Contributions to Quantile and Superquantile Regression

by

Yuanzhi Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2022

Doctoral Committee:

Professor Xuming He, Chair
Professor Moulinath Banerjee
Professor Roderick J.A. Little
Assistant Professor Kean Ming Tan

Yuanzhi Li

yzli@umich.edu

ORCID iD: 0000-0002-5522-3864

ACKNOWLEDGEMENTS

First, my deepest gratitude goes to my advisor, Professor Xuming He, for his continuous support, encouragement, and guidance. During my five years at Michigan, Xuming has been a great academic advisor and role model to me. I have learned tremendously from him, not just about statistical wisdom, but also about being a responsible researcher in general. I will certainly miss our meetings and conversations.

In addition, I would like to thank Professors Moulinath Banerjee, Kean Ming Tan, and Roderick Little for serving on my dissertation committee. I am also grateful for all the help from the department staff and faculty members, who have made my experience at Michigan very smooth.

Lastly, I would like to thank my parents for their constant care and support overseas. I would also like to thank my partner, Zexi Li, for her love that makes the Ph.D. journey much more enjoyable.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Understanding the heterogeneous covariate-response relationship is central to modern data analysis. Beyond the usual descriptors such as the mean and variance, quantile and superquantile (also known as the expected shortfall or conditional value-at-risk) can capture the differential covariate effects on the upper or lower tails of the response distribution. This dissertation studies some fundamental aspects of the statistical inference of quantile and superquantile regression.

In the first part of the dissertation, we propose a novel approach to estimating the superquantile regression. Superquantile measures the average of a response given that it exceeds a certain quantile, and is widely used as a risk measure in financial and engineering applications to quantify the expected outcome in a given percentage of the worst-case scenarios. Most existing approaches for superquantile regression rely explicitly on the modeling of the conditional quantile functions. In this dissertation, we offer new insights into an optimization formulation for the superquantile in the recent literature, based on which we provide and validate a direct approach to superquantile regression estimation without relying on additional quantile regression modeling. Operationally, the approach can be well approximated by fitting a linear quantile regression to an array of pre-estimated conditional superquantile processes. With certain initial estimators based on binning of the covariate space, we show that the proposed superquantile regression estimator is consistent and asymptotically normal. This approach achieves implicit weighting of the data, which is found to be automatically adaptive to data heterogeneity and offers efficiency gain in various scenarios. Via theoretical and numerical comparisons show that the proposed

approach has competitive, and often superior, performance relative to other common approaches in the literature.

In the second part of the dissertation, we study pseudo-Bayesian inference for possibly sparse quantile regression models. We find that by coupling the asymmetric Laplace working likelihood with appropriate shrinkage priors, we can deliver pseudo-Bayesian inference that adapts automatically to the possible sparsity in quantile regression analysis. After a suitable adjustment on the posterior variance, the proposed method provides asymptotically valid inference under heterogeneity. Furthermore, the proposed approach leads to oracle asymptotic efficiency for the active (nonzero) quantile regression coefficients and super-efficiency for the non-active ones. We also discuss the theoretical extension when the covariate dimension increases with the sample size at a controlled rate. By avoiding the need to pursue dichotomous variable selection as well as nuisance parameter estimation, the Bayesian computational framework demonstrates desirable inferential stability.

CHAPTER I

Introduction and Preliminaries

1.1 Superquantile, from one-sample to regression

Superquantile (SQ), also known as the conditional value-at-risk (CVaR), or the expected shortfall (ES), measures the conditional mean of an outcome above certain quantile level. Specifically, for a random variable Y, its τ-th ($0 < \tau < 1$) quantile and SQ are defined as

$$q(\tau) = \inf\{u : \Pr(Y \leq u) \geq \tau\}, \qquad v(\tau) = \mathrm{E}[Y \mid Y \geq q(\tau)], \tag{1.1}$$

respectively. If the distribution of Y is continuous, then the SQ can be expressed as the following alternative form:

$$v(\tau) = \frac{1}{1-\tau} \int_{\tau}^{1} q(\alpha)\, \mathrm{d}\alpha. \tag{1.2}$$

In this dissertation, we focus on the case with continuous outcomes, hence we use the definitions in (1.1) and (1.2) interchangeably. While we focus on the upper-tail average, our discussion easily applies to the lower-superquantile, which is more commonly used in the literature (*Artzner et al.*, 1999; *Acerbi and Tasche*, 2002).

Superquantile plays an important role in a wide range of applications. In par-

ticular, SQ is a popular risk measure in financial applications to quantify the loss in extreme cases. Replacing the quantile, *Basel Committee on Banking Supervision* (2013) has made superquantile the official metric for market risk capital requirements. Such transition has provoked the recent development of novel methods for estimating, forecasting, and backtesting the SQ in the finance industry (*Nolde and Ziegel*, 2017; *Bercu et al.*, 2021; *Deng and Qiu*, 2021). Beyond financial applications, superquantile is also useful in other disciplines such as supply chain management (*Soleimani and Govindan*, 2014), treatment effect detection (*He et al.*, 2010; *Chen and Yen*, 2021), robust machine learning (*Laguel et al.*, 2021a,b), as well as quality control engineering (*Rockafellar and Royset*, 2010).

Compared to the quantile, the SQ has two distinct advantages (*Emmer et al.*, 2015; *Yamai and Yoshiba*, 2005). First, it provides a more informative summary of the upper tail in a distribution. Focusing only on the τ-th quantile overlooks the extreme loss that might occur beyond that quantile. On the other hand, the τ-th SQ quantifies the expected loss in the worst $100\tau\%$ scenarios; it also follows from (1.2) that the SQ takes the entire tail distribution into account. Second, the superquantile is a coherent measure of risk in the sense of *Artzner et al.* (1999), while the quantile is not. Specifically, the SQ satisfies the celebrated *sub-additivity* property of (*Acerbi and Tasche*, 2002), i.e., for two random variables X and Y with finite expectations, we must have

$$v_{X+Y}(\tau) \leq v_X(\tau) + v_Y(\tau).$$

Such *sub-additivity* echoes the principle of diversification in financial portfolio management (*Koumou*, 2020). On the contrary, the lacking of *sub-additivity* makes the quantile less suitable for financial risk management.

