Statistica Sinica

Supplementary Material for "Sparse Composite Quantile Regression with Ultrahigh-dimensional Heterogeneous Data

Lianqiang Qu[1], Meiling Hao[2] and Liuquan Sun[3]

[1]*Central China Normal University*,[2]*University of International*

[3]*Business and Economics and Chinese Academy of Sciences*

The online Supplementary Material includes simulation studies, some sufficient conditions for (C3), and the proofs of Proposition 1 and Theorems 1−6 in the main text.

1. Simulation Studies

In this section, simulation studies are conducted to examine the finite sample performance of the SCQR method. For comparison, several other methods are also considered: SQR(τ) and USQR; the ℓ_1-penalized quantile regression in high-dimensional sparse models in Belloni and Chernozhukov (2011) over Θ, denoted by L1QR; the method of simply taking the union of active covariate sets identified by quantile-adaptive model-free variable screening method (He, Wang, and Hong (2013)), denoted by QaSIS; the method of Xu and Chen (2014) with a ℓ_1-penalized procedure, denoted by PXC; and the sure screening method (Fan and Lv (2008)), denoted by SIS. The following criteria are used to compare the performance of different methods:

Cor: mean number of correctly identified variables (with nonzero coefficient functions);

Inc: mean number of incorrectly selected variables;

UF: a proportion of under-fitted models;

CF: a proportion of correctly fitted models;

OF: a proportion of over-fitted models;

Err: the average of estimation errors, defined as $[\int_{\Theta} \|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^\star_\tau\|_2^2 d\tau]^{1/2}$.

We set $\Theta = [0.1, 0.9]$ and choose $\tau = 0.25$, 0.5, and 0.75 for the SQR(τ). Typically, we use equally spaced grid points with the step size 0.01 on $[0.1, 0.9]$. The C_n in the EBIC is taken as $\log(p)/2$, which is also used by Lee, Noh, and Park (2014). If not specified, the sample size is $n = 200$, and the number of covariates is $p = 1000$. For the bandwidth, a smaller h leads to a higher computation burden, while a larger one may result in some bias of the estimate, and will make our method difficult to distinguish the important covariates from the unimportant ones. We use the rule of thumb bandwidth and choose $h = 1.9n^{-1/3} \approx 0.32$, which performs well in our simulation studies. All simulation results are based on 500 replications.

Scenario 1. We first consider a random coefficient model to compare the performance of methods based on a continuous range of quantile levels and that with a single or multiple quantile levels. Let $Z_1, \ldots, Z_p$ be independent standard normal random variables, and U be a uniform random variable on $(-2, 2)$. Then define $x_j = |Z_j + U|/2$. The relationship between the covariates and the response is $Y = \beta^\star_{\tau,1} x_1 + \beta^\star_{\tau,2} x_2 + \beta^\star_{\tau,3} x_3$. Set $\beta^\star_{\tau,1} = 1$, $\beta^\star_{\tau,2} = 2I(\tau > 0.5)$, and $\beta^\star_{\tau,3} = \exp(\tau^2)$, where τ follows a uniform distribution on $(0, 1)$.

Scenario 2. We consider a homogeneous error model. The covariates $(x_1, \ldots, x_p)^\top$ follow a multivariate normal distribution $N_p(0, \Sigma)$, where $\Sigma = (\sigma_{kl})_{p\times p}$ with $\sigma_{kl} = 0.5$ for $k \neq l$ and $k \neq 5, l \neq 5$, $\sigma_{5l} = \sigma_{l5} = 0$ for $l \neq 5$, and $\sigma_{kl} = 1$ for $k = l$. Thus, x_5 is uncorrelated with x_j $(j \neq 5)$. The response is generated from the following model: $Y = \beta x_1 + \beta x_2 + \beta x_3 - 1.5\beta x_4 + 0.25\beta x_5 + \epsilon$, where ϵ is a random error term, and β is set to be 2.5. Under this setting, the marginal correlation of x_4 and Y is zero, and x_5 has a small contribution to Y without "borrowing" strength from other covariates. This model was also considered by Fan and Lv (2008).

Scenario 3. We consider a heteroscedastic location-scale model, and generate $Z = (z_1, \ldots, z_p)^\top$ from $N_p(0, \Sigma)$ with $\Sigma = (\sigma_{lk})_{p\times p}$ and $\sigma_{lk} = 0.5^{|l-k|}$. Then let $x_j = z_j$ for $j \neq 3$, and $x_3 = |z_3|$. The scalar response is generated from $Y = x_1 + 2x_3 - x_{10} + 2x_3\epsilon$. This implies that the coefficient of x_3 in model (2.1) is $2\{1 + \Phi^{-1}(\tau)\}$, which has a monotone behaviour in τ, governed by the quantile function $\Phi^{-1}(\tau)$.

Scenario 4. We revise a challenging example from Wang (2009). Specifically, we first independently generate Z_j and W_j from the standard multivariate normal distribution. Then set $x_j = (Z_j + W_j)/\sqrt{2}$ for every $1 \leq j \leq 5$, and $x_j = (Z_j + \sum_{j'=1}^{5} Z_{j'})/2$ for every $5 < j \leq p$. The response is generated from $Y = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + \epsilon$. Note that it would be extremely difficult to identify x_1 (for example) as an active covariate, because the correlation of x_1 and Y is much smaller than that of x_j and Y for each $j > 5$.

Scenario 5. We consider the case that the number of the nonzero coefficients is diverging with the sample size n. The covariates $(x_1, \ldots, x_p)^\top$ are generated from $N_p(0, \Sigma)$ with $\Sigma = (\sigma_{lk})_{p\times p}$ and $\sigma_{lk} = 0.5^{|l-k|}$. Let the true model size $s = [\sqrt{n} - 5]$, and $\beta_j^\star = (-1)^{U_{1j}} U_{2j}$ for $1 \leq j \leq s$, where U_{1j} is from a Bernoulli distribution with success probability 0.4, and U_{2j} is from a uniform distribution on $(2\log(n)/\sqrt{n}, 4\log(n)/\sqrt{n})$. Other coefficients are set to be zero. The response is generated from $Y = \sum_{j=1}^{s} x_j \beta_j^\star + \epsilon$. The sample sizes are $n = 200$ and 400.

For Scenarios 2−5, the error term ϵ is independent of the covariates and follows from the standard normal distribution $N(0, 1)$ or a t-distribution with degree of freedom 3, denoted by $t(3)$. The simulation results for Scenarios 1−5 are reported in Tables A.1−A.5. In Scenario 1, SQR(0.25) fails to identify the covariate x_2, yielding Cor= 2.00 and UF= 1.00. SQR(0.5) has a similar performance to SQR(0.25). On the other hand, the covariate x_2 has a strong effect on the response over $\Theta = [0.1, 0.9]$. Therefore, the SCQR successfully identifies the active covariate x_2, yielding CF= 0.82. Under Scenario 2, when the error term follows from $N(0, 1)$, we observe that both the SCQR and SQR(τ), with $\tau \in \{0.25, 0.5, 0.75\}$, have very similar performance in terms of model selection and estimation accuracy. Specifically, the values of Cor are both 5 (the true model size), and the values of Inc are close to 0. Besides, the SCQR has a slightly better performance in estimation accuracy than the other methods. The results also suggest that the USQR and QaSIS methods tend to yield an over-fitted model (2.1) with OF=0.87 and 0.86, respectively. Both the SIS and L1QR tend to yield an under-fitted model with UF= 1.00. Moreover, when the error term follows from a $t(3)$ distribution, SQR(τ) has an unstable performance with CF varying from 0.20 to 0.79. The results of Scenario 3 are similar to those of Scenario 2. Under the challenging case of Scenario 4, the QaSIS and SIS both tend to yield an under-fitted model, while the other methods are all over-fitted. But the SCQR still enjoys the smallest Err (=0.23) and Inc (=0.94) with the error $N(0,1)$, and Err (=0.18) and Inc (=0.97) with the error $t(3)$. For Scenario 5, Table A.5 suggests that our method can also handle the case with the true model size diverging as the sample size n increasing. Based on the results of Tables A.1−A.5, we also observe that the SCQR method performs better than the PXC in terms of correct-fitting. For example, under the setting of Scenario 2 with the error $N(0,1)$, the SCQR method yields CF$=0.99$, while the PXC method only has CF$=0.09$, due to over-fitting. On the other hand, when the data follow from a heteroscedastic location-scale model (Scenario 3), the SCQR procedure is more robust than the PXC, especially for heavy-tailed data.

