Statistica Sinica Preprint No: SS-2024-0351	
Title	Adaptive Estimation for High-Dimensional Quantile Regression with Misspecification and Nonresponse
Manuscript ID	SS-2024-0351
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0351
Complete List of Authors	Wei Xiong, Dianliang Deng, Wanying Zhang and Dehui Wang
Corresponding Authors	Dehui Wang
E-mails	wangdh@jlu.edu.cn
Notice: Accepted author version.	

Adaptive estimation for high-dimensional quantile regression with misspecification and nonresponse

Wei Xiong, Dianliang Deng, Wanying Zhang and Dehui Wang

Liaoning University, University of Regina and Jilin University

Abstract: In high-dimensional data analysis, most sure independence screening (SIS) procedures are significantly affected by both misspecification and missing data, making the results sensitive to the loss of predictive accuracy. On the other hand, classical model averaging methods are typically limited to well-specified structures or imposed restrictive constraints on candidates. To address the gaps, this paper focuses on the conditional quantile estimation in conjunction with inverse probability weighting, the purposes of which are mainly threefold. Firstly, we study the SIS properties under misspecified quantile models. Secondly, we propose an adaptive model averaging algorithm for complex clusters. Thirdly, we develop a robust improvement strategy to enhance asymptotic efficiency with respect to high-dimensional ignorable mechanism. Theoretical properties of the averaging estimator are investigated, including its finite sample performance, the equivalence between adaptation and asymptotic optimality, as well as the consistency of weights. Numerical simulations illustrate the method's ability to efficiently identify the correct specification and maintain resilience against outliers in response probabilities. The real-data example is analyzed to validate our method.

Key words and phrases: Feature screening, Inverse probability weighting, Model averaging, Oracle inequality, Robust inference.

1. Introduction

With the rapid advancement of science and technology, there has been a growing emphasis on high-dimensional statistics. Buhlmann and van de Geer (2011) highlighted the opportu-

nities and challenges in high-dimensional statistical analysis. When the covariate dimension is large, numerous efficient dimension reduction approaches have been developed to address the curse of dimensionality (e.g., regularization methods by Fan and Peng (2003); sufficient dimension reduction techniques by Cook and Forzani (2009), Ma and Zhu (2012); and feature screening procedures by Fan and Lv (2008), Li, Zhong and Zhu (2012)). In recent years, marginal quantile utility based sure independence screening (SIS) method has attracted widespread attention in ultrahigh dimensional quantile regression frameworks. Such approach effectively captures predictor activity patterns across different quantile levels, and has been extended to various model specifications (He, Wang and Hong (2013); Wu and Yin (2015); Kong, Li and Zerom (2019); Jiang, Liang and Wang (2024), among others). Theoretical studies have been thoroughly investigated within fully observed dataset as well.

Nevertheless, data with nonresponse are frequently encountered in many fields such as biostatistics, economics and social science. As introduced by Little and Rubin (2002), ignoring missing data may lead to invalid results, and most of statistical methods rely on efficient estimation of the response mechanism. While progress has been made in high-dimensional settings (Zhao, Wang and Shao (2020); Deng, Yang and Wang (2022); Wang and Liang (2023); etc.), most of studies have primarily focused on correctly specified models (either parametric or nonparametric). If the fitted model involves potential misspecification, most of above conclusions, including SIS properties, are urgent to be adjusted. Consider the τ-th $(0 < \tau < 1)$ conditional quantile function (CQF) of covariates $\boldsymbol{x} = (\boldsymbol{u}^\top, \boldsymbol{z}^\top)^\top$, which takes the following form:

$$Q_\tau(y|\boldsymbol{x}) = q_\tau(\boldsymbol{u}) + \boldsymbol{\beta}_\tau^\top \boldsymbol{z}, \tag{1.1}$$

where $\boldsymbol{z}$ is a high-dimensional vector of linear predictors that may include a number of in-

significant components, y is a scalar-dependent outcome subject to nonresponse, and $q_\tau(\boldsymbol{u})$ is an unknown function. Generally $q_\tau(\boldsymbol{u})$ cannot be precisely identified due to complex externalities: Although nonparametric methods are sometimes efficient, they typically result in higher estimation variance and unsatisfactory finite sample performance if the true structure of $q_\tau(\boldsymbol{u})$ is linear. Conversely, parametric modeling will induce significant bias if $q_\tau(\boldsymbol{u})$ follows a complex nonlinear pattern. Therefore, screening significant parts of $\boldsymbol{z}$ may lead the lack of consistency as they will be affected by misspecification.

The main interest is to estimate (1.1) with multiple candidates of $q_\tau(\boldsymbol{u})$. To properly account for uncertainty associated to model selection when reporting precision estimates, model averaging serves as a feasible scheme to weight competing candidates, further produces a parsimonious combined prediction that guards against overfitting (Ye, Yang and Yang (2018); Zhang et al. (2020)). Over the past decades, numerous studies have explored optimal weighting methods (see early work by Hoeting et al. (1999); Yang (2004); Hansen (2007)), including smoothed criteria that are consistent to model selection (e.g., SAIC, SBIC), Mallows-type information criteria (Zhang et al. (2020); Wang et al. (2024)), cross-validation criteria (Zhang and Liu (2023); Yu, Zhang and Liang (2025)), Kullback-Leibler measures (Ando and Li (2017); Chen et al. (2023)), and so forth. More recently, Xie, Yan and Tang (2021), He et al. (2023), Zeng, Hu and Cheng (2024) developed the method to analyze the incomplete data.

To the best of our knowledge, most existing averaging methods for missing data are developed through weighted loss optimization. However, they exhibit technical challenges when dealing with a complex candidate set. Furthermore, asymptotic results fail to adequately quantify finite sample performance of the weighted estimator with respect to the

optimal candidate estimator. In this paper, we propose an adaptive model averaging algorithm incorporating variable screening and inverse probability weighting (IPW), with the following contributions: (i) We discuss the change of SIS properties under misspecified $q_\tau(\boldsymbol{u})$. Theoretical investigation shows that the number of candidates can be allowed to keep an exponential order. (ii) We develop an IPW-based model aggregation procedure for (1.1) that is applicable for combining complicated models. The predicting accuracy is theoretically established by both oracle inequality and asymptotic optimality, and we survey the relationship between above asymptotic results in our model framework. (iii) We prove the sum of weights corresponding to "quasi-correct" candidates (defined in Section 2.4) converges to one. This is another theoretical development of weight consistency beyond the least squares regression (Zhang and Liu (2019); Fang, Yuan and Tian (2023)),

The last concern is the robust analysis of proposed estimator. Since IPW estimation is sensitive to extreme propensity scores among respondents, the above method will lead to infinite asymptotic variance even the mechanism model is correct (Crump et al. (2009); Busso, DiNardo and McCrary (2014); Austin and Stuart (2017)). As a remedy, this paper explores a strategy of robust improvement. We propose a robust model averaging procedure that incorporates an unbiased restriction, and select the optimal subspace via minimizing asymptotic variance. Theoretical suggestions and simulation studies are both enumerated to promulgate the outperformance against outliers.

The rest of the paper is organized as follows. Section 2.1 provides the investigation of SIS properties. Section 2.2 describes the adaptive model averaging procedure. Sections 2.3–2.4 present the main theoretical properties of adaptive estimator. Section 3 describes an procedure to enhance robustness. Section 4 provides some results of finite sample simulation

studies. Section 5 presents a real-data example, and Section 6 concludes. Lemmas, technical proofs, some further discussions and numerical results are attached in supplementary materials.

2. Methodology and main results

Suppose $\{(\boldsymbol{x}_i^\top, y_i) : i = 1, ..., n\}$ is a set of independent and identically distributed copies of $(\boldsymbol{x}^\top, y)$, where $\boldsymbol{x}_i = (\boldsymbol{u}_i^\top, \boldsymbol{z}_i^\top)^\top$ and $\dim(\boldsymbol{z}) = p$ is divergent with respect to n. Since some of y_i's are subject to missingness, we define r as the response indicator, i.e., $r_i = 1$ if y_i is observed and $r_i = 0$ otherwise. Obviously r is a binary variable with successful probability $\pi(\boldsymbol{x}, y) = \Pr(r = 1|\boldsymbol{x}, y)$, which is named as response mechanism and controls the type of nonresponse: y is missing at random (MAR) when $y \perp r|\boldsymbol{x}$ and y is missing not at random (MNAR) otherwise. Throughout this section we do not investigate the mechanism model, but assume $\hat{\pi}(\boldsymbol{x}, y)$ is a consistent estimator of the response mechanism.

To depict misspecification, let $\Delta = \{\delta_k : k = 1, ..., K\}$ be a set of multiple estimation procedures of $q_\tau(\boldsymbol{u})$, and $\hat{q}_\tau^{(k)}(\boldsymbol{u})$ be the estimator obtained by δ_k. It is worthy recalling that no further constraint will be imposed on δ_k's and they can be given under completely different regularities. Specifically, Δ may contains either parametric or nonparametric procedures (e.g., δ_1 implies a linear model of $q_\tau(\boldsymbol{u})$, δ_2 implies a nonlinear model of $q_\tau(\boldsymbol{u})$), the same procedure with different setting of nuisance parameters (e.g., δ_3 and δ_4 are both nonparametric procedures but with different kernels or bandwidths), and even other prior adaptive procedures.

Note that the proposed methodology depends on a variable screening procedure among multiple $\hat{q}_\tau^{(k)}(\boldsymbol{u})$. We intend to study the alteration of corresponding properties.

2.1 Sure Independence Screening under Misspecified $q_\tau(\mathbf{u})$

We first review the marginal quantile utility based SIS procedure. Consider a special case that $q_\tau^{(k)}(\boldsymbol{u}) = q_\tau(\boldsymbol{u})$ for all δ_k. It is acknowledged that one can adopt varieties of screen criteria to distinguish the significance of components. Denote $\bar{y} = y - q_\tau(\boldsymbol{u})$, the outcome is independent with z_j $(j = 1, ..., p)$ if and only if $Q_\tau(\bar{y}|z_j) = Q_\tau(\bar{y})$, where $Q_\tau(\bar{y}|z_j)$ and $Q_\tau(\bar{y})$ are the τ-th conditional (given z_j) and unconditional quantiles of $\bar{y}$, respectively. However, it is unclear which of $\hat{q}_\tau^{(k)}(\boldsymbol{u})$ provides a better approximation, some of active variables may be excluded from the screened set due to misspecification, potentially leading to theoretical inconsistencies. Let $\xi_\tau^{(k)} = \inf\{t : F_{\bar{y}^{(k)}}(t) \geq \tau\}$ be the τ-th unconditional quantile corresponding to $\bar{y}^{(k)} = y - q_\tau^{(k)}(\boldsymbol{u})$, where $F_y(\cdot)$ denotes the distribution function of y. It can be regarded as a candidate of ξ_τ that is estimated as

$$\hat{\xi}_\tau^{(k)} = \arg\min_{\xi_\tau} n^{-1} \sum_{i=1}^{n} \hat{W}_i \rho_\tau \left(y_i - \hat{q}_\tau^{(k)}(\boldsymbol{u}_i) - \xi_\tau \right),$$

where $\rho_\tau(v) = v(\tau - I(v \leq 0))$ is the check loss function (Koenker (2005)), and $\hat{W}_i = r_i / \hat{\pi}(\boldsymbol{x}_i, y_i)$ is IPW estimator for the i-th copy. Thereby, each z_j has total K pseudo marginal quantile utilities

$$\left\| \hat{f}_j^{(k)}(\tau) \right\|_{1,n} := n^{-1} \sum_{i=1}^{n} \left| \hat{\beta}_{\tau,j}^{(k)} z_{ij} - \hat{\xi}_\tau^{(k)} \right|, \tag{2.2}$$

where $\hat{\beta}_{\tau,j}^{(k)} = \arg\min n^{-1} \sum_{i=1}^{n} \hat{W}_i \rho_\tau(y_i - \hat{q}_\tau^{(k)}(\boldsymbol{u}_i) - \beta_{\tau,j} z_{ij})$. Given a threshold $v_n > 0$, the index set of selected predictors is further estimated by $\hat{\mathcal{I}}_\tau^{(k)} = \{1 \leq j \leq p : \|\hat{f}_j^{(k)}(\tau)\|_{1,n} \geq v_n\}$, which varies across $\delta_k \in \Delta$.