From a statistical perspective, the estimation and inference of SQ in one-sample problems are well-understood. Following (1.1), a simple SQ estimator would be the average of observed data that exceeds the sample quantile. The asymptotic properties

of this empirical estimator can be established via L-statistics theory (*Van der Vaart*, 2000, Chapter 22). In particular, *Chen* (2007) shows that the empirical SQ estimator is asymptotically efficient among a class of kernel-smoothed approaches (*Scaillet*, 2004). More recently, *Zwingmann and Holzmann* (2016) gives a general asymptotic analysis for the empirical SQ estimator under relaxed conditions. There are also several other model-based estimation methods in the literature; See *Nadarajah et al.* (2014) for a review. As for inference for the SQ, the bootstrap method in *Sun and Cheng* (2018) can be helpful.

In this dissertation, we focus on regression modeling of the superquantile. When auxiliary information is available, it is often valuable to study the conditional SQ of Y given a set of predictors X. In parallel to (1.1), we formally define the conditional SQ as:

$$v_{Y|X}(\tau, x) = \frac{1}{1-\tau} \int_{\tau}^{1} q_{Y|X}(\alpha, x)\, \mathrm{d}\alpha = \mathrm{E}\left[Y \mid Y > q_{Y|X}(\tau, x), X = x\right], \tag{1.3}$$

where $q_{Y|X}(\tau, x)$ is the τth quantile of Y given $X = x$. We sometimes omit the subscript $Y \mid X$ in the notation for conditional SQ if there is no confusion. We consider the following linear SQ regression model:

$$v(\tau, x) = x^T \beta \tag{1.4}$$

where β is the SQ regression coefficient. Compared with quantile regression modeling, the SQ regression model (1.4) can better capture the heterogeneous covariate effect in the tail of the response distribution. The focus for the first part of this dissertation is the estimation of β under Model (1.4).

While the formulation of the SQ regression is straightforward, valid estimation under Model (1.4) is non-trivial. The key difficulty is that SQ is *unelicitable* (*Gneiting*, 2011), in the sense that the SQ can not be formulated as the solution to an

M-estimation problem. Even in the one-sample case, *Gneiting* (2011) shows that there does not exist a function ψ such that

$$v(\tau) \in \arg\min_{\theta} \mathrm{E}\,\psi(Y-\theta),$$

for a broad enough class of Y and a given τ. Many classic regression methods, e.g., the least-squares or quantile regression, relies on such *elicitability*. Therefore, estimation of Model (1.4) cannot be achieved via the M-estimation framework, and relatively little has been available in the literature. We review some related methods in the following.

Most approaches in the literature require more modeling assumptions beyond Model (1.4), under which the estimation of SQ regression is more tangible. Assuming a homoscedastic linear model, *Chun et al.* (2012) uses a calibrated composite quantile regression for SQ regression. When the conditional quantile functions are assumed linear at all quantile levels, *Peracchi and Tanase* (2008) proposes to average the estimated quantile regression over a range of quantile levels based on the formula in (1.2) and (1.3). More recently, *Fissler and Ziegel* (2016) recognizes that quantile and SQ are jointly *elicitable* as a pair; Therefore, when both the τ-th quantile and SQ are linear, *Dimitriadis and Bayer* (2019) and *Patton et al.* (2019) develop a joint regression framework that estimates the quantile and SQ regression simultaneously. However, the resulting optimization problem is non-smooth and non-convex. Using the same joint regression model, *Barendse* (2020) and *Peng* (2022) propose two-step procedures that estimate the quantile and SQ regression sequentially, which may improve the computational stability. Importantly, all those approaches rely explicitly on a parametric quantile regression model, in addition to the linear SQ model (1.4).

In the Operations Research literature, *Rockafellar et al.* (2014) and *Rockafellar and Royset* (2018) propose a different superquantile-oriented regression approach (named

RRM hereafter) that does not require quantile regression modeling. However, we demonstrate in the next section that their approach does not deliver the correct SQ regression coefficient in general. Their approach is based on the minimization of a new convex loss function, where the function cannot be written as the expectation of a random function. Though the loss function is valid in the one-sample case without covariates, their regression approach does not work as intended from the statistical perspective. Even on the population level, we show that the minimizer of the RRM loss function may not be the SQ regression coefficient. Therefore, while the RRM approach provides a valuable framework, it requires further study to fully understand its statistical validity.

We hasten to add that non-parametric estimation of the conditional SQ is widely available in the literature. For example, *Cai and Wang* (2008) and *Kato* (2012) consider kernel-based approaches that are generalizations from the one-sample case. *Xiao* (2014) proposes another approach based on the connection between SQ and the check-loss function in quantile regression; and *Martins-Filho et al.* (2018) uses an approach based on the extreme value theory. More recently, *Olma* (2021) considers local-linear estimation based on Neyman-orthogonalized score functions. Another line of work considers aggregating non-parametric quantile regression estimators over a range of quantile levels (*Peracchi and Tanase*, 2008; *Leorato et al.*, 2012). Nonetheless, all these non-parametric SQ regression methods can be less efficient and less interpretable. In this dissertation, we focus on the parametric SQ modeling based on Model (1.4).

Recently, *Chetverikov et al.* (2022) proposes a new approach for average quantile regression estimation, which covers the SQ regression as an example. Their approach is based on integrating non-parametric estimators of the conditional distribution function. By using the idea of debiased machine learning (*Chernozhukov et al.*, 2018), they do not explicitly require a linear quantile regression model.

In this dissertation, we develop new approaches for estimating the linear SQ regression model (1.4). In the remainder of this chapter, we give a more detailed review of the RRM approach and demonstrate its inconsistency for the SQ regression problem. We further give a modified RRM loss function, which we name the *m-Rock* loss function, that correctly identifies the SQ regression coefficients on the population level. Such a modification is critical to its validity in SQ regression.

In Chapter 2, we explore new approaches for SQ regression under the simple yet illustrative scenario with discrete covariates. In particular, we show how the m-Rock loss function can lead to a practical method for SQ regression. We also study two other intuitive methods in the case with discrete covariates. Via practical and asymptotic comparisons, we find that the m-Rock approach is superior to the other two approaches.

In Chapter 3, we focus on the m-Rock approach and seek its extension to the case with general covariates. We give a theoretical analysis of the m-Rock approach based on binning, thereby effectively discretizing the covariate space. Following our analysis, we uncover the principle of the m-Rock approach: First, it needs a set of non-parametric SQ estimators at different quantile levels; Second, it linearizes those initial estimators in an efficient way. We also show that a Neyman-orthogonalized local-linear estimator can be used as an example of the initial SQ estimator, and we demonstrate the merit of the resulting estimator via asymptotic efficiency comparisons. Importantly, the m-Rock approach does not rely on a linear quantile regression model.

In Chapter 4, we discuss the practical applicability of the m-Rock approach. We give a prototype implementation of the m-Rock approach, followed by numerical experiments to demonstrate its performance. We further illustrate the use of the m-Rock approach in two empirical applications related to finance and public health.

