alpha}_{k,\mathrm{I}}(t|t_{k0},x_{k0})\mathbf{B}_{k,\mathrm{A}}^{\tau}(X_k)\theta_k\big]^2 f_{\mathbf{X}|T}(\mathbf{x}|t)\mathbf{dx}\asymp|\boldsymbol{\theta}|^2.\end{aligned}$$

Furthermore, under Assumption (A2), we can show $\|g\|_{n,\mathrm{A}}^2\asymp\|g\|_{\mathrm{A}}^2$ for any $g\in\mathcal{G}_{\mathrm{A}_t}$, which yields $\boldsymbol{\theta}^{\tau}S_{n,\mathrm{A}}^2\boldsymbol{\theta}/n=\|g\|_{n,\mathrm{A}}^2\asymp\|g\|_{\mathrm{A}}^2\asymp|\boldsymbol{\theta}|^2$. □

Along the lines of Lemma 3.2 of He and Shi (1994), we can derive the following lemma.

Lemma 5. *Suppose that Assumptions A1–A5 and M1-M2 hold. Then for any L satisfying $1\le L\le K_{\mathrm{A}}^{\frac{\eta}{10}}$ for some $0<\eta<\frac{r-\frac{1}{2}}{2r+1}$, it holds that $\sup_{\|\boldsymbol{\vartheta}\|\le L}\big|\tilde{K}_{\mathrm{A}}^{-1}\Delta_{n,\mathrm{A}}\big(\tilde{K}_{\mathrm{A}}^{1/2}\boldsymbol{\vartheta}\big)\big|=o_p(1)$. where $\tilde{K}_{\mathrm{A}}=K_{\mathrm{A}}/\bar{N}_{\mathrm{H}}+1-\bar{N}_{\mathrm{H}}^{-1}$.*

Lemma 6. *Under Assumptions A1–A5, M1-M2 or N1-N2, it follows that*

$$P\left(\inf_{\|\boldsymbol{\vartheta}\|\ge L\tilde{K}_{\mathrm{A}}^{1/2}}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho(\varepsilon_{ij}-\tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta}+r_{ij,\mathrm{A}}-\Delta_{ij})>\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho(\varepsilon_{ij}+r_{ij,\mathrm{A}}-\Delta_{ij})\right)\to 1, \tag{S2.12}$$

Proof. It can be shown for the convex and non-convex loss function, respectively.

- For the convex loss function $\rho(\cdot)$, assume that the conditions A5, M1 and M2 hold.

  From Proposition 1, we have that

$$\begin{aligned}\Psi_{ij}^{\tau}\Psi_{ij} &= \sum_{k=1}^{p}\hat{\alpha}_{k,\mathrm{I}}^2(t_{ij}|t_{k0},x_{k_0})\sum_{l}B_{kl,\mathrm{A}}^2(x_{ijk})\\ &\le pK_{\mathrm{A}}+\sup_{i,j}\sum_{k=1}^{p}\big|\hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0,x_{k0}})-\alpha_k(t_{ij})\big|+o(s.o.)=O(K_{\mathrm{A}}),\end{aligned}$$

which yields

$$\max_{i,j} \tilde{K}_{\mathrm{A}} \big(\tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}\big)^2 \leq \max_{i,j} \tilde{K}_{\mathrm{A}} \tilde{\Psi}_{ij}^{\tau} \tilde{\Psi}_{ij} \|\boldsymbol{\vartheta}\|^2 = O_p\big(\tilde{K}_{\mathrm{A}} K_{\mathrm{A}} \|\boldsymbol{\vartheta}\|^2/n\big) = o(1). \tag{S2.13}$$

Similar to the proof of Theorem 1 of Tang and Cheng (2008), we can derive that

$$\begin{aligned}
\tilde{K}_{\mathrm{A}}^{-1} \Phi_{n,\mathrm{A}}\big(\tilde{K}_{\mathrm{A}}^{1/2} \boldsymbol{\vartheta}\big) &= \sum_{i=1}^{n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(t_{ij}) \Big[\frac{1}{2}\big(\tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}\big)^2 - \tilde{K}_{\mathrm{A}}^{-1/2} \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}\big(r_{ij,\mathrm{A}} - \Delta_{ij}\big)\Big] + o_p(1) \\
&\geq \frac{c}{2} \|\boldsymbol{\vartheta}\|^2 - \tilde{K}_{\mathrm{A}}^{-1/2} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(t_{ij}) \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}\big(r_{ij,\mathrm{A}} - \Delta_{ij}\big) + o_p(1)
\end{aligned}$$

from Assumption M1 and (S2.13). Furthermore, noticing that

$$\sup_{i,j} |r_{ij,\mathrm{A}} - \Delta_{ij}|^2 \leq K_{\mathrm{A}}^{-2r} + \underline{h}^{-2r} + \frac{1}{n} + \frac{\bar{h}^2}{n \bar{N}_{\mathrm{H}}} = O(K_{\mathrm{A}}^{-2r})$$

by Proposition 1, we obtain that

$$\sup_{\|\boldsymbol{\vartheta}\| \leq L} \tilde{K}_{\mathrm{A}}^{-1/2} \Big| \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(t_{ij}) \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta} (r_{ij,\mathrm{A}} - \Delta_{ij}) \Big| = O_p\big(\tilde{K}_{\mathrm{A}}^{-1/2} K_{\mathrm{A}}^{-r} \sqrt{n} L\big).$$

On the other hand,

$$\mathrm{Var}\big(\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta} \phi(\boldsymbol{\varepsilon}_{ij})\big) \leq C \mathrm{E}\big[\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \big(\tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}\big)^2\big] = C \|\boldsymbol{\vartheta}\|^2$$

implies that $\sup_{\|\boldsymbol{\vartheta}\| \leq L} \tilde{K}_{\mathrm{A}}^{-1/2} \big| \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta} \phi(\boldsymbol{\varepsilon}_{ij}) \big| = O_p(\tilde{K}_{\mathrm{A}}^{-1/2} L) = o_p(1)$.

Hence, from (S2.15), Lemma 5 and the convexity of ρ, we can show (S2.12).

- For the non-convex loss function $\rho(\cdot)$, assume that the conditions A5, N1 and N2 hold.

