Graduate School of Economics, Hitotsubashi University
Discussion Paper Series No. 2017-04

Adaptively weighted group Lasso for semiparametric quantile regression models

Toshio HONDA, Ching-Kang ING, and Wei-Ying WU

April 2017

Adaptively weighted group Lasso for semiparametric quantile regression models

Toshio Honda, Ching-Kang Ing, and Wei-Ying Wu

Abstract

We propose an adaptively weighted group Lasso procedure for simultaneous variable selection and structure identification for varying coefficient quantile regression models and additive quantile regression models with ultra-high dimensional covariates. Under a strong sparsity condition, we establish selection consistency of the proposed Lasso procedure when the weights therein satisfy a set of general conditions. This consistency result, however, is reliant on a suitable choice of the tuning parameter for the Lasso penalty, which can be hard to make in practice. To alleviate this difficulty, we suggest a BIC-type criterion, which we call high-dimensional information criterion (HDIC), and show that the proposed Lasso procedure with the tuning parameter determined by HDIC still achieves selection consistency. Our simulation studies support strongly our theoretical findings.

Keywords: Additive models; B-spline; high-dimensional information criteria; Lasso; structure identification; varying coefficient models.

Toshio Honda is Professor, Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan (Email: t.honda@r.hit-u.ac.jp). Ching-Kang Ing is Professor, Institute of Statistics, National Tsing Hua University, 101, Sec. 2, Kuang-Fu Rd., Hsinchu 30013, Taiwan (Email: cking@stat.nthu.edu.tw). Wei-Ying Wu is Assistant Professor, Department of Applied Mathematics, National Dong Hwa University, Taiwan (wuweiying1011@gmail.com). Honda's research was supported by JSPS KAKENHI Grant Number JP 16K05268. Ing's research was supported in part by the Science Vanguard Research Program of the Ministry of Science and Technology (MOST), Taiwan.

1 Introduction

We propose adaptively weighted group Lasso (AWG-Lasso) procedures for simultaneous variable selection and structure identification for varying coefficient quantile regression models and additive quantile regression models with ultra-high dimension covariates. Let the number of covariates be denoted by p. Throughout this paper, we assume $p = O(\exp(n^\iota))$, where n is the sample size and ι is a positive constant specified later in Assumption A4 and A4' of Section 5. Under a strong sparsity condition, we establish selection consistency of AWG-Lasso when its weights, determined by some initial estimates, e.g., Lasso and group Lasso, obey a set of general conditions. This consistency result, however, is reliant on a suitable choice for the tuning parameter for the Lasso penalty, which can be hard to make in practice. To alleviate this difficulty, we suggest a BIC-type criterion, which we call high-dimensional information criterion (HDIC), and show that AWG-Lasso with the penalty determined by HDIC (denoted by AWG-Lasso+HDIC hereafter) still achieves selection consistency. This latter result improves previous ones in [20] and the BIC results in [36] since the former does not deal with semiparametric models and the latter concentrates on linear models. See also [4] and [18] for recent developments in BIC-type model selection criteria. With the selected model, one can conduct final statistical inference by appealing to the results in [32] or [26]. Moreover, our approach can be implemented at several different quantiles, thereby leading to a deeper understanding of the data in hand.

High dimensional covariate issues have been important and intractable ones. However, some useful procedures have been proposed, for example, the SCAD in [9], the Lasso in [28], and the group Lasso in [34] and [24]. The properties of the Lasso were studied in [37] and [2]. The adaptive Lasso was proposed by [37] and it has the selection consistency property. The SCAD cannot deal with too many covariates and needs some screening procedures such as the SIS procedure in [11]. [14] proposed a quantile based screening procedure. There are some papers on screening procedures for varying coefficient and additive models, for example, [8], [10], and [19]. Forward type selection procedures are considered in [31], [16], and [6]. We name [3], [13], and [30] as general references on high-dimensional issues.

Because parsimonious modelling is crucial for statistical analysis, simultaneous variable selection and structure identification in semiparametric regression models has been studied by many authors, see, among others, [35], [21], [33], [5], [22], and [15]. Another important reason to attain this purpose is that in some high-dimensional situations, there may be a lack of priori knowledge on how to decide which covariates to be included in the parametric part and which covariates to be included in the nonparametric part. On the other hand, to the best of our knowledge, no theoretical sound procedure has been proposed to achieve the aforementioned goal in the high-dimensional quantile regression setups. Note that [21] and [22] proposed using the estimated derivatives of coefficient functions to identify the structures of additive models. These estimated derivatives, however, usually have slow convergence rates. Moreover, as shown in Section S.2 of the supplementary document, the conditions imposed on the B-spline basis functions in [21] and [22] seem too stringent to be satisfied in practice. Instead of relying on the estimated derivatives of coefficient functions, we appeal to the orthogonal decomposition method through introducing an orthonormal spline basis with desirable properties as in [15], which is devoted to the study of Cox regression models. Our approach not only can be justified theoretically under a set of reasonable assumptions, but also enables a unified analysis of varying coefficient models and additive models.

The Lasso for quantile linear regression is considered in [1] and the adaptively weighted Lasso for quantile linear regression are considered in [7] and [36]. Some authors such as [17] and [27] deal with group Lasso procedures for additive models and varying coefficient models, respectively. [23] applied a reproducing kernel Hilbert space approach to additive models. [26] deals with SCAD type variable selection for parametric part. In [26], the authors applied the adaptively weighted Lasso iteratively to obtain their SCAD estimate starting from the Lasso estimate. However, in the quantile regression setup, there doesn't seem to exist any theoretical or numerical result for simultaneous variable selection and structure identification based on the adaptively weighted group Lasso, in particular when its penalty is determined by a data-driven fashion. To fill this gap, we establish selection consistency of AWG-Lasso and AWG-Lasso+HDIC in Section 3, and illustrate the finite sample performance of AWG-Lasso+HDIC through a simulation study in Section 4. Our simulation study reveals that AWG-Lasso+HDIC performs satisfactorily in terms of true positive and true negative rates.

This paper is organized as follows: We describe our procedures in Section 2. We present our theoretical results in Section 3. The results of numerical studies are given in Section 4. We state assumptions and prove our main results in Section 5 and describe some important properties of B-spline bases in Section S.2 of the supplement. Technical lemmas and the proofs are also given in the supplement.

We end this section with some notation used throughout the paper. $\overline{A}$ and $|A|$ stand for the complement and the number of the elements of a set A, respectively. For a vector a, $|a|$ and a^T are the Euclidean norm and the transpose, respectively. For a function g on the unit interval, $\|g\|$ and $\|g\|_\infty$ stand for the L_2 and sup norms, respectively. We denote the maximum and minimum eigenvalues of a matrix A by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. Besides, C, C_1, C_2, ..., are generic positive constants and their values may change from line to line. Note that $a_n \sim b_n$ means $C_1 < a_n/b_n < C_2$ and that $a \vee b$ and $a \wedge b$ stand for the maximum and the minimum of a and b, respectively. Convergence in probability is denoted by $\xrightarrow{p}$.

2 Simultaneous variable selection and structure identification

2.1 Varying coefficient models

Suppose that we have n i.i.d. observations $\{(Y_i, \boldsymbol{X}_i, Z_i)\}_{i=1}^n$, where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$ is a p-dimensional covariate vector and Z_i is a scalar index covariate. Then we assume a quantile varying coefficient model holds for these observations. First we define the τ-th quantile check function $\rho_\tau(u)$ and its derivative $\rho'_\tau(u)$ by

$$\rho_\tau(u) = u(\tau - I\{u \le 0\}) \quad \text{and} \quad \rho'_\tau(u) = \tau - I\{u \le 0\}.$$

Then our model in this subsection is

$$Y_i = \sum_{j=1}^p X_{ij} g_j(Z_i) + \epsilon_i, \tag{1}$$

where $Z_i \in [0,1]$ and $\mathrm{E}\{\rho'_\tau(\epsilon_i) \,|\, \boldsymbol{X}_i, Z_i\} = 0$. Usually we take $X_{i1} \equiv 1$ for varying coefficient models.

To deal with partially linear varying coefficient models, we decompose $g_j(z)$ as $g_j(z) = g_{cj} + g_{vj}(z)$, where

$$g_{cj} = \int_0^1 g_j(z)dz \quad \text{and} \quad g_{vj}(z) = g_j(z) - g_{cj}.$$

We define the index set, $\mathcal{S}^0 = (\mathcal{S}_c^0, \mathcal{S}_v^0)$, for the true model, where

$$\mathcal{S}_c^0 = \{j \,|\, g_{cj} \ne 0\} \quad \text{and} \quad \mathcal{S}_v^0 = \{j \,|\, g_{cj}(z) \not\equiv 0\}.$$

The index set for a candidate model can be similarly given by $\mathcal{S} = (\mathcal{S}_c, \mathcal{S}_v)$. In the following, we refer to $\mathcal{S}^0$ and $\mathcal{S}$ as the true model and the candidate model, respectively whenever confusion is unlikely. When some j's satisfy both $j \in \mathcal{S}_c^0$ and $j \notin \mathcal{S}_v^0$ simultaneously, our true model is a partially linear varying coefficient model, for example, $\mathcal{S}^0 = (\{1,2,3\},\{1,2\})$ with $\mathcal{S}_c^0 = \{1,2,3\}$ and $\mathcal{S}_v^0 = \{1,2\}$. Moreover, $\mathcal{S}_1 \supset \mathcal{S}_2$ means $\mathcal{S}_{c1} \supset \mathcal{S}_{c2}$ and $\mathcal{S}_{v1} \supset \mathcal{S}_{v2}$, where $\mathcal{S}_j = (\mathcal{S}_{cj}, \mathcal{S}_{vj})$, $j = 1,2$. In addition, $\mathcal{S}_1 \cup \mathcal{S}_2 = (\mathcal{S}_{c1} \cup \mathcal{S}_{c2}, \mathcal{S}_{v1} \cup \mathcal{S}_{v2})$.

We use the regression spline method to estimate coefficient functions and the covariates for regression spline are defined by

$$\boldsymbol{W}_i = \boldsymbol{X}_i \otimes \boldsymbol{B}(Z_i), \tag{2}$$

where $\boldsymbol{B}(z) = (B_1(z), B_2(z), \dots, B_L(z))^T$ is an orthonormal basis constructed from the equispaced B-spline basis $\boldsymbol{B}_0(z) = (B_{01}(z), \dots, B_{0L}(z))^T$ on $[0,1]$ and $\otimes$ is the Kroneker product. We can represent $\boldsymbol{B}(z)$ as $\boldsymbol{B}(z) = A_0\boldsymbol{B}_0(z)$ and we calculate the $L \times L$ matrix A_0 numerically. As in [15], let $\boldsymbol{B}(z)$ satisfy $B_1(z) = 1/\sqrt{L}$, $B_2(z) = \sqrt{12/L}(z - 1/2)$, and

$$\int_0^1 \boldsymbol{B}(z)(\boldsymbol{B}(z))^T dz = L^{-1} I_L. \tag{3}$$

We denote the $L \times L$ identity matrix by I_L. Note that $B_1(z)$ is for g_{cj} (the j-th constant component) and $\boldsymbol{B}_{-1}(z) = (B_2(z), \dots, B_L(z))^T$ is for $g_{vj}(z)$ (the j-th non-constant component). More details are given in Section S.2 of the supplement.

To carry out simultaneous variable selection and structure identification, we apply AWG-Lasso to

$$Y_i = \boldsymbol{W}_i^T \boldsymbol{\gamma} + \epsilon_i', \tag{4}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_p^T)^T$. For a given $\lambda > 0$, the corresponding objective function is given by

$$Q_V(\boldsymbol{\gamma}; \lambda) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - \boldsymbol{W}_i^T\boldsymbol{\gamma}) + \lambda \sum_{j=1}^p (w_{1j}|\gamma_{1j}| + w_{-1j}|\boldsymbol{\gamma}_{-1j}|), \tag{5}$$

where $\{(w_{1j}, w_{-1j})\}_{j=1}^p$ is obtained from some initial estimates such as Lasso and group Lasso, and $(\gamma_{1j}, \boldsymbol{\gamma}_{-1j}^T)^T = \boldsymbol{\gamma}_j$, noting that γ_{1j} is for $B_1(z)$ and $\boldsymbol{\gamma}_{-1j}$ is for $\boldsymbol{B}_{-1}(z)$. Minimizing $Q_V(\boldsymbol{\gamma}; \lambda)$ w.r.t. $\boldsymbol{\gamma}$, one gets

$$\widehat{\boldsymbol{\gamma}}^\lambda = \operatorname*{argmin}_{\boldsymbol{\gamma} \in R^{pL}} Q_V(\boldsymbol{\gamma}; \lambda).$$

Denote $\widehat{\boldsymbol{\gamma}}^\lambda$ by $(\widehat{\gamma}_{11}^\lambda, \widehat{\boldsymbol{\gamma}}_{-11}^{\lambda T}, \dots, \widehat{\gamma}_{1p}^\lambda, \widehat{\boldsymbol{\gamma}}_{-1p}^{\lambda T})^T$. Then, the model selected by AWG-Lasso is $\widehat{\mathcal{S}}^\lambda = (\widehat{\mathcal{S}}_c^\lambda, \widehat{\mathcal{S}}_v^\lambda)$, where $\widehat{\mathcal{S}}_c^\lambda = \{j \,|\, \widehat{\gamma}_{1j}^\lambda \neq 0\}$ and $\widehat{\mathcal{S}}_v^\lambda = \{j \,|\, \widehat{\boldsymbol{\gamma}}_{-1j}^\lambda \neq \mathbf{0}\}$, and this enables us to identify variables and structures simultaneously.

Theorem 1 in Section 3 establishes the selection consistency of $\widehat{\mathcal{S}}^\lambda$ under a set of general conditions on $\{(w_{1j}, w_{-1j})\}_{j=1}^p$ and a strong sparsity condition on the regression coefficients that $|\mathcal{S}_c^0|$ and $|\mathcal{S}_v^0|$ are bounded. Theorem 1, however, also requires that λ falls into a suitable interval, which can sometimes be hard to decide in practice. We therefore introduce a BIC-type criterion, HDIC, to choose a λ in a data-driven fashion. Express $\boldsymbol{W}_i$ as $(v_{11i}, \boldsymbol{v}_{-11i}^T, \cdots, v_{1pi}, \boldsymbol{v}_{-1pi}^T)^T$, where $(v_{1ji}, \boldsymbol{v}_{-1ji}^T)^T$ is the regressor vector corresponding to $\boldsymbol{\gamma}_j$. For a given model $\mathcal{S} = (\mathcal{S}_c, \mathcal{S}_v)$, define $R_V(\boldsymbol{\gamma}_\mathcal{S})$ and $\widetilde{\boldsymbol{\gamma}}_\mathcal{S}$ by

$$R_V(\boldsymbol{\gamma}_\mathcal{S}) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - \boldsymbol{W}_{i\mathcal{S}}^T \boldsymbol{\gamma}_\mathcal{S}) \quad \text{and} \quad \widetilde{\boldsymbol{\gamma}}_\mathcal{S} = \mathop{\mathrm{argmin}}_{\boldsymbol{\gamma}_\mathcal{S} \in R^{|\mathcal{S}_c| + (L-1)|\mathcal{S}_v|}} R_V(\boldsymbol{\gamma}_\mathcal{S}), \tag{6}$$

where $\boldsymbol{W}_{i\mathcal{S}} \in R^{|\mathcal{S}_c|+(L-1)|\mathcal{S}_v|}$ consists of $\{v_{1ji} \,|\, j \in \mathcal{S}_c\}$ and $\{\boldsymbol{v}_{-1ji} \,|\, j \in \mathcal{S}_v\}$. The corresponding coefficient vector $\boldsymbol{\gamma}_\mathcal{S}$ consists of $\{\gamma_{1ji} \,|\, j \in \mathcal{S}_c\}$ and $\{\boldsymbol{\gamma}_{-1ji} \,|\, j \in \mathcal{S}_v\}$ as well. The elements of these vectors are suitably arranged. In this paper, we sometimes take two index sets $\mathcal{S}_1$ and $\mathcal{S}_2$ satisfying $\mathcal{S}_1 \subset \mathcal{S}_2$ and compare $\boldsymbol{\gamma}_{\mathcal{S}_1}$ and $\boldsymbol{\gamma}_{\mathcal{S}_2}$ by enlarging $\boldsymbol{\gamma}_{\mathcal{S}_1}$ with 0 elements or something, for example, $(\boldsymbol{\gamma}_{\mathcal{S}_1}^T, \mathbf{0}^T)^T$. Then $(\boldsymbol{\gamma}_{\mathcal{S}_1}^T, \mathbf{0}^T)^T$ and $\boldsymbol{\gamma}_{\mathcal{S}_2}$ have the same dimension and the elements of these vectors are assumed to be conformably rearranged.

The HDIC value for model $\mathcal{S}$ is stipulated by

$$\mathrm{HDIC}(\mathcal{S}) = \log R_V(\widetilde{\boldsymbol{\gamma}}_\mathcal{S}) + (|\mathcal{S}_c| + (L-1)|\mathcal{S}_v|)\frac{q_n \log p_n}{2n}, \tag{7}$$

where $p_n = p \vee n$ and $q_n \to \infty$ at a slow rate described in Section 5. We consider a set of models $\{\widehat{\mathcal{S}}^\lambda\}$ chosen by AWG-Lasso, where $\lambda \in \Lambda$ with Λ being a prescribed set of positive numbers, and select $\widehat{\mathcal{S}}^{\hat{\lambda}}$ among $\{\widehat{\mathcal{S}}^\lambda\}$, where

$$\hat{\lambda} = \mathop{\mathrm{argmin}}_{\lambda \in \Lambda, |\widehat{\mathcal{S}}_c^\lambda| \le M_c, |\widehat{\mathcal{S}}_v^\lambda| \le M_v} \mathrm{HDIC}(\widehat{\mathcal{S}}^\lambda),$$

with M_c and M_v being known upper bounds for $|\mathcal{S}_c^0|$ and $|\mathcal{S}_v^0|$, respectively. Under some regularity conditions, the consistency of $\widehat{\mathcal{S}}^{\hat{\lambda}}$ is established in Corollary 1.

Note that in the case of high-dimensional sparse linear models, it is shown in [16] that (7) with $\rho_\tau(\cdot)$ replaced by the squared loss $(\cdot)^2$ can be used in conjunction with the orthogonal greedy algorithm (OGA) to yield selection consistency. The major difference between (7) and the BIC-type criteria considered in [20] is that we deal with semiparametric models in this paper. It seems difficult to derive the consistency of $\widehat{\mathcal{S}}^{\hat{\lambda}}$ in any high-dimensional regression setups without the additional penalty term q_n in (7).