Recall that the inefficient $\hat{q}_\tau^{(k)}(\boldsymbol{u})$ will render the marginal information unreliable through (2.2), this forces us to make the rectification of SIS properties for $\hat{\mathcal{I}}_\tau^{(k)}$'s. Lemma 1 indicates that the convergence of marginal coefficient estimators resolves the problem, thus restrictions

on $\hat{\beta}_{\tau,j}^{(k)}$'s are crucial. Suppose $\mathcal{I}_\tau = \{1 \le j \le p : \beta_j = 0\}$ and $\mathcal{I}_\tau^c = \{1 \le j \le p : \beta_j \ne 0\}$ are true index sets of informative and noninformative components of $\boldsymbol{z}$, respectively, and $\|v\|_{L_q} = \{E|v|^q\}^{1/q}$ is L_q norm $(1 \le q < \infty)$ of the random variable v. The following assumptions are required:

(C.1) For each $\delta_k \in \Delta$, there exists a nonstochastic $\bar{q}_\tau^{(k)}(\boldsymbol{u})$ such that (i) $\sup_{\delta_k \in \Delta} \|\hat{q}_\tau^{(k)}(\boldsymbol{u}) - \bar{q}_\tau^{(k)}(\boldsymbol{u})\|_{L_2} = O(n^{-\alpha_1})$ for $\alpha_1 > 0$; (ii) $\sup_{\delta_k \in \Delta} |E\bar{q}_\tau^{(k)}(\boldsymbol{u})|$ is bounded.

(C.2) There exists $0 < A_\pi < 1$ such that $A_\pi \le \pi(\boldsymbol{x}, y) \le 1 - A_\pi$, and $\|\hat{\pi}(\boldsymbol{x}, y) - \pi(\boldsymbol{x}, y)\|_{L_2} = O(n^{-\alpha_2})$ for $\alpha_2 > 0$;

(C.3) $\max_{1 \le j \le p} \|z_j\|_{L_2}$ is bounded away from 0 and infinity;

(C.4) $(\sup_{\delta_k \in \Delta} \min_{j \in \mathcal{I}_\tau} - \inf_{\delta_k \in \Delta} \max_{j \in \mathcal{I}_\tau^c}) n^{-1} \sum_{i=1}^n |\beta_{\tau,j,0}^{(k)} z_{ij} - \xi_\tau^{(k)}| \ge c > 0$, where $\beta_{\tau,j,0}^{(k)}$ is the unique minimizer of $E\{\rho_\tau(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j} z_j)\}$;

(C.5) $\log(pK) = o(n^{1-4\alpha})$, where $0 < \alpha < 1/2 \cdot \min\{1/2, \alpha_1, \alpha_2\}$;

(C.6) The conditional density function of $y - \bar{q}_\tau^{(k)}(\boldsymbol{u})$ among z_j is uniformly bounded away from 0 and infinity in a neighborhood of $\beta_{\tau,j,0}^{(k)} z_j$. Besides, $|Eq_\tau(\boldsymbol{u})|$ and the marginal density functions of z_j are uniformly bounded away from 0 and infinity.

Conditions (C.1)–(C.2) are employed in model misspecification and missing data analysis. They serve to determine the growing rate of pK. Specifically, (C.1) generalizes the linear misspecification in Theorem 1 of Angrist et al. (2006), and is similar to assumption A3 of White (1982), assumptions in Theorem 1 of Li, Gu and Liu (2020) in other model frameworks (e.g., $\alpha_1 = 1/2$ when $q_\tau(\boldsymbol{u})$ is linear with fixed dimension, see Theorem 3.1 of Lu and Su (2015)). (C.2) implies the uniform boundedness of $\pi(\boldsymbol{x}, y)$, and the order of estimator is

determined by its specification. The related discussion is given in Part A.4 of supplementary materials. Condition (C.3) restricts the moments of high-dimensional variable. Condition (C.4) guarantees the identification for marginal utilities of active components. Since some of insignificant variables exhibit counterfactual signals under misspecification, this condition extends the constraint to the entire Δ. Therefore, (C.4) is generally weaker than that in Thoerem 1 of Mai and Zou (2013), assumption 1 of Wang et al. (2023), and will degenerate into the common case when $q_\tau(\boldsymbol{u})$ is well-specified (see A.4 of supplementary materials). Condition (C.5) shows the major difference from conventional dimensionality of $\boldsymbol{z}$ by connecting multiple candidates. Despite a more stringent constraint than countability, it is by no means mutually exclusive with the finite assumption on Δ in practice. On the other hand, (C.5) shows that the method can handle the model with an exponential-growing number as well. The boundedness in condition (C.6) is adopted in literature of variable screening (He, Wang and Hong (2013)) and model aggregation (Shan and Yang (2009); Gu and Zou (2019)), which implies exponential bounds of $\hat{\beta}_{\tau,j}^{(k)}$'s by concentration inequality.

Theorem 1. *Under conditions* (C.1)–(C.6)*, we have for* $n \to \infty$*,*

(i) $\Pr\left\{\inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}_\tau^c}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n} \geq \sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}_\tau}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\right\} \to 0.$

(ii) (Sure independence screening property) *If* $\bar{q}_\tau^{(k)}(\boldsymbol{u}) = q_\tau(\boldsymbol{u})$ *for some* $\delta_k \in \Delta$ *and* $\min_{j\in\mathcal{I}_\tau}\|f_j(\tau)\|_{1,n} \geq v_n$*,*

$$\Pr\left\{\mathcal{I}_\tau \subset \bigcup_{\delta_k\in\Delta}\hat{\mathcal{I}}_\tau^{(k)}\right\} \to 1,$$

where $\|f_j(\tau)\|_{1,n} = n^{-1}\sum_{i=1}^n|\beta_{\tau,j}z_{ij} - \xi_\tau|$ *and* $\xi_\tau = \inf\{t : F_{\bar{y}}(t) \geq \tau\}$.

Theorem 1 reveals two key points. Firstly, the identification of the utility estimator corresponding to the same predictor is tangled among δ_k's. This implies that not all active

components in $\boldsymbol{z}$ perform higher utilities than noninformative components due to the weaker regularity in (C.4). Therefore, it is unsuitable to select δ_k by (2.2) without further structural assumptions on Δ. Secondly, one can prove that $\beta^{(k)}_{\tau,j,0} = \beta_{\tau,j}$ if $q^{(k)}_\tau(\boldsymbol{u})$ has a consistent estimator of $q_\tau(\boldsymbol{u})$ (see Lemma 3), the screening bias is tolerated and the SIS property is established. Further discussions of constraints in Theorem 1 (ii) are given in A.4 of supplementary materials. Of this view, it provides as one of sufficient conditions for weight convergence in subsequent sections. From the theorem, estimating (1.1) is equivalent to matching propensities among δ_k's on subsets of $\boldsymbol{z}$ with mild dimensionality, which aids the theoretic support to conduct model averaging for reformed candidate clusters. Since the weighting process will not be quantified by marginal utilities, it is convenient to employ a common size of $\hat{\mathcal{I}}^{(k)}_\tau$'s to simplify the construction. This approach also helps guard against the misjudgment of the best model. Therefore, common thresholds (e.g., Fan and Lv (2008); Kong, Li and Zerom (2019); Wang et al. (2023)) are practically feasible to implement the screening procedure.

2.2 Exponential Aggregation

To pursue the global optimality of fitting-prediction accuracy, the next step focuses on an eligible model averaging scheme across all dimension reduced candidates. According to SIS, one natural inspiration is to identify weights of models that are reconstructed by combining $q^{(k)}_\tau(\boldsymbol{u})$'s with survived linear predictors. Generally, for each $\delta_k \in \Delta$, denote the (k,m)-th candidate of CQF as

$$Q^{(k,m)}_\tau(y|\boldsymbol{x}) = q^{(k)}_\tau(\boldsymbol{u}) + \sum_{j\in\hat{\mathcal{I}}^{(k)}_{\tau,m}} \beta_{\tau,j} z_j, \quad m = 1, ..., M_k,$$

where $\hat{\mathcal{I}}_{\tau,m}^{(k)} \subseteq \hat{\mathcal{I}}_{\tau}^{(k)}$. Moreover, let $\hat{\mathcal{M}}_{\tau}^{(k)} = \{Q_{\tau}^{(k,m)}(y|\boldsymbol{x}) : m = 1, ..., M_k\}$ be the set formed by merging M_k submodels. The target cluster, which is the union over all k, is then written as $\hat{\mathcal{S}}_{\tau} = \bigcup_{k=1}^{K} \hat{\mathcal{M}}_{\tau}^{(k)} = \{Q_{\tau}^{(k,m)}(y|\boldsymbol{x}) : m = 1, ..., M_k \text{ and } k = 1, ..., K\}$ and contains a total of $\sum_{k=1}^{K} M_k$ candidates.

Algorithm 1: Adaptive model averaging procedure for CQF

Input: Covariates $\boldsymbol{x} = (\boldsymbol{u}^{\top}, \boldsymbol{z}^{\top})^{\top}$. Outcome y. Response indicator r. Number of splits B. Split ratio c $(0 < c < 1)$. Tuning parameters a_n, λ.

Output: Conditional quantile function $\hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x})$.

Evaluate $\hat{\mathcal{S}}_{\tau}$ and $\hat{W}_i$ by $\mathbb{N} = \{(\boldsymbol{x}_i^{\top}, y_i, r_i)^{\top} : i = 1, ..., n\}$.

Set $b = 1$.

repeat

  Randomly permute $\{1, ..., n\}$ as $\{b_1, ..., b_n\}$.

  Set $N = \max\{1, \lfloor cn \rfloor\}$, randomly split $\mathbb{N}$ into a training set $\mathbb{N}_1^b = \{(\boldsymbol{x}_i^{\top}, y_i, \hat{W}_i)^{\top} : i = b_1, ...b_N\}$ and a testing set $\mathbb{N}_2^b = \{(\boldsymbol{x}_i^{\top}, y_i, \hat{W}_i)^{\top} : i = b_{N+1}, ...b_n\}$.

  Fit all $Q_{\tau}^{(k,m)}(y|\boldsymbol{x}) \in \hat{\mathcal{S}}_{\tau}$ by $\mathbb{N}_1^b$, denote as $\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x})$.

  Calculate aggregated weights in terms of the i-th copy from $\mathbb{N}_2^b$ that $\hat{\Omega}_{(k,m),i} = \omega_{(k,m)}$ for $i = b_{N+1}$, and

$$\hat{\Omega}_{(k,m),i}^{b} = \frac{\omega_{(k,m)} \exp\left\{-\lambda \sum_{l=b_{N+1}}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}_l))\right\}}{\sum_{(k,m)\in\hat{\mathcal{S}}_{\tau}} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=b_{N+1}}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}_l))\right\}}$$

  for $b_{N+2} \le i \le b_n$, where $\{0 \le \omega_{(k,m)} \le 1 : \sum_{(k,m)\in\hat{\mathcal{S}}_{\tau}} \omega_{(k,m)} = 1\}$ is a sequence of prior weights.

until $b = B$

Calculate the adaptive estimator

$$\hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{(k,m)\in\hat{\mathcal{S}}_{\tau}} \hat{\Omega}_{(k,m)}^{b} \hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}), \tag{2.3}$$

where $\hat{\Omega}_{(k,m)}^{b} = (n - b_N)^{-1} \sum_{i=b_{N+1}}^{b_n} \hat{\Omega}_{(k,m),i}^{b}$.

Note that $\hat{\mathcal{S}}_{\tau}$ is composed of models with diverse structures corresponding to δ_k's, and

is contaminated by numerous ineffective candidates or insignificant predictors. Although cross-validation model averaging performs satisfactorily under harsh constraints for Δ and $\hat{\mathcal{I}}_\tau^{(k)}$'s, it encounters technical challenges when applied to a rather general candidate set. We develop an IPW-based exponential aggregating procedure in the hope of achieving the ideal prediction risk. To overcome the theoretical predicament caused by check loss function, we employ a surrogate structure as in Shan and Yang (2009), that is

$$\rho_{\tau,a}(v) = \rho_\tau(v) + a_n v^2,$$

where a_n is a nonnegative scale and shrinks to 0 as n increases. In contrast to Shan and Yang (2009), the order of a_n impacts the finite sample performance, and will not undermine the robustness of quantile regression when n is large. Thereby, the procedure is implemented by Algorithm 1.