Furthermore, to gain a direct insight into the efficiency of our algorithm, the average computational times (in seconds) over 500 replications for the SCQR, USQR, QaSIS and L1QR are reported in Table A.6. The results indicate that the SCQR costs the least time among the four robust methods, namely our method is fast to choose the correct model and estimate the coefficients. We also conduct simulation studies to assess the impact of h in practice. The simulation results under the setting of Scenario 2 are reported in Table A.7. The results demonstrate that the model selection results of our method are robust with the bandwidth h varying in a proper interval. The results for the other scenarios are similar and are not shown here to save space.

Scenario 6. We assess the performance of the proposed method in model selection and estimation accuracy at tail quantiles. The covariates $(x_1, \ldots, x_p)^\top$ are generated from $N_p(0, \Sigma)$ with $\Sigma = (\sigma_{lk})_{p\times p}$ and $\sigma_{lk} = 0.5^{|l-k|}$. The nonzero coefficients are set to be $\beta^\star_{\tau,1} = 1$, $\beta^\star_{\tau,3} = 4/3$, $\beta^\star_{\tau,5} = 1$ and $\beta^\star_{\tau,10} = 2$. Other coefficients are set to be zero. The response is generated from $Y = \beta^\star_{\tau,1}x_1 + \beta^\star_{\tau,3}x_3 + \beta^\star_{\tau,5}x_5 + \beta^\star_{\tau,10}x_{10} + \epsilon$, where ϵ follows from $N(0, 1)$. We implement the SQR(0.06) as well as the SCQR, USQR, L1QR, and QaSIS with $\Theta = [0.05, 0.1]$ and $\Theta = [0.04, 0.09]$, respectively. The results are reported in Table A.8. We observe that there is a slight variability in model selection for the SQR when Θ varies from $[0.05, 0.1]$ to $[0.04, 0.09]$. For the SCQR, the results are very similar for the two choices of Θ. These suggest that the method focusing on examination of model sparsity at a single or at multiple quantile levels may be sensitive to the choice of quantile levels in model selection, while the SCQR approach is robust.

Table A.1. *Simulation results for Scenario 1.*

Method	Cor	Inc	UF	CF	OF	Err
SCQR	2.83	0.02	0.16	0.82	0.02	0.31
USQR	2.99	1.90	0.01	0.31	0.68	0.37
SQR(0.25)	2.00	0.00	1.00	0.00	0.00	0.88
SQR(0.5)	1.82	0.06	0.98	0.01	0.01	1.06
SQR(0.75)	2.88	0.34	0.07	0.72	0.20	1.05
L1QR	2.99	93.63	0.01	0.00	0.99	1.61
QaSIS	2.98	201.4	0.02	0.00	0.98	-
PXC	2.71	4.56	0.26	0.03	0.70	-
SIS	1.58	13.42	0.86	0.00	0.14	-

Table A.2. *Simulation results for Scenario 2.*

	$N(0,1)$						$t(3)$					
Method	Cor	Inc	UF	CF	OF	Err	Cor	Inc	UF	CF	OF	Err
SCQR	5.00	0.01	0.00	0.99	0.01	0.19	4.77	0.00	0.23	0.77	0.00	0.31
USQR	5.00	3.07	0.00	0.13	0.87	0.23	4.94	2.28	0.06	0.17	0.77	0.37
SQR(0.25)	5.00	0.05	0.01	0.94	0.05	0.24	4.50	0.25	0.49	0.20	0.31	0.82
SQR(0.5)	5.00	0.01	0.00	0.99	0.01	0.21	4.74	0.01	0.26	0.70	0.04	0.37
SQR(0.75)	4.99	0.04	0.01	0.95	0.04	0.24	4.81	0.01	0.21	0.79	0.00	0.31
L1QR	4.00	87.6	1.00	0.00	0.00	3.28	3.92	93.2	1.00	0.00	0.00	3.46
QaSIS	4.85	178.4	0.14	0.00	0.86	-	4.87	179.8	0.12	0.00	0.88	-
PXC	5.00	2.97	0.00	0.09	0.91	-	4.92	2.89	0.07	0.09	0.84	-
SIS	3.00	12.00	1.00	0.00	0.00	-	2.99	12.01	1.00	0.00	0.00	-

Table A.3. *Simulation results for Scenario 3.*

	$N(0,1)$						$t(3)$					
Method	Cor	Inc	UF	CF	OF	Err	Cor	Inc	UF	CF	OF	Err
SCQR	2.90	0.07	0.08	0.87	0.05	0.34	2.43	0.02	0.27	0.71	0.02	0.72
USQR	2.99	2.02	0.01	0.20	0.79	0.64	2.92	0.98	0.06	0.38	0.55	1.02
SQR(0.25)	1.96	0.03	0.99	0.01	0.00	1.67	1.76	0.00	1.00	0.00	0.00	1.72
SQR(0.5)	2.60	0.14	0.27	0.63	0.10	1.17	2.17	0.01	0.46	0.53	0.01	1.36
SQR(0.75)	2.72	0.29	0.17	0.69	0.14	1.35	2.22	0.15	0.35	0.56	0.09	1.55
L1QR	1.99	90.58	1.00	0.00	0.00	1.98	2.00	106.0	1.00	0.00	0.00	1.61
QaSIS	2.97	147.7	0.04	0.00	0.96	-	2.96	152.7	0.03	0.00	0.97	-
PXC	2.88	0.41	0.06	0.62	0.32	-	1.06	0.06	0.79	0.17	0.04	-
SIS	2.44	12.56	0.46	0.00	0.54	-	2.68	13.3	0.28	0.00	0.72	-

Table A.4. *Simulation results for Scenario 4.*

	$N(0,1)$						$t(3)$					
Method	Cor	Inc	UF	CF	OF	Err	Cor	Inc	UF	CF	OF	Err
SCQR	5.00	0.97	0.00	0.03	0.97	0.18	5.00	0.94	0.00	0.06	0.94	0.23
USQR	5.00	11.51	0.00	0.00	1.00	0.33	5.00	7.66	0.00	0.00	1.00	0.34
SQR(0.25)	5.00	1.13	0.00	0.04	0.96	0.36	5.00	0.96	0.00	0.08	0.92	0.37
SQR(0.5)	5.00	1.03	0.00	0.03	0.97	0.29	5.00	0.96	0.00	0.05	0.95	0.29
SQR(0.75)	5.00	1.12	0.00	0.04	0.96	0.34	4.99	0.95	0.01	0.08	0.91	0.36
L1QR	5.00	88.8	0.00	0.00	1.00	1.00	5.00	99.03	0.00	0.00	1.00	5.87
QaSIS	2.77	184.6	0.87	0.00	0.13	-	1.25	282.5	1.00	0.00	0.00	-
PXC	5.00	2.52	0.00	0.01	0.99	-	4.99	2.43	0.01	0.00	0.98	-
SIS	0.11	14.89	1.00	0.00	0.00	-	0.09	15.9	1.00	0.00	0.00	-

Table A.5. *Simulation results for Scenario 5.*

		$n = 200$						$n = 400$					
Error	Method	Cor	Inc	UF	CF	OF	Err	Cor	Inc	UF	CF	OF	Err
$N(0, 1)$	SCQR	8.87	0.17	0.04	0.83	0.13	0.31	14.99	0.03	0.01	0.97	0.02	0.26
	USQR	8.91	7.60	0.03	0.01	0.96	0.91	14.99	6.49	0.00	0.00	1.00	0.46
	SQR(0.25)	7.74	0.67	0.29	0.38	0.33	0.89	14.82	0.31	0.05	0.76	0.19	0.38
	SQR(0.5)	8.39	0.40	0.15	0.59	0.26	0.59	14.97	0.10	0.01	0.90	0.09	0.30
	SQR(0.75)	7.62	0.65	0.30	0.35	0.34	0.94	14.88	0.36	0.04	0.71	0.25	0.37
	L1QR	3.49	81.87	1.00	0.00	0.00	2.39	6.11	166.3	1.00	0.00	0.00	2.49
	QaSIS	8.65	171.0	0.26	0.00	0.74	-	14.50	163.5	0.37	0.00	0.63	-
	PXC	8.98	5.83	0.01	0.02	0.98	-	15.00	4.00	0.00	0.00	1.00	-
	SIS	6.65	8.35	0.95	0.00	0.05	-	10.96	8.28	0.99	0.00	0.01	-
$t(3)$	SCQR	5.88	0.02	0.63	0.35	0.02	1.32	13.12	0.03	0.32	0.66	0.02	0.70
	USQR	7.23	1.47	0.48	0.11	0.41	1.93	14.21	2.10	0.18	0.14	0.68	1.45
	SQR(0.25)	4.13	0.09	0.90	0.07	0.03	2.35	9.53	0.17	0.74	0.18	0.08	1.80
	SQR(0.5)	5.57	0.09	0.66	0.26	0.07	1.73	2.57	0.10	0.37	0.56	0.07	0.96
	SQR(0.75)	4.12	0.08	0.90	0.07	0.03	2.38	9.93	0.22	0.68	0.19	0.13	1.70