2.2 Additive models

Next we deal with additive models. Recall we assume some initial estimates are available here, too. We have no index variable and assume the additivity and $X_{ij} \in [0,1]$ for $j = 1, \ldots, p$. Hence our model is

$$Y_i = \mu + \sum_{j=1}^{p} g_j(X_{ij}) + \epsilon_i, \tag{8}$$

where $X_{ij} \in [0,1]$, $\int_0^1 g_j(x)dx = 0$, and $\mathrm{E}\{\rho'_\tau(\epsilon_i) \,|\, \boldsymbol{X}_i\} = 0$. To deal with partially linear additive coefficient models, we decompose $g_j(x)$ as $g_j(x) = g_{lj}(x) + g_{aj}(x)$, where $g_{lj}(x) = c_{lj}B_2(x)$ (the j-th linear component) and $g_{aj}(x)$ (the j-th nonlinear component) satisfies

$$\int_0^1 g_{lj}(x) g_{aj}(x) dx = 0.$$

Our regression spline model is given by

$$Y_i = \mu + \boldsymbol{W}_i^T \boldsymbol{\gamma}_{-1} + \epsilon'_i, \tag{9}$$

where $\boldsymbol{\gamma}_{-1} = (\boldsymbol{\gamma}_{-11}^T, \ldots, \boldsymbol{\gamma}_{-1p}^T)^T$ and $\boldsymbol{W}_i = (\boldsymbol{B}_{-1}^T(X_{i1}), \ldots, \boldsymbol{B}_{-1}^T(X_{ip}))^T$, with $\boldsymbol{\gamma}_{-1j}$ and $\boldsymbol{B}_{-1}(z)$ defined as in Subsection 2.1. Denote the true model by $\mathcal{S}^0 = (\mathcal{S}_l^0, \mathcal{S}_a^0)$, where

$$\mathcal{S}_l^0 = \{j \,|\, g_{lj} \neq 0\} \quad \text{and} \quad \mathcal{S}_a^0 = \{j \,|\, g_{aj}(x) \not\equiv 0\}.$$

When some j's satisfy both $j \in \mathcal{S}_l^0$ and $j \notin \mathcal{S}_a^0$ simultaneously, our true model is a partially linear additive model.

We describe the details of our simultaneous variable selection and structure identification procedure for additive models. First express $\boldsymbol{\gamma}_{-1j}$ as $\boldsymbol{\gamma}_{-1j} = (\gamma_{2j}, \boldsymbol{\gamma}_{-2j}^T)^T$, noting that γ_{2j} is for $B_2(X_{ij}) = \sqrt{12/L}(X_{ij} - 1/2)$ and $\boldsymbol{\gamma}_{-2j}$ is for $\boldsymbol{B}_{-2}(X_{ij}) = (B_3(X_{ij}), \ldots, B_L(X_{ij}))^T$. For a given λ, the AWG-Lasso objective function is

$$Q_A(\boldsymbol{\gamma}_{-1}; \lambda) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \mu - \boldsymbol{W}_i^T \boldsymbol{\gamma}_{-1}) + \lambda \sum_{j=1}^{p} (w_{2j}|\gamma_{2j}| + w_{-2j}|\boldsymbol{\gamma}_{-2j}|), \tag{10}$$

where $\{(w_{2j}, w_{-2j})\}_{j=1}^p$ are obtained from some initial estimates. Minimizing $Q_A(\boldsymbol{\gamma}_{-1}; \lambda)$ w.r.t. $\boldsymbol{\gamma}_{-1}$, one gets

$$\widehat{\boldsymbol{\gamma}}_{-1}^\lambda = \underset{\boldsymbol{\gamma}_{-1} \in R^{p(L-1)}}{\mathrm{argmin}} \; Q_A(\boldsymbol{\gamma}_{-1}; \lambda),$$

where $\widehat{\boldsymbol{\gamma}}_{-1}^\lambda = (\widehat{\gamma}_{21}^\lambda, \widehat{\boldsymbol{\gamma}}_{-21}^{\lambda T}, \ldots, \widehat{\gamma}_{2p}^\lambda, \widehat{\boldsymbol{\gamma}}_{-2p}^{\lambda T})^T$. Then, the model selected by AWG-Lasso is $\widehat{\mathcal{S}}^\lambda = (\widehat{\mathcal{S}}_l^\lambda, \widehat{\mathcal{S}}_a^\lambda)$, where $\widehat{\mathcal{S}}_l^\lambda = \{j \,|\, \widehat{\gamma}_{2j}^\lambda \neq 0\}$ and $\widehat{\mathcal{S}}_a^\lambda = \{j \,|\, \widehat{\boldsymbol{\gamma}}_{-2j}^\lambda \neq \boldsymbol{0}\}$. Like Subsection

2.1, this subsection also considers using HDIC to choose a suitable λ from a prescribed set Λ of positive numbers. Denote $\boldsymbol{W}_i$ in (9) by $(v_{21i}, \boldsymbol{v}_{-21i}^T, \ldots, v_{2pi}, \boldsymbol{v}_{-2pi}^T)^T$, where $(v_{2ji}, \boldsymbol{v}_{-2ji}^T)^T$ is the regressor vector corresponds to $\boldsymbol{\gamma}_{-1j}$. For a given model $\mathcal{S} = (\mathcal{S}_l, \mathcal{S}_a)$, define

$$R_A(\boldsymbol{\gamma}_{\mathcal{S}}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \mu - \boldsymbol{W}_{i\mathcal{S}}^T \boldsymbol{\gamma}_{\mathcal{S}}) \quad \text{and} \quad \widetilde{\boldsymbol{\gamma}}_{\mathcal{S}} = \underset{\boldsymbol{\gamma}_{\mathcal{S}} \in R^{|\mathcal{S}_l| + (L-2)|\mathcal{S}_a|}}{\operatorname{argmin}} R_A(\boldsymbol{\gamma}_{\mathcal{S}}), \tag{11}$$

where $\boldsymbol{W}_{i\mathcal{S}} \in R^{|\mathcal{S}_l|+(L-2)|\mathcal{S}_a|}$ consists of $\{v_{2ji} \,|\, j \in \mathcal{S}_l\}$ and $\{\boldsymbol{v}_{-2ji} \,|\, j \in \mathcal{S}_a\}$ and the corresponding coefficient $\boldsymbol{\gamma}_{\mathcal{S}} \in R^{|\mathcal{S}_l|+(L-2)|\mathcal{S}_a|}$ is conformably defined as in (6).

The HDIC value for model $\mathcal{S}$ is stipulated by

$$\mathrm{HDIC}(\mathcal{S}) = \log R_A(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) + (|\mathcal{S}_l| + (L-2)|\mathcal{S}_a|)\frac{q_n \log p_n}{2n}, \tag{12}$$

where p_n and q_n are defined as in Subsection 2.1. Let M_l and M_a be some known upper bounds for $|\mathcal{S}_l^0|$ and $|\mathcal{S}_a^0|$, respectively. We suggest choosing model $\hat{\mathcal{S}}^{\hat{\lambda}}$, where

$$\hat{\lambda} = \underset{\lambda \in \Lambda, |\widehat{\mathcal{S}}_l^{\lambda}| \le M_l, |\widehat{\mathcal{S}}_a^{\lambda}| \le M_a}{\operatorname{argmin}} \mathrm{HDIC}(\widehat{\mathcal{S}}^{\lambda}).$$

3 Consistency results

We prove the consistency of AWG-Lasso and AWG-Lasso+HDIC separately in Subsection 3.1 and 3.2. It is worth pointing out that due to the similarity between (4)-(7) and (9)-(12), the theoretical treatment is almost the same for the two types of models considered in this paper. Therefore, this section concentrates only on the varying coefficient model. On the other hand, our numerical studies are conducted for both types of models, see Section 4.

3.1 Adaptively weighted group Lasso

The consistency of AWG-Lasso for suitably chosen λ and weights is stated in Theorem 1. The proof of Theorem 1 is reliant on the methods of [7], [36], and [26] subject to nontrivial modifications. The details are deferred to Section 5. For clarity of presentation, all the technical assumptions of Theorem 1 are also given in Section 5. Roughly speaking, we assume that the coefficient functions have second order derivatives and we put $L = c_L n^{1/5}$. More smoothness is necessary for Theorem 2. If X_{ij} is uniformly bounded, the Hölder continuity of the second order derivatives with exponent $\alpha = 1/2$ is sufficient for Theorem 2.

Define $d_V(\mathcal{S}) = |\mathcal{S}_c| + (L-1)|\mathcal{S}_v|$ and let $w_{\mathcal{S}^0}$ denote a weight vector consisting of $\{w_{1j} \,|\, j \in \mathcal{S}_c^0\}$ and $\{w_{-1j} \,|\, j \in \mathcal{S}_v^0\}$. For an index set $\mathcal{S}$, we define $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}^{\lambda}$ by

$$\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}^{\lambda} = \operatorname*{argmin}_{\boldsymbol{\gamma}_{\mathcal{S}} \in R^{d_V(\mathcal{S})}} Q_V(\boldsymbol{\gamma}_{\mathcal{S}}; \lambda).$$

Then $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}^{\lambda}$ is an oracle estimator on $R^{d_V(\mathcal{S}^0)}$ with the knowledge of $\mathcal{S}^0$. Assumption A2 assumes that the relevant coefficients and the coefficient functions are large enough to be detected.

Theorem 1 *Assume that Assumptions A1, A3-5 and B1–4 in Section 5 hold. Moreover, assume*

$$\max_{j \in \mathcal{S}_c^0} w_{1j} \vee \max_{j \in \mathcal{S}_v^0} w_{-1j} = O_p(1), \tag{13}$$

and for some $0 < a_1, a_2 < \infty$,

$$\min_{j \notin \mathcal{S}_c^0} w_{1j} \geq (a_1 |w_{\mathcal{S}^0}|) \vee 1 \quad \textit{and} \quad \min_{j \notin \mathcal{S}_v^0} w_{-1j} \geq (a_2 |w_{\mathcal{S}^0}|) \vee 1, \tag{14}$$

with probability tending to 1. We enlarge $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}$ *by adding* 0 *elements for the* $\mathcal{S}^{0c}$ *part so that* $(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}^{\lambda T}, \mathbf{0}^T)^T \in R^{pL}$ *and define* $\widehat{\mathcal{S}}^{\lambda}$ *from this* $(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}^{\lambda T}, \mathbf{0}^T)^T$. *Then for any* λ *satisfying*

$$a_3 \frac{(\log p_n)^{1/2}}{n^{1/2}} \leq \lambda \leq (\log n)^{\kappa} \frac{(\log p_n)^{1/2}}{n^{1/2}} \tag{15}$$

asymptotically, where a_3 *is a sufficiently large constant and* κ *is any positive constant,* $(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}^{\lambda T}, \mathbf{0}^T)^T$ *is actually an optimal solution to minimizing* $Q_V(\boldsymbol{\gamma}; \lambda)$ *w.r.t.* $\boldsymbol{\gamma} \in R^{pL}$ *with probability tending to 1. If Assumption A2 also holds, we have for* $\widehat{\mathcal{S}}^{\lambda}$ *defined here that*

$$\lim_{n \to \infty} \mathrm{P}(\widehat{\mathcal{S}}^{\lambda} = \mathcal{S}^0) = 1.$$

The order of $L^{1/2}\lambda$ in (15) is the standard one in the literature since $(\log p_n)^{1/2}$ is from the large number of covariates and $(L/n)^{1/2}$ is the standard rate for regression spline estimation. Recall that our normalization factor of the orthonormal basis is $1/L$. The upper bound of λ in Theorem 1 is a technical one since we approximate $R_V(\boldsymbol{\gamma})$ by a quadratic function in $\boldsymbol{\gamma}$ on a suitable bounded region.

We need an initial estimate $\overline{\boldsymbol{\gamma}} = (\overline{\gamma}_{11}, \overline{\boldsymbol{\gamma}}_{-11}^T, \ldots, \overline{\gamma}_{1p}, \overline{\boldsymbol{\gamma}}_{-1p}^T)^T$ from the group Lasso as in [27] and [17] to construct weights for AWG-Lasso. Note that $L^{-1/2}|\overline{\gamma}_{1j}|$ and $L^{-1/2}|\overline{\boldsymbol{\gamma}}_{-1j}|$ are estimates of $|g_{cj}|$ and $\|g_{vj}\|$, respectively. They have the convergence rates smaller than $CL^{1/2}\lambda$ for some sufficiently large C and λ in Theorem 1. Hence

$$w_{1j} = (L^{-1/2}|\overline{\gamma}_{1j}|)^{-\eta} \quad \text{and} \quad w_{-1j} = (L^{-1/2}|\overline{\boldsymbol{\gamma}}_{-1j}|)^{-\eta} \tag{16}$$

satisfy the conditions (13) and (14) for any positive η when the norms of the relevant coefficients and the relevant functions are larger than a fixed non-zero constant. Otherwise we should adjust the range of λ by multiplying λ by a suitable constant and dividing the weights in (16) by the suitable constant, respectively so that the assumption on λ, (13), and (14) can hold for these adjusted ones. However, we usually have no knowledge of $|g_{cj}|$ and $\|g_{vj}\|$ in advance and this kind of adjustment is infeasible. Then we should carry out search for an optimal λ on a larger interval than specified by Theorem 1 in practical situations.

When Assumption A2 holds and we use the wights based on the local linear approximation (LLA) to the SCAD penalty as in Section 4, the weights in (19) and (20) satisfy (13) and (14) due to the properties of the SCAD penalty. Some authors as [26] applied this kind of AGW-Lasso iteratively to calculate their SCAD estimates.

3.2 Consistency of AWG-Lasso+HDIC

To state the main result of this subsection, we need to introduce Assumption A1, which assumes that $|\mathcal{S}_c^0| \leq C_c$ and $|\mathcal{S}_v^0| \leq C_v$ for some fixed C_c and C_v. Let M_c and M_v be known positive integers fixed with n such that $C_c < M_c$ and $C_v < M_v$. Define

$$\widehat{\mathcal{S}} = \operatorname*{argmin}_{|\mathcal{S}_c| \leq M_c \text{ and } |\mathcal{S}_v| \leq M_v} \mathrm{HDIC}(\mathcal{S}).$$

Under certain regularity conditions, the next theorem and corollary show that both $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}^{\hat{\lambda}}$ are consistent estimates of $\mathcal{S}^0$. We need to replace Assumptions A2–5 and B1–4 with Assumptions A2'–A5' and B1'–B4' to carry out subtle evaluations of $R_V(\boldsymbol{\gamma}_{\mathcal{S}})$ in the proof since we deal with high-dimensional semiparametric models. All the technical assumptions of Theorem 2 are also given in Section 5.

Theorem 2 *Assume that Assumptions A1,A2'–A5', B1'–B4' and B5 in Section 5 hold. Then,*

$$\lim_{n\to\infty} P(\widehat{\mathcal{S}} = \mathcal{S}^0) = 1.$$

Theorem 1 gives a suitable set of λ, Λ, as in Corollary 1 for which $\{\widehat{\mathcal{S}}_\lambda \mid \lambda \in \Lambda\}$ includes $\mathcal{S}^0$ with probability tending to 1. Thus the consistency of the proposed AWG-Lass+HDIC procedure immediately follows from Theorems 1 and 2.

Corollary 1 *We assume the same assumptions as in Theorem 2 and that* (13) *and* (14) *hold true. Then for* Λ *satisfying* $\Lambda \subset [c_n^{-1}\sqrt{\log p_n/n}, c_n\sqrt{\log p_n/n}]$ *and* $\{c_n\sqrt{\log p_n/n}\} \in \Lambda$*, where* $c_n \to \infty$ *and* $c_n/(\log n)^\kappa \to 0$ *for some* $\kappa > 0$*, we have*

$$\lim_{n\to\infty} P(\widehat{\mathcal{S}}^{\hat{\lambda}} = \mathcal{S}^0) = 1.$$

Some comments are in order. While $\hat{\mathcal{S}}$ can achieve selection consistency without the help of AWG-Lasso, it seems difficult to obtain $\widehat{\mathcal{S}}$ directly when p is large and M_c and M_v are not very small. On the other hand, $\widehat{\mathcal{S}}^{\hat{\lambda}}$ is applicable in most practical situations. We also note that Theorem 2 extends the result in [20] and can be viewed as a generalization of the BIC result in [36] to the semiparametric setup, which is of fundamental interest from both theoretical and practical perspectives. Like [36], [18] also confines its attention to linear quantile models. Moreover, it seems difficult to extend the proof in [18] to situations where the dimension of the true model tends to infinity. Finally, we mention that there is another version of HDIC,

$$\mathrm{HDIC_{II}}(\mathcal{S}) = R_V(\widetilde{\gamma}_{\mathcal{S}}) + (|\mathcal{S}_c| + (L-1)|\mathcal{S}_v|)\frac{q_n \log p_n}{2n}, \tag{17}$$

which becomes

$$\mathrm{HDIC_{II}}(\mathcal{S}) = R_V(\widetilde{\gamma}_{\mathcal{S}}) + (|\mathcal{S}_l| + (L-2)|\mathcal{S}_a|)\frac{q_n \log p_n}{2n} \tag{18}$$

in the case of additive models. It can be shown that $\mathrm{HDIC_{II}}$ and HDIC share the same asymptotic properties and their finite sample performance will be compared in the next section.

4 Numerical studies

In this section, we evaluate the performance of AWG-Lasso+HDIC and AWG-Lasso+$\mathrm{HDIC_{II}}$ using one varying coefficient model and two additive models in the case of $pL > n$. We set $q_n = 1$ in these numerical studies since the optimal choice of q_n in finite sample remains unsettled and is worth further investigation.

We start by assigning $\{(w_{1j}, w_{-1j})\}$ and $\{(w_{2j}, w_{-2j})\}$ used for (5) and (10), respectively. We only focus on $\{(w_{1j}, w_{-1j})\}$ because $\{(w_{2j}, w_{-2j})\}$ can be assigned in a similar fashion. With the initial estimates $(\overline{\gamma}_{1j}, \overline{\boldsymbol{\gamma}}_{-1j})$ obtained from the quantile regression with the Lasso penalty, we apply one-step LLA (see [12]) based on the SCAD penalty as in [26] to obtain $\{(w_{1j}, w_{-1j})\}$. More specifically, we set

$$\lambda w_{1j}|\gamma_{1j}| = p'_{\lambda L^{1/2}}(L^{-1/2}|\overline{\gamma}_{1j}|)(L^{-1/2}|\gamma_{1j}|), \tag{19}$$

and

$$\lambda w_{-1j}|\boldsymbol{\gamma}_{-1j}| = p'_{\lambda L^{1/2}}(L^{-1/2}|\overline{\boldsymbol{\gamma}}_{-1j}|)(L^{-1/2}|\boldsymbol{\gamma}_{-1j}|), \tag{20}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function. Recall the definition of the spline basis in (3), whose normalization factor is L^{-1}, and see a comment after Theorem 1 and Assumption A2 to get a better understanding of the role played by $L^{1/2}$ in (19) and (20). Note that a standard theory for the group Lasso as in [17] and [27] ensures that the weights imposed in (19) and (20) obey (13) and (14).