In the initial step of the algorithm, it is reasonable to compute propensity scores and identify candidate cluster as latent variables corresponding to subjects so that redundant errors are mitigated during aggregation. Compared with classical model averaging based on smoothed information criteria, above algorithm can be adaptively implemented through a multiple data-splitting process, which effectively reduces estimation variance. Moreover, the weight has an analytical form that is computationally convenient. When we obtain any prior information of candidates, they are incorporated into $\omega_{(k,m)}$'s to adjust baseline weights. For example, it can be adopted by a normalized exponential risk such as SAIC, SBIC, or a uniform propensity $\omega_{(k,m)} = 1/|\hat{\mathcal{S}}_\tau|$. The clustering procedure is driven by variable screening, which can generate nested, non-nested, or hybrid structures. Consequently, a data-driven process is employed to thoroughly search for the optimal cluster configuration among these complex possibilities, as demonstrated at the end of next subsection.

2.3 Adaptation and asymptotic optimality

In this subsection we study theoretical properties of the averaging estimator. As discussed in Yang (2004), Gu and Zou (2019) and others, the adaptation is incarnated through the oracle inequality under desired supervisor and is generally distinguished from the asymptotic optimality in model averaging, while the latter performs as an asymptotic accuracy of combing forecasting. Fortunately, the feature of check loss function has built a bridge between aforementioned properties without excessive constraints. In what follows we also investigate the equivalence.

To make a detailed description, we impose general conditions to explore the performance of the risk.

(C.7) (i) For the (k,m)-th candidate, there exists a nonstochastic $Q_\tau^{(k,m)}(y|\boldsymbol{x})$ such that $\|\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau^{(k,m)}(y|\boldsymbol{x})\|_{L_2} = O(n^{-\alpha_{km}})$; (ii) $\lim_{n\to\infty} a_n = 0$.

(C.8) With probability 1, $\sup_{(k,m)} \sup_{1\le i\le n} |\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}_i) - Q_\tau(y|\boldsymbol{x})| \le C_{1,\tau} < \infty$.

(C.9) $\|e_\tau\|_{\mathrm{SEXP}} := \sup_{k\ge1} k^{-1}\{E|e_\tau|^k\}^{1/k}$ is upper-bounded by a finite $C_{2,\tau}$, where $e_\tau = y - Q_\tau(y|\boldsymbol{x})$.

(C.10) The conditional density function of y among $\boldsymbol{x}$ is uniformly bounded away from 0 and infinity in a neighborhood of $Q_\tau(y|\boldsymbol{x})$.

Condition (C.7) (i) is similar to assumption 1 of Yu, Zhang and Liang (2025), and provides the convergence rate of candidate estimators. It is useful for establishing the asymptotic optimality. (C.7) (ii) shows that the quadratic surrogate has a negligible affect in the weighting process, as it mainly serves as a technical tool. Condition (C.8) is fairly common in the literature of adaptive estimation (Yang (2004); Shan and Yang (2009); Gu and

Zou (2019)) and is less stringent than the boundedness of outcome. This condition requires that all candidates should not deviate excessively from the true model. Besides, (C.7)–(C.8) hold uniformly over b. As indicated in Section 5.2.4 of Vershynin (2010), $\|\cdot\|_{\text{SEXP}}$ is the sub-exponential norm. Condition (C.9) implies that the residual is sub-exponential with a uniformly bounded moment generating function (Lemma 5.15 of Vershynin (2010)). This condition is mild as it depicts a broad spectrum of tail behavior beyond sub-gaussianity, and has extensive application in machine learning theory. Condition (C.10) is quite standard for the inference of quantile regression.

The following theorem shows the adaptation of our estimator in terms of the global risk:

Theorem 2. *Under conditions* (C.1)–(C.10)*, if*

$$0 < \lambda \leq \min\left\{(8e\mathcal{M}_{1,\tau})^{-1}, (\overline{\mathcal{K}}_{1,\tau} + \overline{\mathcal{K}}_{2,\tau})^{-1}\right\},$$

the risk bound of (2.3) *under check loss satisfies*

$$E\rho_\tau\left(y - \hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})\right) \leq \inf_{\substack{(k,m)\in\hat{S}_\tau \\ 1\leq b\leq B}} \left\{E\rho_\tau\left(y - \hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x})\right) + \frac{\log(1/\omega_{(k,m)})}{\lambda(n-b_N)} + O\left(n^{-\alpha_2-2\alpha_{km}}\right)\right\}$$
$$+ a_n(C_{1,\tau}^2 + 2C_{2,\tau}^2)\left\{1 + O(n^{-\alpha_2})\right\}. \tag{2.4}$$

where $\overline{\mathcal{K}}_{1,\tau}$, $\overline{\mathcal{K}}_{2,\tau}$ *and* $\mathcal{M}_{1,\tau}$ *are presented in the proof.*

Remark 1. Through the proof we have $8e\mathcal{M}_{1,\tau} = O(a_n)$ and $\overline{\mathcal{K}}_{1,\tau}+\overline{\mathcal{K}}_{2,\tau} = O(a_n+a_n^{-1})\exp(O(\lambda+a_n\lambda+a_n^2\lambda^2))$, respectively. Set $\log(\lambda) = O(\alpha\log(a_n))$, simple derivation shows that the oracle inequality holds consistently when $1-\alpha \leq 0$. Of such view, the order of tuning λ has the same bound with that be implied in full dataset. (see in, Shan and Yang (2009))

Remark 2. Above theorem derives a simple corollary: the estimator is adaptive in the sense

of L_2 loss such that

$$\left\|\hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right\|_{L_2}^2 \le \inf_{\substack{(k,m)\in\hat{S}_\tau \\ 1\le b\le B}} \left\{ C \cdot \left\|\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right\|_{L_2}^2 + \frac{\log(1/\omega_{(k,m)})}{\lambda(n-b_N)} + O\left(n^{-\alpha_2-2\alpha_{km}}\right) \right\} + O(a_n),$$

where C is calculated by the ratio of the upper-bound to the lower-bound of conditional density. This property reflects the consistency if $\hat{\mathcal{S}}_\tau$ contains correct specification, and will be flexible for robust inference.

Theorem 2 reveals the risk of averaging estimator is governed by the optimal candidate, with excess risk incorporated as a penalty term. It guarantees the finite sample performance of the result. Since there is no further restriction on $|\hat{\mathcal{S}}_\tau|$, the number of candidates can be kept as the same order as (C.5), hence is more general than the polynomial order of Wang et al. (2023). On the other side, the lower bound of $E\rho_\tau\left(y - \hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})\right)$ is asymptotically controlled by the infimum among candidates (see Lemma 5). When the order in (C.7) (i) is minimax optimal, (2.4) can be shaped as an asymptotic risk optimality by shrinking the penalty to 0. We conclude the following result:

Theorem 3. *Let the (k_0, m_0)-th candidate have the infimum L_2 norm over $\hat{\mathcal{S}}_\tau$, and $1/\omega_{(k,m)} = O(|\hat{\mathcal{S}}_\tau|)$. Under conditions* (C.1)–(C.10)*:*

(i) *When $\|Q_{\tau,N}^{(k_0,m_0)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\|_{L_2} > 0$, if $\log|\hat{\mathcal{S}}_\tau| = o(\lambda n)$, the adaptive estimator is asymptotically optimal in the sense that*

$$\frac{E\rho_\tau\left(y - \hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})\right)}{\inf_{\substack{(k,m)\in\hat{S}_\tau \\ 1\le b\le B}} E\rho_\tau\left(y - \hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x})\right)} \le 1 + o(1).$$

(ii) *When $\|Q_{\tau,N}^{(k_0,m_0)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\|_{L_2} = 0$, if $\log|\hat{\mathcal{S}}_\tau| = o\left(\lambda n^{1-2\alpha_{k_0m_0}}\right)$ and $a_n = o\left(n^{-2\alpha_{k_0m_0}}\right)$,*

the adaptive estimator is asymptotically optimal in the sense that

$$\frac{E\rho_\tau\left(y-\hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})\right)}{\inf_{\substack{(k,m)\in\hat{S}_\tau\\1\le b\le B}} E\rho_\tau\left(y-\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x})\right)}=1+o(1).$$

Remark 3. Generally $0<\alpha_{km}\le 1/2$ and achieve its upper-bound under a linear specification with fixed dimension. Summarizing the condition in Theorem 2, our asymptotic risk optimality requires that all candidates are misspecified almost surely, or the optimal convergence rate is slower than 1/4. The latter further restricts the true model should not be fully linear with fixed dimension (e.g., Theorem 3.2 of Lu and Su (2015)), also excludes some nonparamtric estimators. Through the proof, it is easy to explain that when such a candidate is optimal, the risk of the estimator decays faster than the excess risk.

Theorem 3 demonstrates that the global risk of averaging estimator asymptotically approximates the best candidate, performing as a computable measure in practice. Theoretically, it performs better than selecting the single model in the sense of all-misspecification. This result differs from the asymptotic risk optimality as Theorem 3.3 of Lu and Su (2015) and Theorem 2 of Wang et al. (2023), because latter shown the prediction risk is asymptotically identical to the optimal convex combination of candidates. Generally one can hardly show the exact equivalence between above optimalities, unless $\hat{\mathcal{S}}_\tau$ satisfies certain specific conditions. Related discussions are sketchily presented in Part B of supplementary materials.

The remaining challenge for application lies in determining a practical range of λ. Since λ depends on the distribution of random error, which may vary across probability levels, $C_{1,\tau}$ and $C_{2,\tau}$ cannot be easily determined even if variance estimation is avoided. Besides, theoretical bounds are inevitably influenced by prior criteria, and explicit solutions are difficult to compute when $\overline{\mathcal{K}}_{1,\tau}$ and $\overline{\mathcal{K}}_{2,\tau}$ involve exponential-type formulas. Gu and Zou (2019)

suggested an empirical range to alleviate the affect for convergence. Alternatively, the idea can be substantively achieved through a data-driven strategy. Note that although a correct δ_k is necessary to guarantee large sample properties of dimension reduction, potential underfitting for $\boldsymbol{z}$ frequently occurs with a single clustering scheme. It is essential to treat the clustering strategy and λ as "tuning parameters" and optimize them jointly. Fix $M_k = 1$ and $\hat{\mathcal{I}}_{\tau,m}^{(k)} = \hat{\mathcal{I}}_{\tau}^{(k)}$, the proposed estimator reaches a relatively minimal risk for correct δ_k's, and we thus regard the clustering scheme as a baseline case. Denote $\mathbb{S}$ as a collection of all possible clustering strategies that encompass the baseline scheme, and Λ as a subset of the support of λ. In this sense, an advisable program for evaluating (1.1) is to compare multiple classifiers that traverse all elements on Λ and $\mathbb{S}$ such that

$$\hat{Q}_{\tau}^{\text{opt}}(y|\boldsymbol{x}) = \underset{(\lambda,\hat{\mathcal{S}}_{\tau})\in(\Lambda,\mathbb{S})}{\arg\min}\ n^{-1}\sum_{i=1}^{n}\hat{W}_i\rho_{\tau}\left(y_i - \hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x}_i)\right). \tag{2.5}$$

The estimator with the optimal candidate cluster satisfies asymptotic properties as well.

2.4 Weight convergence

Another principal aspect pertains to the asymptotic behavior of weights associated with candidate clusters. Given the discussion in Section 2.1 that only predictors within $\mathcal{I}_{\tau}$ are beneficial for predicting, we wish to assign maximal weights to candidates that leverage the best $q_{\tau}^{(k)}(\boldsymbol{u})$ and include a wide array of significant predictors. Such assignment ensures that all underfitted candidates (including those with misspecification) become negligible, due to the implicit constraint that weights sum to one. If above property holds in probability, the weight estimator is said to be "over-consistent". Drawing on the conception of weak inclusion property defined by Chen et al. (2023), we demonstrate the convergence for the weight estimator.