  Note that

$$\begin{aligned}
\tilde{K}_{\mathrm{A}}^{-1} \Gamma_{n,\mathrm{A}}\big(\tilde{K}_{\mathrm{A}}^{1/2} \boldsymbol{\vartheta}\big) &= \tilde{K}_{\mathrm{A}}^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \int_{r_{ij,\mathrm{A}} - \Delta_{ij}}^{r_{ij,\mathrm{A}} - \Delta_{ij} - \tilde{K}_{\mathrm{A}}^{1/2} \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}} \phi(\varepsilon_{ij} + u) \mathrm{d}u \\
&= \tilde{K}_{\mathrm{A}}^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \int_{r_{ij,\mathrm{A}} - \Delta_{ij}}^{r_{ij} - \Delta_{ij} - \tilde{K}_{\mathrm{A}}^{1/2} \tilde{\Psi}_{ij}^{\tau} \boldsymbol{\vartheta}} \big\{\phi(\varepsilon_{ij}) + \phi'(\varepsilon_{ij}) u \\
&\quad + \big[\phi(\varepsilon_{ij} + u) - \phi(\varepsilon_{ij}) - \phi'(\varepsilon_{ij}) u\big] \big\} \mathrm{d}u \\
&=: I_1 + I_2 + I_3,
\end{aligned}$$

where

$$I_1 = -\tilde{K}_{\rm A}^{-1/2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi(\varepsilon_{ij})\tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta},$$

$$I_2 = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi_{\varepsilon}(t_{ij})(1+o_p(1))\big(\tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta}\big)^2 - \tilde{K}_{\rm A}^{-1/2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\phi_{\varepsilon}(t_{ij})\tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta}$$
$$(r_{ij,\rm A}-\Delta_{ij})(1+o_p(1)),$$

and

$$I_3 = \tilde{K}_{\rm A}^{-1}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\int_{r_{ij,\rm A}-\Delta_{ij}}^{r_{ij,\rm A}-\Delta_{ij}-\tilde{K}_{\rm A}^{1/2}\tilde{\Psi}_{ij}^{\tau}\boldsymbol{\vartheta}}\big[\phi(\varepsilon_{ij}+u)-\phi(\varepsilon_{ij})-\phi'(\varepsilon_{ij})u\big]\,{\rm d}u.$$

From Proposition 1, we have that

$$\begin{aligned}\hat{\Psi}_{ij}^{\tau}\hat{\Psi}_{ij} &= \sum_{k=1}^{p}\hat{\alpha}_{k,\rm I}^2(t_{ij}|t_{k0},x_{k_0})\sum_{l}B_{kl,\rm A}^2(x_{ijk})\\ &\le pK_{\rm A} + \sup_{i,j}\sum_{k=1}^{p}\big|\hat{\alpha}_{k,\rm I}(t_{ij}|t_{k0,x_{k0}})-\alpha_k(t_{ij})\big| + o(s.o.) = O(K_{\rm A}),\end{aligned}$$

and

$$\sup_{i,j}|r_{ij,\rm A}-\Delta_{ij}|^2 = O_p\Big(K_{\rm A}^{-2r}+\underline{h}^{-2r}+\frac{1}{n}+\frac{\bar{h}^2}{n\bar{N}_{\rm H}}\Big) = O_p\left(K_{\rm A}^{-2r}\right).$$

Similar to the proof of Proposition 1, we can show that

$$|I_1| = O_p\big(\tilde{K}_{\rm A}^{-1/2}\left\|\boldsymbol{\vartheta}\right\|\big)\quad\text{and}\quad |I_3| = o\big(\sqrt{n}\tilde{K}_{\rm A}^{-1/2}K_{\rm A}\left\|\boldsymbol{\vartheta}\right\|\big)+o\big(\left\|\boldsymbol{\vartheta}\right\|^2\big),$$

and $I_2 > \frac{c}{2}\left\|\boldsymbol{\vartheta}\right\|^2 + O\big(\sqrt{n}\tilde{K}_{\rm A}^{-1/2}K_{\rm A}^{-r}\left\|\boldsymbol{\vartheta}\right\|\big)$. Therefore, for any sufficient large L, (S2.12) holds. □

Proof of Theorem 1

Proof. By Lemma 6, we have that $\|\hat{\boldsymbol{\vartheta}}\| = O(\tilde{K}_{\rm A}^{1/2})$. Furthermore, Lemma 4 gives

$$\|\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}\|^2 \asymp (\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})^{\tau}S_{n,\rm A}^2(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})/n = \|\hat{\boldsymbol{\vartheta}}\|^2/n = O_p(\tilde{K}_{\rm A}/n) = o_p(1).$$

Employing Cauchy-Schwartz inequality, we have $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[\hat{\beta}_k(x_{ijk})-\beta_k(x_{ijk})\right]^2 \preceq I_1+I_2+I_3$, where

$$\begin{aligned}
I_1 =& \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[\check{\beta}_k(x_{ijk})-\beta_k(x_{ijk})\right]^2,\\
I_2 =& \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[\check{\beta}_k(x_{ijk})-\beta_k(x_{ijk})\right]\Big]^2,\\
I_3 =& \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\beta_k(x_{ijk})\Big]^2.
\end{aligned}$$

Since $\mathrm{E}[\beta_k(X_k)]=0$, we have $I_3=O_p(N^{-1})$. It is sufficient to deal with I_1.

Under Assumption A2, we can show the largest eigenvalue of $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{B}_k(x_{ijk})\mathbf{B}_k^{\tau}(x_{ijk})$ is bounded, which leads

$$\begin{aligned}
I_1 &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\hat{\theta}_k-\tilde{\theta}_k)^{\tau}\mathbf{B}_k(x_{ijk})\mathbf{B}_k^{\tau}(x_{ijk})(\hat{\theta}_k-\tilde{\theta}_k)+c_{k,\mathrm{A}}K_{\mathrm{A}}^{-2r}\\
&= O_p\left(\frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}}+\frac{1}{n}+K_{\mathrm{A}}^{-2r}\right).
\end{aligned}$$

Hence the rate of convergence in the sense of MSE.

Next, we show L_2 convergence rate of M-estimators of β_k. Again by Cauchy-Schwartz inequality,

$$\left\|\hat{\beta}_k-\beta_k\right\|_{L_2}^2 \preceq \left\|\check{\beta}_k-\beta_k\right\|_{L_2}^2+\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[\check{\beta}_k(x_{ijk})-\beta_k(x_{ijk})\right]^2+\Big|\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\beta_k(x_{ijk})\Big|^2.$$

It is sufficient to deal with $\left\|\check{\beta}_k-\beta_k\right\|_{L_2}^2$. From (S2.1), we get

$$\begin{aligned}
\left\|\check{\beta}_k-\beta_k\right\|_{L_2}^2 &= \int_0^1\left[\check{\beta}_k(x_k)-\beta_k(x_k)\right]^2\mathrm{d}x_k\\
&\le \int_0^1\left[(\hat{\theta}_k-\tilde{\theta}_k)^{\tau}\mathbf{B}_{k,\mathrm{A}}(x_k)\right]^2\mathrm{d}x_k+c_{k,\mathrm{A}}K_{\mathrm{A}}^{-2r}\\
&\asymp \|\hat{\theta}_k-\tilde{\theta}_k\|^2+c_{k,\mathrm{A}}K_{\mathrm{A}}^{-2r}\\
&= O_p(\tilde{K}_{\mathrm{A}}/n+K_{\mathrm{A}}^{-2r}),
\end{aligned}$$

which to lead $\|\hat{\beta}_k-\beta_k\|_{L_2}^2=O_p(\tilde{K}_{\mathrm{A}}/n+K_{\mathrm{A}}^{-2r})$. □