Table A.6. *Average computational times (in seconds) for robust methods.*

	$N(0, 1)$				$t(3)$			
Scenario	SCQR	USQR	L1QR	QaSIS	SCQR	USQR	L1QR	QaSIS
1	35.9	58.5	485.4	129.0	-	-	-	-
2	45.7	176.6	472.2	123.8	68.7	205.6	503.9	129.0
3	65.0	101.4	297.0	90.3	82.9	121.2	295.1	90.4
4	50.2	182.1	511.5	131.0	58.4	199.6	499.7	133.1
5	54.2	124.0	295.3	93.0	72.1	129.5	319.2	100.0

Table A.7. *Simulation results of the SCQR method for Scenario 2 with different bandwidths.*

	$N(0,1)$						$t(3)$					
h	Cor	Inc	UF	CF	OF	Err	Cor	Inc	UF	CF	OF	Err
0.16	4.98	0.00	0.02	0.98	0.00	0.21	4.66	0.00	0.34	0.66	0.00	0.37
0.32	5.00	0.01	0.00	0.99	0.01	0.19	4.77	0.00	0.23	0.77	0.00	0.31
0.48	4.99	0.00	0.01	0.99	0.00	0.19	4.69	0.00	0.31	0.69	0.00	0.35

Table A.8. *Simulation results for Scenario 6.*

	$\Theta = [0.04, 0.09]$						$\Theta = [0.05, 0.1]$					
Method	Cor	Inc	UF	CF	OF	Err	Cor	Inc	UF	CF	OF	Err
SCQR	3.99	7.66	0.01	0.01	0.98	0.02	4.00	4.60	0.00	0.07	0.93	0.02
USQR	4.00	23.16	0.00	0.00	1.00	0.02	4.00	17.63	0.00	0.00	1.00	0.02
SQR(0.06)	3.97	10.48	0.02	0.00	0.97	0.73	3.98	8.95	0.02	0.01	0.98	0.69
L1QR	4.00	81.99	0.00	0.00	1.00	0.04	4.00	82.30	0.00	0.00	1.00	0.04
QaSIS	3.34	67.37	0.51	0.00	0.49	-	3.42	65.76	0.48	0.00	0.52	-

2. Sufficient Conditions for (C3)

Some sufficient conditions for (C3) are presented in the following examples, and the proofs are included at the end of the Supplementary. Let $M \in M_{+}^{2t}$. *Example 1.* Suppose that $n^{-1}\sum_{i=1}^{n} f(0|\mathbf{X}_i)\mathbf{X}_{i,M}\mathbf{X}_{i,M}^{\top}$ is positive define. This condition is considered by Koenker (2005) Then condition (C3) is satisfied for sufficiently large n. *Example 2.* By *Example 1*, we can also assume that $f(0|\mathbf{X}_i) > \underline{f}$ for a constant $\underline{f} > 0$, and $n^{-1}\sum_{i=1}^{n} \mathbf{X}_{i,M}\mathbf{X}_{i,M}^{\top}$ is bounded away from 0. Then condition (C3) holds.

Example 3. Suppose that the following conditions hold: (I) there exist positive constants $\underline{f}$ and $\bar{f}$, which are independent of τ and x, such that $\underline{f} < f(0|x) < \bar{f}$. (II) $\sup_{u,x} |f(u|x) - f(0|x)| \leq A_0|u|$ for a constant $A_0 > 0$. (III) The eigenvalues of $n^{-1}\sum_{i=1}^{n} \mathbf{X}_{i,M}\mathbf{X}_{i,M}^{\top}$ are bounded from below and above by some positive constants $\tilde{m}_1$ and $\tilde{m}_2$, respectively. (IV) Assume that $\{\sum_{i=1}^{n}(\mathbf{X}_i^{\top}\boldsymbol{u})^2\}^{3/2}/(\sqrt{n}\sum_{i=1}^{n}|\mathbf{X}_i^{\top}\boldsymbol{u}|^3) = \tilde{m}_3 > 0$, where the support of $\boldsymbol{u}$ belongs to M_{+}^{2t} and $\boldsymbol{u} \neq 0$, and $\tilde{m}_3$ is a constant free of $\boldsymbol{u}$. Conditions (I) and (II) are considered by Zheng, Peng, and He (2015), while condition (III) is a variant version of their condition (C4). Condition (IV) is similar to condition (D.5) of Belloni and Chernozhukov (2011). If $\|\boldsymbol{u}\|_2 \leq \underline{f}\tilde{m}_1\tilde{m}_3/(2A_0\tilde{m}_2^{3/2})$, condition (C3) is satisfied for sufficiently large

n.

3. Technical Proofs

In this section, we establish the proofs of Proposition 1 and Theorems 1−6 in the main text. For any two numbers, $a \lesssim (\gtrsim)\ b$ means that there exists a constant $c > 0$, satisfying $a \leq (\geq)\ cb$.

Lemma 1. *If $h \to 0$ and $nh^4 \to 0$, then we have that uniformly in τ and $\boldsymbol{\beta}_\tau$ with supp$(\boldsymbol{\beta}_\tau) \in M_+^{2t}$,*

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau) = \frac{1}{n}\sum_{i=1}^{n} \psi_{\tau,h}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau) + o_p(n^{-1/2}).$$

Proof. Let $\tilde{\epsilon}_i = Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau$ for the ith subject. With a slight abuse of notation, denote $F(\cdot)$ as the cumulative distribution function of $\tilde{\epsilon}_i$. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau) - \frac{1}{n}\sum_{i=1}^{n} \psi_{\tau,h}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau)\\
=&\frac{1}{n}\sum_{i=1}^{n} \tilde{\epsilon}_i \big[\Phi(-\tilde{\epsilon}_i/h) - I(\tilde{\epsilon}_i < 0)\big]\\
=&\int s\Phi(-s/h)[d\hat{F}(s) - dF(s)] - \int sI(s<0)[d\hat{F}(s) - dF(s)]\\
&+\int s[\Phi(-s/h) - I(s<0)]dF(s|X)\\
\equiv&U_1 + U_2 + U_3,
\end{aligned}$$

where $\hat{F}(\cdot)$ is the empirical cumulative distribution of $\tilde{\epsilon}_i, i = 1, \ldots, n$. Let

$u = sh^{-1}$. For U_1 and U_2, the integration by parts gives

$$\begin{aligned} U_1 &= -h \int [\hat{F}(uh) - F(uh)] d\{u\Phi(-u)\} \\ &= -h \int [\hat{F}(uh) - F(uh)]\Phi(-u)du + h \int [\hat{F}(uh) - F(uh)]u\phi(-u)du \\ &\equiv U_{11} + U_{12}, \\ U_2 &= \int [\hat{F}(u) - F(u)] I(u < 0)du. \end{aligned}$$

Using oscillations of the empirical process (Shorack and Wellner (1986), p.531), we have that uniformly in $\boldsymbol{\beta}_\tau$ and τ,

$$|U_{11} + U_2| = O_p\Big(\Big[\frac{h\log(n)}{n}\log\Big(\frac{1}{h\log(n)}\Big)\Big]^{1/2}\Big),$$

and

$$|U_{12}| = O_p\Big(\Big[\frac{h\log(n)}{n}\log\Big(\frac{1}{h\log(n)}\Big)\Big]^{1/2}\Big).$$

Applying the integration by parts and the Taylor's expansion, we obtain that $|U_3| = O_p(h^2)$, which is free of $\boldsymbol{\beta}_\tau$ and τ. This, together with the results of U_1 and U_2, gives

$$\Big|\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau) - \frac{1}{n}\sum_{i=1}^{n} \psi_{\tau,h}(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}_\tau)\Big| \le |U_1 + U_2| + |U_3|,$$

which is of order

$$O_p\Big(\Big[\frac{h\log(n)}{n}\log\Big(\frac{1}{h\log(n)}\Big)\Big]^{1/2} + h^2\Big).$$

Thus, the conclusion follows from $h \to 0$ and $nh^4 \to 0$. □

3.1 Proofs of Proposition 1 and Theorems 1−2

Proof of Proposition 1. Suppose that B is an optimal solution, and $\bar{M} = \{j : \|\boldsymbol{b}_j\|_2 \neq 0\}$ is the support of B. We can rewrite the objective function as

$$\|B-D\|_F^2 = \sum_{j\in\bar{M}} \|\boldsymbol{b}_j - \boldsymbol{d}_j\|_2^2 + \sum_{j\notin\bar{M}} \|\boldsymbol{d}_j\|_2^2.$$

Note that each term in the right hand side of the above equation is not negative. Then set $\boldsymbol{b}_j = \boldsymbol{d}_j$ for $j \in \bar{M}$, and the objective function becomes $\sqrt{\sum_{j\notin\bar{M}} \|\boldsymbol{d}_j\|_2^2}$. Thus, to minimize the objective function, $\bar{M}$ must correspond to the indices of the largest t values of d_j^*.