1.2 RRM regression revisited

In this section, we review the RRM approach in *Rockafellar et al.* (2014), and we use a toy example to demonstrate that it does not deliver the true SQ regression coefficient in general. We present both analytical and numerical evidence.

1.2.1 Original formulation

Rockafellar et al. (2014) proposed a superquantile-oriented regression approach that can be transformed into a convex optimization problem. Key to the approach is a novel loss function induced by the following RRM formula. Denote by $v_{[Y]}(\alpha)$ as the α-th superquantile function of Y, Theorem 1 of *Rockafellar et al.* (2014) shows that

$$\begin{aligned} v_{[Y]}(\tau) &= \arg\min_{C} \left\{ C + \frac{1}{1-\tau} \int_0^1 \max\{0, v_{[Y-C]}(\alpha)\} \, \mathrm{d}\alpha \right\} \\ &\triangleq \arg\min_{C} \ \mathcal{L}_\tau(C), \end{aligned} \tag{1.5}$$

and that $\mathcal{L}_\tau(C)$ is a convex function of C; we sometimes omit the index τ and write $\mathcal{L}(C)$. Equation (1.5) is a population level formula since it relies on the true yet unknown superquantile function of Y, and direct calculation/optimization of $\mathcal{L}(C)$ is infeasible. However, *Rockafellar et al.* (2014) and *Rockafellar and Royset* (2018) show that the empirical RRM problem based on the observed data can be transformed into an alternate form, which can then be solved without knowing the superquantile of Y in advance. Therefore, in the one-sample case without covariates, the RRM formula provides a valuable alternative to approximating the superquantile, and is useful for applications that involve large-scale optimization of the SQ (*Xu et al.*, 2016; *Rockafellar and Royset*, 2018).

As a direct extension, *Rockafellar et al.* (2014, Section 3.1) proposes a regression approach by exploiting the same loss function. Targeting the τ-th SQ of Y given X,

the RRM regression solves:

$$\min_{\theta_1,\theta_0}\left\{\theta_0 + \mathrm{E}X^T\theta_1 + \frac{1}{1-\tau}\int_0^1 \max\{0, v_{[Y-\theta_0-X^T\theta_1]}(\alpha)\}\,\mathrm{d}\alpha\right\}, \tag{1.6}$$

where θ_1 is the slope and θ_0 is the intercept; Note θ_1 is generally a vector. It is important that Equation (1.6) relies only on the marginal superquantile of $Y - \theta_0 - X^T\theta_1$, not the conditional superquantile. In parallel to (1.5), the RRM regression problem (1.6) can be solved via convex optimization algorithms (*Rockafellar and Royset*, 2018). Therefore, the RRM approach offers a computationally efficient regression technique for superquantile-based modeling, and has been used in many applications ever since (*Xu et al.*, 2016; *Laguel et al.*, 2021b).

To facilitate subsequent analysis, we review some additional results regarding the optimization problem (1.6). Proposition 3 of *Rockafellar et al.* (2014) shows that solving (1.6) is equivalent to the following two-step procedure:

$$\theta_1^* \leftarrow \arg\min_{\theta_1}\left\{\mathrm{E}X^T\theta_1 + \frac{1}{1-\tau}\int_\tau^1 v_{[Y-X^T\theta_1]}(\alpha)\,\mathrm{d}\alpha\right\}, \tag{1.7}$$

$$\theta_0^* \leftarrow v_{[Y-X^T\theta_1^*]}(\tau), \tag{1.8}$$

where θ_0^* and θ_1^* are the population-level minimizers, i.e., the estimands for the RRM approach. In what follows, we shall work with Equations (1.7) and (1.8), instead of the original formulation in (1.6). Note, we use the notations θ_0^* and θ_1^* to emphasize that they may be different than the SQ regression coefficient in Model (1.4).

1.2.2 A counter-example

Here we illustrate that θ_0^* and θ_1^* from the RRM approach do not coincide with the true SQ regression coefficients under Model (1.4). Consider the following data generating model:

$$Y = 1 + X\,\varepsilon, \tag{1.9}$$

where $X \sim \Gamma(2,1)$ with $\mathrm{E}X = 2$, and the error term $\varepsilon \sim U(-1,1)$ independent of X. We aim to estimate the 50% SQ regression, where the true SQ regression coefficients are $\beta_0 = 1$, $\beta_1 = 0.5$.

We find the minimizer to the population level RRM loss function (1.7) with $\tau = 0.5$, where we label the loss function as $L_1(\theta_1)$. To this end, we compute the analytical expression for the marginal superquantile of $Z(\theta_1) = (Y - \theta_1 X)$ for any θ_1 and any quantile level. For all $\theta_1 \in (-1,1)$, $Z(\theta_1)$ follows tilted double exponential distribution with density function:

$$f_{Z(\theta_1)}(z;\theta_1) = \begin{cases} \frac{1}{2}\exp\left\{\frac{z}{1+\theta_1}\right\}, & z < 0, \\ \\ \frac{1}{2}\exp\left\{\frac{-z}{1-\theta_1}\right\}, & z \geq 0. \end{cases}$$

Straightforward probabilistic calculation shows that the marginal super quantile of $Z(\theta_1) = Y - \theta X$ is:

$$v_{[Z(\theta_1)]}(\alpha) = \begin{cases} 1 + \frac{1}{1-\alpha}\left(\alpha(1+\theta_1)(1 - \log\left[\frac{2\alpha}{1+\theta_1}\right]) - 2\theta_1\right), & 0 \leq \alpha \leq \frac{1+\theta_1}{2}, \\ \\ 1 + (\theta_1 - 1)\left(\log\left[\frac{2(1-\alpha)}{1-\theta_1}\right] - 1\right), & \frac{1+\theta_1}{2} < \alpha < 1. \end{cases} \tag{1.10}$$

Substituting Equation (1.10) into the the RRM loss function (1.7), we can obtain an analytical expression for the RRM loss function $L_1(\theta_1)$. Moreover, we can compute

the first-order derivative to the loss function $L_1(\theta_1)$ as:

$$\begin{aligned}\frac{\partial L_1(\theta_1)}{\partial \theta_1} &= EX + \frac{1}{1-1/2}\int_{1/2}^{1} \frac{\partial v_{[Y-X^T\theta_1]}(\alpha)}{\partial \theta_1}\, d\alpha \\ &= \begin{cases} -1 - \log(1-\theta_1), & -1 \le \theta_1 \le 0, \\ 2\left\{-\frac{1}{2} - \mathrm{Li}_2(\frac{1}{2}) + \mathrm{Li}_2(\frac{\theta_1+1}{2}) + (\frac{1}{2} - \log 2)\log(1+\theta_1)\right\}, & 0 < \theta_1 \le 1, \end{cases}\end{aligned}$$

where $\mathrm{Li}_2(x) = -\int_0^x \log(1-z)/z\, dz$.