S2.3 Proof of Theorem 2

Under Assumption (A3), there exists positive constants $d_{k,\mathrm{C}}$ $(k = 0, ..., p)$, such that

$$\begin{cases} \alpha_k(t) = \tilde{\alpha}_k(t) + r_{k,\mathrm{C}}(t) = \tilde{h}_k^\tau \mathbf{b}_\mathrm{C}(t) + r_{k,\mathrm{C}}(t), \\ \sup_{t\in[0,1]} |r_{k,\mathrm{C}}(t)| \le d_{k,\mathrm{C}} K_\mathrm{C}^{-r}. \end{cases}$$

Let $r_{ij,\mathrm{C}} = r_{0,\mathrm{C}}(t_{ij}) + \sum_{k=1}^p r_{k,\mathrm{C}}(t_{ij})\hat{\beta}_k(x_{ijk})$, then

$$\max_{i,j} \left| r_{ij,\mathrm{C}} \right| = O_p(K_\mathrm{C}^{-r}) \tag{S2.14}$$

by Theorem 1 and Assumption A3.

Let $\delta_{ij} = \sum_{k=1}^p \alpha_k(t_{ij}) \big[\hat{\beta}_k(x_{ijk}) - \beta_k(x_{ijk}) \big]$, $\hat{\Phi}_{ij} = \{1, \hat{\beta}_1(x_{ij1}), ..., \hat{\beta}_p(x_{ijp})\}^\tau \otimes \mathbf{b}_\mathrm{C}(t_{ij})^\tau$. and $\tilde{\Phi}_{ij} = S_{n,\mathrm{C}}^{-1} \hat{\Phi}_{ij}$ with $S_{n,\mathrm{C}}^2 = \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\Phi}_{ij} \hat{\Phi}_{ij}^\tau$. Denote $\boldsymbol{\varsigma} = S_{n,\mathrm{C}}(\boldsymbol{h} - \tilde{\boldsymbol{h}})$ and $\hat{\boldsymbol{\varsigma}} = S_{n,\mathrm{C}}(\hat{\boldsymbol{h}} - \tilde{\boldsymbol{h}})$. The minimizing problem (2.7) can be equivalently written as

$$\operatorname*{argmin}_{\boldsymbol{\varsigma}} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\rho\big(\varepsilon_{ij} - \tilde{\Phi}_{ij}^\tau \boldsymbol{\varsigma} + r_{ij,\mathrm{C}} - \delta_{ij}\big) - \rho\big(\varepsilon_{ij} + r_{ij,\mathrm{C}} - \delta_{ij}\big) \right].$$

Denote $\Gamma_{n,\mathrm{C}}(\boldsymbol{\varsigma})$ be the objective function above, $\Phi_{n,\mathrm{C}}(\boldsymbol{\varsigma}) = \mathrm{E}\big[\Gamma_{n,\mathrm{C}}(\boldsymbol{\varsigma})|\mathcal{J}\big]$, and $\Delta_{n,\mathrm{C}}(\boldsymbol{\varsigma}) = \Gamma_{n,\mathrm{C}}(\boldsymbol{\varsigma}) - \Phi_{n,\mathrm{C}}(\boldsymbol{\varsigma}) + \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Phi}_{ij}^\tau \boldsymbol{\varsigma} \phi(\varepsilon_{ij})$. Then,

$$\Gamma_{n,\mathrm{A}}(\boldsymbol{\varsigma}) = \Phi_{n,\mathrm{A}}(\boldsymbol{\varsigma}) - \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Phi}_{ij}^\tau \boldsymbol{\varsigma} \phi(\varepsilon_{ij}) + \Delta_n(\boldsymbol{\varsigma}). \tag{S2.15}$$

Similar to Lemma 4—6, we have the following lemmas.

Lemma 7. *Under Assumption A1 and A3, except on an event whose probability tends to zero, the eigenvalues of $S_{n,\mathrm{C}}^2/n$ has positive lower bound $L_{1,\mathrm{C}}$ and upper bound $L_{2,\mathrm{C}}$.*

Lemma 8. *Suppose that conditions A1–A5 and M1-M2 hold. Then for any L satisfying $1 \le L \le K_\mathrm{C}^{\frac{\eta}{10}}$ for some $0 < \eta < \frac{r-\frac{1}{2}}{2r+1}$, it holds that $\sup_{\|\boldsymbol{\vartheta}\| \le L} \big| \tilde{K}_\mathrm{C}^{-1} \Delta_{n,\mathrm{C}}\big(\tilde{K}_\mathrm{C}^{1/2} \boldsymbol{\vartheta}\big) \big| = o_p(1)$, where $\tilde{K}_\mathrm{C} = K_\mathrm{C}/\bar{N}_\mathrm{H} + 1 - \bar{N}_\mathrm{H}^{-1}$.*

Proof of Theorem 2.

Proof. In the same vein of Theorem 1, we can show that, for sufficiently large L,

$$P\Big(\inf_{\|\boldsymbol{\varsigma}\|\geq L\tilde{K}_{\mathrm{C}}^{1/2}}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho\big(\varepsilon_{ij}-\tilde{\Phi}_{ij}^{\tau}\boldsymbol{\varsigma}+r_{ij,\mathrm{C}}-\delta_{ij}\big)>\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho\big(\varepsilon_{ij}+r_{ij,\mathrm{C}}-\delta_{ij}\big)\Big)\to 1,$$

which implies that $\|\boldsymbol{\varsigma}\|=O\big(\tilde{K}_{\mathrm{C}}^{1/2}\big)$. Furthermore, Lemma 7 gives

$$\|\hat{\boldsymbol{h}}-\tilde{\boldsymbol{h}}\|^2\asymp(\hat{\boldsymbol{h}}-\tilde{\boldsymbol{h}})^{\tau}S_{n,\mathrm{C}}^2(\hat{\boldsymbol{h}}-\tilde{\boldsymbol{h}})/n=\|\boldsymbol{\varsigma}\|^2/n=O_p(\tilde{K}_{\mathrm{C}}/n)=o_p(1).$$

Denote $\hat{\alpha}_k=\check{\alpha}_k/\left\|\check{\alpha}_k\right\|_{L_1}$, where $\check{\alpha}_k=\mathbf{b}_{\mathrm{C}}^{\tau}(t)\hat{h}_k$. Then, by the identification condition $\int_0^1\alpha_k(t)\mathrm{d}t=1$ and triangular inequality, we have

$$\begin{aligned}\|\hat{\alpha}_k-\alpha_k\|_{L_2}^2&=\left\|\frac{\check{\alpha}_k}{\|\check{\alpha}_k\|_{L_1}}-\alpha_k\right\|_{L_2}^2\leq 2\left\|\check{\alpha}_k(t)-\alpha_k(t)\right\|_{L_2}^2+2\left\|\frac{\check{\alpha}_k}{\|\check{\alpha}_k\|_{L_1}}-\check{\alpha}_k\right\|_{L_2}^2\\&=2\left\|\check{\alpha}_k-\alpha_k\right\|_{L_2}^2+4\left(\|\check{\alpha}_k-\alpha_k\|_{L_2}^2++\|\alpha_k\|_{L_2}^2\right)\left(\frac{1-\|\check{\alpha}_k\|_{L_1}}{\|\check{\alpha}_k\|_{L_1}}\right)^2.\end{aligned}$$