Proof of Theorem 1. To prove Theorem 1, we first show the convergence of $\{\tilde{\mathcal{U}}(D^{[l]})\}$. For this, it suffices to show that the inequality (3.6) in the main text holds since $\{\tilde{\mathcal{U}}(D^{[l]})\}$ is bounded above. An application of Lipschitz condition of $\ell_\tau(\boldsymbol{\beta})$ gives that for any $\lambda \geq \phi$,

$$\tilde{\mathcal{U}}(B) \leq \tilde{\mathcal{U}}(D) + \frac{1}{K}\sum_{k=1}^{K} \left\langle \check{\boldsymbol{\beta}}_{\tau_k} - \boldsymbol{\beta}_{\tau_k}, \dot{\ell}_{\tau_k}(\boldsymbol{\beta}_{\tau_k}) \right\rangle + \frac{\lambda}{2K}\|B-D\|_F^2 = \mathcal{Q}_\lambda(B|D). \tag{C.1}$$

By the definitions of $\mathcal{Q}_{\lambda}(B|D)$ and $D^{[l+1]}$, we have

$$\begin{aligned}
\tilde{\mathcal{U}}(D^{[l]}) =&\mathcal{Q}_{\lambda^{[l]}}(D^{[l]}|D^{[l]}) \geq \mathcal{Q}_{\lambda^{[l]}}(D^{[l+1]}|D^{[l]}) \\
=&\tilde{\mathcal{U}}(D^{[l]}) + \frac{1}{K}\sum_{k=1}^{K}\big\langle \boldsymbol{\beta}_{\tau_k}^{[l+1]} - \boldsymbol{\beta}_{\tau_k}^{[l]}, \dot{\ell}_{\tau_k}(\boldsymbol{\beta}_{\tau_k}^{[l]})\big\rangle + \frac{\lambda^{[l]}}{2K}\|D^{[l+1]} - D^{[l]}\|_F^2 \\
=&\tilde{\mathcal{U}}(D^{[l]}) + \frac{1}{K}\sum_{k=1}^{K}\big\langle \boldsymbol{\beta}_{\tau_k}^{[l+1]} - \boldsymbol{\beta}_{\tau_k}^{[l]}, \dot{\ell}_{\tau_k}(\boldsymbol{\beta}_{\tau_k}^{[l]})\big\rangle + \frac{\phi}{2K}\|D^{[l+1]} - D^{[l]}\|_F^2 \\
&+ \frac{\lambda^{[l]} - \phi}{2K}\|D^{[l+1]} - D^{[l]}\|_F^2 \\
=&\mathcal{Q}_{\phi}(D^{[l+1]}|D^{[l]}) + \frac{\lambda^{[l]} - \phi}{2K}\|D^{[l+1]} - D^{[l]}\|_F^2. \qquad \text{(C.2)}
\end{aligned}$$

By (C.1) and (C.2), we obtain

$$\tilde{\mathcal{U}}(D^{[l]}) \geq \tilde{\mathcal{U}}(D^{[l+1]}) + \frac{\lambda^{[l]} - \phi}{2K}\|D^{[l+1]} - D^{[l]}\|_F^2, \qquad \text{(C.3)}$$

which implies that the inequality (3.6) in the main text holds whenever $\lambda^{[l]} \geq \phi/(1-\varrho)$. Because $\tilde{\mathcal{U}}(D^{[l]})$ has an upper bound, (C.3) implies that $\{\tilde{\mathcal{U}}(D^{[l]})\}$ has at least one limiting point in the feasible region.

We now show that there exists a subsequence such that $D^{[l]}$ is convergent. When $\lambda^{[l]} \to \infty$ as $l \to \infty$, the result is trivial. Next, we assume that $\{\lambda^{[l]}\}$ is bounded, and hence there exists a subsequence $\mathcal{S}$ such that $\lambda^{[l]} \to \tilde{\lambda}$. For each $l \in \mathcal{S}$, denote the support of $D^{[l]}$ by $\mathcal{M}^{[l]}$, that is, $\mathcal{M}^{[l]} = \{j : K^{-1}\sum_{k=1}^{K}(\beta_{\tau_k,j}^{[l]})^2 \neq 0\}$. By (C.3) and the fact that $\tilde{\mathcal{U}}(D^{[l]})$ is convergent, we know that $\|D^{[l+1]} - D^{[l]}\|_F^2 \to 0$ as $l \in \mathcal{S}$ goes to infinity, which implies that $\mathcal{M}^{[l]}$ is also convergent. Also, $\mathcal{M}^{[l]}$ is a discrete sequence, and hence there exists a constant $l^* \in \mathcal{S}$ such that $\mathcal{M}^{[l]} = \mathcal{M}^{[l^*]}$ for all $l \in \mathcal{S}$ and $l \geq l^*$. Thus, Algorithm 1 becomes a gradient descent algorithm on the space $\mathcal{M}^{[l]}$ for all $l \in \mathcal{S}$ and $l \geq l^*$. Since a gradient descent algorithm for minimizing a convex function over a closed convex set yields a sequence of iterations that converges (Nestenrov (2004)), we conclude that the subsequence $\{D^{[l]} : l \in \mathcal{S}\}$ is convergent.

Next, we prove that Algorithm 1 will stop in a finite number of steps. By (C.3) and assuming that $\lambda^{[l]} \geq \phi/(1-\varrho)$, we get

$$\begin{aligned}\sum_{l=0}^{L}\{\tilde{\mathcal{U}}(D^{[l]}) - \tilde{\mathcal{U}}(D^{[l+1]})\} \geq & \frac{\lambda^{[l]} - \phi}{2K}\sum_{l=0}^{L}\|D^{[l+1]} - D^{[l]}\|_F^2 \\ \geq & \frac{\varrho\phi}{2K(1-\varrho)}\sum_{l=0}^{L}\|D^{[l+1]} - D^{[l]}\|_F^2,\end{aligned}$$

which yields that

$$\min_{0\leq l\leq L}\|D^{[l+1]} - D^{[l]}\|_F^2 \leq \frac{2K(1-\varrho)}{\varrho\phi}\frac{\{\tilde{\mathcal{U}}(D^{[0]}) - \tilde{\mathcal{U}}(D^{[l+1]})\}}{L}.$$

Let $\lambda^\star = 2(1-\varrho)/(\varrho\phi)$. By the above proofs, we know that the decreasing sequence $\tilde{\mathcal{U}}(D^{[l]})$ has at least one limiting point, denoted by $\tilde{\mathcal{U}}(D^\star)$. Then we have

$$\begin{aligned}\min_{0\leq l\leq L}\|D^{[l+1]} - D^{[l]}\|_F^2 \leq & \frac{\lambda^\star K\{\tilde{\mathcal{U}}(D^{[0]}) - \tilde{\mathcal{U}}(D^{[L+1]})\}}{L} \\ \leq & \frac{\lambda^\star K\{\tilde{\mathcal{U}}(D^{[0]}) - \tilde{\mathcal{U}}(D^\star)\}}{L}.\end{aligned}$$

Note that $\lambda^\star$ and $\tilde{\mathcal{U}}(\cdot)$ are bounded above. Thus, for any $\varepsilon > 0$, there exists $L = O(1/\varepsilon^2)$ such that for some $1 \le \tilde{l} \le L$, $K^{-1}\|D^{[\tilde{l}+1]} - D^{[\tilde{l}]}\|_F^2 \lesssim \varepsilon^2$. In other words, Algorithm 1 stops in a finite number of steps with the stopping criteria being $K^{-1/2}\|D^{[l+1]} - D^{[l]}\|_F \lesssim \varepsilon$.