Figure 1.1 below shows the RRM loss function in (1.7) and its derivative under Model (1.9). Since the RRM loss function is convex and differentiable (*Rockafellar et al.*, 2008; *Rockafellar and Uryasev*, 2013), we can use a first-order method, e.g., the Newton-Raphson method, to solve the minimization problem (1.7). We use the convex optimization toolbox in MATLAB and obtain the population-level minimizer as $\theta_1^* = 0.7041$, marked by the red line in Figure 1.1; The true SQ regression coefficient $\beta_1 = 0.5$ is marked by the blue line. Note we focus on the population level loss function, therefore the clear discrepancy between θ_1^* and β_1 shows the RRM approach fails to give the targeted coefficients for the superquantile regression. [1]

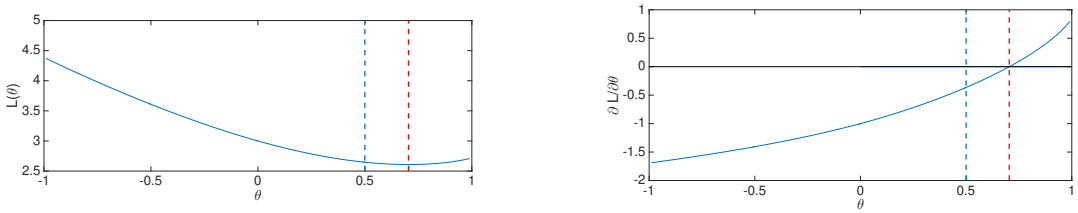


Figure 1.1: The population level loss function $L_1(\theta_1)$ (left panel) and its derivative (right panel). The blue dashed line marks the true SQ regression coefficient β_1, while the red one marks the minimizer of $L_1(\theta_1)$.

We further demonstrate the inconsistency of the RRM approach by a numerical

[1] Although the foregoing derivation for $L_1(\theta_1)$ is only valid for $\theta \in (0,1)$, it does not affect our conclusion. This is due to the global convexity of the RRM loss function (*Rockafellar et al.*, 2008): A local minimizer within $[-1,1]$ must also be the global minimizer.

experiment. We generate 200 Monte Carlo datasets from Model (1.9), and we consider sample sizes at $n = 100$ or $n = 1000$. Setting $\tau = 0.5$, Figure 1.2 shows the histogram of the estimated slope term $\hat{\theta}_1$ among the 200 Monte Carlo datasets. The empirical RRM problem is solved by the numerical integration method in Section 5.2 of *Rockafellar et al.* (2014) with 100 grid points. We can see the histograms are clearly concentrating toward $\theta_1^* = 0.704$, instead of the true SQ regression coefficient $\beta_1 = 0.5$.

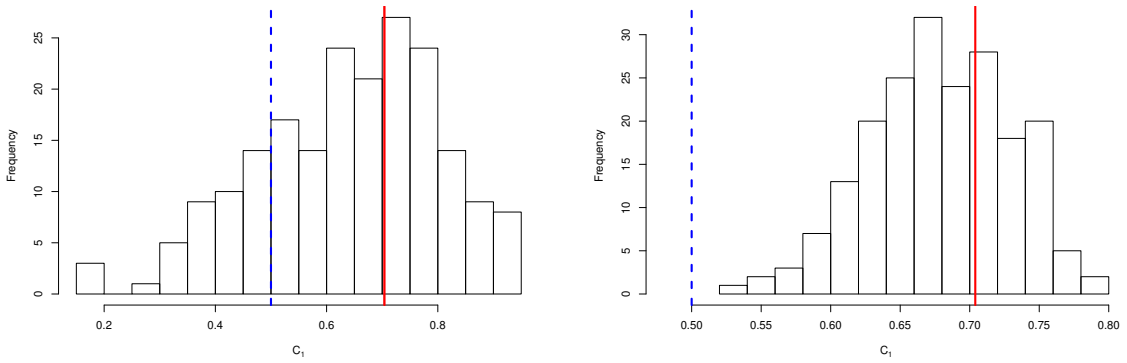


Figure 1.2: The empirical distribution of the RRM estimator $\widehat{\theta}_1$ at sample size $n = 100$ (left) and $n = 1000$ (right). The blue line marks the true SQ regression coefficient $\beta_1 = 0.5$, while the red one marks the RRM estimand $\theta_1^* = 0.704$.

Although the RRM approach is not valid for superquantile regression, it can still be valuable as a generalized regression technique. As shown in *Rockafellar and Uryasev* (2013), the RRM approach finds the best linear approximation to the response Y using the covariates X, in the sense that the residuals minimize the superquantile-oriented loss function (1.6). Furthermore, *Rockafellar and Royset* (2018) shows that the RRM approach is consistent for the SQ regression in homoscedastic linear models; and *Golodnikov et al.* (2019) shows that the RRM approach is equivalent to a composite quantile regression under certain scenarios. Therefore, the RRM approach can be useful for risk tuning and optimization that incorporates covariate information (*Miranda*, 2014).

1.3 A modified loss function

In this section, we provide a corrected RRM loss function based on (1.5) and (1.6) that can be used for superquantile regression. Here we stay on the population level, where the modified loss function depends on unknown distributional quantities and is not directly useful for empirical estimation; We shall discuss the implementation and finite-sample properties of the modified loss function in Chapter 2. The proofs of the results in this section are relegated to Section 2.5.3 in Chapter 2.

1.3.1 The one-sample RRM formula revisited

Before extending to the regression case, we first investigate the RRM formula (1.5) in the one-sample case. The mathematical correctness of the formula follows from Theorem 1 in *Rockafellar et al.* (2014). We provide some statistical insight behind the RRM formula in the following Proposition.

Proposition 1. *The loss function $\mathcal{L}_\tau$ in (1.5) can be written as*

$$\mathcal{L}_\tau(C) - \int_0^1 v_{[Y]}(\alpha)\,\mathrm{d}\alpha = \frac{1}{1-\tau}\int_0^1 \rho_\tau\left(v_{[Y]}(\alpha) - C\right)\,\mathrm{d}\alpha = \frac{1}{1-\tau}\,\mathrm{E}_\xi\left[\rho_\tau\left(g(\xi) - C\right)\right], \tag{1.11}$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}[u \le 0])$ is the check-loss function, $g(z) = v_{[Y]}(z)$, and ξ follows a uniform distribution on $(0,1)$.