Note that $\left|\,\|\check{\alpha}_k\|_{L_1}-1\right|=\left|\int_0^1\{\check{\alpha}_k(t)-\alpha_k(t)\}\mathrm{d}t\right|\leq\|\check{\alpha}_k-\alpha_k\|_{L_2}$ and

$$\|\check{\alpha}_k-\alpha_k\|_{L_2}^2\leq 2\left\|\check{\alpha}_k-\tilde{\alpha}_k\right\|_{L_2}^2+2d_{k,\mathrm{C}}K_{\mathrm{C}}^{-2r}=O_p(\tilde{K}_{\mathrm{C}}/n+K_{\mathrm{C}}^{-2r}),\tag{S2.16}$$

we have $\|\hat{\alpha}_k-\alpha_k\|_{L_2}^2=O_p(\tilde{K}_{\mathrm{C}}/n+K_{\mathrm{C}}^{-2r})$.

It is not difficult to show the largest eigenvalue of $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{b}_{\mathrm{C}}(t_{ij})\mathbf{b}_{\mathrm{C}}^{\tau}(t_{ij})$ is bounded, which implies

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big[\hat{\alpha}_0(t_{ij})-\alpha_0(t_{ij})\big]^2=O_p(\tilde{K}_{\mathrm{C}}/n+K_{\mathrm{C}}^{-2r}).$$

Furthermore,

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big[\hat{\alpha}_k(t_{ij})-\alpha_k(t_{ij})\big]^2\preceq W_1+W_2+W_3,$$

where

$$\begin{aligned}W_1&=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big\{\check{\alpha}_k(t_{ij})-\alpha_k(t_{ij})\big\}^2,\\W_2&=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\alpha_k^2(t_{ij})\left(\frac{1-\|\check{\alpha}_k\|_{L_1}}{\|\check{\alpha}_k\|_{L_1}}\right)^2,\\W_3&=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big\{\check{\alpha}_k(t_{ij})-\alpha_k(t_{ij})\big\}^2\left(\frac{1-\|\check{\alpha}_k\|_{L_1}}{\|\check{\alpha}_k\|_{L_1}}\right)^2.\end{aligned}$$

It is easy to derive that $W_1 = O_p(\tilde{K}_{\mathrm{C}}/n + K_{\mathrm{C}}^{-2r})$. Again by (S2.16), we can complete the proof. □

S2.4 Proof of Theorem 3 and 4

Proof. Let $\tilde{W}_{n,\mathrm{A}} = S_{n,\mathrm{A}}^{-1} W_{n,\mathrm{A}} S_{n,\mathrm{A}}^{-1}$ and $\tilde{\boldsymbol{\vartheta}} = \tilde{W}_{n,\mathrm{A}}^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Psi}_{ij}\phi(\varepsilon_{ij})$. From Assumption M1 and Lemma 4, we can show $\tilde{W}_{n,\mathrm{A}}$ is invertible and its smallest eigenvalue is positive. Using Assumption A5, Proposition 1 and Lemma 4, we obtain that $\|\tilde{\boldsymbol{\vartheta}}\|^2 = O_p(\tilde{K}_{\mathrm{A}})$. Notice that $\Phi_{n,\mathrm{A}}(\boldsymbol{\vartheta}) = \boldsymbol{\vartheta}^\tau \tilde{W}_{n,\mathrm{A}} \boldsymbol{\vartheta}$ and $\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\Psi}_{ij}^\tau \vartheta \phi(\varepsilon_{ij}) = \tilde{\boldsymbol{\vartheta}}^\tau \tilde{W}_{n,\mathrm{A}} \boldsymbol{\vartheta}$, we have

$$\tilde{K}_{\mathrm{A}}^{-1} \Gamma_{n,\mathrm{A}}(\boldsymbol{\vartheta}) = \tilde{K}_{\mathrm{A}}^{-1} \big\{ \boldsymbol{\vartheta}^\tau \tilde{W}_{n,\mathrm{A}} \boldsymbol{\vartheta}/2 - \tilde{\boldsymbol{\vartheta}}^\tau \tilde{W}_{n,\mathrm{A}} \boldsymbol{\vartheta} + \Delta_{n,\mathrm{A}}(\boldsymbol{\vartheta}) \big\} + o_p(1),$$

since $K_{\mathrm{A}}^{2r} \tilde{K}_{\mathrm{A}}/n \to \infty$. Similar to the proof of Lemma 3 of Tang and Cheng (2008), we can derive that $\tilde{K}_{\mathrm{A}}^{-1/2} \|\hat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}\| = o_p(1)$, which implies $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} - S_{n,\mathrm{A}}^{-1} \tilde{\boldsymbol{\vartheta}}\| = O_p\big(\|\hat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}\|/\sqrt{n}\big) = o_p\big(\{\tilde{K}_{\mathrm{A}}/n\}^{1/2}\big)$ from Lemma 4. Therefore,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} &= S_{n,\mathrm{A}}^{-1} \tilde{\boldsymbol{\vartheta}} + o_p\big(\{\tilde{K}_{\mathrm{A}}/n\}^{1/2}\big) \\
&= W_{n,\mathrm{A}}^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \Psi_{ij}\phi(\varepsilon_{ij}) + W_{n,\mathrm{A}}^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} (\hat{\Psi}_{ij} - \Psi_{ij})\phi(\varepsilon_{ij}) + o_p(1).
\end{aligned}$$

Employing Proposition 1 and the conditions of Theorem 3, we get

$$\begin{aligned}
&\mathrm{E}\Big[\Big\| \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \big(\hat{\Psi}_{ij} - \Psi_{ij}\big)\phi(\varepsilon_{ij}) \Big\|^2\Big] \\
\le &\sum_{i=1}^{n} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \mathrm{E}\big[\|\hat{\Psi}_{ij} - \Psi_{ij}\|^2\big] \\
= &\,\mathrm{E}\Big[\sum_{i=1}^{n_i} \frac{1}{n_i^2} \sum_{j=1}^{n_i} [\hat{\alpha}_{k,\mathrm{I}}(t_{ij}) - \alpha_k(t_{ij})]^2 \sum_{l} B_{lk,\mathrm{A}}^2(t_{ij})\Big] \\
= &\,O_p\left(K_{\mathrm{A}}/\underline{h}^{2r} + K_{\mathrm{A}}\bar{h}^2/(n\bar{N}_{\mathrm{H}}) + K_{\mathrm{A}}/n\right) = o_p(1),
\end{aligned}$$

which implies $\big\|W_{n,\mathrm{A}}^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} (\hat{\Psi}_{ij} - \Psi_{ij})\phi(\varepsilon_{ij})\big\|^2 = o_p(1)$.