Proof of Theorem 2. Note that

$$\begin{aligned}\|D^{[l+1]} - D^\star\|_F =& \left\|\left\{D^{[l+1]} - D^{[l]} + \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]})\right\} + \left\{D^{[l]} - \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]}) - D^\star\right\}\right\|_F \\ \le& \left\|D^{[l+1]} - D^{[l]} + \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]})\right\|_F + \left\|D^{[l]} - \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]}) - D^\star\right\|_F.\end{aligned}$$

By the definition of $D^{[l+1]}$ and $s \le t$, we have

$$\left\|D^{[l+1]} - \left\{D^{[l]} - \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]})\right\}\right\|_F \le \left\|D^\star - \left\{D^{[l]} - \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]})\right\}\right\|_F.$$

Thus,

$$\|D^{[l+1]} - D^\star\|_F \le 2\left\|D^{[l]} - \frac{1}{\lambda^{[l]}}\dot{\Psi}(D^{[l]}) - D^\star\right\|_F.$$

The Taylor's expansion yields that

$$\begin{aligned}&\|D^{[l]} - D^\star - (\lambda^{[l]})^{-1}\dot{\Psi}(D^{[l]})\|_F \\ =& \sqrt{\sum_{k=1}^K \|\boldsymbol{\beta}^{[l]}_{\tau_k} - \boldsymbol{\beta}^\star_{\tau_k} - \frac{1}{\lambda^{[l]}}\dot{\ell}_{\tau_k}(\boldsymbol{\beta}^\star_{\tau_k}) - \frac{1}{\lambda^{[l]}}\ddot{\ell}_{\tau_k}(\tilde{\boldsymbol{\beta}}_{\tau_k})(\boldsymbol{\beta}^{[l]}_{\tau_k} - \boldsymbol{\beta}^\star_{\tau_k})\|_2^2} \\ \le& \sqrt{2}\sqrt{\sum_{k=1}^K \left[\|(\mathbf{I} - \frac{1}{\lambda^{[l]}}\ddot{\ell}_{\tau_k}(\tilde{\boldsymbol{\beta}}_{\tau_k}))(\boldsymbol{\beta}^{[l]}_{\tau_k} - \boldsymbol{\beta}^\star_{\tau_k})\|_2^2 + \frac{1}{\lambda^{[l]}}\|\dot{\ell}_{\tau_k}(\boldsymbol{\beta}^\star_{\tau_k})\|_2^2\right]} \\ \le& \sqrt{2}\sqrt{\sum_{k=1}^K \|(\mathbf{I} - \frac{1}{\lambda^{[l]}}\ddot{\ell}_{\tau_k}(\tilde{\boldsymbol{\beta}}_{\tau_k}))(\boldsymbol{\beta}^{[l]}_{\tau_k} - \boldsymbol{\beta}^\star_{\tau_k})\|_2^2} + \sqrt{\frac{2}{\lambda^{[l]}}}\sqrt{\sum_{k=1}^K \|\dot{\ell}_{\tau_k}(\boldsymbol{\beta}^\star_{\tau_k})\|_2^2},\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{\tau_k}$ lies between $\boldsymbol{\beta}_{\tau_k}^{[l]}$ and $\boldsymbol{\beta}_{\tau_k}^{\star}$, and $\mathbf{I}$ denotes the identity matrix. The last two inequalities are due to the fact that $(a+b)^2 \leq 2(a^2+b^2)$ and $\sqrt{\sum_{k=1}^{K}(a_i^2+b_i^2)} \leq \sqrt{\sum_{k=1}^{K} a_i^2} + \sqrt{\sum_{k=1}^{K} b_i^2}$, respectively. This, combining with $\phi < \lambda^{[l]} < \tilde{\phi}/[1-1/(4\sqrt{2})]$, yields that

$$\|D^{[l+1]} - D^{\star}\|_F \leq 2^{-1}\|D^{[l]} - D^{\star}\|_F + \sqrt{\frac{2}{\phi}}\|\dot{\Psi}(D^{\star})\|_F.$$

Iterating this relationship, we obtain

$$\|D^{[l]} - D^{\star}\|_F \leq 2^{-l}\|D^{[0]} - D^{\star}\|_F + \sqrt{\frac{8}{\phi}}\|\dot{\Psi}(D^{\star})\|_F.$$

This completes the proof.

3.2 Proofs of Theorems 3–6

Lemma 2. *Under condition (C4), there exist positive constants c_1 and b such that for $1 \leq j \leq p$,*

$$P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}[I(\epsilon_i \leq 0) - \tau]\Big| \geq b\Big\} \leq 2\exp\{-c_1 n b^2\}.$$

Proof. Since $\{I(\epsilon_i \leq 0) : i = 1, \ldots, n\}$ are i.i.d. Bernoulli random variables with mean τ, it follows from the Hoeffding's inequality (Hoeffding (1963)) that

$$P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}[I(\epsilon_i \leq 0) - \tau]\Big| \geq b\Big\} \leq 2\exp\{-c_1 n b^2\}.$$

This completes the proof. □

Proof of Theorem 3. Without loss of generality, we assume that $\Theta = (0, 1)$. Define the collections of under-fitted models with model size t as

$M_-^t = \{M : M^\star \nsubseteq M_t\}$. For any $M \in M_-^t$, define $M' = M \cup M^\star \in M_+^{2t}$. Consider $\boldsymbol{\beta}_{\tau,M}$ close to $\boldsymbol{\beta}^\star_{\tau,M}$ such that $\int_\Theta \|\boldsymbol{\beta}_{\tau,M'} - \boldsymbol{\beta}^\star_{\tau,M'}\|_2^2 d\tau = (\omega_1 n^{-\kappa_1})^2$ for some ω_1 and $\kappa_1 > 0$. Let $\boldsymbol{u}_{k,M'} = \hat{\boldsymbol{\beta}}_{\tau_k,M'} - \boldsymbol{\beta}^\star_{\tau_k,M'}$. Then the definition of integration implies that $(\omega_1 n^{-\kappa_2})^2/2 \leq K^{-1}\sum_{k=1}^K \|\boldsymbol{u}_{k,M'}\|_2^2 \leq (2\omega_1 n^{-\kappa_1})^2$ for sufficiently large K. Set $\Delta = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K) = \hat{D} - D^\star$. Further define the following criterion:

$$\mathcal{O}_n(\Delta_{M'}) = \frac{1}{nK}\sum_{k=1}^K\sum_{i=1}^n \Big[\rho_{\tau_k}\big(\epsilon_{ik} - \mathbf{X}_{i,M'}^\top \boldsymbol{u}_{k,M'}\big) - \rho_{\tau_k}\big(\epsilon_{ik}\big)\Big],$$

where $\epsilon_{ik} = Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}^\star_{\tau_k}$, and $\Delta_{M'}$ is the sub-matrix of Δ associated with the index of rows in M'. By the identity (Knight (1998))

$$|r-s| - |r| = -s\big\{I(r>0) - I(r<0)\big\} + 2\int_0^s \big[I(r \leq t) - I(r \leq 0)\big]dt,$$

we have

$$\rho_\tau(r-s) - \rho_\tau(r) = s\big\{I(r<0) - \tau\big\} + \int_0^s \big[I(r \leq t) - I(r \leq 0)\big]dt.$$

Thus, $\mathcal{O}_n(\Delta_{M'})$ can be written as follows:

$$\begin{aligned}\mathcal{O}_n(\Delta_{M'}) =& \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{X}_{i,M'}^{\top}\boldsymbol{u}_{k,M'}\big[I(\epsilon_{ik}\le 0)-\tau_k\big] \\ &+\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_{0}^{\mathbf{X}_{i,M'}^{\top}\boldsymbol{u}_{k,M'}}\big[I(\epsilon_{ik}\le t)-I(\epsilon_{ik}\le 0)\big]dt \\ \ge& -\frac{1}{K}\sum_{k=1}^{K}\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,M'}[I(\epsilon_{ik}\le 0)-\tau_k]\Big\|_2\|\boldsymbol{u}_{k,M'}\|_2 \\ &+\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_{0}^{\mathbf{X}_{i,M'}^{\top}\boldsymbol{u}_{k,M'}}\big[I(\epsilon_{ik}\le t)-I(\epsilon_{ik}\le 0)\big]dt \\ =& -I_1+I_2. \end{aligned} \tag{C.6}$$

Let

$$I_{2k}=n^{-1}\sum_{i=1}^{n}\int_{0}^{\mathbf{X}_{i,M'}^{\top}\boldsymbol{u}_{k,M'}}\big[I(\epsilon_{ik}\le t)-I(\epsilon_{ik}\le 0)\big]dt.$$