By recasting the original RRM loss function into an alternate form, it is now easier to see why the one-sample RRM formula (1.5) is correct. By the property of the check-loss function $\rho_\tau(\cdot)$, minimizing $\mathcal{L}(C)$ finds the τ-th quantile of the random variable $g(\xi)$; See e.g., *Koenker* (2005, Section 2). Since the function g is monotonically increasing, it is apparent that the τ-th quantile of $g(\xi)$, also the minimizer of $\mathcal{L}_\tau(C)$, is simply $g(\tau) = v_{[Y]}(\tau)$.

In fact, we can think of the RRM formula as finding the τth quantile of the

superquantile process. For a sufficiently fine grid $\alpha_1 < \ldots < \alpha_J$ that spans the interval $(0, 1)$, finding the τth sample quantile of the set $\{v_{[Y]}(\alpha_1), \ldots, v_{[Y]}(\alpha_J)\}$ would approximately recover the targeted τth SQ $v_{[Y]}(\tau)$. Therefore, the one-sample RRM formula works by leveraging the *monotonicity* of the superquantile process $v_{[Y]}(\alpha)$ over $\alpha \in (0, 1)$.

1.3.2 A suitable loss function for the regression setting

While the RRM formula (1.5) is correct in the one-sample case, its direct extension as in (1.6) is not valid for the SQ regression setting. The main reason is as follows. The loss function (1.6) involves only the marginal SQ of the residual, but not the conditional SQ given the covariates. Moreover, the marginal SQ in the RRM loss function cannot be directly connected with the conditional SQ since the law of total expectation does not apply for (1.6).[2] Therefore, the RRM loss function (1.6) is not suitable for modeling the conditional SQ of Y given X.

To properly model the conditional SQ, we propose the following population-level loss function in place of (1.6). In the following, we shall write $v_{[Z|X]}(\alpha, x)$ as the conditional α-th SQ of Z given $X = x$. Given the covariate vector X (which includes an intercept term) and the response Y, we define the modified RRM function as:

$$L(\theta) = \mathrm{E} X^T\theta + \frac{1}{1-\tau}\int_0^1 \mathrm{E}\,\max\{0,\, v_{[Y-X^T\theta|X]}(\alpha, X)\}\,\mathrm{d}\alpha, \tag{1.12}$$

which simply substitutes the conditional SQ function for the marginal SQ in (1.6); the expectation in (1.12) is for X only. We name (1.12) as the modified Rockafellar (*m-Rock*) loss function hereafter. Though the modification from (1.6) to (1.12) seems straightforward, it is a substantial step to ensure identification of the SQ regression coefficients.

[2]For a classic M-estimation problem with loss function $\ell(\cdot)$, it follows that $\mathrm{E}[\ell(Y - X^T\theta)] = \mathrm{E}\left\{\mathrm{E}[\ell(Y - x^T\theta)] \mid X = x\right\}$. However the RRM loss function cannot be written as an expectation.

Conditioning on $X = x$, the m-Rock loss function reduces to the one-sample RRM loss function in (1.5); it is then intuitive that the m-Rock loss function is minimized at the conditional SQ The result below formally establishes the validity of (1.12) in SQ regression.

Theorem I.1. *There exists a constant C_0, such that the population level m-Rock loss function $L(\theta)$ can be written as:*

$$\begin{aligned} L(\theta) &= C_0 + \frac{1}{1-\tau}\mathrm{E}_X\left[\int_0^1 \rho_\tau\left(v_{[Y|X]}(\alpha, X) - X^T\theta\right)\,\mathrm{d}\alpha\right] \\ &= C_0 + \frac{1}{1-\tau}\mathrm{E}_{(X,\xi)}\left[\rho_\tau\left(g(X,\xi) - X^T\theta\right)\right], \end{aligned} \tag{1.13}$$

where $g(x,z) = v_{[Y|X]}(z,x)$, and ξ follows a uniform distribution on $(0,1)$ independent of X. Furthermore, suppose: (i) for some $c_0 > 0$, the function $g(x,z)$ is differentiable with respect to z for all x and $|z-\tau| < c_0$, and the derivative is uniformly (in x) bounded; and (ii) the matrix $E[XX^T]$ is positive definite. Then, we have:

$$\beta = \arg\min_\theta\ L(\theta),$$

under Model (1.4), where β is the true SQ regression coefficient and the minimizer is uniquely identified[3]*.*

In parallel to Proposition 1, the first part of Theorem I.1 translates the m-Rock loss function into a more useful form (1.13). Minimizing $L(\theta)$ solves the τth linear quantile regression problem of Z versus X, where Z is distributed as (conditional on $X = x$):

$$Z \mid X = x \sim g(x,\xi), \qquad \xi \sim U(0,1).$$

Similar to the arguments following Proposition 1, finding the conditional quantile of

[3]In the case where $|L(\theta)|$ may not be finite, we can consider the minimization of $L(\theta) - L(\beta)$, which is guaranteed to take finite values under the conditions of Theorem I.1.

Z is equivalent to finding the conditional SQ of Y. Thus, β from Model (1.4) is identifiable from the m-Rock loss function $L(\theta)$. The second part of Theorem II.1 provides some sufficient conditions for the uniqueness of the minimizer. In light of Theorem I.1, we shall use the expressions (1.12) and (1.13) interchangeably in the remainder of this dissertation. We emphasize that our modification to the RRM loss function is critical to its validity in SQ regression.

Here we provide some further comments about the m-Rock loss function. First, the conditions for unique identifiability in Theorem I.1 are relatively weak, and are also required for quantile regression analysis (*Koenker*, 2005, Section 2). Second, Theorem I.1 is for population-level identifiability. The loss function $L(\theta)$ involves the unknown conditional SQ of Y given X, therefore Theorem I.1 is not directly useful for SQ estimation in a finite sample. Third, the domain of integration in $L(\theta)$ can be shortened without jeopardizing the identification, as shown in the following Corollary.

Corollary 1. *For any* $0 < \delta \leq 1$, *define*

$$L^{(\delta)}(\theta) = \mathrm{E}_X \left[\int_{\tau-\delta\tau}^{\tau+\delta(1-\tau)} \rho_\tau \left(v_{[Y|X]}(\alpha, X) - X^T\theta \right) \, \mathrm{d}\alpha \right].$$

Under the same conditions of Theorem I.1, the minimizer to $L^{(\delta)}(\theta)$ *is identical to that of* $L(\theta)$.