Suppose there are $1 \le K_1 \le K$ consistent procedures for $q_\tau(\boldsymbol{u})$. Without loss of generality they are labeled as the first K_1 elements in Δ (i.e., $\bar{q}_\tau^{(k)}(\boldsymbol{u}) = q_\tau(\boldsymbol{u})$ for $k = 1, ..., K_1$). Let $\mathcal{K}_{\tau,T} = \{q_\tau^{(k)}(\boldsymbol{u}) : k = 1, ..., K_1\}$, we refer to $Q_\tau^{(k,m)}(y|\boldsymbol{x})$ as a "quasi-correct" candidate of (1.1) if $q_\tau^{(k)}(\boldsymbol{u}) \in \mathcal{K}_{\tau,T}$ and $\mathcal{I}_\tau \subset \hat{\mathcal{I}}_{\tau,m}^{(k)}$. The weak inclusion property of the target cluster is equivalent to the inclusion of at least one quasi-correct candidate in its limit version, that is, $Q_\tau^{(k,m)}(y|\boldsymbol{x}) \in \mathcal{S}_\tau$. According to Theorem 1 (ii), we claim $\hat{\mathcal{S}}_\tau$ has the limit weak inclusion property as $n \to \infty$ when Δ contains consistent procedures. For this scenario the specification of Chen et al. (2023) could be regarded as an exceptional instance where δ_k implies an additive structure and $K = 1$.

Denote $\mathcal{S}_{\tau,\text{cor}}$ as the set of quasi-correct candidates, and let $\hat{\Omega}_{\text{cor}} = \sum_{(k,m)\in\mathcal{S}_{\tau,\text{cor}}} \hat{\Omega}_{(k,m)}$, where $\hat{\Omega}_{(k,m)}$ is the weight estimator through a single random permutation. In addition, let $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{\tau,1}^\top, \boldsymbol{\beta}_{\tau,0}^\top)^\top$ such that $\boldsymbol{\beta}_{\tau,1}$ contains all significant components. Note that the convergence of $\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x})$'s and the response probability have been previously assumed, one can verify that the prediction risk of candidate estimators in $\mathcal{S}_{\tau,\text{cor}}$ is consistent to $E\rho_\tau(y - q_\tau(\boldsymbol{u}) - \boldsymbol{\beta}_{\tau,1}^\top \boldsymbol{z}_1)$ under common constraints of the conditional distribution (e.g., Knight (1998); Wang, Wu and Li (2012); He, Wang and Hong (2013)), where each component of $\boldsymbol{z}_1$ corresponds one-to-one to the elements in $\mathcal{I}_\tau$. For this aspect, condition (C.10) is transformed into the uniform boundedness in the neighborhood of $q_\tau(\boldsymbol{u}) + \boldsymbol{\beta}_{\tau,1}^\top \boldsymbol{z}_1$, which instrumentally implies the follow result:

Theorem 4 (Over-consistency)**.** *Under conditions* (C.1)–(C.10)*, if $\hat{\mathcal{S}}_\tau$ has the limit weak inclusion property, then $\hat{\Omega}_{\text{cor}} \xrightarrow{p} 1$ as $n \to \infty$.*

Theorem 4 indicates that the convergence is independent of the restriction on the number of candidates as well. This property ensures the feasibility of the suggested data-driven

optimization. Indeed, dimensional requirements for linear predictors of candidates are scrutinized by the methodology in δ_k and have been numerous fruitful research for linear and nonlinear specifications. This will not emerge as a dominant concern of the paper.

3. Improving Robustness

It is typical to encounter outliers entering the IPW estimator, and the sensitivity to near-zero response probabilities becomes a notable critique. Note that extremely large values of $\hat{W}_i$ may emerge under mild misspecification for the response mechanism. Furthermore, above problem will be amplified under dimensional reduction and sub-exponential assumptions common in literature of ultrahigh dimensional analysis (Ma and Wang (2020)). In this section, we explore an alternative method to enhance the robustness of the adaptive estimator.

We illustrate the idea with a parametric MAR mechanism model. Denote $\boldsymbol{x} = (1, \boldsymbol{u}^\top, \boldsymbol{z}^\top)^\top$ and $\pi(\boldsymbol{x}, y) = g^{-1}(\boldsymbol{\phi}^\top \boldsymbol{x}) := \pi(\boldsymbol{x}; \boldsymbol{\phi})$, where $g(\cdot)$ is a well-specified canonical link function and $\boldsymbol{\phi}$ is the unknown parameter vector. As we can see, the estimation of $\pi(\boldsymbol{x}; \boldsymbol{\phi})$ can be accomplished by variable screening (e.g., Fan and Song (2010); Mai and Zou (2013)) and model averaging in the high-dimensional framework. Let $\mathcal{L}$ be the index set of screened variables. The original model is partitioned into total L candidates $\pi_{(1)}, ..., \pi_{(L)}$ such that

$$\pi_{(l)} = g^{-1}\left(\phi_0 + \sum_{m \in \mathcal{L}_l} \phi_m x_m\right) := g^{-1}\left(\boldsymbol{\phi}_{(l)}^\top \boldsymbol{x}_{(l)}\right), \quad \mathcal{L}_l \subset \mathcal{L}, \quad l = 1, ..., L.$$

By Ando and Li (2017), the jackknife model averaging (JMA) procedure is to assign weights by maximizing

$$\mathrm{CV}(\boldsymbol{\omega}) = \sum_{i=1}^{n}\left[r_i \log g^{-1}\left(\sum_{l=1}^{L} \omega_l \hat{\boldsymbol{\phi}}_{(l)[-i]}^\top \boldsymbol{x}_{(l)i}\right) + (1 - r_i)\log\left\{1 - g^{-1}\left(\sum_{l=1}^{L} \omega_l \hat{\boldsymbol{\phi}}_{(l)[-i]}^\top \boldsymbol{x}_{(l)i}\right)\right\}\right],$$

where $\hat{\boldsymbol{\phi}}_{(l)[-i]}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\phi}_{(l)}$ with the i-th observation deleted, and the support of weights is relaxed by removing the constraint $\sum_{l=1}^{L} \omega_l = 1$.

However, one challenge arises when applying JMA for evaluating propensity scores: extreme likelihoods are inevitable among candidates due to the difficulty in achieving a balance between appropriate weights and candidate sets. The estimator is probably sensitive to the failure of efficiency in the sense that a quantity of propensities is computed close to 0 even if $g(\cdot)$ is correct. Note that $E(r/\pi(\boldsymbol{x};\boldsymbol{\phi})-1)=0$, it is acknowledged to portray a critical role in counteracting the misspecification introduced by Kullback-Leibler divergence. Therefore, its empirical version is referred to as an "unbiased constraint" to guarantee the consistency. The JMA estimator of the response mechanism is modified as

$$\hat{\pi}(\boldsymbol{x})=g^{-1}\left(\sum_{l=1}^{L}\hat{\omega}_l\hat{\boldsymbol{\phi}}_{(l)}^{\top}\boldsymbol{x}_{(l)}\right), \tag{3.6}$$

coupling all candidates with weights chosen by the following constrained optimization:

$$\begin{aligned}&\hat{\boldsymbol{\omega}}=\underset{\boldsymbol{\omega}\in H_\omega}{\arg\max}\,\mathrm{CV}(\boldsymbol{\omega});\\ &\text{s.t.}\ \sum_{i=1}^{n}r_i/g^{-1}\left(\sum_{l=1}^{L}\omega_l\hat{\boldsymbol{\phi}}_{(l)}^{\top}\boldsymbol{x}_{(l)i}\right)=n,\end{aligned}$$

where $H_\omega=\{\boldsymbol{\omega}:(\omega_1,...,\omega_L)\in[0,1]^L\}$, and $\hat{\boldsymbol{\phi}}_{(l)}$ is the full-data MLE corresponding to the l-th candidate.

Remark 4. To efficiently predict CQF we only focus on the unbiasness of IPW. Therefore, the prediction of response mechanism is not our main objective. Numerical results in supplementary materials also show that high prediction precision for $\hat{\pi}(\boldsymbol{x})$ may not lead to improved performance for (1.1). Note that the asymptotic optimality of (3.6) is not affected by either the unbiased constraint or the nonnegative boundary of ω_l's (Ando and Li (2017)). This inspires us to generalize the weight space as $H_\omega=\{\boldsymbol{\omega}:(\omega_1,...,\omega_L)\in[-c,c]^L,\ 0<c<\infty\}$ to tolerate misspecification.

Although the constraint fits the unbiasedness of IPW estimation, it still falls short in

balancing certain subjects with inherently teeny response probabilities, which primarily affect SIS and exponential aggregation. A nature idea is to exclude outliers to achieve the asymptotically efficient estimator, which is the other modification of the procedure.

We invoke a sketchy conditional quantile estimation to examine the methodology. Let $\bar{\mathbb{X}}$ be the support of $\bar{\boldsymbol{x}}_i = (\boldsymbol{x}_i^\top, y_i)^\top$ for $i = 1, ..., n$, and $\hat{Q}_\tau^{\bar{\mathbb{X}}}(y|\boldsymbol{x})$ be the general IPW estimator on $\bar{\mathbb{X}}$ across previous sections. The result of Chen, Wan and Zhou (2015) implied the asymptotic variance of $\hat{Q}_\tau^{\bar{\mathbb{X}}}(y|\boldsymbol{x})$ has the form

$$\boldsymbol{V}_{\text{ipw}}(\bar{\mathbb{X}}) = \left(E f_{\varepsilon|\boldsymbol{x}}(0)\right)^{-2} E\left\{\frac{\left(I(\varepsilon<0)-\tau\right)^2}{\pi(\boldsymbol{x})}\right\} \le \max\left\{\tau^2, (1-\tau)^2\right\} \frac{E\pi^{-1}(\boldsymbol{x})}{\left(E f_{\varepsilon|\boldsymbol{x}}(0)\right)^2} := \overline{\boldsymbol{V}}_{\text{ipw}}(\bar{\mathbb{X}}),$$

where $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is the conditional density function of $\varepsilon = y - Q_\tau(y|\boldsymbol{x})$ given $\boldsymbol{x}$. The purpose is to select a subspace $\bar{\mathbb{X}}_S \subset \bar{\mathbb{X}}$ such that the estimator achieves the asymptotic variance to be controlled by $\overline{\boldsymbol{V}}_{\text{ipw}}(\bar{\mathbb{X}})$. Comparing to Crump et al. (2009), the unquantifiable $f_{\varepsilon|\boldsymbol{x}}(0)$ poses challenges during searching $\bar{\mathbb{X}}_S$. To conquer the problem, consider an "oracle" weighted estimator that is optimized by

$$\min \sum_{i=1}^n \frac{\omega_S(\bar{\boldsymbol{x}}_i)}{\sum_{i=1}^n \omega_S(\bar{\boldsymbol{x}}_i)} \hat{W}_i \rho_\tau\left(y_i - Q_\tau(y|\boldsymbol{x}_i)\right), \tag{3.7}$$

where $\omega_S(\bar{\boldsymbol{x}}) = \omega(\bar{\boldsymbol{x}}) I(\bar{\boldsymbol{x}} \in \bar{\mathbb{X}}_S) f_{\varepsilon|\boldsymbol{x}}^{-1}(0)$ for $\omega : \bar{\mathbb{X}} \to [0, +\infty)$. Since $f_{\varepsilon|\boldsymbol{x}}^{-1}(0)$ does not involve the information for optimizing, asymptotic variance of the estimator in (3.7) is upper-bounded by

$$\overline{\boldsymbol{V}}_\omega(\bar{\mathbb{X}}) = \max\left\{\tau^2, (1-\tau)^2\right\} \frac{E\left\{\omega_S^2(\bar{\boldsymbol{x}})\pi^{-1}(\boldsymbol{x})\right\}}{\left[E\left\{\omega(\bar{\boldsymbol{x}}) I(\bar{\boldsymbol{x}} \in \bar{\mathbb{X}}_S)\right\}\right]^2}.$$

It can be seen that $f_{\varepsilon|\boldsymbol{x}}(0)$ is canceled in the denominator. Consequently, above question is transformed into the minimization of $\overline{\boldsymbol{V}}_\omega(\bar{\mathbb{X}}_S)$ to trim subjects with small response probabilities. For the concern, we have the following result:

Corollary 1. *Suppose $f_{\varepsilon|\boldsymbol{x}}(0)$ is bounded away from 0 and infinity, $f_{\varepsilon|\boldsymbol{x}}(0)$ and $\omega(\bar{\boldsymbol{x}})$ are*

continuously differentiable. When $0 < |E\{\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\}| < \infty$ *and condition* (C.2) *holds, the optimal* $\bar{\mathbb{X}}_S$ *is equal to* $\bar{\mathbb{X}}$ *if*

$$\sup_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}} \frac{\omega(\bar{\boldsymbol{x}})}{\pi(\boldsymbol{x})f_{\varepsilon|\boldsymbol{x}}(0)} \le 2\frac{E\left\{\omega^2(\bar{\boldsymbol{x}})\pi^{-1}(\boldsymbol{x})f^{-2}_{\varepsilon|\boldsymbol{x}}(0)\right\}}{E\left\{\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\right\}}.$$

Otherwise, $\bar{\mathbb{X}}_S = \left\{\bar{\boldsymbol{x}} \in \bar{\mathbb{X}} \middle| \omega(\bar{\boldsymbol{x}})\pi^{-1}(\boldsymbol{x})f^{-1}_{\varepsilon|\boldsymbol{x}}(0) \le \gamma\right\}$ *with a positive* γ *solved as*

$$\gamma = 2\frac{E\left\{\omega^2(\bar{\boldsymbol{x}})\pi^{-1}(\boldsymbol{x})f^{-2}_{\varepsilon|\boldsymbol{x}}(0)|\omega(\bar{\boldsymbol{x}})\pi^{-1}(\boldsymbol{x})f^{-1}_{\varepsilon|\boldsymbol{x}}(0) < \gamma\right\}}{E\left\{\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)|\omega(\bar{\boldsymbol{x}})\pi^{-1}(\boldsymbol{x})f^{-1}_{\varepsilon|\boldsymbol{x}}(0) < \gamma\right\}}. \tag{3.8}$$

From the application perspective (3.8) is useless. Since $\omega(\bar{\boldsymbol{x}})$ acts merely as a connector via a nonnegative function of covariates, it can be replaced by $\omega(\bar{\boldsymbol{x}}) = f_{\varepsilon|\boldsymbol{x}}(0)$ to satisfy the condition of Corollary 1. This substitution also guides the expected estimator on $\bar{\mathbb{X}}_S$. Of this view, the optimal subspace is rewritten as $\bar{\mathbb{X}}_S = \{\boldsymbol{x} \in \bar{\mathbb{X}}|\hat{\pi}(\boldsymbol{x}) \ge \hat{\alpha}\}$ for $\hat{a} = 1/\hat{\gamma}$, where $\hat{\gamma}$ is calculated from (3.8) and is restricted on $[2, +\infty)$. The strategy of robust improvement is implemented by the following algorithm:

Algorithm 2: Robust adaptive model averaging procedure for CQF

Input: Covariates $\boldsymbol{x} = (\boldsymbol{u}^\top, \boldsymbol{z}^\top)^\top$. Outcome y. Response indicator r. Number of splits B. Split ratio c. Tuning parameters a_n, λ.

Output: Conditional quantile function $\hat{Q}^{\text{mix}}_\tau(y|\boldsymbol{x})$.

1 Evaluate the response mechanism by (3.6) with $\{(\boldsymbol{x}_i^\top, r_i)|\boldsymbol{x}_i \in \bar{\mathbb{X}}\}$, denote as $\hat{\pi}(\boldsymbol{x})$.

2 Select the smallest $\hat{\alpha}$ on $[0, 1/2]$ subject to

$$\alpha \ge 2\frac{\sum_{i:\boldsymbol{x}_i\in\bar{\mathbb{X}}}\left\{I(\hat{\pi}(\boldsymbol{x}_i) \ge \alpha)\hat{\pi}^{-1}(\boldsymbol{x}_i)\right\}}{\sum_{i:\boldsymbol{x}_i\in\bar{\mathbb{X}}} I(\hat{\pi}(\boldsymbol{x}_i) \ge \alpha)}.$$

The optimal subspace is estimated as $\bar{\mathbb{X}}_S = \{\boldsymbol{x} \in \bar{\mathbb{X}}|\hat{\pi}(\boldsymbol{x}) \ge \hat{\alpha}\}$.

3 Implement Algorithm 1 by replacing $\mathbb{N} = \{(\boldsymbol{x}_i^\top, y_i, r_i)^\top : i = 1, ..., n\}$ as $\mathbb{N} = \{(\boldsymbol{x}_i^\top, y_i, r_i)^\top : \bar{\boldsymbol{x}}_i \in \bar{\mathbb{X}}_S\}$.

4. Simulations

In this section, we implement several simulation studies to investigate the performance of proposed adaptive model averaging algorithms in Sections 2–3. All the simulation procedures are independently replicated by $R = 500$ times. The criterion to quantify the predicting accuracy is out-of-sample final prediction risk (FPR), which is defined as

$$\mathrm{FPR}(\tau) = \frac{1}{Rn_t}\sum_{r=1}^{R}\sum_{i=1}^{n_t}\rho_\tau\left(\dot{y}_i - \hat{Q}_{\tau,r}(y|\dot{\boldsymbol{x}}_i)\right),$$

where $\{(\dot{\boldsymbol{x}}_i^\top, \dot{y}_i)^\top : i = 1, ..., n_t\}$ is the set of out-of-sample observations, and $\hat{Q}_{\tau,r}(y|\boldsymbol{x})$ is the CQF estimator in the r-th replication. Through the section we unite $n_t = 100$.

4.1 Some basic comparisons

We use the following data generating process (DGP):

$$y = u_1 + (\tau - 0.5)u_3^2 + \boldsymbol{\beta}^\top \boldsymbol{z} + \exp(0.3u_2 - 0.5u_3)\varepsilon, \tag{4.9}$$

where $\varepsilon \sim N(0,1)$. The covariates $(u_1, u_2, u_3, \boldsymbol{z}^\top)^\top$ are generated from a multivariate normal distribution with mean $\mathbf{0}_{1\times(p+3)}$ and variance-covariance matrix $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{p+3}$. The coefficient $\beta_j = (-1)^j(3+\gamma)$ for $j = 1, ..., 5$ and $\beta_j = 0$ otherwise, where $\gamma \sim N(0,1)$ and is fixed across replications. A simple MAR mechanism is specified as $\pi(\boldsymbol{u};\boldsymbol{\theta}) = 1/(1+\exp(\theta_1+\theta_2u_1+\theta_3u_2))$, such that the missing rate can be adjusted by varying $\boldsymbol{\theta}$. The dimension of $\boldsymbol{z}$ is $p = n^2/8 - 55n/2+2000$, which implies $n = \{100, 200, 300\}$ for $p = \{500, 1500, 5000\}$, respectively. Besides, the quantile level is chosen across $\tau = \{0.05, 0.5, 0.75\}$.

According to (4.9) one can see $q_\tau(\boldsymbol{u}) = u_1 + (\tau - 0.5)u_3^2 + \exp(0.3u_2 - 0.5u_3)b_\tau$, where b_τ is the τ-th quantile of $N(0,1)$. Obviously $q_\tau(\boldsymbol{u})$ has a linear structure at $\tau = 0.5$ and has

a measurably nonlinear form otherwise. To process the method, we use the following two candidates for fitting $q_\tau(\boldsymbol{u})$:

1. Linear model (q_τ-LM): $q_\tau^{(1)}(\boldsymbol{u}) = \alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3$;

2. Nonparametric additive model (q_τ-NAM): $q_\tau^{(2)}(\boldsymbol{u}) = f_{1,\tau}(u_1) + f_{2,\tau}(u_2) + f_{3,\tau}(u_3)$.

The second candidate is approximated by b-spline with 5 degrees of freedom, which is performed by `R` package `bs`. We employ MLE for the true response mechanism, and preliminary estimators of $q_\tau^{(k)}(\boldsymbol{u})$'s are implemented by a true specification of the linear part $\boldsymbol{\beta}^\top \boldsymbol{z} = \sum_{j=1}^5 \beta_j z_j$ for simplicity.

We only investigate the efficiency of adaptive estimation. The results of variable screening are attached in Part C.1 of supplementary materials. To check the adaptation with different structures of clustering, we design the following schemes to construct candidate clusters of screened explanatory variables:

Scheme 1 (Non-nested clustering): For each $q_\tau^{(k)}(\boldsymbol{u})$ $(k = 1, 2)$, set $|\hat{\mathcal{M}}_\tau^{(k)}| = \lfloor 2n/(3\log(n)) \rfloor$ and the grid points $h = 5n^{-1}J$ with $J = 1, ..., \lceil |\hat{\mathcal{M}}_\tau^{(k)}|/5 \rceil$, where $\lceil x \rceil$, $\lfloor x \rfloor$ are round up and down of x respectively. The number of candidate clusters $M = |\hat{\mathcal{M}}_\tau^{(k)}|/(nh)$ if $|\hat{\mathcal{M}}_\tau^{(k)}|/(nh)$ is an integer, and $M = \lfloor |\hat{\mathcal{M}}_\tau^{(k)}|/(nh) \rfloor + 1$ otherwise. In the first case, each candidate model allocates nh screened explanatory variables. In the second case, there are nh predictors in the top $M-1$ candidates and the other $|\hat{\mathcal{M}}_\tau^{(k)}| - (M-1)nh$ parts in the last candidate. The linear parts of all submodels are disjoint.

Scheme 2 (Nested clustering): Let $|\hat{\mathcal{M}}_\tau^{(k)}|$, h, J and M be same as those in Scheme 1. If $|\hat{\mathcal{M}}_\tau^{(k)}|/(nh)$ is an integer, the m-th candidate model $(1 \le m \le M)$ allocates the top mnh survived explanatory variables. Otherwise, the m-th candidate model $(1 \le m \le M-1)$ contains the top mnh survived explanatory variables and the last model contains all variables.

The linear part of the latter candidate embraces all in the former.

Table 1: Out-of-sample FPR and its standard deviation (in parentheses) of CQF estimator via NON-NESTED candidate clusters. †

(n,ρ)	τ	SSIC			JMA			AMA		
		q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA
(100, 0)	0.05	4.56(1.11)	4.05(3.83)	4.05(3.83)	3.60(0.60)	3.48(4.41)	3.48(4.31)	4.06(0.88)	2.45(1.69)	**2.61**(**1.14**)
	0.5	2.09(0.66)	3.08(1.76)	2.53(1.52)	2.05(0.61)	2.76(1.43)	2.19(0.93)	2.02(0.63)	2.83(1.39)	**2.06**(**0.72**)
	0.75	2.83(0.75)	3.29(2.14)	3.23(2.14)	2.67(0.62)	2.97(1.62)	2.68(1.32)	2.75(0.71)	2.80(1.33)	**2.47**(**0.74**)
(100, 0.5)	0.05	3.74(0.91)	5.40(8.61)	5.40(8.61)	2.72(0.44)	4.58(8.25)	4.67(8.61)	3.25(0.79)	3.57(7.70)	**2.78**(**3.90**)
	0.5	1.88(0.29)	4.07(3.80)	2.79(1.93)	1.79(0.25)	3.04(2.71)	**2.13**(1.12)	1.82(0.28)	3.89(3.63)	**2.13**(**0.84**)
	0.75	2.36(0.41)	4.55(7.42)	4.49(7.40)	2.14(0.31)	3.50(4.52)	3.00(3.68)	2.27(0.37)	3.76(4.60)	**2.41**(**1.69**)
(200, 0)	0.05	4.02(0.76)	2.60(1.57)	2.60(1.57)	3.38(0.51)	2.23(1.50)	2.25(1.52)	3.69(0.63)	1.61(0.97)	**1.63**(**0.90**)
	0.5	1.28(0.63)	1.56(0.81)	1.31(0.72)	1.27(0.62)	1.53(0.77)	1.26(0.66)	1.24(0.60)	1.47(0.75)	**1.20**(**0.60**)
	0.75	2.13(0.73)	1.72(1.06)	1.67(1.05)	2.05(0.66)	1.65(0.93)	1.54(0.87)	2.09(0.72)	1.55(0.84)	**1.43**(**0.71**)
(200, 0.5)	0.05	3.28(0.63)	2.72(3.39)	2.72(3.39)	2.60(0.36)	2.43(3.32)	2.35(2.71)	2.89(0.48)	1.44(1.67)	**1.46**(**1.66**)
	0.5	1.75(0.22)	2.67(1.49)	2.09(0.95)	1.68(0.20)	2.32(0.99)	1.88(0.54)	1.69(0.21)	2.37(1.05)	**1.83**(**0.37**)
	0.75	2.12(0.35)	2.44(1.42)	2.43(1.42)	1.97(0.29)	2.19(1.16)	2.10(1.00)	2.07(0.33)	2.08(1.13)	**1.84**(**0.54**)

† The missing rate is about 32.42% by setting $\boldsymbol{\theta} = (-1, 1, 1)^\intercal$. In each row, the best output among "MA" columns is labeled in **bold**.