Hence, for any vector $\mathbf{h}$, it follows that

$$\mathbf{h}^\tau(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = \mathbf{h}^\tau W_{n,\mathrm{A}}^{-1} \sum_{i=1}^{n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} \Psi_{ij}\phi(\varepsilon_{ij}) + o_p(1) = \sum_{i=1}^{n} \frac{1}{n_i} \nu_i^\tau \phi(\varepsilon_i) + o_p(1),$$

where $\nu_i = \mathbf{h}^\tau W_{n,\mathrm{A}}^{-1} \sum_{j=1}^{n_i} \Psi_{ij}\phi(\varepsilon_{ij})$.

Denote $\psi_{k,ijl}$ as the l-th element of vector $\psi_k(x_{ijk})$ and we suppressed the subscript 'k' for the sake of compact notation. By Assumption A4 and the properties of B-spline functions, we have $\mathrm{E}\big[\sum_{j=1}^{n_i} \psi_{ijl}\psi_{ij'l}\phi^2(\varepsilon_{ij})\big] = O(n_i)$, $\mathrm{E}\big[\sum_{j\neq j'} \psi_{ijl}\psi_{ij'l}\phi(\varepsilon_{ij})\phi(\varepsilon_{ij'})\big] = O(n_i(n_i-1)/K_\mathrm{A})$ and $\mathrm{E}\big[\sum_{j\neq j'} \psi_{ijl}\psi_{ij'l'}\phi(\varepsilon_{ij})\phi(\varepsilon_{ij'})\big] = O(n_i(n_i-1)/K_\mathrm{A})$, which yields

$$\begin{aligned}\mathrm{Var}\Big(\sum_{i=1}^{n_i}\frac{1}{n_i}\nu_i\Big) &= \sum_{i=1}^{n}\frac{1}{n_i^2}\mathbf{h}^\tau W_{n,\mathrm{A}}^{-1}\sum_{j=1}^{n_i}\mathrm{E}\Big[\sum_{j=1}^{n_i}\Psi_{ij}\phi(\varepsilon_{ij})\sum_{j=1}^{n_i}\Psi_{ij}^\tau\phi(\varepsilon_{ij})\Big]W_{n,\mathrm{A}}^{-1}\mathbf{h}\\ &= K_\mathrm{A}\Big(\sum_{i=1}^{n}\frac{1}{n^2n_i^2}\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-2}\mathbf{h}\Big)O(n_i+n_i(n_i-1)/K_\mathrm{A})\\ &\geq O\Big(\frac{K_\mathrm{A}}{n\bar{N}_\mathrm{H}}+\frac{1}{n}\Big).\end{aligned}$$

Furthermore, we write $\sum_{i=1}^{n}\mathrm{E}\big|\frac{1}{n_i}\nu_i\big|^3 = I_1+I_2+I_3$, where

$$\begin{aligned}I_1 &= \frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{n_i^3}\sum_{j=1}^{n_i}\mathrm{E}\big[\big|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\Psi_{ij}\phi(\varepsilon_{ij})\big|^3\big],\\ I_2 &= \frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{n_i^3}\sum_{j\neq j'}\mathrm{E}\big[\big|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\Psi_{ij}\phi(\varepsilon_{ij})\big|^2\cdot\big|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\Psi_{ij'}\phi(\varepsilon_{ij'})\big|\big],\quad\text{and}\\ I_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{n_i^3}\sum_{j_1\neq j_2\neq j_3}\mathrm{E}\big[\big|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\psi_{ij1}\phi(\varepsilon_{ij1})\big|\cdot\big|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\psi_{ij2}\phi(\varepsilon_{ij2})\big|\cdot\\ &\qquad\big|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\psi_{ij3}\phi(\varepsilon_{ij3})\big|\big].\end{aligned}$$

Notice that $\big\|\mathbf{h}^\tau(W_{n,\mathrm{A}}/n)^{-1}\big\|\leq C$, $\mathrm{E}|\psi_{ijl}^3\phi^3(\varepsilon_{ij})| = O(K_\mathrm{A}^{1/2})$ and $\mathrm{E}|\psi_{ijl}^2\psi_{ijl'}| = O(K_\mathrm{A}^{1/2})$, we obtain

$$I_1 = \frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{n_i^2}O\big(K_\mathrm{A}^{3/2}\big) = O\left(\sum_{i=1}^{n}\frac{K_\mathrm{A}^{3/2}}{n^3n_i^2}\right). \tag{S2.17}$$

Next, from the fact that $\mathrm{E}|\psi_{ijl}^2\psi_{ij'l}| = O\big(K_\mathrm{A}^{-1/2}\big)$, $\mathrm{E}|\psi_{ijl}^2\psi_{ij'l'}| = O\big(K_\mathrm{A}^{-1/2}\big)$ and $\mathrm{E}|\psi_{ijl}\psi_{ij'l}\psi_{ijl'}| = O\big(K_\mathrm{A}^{-1/2}\big)$, we have

$$I_2 = \frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{n_i^3}n_i(n_i-1)O\big(K_\mathrm{A}^{1/2}\big) = O\left(\sum_{i=1}^{n}\frac{1}{n^3n_i^2}(n_i-1)K_\mathrm{A}^{1/2}\right). \tag{S2.18}$$

It is easy to see that $I_3 = O\big(\sum_{i=1}^{n}\frac{1}{n^3n_i^2}(n_i-1)(n_i-2)\big)$. Combining with (S2.17) and (S2.18), we get

$$\sum_{i=1}^{n} \mathrm{E}\Big|\frac{1}{n_i}\nu_i\Big|^3 = O\left(\sum_{i=1}^{n}\frac{K_{\mathrm{A}}^{3/2}}{n^3 n_i^2} + \sum_{i=1}^{n}\frac{1}{n^3 n_i^2}(n_i-1)K_{\mathrm{A}}^{1/2} + \sum_{i=1}^{n}\frac{1}{n^3 n_i^2}(n_i-1)(n_i-2)\right).$$

Obviously, the condition given in (3.2) of the main body implies the Lyapunov conditions hold, and hence

$$\frac{\sum_{i=1}^{n}\frac{1}{n_i}\nu_i}{\sqrt{\mathrm{Var}\left(\sum_{i=1}^{n}\frac{1}{n_i}\nu_i\right)}} \xrightarrow{L} N(0,1)$$

from Lyapunov central limit theorem. Under Assumption A5, we derive that

$$\mathrm{Var}\left(\sum_{i=1}^{n}\frac{1}{n_i}\nu_i\right) = \mathbf{h}^\tau W_{n,\mathrm{A}}^{-1}\sum_{i=1}^{n}\frac{1}{n_i^2}\boldsymbol{\Psi}_i^\tau G_i \boldsymbol{\Psi}_i W_{n,\mathrm{A}}^{-1}\mathbf{h}$$

Let $\mathbf{h} = A_k(x)$, we have the asymptotic distribution $\hat{\beta}_k(x) - \beta_k(x) \xrightarrow{D} N(0, D_{n,\mathrm{A}}(x))$. □

Theorem 4 can be proceeded in the sam vein of Theorem 3, and we omit the details here.