The upper and lower end of the integral in $L^{(\delta)}$ matches in a suitable way, where taking $\delta = 1$ recovers $L(\theta)$. By taking a smaller δ, Corollary 1 shows that only the conditional SQ at levels near τ are relevant for the m-Rock loss function. Corollary 1 is beneficial when we discuss the practical implementation of the m-Rock SQ regression approach in Chapter 4.

CHAPTER II

Superquantile Regression with Discrete Covariates

In this chapter, we explore new approaches for superquantile regression in the case with discrete covariates. In particular, we devote much of our focus to a new approach based on the m-Rock loss function, which we name the m-Rock approach. We show how the population-level formula can be used to obtain an empirical estimator in a finite sample. As benchmarks, we also investigate two other methods for SQ regression; Both of these methods are intuitive in the setting with discrete covariates, yet they have not been thoroughly studied in the literature.

We begin with the simple setting with discrete covariates because it greatly simplifies the theory and methodology. The main goal of this chapter is to provide some understanding on the validity and applicability m-Rock approach, and to compare it with other intuitive alternatives. We find that the m-Rock approach is the most flexible and efficient among the three new approaches, which confirms its value and potential. These results can be seen as a preliminary and a foundation before further discussion with continuous covariates.

To further simplify the notations, in this section we shall assume a fixed design where the covariates are equally distributed on a few distinct values. We suppose we have M fixed covariate values (including the intercept term) $\{x_1, \ldots, x_M\}$, where M is a fixed number that does not depend on the sample size; And at each covariate

value, we have the same number of *i.i.d.* observations for the response. Therefore, with a total of n samples, the observed data can be written as[1]

$$\{(x_m, Y_{mj}) : \ m = 1, \ldots, M; \ j = 1 \ldots, n/M\}.$$

At each m, the responses Y_{mj} $(j = 1 \ldots, n/M)$ are a random sample from the distribution Y_m. Note, our discussion in this section easily extends to the case with random discrete design and/or unbalanced covariates, but at the cost of more complicated notations.

We fix some other notations here. At each covariate value x_m, let $q_m(s)$ and $v_m(s)$ as the sth $(0 < s < 1)$ quantile and SQ for Y_m, and the linear SQ model (1.4) simplifies to

$$v_m(\tau) = x_m^T \beta, \quad m = 1, \ldots, M,$$

in our setting. We use $\hat{q}_m(s)$ and $\hat{v}_m(s)$ for empirical estimators for the quantile and SQ. When s varies within a range $\mathcal{I}$, we call $\{\hat{v}_m(s) : s \in \mathcal{I}\}$ the empirical SQ process. Let F_m and f_m be the distribution and density function for Y_m. For any vector a, let $\|a\|$ be its ℓ_2 norm.

2.1 The m-Rock approach

2.1.1 A practical implementation

Here we introduce the new m-Rock approach. The population-level loss function $L(\theta)$ in Theorem I.1 of Chapter 1 is not directly feasible for empirical estimation, because it involves the unknown conditional SQ process of Y given X, which includes the parameter of interest itself. To make it practical, we use an initial estimator for the conditional SQ to obtain a plug-in version of $L(\theta)$, which is relatively simple when

[1]Without loss of generality, we assume n is divisible by M.

the covariates are discrete. We give the details below.

With discrete covariates, it is natural to use the empirical SQ (*Scaillet*, 2004; *Chen*, 2007) at each covariate value as an initial estimator. To be more specific, for each $m = 1, \ldots, M$ and each quantile level $s \in (0, 1)$, we first find the sample quantile of the response at x_m, denoted by $\hat{q}_m(s)$; then we simply average the response at x_m that are above the estimated quantile:

$$\hat{v}_m(s) = \frac{\sum_{j=1}^{n/M} Y_{mj} \mathbf{1}[Y_{mj} \geq \hat{q}_m(s)]}{(1-s)n/M},$$

since there are $(1-s)n/M$ observations above the quantile at each x_m. The empirical m-Rock loss function is then given by:

$$L_n(\theta) = \frac{1}{M} \sum_{m=1}^{M} \left[\int_0^1 \rho_\tau \left(\hat{v}_m(\alpha) - x_m^T \theta \right) \, \mathrm{d}\alpha \right], \tag{2.1}$$

which is an approximation for the population-level loss function (1.13).

Correspondingly, the m-Rock estimator is defined to be the minimizer of $L_n(\theta)$. While direct optimization of (2.1) is possible, numerically it may be more convenient to consider the following approximation:

$$\begin{aligned} \widehat{\beta} &= \arg\min_\theta \; L_n(\theta) \\ &\approx \arg\min_\theta \; (MT)^{-1} \sum_{m=1}^{M} \sum_{t=1}^{T} \rho_\tau \left(\hat{v}_m(\alpha_t) - x_m^T \theta \right), \end{aligned} \tag{2.2}$$

where $\alpha_1, \ldots, \alpha_T$ is a fine enough equally-spaced grid over the interval $(0, 1)$. Therefore, computation of the m-Rock SQ regression can be reduced to a quantile regression problem, for which efficient numerical algorithms exist (*Koenker*, 2005, Section 6).[2] This computational trick is due to our Theorem II.1 in Chapter 1.

[2]Note the minimizer to (2.2) may not be unique, in that case $\widehat{\beta}$ refers to any such minimizers.

Several comments are in place for the implementation of the m-Rock approach. First, the grid-based approximation for $L_n(\theta)$ in (2.2) is only for computational purposes. In our theoretical analysis, we still focus on the loss function $L_n(\theta)$ in (2.1) that involves integration. Since $\hat{v}_m(\alpha)$ is a piece-wise constant function in α, the approximation of the integral can be made exact by choosing a sufficiently fine grid, though it may not be necessary in practice. Second, on top of (2.1), we can further exploit Corollary 1 and use a truncated range of integration. In this way, we can avoid estimating the empirical SQ at extreme levels, which may help to stabilize the numerical performance.

2.1.2 Statistical properties of the m-Rock estimator

Here we study the statistical properties of the m-Rock estimator. To this end, we need several technical conditions on the data generating mechanism, which are also needed for other methods later in this Chapter.

Condition R-X. The Gram matrix $D_0 = \sum_{m=1}^{M} x_m x_m^T / M$ is positive definite.

Condition R-Y1. At each x_m $(m = 1, \ldots, M)$, the distribution of Y_m is continuous with density function $f_m(y)$. Furthermore, $f_m(y)$ satisfies: (i) $f_m(y)$ is finite and continuous at each y; (ii) $f_m(q_m(\tau)) > 0$.