We make comparisons of Algorithm 1 (denoted as AMA) with: (i) Model averaging by smoothed Schwarz information criterion (denoted as SSIC):

$$\text{SSIC} = \frac{\exp(-0.5\text{SIC}_k)}{\sum_{k=1}^{2}\exp(-0.5\text{SIC}_k)},$$

where $\text{SIC}_k = 2n\log\{n^{-1}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{Q}_\tau^{(k,m)}(y|\boldsymbol{x}_i))\} + \text{df}_k\log(n)$ for the k-th candidate model with respect to the degree of freedom df_k; (ii) Jackknife model averaging (denoted as JMA) of Wang et al. (2023). In AMA, we set $\Lambda = \{n^{-1}, n^{-2/3}, n^{-1/2}, n^{-1/3}, 1\}$ to construct a rough

grid and $c = 1/2$, $a_n = n^{-1}$, $B = 10$. Results of each methodology are optimized by (2.5) across grid points.

Table 2: Out-of-sample FPR and its standard deviation (in parentheses) of CQF estimator via NESTED candidate clusters. †

(n, ρ)	τ	SSIC			JMA			AMA		
		q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA
(100, 0)	0.05	4.39(0.97)	4.15(5.02)	4.15(5.02)	3.46(0.54)	3.40(3.65)	3.41(3.58)	4.18(0.94)	2.75(2.44)	**2.84**(**1.60**)
	0.5	2.24(0.73)	3.33(2.17)	2.71(1.39)	2.02(0.61)	2.70(1.42)	2.13(**0.82**)	1.99(0.62)	2.91(2.17)	**2.07**(0.89)
	0.75	2.95(0.74)	3.44(2.13)	3.32(1.98)	2.66(0.61)	2.85(1.39)	2.61(1.07)	2.75(0.69)	2.91(1.65)	**2.46**(**0.80**)
(100, 0.5)	0.05	3.75(1.02)	5.50(9.99)	5.50(9.99)	2.70(0.45)	4.12(8.26)	4.10(8.17)	3.47(0.85)	3.40(6.19)	**2.84**(**2.60**)
	0.5	1.89(0.29)	3.90(3.61)	2.94(3.40)	1.77(0.25)	2.72(2.38)	**2.08**(1.30)	1.80(0.27)	4.12(4.30)	2.12(**1.06**)
	0.75	2.37(0.44)	4.64(6.33)	4.44(6.26)	2.12(0.35)	3.17(3.30)	2.72(2.19)	2.24(0.40)	4.33(7.50)	**2.31**(**0.96**)
(200, 0)	0.05	4.11(0.82)	2.91(2.37)	2.91(2.37)	3.39(0.54)	2.43(1.82)	2.42(1.79)	3.93(0.83)	1.74(1.14)	**1.76**(**1.13**)
	0.5	1.90(1.25)	2.24(1.32)	1.85(1.27)	1.25(0.61)	1.57(0.80)	1.27(0.65)	1.23(0.61)	1.48(0.77)	**1.22**(**0.62**)
	0.75	2.50(0.99)	2.21(1.31)	2.19(1.34)	2.02(0.67)	1.70(1.03)	1.57(0.88)	2.04(0.71)	1.54(0.77)	**1.45**(**0.69**)
(200, 0.5)	0.05	3.32(0.62)	2.81(2.69)	2.81(2.69)	2.63(0.38)	2.48(2.60)	2.49(2.60)	3.08(0.56)	1.77(2.30)	**1.69**(**1.64**)
	0.5	1.76(0.25)	2.78(1.35)	2.19(1.08)	1.64(0.20)	2.31(0.99)	1.89(0.60)	1.66(0.21)	2.35(0.95)	**1.84**(**0.45**)
	0.75	2.14(0.37)	2.66(1.82)	2.66(1.82)	1.94(0.30)	2.24(1.18)	2.11(1.09)	2.03(0.33)	2.15(0.18)	**1.89**(**0.65**)

† The missing rate is about 32.42% by setting $\boldsymbol{\theta} = (-1, 1, 1)^\intercal$. In each row, the best output among "MA" columns is labeled in **bold**.

Tables 1–2 summarize the results of CQF estimators. It is evident to see that FPR and standard deviation (SD) of AMA are uniformly smaller than others, with regardless of the construction for candidate clusters. This implies that our method is not constrained by a specific scheme and is workable on a varieties of candidates. Although q_τ-NAM provides a better approximation when τ deviates from the median, its FPR deteriorates significantly

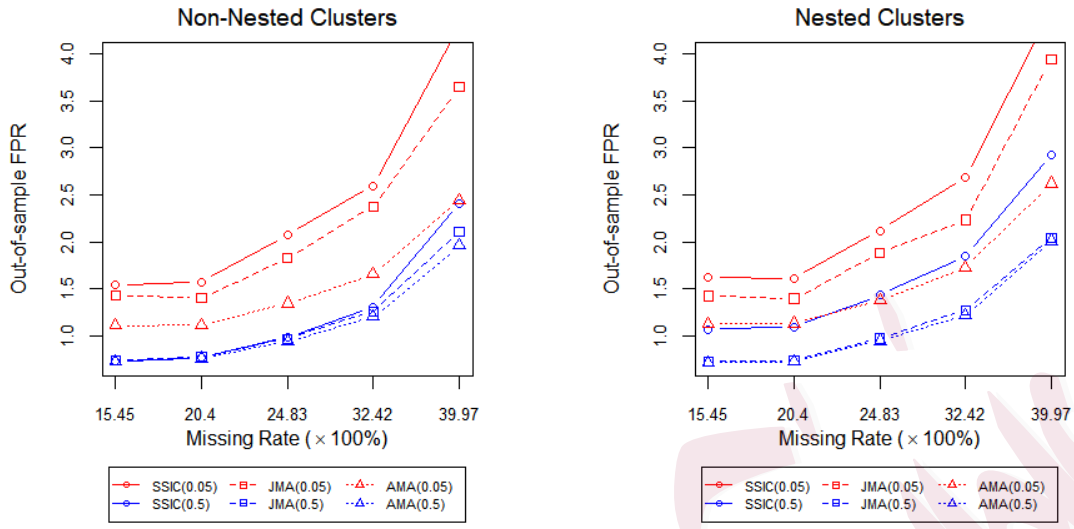

Figure 1: FPR of the MA estimator for CQF across a range of missing rates $(n = 200, \rho = 0)$.

compared to q_τ-LM under high correlation among z_j's. This can be explained by two aspects: Firstly, the correlation in (4.9) degrades the performance of variable screening, resulting in unsuitable candidates. Secondly, the efficiency of nonparametric estimation requires a large sample size, while it may be unsatisfied for only less than $0.7n$ pairs of available data. In this sense, SSIC and JMA naively incline weights to q_τ-NAM, while AMA adaptively compares the exponential loss between q_τ-NAM and q_τ-LM. Therefore, AMA exhibits data-adaptation analogous to multi-folds cross-validation due to the randomly splitting process (see Part C.1 of supplementary materials for more simulation results). It is beneficial to distinguish overfitting of the additive candidate ($\tau = 0.5$) as well.

Following one referee's suggestion, we compare three "MA" estimators under dynamic missing rates. Figure 1 reveals a consistent truth: AMA almost has the best predicting accuracy and the lowest increasing rate of FPR as the missing rate grows up. Particularly, AMA keeps the outperformance to aggregate misspecified models (i.e., $\tau = 0.05$). Above

findings further confirm the model-adaptation of our method.

4.2 Performance of over-consistency

In this subsection we organize another simulation setting to evaluate the over-consistency of adaptive weights in the sense of weakly including candidate set. To eliminate the convergence uncertainty of SIS for finite sample size, we employ the following model for generating observations:

$$y = u_1 + u_2 + 2(\tau - 0.5)u_3^2 + \sum_{j=1}^{3} \beta_j z_j + 0 \times z_4 + 0 \times z_5 + \exp(-0.5u_1)\varepsilon,$$

where $\varepsilon \sim N(0,1)$, $(u_1, ..., u_3, z_1, ..., z_5)^\top$ is sampled from the multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{I}_8$, $(\beta_1, \beta_2, \beta_3)^\top = (3,3,3)^\top$. Besides, we use the same response mechanism as in Section 4.1, and set $\boldsymbol{\theta} = (-1.5, 1, 1)^\top$ to cause around 25% nonresponse. Other settings in the adaptive algorithm remain unchanged.

We also use LM ($k = 1$) and NAM ($k = 2$) to fit $q_\tau(\boldsymbol{u})$ and the following digit sets to symbolize the choice of z_j ($j = 1, ..., 5$) in the linear part: $\mathcal{I}_{\tau,1} = \{1,2,3\}$; $\mathcal{I}_{\tau,2} = \{1,2,3,4,5\}$; $\mathcal{I}_{\tau,3} = \{2,4\}$; $\mathcal{I}_{\tau,4} = \{1,3,5\}$; $\mathcal{I}_{\tau,5} = \{3,4,5\}$, which constitute total 10 candidate models. It can be seen that $\mathcal{I}_{\tau,1}$, $\mathcal{I}_{\tau,2}$ contains all active components, and quasi-correct candidates of conditional quantile are $\{k = 1,2; m = 1,2\}$ with $\tau = 0.5$ and $\{k = 2; m = 1,2\}$ otherwise.

Figure 2 depicts the mean of out-of-sample FPR, mean of $\sum_{m\in\{1,2\}} \hat{\Omega}_{(k,m)}$ with respect to $k = 1,2$, and $\hat{\Omega}_{\text{cor}}$. Specifically, aggregated weights grow more rapidly when $\tau = 0.5$. This is because the number of quasi-correct models is double that of $\tau = 0.05$, and the linear specification has a faster convergence rate to potentially influence the convergence of $\hat{\Omega}_{\text{cor}}$. Note that when $\tau = 0.05$, the FPR of NAM is uniformly smaller than that of LM. This performs differently from Section 4.1, because the true model is strictly additive. Moreover,