In our simulation study, (n, p) is set to (500, 400) or $(1500, 2000)$, $L = 6$, $\tau = 0.5$, $M_c = M_v = M_l = M_a = 20$ and

$$\Lambda = \left\{c_n^{-1}\sqrt{\log p/n} + kd_n, k = 1, \dots, 50\right\},$$

where $c_n = 2\log n$ and

$$d_n = \frac{(c_n - c_n^{-1})\sqrt{\log p/n}}{50}.$$

Based on a $\lambda \in \Lambda$ and the weights described above, we employ the alternating direction method of multipliers (ADMM) to minimize (5) ((10)) over $\boldsymbol{\gamma}$ ($\boldsymbol{\gamma}_{-1}$), and then choose the λ minimizing HDIC($\hat{\mathcal{S}}^\lambda$) defined in (7) ((12)) over $\lambda \in \Lambda$, and the λ minimizing HDIC_{II} ($\hat{\mathcal{S}}^\lambda$) defined in (17) ((18)) over the same set. For each of the following three examples, we conduct 50 simulations and record the performance of AWG-Lasso+HDIC and AWG-Lasso+HDIC_{II} in Tables 1–3.

Example 1. We generate the output variables $Y_1, \dots, Y_n$ using the varying coefficient model,

$$Y_i = \sum_{j=1}^{p} X_{ij} g_j(Z_i) + \epsilon_i,$$

where ϵ_i, Z_i and $\{X_{ij}\}_{j=1}^p$ are independently generated from $N(0, 0.5^2)$, $U(0,1)$ and $U(0, 100)$ distributions, respectively. Following [15], the coefficient functions $g_j(z)$ are set to

$$g_1(z) = g_2(z) = 1, g_3(z) = 4z, g_4(z) = 4z^2, g_j(z) = 0,\ 5 \le j \le p.$$

Therefore, X_{i1} and X_{i2} are relevant covariates with constant coefficients, X_{i3} and X_{i4} are relevant covariates with non-constant coefficients, whereas $X_{i,5}, \dots, X_{i,p}$, are irrelevant variables. Note that our goal is to identify not only relevant variables, but also the structures of relevant coefficients. To this aim, we first define

$$\begin{aligned}
\mathrm{C}_{sj} &= I_{\{g_j(\cdot) \text{ is identified as a constant function at the } s\text{th replication}\}}, \\
\mathrm{NC}_{sj} &= I_{\{g_j(\cdot) \text{ is identified as a non-constant function at the } s\text{th replication}\}}, \\
\mathrm{NS}_{sj} &= I_{\{g_j(\cdot) \text{ is identified as a zero function at the } s\text{th replication}\}}.
\end{aligned}$$

It is clear that $\mathrm{C}_{sj} + \mathrm{NC}_{sj} + \mathrm{NS}_{sj} = 1$ for each $1 \le j \le p$. We further define the true negative rate (TNR) and the strictly true positive rate (STPR),

$$\begin{aligned}
\mathrm{TNR}_s &= \frac{\sum_{j=5}^{p} I_{\{\mathrm{NS}_{sj}=1\}}}{p-4}, \\
\mathrm{STPR}_s &= \frac{\sum_{j=1}^{2} I_{\{\mathrm{C}_{sj}=1\}} + \sum_{j=3}^{4} I_{\{\mathrm{NC}_{sj}=1\}}}{4},
\end{aligned}$$

noting that $\mathrm{STPR}_s = 1$ if at the sth replication, X_{i1} and X_{i2} are identified as relevant variables with constant coefficients and X_{i3} and X_{i4} are identified as relevant variables with non-constant coefficients. Therefore, STPR_s can be viewed as a stringent version of the conventional true positive rate, which treats constant and non-constant coefficient functions indifferently. Now, the performance measures of the proposed methods are specified as follows:

$$\begin{aligned}
\mathrm{C}_j &= \frac{1}{50}\sum_{s=1}^{50} \mathrm{C}_{sj},\ \mathrm{NC}_j = \frac{1}{50}\sum_{s=1}^{50} \mathrm{NC}_{sj}, \mathrm{NS}_j = \frac{1}{50}\sum_{s=1}^{50} \mathrm{NS}_{sj}, \\
\mathrm{TNR} &= \frac{1}{50}\sum_{s=1}^{50} \mathrm{TNR}_s,\ \mathrm{STPR} = \frac{1}{50}\sum_{s=1}^{50} \mathrm{STPR}_s.
\end{aligned}$$

The performance of AWG-Lasso+HDIC and AWG-Lasso+HDIC$_{\mathrm{II}}$ on $(\mathrm{C}_j, \mathrm{NC}_j, \mathrm{NS}_j)$, $j = 1, \dots, 4$, STPR and TNR is demonstrated in Table 1. Table 1 shows that AWG-Lasso+HDIC and AWG-Lasso+HDIC$_{\mathrm{II}}$ have high capability in identifying the true variables and true structures in the sense that C_1=C_2=NC_3=NC_4=STPR=1 hold for every method and every (n, p) pair. Table 1 also reveals that AWG-Lasso+HDIC$_{\mathrm{II}}$ performs quite satisfactorily in identifying irrelevant variables since all of its TNR values are equal to 1. On the other hand, AWG-Lasso+HDIC tends to erroneously choose some irrelevant variables in the case of $(n, p) = (500, 400)$. This situation, however, is somewhat alleviated when (n, p) becomes (1500, 2000).

Example 2. We generate $Y_1, \dots, Y_n$ from the following additive model,

$$Y_i = \mu + \sum_{j=1}^{p} g_j(X_{ij}) + \epsilon_i, \tag{21}$$

where $\mu = 0$, ϵ_i and $\{X_{ij}\}_{j=1}^p$ follow $N(0, 0.5^2)$ and $U(0,1)$, respectively. Following [15] again, we set

$$\begin{aligned} &g_1(x) = g_2(x) = 2^{1/2}(x - 1/2),\ g_3(x) = 2^{-1/2}\cos(2\pi x) + (x - 1/2), \\ &g_4(x) = \sin(2\pi x),\ g_i(x) = 0,\ 5 \leq i \leq p, \end{aligned} \tag{22}$$

noting that X_{i1} and X_{i2} are relevant through the linear functions $g_1(\cdot)$ and $g_2(\cdot)$, whereas X_{i3} and X_{i4} are relevant through the nonlinear functions $g_3(\cdot)$ and $g_4(\cdot)$. Let NS_{sj} and TNR_s be defined as in Example 1, and define

$$\begin{aligned} \mathrm{L}_{sj} &= I_{\{g_j(\cdot) \text{ is identified as a linear function at the } s\text{th replication}\}}, \\ \mathrm{NL}_{sj} &= I_{\{g_j(\cdot) \text{ is identified as a non-linear function at the } s\text{th replication}\}}, \\ \mathrm{STPR}_s &= \frac{\sum_{j=1}^2 I_{\{\mathrm{L}_{sj}=1\}} + \sum_{j=3}^4 I_{\{\mathrm{NL}_{sj}=1\}}}{4}. \end{aligned}$$

Then, the performance measures of the proposed methods in this example are given by

$$\begin{aligned} \mathrm{L}_j &= \frac{1}{50}\sum_{s=1}^{50}\mathrm{L}_{sj},\ \mathrm{NL}_j = \frac{1}{50}\sum_{s=1}^{50}\mathrm{NL}_{sj}, \mathrm{NS}_j = \frac{1}{50}\sum_{s=1}^{50}\mathrm{NS}_{sj}, \\ \mathrm{TNR} &= \frac{1}{50}\sum_{s=1}^{50}\mathrm{TNR}_s,\ \mathrm{STPR} = \frac{1}{50}\sum_{s=1}^{50}\mathrm{STPR}_s, \end{aligned}$$

and summarized in Tables 2. Table 2 shows that $\mathrm{L}_1 = \mathrm{L}_2 = 1$ hold for every method and every (n, p) pair, implying that AWG-Lasso+HDIC and AWG-Lasso+HDIC$_{\mathrm{II}}$ can easily identify relevant linear functions. On the other hand, when $(n, p) = (500, 400)$, AWG-Lasso+HDIC$_{\mathrm{II}}$ tends to be more conservative in choosing nonlinear structures than AWG-Lasso+HDIC because while the NL_3 and NL_4 of the latter still achieve the highest possible value of 1, the NL_3 and NL_4 of the former are slightly less than 1. However, as (n, p) becomes (1500, 2000), $\mathrm{NL}_3 = \mathrm{NL}_4 = 1$ are attained by both methods. These results also coincide with the corresponding results for STPR. The TNR values of AWG-Lasso+HDIC$_{\mathrm{II}}$ (AWG-Lasso+HDIC) increase (decrease) from 0.994 to 0.998 (0.989) when (n, p) changes from (500, 400) to (1500, 2000), revealing that both methods tend to include a few irrelevant functions. Moreover, the false positive problem of AWG-Lasso+HDIC appears to have slightly worsened when p grows faster than n.

Example 3. Suppose that $Y_1, \ldots, Y_n$ are still generated from model (21), but with (22) replaced by

$$\begin{aligned} &g_1(x) = \frac{3\sin(2\pi x)}{(2 - \sin(2\pi x))} - 0.4641016, g_2(x) = 6x(1-x) - 1, g_3(x) = 2x - 1, \\ &g_4(x) = x - 0.5, g_5(x) = -x + 0.5, g_i(x) = 0,\ 6 \leq i \leq p, \end{aligned} \tag{23}$$

Table 1: $(\mathrm{C}_i, \mathrm{NC}_i, \mathrm{NS}_i), i = 1, \ldots, 4$, STPR, and TNR in Example 1

	$(n,p)=(500,400)$					
	$(\mathrm{C}_1,\mathrm{NC}_1,\mathrm{NS}_1)$	$(\mathrm{C}_2,\mathrm{NC}_2,\mathrm{NS}_2)$	$(\mathrm{C}_3,\mathrm{NC}_3,\mathrm{NS}_3)$	$(\mathrm{C}_4,\mathrm{NC}_4,\mathrm{NS}_4)$	STPR	TNR
AWG-Lasso+HDIC	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	1.0	0.948
AWG-Lasso+HDIC$_{\mathrm{II}}$	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	1.0	1.0
	$(n,p)=(1500,2000)$					
	$(\mathrm{C}_1,\mathrm{NC}_1,\mathrm{NS}_1)$	$(\mathrm{C}_2,\mathrm{NC}_2,\mathrm{NS}_2)$	$(\mathrm{C}_3,\mathrm{NC}_3,\mathrm{NS}_3)$	$(\mathrm{C}_4,\mathrm{NC}_4,\mathrm{NS}_4)$	STPR	TNR
AWG-Lasso+HDIC	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	1.0	0.999
AWG-Lasso+HDIC$_{\mathrm{II}}$	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	1.0	1.0

which are suggested in [21]. As observed in (23), X_{i1} and X_{i2} are relevant through the nonlinear functions $g_1(\cdot)$ and $g_2(\cdot)$, and $X_{i3} \sim X_{i5}$ are relevant through the linear functions $g_3(\cdot) \sim g_5(\cdot)$. With

$$\begin{aligned} \mathrm{TNR}_s &= \frac{\sum_{j=6}^{p} I_{\{\mathrm{NS}_{sj}=1\}}}{p-5}, \\ \mathrm{STPR}_s &= \frac{\sum_{j=1}^{2} I_{\{\mathrm{NL}_{sj}=1\}} + \sum_{j=3}^{5} I_{\{\mathrm{L}_{sj}=1\}}}{5}, \end{aligned}$$

the performance measures of the proposed methods in this example are given by

$$\begin{aligned} \mathrm{L}_j &= \frac{1}{50}\sum_{s=1}^{50} \mathrm{L}_{sj},\ \mathrm{NL}_j = \frac{1}{50}\sum_{s=1}^{50} \mathrm{NL}_{sj}, \mathrm{NS}_j = \frac{1}{50}\sum_{s=1}^{50} \mathrm{NS}_{sj}, \\ \mathrm{TNR} &= \frac{1}{50}\sum_{s=1}^{50} \mathrm{TNR}_s,\ \mathrm{STPR} = \frac{1}{50}\sum_{s=1}^{50} \mathrm{STPR}_s, \end{aligned}$$

and summarized in Table 3. As observed in Table 3, $\mathrm{NL}_1 = \mathrm{NL}_2 = \mathrm{L}_3 = \mathrm{L}_4 = \mathrm{L}_5 = 1$ hold for every method and every (n, p) pair, suggesting that AWG-Lasso+HDIC and AWG-Lasso+HDIC$_{\mathrm{II}}$ perform perfectly in identifying the relevant variables as well as the corresponding functional structures. The performance of the two methods on TNR in this example is similar to that in example 2.

In conclusion, we note that the results of this section, together with those obtained in the previous sections, demonstrate that AWG-Lasso+HDIC and AWG-Lasso+HDIC$_{\mathrm{II}}$ have a strong ability to simultaneously identify the relevant variables and their corresponding structures in the high-dimensional quantile regression setup, a feature rarely reported in the literature. Moreover, while AWG-Lasso+HDIC seems to have a better STPR than AWG-Lasso+HDIC$_{\mathrm{II}}$, the latter tends to outperform the former in terms of TNR.

Table 2: $(\mathrm{L}_i, \mathrm{NL}_i, \mathrm{NS}_i), i = 1, \ldots, 4$, STPR, and TNR in Example 2

	$(n, p) = (500, 400)$					
	$(\mathrm{L}_1, \mathrm{NL}_1, \mathrm{NS}_1)$	$(\mathrm{L}_2, \mathrm{NL}_2, \mathrm{NS}_2)$	$(\mathrm{L}_3, \mathrm{NL}_3, \mathrm{NS}_3)$	$(\mathrm{L}_4, \mathrm{NL}_4, \mathrm{NS}_4)$	STPR	TNR
AWG-Lasso+HDIC	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	1.0	0.994
AWG-Lasso+HDIC$_{\mathrm{II}}$	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 0.94, 0.06)	(0.06, 0.94, 0.0)	0.97	0.994
	$(n, p) = (1500, 2000)$					
	$(\mathrm{L}_1, \mathrm{NL}_1, \mathrm{NS}_1)$	$(\mathrm{L}_2, \mathrm{NL}_2, \mathrm{NS}_2)$	$(\mathrm{L}_3, \mathrm{NL}_3, \mathrm{NS}_3)$	$(\mathrm{L}_4, \mathrm{NL}_4, \mathrm{NS}_4)$	STPR	TNR
AWG-Lasso+HDIC	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 1.0)	(0.0, 1.0, 1.0)	1.0	0.989
AWG-Lasso+HDIC$_{\mathrm{II}}$	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 1.0, 1.0)	(0.0, 1.0, 1.0)	1.0	0.998

Table 3: $(\mathrm{L}_i, \mathrm{NL}_i, \mathrm{NS}_i), i = 1, \ldots, 5$, STPR, and TNR in Example 3

	$(n, p) = (500, 400)$						
	$(\mathrm{L}_1, \mathrm{NL}_1, \mathrm{NS}_1)$	$(\mathrm{L}_2, \mathrm{NL}_2, \mathrm{NS}_2)$	$(\mathrm{L}_3, \mathrm{NL}_3, \mathrm{NS}_3)$	$(\mathrm{L}_4, \mathrm{NL}_4, \mathrm{NS}_4)$	$(\mathrm{L}_5, \mathrm{NL}_5, \mathrm{NS}_5)$	STPR	TNR
AWG-Lasso+HDIC	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	1.0	0.993
AWG-Lasso+HDIC$_{\mathrm{II}}$	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	1.0	0.993
	$(n, p) = (1500, 2000)$						
	$(\mathrm{L}_1, \mathrm{NL}_1, \mathrm{NS}_1)$	$(\mathrm{L}_2, \mathrm{NL}_2, \mathrm{NS}_2)$	$(\mathrm{L}_3, \mathrm{NL}_3, \mathrm{NS}_3)$	$(\mathrm{L}_4, \mathrm{NL}_4, \mathrm{NS}_4)$	$(\mathrm{L}_5, \mathrm{NL}_5, \mathrm{NS}_5)$	STPR	TNR
AWG-Lasso+HDIC	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	1.0	0.991
AWG-Lasso+HDIC$_{\mathrm{II}}$	(0.0, 1.0, 0.0)	(0.0, 1.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	1.0	0.997

5 Proofs of the main theorems

First we introduce notation and assumptions. Then we prove Theorems 1 and 2. All the technical proofs are given in the supplement.

We denote the conditional probability and expectation on $\{(\boldsymbol{X}_i, Z_i)\}_{i=1}^n$ by $\mathrm{P}_\epsilon(\cdot)$ and $\mathrm{E}_\epsilon(\cdot)$, respectively.

Assumption A1 is about $|\mathcal{S}_c^0|$ and $|\mathcal{S}_v^0|$.

Assumption A1: There are bounded constants C_c, C_v, M_c, and M_v such that

$$|\mathcal{S}_c^0| \leq C_c < M_c \quad \text{and} \quad |\mathcal{S}_v^0| \leq C_v < M_v.$$

We know M_c and M_v in advance.

This assumption looks restrictive and we may be able to relax this assumption slightly. However, there are still many assumptions and parameters and we decided not to introduce more complications to relax Assumption A1. Note that we can easily relax the conditions on C_c only for Theorem 1 if

$$\sum_{j \in \mathcal{S}_c^0} w_{1j}^2 = O_p(1).$$

Assumptions A2 and A2' are about the relevant non-zero coefficients and coefficient functions. We need to assume that they are large enough to be detected for our consistency results. Recall that L is the dimension of the spline basis and referred to in Assumption A3 and that q_n appeared in (7).

Assumption A2: We have in probability

$$\frac{\min_{j\in\mathcal{S}_c^0}|g_{cj}| \wedge \min_{j\in\mathcal{S}_v^0}\|g_{vj}\|}{L^{1/2}\{(n^{-1}\log p_n)^{1/2}+\lambda|w_{\mathcal{S}^0}|\}} \to \infty.$$

Assumption A2': We have

$$\frac{\min_{j\in\mathcal{S}_c^0}|g_{cj}| \wedge \min_{j\in\mathcal{S}_v^0}\|g_{vj}\|}{q_n^{1/2}(n^{-1}L\log p_n)^{1/2}} \to \infty.$$

Next we consider the smoothness of relevant non-zero coefficient functions and spline approximation.