S2.5 Proof of Theorem 5

Proof. We only give the proof of part(i), and (ii) can be proceeded similarly.

Let $\hat{\nu}_i^\tau = A_k(x)^\tau \hat{W}_{n,\mathrm{A}}^{-1}\hat{\boldsymbol{\Psi}}_i^\tau$, we can write $\hat{D}_{n,\mathrm{A}}(x) = \sum_{i=1}^{n}\frac{1}{n_i^2}\hat{\nu}_i^\tau \hat{G}_i \hat{\nu}_i^\tau = I_1 + I_2 + I_3 + I_4$, where

$$I_1 = \sum_{i=1}^{n}\frac{1}{n_i^2}\nu_i^\tau \hat{G}_i \nu_i, \qquad I_2 = \sum_{i=1}^{n}\frac{1}{n_i^2}(\hat{\nu}_i - \nu_i)^\tau \hat{G}_i \nu_i,$$

$$I_3 = \sum_{i=1}^{n}\frac{1}{n_i^2}\nu_i^\tau \hat{G}_i(\hat{\nu}_i - \nu_i), \qquad I_4 = \sum_{i=1}^{n}\frac{1}{n_i^2}(\hat{\nu}_i - \nu_i)^\tau \hat{G}_i(\hat{\nu}_i - \nu_i).$$

Denote that $\tilde{G}_i = \phi(\varepsilon_i)\phi(\varepsilon_i)^\tau$, we further represent $I_1 = D_{n,\mathrm{A}} + I_{11} + I_{12}$, where

$$I_{11} = \sum_{i=1}^{n}\frac{1}{n_i^2}\nu_i^\tau(\hat{G}_i - \tilde{G}_i)\nu_i, \quad \text{and} \quad I_{12} = \sum_{i=1}^{n}\frac{1}{n_i^2}\nu_i^\tau(\tilde{G}_i - G_i)\nu_i.$$

Similar to the proof of Theorem 3 in Tang and Cheng (2008), we have

$$\mathrm{Var}(I_{12}|\mathcal{J}) \le \tilde{C}\max_i \nu_i^\tau \nu_i \sum_{i=1}^{n}\frac{1}{n_i^2}\nu_i^\tau \nu_i.$$

Since $\nu_i^\tau \nu_i = \frac{1}{n^2} A_k(x)^\tau (W_{n,\mathrm{A}}/n)^{-1} \mathbf{\Psi}_i^\tau \mathbf{\Psi}_i (W_{n,\mathrm{A}}/n)^{-1} A_k(x)$ and $\mathbf{\Psi}_i^\tau \mathbf{\Psi}_i = \sum_{j=1}^{n_i} \Psi_{ij}\Psi_{ij}^\tau$, it is straight to derive that $\|\mathbf{\Psi}_i^\tau \mathbf{\Psi}_i\|_F = O(\sqrt{n_i} K_\mathrm{A})$ and $\|\nu_i^\tau \nu_i\| \le \sqrt{n_i} K_\mathrm{A}^2/n^2$. Therefore,

$$\mathrm{Var}(I_{12}|\mathcal{J}) \le \tilde{C} \max_i \sqrt{n_i} K_\mathrm{A}^2/n^2 \cdot K_\mathrm{A}^2/(n\bar{N}_\mathrm{H}) = o_p(1)$$

under the conditions of Theorem 5(i).

On the other hand, it follows that

$$\begin{aligned}
\mathrm{E}\Big(\sum_{i=1}^n \frac{1}{n_i^2} \nu_i^\tau (\hat{G}_i - \tilde{G}_i)\nu_i|\mathcal{J}\Big) &\le \max_i n_i \sup_{i,j,l} \mathrm{E}\left(|\phi(\hat{\varepsilon}_{ij})\phi(\hat{\varepsilon}_{il}) - \phi(\varepsilon_{ij})\phi(\varepsilon_{il})||\mathcal{J}\right) \times \sum_{i=1}^n \frac{1}{n_i^2}\nu_i^\tau \nu_i \\
&= o_p(1) \max_i n_i O_p\Big(\frac{K_\mathrm{A}^2}{n^2} \sum_{i=1}^n n_i^{-3/2}\Big) \\
&= o_p\big(\max_i n_i K_\mathrm{A}^2/(n\bar{N}_\mathrm{H})\big), \\
&= o_p(1).
\end{aligned}$$

which yields that $I_1 = D_{n,\mathrm{A}}(x) + o_p(1)$.

Furthermore, note that $I_4 = o(I_2)$, it remains to bound I_2, which can be written as $I_2 = I_{21} + I_{22}$ with

$$I_{21} = \sum_{i=1}^n \frac{1}{n_i^2} (\hat{\nu}_i - \nu_i)^\tau G_i \nu_i \quad \text{and} \quad I_{22} = \sum_{i=1}^n \frac{1}{n_i^2} (\hat{\nu}_i - \nu_i)^\tau (\hat{G}_i - G_i)\nu_i.$$

On the one hand, it holds that $\|\nu_i\| = \frac{1}{n}\|A_k(x)\| \cdot \left\|(W_{n,\mathrm{A}}/n)^{-1}\right\| \cdot \|\mathbf{\Psi}_i\| = O(\sqrt{n_i} K_\mathrm{A}/n)$. Moreover, we can write

$$\begin{aligned}
(\hat{\nu}_i^\tau - \nu_i)^\tau &= A_k(x)^\tau \hat{W}_{n,\mathrm{A}}^{-1} \hat{\mathbf{\Psi}}_i^\tau - A_k(x)^\tau W_{n,\mathrm{A}}^{-1} \mathbf{\Psi}_i^\tau \\
&= A_k(x)^\tau W_{n,\mathrm{A}}^{-1} (\hat{\mathbf{\Psi}}_i - \mathbf{\Psi}_i)^\tau + A_k(x)^\tau (\hat{W}_{n,\mathrm{A}}^{-1} - W_{n,\mathrm{A}}^{-1})\mathbf{\Psi}_i + o(s.o.).
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
\left\|\hat{\mathbf{\Psi}}_i - \mathbf{\Psi}_i\right\|_\mathrm{F}^2 &= \sum_{k=1}^p \sum_{j=1}^{n_i} \sum_{l=1}^{J_\mathrm{A}} [\hat{\alpha}_{k,\mathrm{I}}(t_{ij} - \alpha_k(t_{ij})]^2 B_{k,l,\mathrm{A}}^2(x_{ijk}) \\
&= O\Big(n_i K_\mathrm{A}\Big(\bar{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_\mathrm{H}} + \frac{1}{n}\Big)\Big).
\end{aligned}$$