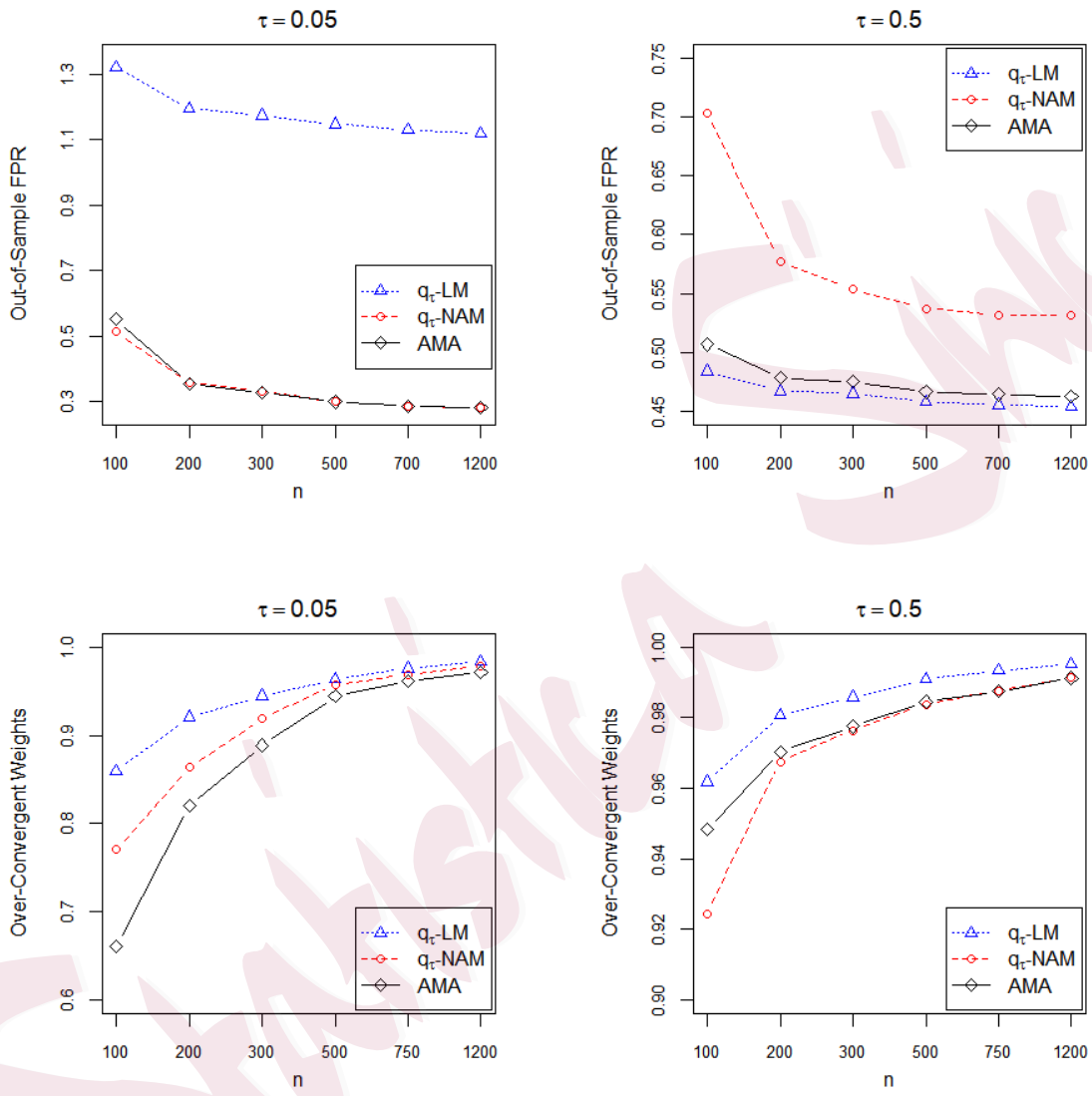

Figure 2: Out-of-sample FPR (the first row), sum of weights (the second row) of quasi-correct candidates with $\tau = 0.05$ and 0.5, respectively.

one can see an interesting phenomenon that when $\tau = 0.5$, FPR of AMA is higher than that of LM. Such finding is helpful for understanding (ii) of Theorem 3.

4.3 Robust analysis

In this subsection, we implement another simulation design to check the robustness of Algorithm 2. Set $(n, p) = (200, 2000)$, the quantile regressive model is employed as (4.9) but with distinct covariates: Given $(\widetilde{u}_1, ..., \widetilde{u}_3, \widetilde{z}_1, ..., \widetilde{z}_p)^\top$ from $N(\mathbf{0}_{p+3}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\rho^{|i-j|})_{p+3}$, we adjust that u_1 is generated by a student's t-distribution with df = 5, $z_1 = \widetilde{z}_1 \exp(0.3v_1)$ with $v_1 \sim N(0, 1)$, $z_{10} = \widetilde{z}_{10} v_2$ with v_2 generated from Pareto distribution with the shape parameter 2 and the scale parameter 0.1, and $(u_2, u_3) = (\widetilde{u}_2, \widetilde{u}_3)$, $z_j = \widetilde{z}_j$ for $j = 2, 3, ...9, 11, ..., p$ respectively. $q_\tau(\boldsymbol{u})$ and $\boldsymbol{\beta}$ are set as same as those in Section 4.1.

We specify the response model as $\text{logit}\{\pi(\boldsymbol{x})\} = 1 + 1.3u_1 + 0.8(z_1 - z_{10} + z_{20})$, which leads the missing rate at around 34%–35% across $\rho \in \{0, 0.5, 0.8\}$. Several competitors are considered: (i) G-JMA: group-lasso penalized likelihood estimation for $\pi(\boldsymbol{x})$ and jackknife model averaging for CQF; (ii) 5-CVMA: 5-folds cross-validation model averaging for both $\pi(\boldsymbol{x})$ and CQF. (iii) T-AMA: MLE for the true $\pi(\boldsymbol{x})$ and adaptive model averaging (Algorithm 1) for CQF. (iv) G-AMA: group-lasso penalized likelihood estimation for $\pi(\boldsymbol{x})$ and adaptive model averaging (Algorithm 1) for CQF. (v) J-AMA: jackknife model averaging for $\pi(\boldsymbol{x})$ and adaptive model averaging (Algorithm 1) for CQF. (vi) R-AMA: robust adaptive model averaging (Algorithm 2). (vii) trm-AMA: robust adaptive model averaging with the fixed threshold $\hat{\alpha} = 0.1$. To estimate $\pi(\boldsymbol{x})$, We adopt the prior screening procedure in Ando and Li (2017) to select covariates with their p-values smaller than 0.05 (denotes the number as $|\hat{\mathcal{N}}_\pi|$), and cluster total $L + 1$ non-nested candidates ($L = \lfloor |\hat{\mathcal{N}}_\pi|/d \rfloor$) such that the

first L models cooperate with d predictors and the last cooperates with remainders, where $d \in \{5, 10\}$ in others that to be optimized by the largest log-likelihood. The group-lasso is implemented by `R` package `grplasso` with BIC optimized regularization parameter prepared by `lambdamax()*0.5^(1:5)`. To estimate CQF, the candidate set is constructed as same as scheme 1 of Section 4.1. Nuisance parameters are set as $\Lambda = \{1, ..., 5\}$, $c = 0.8$, $a_n = n^{-1}$ and $B = 10$, respectively.

Table 3: Out-of-sample FPR and its standard deviation (in parentheses) of averaged estimators for CQF.†

ρ	τ	Methods						
		G-JMA	5-CVMA	T-AMA	G-AMA	J-AMA	R-AMA	trm-AMA
0	0.05	1.666(1.230)	1.707(0.688)	1.456(0.659)	1.419(1.110)	1.348(0.692)	**1.221**(**0.420**)	*1.291*(*0.597*)
	0.5	1.176(0.374)	1.204(0.365)	1.484(0.466)	1.144(*0.349*)	1.149(0.357)	**1.141**(**0.341**)	1.176(0.356)
	0.75	1.412(0.459)	1.476(0.433)	1.651(0.593)	1.345(0.396)	1.356(0.405)	**1.211**(**0.343**)	*1.318*(*0.364*)
0.5	0.05	1.277(0.886)	1.394(0.688)	1.188(0.603)	1.081(0.760)	1.095(0.753)	**0.926**(**0.339**)	*1.019*(*0.365*)
	0.5	1.427(0.328)	1.436(*0.319*)	1.583(0.399)	1.428(0.337)	1.438(0.361)	**1.405**(**0.314**)	*1.411*(*0.319*)
	0.75	1.445(0.559)	1.462(*0.347*)	1.606(0.449)	1.403(0.508)	1.422(0.469)	**1.306**(**0.296**)	*1.385*(0.436)
0.8	0.05	0.937(0.566)	1.002(0.470)	0.821(0.474)	*0.744*(0.401)	0.745(0.431)	**0.702**(**0.284**)	0.757(*0.372*)
	0.5	0.817(0.275)	**0.796**(0.278)	0.893(*0.270*)	0.813(0.274)	0.817(0.275)	0.820(**0.266**)	*0.800*(0.272)
	0.75	0.984(0.309)	1.010(0.288)	1.025(0.319)	0.952(0.295)	0.957(*0.284*)	**0.848**(**0.231**)	*0.935*(0.288)

† The best and the second best outputs are in **bold** and *italic*, respectively.

Table 3 reports predicting FPR and SD of weighted estimators. One can see that most of them exhibit stable performance in median, mainly because the true $q_\tau(\boldsymbol{u})$ is linear and mitigates the impact of extreme propensity scores. However, when an additive structure operates

as the "best approximation", such issue is magnified and directly leads to outliers for the first five methods, including robust estimations for response probability (5-CVMA and T-AMA). Compared with the last 2 methods, trm-AMA is unstable (e.g., $(\tau, \rho) = (0.5, 0)$, $(0.75, 0.5)$ and $(0.05, 0.8)$) because the fixed threshold may insufficiently or excessively remove outliers. R-AMA overcomes the problem by adaptively selecting a trimming threshold. In summary, our estimation guarantees the robustness of CQF prediction on the entire sample space.

5. Real-data example

In this section, we apply the proposed method to salary data of 322 baseball hitters (the dataset is available at http://lib.stat.cmu.edu/datasets/). The dataset consists of the annual salary in 1987 (denoted as y) and 16 other fully-observed characteristics (denoted as $\boldsymbol{x}$): x_1: Times at bat (in 1986); x_2: Hits (in 1986); x_3: Home runs (in 1986); x_4: Runs (in 1986); x_5: Runs batted in (in 1986); x_6: Walks (in 1986); x_7: Years in major leagues until 1986; x_8: Times at bat (during their entire career up to 1986); x_9: Hits (up to 1986); x_{10}: Home runs (up to 1986); x_{11}: Runs (up to 1986); x_{12}: Runs batted in (up to 1986); x_{13}: Walks (up to 1986); x_{14}: Put-outs (in 1986); x_{15}: Assists (in 1986); x_{16}: Errors (in 1986). Besides, there are total 59 nonresponse of the annual salary. As pointed by Deng, Yang and Wang (2022), the dataset needs to be prepared as: (i) $x_{7+j} := \sqrt{x_{7+j} - x_j}$ for $j = 1, ..., 6$, and $x_j := \sqrt{x_j}$ otherwise; (ii) $y := \log(y)$. Moreover, we standardize all pre-processed covariates to mitigate affect of heterogeneous measurement scales across features. We randomly sample 100 pairs from the 263 observed data points for validation. Hence the size of training sample is $n_1 = 222$, the nonresponse rate of the training outcome is approximated 26.58%.

Note that the data originate from Sports Illustrated rather than an authoritative insti-

tution, and the dataset lacks observable characteristics to distinguish between respondents and non-respondents. It prevents us from empirically verifying whether the missingness is MAR or MNAR. We consider the linear logistic model to fit the mechanism, that is, $\text{logit}\{\pi(\boldsymbol{x})\} = \theta_0 + \boldsymbol{\theta}^\top \boldsymbol{x}$ for MAR, and $\text{logit}\{\pi(\boldsymbol{x}, y)\} = \theta_0 + \boldsymbol{\theta}^\top \boldsymbol{x} + \gamma y$ for MNAR. To estimate $\pi(\boldsymbol{x})$, we sort components by their p-values in a decreasing order, and cluster total $L+1$ non-nested candidates ($L = \lfloor 16/d \rfloor$) such that the first L models cooperate with d predictors and the last cooperates with remainders, where $d \in \{1, ..., 5\}$ during each procedure (group-lasso, 5-CVMA, and robust JMA in Algorithm 2). To estimate $\pi(\boldsymbol{x}, y)$, we apply the robust CVMA procedure for semiparametric likelihood (Xiong, Deng and Wang (2025)), where $\boldsymbol{\omega} \in [-10, 10]^4$. The candidate set is constructed through the decreasing order of the distance-correlation SIS procedure (Li, Zhong and Zhu (2012)) between ry and x_j's, and contains total 4 non-nested models that the first three models cooperate with 5 predictors and the last cooperates with remainders. In this case, $E(\exp(\gamma y)|\boldsymbol{x}, r = 1)$ is evaluated by nonparametric sampling and sufficient dimension reduction.

We employ (1.1) to predict the conditional quantile of salary of hitters, where $q_\tau(u)$ is specified by either $q_\tau^{(1)}(u) = \alpha_\tau u$ and $q_\tau^{(2)}(u) = f_\tau(u)$ with $u = x_9$ that has the strongest correlation corresponding to y, and $\boldsymbol{z} = (x_1, ..., x_8, x_{10}, ..., x_{16})^\top$. We adopt the same clustering procedure in Section 4.1 to establish either non-nested or nested candidate set by $|\hat{\mathcal{M}}_\tau^{(k)}| = 15$, $h = 2n^{-1}J$ and $J = 1, ..., \lceil |\hat{\mathcal{M}}_\tau^{(k)}|/2 \rceil$. Other settings are same as in Section 4.3. The averaging procedures are employed by: (i) G-JMA; (iii) 5-CVMA; (ii) R-AMA; (iv) R-AMA (MNAR), where the last method is based on the MNAR response mechanism.