Assumption A3: We take $L = c_L n^{1/5}$ and use linear or smoother splines. Besides, we have for some positive C_g,

$$\sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0}(\|g_j\|_\infty + \|g_j'\|_\infty + \|g_j''\|_\infty) \le C_g.$$

When Assumption A3 holds, there exists $\boldsymbol{\gamma}_j^* = (\gamma_{1j}^*, \boldsymbol{\gamma}_{-1j}^{*T})^T \in R^L$ for every $j \in \mathcal{S}_c^0 \cup \mathcal{S}_v^0$ such that

$$\sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0}\|g_j - \boldsymbol{\gamma}_j^{*T}\boldsymbol{B}\|_\infty \le C_1 L^{-2},$$
$$\gamma_{1j}^* = L^{1/2}g_{cj}, \quad \text{and} \quad \sum_{j\in\mathcal{S}_v^0}\|g_{vj} - \boldsymbol{\gamma}_{-1j}^{*T}\boldsymbol{B}_{-1}\|_\infty \le C_2 L^{-2},$$

where C_1 and C_2 depend only on C_g and the order of the spline basis. Let $\boldsymbol{\gamma}_{\mathcal{S}^0}^*$ consist of γ_{1j}^*, $j \in \mathcal{S}_c^0$, and $\boldsymbol{\gamma}_{-1j}^*$, $j \in \mathcal{S}_v^0$. For $\mathcal{S}$ including the true $\mathcal{S}^0$, $\boldsymbol{\gamma}_{\mathcal{S}}^*$ means a vector of coefficients for our spline basis to approximate g_j up to the order of L^{-2}. When $j \in \mathcal{S}_c \cap \overline{\mathcal{S}_c^0}$ or $j \in \mathcal{S}_v \cap \overline{\mathcal{S}_v^0}$, the corresponding elements are put to 0. The other elements are γ_{1j}^*, $j \in \mathcal{S}_c^0$, and $\boldsymbol{\gamma}_{-1j}^*$, $j \in \mathcal{S}_v^0$. See section S.2 in the supplement for more details on the above approximations.

We define some notation related to spline approximation, δ_i, δ_{ij}, ϵ_i', and τ_i, by $\delta_{ij} = g_j(Z_j) - \boldsymbol{\gamma}_j^{*T}\boldsymbol{B}(Z_i)$,

$$\delta_i = \sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0} X_{ij}(g_j(Z_i) - \boldsymbol{\gamma}_j^{*T}\boldsymbol{B}(Z_i)) = \sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0} X_{ij}\delta_{ij},$$
$$\epsilon_i' = \epsilon_i + \delta_i, \quad \text{and} \quad \tau_i = \mathrm{P}_\epsilon(\epsilon_i' \le 0). \tag{24}$$

Under Assumptions A3 and A4 below, we have uniformly in i and j,

$$|\delta_{ij}| = O(L^{-2}) \quad \text{and} \quad |\delta_i| \leq C_1 X_M L^{-2} \to 0$$

for some positive C_1, where let X_M be a constant satisfying

$$\max_{i,j} |X_{ij}| \leq X_M.$$

We allow X_M to diverge as in Assumptions A4 and A4'. Note that

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i^2 \leq \Big\{n^{-1}\sum_{i=1}^{n}\Big(\sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0} X_{ij}^2\Big)^2\Big\}^{1/2}\Big\{n^{-1}\sum_{i=1}^{n}\Big(\sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0} \delta_{ij}^2\Big)^2\Big\}^{1/2}. \tag{25}$$

When we examine the properties of our BIC type criteria, we need more smoothness of the coefficient functions to evaluate the approximation bias. We replace Assumption A3 with Assumption A3' for simplicity of presentation. In fact, the Hölder continuity of g_j'' with exponent $\alpha \geq 1/2$ is sufficient if $X_M^4 L^{-2\alpha} = O(L^{-1})$. If X_M is bounced, the proof of Theorem 2 will work if $\alpha = 1/2$. See Lemma 3 at the end of this section. When we assume Assumption A3', we can replace L^{-2} with L^{-3} in the above approximations.
Assumption A3': We take $L = c_L n^{1/5}$ and use quadratic or smoother splines. Besides, we have for some positive C_g,

$$\sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0} (\|g_j\|_\infty + \|g_j'\|_\infty + \|g_j''\|_\infty + \|g_j^{(3)}\|_\infty) \leq C_g.$$

Next we state assumptions on X_M, p, and q_n. When we consider additive models, we can take $X_M = 1$. Assumptions A4 and A4' imply that ι in $p = O(\exp(n^\iota))$ is less than 1/5.
Assumption A4: For any positive k,

$$X_M(\log p_n)^{1/2} n^{-1/10} (\log n)^k \to 0. \tag{26}$$

Besides, $\mathrm{E}\{B_{0l}^2(Z_1)X_{1j}^2\} = O(L^{-1})$ and $\mathrm{E}\{B_{0l}(Z_1)|X_{1j}\} = O(L^{-1})$ uniformly in l and j. Recall that $B_{0l}(z)$ is the l-th element of the B-spline basis.
Assumption A4': In Assumption A4, (26) is replaced with

$$X_M(\log p_n)^{1/2} q_n^{3/2} n^{-1/10} (\log n)^k \to 0.$$

Next we state assumptions on the conditional distribution of ϵ_i on $(\boldsymbol{X}_i, Z_i)$. We denote the conditional distribution function by $F_i(\epsilon)$ and the conditional density function by $f_i(\epsilon)$.

Assumption A5: There exist positive C_{f1}, C_{f2}, and C_{f3} such that uniformly in i,

$$|F_i(u+\delta) - F_i(\delta) - uf_i(\delta)| \le C_{f1}u^2 \quad \text{and} \quad f_i(\delta) \le C_{f2}$$

when $|\delta| + |u| \le C_{f3}$.

Assumption A5': In addition to Assumption A5, $\mathrm{E}\{|\epsilon_i|\} < \infty$ and when $|a| \to 0$, we have uniformly in i,

$$\mathrm{E}_\epsilon[(a - \epsilon_i - \delta_i)I\{0 < \epsilon_i + \delta_i \le a\}] = \frac{a^2}{2} f_i(-\delta_i) + O(|a|^3) \quad \text{for } a > 0,$$

and

$$\mathrm{E}_\epsilon[(\epsilon_i + \delta_i - a)I\{a < \epsilon_i + \delta_i \le 0\}] = \frac{a^2}{2} f_i(-\delta_i) + O(|a|^3) \quad \text{for } a < 0.$$

Actually, when $a > 0$ and $a \to 0$, we have under some regularity conditions that

$$\int_{-\delta_i}^{a-\delta_i} (a - \epsilon_i - \delta_i) f_i(\epsilon) d\epsilon = \frac{a^2}{2} f_i(-\delta_i) + O(a^3).$$

We introduce some more notation and another kind of assumptions to describe properties of the adaptively weighted Lasso estimators.

We define two index sets $\boldsymbol{S}_M$ and $\boldsymbol{S}_{C+M}$. These index sets are defined for Theorem 2 and they are related to Assumption A1.

$$\boldsymbol{S}_M = \{\mathcal{S} \,|\, \mathcal{S}^0 \subset \mathcal{S},\ |\mathcal{S}_c| \le M_c,\ \text{and } |\mathcal{S}_v| \le M_v\} \tag{27}$$

and

$$\boldsymbol{S}_{C+M} = \{\mathcal{S} \,|\, \mathcal{S}^0 \subset \mathcal{S},\ |\mathcal{S}_c| \le C_c + M_c,\ \text{and } |\mathcal{S}_v| \le C_v + M_v\} \tag{28}$$

We define some random variables related to $\boldsymbol{W}_{i\mathcal{S}}$ and describe assumptions on those random variables. The assumptions on those random variables follow from similar assumptions on their population versions and standard technical arguments. We omit the assumptions on the population versions and standard technical arguments here since they are just standard ones in the literature.

We define $\Theta_1(\mathcal{S})$ by

$$\Theta_1(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^n |\boldsymbol{W}_{i\mathcal{S}}|^2 = \frac{1}{n}\sum_{i=1}^n L^{-1} \sum_{j\in\mathcal{S}_c} |X_{ij}|^2 + \frac{1}{n}\sum_{i=1}^n |\boldsymbol{B}_{-1}(Z_i)|^2 \sum_{j\in\mathcal{S}_v} |X_{ij}|^2.$$

For technical and notational convenience, we redefine $\Theta_1(\mathcal{S})$ by $\Theta_1(\mathcal{S}) \vee 1$.

Assumption B1: For some positive C_{B1}, we have $\Theta_1(\mathcal{S}^0) \leq C_{B1}$ with probability tending to 1,

Assumption B1 follows from some mild moment conditions under Assumption A1.

We define $\Theta_2(\mathcal{S})$ and $\Theta_3(\mathcal{S})$ by

$$\Theta_2(\mathcal{S}) = L\lambda_{\min}(\widehat{\Sigma}_{\mathcal{S}}) \quad \text{and} \quad \Theta_3(\mathcal{S}) = L\lambda_{\max}(\widehat{\Sigma}_{\mathcal{S}}),$$

where $\widehat{\Sigma}_{\mathcal{S}} = n^{-1}\sum_{i=1}^{n} f_i(-\delta_i)\boldsymbol{W}_{i\mathcal{S}}\boldsymbol{W}_{i\mathcal{S}}^T$. The following assumptions are about their eigenvalues. Recall that our normalization factor of the basis is L^{-1}.

Assumption B2: For some positive C_{B2}, we have $\Theta_2(\mathcal{S}^0) \geq C_{B2}$ with probability tending to 1.

Assumption B2': For some positive C'_{B2}, we have $\Theta_2(\mathcal{S}) \geq C'_{B2}$ uniformly in $\mathcal{S} \in \boldsymbol{S}_{C+M}$ with probability tending to 1.

Assumption B3: For some positive C_{B3}, we have with probability tending to 1

$$\begin{aligned} &\Theta_3(\mathcal{S}^0 \cup (\{j\}, \phi)) \leq C_{B3} \quad \text{uniformly in } j \in \overline{\mathcal{S}_c^0} \qquad \text{and} \\ &\Theta_3(\mathcal{S}^0 \cup (\phi, \{j\})) \leq C_{B3} \quad \text{uniformly in } j \in \overline{\mathcal{S}_v^0}. \end{aligned}$$

Assumption B3': For some positive C'_{B3}, we have with probability tending to 1

$$\Theta_3(\mathcal{S}) \leq C'_{B3} \quad \text{uniformly in } \mathcal{S} \in \boldsymbol{S}_{C+M}.$$

We define Θ_4 by

$$\Theta_4 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{S}_v^0} X_{ij}^2.$$

Assumption B4: For some positive C_{B4}, we have $\Theta_4 \leq C_{B4}$ with probability tending to 1.

Assumption B4': In addition to Assumption B4, we have for some positive C'_{B4},

$$n^{-1}\sum_{i=1}^{n}\Big(\sum_{j\in\mathcal{S}_c^0\cup\mathcal{S}_v^0} X_{ij}^2\Big)^2 \leq C'_{B4} \quad \text{with probability tending to 1.}$$

Assumption B4' is used to control (25). Assumptions B4 and B4' follow from mild moment conditions under Assumption A1.

We define $\Theta_5(\mathcal{S})$ by $\Theta_5(\mathcal{S}) = \max_{1\le i\le n} |\boldsymbol{W}_{i\mathcal{S}}|^2$. Notice that there are positive constants C_1 and C_2 such that

$$|\boldsymbol{W}_{i\mathcal{S}}|^2 = L^{-1}\sum_{j\in\mathcal{S}_c} X_{ij}^2 + |\boldsymbol{B}_{-1}(Z_i)|^2 \sum_{j\in\mathcal{S}_v} X_{ij}^2 \tag{29}$$
$$\le C_1 X_M^2 (L^{-1}|\mathcal{S}_c| + |\mathcal{S}_v|) \le C_2 X_M^2$$

for any $\mathcal{S} \in \boldsymbol{S}_{C+M}$ under Assumption A1.

We define $\widehat{\Omega}_{\mathcal{S}}$ by $\widehat{\Omega}_{\mathcal{S}} = n^{-1}\sum_{i=1}^n \tau_i(1-\tau_i)\boldsymbol{W}_{i\mathcal{S}}\boldsymbol{W}_{i\mathcal{S}}^T$. The last assumption is about its eigenvalues. Recall that τ_i is defined in (24).

Assumption B5: There is a positive constant C_{B5} such that uniformly in $\mathcal{S} \in \boldsymbol{S}_{C+M}$,

$$\frac{1}{C_{B5}} \le L\lambda_{\min}(\widehat{\Omega}_{\mathcal{S}}) \le L\lambda_{\max}(\widehat{\Omega}_{\mathcal{S}}) \le C_{B5} \quad \text{with probability tending to 1.}$$

We state Proposition 1 before we prove Theorem 1. The proposition gives the convergence rate of the AWG-Lasso estimator. We prove this proposition by following that of Theorem 1 in [7] in the supplement.

We use the proposition with $\mathcal{S} = \mathcal{S}^0$ or with $\mathcal{S} \in \boldsymbol{S}_{C+M}$ and $\lambda = 0$. Let $w_{\mathcal{S}}$ be a vector consisting of $\{w_{1j} \,|\, j \in \mathcal{S}_c\}$ and $\{w_{-1j} \,|\, j \in \mathcal{S}_v\}$. Then we define $|w_{\mathcal{S}}|$ and K_n by

$$|w_{\mathcal{S}}|^2 = \sum_{j\in\mathcal{S}_c} w_{1j}^2 + \sum_{j\in\mathcal{S}_v} w_{-1j}^2 \quad \text{and} \quad K_n(\mathcal{S}) = \sqrt{n^{-1}\Theta_1(\mathcal{S})\log p_n} + \lambda|w_{\mathcal{S}}|.$$

Tentatively we assume the weights are constants, not random variables.

Proposition 1 *Suppose that $\mathcal{S}^0 \subset \mathcal{S}$ and Assumptions A1 and A3-5 hold. Besides we assume*

$$\Big(\frac{\Theta_5(\mathcal{S})}{\Theta_2(\mathcal{S})}\Big)^{1/2} (\Theta_2^{-1/2}(\mathcal{S}) \vee \Theta_4^{1/2}) K_n(\mathcal{S}) L \to 0 \tag{30}$$

and we define η_n by $\eta_n = C_M L K_n(\mathcal{S})$, where C_M satisfies

$$C_M \ge b_1 \Big\{ \frac{1}{\Theta_2(\mathcal{S})} \vee \Big(\frac{\Theta_4}{\Theta_2(\mathcal{S})}\Big)^{1/2} \Big\} \tag{31}$$

for sufficiently large b_1 depending on b_2 in (32). Then we have for any fixed positive b_2 that

$$\mathrm{P}_\epsilon(|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}^{\lambda} - \boldsymbol{\gamma}_{\mathcal{S}}^*| \ge \eta_n) \le \exp(-b_2 \log p_n). \tag{32}$$

Later we use Assumptions B1-4 to control random variables in (30) and (31) in Proposition 1. Here some remarks on Proposition 1 are in order.

Remark 1 When $w_{\mathcal{S}}$ is a random vector and $\lambda > 0$, "$\to 0$" in (30) should be replaced with "$\xrightarrow{p} 0$." Besides, when for some positive C_1, C_2, and C_3,

$$\mathrm{P}(C_1 \leq \Theta_2(\mathcal{S}),\ \Theta_1(\mathcal{S}) \leq C_2,\ \Theta_4 \leq C_3) \to 1,$$

the RHS of (31) is bounded from above in probability and $\Theta_1(\mathcal{S})$ in $K_n(\mathcal{S})$ can be replaced with a constant. Thus we have

$$\mathrm{P}(|\widehat{\boldsymbol{\gamma}}^{\lambda}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| \geq \eta_n) \to 0$$

under (30) in probability with a fixed C_M. Especially when $\mathcal{S} = \mathcal{S}^0$,

$$\eta_n \sim L\{(n^{-1}\log p_n)^{1/2} + \lambda|w_{\mathcal{S}^0}|\}.$$

Remark 2 Since $\Theta_5(\mathcal{S}^0) \leq C_4 X_M^2$ for some positive C_4 under Assumption A1, (30) reduces to $X_M L\{(n^{-1}\log p_n)^{1/2} + \lambda|w_{\mathcal{S}^0}|\} \xrightarrow{p} 0$ in the setup of Remark 1 with $\mathcal{S} = \mathcal{S}^0$ and this is not a restrictive condition.

Remark 3 When $\lambda = 0$ and the assumptions in Theorem 2 hold, we have for $\widehat{\boldsymbol{\gamma}}^{\lambda}_{\mathcal{S}} = \widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}$ that

$$|\widehat{\boldsymbol{\gamma}}^{\lambda}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| = |\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| \leq C_5 L(n^{-1}\log p_n)^{1/2}$$

uniformly in $\mathcal{S} \in \boldsymbol{S}_{C+M}$ with probability tending to 1 for some positive C_5. We use this result in the proof of Theorem 2.

We provide the proof of Theorem 1. We define $\Gamma_{\mathcal{S}}(M)$ by

$$\Gamma_{\mathcal{S}}(M) = \{\boldsymbol{\gamma}_{\mathcal{S}} \in R^{d_V(\mathcal{S})} \,|\, |\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| \leq M\} \tag{33}$$

Proof of Theorem 1) First we prove $(\widehat{\boldsymbol{\gamma}}^{\lambda}_{\mathcal{S}^0}, \mathbf{0}^T)^T \in R^{pL}$ is an global minimizer of (5) by checking the following conditions (34) and (35). These conditions follow from the standard optimization theory as in [36] and [26]. In addition to (34) as in [36] and [26], we should deal with (35) since we are employing group penalties. Hereafter in this proof, we omit the superscript λ and write $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}$ for $\widehat{\boldsymbol{\gamma}}^{\lambda}_{\mathcal{S}^0}$

With probability tending to 1, we have

$$\Big|\frac{1}{n}\sum_{i=1}^{n} L^{-1/2} X_{ij}\rho'_{\tau}(Y_i - \boldsymbol{W}^T_{i\mathcal{S}^0}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0})\Big| \leq \lambda w_{1j} \text{ for any } j \in \overline{\mathcal{S}^0_c} \tag{34}$$

and

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0})\Big|\le\lambda w_{-1j}\text{ for any }j\in\overline{\mathcal{S}_v^0}.\tag{35}$$

We verify only (35) since (34) is easier.