Condition R-Y2. At each each x_m $(m = 1, \ldots, M)$, we have:

$$\mathrm{E}[(Y_m^+)^2] < +\infty,$$

where $Y_m^+ = \max\{Y_m, 0\}$.

Condition R-X ensures the m different covariate values are non-degenerate. Conditions R-Y1 and R-Y2 are relatively weak for the response distribution. In fact, these conditions are even weaker than those required by quantile and least-squares regression with fixed design. Under Conditions R-Y1 and R-Y2, the superquantile

function $v_m(\alpha)$ is strictly increasing and continuously differentiable with respect to α.

We begin by giving some finite-sample properties of the empirical m-Rock loss function (2.1). We suppose that $|L_n(\theta)| < +\infty$ for all θ; otherwise we can restrict the domain of interest to $\{\theta : |L_n(\theta)| < +\infty\}$.

Proposition 2. *Under Condition R-X, the following holds true for the empirical m-Rock loss function $L_n(\theta)$:*

1. *$L_n(\theta)$ is convex and Lipshitz continuous.*

2. *The directional derivative of $L_n(\theta)$ exists at any θ and along any direction.*

3. *Suppose there are no ties among the response at each covariate value, then any minimizer of $L_n(\theta)$, denoted by $\widehat{\beta}$, satisfies:*

$$\left\| M^{-1} \sum_{m=1}^{M} x_m \left[\tau - \hat{h}_m(x_m^T \widehat{\beta}) \right] \right\| \leq \frac{C_1 M}{n}, \tag{2.3}$$

for some universal constant $C_1 > 0$, where $\hat{h}_m(z)$ is the empirical inverse[3] of the SQ function:

$$\hat{h}_m(z) = \inf\{s \in [0,1] : \hat{v}_m(s) \geq z\}.$$

Proposition 2 shows that the function $L_n(\theta)$ enjoys some desirable properties. Therefore, theoretical and computational tools from convex optimization apply to the analysis of the m-Rock approach. With convexity, (2.3) gives the necessary first-order optimality condition for the m-Rock estimator. Though $L_n(\theta)$ is not everywhere differentiable, optimality requires all the directional derivatives of $L_n(\theta)$ to be non-negative at $\widehat{\beta}$, which leads to (2.3).

[3]The inverse is well-defined as $\hat{v}_m(\alpha)$ is monotonically increasing in α; see Lemma 3 in Section 2.5.2.

Remark 1. Proposition 2 is a general result that does not depend on Conditions R-Y1 and R-Y2; nor does it depend on the choice of $\hat{v}_m(\alpha)$ in the loss function. The conclusions take hold for any estimator $\hat{v}_m(\alpha)$ that is (i) monotonic in α, and (ii) not flat over α for any interval of length M/n. However, our subsequent asymptotic analysis may depend on the sampling properties of $\hat{v}_m(\alpha)$.

Our subsequent theoretical analysis builds upon the generalized Z-estimation framework[4] from Proposition 2. In particular, (2.3) suggests that the property of the m-Rock estimator is closely tied to that of $\hat{h}_m(\cdot)$, the inverse empirical SQ. The following Lemma establishes a key asymptotic result for $\hat{h}_m$, where we define

$$h_m(z) = \inf\{s \in [0,1] : v_m(s) \geq z\} = v_m^{-1}(z),$$

since $v_m(\alpha)$ is strictly increasing with respect to α.

Lemma 1. *Under a fixed discrete design and Conditions R-Y1 and R-Y2, the inverse empirical SQ satisfies:*

$$\sqrt{\frac{n}{M}}\left(\hat{h}_m[v_m(\tau)] - \tau\right) \xrightarrow{d} \mathrm{N}\left(0,\ \frac{(1-\tau)^2\sigma_m^2(\tau)}{[v_m(\tau) - q_m(\tau)]^2}\right),$$

for each $m = 1, \ldots, M$, *with* $(1-\tau)\sigma_m^2(\tau) = var[Y_m \mid Y_m \geq q_m(\tau)] + \tau[v_m(\tau) - q_m(\tau)]^2$.

Lemma 1 establishes the asymptotic normality of the inverse SQ estimator. It serves as an important technical tool to understand the m-Rock estimator via (2.21). While the asymptotic properties of the SQ estimator $\hat{v}_m(\tau)$ has been well studied in the literature (*Chen*, 2007; *Nadarajah et al.*, 2014; *Zwingmann and Holzmann*, 2016), our Lemma 1 gives the first asymptotic analysis of $\hat{h}_m$. In fact, we obtain more asymptotic results for both $\hat{v}_m$ and $\hat{h}_m$ that complement the literature, but we

[4]We use the word 'generalized' because the estimating equation (2.3) is not an empirical average over each data point.

relegate those discussions to Section 2.5.1; Lemma 1 here is a simple corollary from Lemma 2 therein.

With the help of Proposition 2 and Lemma 1, we are now ready to state the main result for the m-Rock estimator.

Theorem II.1. *Under a fixed discrete design and suppose Conditions R-X, R-Y1 and R-Y2 hold. In addition, if the matrix*

$$D_1 = M^{-1} \sum_{m=1}^{M} \frac{x_m x_m^T}{v_m(\tau) - q_m(\tau)}$$

is positive definite, then the m-Rock estimator $\widehat{\beta}$ is consistent for β in Model (1.4), and it holds that

$$(1-\tau) D_1 \left(\widehat{\beta} - \beta \right) = \frac{1}{M} \sum_{m=1}^{M} x_m \left\{ \tau - \hat{h}_m[v_m(\tau)] \right\} + o_P\left(\frac{1}{\sqrt{n}} \right).$$

In particular,

$$\sqrt{n} \left(\widehat{\beta} - \beta \right) \xrightarrow{d} \mathrm{N}\left(0, \;\; D_1^{-1} \Omega_1 D_1^{-1}\right),$$

where

$$\Omega_1 = M^{-1} \sum_{m=1}^{M} \left\{ \frac{\sigma_m^2}{[v_m(\tau) - q_m(\tau)]^2} x_m x_m^T \right\},$$

and σ_m is defined in Lemma 1.

Theorem II.1 uncovers the main statistical properties of the m-Rock estimator: consistency and asymptotically normality. It also gives an explicit connection between $\widehat{\beta}$ and $\hat{h}_m$ in Lemma 1 via a Bahadur-type representation. While implementation of the m-Rock approach only depends on $\hat{v}_m$ as in (2.2), the first-order asymptotic property of $\widehat{\beta}$ depends directly on $\hat{h}_m$, the inverse function of $\hat{v}_m$.