We also have $\|\Psi_{ij}\|_{\rm F}^2 = \sum_{k=1}^{p}\sum_{l=1}^{J_{\rm A}} \alpha_k^2(t_{ij}) B_{k,l,{\rm A}}^2(x_{ijk}) = O(K_{\rm A})$ and

$$\left\|\hat{\Psi}_{ij} - \Psi_{ij}\right\|_{\rm F}^2 = \sum_{k=1}^{p}\sum_{l=1}^{J_{\rm A}} [\hat{\alpha}_{k,{\rm I}}(t_{ij} - \alpha_k(t_{ij})]^2 B_{k,l,{\rm A}}^2(x_{ijk})$$
$$= O\Big(K_{\rm A}\Big(\bar{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\rm H}} + \frac{1}{n}\Big)\Big).$$

Since $\left\|\hat{\Psi}_{ij}\hat{\Psi}_{ij}^{\tau} - \Psi_{ij}\Psi_{ij}^{\tau}\right\|_{\rm F}^2 \le \left\|\hat{\Psi}_{ij} - \Psi_{ij}\right\|_{\rm F}^2 + 2\left\|\Psi_{ij}\right\|_{\rm F} \cdot \left\|\hat{\Psi}_{ij} - \Psi_{ij})^{\tau}\right\|_{\rm F}$, we obtain that

$$\left\|(\hat{W}_{n,{\rm A}} - W_{n,{\rm A}})/n\right\|_{\rm F}^2 = O\Big(K_{\rm A}\Big(\bar{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\rm H}} + \frac{1}{n}\Big)\Big) = o_p(1),$$

which implies $(\hat{W}_{n,{\rm A}}/n)^{-1} \xrightarrow{p} (W_{n,{\rm A}}/n)^{-1}$. Therefore,

$$\|I_{21}\| \le \sum_{i=1}^{n} \frac{1}{n_i^2}\|\hat{\nu}_i - \nu_i\| \cdot \|\nu_i\| = O_p\left(\frac{K_{\rm A}^2}{n\bar{N}_{\rm H}}\left(\bar{h}^{-2r} + \frac{\bar{h}^2}{n\bar{N}_{\rm H}} + \frac{1}{n}\right)\right) = o_p(1).$$

In the same vein, we can show $\|I_{22}\| = o_p(1)$, which completes the proof of part (i). □

S2.6 Proof of Theorem 6

Proof. We only give the proof of (i), and (ii) can be proceeded in a similar way. Without loss of generality, we assume the first l terms are additive terms, i.e.,

$$m_0(t, \mathbf{x}) = \alpha_0(t) + \sum_{k=1}^{l} c_k\beta_k(x_k) + \sum_{k=l+1}^{p} \alpha_k(t)\beta_k(x_k).$$

According to the proposed model identification procedure, the approximation space of m_0 is given by $\mathcal{M}_0 = \big\{p(t, \mathbf{x}; \eta) = \eta_0^{\tau}\mathbf{b}_{\rm CP}(t) + \sum_{k=1}^{l} x_k\eta_k^{\tau}\tilde{M}_k + \sum_{k=l+1}^{p} x_k\eta_k^{\tau}\tilde{Z}_k(t, x_k)\big\}$, where $\tilde{M}_k = \{1, \mathbf{0}_{J_{\rm CP}}^{\tau}, 1, \mathbf{0}_{J_{\rm CP}}^{\tau}, ..., 1, \mathbf{0}_{J_{\rm CP}}^{\tau}\}^{\tau}$, and $\tilde{Z}(t, x_k) = M_k(t, x_k) + \tilde{M}_k$. Let $m_{n,0}$ be any function in $\mathcal{M}_0$ with $\|m_{n,0} - m_0\|_{L_2} = O_p(\varrho_n)$, $\nu(t, x) = x_{\iota}M_{\iota}^{\tau}(t, x)\eta_{\iota}$ satisfies $0 < \|\nu\|_{L_2} \le c\varrho_n$ for some positive constant c, where $\eta_{\iota} \ne 0$ and ι is any element of $\mathcal{I}_{\rm A}$.

It is sufficient to show that $Q(m_{n,0}) \le Q(m_{n,0} + \nu)$, where Q is the objection function of (3.1). For the sake of convenience, we denote $m_{0,ij} = m_0(t_{ij}, \mathbf{x}_{ij})$, $m_{n,ij} = m_{n,0}(t_{ij}, \mathbf{x}_{ij})$ and

$\nu_{ij} = \nu(t_{ij}, x_{ij\iota})$. Employing mean value theorem, we have

$$\begin{aligned}
&Q(m_{n,0}) - Q(m_{n,0} + \nu) \\
&= \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\rho(y_{ij} - m_{n,ij}) - \rho(y_{ij} - m_{n,ij} - \nu_{ij})\right] - np_{\lambda_1}\left(\|\nu\|_{L_2} \right) \\
&= n\Big\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(\varepsilon_{ij} + m_{0,ij} - m_{n,ij} + \delta\nu_{ij})\nu_{ij} - p_{\lambda_1}\left(\|\nu\|_{L_2} \right) \Big\} \\
&= n\big\{ I_1 + I_2 - p_{\lambda_1}\left(\|\nu\|_{L_2} \right) \big\},
\end{aligned}$$

where $\delta \in [0,1]$, $I_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\phi(\varepsilon_{ij} + m_{0,ij} - m_{n,ij} + \delta\nu_{ij}) - \phi(\varepsilon_{ij})\right]\nu_{ij}$, and $I_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(\varepsilon_{ij})\nu_{ij}$.