Proposition 1, Remark 1, and the conditions of the theorem imply that

$$|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}-\boldsymbol{\gamma}_{\mathcal{S}^0}^*|\le C_1L\{(n^{-1}\log p_n)^{1/2}+\lambda|w_{\mathcal{S}^0}|)\le C_2L(n^{-1}\log p_n)^{1/2}(\log n)^{k_\lambda}\tag{36}$$

with probability tending to 1 for some positive C_1 and C_2. We define $V_j(\boldsymbol{\gamma}_{\mathcal{S}^0})$ by

$$\begin{aligned}V_j(\boldsymbol{\gamma}_{\mathcal{S}^0})&=n^{-1}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\Big\{\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0})-\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)\Big\}\\&\quad-\mathrm{E}_\epsilon\Big[n^{-1}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\Big\{\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0})-\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)\Big\}\Big]\end{aligned}$$

By considering the upper bounds given in (36), we can take a positive constant C_ξ for any small positive ξ such that with probability larger than $1-\xi$,

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0})\Big|\\&\le\ \Big|\mathrm{E}_\epsilon\Big[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\{\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0})-\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)\}\Big]_{\boldsymbol{\gamma}_{\mathcal{S}^0}=\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}}\Big|\\&+\ \Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)\Big|+\max_{\boldsymbol{\gamma}_{\mathcal{S}^0}\in\Gamma_{\mathcal{S}^0}(C_\xi L(n^{-1}\log p_n)^{1/2}(\log n)^{k_\lambda})}|V_j(\boldsymbol{\gamma}_{\mathcal{S}^0})|.\end{aligned}\tag{37}$$

We use the following two lemmas to evaluate (37). These lemmas are to be proved in the supplement.

Lemma 1 *For some positive C_1, we have*

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{-1}(Z_i)X_{ij}\rho_\tau'(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)\Big|\le C_1(n^{-1}\log p_n)^{1/2}$$

uniformly in $j\in\overline{\mathcal{S}_v^0}$ with probability tending to 1

Lemma 2 *Take any fixed positive C and k and fix them. Then we have*

$$\max_{\boldsymbol{\gamma}_{\mathcal{S}^0}\in\Gamma_{\mathcal{S}^0}(CL(n^{-1}\log p_n)^{1/2}(\log n)^{k})}|V_j(\boldsymbol{\gamma}_{\mathcal{S}^0})|=o_p(\lambda)$$

uniformly in $j\in\overline{\mathcal{S}_v^0}$.

Finally we evaluate

$$\mathrm{E}_\epsilon\Big[\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_{-1}(Z_i)X_{ij}\{\rho'_\tau(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0})-\rho'_\tau(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}^*_{\mathcal{S}^0})\}\Big]_{\gamma_{\mathcal{S}^0}=\widehat{\gamma}_{\mathcal{S}^0}} \tag{38}$$
$$= \frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_{-1}(Z_i)X_{ij}\{F_i(-\delta_i)-F_i(-\delta_i+\boldsymbol{W}_{i\mathcal{S}^0}^T(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}-\boldsymbol{\gamma}^*_{\mathcal{S}^0}))\}.$$

Setting $\widehat{\Delta}^0=\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0}-\boldsymbol{\gamma}^*_{\mathcal{S}^0}$ and recalling Assumption A5, we find that (38) is rewritten as

$$-\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_{-1}(Z_i)X_{ij}f_i(-\delta_i)\boldsymbol{W}_{i\mathcal{S}^0}^T\widehat{\Delta}^0+o_p((n^{-1}\log p_n)^{1/2})=-D_j\widehat{\Delta}^0+o_p((n^{-1}\log p_n)^{1/2}) \tag{39}$$

uniformly in $j\in\overline{\mathcal{S}_v^0}$, where D_j is clearly defined in the above equation.

Assumption B3 implies that for some positive C_1,

$$\lambda_{\max}(D_j^TD_j)\le C_1L^{-2} \tag{40}$$

uniformly in $j\in\overline{\mathcal{S}_v^0}$ with probability tending to 1. This is because D_j is part of $\widehat{\Sigma}_{\mathcal{S}^0\cup(\phi,\{j\})}$. Thus (36) and (40) yield that for some positive C_2,

$$|D_j\widehat{\Delta}^0|\le C_2\{(n^{-1}\log p_n)^{1/2}+\lambda|w_{\mathcal{S}^0}|\} \tag{41}$$

uniformly in $j\in\overline{\mathcal{S}_v^0}$ with probability tending to 1.

By combining (37), Lemmas 1 and 2, (39), and (41), we obtain

$$\Big|\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_{-1}(Z_i)X_{ij}\rho'_\tau(Y_i-\boldsymbol{W}_{i\mathcal{S}^0}^T\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^0})\Big|\le\lambda w_{-1j}$$

uniformly in $j\in\overline{\mathcal{S}_v^0}$ with probability tending to 1. Hence (35) is established.

As for the latter part of the theorem, Assumption A2 implies that γ^*_{1j}, $j\in\mathcal{S}_c^0$, and $\boldsymbol{\gamma}^*_{-1j}$, $j\in\mathcal{S}_v^0$, are large enough to be detected due to Proposition 1 with $\mathcal{S}=\mathcal{S}^0$.

Hence the proof of the theorem is complete.

Now we state the proof of Theorem 2

Proof of Theorem 2) First we deal with the overfitting case. Then let $\mathcal{S}$ satisfy $\mathcal{S}\in\boldsymbol{S}_M$ and $\mathcal{S}\ne\mathcal{S}^0$. See (27) for the definition of $\boldsymbol{S}_M$. "Uniformly in $\mathcal{S}$" means "uniformly in $\mathcal{S}$ satisfying $\mathcal{S}\in\boldsymbol{S}_M$ and $\mathcal{S}\ne\mathcal{S}^0$". We have replaced Assumption A3 with Assumption A3'. We use Assumption A3' only once in the proof (Lemma 3) and we use Assumption A3 in the other part. Assumption A3' can be relaxed in some cases. See Lemma 3 at the end of this section for more details.

If we verify

$$R_V(\boldsymbol{\gamma}^*_{\mathcal{S}^0}) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(\epsilon_i) + O(X_M L^{-2}) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\{\rho_\tau(\epsilon_i)\} + o_p(1), \tag{42}$$

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) = R_V(\boldsymbol{\gamma}^*_{\mathcal{S}^0}) + o_p(1), \quad \text{and} \tag{43}$$

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) - R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) = (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p(n^{-1}\{(\log p_n) \vee (q_n \log p_n)^{1/2}\}) \quad \text{uniformly in } \mathcal{S}, \tag{44}$$

then we have for some positive C_1,

$$\begin{aligned} 0 &\le \log R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) - \log R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) = -\log\Big\{1 + \frac{R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) - R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0})}{R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0})}\Big\} \\ &\le \frac{1}{C_1}\{R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) - R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}})\} \end{aligned} \tag{45}$$

uniformly in $\mathcal{S}$ with probability tending to 1. By (44) and (45), we obtain

$$\begin{aligned} \log R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) - \log R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) &= (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p(n^{-1}\{\log p_n \vee (q_n \log p_n)^{1/2}\}) \\ &< (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))\frac{\log p_n}{2n} q_n \end{aligned}$$

uniformly in $\mathcal{S}$ with probability tending to 1. Hence the proof for the overfitting case is complete.

Thus we have only to prove (42)-(44). We prove only (44) since (42) and (43) are easy to deal with.

(63), (64), and (67) are important when we prove (44). To verify (63), we take a positive M_1 and consider

$$\begin{aligned} & R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}}) \\ + \ & \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}^T_{i\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}})(\tau - I\{\epsilon'_i \le 0\}) - \mathrm{E}_\epsilon\{R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}})\} \\ - \ & \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}^T_{i\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}})(\tau - \tau_i) \\ = \ & \frac{1}{n}\sum_{i=1}^n D_i(\boldsymbol{\gamma}_{\mathcal{S}}), \end{aligned} \tag{46}$$

where $D_i(\boldsymbol{\gamma}_{\mathcal{S}})$ is clearly defined in the above equation, $\tau_i = \mathrm{P}_\epsilon(\epsilon'_i \le 0)$, and $|\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| \le M_1 L(q_n n^{-1}\log p_n)^{1/2}$.

We show that

$$\frac{1}{n}\sum_{i=1}^n D_i(\boldsymbol{\gamma}_{\mathcal{S}}) = O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big) \tag{47}$$

uniformly in $\boldsymbol{\gamma}_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M_1 L(q_n n^{-1} \log p_n)^{1/2})$ and $\mathcal{S}$ for any fixed M_1. To verify (47), we should note that

$$D_i(\boldsymbol{\gamma}_{\mathcal{S}}) = \overline{D}_i(\boldsymbol{\gamma}_{\mathcal{S}}) - \mathrm{E}_\epsilon\{\overline{D}_i(\boldsymbol{\gamma}_{\mathcal{S}})\}, \tag{48}$$

where

$$\overline{D}_i(\boldsymbol{\gamma}_{\mathcal{S}}) = \rho_\tau(Y_i - \boldsymbol{W}_{i\mathcal{S}}^T \boldsymbol{\gamma}_{\mathcal{S}}) - \rho_\tau(\boldsymbol{W}_{i\mathcal{S}}^T \boldsymbol{\gamma}_{\mathcal{S}}^*) + \boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*)(\tau - I\{\epsilon_i' \leq 0\})$$

and that

$$\rho_\tau(\epsilon_i' - a_i) - \rho_\tau(\epsilon_i') = -a_i(\tau - I\{\epsilon_i' \leq 0\}) - (\epsilon_i' - a_i)[I\{\epsilon_i' \leq a_i\} - I\{\epsilon_i' \leq 0\}], \tag{49}$$

where $a_i = \boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*)$.

By using (49), we can obtain the following three facts (50)-(52) uniformly in $\boldsymbol{\gamma}_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M_1 L(q_n n^{-1} \log p_n)^{1/2})$ and $\mathcal{S}$. Note that $C_2, \dots, C_7$ are some positive constants.

$$\max_{1 \leq i \leq n} |\boldsymbol{W}_{i\mathcal{S}}| \leq C_2 X_M (M_c^{1/2} L^{-1/2} + M_v^{1/2}) \leq C_3 X_M \tag{50}$$

$$\max_{1 \leq i \leq n} |\overline{D}_i(\boldsymbol{\gamma}_{\mathcal{S}})| \leq \max_{1 \leq i \leq n} |\boldsymbol{W}_{i\mathcal{S}}| M_1 L(q_n n^{-1} \log p_n)^{1/2} \leq C_4 X_M M_1 L(q_n n^{-1} \log p_n)^{1/2} \tag{51}$$

$$\begin{aligned} \frac{1}{n^2}\sum_{i=1}^n \mathrm{E}_\epsilon[\{\overline{D}_i(\boldsymbol{\gamma}_{\mathcal{S}})\}^2] &\leq \frac{C_5}{n^2}\sum_{i=1}^n |\boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*)|^3 \\ &\leq \frac{C_6}{n} \max_{1 \leq i \leq n} |\boldsymbol{W}_{i\mathcal{S}}| \{M_1 L(q_n n^{-1} \log p_n)^{1/2}\}^3 \lambda_{\max}\Big(n^{-1}\sum_{i=1}^n \boldsymbol{W}_{i\mathcal{S}} \boldsymbol{W}_{i\mathcal{S}}^T\Big) \\ &\leq \frac{C_7 M_1^3 M_2}{n} L^2 X_M (q_n n^{-1} \log p_n)^{3/2} \end{aligned} \tag{52}$$

if

$$\lambda_{\max}\Big(n^{-1}\sum_{i=1}^n \boldsymbol{W}_{i\mathcal{S}} \boldsymbol{W}_{i\mathcal{S}}^T\Big) \leq \frac{M_2}{L}. \tag{53}$$

By using (50)-(52) and Bernstein's inequality, we have

$$\mathrm{P}_\epsilon\Big(\Big|n^{-1}\sum_{i=1}^n D_i(\boldsymbol{\gamma}_{\mathcal{S}})\Big| \geq \frac{\log p_n}{n(\log n)^2}\Big) \leq C_8 \exp\Big\{-\frac{C_9 n^{1/10}(\log p_n)^{1/2}}{M_1^3 M_2 q_n^{3/2} X_M (\log n)^4}\Big\} \tag{54}$$

for any fixed $\boldsymbol{\gamma}_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M_1 L(q_n n^{-1} \log p_n)^{1/2})$ and $\mathcal{S}$ if (53) holds. Note that C_8 and C_9 are some positive constants.

By appealing to the standard argument based on the Lipschitz continuity and (54) and using Assumptions A4' and B5, we obtain (47) uniformly in $\boldsymbol{\gamma}_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M_1 L(q_n n^{-1} \log p_n)^{1/2})$ and $\mathcal{S}$ for any fixed M_1.

We evaluate $\mathrm{E}_\epsilon\{R_V(\boldsymbol{\gamma}_\mathcal{S})-R_V(\boldsymbol{\gamma}^*_\mathcal{S})\}$ in (46) by using (49) and Assumption A5'. Since

$$\mathrm{E}_\epsilon\{\rho_\tau(\epsilon'_i-a_i)-\rho_\tau(\epsilon'_i)\}=\frac{1}{2}f_i(-\delta_i)a_i^2+a_i(\tau_i-\tau)+O(|a_i|^3),$$

where $a_i=\boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})$, we have

$$\mathrm{E}_\epsilon\{R_V(\boldsymbol{\gamma}_\mathcal{S})-R_V(\boldsymbol{\gamma}^*_\mathcal{S})\}=\frac{1}{2n}\sum_{i=1}^n f_i(-\delta_i)a_i^2+\frac{1}{n}\sum_{i=1}^n a_i(\tau_i-\tau)+O\Big(\frac{1}{n}\sum_{i=1}^n|a_i|^3\Big) \tag{55}$$

uniformly in $\boldsymbol{\gamma}_\mathcal{S}\in\Gamma_\mathcal{S}(M_1L(q_nn^{-1}\log p_n)^{1/2})$ and $\mathcal{S}$. Assumption A4' implies that uniformly in $\boldsymbol{\gamma}_\mathcal{S}\in\Gamma_\mathcal{S}(M_1L(q_nn^{-1}\log p_n)^{1/2})$ and $\mathcal{S}$ for any fixed M_1,

$$\frac{1}{n}\sum_{i=1}^n|a_i|^3\le\frac{\max_{i=1}^n|a_i|}{n}\sum_{i=1}^n|a_i|^2=O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big). \tag{56}$$

By (55) and (56), we obtain

$$\begin{aligned}\mathrm{E}_\epsilon\{R_V(\boldsymbol{\gamma}_\mathcal{S})-R_V(\boldsymbol{\gamma}^*_\mathcal{S})\}=&\frac{1}{2}(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})^T\widehat{\Sigma}_\mathcal{S}(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})\\&+\frac{1}{n}\sum_{i=1}^n a_i(\tau_i-\tau)+O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big)\end{aligned} \tag{57}$$

uniformly in $\boldsymbol{\gamma}_\mathcal{S}\in\Gamma_\mathcal{S}(M_1L(q_nn^{-1}\log p_n)^{1/2})$ and $\mathcal{S}$ for any fixed M_1.

By combining (46), (47), and (57), we obtain

$$\begin{aligned}R_V(\boldsymbol{\gamma}_\mathcal{S})-R_V(\boldsymbol{\gamma}^*_\mathcal{S})=&-(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})^T\frac{1}{n}\sum_{i=1}^n\boldsymbol{W}_{i\mathcal{S}}(\tau_i-I\{\epsilon'_i\le 0\})+\frac{1}{2}(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})^T\widehat{\Sigma}_\mathcal{S}(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})\\&+(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})^T\frac{1}{n}\sum_{i=1}^n\boldsymbol{W}_{i\mathcal{S}}(\tau_i-\tau)+O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big)\end{aligned} \tag{58}$$

uniformly in $\boldsymbol{\gamma}_\mathcal{S}\in\Gamma_\mathcal{S}(M_1L(q_nn^{-1}\log p_n)^{1/2})$ and $\mathcal{S}$ for any fixed M_1.

We use (58) to derive a useful expression of $R_V(\widetilde{\boldsymbol{\gamma}}_\mathcal{S})$. Put

$$\boldsymbol{a}_\mathcal{S}=\frac{1}{n}\sum_{i=1}^n\boldsymbol{W}_{i\mathcal{S}}(\tau_i-I\{\epsilon'_i\le 0\}),\ \boldsymbol{b}_\mathcal{S}=\frac{1}{n}\sum_{i=1}^n\boldsymbol{W}_{i\mathcal{S}}(\tau_i-\tau),\ \text{and}\ \overline{\boldsymbol{\gamma}}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S}=\widehat{\Sigma}_\mathcal{S}^{-1}\boldsymbol{a}_\mathcal{S}. \tag{59}$$

According to (87) in Lemma 3 at the end of this section,

$$(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})^T\frac{1}{n}\sum_{i=1}^n\boldsymbol{W}_{i\mathcal{S}}(\tau_i-\tau)=(\boldsymbol{\gamma}_\mathcal{S}-\boldsymbol{\gamma}^*_\mathcal{S})^T\boldsymbol{b}_\mathcal{S}=O_p\Big(\frac{(q_n\log p_n)^{1/2}}{n}\Big) \tag{60}$$

and this term in (58) is negligible uniformly in $\boldsymbol{\gamma}_\mathcal{S}\in\Gamma_\mathcal{S}(M_1L(q_nn^{-1}\log p_n)^{1/2})$ and $\mathcal{S}$ for any fixed M_1.

By applying Bernstein's inequality conditionally on $\{(\boldsymbol{X}_i, Z_i)\}_{i=1}^n$ first and using Assumption B5, we have

$$|\boldsymbol{a}_{\mathcal{S}}|^2 = O_p\Big(\frac{\log p_n}{n}\Big) \tag{61}$$

uniformly in $\mathcal{S}$. Thus we have from Assumption Assumption B2'

$$\overline{\boldsymbol{\gamma}}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}} = O_p(L(n^{-1}\log p_n)^{1/2}) \tag{62}$$

uniformly in $\mathcal{S}$.