Tables 4–5 present the results over 100 replications. It is observed that (i) the spline fitting of $q_\tau(u)$ at $\tau = 0.75$ performs better than the linear model, whereas the latter performs

Table 4: Hitter salary data analysis: out-of-sample FPR and SD (in parentheses) via NON-NESTED clusters. †

τ	G-JMA			5-CVMA			R-AMA			R-AMA (MNAR)		
	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA
0.05	0.069	0.400	0.070	0.069	0.411	0.065	0.062	0.410	**0.062**	0.063	0.407	0.063
	(0.017)	(0.037)	(0.023)	(0.015)	(0.042)	(0.015)	(0.014)	(0.041)	(**0.015**)	(0.015)	(0.039)	(0.016)
0.25	0.158	0.324	0.157	0.157	0.326	0.154	0.154	0.324	**0.147**	0.155	0.323	0.148
	(0.017)	(0.029)	(0.017)	(0.016)	(0.029)	(**0.016**)	(0.016)	(0.028)	(0.017)	(0.016)	(0.028)	(**0.016**)
0.5	0.194	0.236	0.186	0.192	0.236	0.183	0.187	0.234	0.169	0.187	0.234	**0.168**
	(0.031)	(0.016)	(0.028)	(0.031)	(0.017)	(0.027)	(0.029)	(0.017)	(0.018)	(0.029)	(0.017)	(**0.017**)
0.75	0.181	0.159	0.155	0.180	0.157	0.153	0.173	0.155	0.138	0.172	0.155	**0.136**
	(0.047)	(0.012)	(0.035)	(0.048)	(0.012)	(0.034)	(0.043)	(0.012)	(**0.020**)	(0.043)	(0.012)	(**0.020**)
0.95	0.088	0.105	0.077	0.095	0.104	0.073	0.086	0.103	**0.065**	0.086	0.100	**0.065**
	(0.040)	(0.017)	(0.035)	(0.045)	(0.019)	(0.036)	(0.041)	(0.018)	(0.026)	(0.040)	(0.018)	(**0.023**)

† In each row, the best output among "MA" columns is labeled in **bold**.

better at other quantile levels. This suggests the historical number of hits may have a non-linear relationship corresponding to the annual salary. (ii) Model averaging across both nested and non-nested clusters shows negligible variation. (iii) FPR of R-AMA is lower than that of G-JMA and 5-CVMA, with increasing stability as τ grows up. (iv) The difference between two R-AMA methods is insignificant for a given τ, indicating that the annual salary prediction of athletes is unaffected by the assumed response mechanism. These results also confirm conclusions consistent with our simulations.

Table 5: Hitter salary data analysis: out-of-sample FPR and SD (in parentheses) via NEST-ED clusters. †

τ	G-JMA			5-CVMA			R-AMA			R-AMA (MNAR)		
	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA
0.05	0.066	0.418	0.066	0.062	0.410	0.063	0.059	0.407	0.058	0.059	0.403	**0.057**
	(0.014)	(0.038)	(0.014)	(0.009)	(0.042)	(**0.009**)	(0.010)	(0.041)	(0.011)	(0.008)	(0.039)	(0.010)
0.25	0.155	0.333	0.152	0.154	0.323	0.150	0.151	0.320	**0.146**	0.153	0.317	0.147
	(0.018)	(0.025)	(0.018)	(0.018)	(0.027)	(0.018)	(0.016)	(0.027)	(**0.017**)	(0.017)	(0.026)	(**0.017**)
0.5	0.199	0.242	0.191	0.197	0.237	0.187	0.195	0.235	0.178	0.194	0.234	**0.176**
	(0.030)	(0.017)	(0.027)	(0.028)	(0.018)	(0.027)	(0.027)	(0.018)	(**0.021**)	(0.027)	(0.017)	(**0.021**)
0.75	0.175	0.161	0.151	0.171	0.157	0.152	0.167	0.155	**0.136**	0.166	0.154	**0.136**
	(0.048)	(0.012)	(0.036)	(0.047)	(0.011)	(0.039)	(0.045)	(0.010)	(**0.022**)	(0.044)	(0.010)	(0.023)
0.95	0.093	0.111	0.077	0.090	0.105	0.079	0.080	0.103	**0.064**	0.082	0.100	**0.064**
	(0.049)	(0.018)	(0.036)	(0.048)	(0.016)	(0.039)	(0.040)	(0.015)	(0.025)	(0.042)	(0.015)	(**0.024**)

† In each row, the best output among "MA" columns is labeled in **bold**.

6. Conclusion

In the context of high-dimensional quantile regression, we propose an adaptive estimation to weight candidates with missing data, along with conducting robust inference. The adaptation is methodologically manifested through variance reduction via random partition of dataset, and is theoretically established by the oracle inequality. Large sample properties of our estimator reveal similarities to asymptotic results of CVMA as well. Numerical comparisons with existing model averaging estimation demonstrate the outperformance of our method in evaluating a variety of candidate model structures, which particularly exhibites robustness in the sense of high-dimensional response mechanism.

There remains several points to be clarified. Firstly, the robust inference is discussed under the MAR mechanism, it could be naturally generalized to MNAR frameworks, as demonstrated in our real-data analysis. Secondly, the discussion on the relationship between oracle inequality and asymptotic risk optimality depends on the loss function as well. When the adaptive procedure is implemented under a strongly convex loss, Theorem 3 can be established through a more general convergence rate. Thirdly, while the proposed method allows an exponential growing number of candidates, it brings the computational burden among massive models. Of this view, constructing an optimal cluster by desired variable importance measure is a considerable strategy to enhance the efficiency of weighting process. Besides, asymptotic properties are uniformly constructed under the consistent $\hat{\pi}(\boldsymbol{x}, y)$, the properties in the sense of misspecified mechanism need to be considered as well. Inference for above problems will become valuable in our further study.

Acknowledgements

The authors would like to thank the co-editor and anonymous referees for their constructive comments, which substantially improved the earlier version of the paper. We also appreciate Prof. Xinyu Zhang for suggestions on model averaging theory. Xiong's work is supported by National Natural Science Foundation of China (No.12401352), Postdoctoral Fellowship Program of CPSF (No.GZC20231022), and China Postdoctoral Science Foundation (No.2025T180847). Deng's work is supported by Natural Sciences and Engineering Research Council of Canada (NSERC). Wang's work is supported by National Natural Science Foundation of China (No.12271231, 12001229). The usual disclaimer applies.

References

Angrist, J., Chernozhukov, V., Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* **74**, 539–563.

Ando, T., Li, K. C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *Annals of Statistics* **45**, 2654–2679.

Austin, P. C., Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* **2017**, 1654–1670.

Bühlmann, P., van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Berlin: Springer-Verlag.

Busso, M., DiNardo, J., McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* **96**, 885–897.

Chen, Z., Liao, J., Xu, W., Yang, Y. (2023). Multifold cross-validation model averaging for generalized additive partial linear models. *Journal of Computational and Graphical Statistics* **32**, 1649–1659.

Chen, X., Wan A. T. K., Zhou, Y. (2015). Efficient quantile regression analysis With missing observations. *Journal of the American Statistical Association* **110**, 723–741.

Cook, R. D., Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association* **104**, 197–208.

Crump, R. K., Imbens, G. W., Mitnik, O. A., Hotz, V. J. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–199.

Deng, J., Yang, X., Wang, Q. (2022). Surrogate space based dimension reduction for nonignorable nonresponse. *Computational Statistics and Data Analysis* **168**, 107374.

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

Fan, J., Peng, H. (2003). Non-concave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* **38**, 3567–3604.

Fang, F., Yuan, C., Tian, W. (2023). An asymptotic theory for least squares model averaging with nested models. *Econometric Theory* **39**, 412–441.

Gu, Y., Zou, H. (2019). Aggregated expectile regression by exponential weighting. *Statistica Sinica* **29**, 671–692.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica* **75**, 1175–1189.

He, B., Ma, S., Zhang, X., Zhu, L. X. (2023). Rank-based greedy model averaging for high-dimensional survival data. *Journal of the American Statistical Association* **118**, 2658–2670.

He, X., Wang, L., Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics* **41**, 342–369.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–401.

Jiang, X., Liang, Y., Wang, H. (2024). Screen then select: a strategy for correlated predictors in high-dimensional quantile regression. *Statistics and Computing* **34**, 112.

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

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

Kong, Y., Li, Y., Zerom, D. (2019). Screening and selection for quantile regression using an alternative measure of

variable importance. *Journal of Multivariate Analysis* **173**, 435–455.

Li, W., Gu, Y., Liu, L. (2020). Demystifying a class of multiply robust estimators. *Biometrika* **107**, 919–933.

Li, R., Zhong, W., Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

Little, R. J., Rubin, D. B. (2002). *Statistical analysis with missing data.* Wiley: New York.

Lu, X., Su, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics* **188**, 40–58.

Ma, X., Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association* **115**, 1851–1860.

Ma, Y., Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179.

Mai, Q., Zou, H. (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–234.

Shan, K., Yang, Y. (2009). Combining regression quantile estimators. *Statistica Sinica* **19**, 1171–1191.

Vershynin, R. (2010). *Introduction to the non-asymptotic analysis of random matrices. arXiv preprint* arXiv:1011.3027v7.

Wang, B., Liang, H. (2023). Quantile regression of ultra-high dimensional partially linear varying-coefficient model with missing observations. *Acta Mathematica Sinica: English Series* **39**, 1701–1726.

Wang, L., Wu, Y., Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214–222.

Wang, M., You, K., Zhu, L., Zou, G. (2024). Robust model averaging approach by mallows-type criterion. *Biometrics* **80**, ujae128.

Wang, M., Zhang, X., Wan, A. T. K., You, K., Zou, G. (2023). Jackknife model averaging for high-dimensional

quantile regression. *Biometrics* **79**, 178–189.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Wu, Y., Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102**, 65–76.

Xie, J., Yan, X., Tang, N. (2021). A model-averaging method for high-dimensional regression with missing responses at random. *Statistica Sinica* **31**, 1005–1026.

Xiong, W., Deng, D., Wang, D. (2025). Semiparametric model averaging for high-dimensional quantile regression with nonignorable nonresponse. *arXiv preprint* arXiv:2509.00464.

Yang, Y. (2004). Combining Forecasting procedures: some theoretical results. *Econometric Theory* **20**, 176–222.

Ye, C., Yang, Y., Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association* **113**, 1797–1812.

Yu, D., Zhang, X., Liang, H. (2025). Unified optimal model averaging with a general loss function based on cross-validation. *Journal of the American Statistical Association*, 1–12.

Zeng, J., Hu, G., Cheng, W. (2024). A Mallows-type model averaging estimator for ridge regression with randomly right censored data. *Statistics and Computing* **34**, 159.

Zhang, X., Liu C. A. (2019). Inference after model averaging in linear regression models. *Econometric Theory* **35**, 816–841.

Zhang, X., Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. *Journal of Econometrics* **235**, 280–301.

Zhang, X., Zou, G., Liang, H., Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association* **115**, 972–984.

Zhao, P., Wang, L., Shao, J. (2020). Sufficient dimension reduction and instrument search for data with nonignorable

nonresponse. *Bernoulli* **2**, 930–945.

Wei Xiong

School of Mathematics and Statistics, Liaoning University, Shenyang, China

E-mail: (xiongwei16@mails.jlu.edu.cn)

Dianliang Deng

Department of Mathematics and Statistics, University of Regina, SK, Canada

E-mail: (deng@uregina.ca)

Wanying Zhang

School of Mathematics, Jilin University, Changchun, China

E-mail: (wyzhang20@mails.jlu.edu.cn)

Dehui Wang

School of Mathematics and Statistics, Liaoning University, Shenyang, China

E-mail: (wangdehui@lnu.edu.cn)