An application of Hoeffding's inequality (Hoeffding (1963)) yields that for some c_2 and $\nu_1>0$,

$$P\{|I_{2k}-EI_{2k}|>(\nu_1/2)n^{-2\xi_2}\}\le 2\exp\{-c_2n^{1-4\xi_2}\}.$$

This, combining with conditions (C2) and (C4), implies that with probability greater than $1-2n\exp\{-c_2n^{1-4\xi_2}\}$,

$$\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_{0}^{\mathbf{X}_{i,M'}^{\top}\boldsymbol{u}_{k,M'}}\big[I(\epsilon_{ik}\le t)-I(\epsilon_{ik}\le 0)\big]dt\ge\frac{\nu_1}{2K}\sum_{k=1}^{K}\|\boldsymbol{u}_{k,M'}\|_2^2.$$

Define the event

$$\Omega=\Big\{I_2>\frac{\nu_1}{2K}\sum_{k=1}^{K}\|\boldsymbol{u}_{k,M'}\|_2^2\Big\}.$$

In view of (C.6), to prove $P\{\mathcal{O}_n(\Delta_{M'}) < 0\} \to 0$, it suffices to establish $P\{I_1 > I_2\} \to 0$. On the event Ω, we obtain that for some positive constant ν_2,

$$
\begin{aligned}
& P\{I_1 > I_2\} \\
\leq & P\Big\{\frac{1}{K}\sum_{k=1}^{K} \Big\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,M'}[I(\epsilon_{ik} \leq 0) - \tau_k]\Big\|_2 \|\boldsymbol{u}_{k,M'}\|_2 \geq \frac{\nu_1}{2K}\sum_{k=1}^{K} \|\boldsymbol{u}_{k,M'}\|_2^2\Big\} \\
\leq & P\Big\{\Big[\max_{1\leq k\leq K} \Big\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,M'}[I(\epsilon_{ik} \leq 0) - \tau_k]\Big\|_2\Big]\Big[\frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{u}_{k,M'}\|_2\Big] \geq \frac{\nu_1}{2K}\sum_{k=1}^{K} \|\boldsymbol{u}_{k,M'}\|_2^2\Big\} \\
\leq & P\Big\{\Big[\max_{1\leq k\leq K} \Big\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,M'}[I(\epsilon_{ik} \leq 0) - \tau_k]\Big\|_2\Big]\Big[\frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{u}_{k,M'}\|_2^2\Big]^{1/2} \geq \frac{\nu_1}{2K}\sum_{k=1}^{K} \|\boldsymbol{u}_{k,M'}\|_2^2\Big\} \\
\leq & P\Big\{\Big[\max_{1\leq k\leq K} \Big\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,M'}[I(\epsilon_{ik} \leq 0) - \tau_k]\Big\|_2\Big] \geq \frac{\nu_1\omega_1}{4} n^{-\xi_2}\Big\} \\
\leq & \sum_{k=1}^{K}\sum_{j\in\mathcal{M}'} P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}[I(\epsilon_{ik} \leq 0) - \tau_k]\Big| \geq \nu_2 t^{-1/2} n^{-\xi_2}\Big\}. \qquad \text{(C.7)}
\end{aligned}
$$

By Lemma 2 and condition (C4), there exist some positive constants c_3 and c_4 such that

$$
P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}[I(\epsilon_{ik} \leq 0) - \tau_k]\Big| \geq \nu_2 t^{-1/2} n^{-\xi_2}\Big\} \leq c_3 \exp\big\{-c_4 n^{1-\xi_1-2\xi_2}\big\}. \qquad \text{(C.8)}
$$

Note that $P\{\Omega^c\} \leq 2n\exp\{-c_2 n^{1-4\xi_2}\}$. This, together with (C.7) and (C.8), yields that

$$
P\{\mathcal{U}(D^{\star}_{M'}) \geq \mathcal{U}(D_{M'})\} \leq 2n\exp\{-c_2 n^{1-4\xi_2}\} + c_3 K t \exp\big\{-c_4 n^{\frac{(1-\xi_1-2\xi_2)\alpha}{\alpha+2}}\big\}. \qquad \text{(C.9)}
$$

Therefore, the Bonferroni inequality implies that

$$
\begin{aligned}
&P\big\{\mathcal{U}(D^{\star}_{M'}) \geq \min_{M\in M^t_-} \mathcal{U}(D_M)\big\}\\
\leq& 2np^t \exp\{-c_2 n^{1-4\xi_2}\} + c_3 K n^{\xi_1} p^t \exp\big\{-c_4 n^{1-\xi_1-2\xi_2}\big\}.
\end{aligned}
$$

Note that $\mathcal{U}(D)$ is strictly convex. Thus, the above results hold for any $\boldsymbol{\beta}_\tau$ such that $\sqrt{\int_\Theta \|\boldsymbol{\beta}_{\tau,M'} - \boldsymbol{\beta}^\star_{\tau,M'}\|_2^2 d\tau} \geq \omega_2 n^{-\xi_2}$. For any $M \in M^t_-$, let $\tilde{\boldsymbol{\beta}}_{\tau,M'}$ be the augmented vector of $\beta_{\tau,M}$ with zeros corresponding to the elements in $M'/M^\star$, where $M'/M^\star$ is the complement set of $M^\star$ in M'. Since $M' = \{M \cup (M^\star/M)\} \cup \{M'/M^\star\}$, it follows from condition (C2) that $\|\tilde{\boldsymbol{\beta}}_{\tau,M'} - \boldsymbol{\beta}^\star_{\tau,M'}\|_2 \geq \|\boldsymbol{\beta}_{\tau,M'/M^\star}\|_2$, and hence $\sqrt{\int_\Theta \|\tilde{\boldsymbol{\beta}}_{\tau,M'} - \boldsymbol{\beta}^\star_{\tau,M'}\|_2^2 d\tau} \geq \omega_2 n^{-\xi_2}$. Then the definition of integration implies that $(\omega_1 n^{-\kappa_2})^2/2 \leq K^{-1}\sum_{k=1}^K \|\tilde{\boldsymbol{\beta}}_{\tau_k,M'} - \boldsymbol{\beta}^\star_{\tau_k,M'}\|_2^2 \leq (2\omega_1 n^{-\kappa_1})^2$ for sufficiently large K. Consequently, we have that under condition $\max\{\xi_1+\xi_2, \xi_1/2+2\xi_2\} < (1-\xi_0)/2$,

$$
\begin{aligned}
&P\big\{\max_{M\in M^t_+} \mathcal{U}(D_M) \geq \min_{M\in M^t_-} \mathcal{U}(D_M)\big\}\\
\leq& P\big\{\mathcal{U}(D^{\star}_{M'}) \geq \min_{M\in M^t_-} \mathcal{U}(\tilde{D}_M)\big\}\\
\leq& 2np^t \exp\{-c_2 n^{1-4\xi_2}\} + c_3 K n^{\xi_1} p^t \exp\big\{-c_4 n^{1-\xi_1-2\xi_2}\big\} \to 0.
\end{aligned}
$$

This completes the proofs.

Proof of Corollary 1. If $\hat{M}_s \not\subset M^\star$, then there exists some $j \in \hat{M}_s$ but $j \notin M^\star$. Thus, it follows from Theorem 3 that $|M^\star| < |\hat{M}_s|$ with probability tending to 1. This is contradictory to the fact $|M^\star| = |\hat{M}_s|$. Hence, we must have $\hat{M}_s \subset M^\star$ with probability tending to 1. This, together with Theorem 3, yields the conclusion.