We take some $\boldsymbol{\delta}_{\mathcal{S}} \in R^{d_V(\mathcal{S})}$. If $\overline{\boldsymbol{\gamma}}_{\mathcal{S}} + \boldsymbol{\delta}_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M_1 L(q_n n^{-1}\log p_n)^{1/2})$, we have from (58) and (60) that

$$R_V(\overline{\boldsymbol{\gamma}}_{\mathcal{S}}+\boldsymbol{\delta}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}}) = -\frac{1}{2}\boldsymbol{a}_{\mathcal{S}}^T\widehat{\Sigma}_{\mathcal{S}}^{-1}\boldsymbol{a}_{\mathcal{S}} + \frac{1}{2}\boldsymbol{\delta}_{\mathcal{S}}^T\widehat{\Sigma}_{\mathcal{S}}\boldsymbol{\delta}_{\mathcal{S}} + O_p\Big(\frac{(q_n\log p_n)^{1/2}}{n}\Big) + O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big) \tag{63}$$

uniformly in $\boldsymbol{\delta}_{\mathcal{S}}$ and $\mathcal{S}$.

Because of the optimality of $R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}})$ and (63), we should have

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}}) = -\frac{1}{2}\boldsymbol{a}_{\mathcal{S}}^T\widehat{\Sigma}_{\mathcal{S}}^{-1}\boldsymbol{a}_{\mathcal{S}} + O_p\Big(\frac{(q_n\log p_n)^{1/2}}{n}\Big) + O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big) \tag{64}$$

uniformly in $\mathcal{S}$. This expression also holds for $\mathcal{S}^0$. By combining (63) and (64) and setting $\boldsymbol{\delta}_{\mathcal{S}} = \widetilde{\boldsymbol{\gamma}}_{\mathcal{S}} - \overline{\boldsymbol{\gamma}}_{\mathcal{S}}$, we also obtain

$$|\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}} - \overline{\boldsymbol{\gamma}}_{\mathcal{S}}|^2 = O_p\Big(\frac{L(q_n\log p_n)^{1/2}}{n}\Big) + O_p\Big(\frac{L\log p_n}{n(\log n)^2}\Big) \tag{65}$$

uniformly in $\mathcal{S}$. Note that these expressions also hold for $\mathcal{S}^0$. This equation is used later in the underfitting case.

We evaluate the difference between $R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}})$ and $R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0})$. Now write

$$\widehat{\Sigma}_{\mathcal{S}} = \begin{pmatrix} \widehat{\Sigma}_{\mathcal{S}^0} & \widehat{\Sigma}_{\mathcal{S}12} \\ \widehat{\Sigma}_{\mathcal{S}21} & \widehat{\Sigma}_{\mathcal{S}22} \end{pmatrix} \quad \text{and} \quad \boldsymbol{a}_{\mathcal{S}} = \begin{pmatrix} \boldsymbol{a}_{\mathcal{S}^0} \\ \boldsymbol{a}_{\mathcal{S}2} \end{pmatrix} \tag{66}$$

and note that $R_V(\boldsymbol{\gamma}^*_{\mathcal{S}}) = R_V(\boldsymbol{\gamma}^*_{\mathcal{S}^0})$. Thus due to (64), we have only to consider the difference

$$\begin{aligned} \boldsymbol{a}_{\mathcal{S}}^T\widehat{\Sigma}_{\mathcal{S}}^{-1}\boldsymbol{a}_{\mathcal{S}} - \boldsymbol{a}_{\mathcal{S}^0}^T\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{a}_{\mathcal{S}^0} &= \boldsymbol{a}_{\mathcal{S}^0}^T\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Sigma}_{\mathcal{S}12}\widehat{F}_{\mathcal{S}2}\widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{a}_{\mathcal{S}^0} \\ &\quad - 2\boldsymbol{a}_{\mathcal{S}^0}^T\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Sigma}_{\mathcal{S}12}\widehat{F}_{\mathcal{S}2}\boldsymbol{a}_{\mathcal{S}2} + \boldsymbol{a}_{\mathcal{S}2}^T\widehat{F}_{\mathcal{S}2}\boldsymbol{a}_{\mathcal{S}2}, \end{aligned} \tag{67}$$

where

$$\widehat{F}_{\mathcal{S}2} = (\widehat{\Sigma}_{\mathcal{S}22} - \widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Sigma}_{\mathcal{S}12})^{-1}.$$

We will demonstrate that the RHS of (67) has the stochastic order of $(d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p(n^{-1}\log p_n)$ uniformly in $\mathcal{S}$.

From Assumptions B2' and B3', we have for some positive C_1, C_2, and C_3,

$$C_1 L \leq \lambda_{\max}(\widehat{F}_{\mathcal{S}2}) \leq \lambda_{\max}(\widehat{F}_{\mathcal{S}2}) \leq C_2 L \text{ and } \lambda_{\max}(\widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}12}) \leq C_3 L^{-2} \tag{68}$$

uniformly in $\mathcal{S}$ with probability tending to 1.

By applying Bernstein's inequality conditionally on $\{(\boldsymbol{X}_i, Z_i)\}_{i=1}^n$ first and using Assumption B5, we have

$$|\boldsymbol{a}_{\mathcal{S}2}|^2 = (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p\Big(\frac{\log p_n}{nL}\Big) \tag{69}$$

uniformly in $\mathcal{S}$. Hence (68) and (69) imply that the third term on the RHS of (67) satisfies

$$\boldsymbol{a}_{\mathcal{S}2}^T \widehat{F}_{\mathcal{S}2}\boldsymbol{a}_{\mathcal{S}2} = (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p(n^{-1}\log p_n) \tag{70}$$

uniformly in $\mathcal{S}$.

To evaluate the first and second terms on the RHS of (67),

$$(\boldsymbol{a}_{\mathcal{S}^0}^T \widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Sigma}_{\mathcal{S}12})\widehat{F}_{\mathcal{S}2}(\widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{a}_{\mathcal{S}^0}) \quad \text{and} \quad (\boldsymbol{a}_{\mathcal{S}^0}^T \widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Sigma}_{\mathcal{S}12})\widehat{F}_{\mathcal{S}2}\boldsymbol{a}_{\mathcal{S}2}, \tag{71}$$

we evaluate

$$\widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{a}_{\mathcal{S}^0} = \widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_{i\mathcal{S}^0}(\tau_i - I\{\epsilon_i' \leq 0\}) \tag{72}$$

to obtain (76) as well as (69).

Write

$$\widehat{\Sigma}_{\mathcal{S}12} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{d_V(\mathcal{S})-d_V(\mathcal{S}^0)})$$

and note that (68) implies

$$\boldsymbol{s}_j^T\boldsymbol{s}_j = O_p(L^{-2}) \text{ and } \lambda_{\max}(\widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Omega}_{\mathcal{S}^0}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\widehat{\Sigma}_{\mathcal{S}12}) = O_p(L^{-1}) \tag{73}$$

uniformly in j and $\mathcal{S}$ with probability tending to 1. Besides, we have for some positive C_4 and C_5,

$$\max_j |\boldsymbol{s}_j^T\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{W}_{i\mathcal{S}^0}| \leq C_4 L|\boldsymbol{s}_j||\boldsymbol{W}_{i\mathcal{S}^0}| \leq C_5 L|\boldsymbol{s}_j|X_M = O_p(X_M) \tag{74}$$

uniformly in i and $\mathcal{S}$ with probability tending to 1.

Hence by applying Bernstein's inequality conditionally together with (73) and (74), we obtain

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{s}_j^T\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{W}_{i\mathcal{S}^0}(\tau_i - I\{\epsilon_i' \leq 0\}) = O_p(\{(nL)^{-1}\log p_n\}^{1/2}) \tag{75}$$

uniformly in j and $\mathcal{S}$. Therefore (75) yields

$$|\widehat{\Sigma}_{\mathcal{S}21}\widehat{\Sigma}_{\mathcal{S}^0}^{-1}\boldsymbol{a}_{\mathcal{S}^0}|^2 = (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p((nL)^{-1}\log p_n) \tag{76}$$

uniformly in $\mathcal{S}$.

Thus (68), (69), (71), and (76) imply that the first and second terms on the RHS of (67) have the stochastic order of $(d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p(n^{-1}\log p_n)$ uniformly in $\mathcal{S}$ as in (70). We have demonstrated that the RHS of (67) has the stochastic order of $(d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p(n^{-1}\log p_n)$ uniformly in $\mathcal{S}$.

Hence (44) follows from (64) and this evaluation of (67) and the proof of the overfitting case is complete.

Next we consider the underfitting case. For $\mathcal{S} = (\mathcal{S}_c, \mathcal{S}_v)$ that does not include $\mathcal{S}^0$ and satisfies

$$|\mathcal{S}_c| \le M_c \quad \text{and} \quad |\mathcal{S}_v| \le M_v,$$

we put

$$\mathcal{S}^+ = \mathcal{S} \cup \mathcal{S}^0. \tag{77}$$

Then $\mathcal{S}^+ \in \boldsymbol{S}_{C+M}$ in (28). Note that uniform results proved in the overfitting case still hold for $\mathcal{S}^+$ in (77).

Since

$$\log R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) - \log R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) = \log\left\{1 + \frac{R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) - R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0})}{R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0})}\right\}$$

and

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\epsilon_i) + o_p(1) = \mathrm{E}\{\rho_\tau(\epsilon_i)\} + o_p(1), \tag{78}$$

we have only to demonstrate

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) - R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) > C_1 L\zeta_n^2\frac{\log p_n}{2n} \tag{79}$$

uniformly in $\mathcal{S}$ with probability tending to 1 for some C_1 and ζ_n such that $\zeta_n/q_n^{1/2} = C_\zeta$. Note that we should be able to take and fix any sufficiently large C_ζ and that C_1 has to be independent of C_ζ when C_ζ is large. Then Assumption A1 and (78) assure (79) dominates the penalty terms. Since (78) follows from the argument for the overfitting case and Assumption A5', we consider only (79).

From Assumption A2', we have uniformly in $\mathcal{S}$,

$$\frac{|\boldsymbol{\gamma}^*_{\mathcal{S}^0-\mathcal{S}}|}{L(n^{-1}q_n\log p_n)^{1/2}} \to \infty,$$

where $\boldsymbol{\gamma}^*_{\mathcal{S}^0-\mathcal{S}}$ is obtained by removing all the j-th elements satisfying $j \in \mathcal{S} \cap \mathcal{S}^0$ from $\boldsymbol{\gamma}^*_{\mathcal{S}^0}$.

Since $\mathcal{S}^+$ includes $\mathcal{S}^0$ and $\mathcal{S}$ does not include $\mathcal{S}^0$, Proposition 1 with no penalty implies that

$$|(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}^T, \mathbf{0}^T)^T - \widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}| > L\zeta_n (n^{-1}\log p_n)^{1/2} \tag{80}$$

uniformly in $\mathcal{S}$ with probability tending to 1 for $\zeta_n = C_\zeta q_n^{1/2}$. Note that we can take and fix any large C_ζ here. This also holds with $\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}$ replaced by $\overline{\boldsymbol{\gamma}}_{\mathcal{S}^+}$ due to (65).

Let us follow the standard arguments for general underfitting cases. There is an $0 < \alpha < 1$ such that

$$|\alpha((\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}^T, \mathbf{0}^T)^T - \overline{\boldsymbol{\gamma}}_{\mathcal{S}^+})| = L\zeta_n (n^{-1}\log p_n)^{1/2}$$

and set

$$\Delta_{\mathcal{S}} = \alpha((\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}^T, \mathbf{0}^T)^T - \overline{\boldsymbol{\gamma}}_{\mathcal{S}^+}).$$

The arguments from (58) to (64) imply that

$$\begin{aligned} R_V(\overline{\boldsymbol{\gamma}}_{\mathcal{S}^+} + \Delta_{\mathcal{S}}) &\geq R_V(\overline{\boldsymbol{\gamma}}_{\mathcal{S}^+}) + C_2\zeta_n^2 \frac{L\log p_n}{2n} + O_p\Big(\frac{(q_n \log p_n)^{1/2}}{n}\Big) + O_p\Big(\frac{\log p_n}{n(\log n)^2}\Big) \\ &\geq R_V(\overline{\boldsymbol{\gamma}}_{\mathcal{S}^+}) + C_2\zeta_n^2 \frac{L\log p_n}{4n} \geq R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}) + C_2\zeta_n^2 \frac{L\log p_n}{4n} \end{aligned} \tag{81}$$

uniformly in $\mathcal{S}$ with probability tending to 1 for some positive C_2 independent of C_ζ. We used the optimality of $\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}$ and Assumption B5 here.

Because of (81), the convexity of $R_V(\boldsymbol{\gamma}_{\mathcal{S}^+})$, and the definition of $\Delta_{\mathcal{S}}$, we have

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) \geq R_V(\overline{\boldsymbol{\gamma}}_{\mathcal{S}^+} + \Delta_{\mathcal{S}}) \geq R_V(\overline{\boldsymbol{\gamma}}_{\mathcal{S}^+}) \geq R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}) \tag{82}$$

uniformly in $\mathcal{S}$ with probability tending to 1. From (81) and (82), we obtain

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) \geq R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}) + C_2\zeta_n^2 \frac{L\log p_n}{4n} \tag{83}$$

uniformly in $\mathcal{S}$ with probability tending to 1. Recalling the results for the overfitting case such as (64) and the evaluation of (67), we have

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^+}) \geq R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) + (d_V(\mathcal{S}^0) - d(\mathcal{S}^+))\frac{q_n \log p_n}{2n} \tag{84}$$

uniformly in $\mathcal{S}$ with probability tending to 1.

By combining (83) and (84), we get

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) \geq R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) + C_2\zeta_n^2 \frac{L \log p_n}{4n} + (d_V(\mathcal{S}^0) - d(\mathcal{S}^+))\frac{q_n \log p_n}{2n} \tag{85}$$

uniformly in $\mathcal{S}$ with probability tending to 1. Since $d_V(\mathcal{S}^0) - d(\mathcal{S}^+) = O(L)$ from Assumption A1 and $\zeta_n = C_\zeta q_n^{1/2}$, we have from (85) that

$$R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}) > R_V(\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}^0}) + C_3\zeta_n^2 \frac{L \log p_n}{2n} \tag{86}$$

for any sufficiently large fixed C_ζ uniformly in $\mathcal{S}$ with probability tending to 1. Note that C_3 is independent of C_ζ when C_ζ is larger than some value depending on the assumptions.

Hence we have established (79) and the proof for the case is complete.

We state Lemma 3, which is used to evaluate the bias from $(\tau_i - \tau)$ in the proof of Theorem 2. The proof is given in the supplement. Note that the Hölder continuity of g_j'' with exponent α is almost sufficient for $\tau_i - \tau = O_p(X_M L^{-(2+\alpha)})$.

By using the properties of $\boldsymbol{b}_{\mathcal{S}}$ and $\boldsymbol{b}_{\mathcal{S}2}$ in this lemma and replacing $\boldsymbol{a}_{\mathcal{S}}$ with $\boldsymbol{a}_{\mathcal{S}} + \boldsymbol{b}_{\mathcal{S}}$ in (59), we can prove Theorem 2 in the same way if $X_M^4 L^{-2\alpha} = O(L^{-1})$. Recall that $L = c_L n^{1/5}$ in this paper. Both of $|\boldsymbol{b}_{\mathcal{S}}|^2$ and $|\boldsymbol{b}_{\mathcal{S}2}|^2$ have $O_p\Big(\frac{X_M^4 \log n}{L^{5+2\alpha}}\Big)$ and these are not typos.

Lemma 3 *In the setup of Theorem 2, we have*

$$(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*)^T \boldsymbol{b}_{\mathcal{S}} = O_p\Big(\frac{(q_n \log p_n)^{1/2}}{n}\Big) \tag{87}$$

uniformly in $\boldsymbol{\gamma}_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M_1 L (q_n n^{-1} \log p_n)^{1/2})$ *and* $\mathcal{S}$ *for any fixed* M_1*. Let Assumption A3' be replaced with Assumption A3. If* $\tau_i - \tau = O_p(X_M L^{-(2+\alpha)})$ *uniformly in* i *for some nonnegative* α*, we have*

$$|\boldsymbol{b}_{\mathcal{S}}|^2 = O_p\Big(\frac{X_M^4 \log n}{L^{5+2\alpha}}\Big) \quad \text{and} \quad |\boldsymbol{b}_{\mathcal{S}2}|^2 = (d_V(\mathcal{S}) - d_V(\mathcal{S}^0))O_p\Big(\frac{X_M^4 \log n}{L^{5+2\alpha}}\Big)$$

uniformly in $\boldsymbol{\gamma}_{\mathcal{S}}$*, where* $\boldsymbol{b}_{\mathcal{S}2}$ *is defined as* $\boldsymbol{a}_{\mathcal{S}2}$ *in (66) .*

References

[1] A. Belloni and V. Chernozhukov. $l1$-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.*, 39:82–130, 2011.

[2] P. J. Bickel, y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[3] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods Theory and Applications.* Springer, New York, Dordrecht, Heidelberg, London, 2011.

[4] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.

[5] M.-Y. Cheng, T. Honda, J. Li, and H. Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Ann. Statist.*, 42:1819–1849, 2014.

[6] M.-Y. Cheng, T. Honda, and J.-T. Zhang. Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 111:1209–1221, 2016.

[7] J. Fan, Y. Fan, and E. Barut. Adaptive robust variable selection. *Ann. Statist.*, 42:324–351, 2014.

[8] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.*, 106:544–557, 2011.

[9] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348 – 1360, 2001.

[10] J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 109:1270–1284, 2014.

[11] J. Fan and R. Song. Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.*, 38:3567–3604, 2010.

[12] J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.*, 42:819–849, 2014.

[13] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity.* CRC press, Boca Raton, 2015.

[14] X. He, L. Wang, and H. G. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.*, 41:342–369, 2013.

[15] T. Honda and R. Yabe. Variable selection and structure identification for varying coefficient cox models. *Discussion papers, No. 2016-05, Graduate School of Economics, Hitotsubashi University*, 2016.

[16] C.-K. Ing and T. L. Lai. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, 22:1473–1513, 2011.

[17] K. Kato. Group lasso for high dimensional sparse quantile regression models. *arXiv preprint arXiv:1103.1458*, 2011.

[18] Y. Kim and J.-J. Jeon. Consistent model selection criteria for quadratically supported risks. *Forthcoming in Ann. Statist.*, 2016.

[19] E. R. Lee and E. Mammen. Local linear smoothing for sparse high dimensional varying coefficient models. *Electronic Journal of Statistics*, 10:855–894, 2016.

[20] E. R. Lee, H. Noh, and B. U. Park. Model selection via bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.*, 109:216–229, 2014.

[21] H. Lian. Semiparametric estimation of additive quantile regression models by twofold penalty. *Journal of Business & Economic Statistics*, 30:337–350, 2012.

[22] H. Lian, H. Liang, and D. Ruppert. Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statist. Sinica*, 25:591–607, 2015.

[23] S. Lv, H. Lin, H. Lian, and J. Huang. Oracle inequalities for sparese additive quantile regression in reproducing kernel hilbert space. *Forthcoming in Ann. Statist.*, 2017.