By Assumption A4, $I_2 = o(1)$. Employing Cauthy-Schwartz inequality and Assumption M2, we obtain that

$$\begin{aligned}
I_1 &\leq \Big\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\phi(\varepsilon_{ij} + m_{0,ij} - m_{n,ij} + \delta\nu_{ij}) - \phi(\varepsilon_{ij})\right]^2 \Big\}^{1/2} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \nu_{ij}^2 \Big\}^{1/2} \\
&\leq \Big\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left| m_{0,ij} - m_{n,ij} + \delta\nu_{ij} \right| \Big\}^{1/2} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \nu_{ij}^2 \Big\}^{1/2} (1 + o_p(1)).
\end{aligned}$$

It is routine to show $\frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \nu_{ij}^2 = \|\nu\|_{L_2}^2 (1 + o_p(1))$. Furthermore, the square of the first term in the last inequality can be bounded by

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} |m_{0,ij} - m_{n,ij}| + |v_{ij}| \leq (\|m_{n,0} - m_0\|_{L_2} + \|\nu\|_{L_2})(1 + o_p(1)) = O_p(\varrho_n).$$

Therefore,

$$Q(m_{n,0}) - Q(m_{n,0} + \nu) \leq n\lambda_1 \|\nu\|_{L_2} \left\{ \frac{O_p(\sqrt{\varrho}_n)}{\lambda_1} - \frac{p_{\lambda_1}(\| \nu \|_{L_2})}{\lambda_1 \|\nu\|_{L_2}} \right\} \leq 0,$$

which completes the proof. □

S3 Algorithm

We now formulate the algorithm for optimization problem (4.1) in the main text. Following the LQA procedure introduced by Fan and Li (2001), we have

$$p_{\lambda_1}\left(\|M_k^\tau \eta_k\|_{L_2}\right) \approx p_{\lambda_1}\left(\left\|M_k^\tau \eta_k^0\right\|_{L_2}\right) + \frac{1}{2}\frac{p'_{\lambda_1}\left(\left\|M_k^\tau \eta_k^0\right\|_{L_2}\right)}{\|M_k^\tau \eta_k^0\|_{L_2}}\Big(\|M_k^\tau \eta_k\|_{L_2}^2 - \left\|M_k^\tau \eta_k^0\right\|_{L_2}^2\Big),$$

$$p_{\lambda_2}\left(\|F_k^\tau \eta_k\|_{L_2}\right) \approx p_{\lambda_2}\left(\left\|F_k^\tau \eta_k^0\right\|_{L_2}\right) + \frac{1}{2}\frac{p'_{\lambda_2}\left(\left\|F_k^\tau \eta_k^0\right\|_{L_2}\right)}{\|F_k^\tau \eta_k^0\|_{L_2}}\Big(\|F_k^\tau \eta_k\|_{L_2}^2 - \left\|F_k^\tau \eta_k^0\right\|_{L_2}^2\Big),$$

where η_k^0 is a given initial estimate of η_k such that $\left\|M_k^\tau \eta_k^0\right\|_{L_2} > 0$ and $\left\|F_k^\tau \eta_k^0\right\|_{L_2} > 0$ for each $k = 1, ..., p$. Moreover, we approximate the first term of (4.1) as

$$\rho(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta}) \approx \frac{\rho(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta}^{(0)})}{(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta}^{(0)})^2}(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta})^2 = \omega_{ij}(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta})^2,$$

where $Z_{ij} = \{\mathbf{b}_{\mathrm{C}}^\tau(t_{ij}), x_{ij1}\mathcal{T}_1^\tau(t_{ij}, x_{ij1}), ..., x_{ijp}\mathcal{T}_p^\tau(t_{ij}, x_{ijp})\}^\tau$, with $\mathcal{T}_k^\tau(t, x) = \{1, \mathbf{B}_{k,\mathrm{AP}}^\tau(x)\} \otimes \{1, \mathbf{B}_{\mathrm{CP}}^\tau(t)\}$ and $\omega_{ij} = \rho(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta}^{(0)})/(y_{ij} - Z_{ij}^\tau \boldsymbol{\eta}^{(0)})^2$.

Furthermore, let $\mathcal{Y} = (\mathbf{y}_1^\tau, ..., \mathbf{y}_n^\tau)^\tau$ with $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\tau$, $\mathcal{Z} = (\mathbf{Z}_1^\tau, ..., \mathbf{Z}_n^\tau)^\tau$ with $\mathbf{Z}_i = (Z_{i1}, ..., Z_{in_i})^\tau$, and $W = \mathrm{diag}(W_1, ..., W_n)$ with $W_i = n_i^{-1}\mathrm{diag}(\omega_{i1}, ..., \omega_{in_i})$. Denote

$$\Omega_1 = \mathrm{diag}\Big(0, \frac{p'_{\lambda_1}\left(\left\|M_1^\tau \eta_1^0\right\|_{L_2}\right)}{\|M_1^\tau \eta_1^0\|_{L_2}} M_1 M_1^\tau, ..., \frac{p'_{\lambda_1}\left(\left\|M_p^\tau \eta_p^0\right\|_{L_2}\right)}{\left\|M_p^\tau \eta_p^0\right\|_{L_2}} M_p M_p^\tau\Big),$$

$$\Omega_2 = \mathrm{diag}\Big(0, \frac{p'_{\lambda_2}\left(\left\|F_1^\tau \eta_1^0\right\|_{L_2}\right)}{\|F_1^\tau \eta_1^0\|_{L_2}} F_1 F_1^\tau, ..., \frac{p'_{\lambda_2}\left(\left\|F_p^\tau \eta_p^0\right\|_{L_2}\right)}{\left\|F_p^\tau \eta_p^0\right\|_{L_2}} F_p F_p^\tau\Big).$$

We can then approximate $Q(\boldsymbol{\eta})$ in (4.1), up to a constant, as

$$Q(\boldsymbol{\eta}; \lambda_1, \lambda_2) \approx (\mathcal{Y} - \mathcal{Z}\boldsymbol{\eta})^\tau W(\mathcal{Y} - \mathcal{Z}\boldsymbol{\eta}) + \tfrac{1}{2}n\boldsymbol{\eta}^\tau(\Omega_1 + \Omega_2)\boldsymbol{\eta},$$

which implies that the minimizer of (4.1) can be derived by iteratively computing the estimator $\hat{\boldsymbol{\eta}} = \left(\mathcal{Z}^\tau W \mathcal{Z} + \tfrac{1}{2}n\{\Omega_1 + \Omega_2\}\right)^{-1}\mathcal{Z}^\tau W \mathcal{Y}$ until convergence.

References

Baltagi BH, Li D (2004). *Prediction in the Panel Data Model with Spatial Correlation.* In Advances in Spatial Econometrics: Methodology, Tools and Applications, 1st edition, pp. 283–295.

Fan, J. and Zhang, J. T. (2000a). Functional linear models for longitudinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **62**, 303–322.

Fan, J., and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96(456)**, 1348–1360.

He, X., and Shi, P. (1994). Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, **3**, 299–308.

He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89**, 579–590.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111–128.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**, 763–788.

Schumaker, L. L. (1981). *Spline Functions, Basic Theory.* Cambridge Mathematical Library.

Shi, P., and Li, G. (1995). Global convergence of B-spline M-estimators in nonparametric regression. *Statistica Sinica*, **5**, 303–318.

Stone, Charles J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics*,textbf13, 689–705.

Tang, Q., and Cheng, L. (2008). M-estimation and B-spline approximation for varying coefficient models with longitudinal data. *Journal of Nonparametric Statistics*, **20**, 611–625.

Thomas Scheike (2019). timereg: Flexible regression models for survival data. R package version 2.15.

Zhang, X., and Wang, J.-L. (2015). Varying-coefficient additive models for functional data. *Biometrika*, **102**, 15–32.