Lemma 3. *Suppose that conditions (C1)–(C4) hold. Then for a constant* $c_5 > 0$,

$$\lim_{n\to\infty} P\Big\{\frac{1}{\sqrt{K}}\|\hat{D} - D^\star\|_F \le c_5 n^{-\xi_2}\Big\} = 1.$$

Proof. By the convexity of $\rho_\tau(\cdot)$, it suffices to show that for any given $a > 0$,

$$\liminf_{n\to\infty} P\Big\{\inf_{M\in M_+^t} \inf_{\|\Delta\|_F = b} \mathcal{O}_n(\Delta_M) > 0\Big\} \ge 1 - a,$$

where $b = c_5\sqrt{K}n^{-\xi_2}$. Following similar arguments to the proofs of (C.5), (C.7) and (C.9), we can obtain the conclusion. □

Proof of Theorem 4. For any $M \in M_-^t$, define $M' = M \cup M^\star \in M_+^{2t}$. Similarly to (C.7), we can get that with probability tending to 1,

$$\frac{\nu_1}{4} n^{-2\xi_2} < \mathcal{O}_n(\Delta) = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\big[\rho_{\tau_k}\big(\epsilon_{ik} - \mathbf{X}_{i,M'}^\top \boldsymbol{u}_{k,M'}\big) - \rho_{\tau_k}\big(\epsilon_{ik}\big)\big] \le \nu_1 n^{-2\xi_2}. \tag{C.10}$$

Under the assumption $E(|\epsilon|) < \infty$, we obtain that $n^{-1}\sum_{i=1}^n \rho_{\tau_k}(\epsilon_{ik}) \to_p E\{\rho_{\tau_k}(\epsilon_{ik})\}$, and $\nu_3^{-1} < E\{\rho_{\tau_k}(\epsilon)\} < \nu_3$ for some constant $0 < \nu_3 < \infty$. Therefore, we have that with probability tending to 1,

$$\frac{1}{\nu_3} < \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}\big(\epsilon_{ik} - \mathbf{X}_{i,M'}^\top \boldsymbol{u}_{k,M'}\big) < 2\nu_3. \tag{C.11}$$

In view of Lemma 3, (C.11) and condition (C3), following similar arguments to the proof of (A.20) in the online Supplementary Material of Lee, Noh, and Park (2014), we can show that there exists some constant $\nu_4 > 0$ (independent of $M \in M^t_-$) such that with probability tending to 1,

$$\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\left[\rho_{\tau_k}(Y_i - \mathbf{X}_{i,M}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M}) - \rho_{\tau_k}(Y_i - \mathbf{X}_{i,M'}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M'})\right] > 2\nu_4 > 0. \tag{C.12}$$

Then it follows from (C.11) and (C.12) that with probability tending to 1,

$$\begin{aligned}
&\min_{M\in M^t_-} \mathrm{EBIC}(M) - \mathrm{EBIC}(M') \\
&= \min_{M\in M^t_-}\left[\log\left\{1 + \frac{(nK)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}[\rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M}\big) - \rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M'}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M'}\big)]}{(nK)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M'}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M'}\big)}\right\}\right. \\
&\quad\left. + (|M| - |M'|)C_n\frac{\log(n)}{2n}\right] \\
&\geq \min_{M\in M^t_-}\min\left\{\log(2), \frac{(nK)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}[\rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M}\big) - \rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M'}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M'}\big)]}{2(nK)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M'}^\top\hat{\boldsymbol{\beta}}_{\tau_k,M'}\big)}\right\} \\
&\quad - tC_n\frac{\log(n)}{n} \\
&\geq \min_{M\in M^t_-}\left\{\log(2), \frac{\nu_4}{2\nu_3}\right\} - \omega_1 n^{\xi_1}C_n\frac{\log(n)}{n} > 0,
\end{aligned} \tag{C.13}$$

where the first inequality follows from $\log(1+x) \geq \min\{x/2, \log(2)\}$ for any $x > 0$, and the last inequality follows from the assumption $C_n\log(n)/n^{1-\xi_1} = o(1)$. By (C.13), we have that for any underfitted model M with size t, there exists an overfitted model $M' = M \cup M^\star$ such that $\mathrm{EBIC}(M) > \mathrm{EBIC}(M')$

with probability tending to 1 as $n \to \infty$. Thus, to prove Theorem 4, it suffices to show

$$P\{\min_{M \in M_+^t} \mathrm{EBIC}(M) > \mathrm{EBIC}(M^\star)\} \to 1. \tag{C.14}$$

Note that

$$\begin{aligned}
&\min_{M \in M_+^t} \mathrm{EBIC}(M) - \mathrm{EBIC}(M^\star) \\
=& \min_{M \in M_+^t} \Big[\log\Big\{1 + \frac{(nK)^{-1}\sum_{k=1}^K \sum_{i=1}^n [\rho_{\tau_k}\big(Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{\tau_k,M}\big) - \rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M^\star}^\top \hat{\boldsymbol{\beta}}_{\tau_k,M^\star}\big)]}{(nK)^{-1}\sum_{k=1}^K \sum_{i=1}^n \rho_{\tau_k}\big(Y_i - \mathbf{X}_{i,M^\star}^\top \hat{\boldsymbol{\beta}}_{\tau_k,M^\star}\big)}\Big\} \\
&\quad + (|M| - |M^\star|)C_n \frac{\log(n)}{2n}\Big] \\
\geq& \min_{M \in M_+^t} \min\Big\{ \log(2), \frac{(nK)^{-1}\sum_{k=1}^K \sum_{i=1}^n \big[\rho_{\tau_k}(Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{\tau_k,M}) - \rho_{\tau_k}(Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{\tau_k,M^\star})\big]}{2(nK)^{-1}\sum_{k=1}^K \sum_{i=1}^n \rho_{\tau_k}(Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{\tau_k,M^\star})}\Big\} \\
&\quad + (|M| - |M^\star|)C_n \frac{\log(n)}{n} \\
\geq& \min_{M \in M_+^t} \min\{\log(2), \frac{\nu_1}{4\nu_3} n^{-2\xi_2}\} + C_n \frac{\log(n)}{n} > 0,
\end{aligned}$$

where the second inequality follows from (C.10) and (C.11). This implies that (C.14) holds, and hence completes the proof.

Proof of Theorem 5. Note that for sufficiently large K,

$$\begin{aligned}
\Big[\int_\Theta \|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau^\star\|_2^2 d\tau\Big]^{1/2} \leq& \Big[\frac{1}{K}\sum_{k=1}^K \|\hat{\boldsymbol{\beta}}_{\tau_k} - \boldsymbol{\beta}_{\tau_k}^\star\|_2^2 + (c_5 n^{-\xi_2})^2\Big]^{1/2} \\
\leq& \Big[\frac{1}{K}\sum_{k=1}^K \|\hat{\boldsymbol{\beta}}_{\tau_k} - \boldsymbol{\beta}_{\tau_k}^\star\|_2^2\Big]^{1/2} + c_5 n^{-\xi_2} \\
=& \frac{1}{\sqrt{K}} \|\hat{D} - D^\star\|_F + c_5 n^{-\xi_2},
\end{aligned}$$

where the first inequality follows from the definition of integration, and the second inequality follows from $\sqrt{|a|+|b|} \leq \sqrt{|a|} + \sqrt{|b|}$. Therefore, it follows from Lemma 3 that

$$P\Big\{\big[\int_{\Theta} \|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^\star_\tau\|_2^2 d\tau\big]^{1/2} \geq 2c_5 n^{-\xi_2}\Big\} \leq P\Big\{\frac{1}{\sqrt{K}}\|\hat{D} - D^\star\|_F \geq c_5 n^{-\xi_2}\Big\} \to 0.$$

This completes the proof.

Proof of Theorem 6. Let $\mathcal{M}_\tau = \hat{M}_\tau \cup M^\star$. As in the proof of Theorem 5, it suffices to show that for sufficiently large K,

$$P\Big\{\frac{1}{\sqrt{K}}\|D^{[l]} - D^\star\|_F \geq c_5 n^{-\xi_2}\Big\} \to 0 \quad \text{as} \quad n \to \infty. \tag{C.15}$$

To show (C.15), in view of the proof of Theorem 2, we have

$$\begin{aligned} P\Big\{\frac{1}{\sqrt{K}}\|D^{[l]}_{\mathcal{M}_\tau} - D^\star_{\mathcal{M}_\tau}\|_F \geq c_5 n^{-\xi_2}\Big\} \leq & P\Big\{\frac{1}{\sqrt{K}}\|\dot{\Psi}_{\mathcal{M}_\tau}(D^\star)\|_F \geq c_5 n^{-\xi_2}\Big\} \\ \leq & 2K\tau P\Big\{|\dot{\ell}_{\tau_k,j}(\boldsymbol{\beta}^\star_{\tau_k})| \geq \tilde{c}_5 n^{-0.5\xi_1 - \xi_2}\Big\}, \end{aligned} \tag{C.16}$$

where $\tilde{c}_5 = c_5/(2\omega_1)$. By the definition of $\ell_\tau(\boldsymbol{\beta})$, we get

$$\begin{aligned} |\dot{\ell}_{\tau_k,j}(\boldsymbol{\beta}^\star_{\tau_k})| \leq & \Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\big[\tau_k - 1 + \Phi(\frac{\epsilon_{ik}}{h})\big]\Big| + \Big|\frac{1}{nh}\sum_{i=1}^n x_{ij}\epsilon_{ik}\dot{\Phi}(\frac{\epsilon_{ik}}{h})\Big| \\ = & \Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\big[\tau_k - 1 + \Phi(\frac{\epsilon_{ik}}{h})\big]\Big| + \Big|\frac{1}{\sqrt{2\pi}nh}\sum_{i=1}^n x_{ij}\epsilon_{ik}\exp\{-\frac{\epsilon_{ik}^2}{2h^2}\}\Big|. \end{aligned} \tag{C.17}$$

Note that $\exp\{-\epsilon_{ik}^2/(2h^2)\} = o_p(hn^{-\xi_2})$ as $h \to 0$. Then for sufficiently large n,

$$\Big|\frac{1}{nh}\sum_{i=1}^n x_{ij}\epsilon_{ik}\exp\{-\frac{\epsilon_{ik}^2}{2h^2}\}\Big| = o_p(n^{-\xi_2})\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\epsilon_{ik}\Big|.$$

Moreover, the assumption $E(|\epsilon|) < \infty$ and condition (C4) imply that $n^{-1}|\sum_{i=1}^n x_{ij}\epsilon_{ik}| < \infty$. Therefore, we get that the second term in the right hand side of (C.17) is of order $o_p(n^{-\xi_2})$.