[24] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. Royal Statist. Soc. Ser. B*, 70:53–71, 2008.

[25] L. L. Schumaker. *Spline Functions: Basic Theory 3rd ed.* Cambridge University Press, Cambridge, 2007.

[26] B. Sherwood and L. Wang. Partially linear additive quantile regression in ultra-high dimension. *Ann. Statist.*, 44:288–317, 2016.

[27] Y. Tang, X. Song, H. J. Wang, and Z. Zhu. Variable selection in high-dimensional quantile varying coefficient models. *J. Multivar. Anal.*, 122:115–132, 2013.

[28] R. J. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. Ser. B*, 58:267–288, 1996.

[29] S. van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, Cambridge, 2000.

[30] S. van de Geer. *Estimation and testing under sparsity.* Springer, Switzerland, 2016.

[31] H. Wang. Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.*, 104:1512–1524, 2009.

[32] H. J. Wang, Z. Zhu, and J. Zhou. Quantile regression in partially linear varying coefficient models. *Ann. Statist.*, 37:3841–3866, 2009.

[33] J. Yan and J. Huang. Model selection for cox models with time-varying coefficients. *Biometrics*, 68:419–428, 2012.

[34] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. Ser. B*, 68:49–67, 2006.

[35] H. H. Zhang, G. Cheng, and Y. Liu. Linear or nonlinear? automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc.*, 106:1099–1112, 2011.

[36] Q. Zheng, L. Peng, and X. He. Globally adaptive quantile regression with ultra-high dimensional data. *Ann. Statist.*, 43:2225–2258, 2015.

[37] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101:1418–1429, 2006.

Supplement to "Adaptively weighted group Lasso for semiparametric quantile regression model"

by Toshio Honda, Ching-Kang Ing, and Wei-Ying Wu

S.1 Technical results for Theorems 1 and 2

We provide the proofs of Proposition 1 and Lemmas 1-3 here. We omit λ of $\widehat{\boldsymbol{\gamma}}^{\lambda}_{\mathcal{S}}$ for notational simplicity.

First we state Lemma 4 for Proposition 1 and the notation for the lemma. Then we prove Proposition 1 by following Lemma 1 and Theorem 1 in [7]. Next we present the proofs of Lemmas 4, 1, 2, and 3.

Before we state Lemma 4, we define

$$G_{\mathcal{S}}(M) = \sup_{\boldsymbol{\gamma}_{\mathcal{S}}\in\Gamma_{\mathcal{S}}(M)} |\{R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}})\} - \mathrm{E}_{\epsilon}\{R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}})\}|$$

where $\Gamma_{\mathcal{S}}(M)$ is defined in (33).

Lemma 4 *Assume that Assumption A3 holds. For any fixed M, t, and $\mathcal{S}$, we have*

$$\mathrm{P}_{\epsilon}\Big(G_{\mathcal{S}}(M) \geq 4M\sqrt{\frac{\Theta_1(\mathcal{S})}{n}} + t\Big) \leq \exp\Big\{-\frac{nt^2}{8\Theta_1(\mathcal{S})M^2}\Big\}.$$

When $t = K_0M\{n^{-1}\Theta_1(\mathcal{S})\log p_n\}^{1/2}$, we have from Lemma 4 that

$$\mathrm{P}_{\epsilon}\Big(G_{\mathcal{S}}(M) \geq (4+K_0)M\sqrt{\frac{\Theta_1(\mathcal{S})\log p_n}{n}}\Big) \leq \exp(-K_0^2\log p_n/8).$$

Proof of Proposition 1) We follow that of Theorem 1 in [7]. The following arguments do not depend on $\mathcal{S}$.

Taking $M = C_MLK_n(\mathcal{S})$, we evaluate the following expression on $\Gamma_{\mathcal{S}}(M)$.

$$\mathrm{E}_{\epsilon}\{R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}})\} = \mathrm{E}_{\epsilon}\Big[\frac{1}{n}\sum_{i=1}^{n}\{\rho_{\tau}(\epsilon'_i - a_i) - \rho_{\tau}(\epsilon'_i)\}\Big], \tag{S.1}$$

where we use the notation defined in (24) after Assumption A3 such as $\epsilon'_i = \epsilon_i + \delta_i$ and $a_i = \boldsymbol{W}^T_{i\mathcal{S}}(\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}})$. Note that

$$|a_i| \leq |\boldsymbol{W}_{i\mathcal{S}}|M \leq \Theta_5^{1/2}(\mathcal{S})M \to 0$$

due to the assumption of this proposition.

If $a_i > 0$, we have from the definition of $\rho_\tau(\cdot)$ that

$$\rho_\tau(\epsilon_i' - a_i) - \rho_\tau(\epsilon_i') = \int_0^{a_i} I\{0 < \epsilon_i' \le s\} ds + a_i(I\{\epsilon_i' \le 0\} - \tau).$$

Then from Assumption A5, we obtain

$$\begin{aligned}
&\mathrm{E}_\epsilon\Big[\int_0^{a_i} I\{0 < \epsilon_i' \le s\} ds + a_i(I\{\epsilon_i' \le 0\} - \tau)\Big] \\
&= \int_0^{a_i} (F_i(s - \delta_i) - F_i(-\delta_i)) ds + a_i(\tau_i - \tau) \\
&= \frac{1}{2} f_i(-\delta_i) a_i^2 + o(a_i^2) + O(a_i^2 (\log n)^{-1}) + O(|\tau - \tau_i|^2 \log n).
\end{aligned}$$

uniformly in i. Note that $|\tau - \tau_i|^2 \le C_1 |\delta_i|^2$ for some positive C_1 and that we can deal with the case of $a_i < 0$.

Hence the expression in (S.1) can be represented as

$$\frac{1}{2n} \sum_{i=1}^n f_i(-\delta_i) a_i^2 + o\Big(n^{-1} \sum_{i=1}^n a_i^2\Big) + O\Big(n^{-1} \log n \sum_{i=1}^n \delta_i^2\Big). \tag{S.2}$$

The first term of (S.2) is written as

$$\begin{aligned}
\frac{1}{2n} \sum_{i=1}^n f_i(-\delta_i) a_i^2 &= \frac{1}{2} (\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*)^T \frac{1}{n} \sum_{i=1}^n f_i(-\delta_i) \boldsymbol{W}_{i\mathcal{S}} \boldsymbol{W}_{i\mathcal{S}}^T (\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*) \qquad \text{(S.3)} \\
&\ge \frac{\Theta_2(\mathcal{S})}{2L} |\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*|^2.
\end{aligned}$$

As for the third term of (S.2), we have from Assumption A3 that

$$\begin{aligned}
\frac{\log n}{n} \sum_{i=1}^n \delta_i^2 &= \frac{\log n}{n} \sum_{i=1}^n \Big(\sum_{j \in \mathcal{S}_v^0} X_{ij} \delta_{ij}\Big)^2 \le \frac{\log n}{n} \sum_{i=1}^n \Big(\sum_{j \in \mathcal{S}_v^0} X_{ij}^2\Big)\Big(\sum_{j \in \mathcal{S}_v^0} \delta_{ij}^2\Big) \qquad \text{(S.4)} \\
&\le \frac{C_1 \log n}{n L^4} \sum_{i=1}^n \sum_{j \in \mathcal{S}_v^0} X_{ij}^2 \le \frac{C_1 \log n}{L^4} \Theta_4 \qquad \text{(S.5)}
\end{aligned}$$

for some positive C_1. We defined Θ_4 just before Assumption B4.

By combining (S.2), (S.3), and (S.4), we have

$$\mathrm{E}_\epsilon\{R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}_{\mathcal{S}}^*)\} \ge \frac{\Theta_2(\mathcal{S})}{2L} (1 + o(1)) |\boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^*|^2 + O\Big(\frac{\Theta_4 \log n}{L^4}\Big). \tag{S.6}$$

We define $\boldsymbol{\gamma}_{\mathcal{S}}^\alpha$ by

$$\boldsymbol{\gamma}_{\mathcal{S}}^\alpha = \alpha \widehat{\boldsymbol{\gamma}}_{\mathcal{S}} + (1 - \alpha) \boldsymbol{\gamma}_{\mathcal{S}}^* \tag{S.7}$$

for

$$0 \le \alpha = \frac{M}{M + |\widehat{\boldsymbol{\gamma}}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}|} \le 1.$$

Then

$$\boldsymbol{\gamma}^\alpha_{\mathcal{S}} \in \Gamma_{\mathcal{S}}(M).$$

Since the convexity of $Q_V(\boldsymbol{\gamma}_{\mathcal{S}})$ implies that

$$Q_V(\boldsymbol{\gamma}^\alpha_{\mathcal{S}}) \le \alpha Q_V(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) + (1-\alpha) Q_V(\boldsymbol{\gamma}^*_{\mathcal{S}}) \le Q_V(\boldsymbol{\gamma}^*_{\mathcal{S}}),$$

we have with probability larger than or equal to $1 - \exp(-K_0^2 \log p_n/8)$ that

$$\begin{aligned}
&\mathrm{E}_\epsilon[R_V(\boldsymbol{\gamma}_{\mathcal{S}}) - R_V(\boldsymbol{\gamma}^*_{\mathcal{S}})]_{\boldsymbol{\gamma}_{\mathcal{S}} = \boldsymbol{\gamma}^\alpha_{\mathcal{S}}} \qquad \text{(S.8)}\\
\le\ & \frac{1}{n}\sum_{i=1}^n \rho_\tau(\boldsymbol{\gamma}^*_{\mathcal{S}}) - \mathrm{E}_\epsilon\Big\{\frac{1}{n}\sum_{i=1}^n \rho_\tau(\boldsymbol{\gamma}^*_{\mathcal{S}})\Big\} - \frac{1}{n}\sum_{i=1}^n \rho_\tau(\boldsymbol{\gamma}^\alpha_{\mathcal{S}}) + \mathrm{E}_\epsilon\Big[\frac{1}{n}\sum_{i=1}^n \rho_\tau(\boldsymbol{\gamma}_{\mathcal{S}})\Big]_{\boldsymbol{\gamma}_{\mathcal{S}} = \boldsymbol{\gamma}^\alpha_{\mathcal{S}}}\\
&\quad + Q_V(\boldsymbol{\gamma}^\alpha_{\mathcal{S}}) - Q_V(\boldsymbol{\gamma}^*_{\mathcal{S}})\\
&\quad -\lambda \sum_{j\in\mathcal{S}_c} w_{1j}|\gamma^\alpha_{1j}| - \lambda \sum_{j\in\mathcal{S}_v} w_{-1j}|\boldsymbol{\gamma}^\alpha_{-1j}| + \lambda \sum_{j\in\mathcal{S}_c} w_{1j}|\gamma^*_{1j}| + \lambda \sum_{j\in\mathcal{S}_v} w_{-1j}|\boldsymbol{\gamma}^*_{-1j}|\\
\le\ & G_{\mathcal{S}}(M) + \lambda|w_{\mathcal{S}}||\boldsymbol{\gamma}^\alpha_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}|\\
\le\ & (4+K_0)M\Big\{\sqrt{\frac{\Theta_1(\mathcal{S})\log p_n}{n}} + \lambda|w_{\mathcal{S}}|\Big\} \ = \ (4+K_0)MK_n(\mathcal{S}).
\end{aligned}$$

By (S.6) and (S.8), we have

$$\begin{aligned}
|\boldsymbol{\gamma}^\alpha_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}|^2 &\le \frac{2(4+K_0)L}{\Theta_2(\mathcal{S})}\{MK_n(\mathcal{S}) + O(\Theta_4 L^{-4}\log n)\}\\
&\le \frac{2(4+K_0)L}{\Theta_2(\mathcal{S})}\{C_M K_n^2(\mathcal{S})L + O(\Theta_4 L^{-4}\log n)\}
\end{aligned}$$

with probability larger than or equal to $1 - \exp(-K_0^2 \log p_n/8)$. Hence

$$\begin{aligned}
|\boldsymbol{\gamma}^\alpha_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| &\le \frac{\{2(4+K_0)\}^{1/2}}{\Theta_2^{1/2}(\mathcal{S})}\{C_M^{1/2}K_n(\mathcal{S})L + O(\Theta_4^{1/2}L^{-3/2}(\log n)^{1/2})\} \qquad \text{(S.9)}\\
&\le \frac{1}{2}C_M L K_n(\mathcal{S}) = \frac{1}{2}M
\end{aligned}$$

with probability larger than or equal to $1 - \exp(-K_0^2 \log p_n/8)$.

(S.7), (S.9), and simple algebra yield

$$|\widehat{\boldsymbol{\gamma}}_{\mathcal{S}} - \boldsymbol{\gamma}^*_{\mathcal{S}}| \le M = C_M L K_n(\mathcal{S})$$

with probability larger than or equal to $1 - \exp(-K_0^2 \log p_n/8)$.

Hence the proof of the proposition is complete.

Proof of Lemma 4) We follow that of Lemma 1 in [7].

Due to the Lipschitz continuity of $\rho_\tau(u)$ and application of the concentration inequalities (Theorems 14.3 and 14.4 in [3]), we have

$$\begin{aligned}\mathrm{E}_\epsilon\{G_{\mathcal{S}}(M)\} &\le 2\mathrm{E}_\epsilon\Big[\sup_{\boldsymbol{\gamma}_{\mathcal{S}}\in\Gamma_{\mathcal{S}}(M)}\Big|\frac{1}{n}\sum_{i=1}^n \xi_i\{\rho_\tau(Y_i-\boldsymbol{W}_{i\mathcal{S}}^T\boldsymbol{\gamma}_{\mathcal{S}})-\rho_\tau(Y_i-\boldsymbol{W}_{i\mathcal{S}}^T\boldsymbol{\gamma}_{\mathcal{S}}^*)\}\Big|\Big]\\ &\le 4\mathrm{E}_\epsilon\Big[\sup_{\boldsymbol{\gamma}_{\mathcal{S}}\in\Gamma_{\mathcal{S}}(M)}\Big|\frac{1}{n}\sum_{i=1}^n \xi_i\boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_{\mathcal{S}}-\boldsymbol{\gamma}_{\mathcal{S}}^*)\Big|\Big],\end{aligned}$$

where $\{\xi_j\}_{j=1}^n$ is a Rademacher sequence of and independent of $\{(Y_j,\boldsymbol{X}_j,Z_j)\}_{j=1}^n$. Since

$$\begin{aligned}&\Big|\sum_{i=1}^n \xi_i\boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_{\mathcal{S}}-\boldsymbol{\gamma}_{\mathcal{S}}^*)\Big|\\ &= \Big|\sum_{j\in\mathcal{S}_c}\Big(\sum_{i=1}^n \xi_i X_{ij}L^{-1/2}\Big)(\gamma_{1j}-\gamma_{1j}^*)+\sum_{j\in\mathcal{S}_v}\Big\{\sum_{i=1}^n \xi_i X_{ij}\boldsymbol{B}_{-1}^T(Z_i)(\boldsymbol{\gamma}_{-1j}-\boldsymbol{\gamma}_{-1j}^*)\Big\}\Big|\\ &\le |\boldsymbol{\gamma}_{\mathcal{S}}-\boldsymbol{\gamma}_{\mathcal{S}}^*|\Big\{\sum_{j\in\mathcal{S}_c}\Big|\sum_{i=1}^n \xi_i X_{ij}L^{-1/2}\Big|^2+\sum_{j\in\mathcal{S}_v}\Big|\sum_{i=1}^n \xi_i X_{ij}\boldsymbol{B}_{-1}(Z_i)\Big|^2\Big\}^{1/2},\end{aligned}$$

we have

$$\begin{aligned}&\mathrm{E}_\epsilon\{G_{\mathcal{S}}(M)\} \qquad\qquad (\mathrm{S}.10)\\ &\le \frac{4M}{n^{1/2}}\mathrm{E}_\epsilon\Big[\Big\{\frac{1}{n}\sum_{j\in\mathcal{S}_c}\Big|\sum_{i=1}^n \xi_i X_{ij}L^{-1/2}\Big|^2+\frac{1}{n}\sum_{j\in\mathcal{S}_v}\Big|\sum_{i=1}^n \xi_i X_{ij}\boldsymbol{B}_{-1}(Z_i)\Big|^2\Big\}^{1/2}\Big]\\ &\le \frac{4M}{n^{1/2}}\Big[\mathrm{E}_\epsilon\Big\{\frac{1}{n}\sum_{j\in\mathcal{S}_c}\Big|\sum_{i=1}^n \xi_i X_{ij}L^{-1/2}\Big|^2+\frac{1}{n}\sum_{j\in\mathcal{S}_v}\Big|\sum_{i=1}^n \xi_i X_{ij}\boldsymbol{B}_{-1}(Z_i)\Big|^2\Big\}\Big]^{1/2}\\ &\le \frac{4M}{n^{1/2}}\Big\{\frac{1}{n}\sum_{i=1}^n|\boldsymbol{W}_{i\mathcal{S}}|^2\Big\}^{1/2}\le 4M\sqrt{\frac{\Theta_1(\mathcal{S})}{n}}.\end{aligned}$$

Next we apply Massart's inequality (Theorem 14.2 in [3]) to evaluate the stochastic part $G_{\mathcal{S}}(M)-\mathrm{E}_\epsilon\{G_{\mathcal{S}}(M)\}$. Then noticing

$$|\boldsymbol{W}_{i\mathcal{S}}^T(\boldsymbol{\gamma}_{\mathcal{S}}-\boldsymbol{\gamma}_{\mathcal{S}}^*)|^2\le|\boldsymbol{W}_{i\mathcal{S}}|^2|\boldsymbol{\gamma}_{\mathcal{S}}-\boldsymbol{\gamma}_{\mathcal{S}}^*|^2\le|\boldsymbol{W}_{i\mathcal{S}}|^2M^2$$

and

$$\frac{1}{n}\sum_{i=1}^n|\boldsymbol{W}_{i\mathcal{S}}|^2M^2\le\Theta_1(\mathcal{S})M^2,$$

we have as in Lemma 1 in [7]

$$\mathrm{P}_\epsilon\Big(G_{\mathcal{S}}(M)\ge 4M\sqrt{\frac{\Theta_1(\mathcal{S})}{n}}+t\Big)\le\exp\Big\{-\frac{nt^2}{8\Theta_1(\mathcal{S})M^2}\Big\}.$$

We used (S.10) to evaluate $\mathrm{E}_\epsilon\{G_{\mathcal{S}}(M)\}$ in the conditional probability.

Hence the proof of the lemma is complete.

Proof of Lemma 1) Recall that $\boldsymbol{B}(z) = A_0\boldsymbol{B}_0(z)$ and note (S.20) in section S.2. Thus we have only to demonstrate

$$\Big|\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}\rho'_\tau(\epsilon_i+\delta_i)\Big| \le C_1\{(nL)^{-1}\log p_n\}^{1/2} \tag{S.11}$$

uniformly in l and j with probability tending to 1 for some positive C_1. Recall $B_{0l}(z)$ is the l-th element of the B-spline basis.

Note that

$$\mathrm{E}_\epsilon\Big\{\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}\rho'_\tau(\epsilon_i+\delta_i)\Big\} = \frac{1}{n}\sum_{i=1}^{n} B_{0l}X_{ij}(Z_i)(\tau-\tau_i)$$

and $|\tau-\tau_i| = O(L^{-2})$ uniformly in i.

Since Assumption A4 implies

$$\mathrm{E}\Big\{\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}(\tau-\tau_i)\Big\} = O(L^{-3})$$

and

$$\mathrm{Var}\Big\{\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}(\tau-\tau_i)\Big\} = O(n^{-1}L^{-5}),$$

uniformly in l and j, we apply Bernstein's inequality unconditionally and obtain

$$\Big|\mathrm{E}_\epsilon\Big\{\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}\rho'_\tau(\epsilon_i+\delta_i)\Big\}\Big| \le C_2\{(nL^5)^{-1}\log p_n\}^{1/2} + O(L^{-3}) \tag{S.12}$$

uniformly in l and j with probability tending to 1 for some positive C_2.

Noticing that

$$\frac{1}{n}\sum_{i=1}^{n} B_{0l}^2(Z_i)X_{ij}^2 \le C_3L^{-1}$$

uniformly in l and j with probability tending to 1 for some positive C_3, we apply Bernstein's inequality conditionally and obtain

$$\Big|\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}\rho'_\tau(\epsilon_i+\delta_i) - \mathrm{E}_\epsilon\Big\{\frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}\rho'_\tau(\epsilon_i+\delta_i)\Big\}\Big| \le C_4\{(nL)^{-1}\log p_n\}^{1/2} \tag{S.13}$$

uniformly in l and j with probability tending to 1 for some positive C_4.

Hence (S.11) follows from (S.12) and (S.13) and the proof of the lemma is complete.

Proof of Lemma 2) We can prove this lemma almost in the same way as Lemma B.5 in [26] and the detailed proof is very lengthy. We just outline the proof.

First we define $d_{lj}(\boldsymbol{\gamma}_{\mathcal{S}^0})$ by

$$d_{lj}(\boldsymbol{\gamma}_{\mathcal{S}^0}) = \frac{1}{n}\sum_{i=1}^{n} B_{0l}(Z_i)X_{ij}[\rho_\tau'(Y_i - \boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}) - \rho_\tau'(Y_i - \boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)$$
$$- \mathrm{E}_\epsilon\{\rho_\tau'(Y_i - \boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}) - \rho_\tau'(Y_i - \boldsymbol{W}_{i\mathcal{S}^0}^T\boldsymbol{\gamma}_{\mathcal{S}^0}^*)\}]$$

and take and fix any positive C_0. Then as in the proof of Lemma 1, we have only to prove that

$$|d_{lj}(\boldsymbol{\gamma}_{\mathcal{S}^0})| \le C_1\{(nL\log n)^{-1}\log p_n\}^{1/2}$$

uniformly in l, $j \in \overline{\mathcal{S}_v^0}$, and $\boldsymbol{\gamma}_{\mathcal{S}^0} \in \Gamma_{\mathcal{S}^0}(C_0L(n^{-1}\log p_n)^{1/2}(\log n)^k)$ with probability tending to 1 for some positive C_1 depending on C_0.

Note that the conditional variance of $d_{lj}(\boldsymbol{\gamma}_{\mathcal{S}^0})$ is uniformly bounded by

$$\frac{C_2X_M}{nL}L(n^{-1}\log p_n)^{1/2}(\log n)^k \le C_3X_M\{n^{-3}(\log n)^{2k}\log p_n\}^{1/2}$$

with probability tending to 1 for some positive C_2 and C_3. They depend on C_0. Besides, we can cover $\Gamma_{\mathcal{S}^0}(C_0L(n^{-1}\log p_n)^{1/2}(\log n)^k)$ by N open balls with radius

$$[\{C_0L(n^{-1}\log p_n)^{1/2}(\log n)^k\}n^{-2m}]^{1/2}$$

for any large fixed m and this N satisfies

$$N = O(n^{md_V(\mathcal{S}^0)}).$$

See Lemma 2.5 in [29] for this upper bound of N. We denote the centers of the covering open balls by $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_N$. Note that

$$pLN = O(\exp\{\log p_n + md_V(\mathcal{S}^0)\log n\}).$$

For any $\boldsymbol{\gamma}_s$ among the centers, we have by employing Bernstein's inequality conditionally that

$$\mathrm{P}_\epsilon\Big(|d_{lj}(\boldsymbol{\gamma}_s)| \ge C_4\sqrt{\frac{\log p_n}{nL\log n}}\Big) \le \exp\Big\{-C_3\frac{(\log p_n)^{1/2}n^{3/10}}{X_M(\log n)^{k+1}}\Big\}$$

uniformly in $\boldsymbol{\gamma}_s$ with probability tending to 1 for some positive C_4 and C_5 and we also have from Assumption A4 that

$$pLN\exp\Big\{-C_3\frac{(\log p_n)^{1/2}n^{3/10}}{X_M(\log n)^{k+1}}\Big\}=\exp\Big[C_6\{\log p_n+md_V(\mathcal{S}^0)\log n\}-C_3\frac{(\log p_n)^{1/2}n^{3/10}}{X_M(\log n)^{k+1}}\Big]$$
$$\to 0$$

for some positive C_6. Therefore we successfully evaluated $d_{lj}(\boldsymbol{\gamma}_{\mathcal{S}^0})$ at all the centers.

We can evaluate $d_{lj}(\boldsymbol{\gamma}_{\mathcal{S}^0})$ inside the open balls exactly as in the proof of Lemma B.5 in [26] since we can take any large m. Hence the proof of the lemma is complete.

Proof of Lemma 3) We prove the former half by using Assumption A3'. By exploiting (25) and Assumptions A3' and B4', we have

$$\frac{1}{n}\sum_{i=1}^n|a_i(\tau_i-\tau)|\le\Big(n^{-1}\sum_{i=1}^n a_i^2\Big)^{1/2}\Big(n^{-1}\sum_{i=1}^n(\tau_i-\tau)^2\Big)^{1/2}=O_p\Big(\frac{(q_n\log p_n)^{1/2}}{n}\Big).$$

uniformly in $\boldsymbol{\gamma}_{\mathcal{S}}\in\Gamma_{\mathcal{S}}(M_1L(q_nn^{-1}\log p_n)^{1/2})$ and $\mathcal{S}$ since

$$\frac{1}{n}\sum_{i=1}^n a_i^2=O_p(n^{-1}Lq_n\log p_n)\quad\text{and}\quad\frac{1}{n}\sum_{i=1}^n(\tau_i-\tau)^2=O_p(L^{-6})$$

uniformly as well.

Before we consider the latter, we should recall that $\boldsymbol{B}(z)=A_0\boldsymbol{B}_0(z)$, where $\boldsymbol{B}_0(z)=(B_{01}(z),\dots,B_{0L}(z))^T$ is the equispaced B-spline basis on $[0,1]$, and that the first element of $\boldsymbol{B}(z)$ is $L^{-1/2}$. Therefore we should deal with

$$\frac{X_M}{nL^{1/2}}\sum_{i=1}^n|\tau_i-\tau| \tag{S.14}$$

and

$$\frac{X_M}{n}\sum_{i=1}^n\boldsymbol{B}_0(Z_i)|\tau_i-\tau|. \tag{S.15}$$

As for (S.14), we have

$$\frac{X_M}{nL^{1/2}}\sum_{i=1}^n|\tau_i-\tau|=O_p\Big(\frac{X_M^2}{L^{2+1/2+\alpha}}\Big) \tag{S.16}$$

from the assumption on $(\tau_i-\tau)$.

Since we have $\mathrm{E}\{B_{0j}(Z_i)\}=O(L^{-1})$ uniformly in j, we have

$$\frac{X_M}{n}\sum_{i=1}^n B_{0j}(Z_i)|\tau_i-\tau|=O_p\Big(\frac{X_M^2(\log n)^{1/2}}{L^{3+\alpha}}\Big) \tag{S.17}$$

uniformly in j from the standard argument based on Bernstein's inequality.

(S.16) and (S.17) yields that

$$|\boldsymbol{b}_{\mathcal{S}}|^2 = |\mathcal{S}_v| O_p\Big(\frac{X_M^4 \log n}{L^{5+2\alpha}}\Big) + |\mathcal{S}_c| O_p\Big(\frac{X_M^4}{L^{5+2\alpha}}\Big)$$

uniformly in $\mathcal{S}$. Since

$$d_V(\mathcal{S}) = (L-1)|\mathcal{S}_v| + |\mathcal{S}_c|,$$

the result for $\boldsymbol{b}_{\mathcal{S}2}$ follows from the same argument.

Hence the proof of the lemma is complete.

S.2 Properties of B-spline bases

We describe properties of our basis and give comments on some misleading assumptions on spline bases in the literature for reference.

First we describe how to construct our orthonormal spline basis $\boldsymbol{B}(z) = (B_1(z), \dots, B_L(z))^T$ from the equispaced B-spline basis on $[0,1]$, which is denoted by $\boldsymbol{B}_0(z) = (B_{01}(z), \dots, B_{0L}(z))^T$. Recall that $L = c_L n^{1/5}$ in this paper. We also should recall two well-known facts:

$$\sum_{j=1}^{L} B_{0j}(z) = 1 \qquad \text{and} \qquad B_{0j}(z) \geq 0 \tag{S.18}$$

$$\frac{C_1}{L} \leq \lambda_{\min}(\Omega_0) \leq \lambda_{\max}(\Omega_0) \leq \frac{C_2}{L} \tag{S.19}$$

where $\Omega_0 = \int_0^1 \boldsymbol{B}_0(z)\boldsymbol{B}_0^T(z)dz$ and C_1 and C_2 are positive constants and independent of L.

Therefore there exists an $L \times L$ matrix A_0 such that

$$\boldsymbol{B}(z) = A_0 \boldsymbol{B}_0(z), \quad \int_0^1 \boldsymbol{B}(z)\boldsymbol{B}^T(z)dz = A_0 \Omega_0 A_0^T = L^{-1} I_L,$$
$$B_1(z) = L^{1/2}, \quad \text{and} \quad B_2(z) = \sqrt{\frac{12}{L}}\Big(z - \frac{1}{2}\Big).$$

We denote the $L \times L$ identity matrix by I_L.

We can obtain an A_0 numerically by carrying out the Gram-Schmidt orthonormalization. Notice also that

$$C_3 \leq \lambda_{\min}(A_0 A_0^T) \leq \lambda_{\max}(A_0 A_0^T) \leq C_4, \tag{S.20}$$

where C_3 and C_4 are positive constants and independent of L.

When we deal with varying coefficient models, $B_1(z) = L^{-1/2}$ is used for the constant parts and $\boldsymbol{B}_{-1} = (B_2(z), \dots, B_L(z))^T$ is used for the non-constant parts. When we deal with additive models, $B_2(z) = \sqrt{\frac{12}{L}}\left(z - \frac{1}{2}\right)$ is used for the linear parts and $(B_3(z), \dots, B_L(z))^T$ is used for the nonlinear parts.

Next we consider approximation by our spline basis $\boldsymbol{B}(z) = (B_1(z), \boldsymbol{B}_{-1}^T(z))^T = (B_1(z), B_2(z), \boldsymbol{B}_{-2}^T(z))^T$ under Assumption A3. Assume that

$$\|g\|_\infty + \|g'\|_\infty + \|g''\|_\infty \le C_g.$$

Varying coefficient models: There exists $\boldsymbol{\gamma}_{-1}^* \in R^{L-1}$ such that $\|g_n - \boldsymbol{\gamma}_{-1}^{*T}\boldsymbol{B}_{-1}\|_\infty \le C_1 C_g L^{-2}$. We can take $\gamma_1^* = L^{1/2} g_c$.
Additive models: Let $g(x)$ satisfy $\int_0^1 g(x)dx = 0$. Then there exist $\gamma_2^* \in R$ and $\boldsymbol{\gamma}_{-2}^* \in R^{L-2}$ such that

$$\|g_l - \gamma_2^* B_2\|_\infty + \|g_a - \boldsymbol{\gamma}_{-2}^{*T}\boldsymbol{B}_{-2}\|_\infty \le C_2 C_g L^{-2}.$$

Note that C_1 and C_2 are independent of the specific function. We verify the latter here since the former is easier.

Corollary 6.26 in [25] implies that there is $\boldsymbol{\gamma}^* = (\gamma_1^*, \gamma_2^*, \boldsymbol{\gamma}_{-2}^{*T})^T$ such that

$$\|g - \boldsymbol{\gamma}^{*T}\boldsymbol{B}\|_\infty \le C_3 C_g L^{-2} \tag{S.21}$$

since $\boldsymbol{B}(x)$ is constructed from $\boldsymbol{B}_0(x)$. Noticing

$$\gamma_1^* = L^{1/2}\int_0^1 (\boldsymbol{\gamma}^{*T}\boldsymbol{B}(x) - g(x))dx$$

and $|\gamma_1^*| \le C_3 C_g L^{-3/2}$, we can take $\gamma_1^* = 0$ without affecting (S.21).

Put

$$g^*(x) = \gamma_2^* B_2(x) + \boldsymbol{\gamma}_{-2}^{*T}\boldsymbol{B}_{-2}(x) \quad \text{and} \quad g(x) = \gamma_2' B_2(x) + g_a(x),$$

where γ_2' is defined in the second equation and $g_l(x) = \gamma_2' B_2(x)$. Recalling the decomposition of $g(x)$ and that $\boldsymbol{B}(x)$ is an orthonormal basis with the normalization factor of L^{-1} and $\|B_2\|_\infty = O(L^{-1/2})$, we get

$$L^{-1}|\gamma_2^* - \gamma_2'| = \left|\int_0^1 (g^*(x) - g(x))B_2(x)dx\right| \le C_4 C_g L^{-5/2}.$$

Thus we have $|\gamma_2^* - \gamma_2'| \le C_4 C_g L^{-3/2}$ and

$$\|(\gamma_2^* - \gamma_2')B_2\|_\infty \le C_5 C_g L^{-2}. \tag{S.22}$$

Note that C_3, C_4, and C_5 are independent of the specific function. Hence the desired result follows from (S.21) and (S.22).

Finally we consider

$$\Omega_1 = \int_0^1 \boldsymbol{B}_0'(z)(\boldsymbol{B}_0'(z))^T dz,\ \Omega_2 = \int_0^1 \boldsymbol{B}_0''(z)(\boldsymbol{B}_0''(z))^T dz,\ \text{and } \boldsymbol{B}_0(Z_1) - \mathrm{E}\{\boldsymbol{B}_0(Z_1)\}.$$

We demonstrate that both Ω_1 and Ω_2 does not necessarily have desirable properties for theoretical analysis. This conclusion also applies to $\boldsymbol{B}_0(Z_1) - \mathrm{E}\{\boldsymbol{B}_0(Z_1)\}$.

Take a three times continuously differentiable function $g(z)$. Then Corollary 6.26 in [25] implies that for some $\boldsymbol{\gamma} \in R^L$,

$$\begin{aligned}
\|g - \boldsymbol{\gamma}^T \boldsymbol{B}_0\| &\le C_1 L^{-3} \sum_{j=0}^{3} \|g^{(j)}\|, \\
\|g' - \boldsymbol{\gamma}^T \boldsymbol{B}_0'\| &\le C_2 L^{-2} \sum_{j=0}^{3} \|g^{(j)}\|, \\
\|g'' - \boldsymbol{\gamma}^T \boldsymbol{B}_0''\| &\le C_3 L^{-1} \sum_{j=0}^{3} \|g^{(j)}\|.
\end{aligned}$$

where C_1, C_2, and C_3 are independent of $g(z)$.

Taking $g(z) = \sin(2\pi R z)$ with $R \to \infty$ and $R^3/L \to 0$, we have from the above three inequalities that

$$\begin{gathered}
\|g\| \sim 1, \quad \|g'\| \sim R, \quad \|g''\| \sim R^2, \\
\boldsymbol{\gamma}\Omega_0\boldsymbol{\gamma} \sim 1, \quad (\boldsymbol{\gamma}\Omega_1\boldsymbol{\gamma})^{1/2} \sim R, \quad (\boldsymbol{\gamma}\Omega_2\boldsymbol{\gamma})^{1/2} \sim R^2.
\end{gathered}$$

These and (S.19) imply that Ω_1 and Ω_2 have eigenvalues $\widetilde{\lambda}_1$ and $\widetilde{\lambda}_2$ satisfying $\widetilde{\lambda}_j L \to \infty\, (j = 1, 2)$, respectively. This contradicts some critical assumptions in some papers.

To consider $\boldsymbol{B}_0(Z_1) - \mathrm{E}\{\boldsymbol{B}_0(Z_1)\}$, we note the following equations.

$$\sum_{j=1}^{L} \tau_j = 1 \quad \text{and} \quad \begin{pmatrix} B_{02}(Z_1) - \mathrm{E}\{B_{02}(Z_1)\} \\ \vdots \\ B_{0L}(Z_1) - \mathrm{E}\{B_{0L}(Z_1)\} \end{pmatrix} = D\boldsymbol{B}_0(Z_1), \tag{S.23}$$

where $\tau_j = \mathrm{E}\{B_{0j}(Z_1)\}$ and the $(L-1) \times L$ matrix D is defined by

$$D = (0\, I_{L-1}) - \begin{pmatrix} \tau_2 & \cdots & \tau_2 \\ \cdots & \cdots & \cdots \\ \tau_L & \cdots & \tau_L \end{pmatrix}.$$

When Z_1 has a bounded density function, $\tau_j \sim 1/L$ uniformly in j and we have

$$\boldsymbol{i}_{L-1}^T D = (\tau_1 - 1, \tau_1, \dots, \tau_1) \text{ and } |D^T \boldsymbol{i}_{L-1}| \sim 1$$

for $\boldsymbol{i}_{L-1} = (1, \dots, 1)^T \in R^{L-1}$. This means

$$\lambda_{\min}(DD^T) = O(L^{-1}) \text{ and } \lambda_{\min}(D\Omega_0 D^T) = O(L^{-2}).$$

This implies that the basis in (S.23) is not suitable for additive models for this poor eigenvalue property. That is why we have introduced another basis.