For the first term in the right hand side of (C.17), it can be checked that

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\big[\tau_k - 1 + \Phi(\frac{\epsilon_{ik}}{h})\big]\Big| \\
=&\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\big[(\tau_k - I(\epsilon_{ik} < 0)) + (\Phi(\frac{\epsilon_{ik}}{h}) - I(\epsilon_{ik} \geq 0))\big]\Big| \\
\leq&\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}[\tau_k - I(\epsilon_{ik} < 0)]\Big| + \Big|\frac{1}{n}\sum_{i=1}^n x_{ij}[\Phi(\frac{\epsilon_{ik}}{h}) - I(\epsilon_{ik} \geq 0)]\Big|. \qquad \text{(C.18)}
\end{aligned}$$

Note that $[\Phi(\epsilon_{ik}/h) - I(\epsilon_{ik} \geq 0)] = \Phi(\epsilon_{ik}/h)$ if $\epsilon_{ik} < 0$, and $[\Phi(\epsilon_{ik}/h) - I(\epsilon_{ik} \geq 0)] = -\Phi(-\epsilon_{ik}/h)$ if $\epsilon_{ik} \geq 0$. Thus, we have

$$\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}[\Phi(\frac{\epsilon_{ik}}{h}) - I(\epsilon_{ik} \geq 0)]\Big| = o_p(hn^{-\xi_2})\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\Big|.$$

Then by condition (C4), we obtain that the second term in the right hand side of (C.18) is also of order $o_p(n^{-\xi_2})$. It follows from (C.16)$-$(C.18) and

Lemma 2 that for some positive constants c_6 and c_7,

$$\begin{aligned}
P\Big\{\frac{1}{\sqrt{K}}\|D^{[l]}-D^{\star}\|_F \geq c_5 n^{-\xi_2}\Big\} \leq & KpP\Big\{|\dot{\ell}_{\tau_k,j}(\boldsymbol{\beta}^{\star}_{\tau_k})| \geq c_5 n^{-\xi_2}\Big\} \\
\leq & KpP\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} x_{ij}[\tau_k - I(\epsilon_{ik}<0)]\Big| > c_5 n^{-\xi_2}/8\Big\} \\
\leq & c_6 Kp\exp\{-c_7 n^{1-\xi_1-2\xi_2}\},
\end{aligned}$$

which implies that (C.15) holds with $0<(\xi_1+2\xi_2)<(1-\xi_0)$.

3.3 Proofs of Examples 1−3

Example 1. For simplicity, we assume that for each vector $\boldsymbol{u} \in \{\boldsymbol{u} : \|\boldsymbol{u}_M\|_2 < \delta,\ M \in M_+^{2t}\}$, $\mathbf{X}_i^{\top}\boldsymbol{u} > 0$ for all $1 \leq i \leq n$. Define $B = n^{-1}\sum_{i=1}^{n} f(0|\mathbf{X}_i)\mathbf{X}_{i,M}\mathbf{X}_{i,M}^{\top}$. Note that

$$\begin{aligned}
\frac{1}{n^{1-\xi_2}}\sum_{i=1}^{n}\int_0^{\mathbf{X}_i^{\top}\boldsymbol{u}}\big[F(\frac{s}{n^{\xi_2}}|\mathbf{X}_i)-F(0|\mathbf{X}_i)\big]ds =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\mathbf{X}_i^{\top}\boldsymbol{u}}\big[sf(0|\mathbf{X}_i)+o(s)\big]ds \\
=& \frac{1}{2n}\sum_{i=1}^{n} f(0|\mathbf{X}_i)(\mathbf{X}_{i,M}^{\top}\boldsymbol{u}_M)^2 + o(\boldsymbol{u}^{\top}B\mathbf{u}) \\
\geq & \frac{1}{4}\boldsymbol{u}_M^{\top}B\mathbf{u}_M \geq v\|\boldsymbol{u}\|_2^2,
\end{aligned}$$

where the first inequality follows from the fact that the largest eigenvalue of B is bounded above by some positive constant. The second inequality holds because the smallest eigenvalue of B is bounded away from 0.

Example 2. The proof of *Example 2* can be obtained with a slight modification of the proof of Example 1.

Example 3. Note that

$$
\begin{aligned}
&\frac{1}{n^{1-\xi_2}}\sum_{i=1}^{n}\int_0^{\mathbf{X}_i^\top \boldsymbol{u}}\big[F(\frac{s}{n^{\xi_2}}|\mathbf{X}_i)-F(0|\mathbf{X}_i)\big]ds\\
\geq&\frac{1}{n}\sum_{i=1}^{n}\int_0^{\mathbf{X}_i^\top \boldsymbol{u}}\big(sf(0|\mathbf{X}_i)-A_0s^2\big)ds \geq \frac{1}{2n}\sum_{i=1}^{n} f(0|\mathbf{X}_i)(\mathbf{X}_i^\top\boldsymbol{u})^2-\frac{A_0}{3n}\sum_{i=1}^{n}|\mathbf{X}_i^\top\boldsymbol{u}|^3\\
=&\frac{1}{2}\boldsymbol{u}_M^\top\Big(n^{-1}\sum_{i=1}^{n} f(0|\mathbf{X}_i)\mathbf{X}_{i,M}\mathbf{X}_{i,M}^\top\Big)\boldsymbol{u}_M-\frac{A_0}{3}\frac{\big[\boldsymbol{u}_M^\top(n^{-1}\sum_{i=1}^{n}\mathbf{X}_{i,M}\mathbf{X}_{i,M}^\top)\boldsymbol{u}_M\big]^{3/2}}{\tilde{m}_3}\\
\geq&\frac{\underline{f}\tilde{m}_1}{2}\|u\|_2^2-\frac{A_0}{3}\frac{(\|\boldsymbol{u}\|_2^2)^{3/2}\tilde{m}_2^{3/2}}{\tilde{m}_3}\\
=&\|u\|_2^2\big(\frac{\underline{f}\tilde{m}_1}{2}-\frac{A_0}{3}\frac{\|\boldsymbol{u}\|_2\tilde{m}_2^{3/2}}{\tilde{m}_3}\big),
\end{aligned}
$$

where the first inequality follows from the mean value theorem and condition (II), and the third equality follows from conditions (I), (III) and (IV). If we take $\|u\|_2 \leq \underline{f}\tilde{m}_1\tilde{m}_3/(2A_0\tilde{m}_2^{3/2})$, we obtain that condition (C3) holds with $v = \underline{f}\tilde{m}_1/3$.

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high dimensional sparse models. *The Annals of Statistics* **39**, 82-130.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B* **70**, 849-911.

He, X., Wang, L. and Hong, H. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous Data. *The Annals of Statistics* **41**, 342-69.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**, 13-30.

Knight, K. (1998). Limiting distributions for ℓ_1 regression estimators under general conditions. *The Annals of Statistics* **26**, 755-770.

Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

Lee, E., Noh, H. and Park, B. (2014). Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association* **109**, 216-229.

Nestenrov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course.* Boston Kluwer Academic Publishers.

Shorack, G. and Wellner, J. (1986). *Empirical Processes with Applications to Statistics.* New York: Wiley.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512-1524.

Xu, C. and Chen, J. (2014). The sparse MLE for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association* **109**, 1257-1269.

Zheng, Q., Peng, L. and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics* **43**, 2225-2258.