Our consistency result in Theorem II.1 is different from Theorem 3 of *Rockafellar et al.* (2014). While they show the RRM estimator converges in probability, we

find in Section 1.2 that their convergent limit is not the SQ regression coefficient. Therefore, the modification in the m-Rock approach is necessary for estimating the SQ regression.

2.2 Two other approaches

In this section, we introduce two other methods for linear SQ regression in the case with discrete covariates.

2.2.1 The Linearization approach

The first approach is to linearize the initial SQ estimators. Similar to the m-Rock approach, we start with the empirical SQ estimator $\hat{v}_m(\tau)$ at each covariate value; however, in this method we only use the τth SQ instead of the entire SQ process. While the initial estimates $\hat{v}_m(\tau)$ $(m = 1, \ldots, M)$ are not linear in covariates, we can enforce linearity by fitting a least-squares regression. Specifically, we define the Linearization estimator as follows.

$$\begin{aligned} \widehat{\beta}^{(L)} &= \arg\min_{u} \sum_{m=1}^{M} \left(\hat{v}_m(\tau) - x_m^T u\right)^2 \\ &= \left(\sum_{m=1}^{M} x_m x_m^T\right)^{-1} \left[\sum_{m=1}^{M} x_m \hat{v}_m(\tau)\right]. \end{aligned} \tag{2.4}$$

The Linearization approach effectively starts with a non-parametric estimation of the τth conditional SQ, then it uses those estimators, instead of the original data, to form a parametric estimator under the linear SQ model (1.4). The following Theorem gives the asymptotic properties of the Linearization estimator.

Theorem II.2. *Under a fixed discrete design, and suppose Conditions R-X, R-Y1*

and R-Y2 hold. We have

$$\sqrt{n}\left(\widehat{\beta}^{(L)}-\beta\right)\xrightarrow{d}\mathrm{N}\left(0,\ \ D_0^{-1}\Omega_0 D_0^{-1}\right),$$

where $D_0=\sum_{m=1}^{M}x_m x_m^T/M$, $\Omega_0=\sum_{m=1}^{M}(\sigma_m^2 x_m x_m^T)/M$, *and* σ_m^2 *is defined in Lemma 1.*

Theorem II.2 implies that the Linearization estimator is $\sqrt{n}$-consistent for the SQ regression model, which is typical for parametric estimation and the rate is independent of M. On the contrary, each of the initial estimators $\hat{v}_m(\tau)$ has a variance at the order of M/n. By aggregating a series of non-parametric estimators, we may be able to recover a parametric estimator with a $\sqrt{n}$-convergence rate.

Moreover, the SQ regression model (1.4) can accommodate heterogeneity in data. Motivated by the possible heteroscedasticity, we can use weighted least-squares (WLS) as an alternative approach for the Linearization method in (2.4), which solves:

$$\min_{u}\sum_{m=1}^{M}w_m\left(\hat{v}_m(\tau)-x_m^T u\right)^2, \tag{2.5}$$

where w_m is the weight attached to each covariate value. When heteroscedasticity is present, it is known that WLS with proper weights can achieve better efficiency than OLS in general. For the Linearization method, the optimal (infeasible) weights are given by

$$w_m^*\propto\frac{1}{\sigma_m^2},\quad m=1,\ldots,M. \tag{2.6}$$

However, those optimal weights are generally unknown and have to be estimated in practice. Here we only provide a theoretical guideline for optimal weighting, yet do not further pursue the empirical estimation of WLS Linearization in this chapter.

2.2.2 The Two-Step approach

The next approach is a two-step method motivated by SQ estimation in the one-sample case, where we average the response above the estimated quantile. In the regression setting, we can use quantile regression to estimate the conditional quantile, followed by least-squares regression using the data above the fitted quantile. For this approach, we assume the linearity assumption holds for both the conditional SQ and quantile, which involves the additional assumption that:

$$q_m(\tau) = x_m^T \beta_q, \tag{2.7}$$

on top of the linear SQ model (1.4). With a joint quantile and SQ regression model, the Two-Step procedure is given by

$$\begin{aligned} \hat{\beta}_q &\leftarrow \arg\min_u \sum_{m=1}^{M} \sum_{j=1}^{n/M} \rho_\tau(Y_{mj} - x_m^T u), \\ \widehat{\beta}^{(TS)} &= \arg\min_\theta \sum_{m=1}^{M} \sum_{j=1}^{n/M} \left[(Y_{mj} - x_m^T \theta)^2 \cdot \mathbf{1}\{Y_{mj} \geq x_m^T \hat{\beta}_q\} \right], \end{aligned} \tag{2.8}$$

where $\rho_\tau(\cdot)$ is the check function; here $\widehat{\beta}^{(TS)}$ is the SQ regression estimator of interest and $\hat{\beta}_q$ is the intermediate quantile regression estimator.

The estimation procedure (2.8) falls into the standard framework of two-step M-estimation; See, e.g., Section 12.4 of *Wooldridge* (2010). Accordingly, we can derive the asymptotic property of $\widehat{\beta}^{(TS)}$ in the following result.

Theorem II.3. *Under a fixed discrete design, and suppose Conditions R-X, R-Y1 and R-Y2, In addition, suppose the matrix*

$$D_2 = M^{-1} \sum_{m=1}^{M} f_m(q_m(\tau)) x_m x_m^T,$$

is positive definite. Under a joint model of (1.4) and (2.7), we have

$$\sqrt{n}\left(\widehat{\beta}^{(TS)}-\beta\right)\xrightarrow{d}\mathrm{N}\left(0,\ \ D_0^{-1}\left(V+G\,\Sigma\,G\right)D_0^{-1}\right),$$

where $\Sigma=\tau(1-\tau)D_2^{-1}D_0D_2^{-1}$, *and*

$$V=[M(1-\tau)]^{-1}\sum_{m=1}^{M}\left\{var[Y_m\mid Y_m\geq q_m(\tau)]x_mx_m^T\right\},$$
$$G=[M(1-\tau)]^{-1}\sum_{m=1}^{M}\left\{f_m(q_m(\tau))[v_m(\tau)-q_m(\tau)]x_mx_m^T\right\}.$$

In particular, the property of the Two-Step estimator depends on Σ, which is the asymptotic variance-covariance matrix for the classic quantile regression estimator $\hat{\beta}_q$. This is intuitive since we use the linear quantile regression as the first step in (2.8).

The Two-Step approach is connected to the Linearization approach as follows. In the Two-Step estimation procedure (2.8), we rely on the linear quantile regression model (2.7) to provide an estimator for $q_m(\tau)$. Without relying on the linearity of conditional quantile, we may also choose to use the empirical estimator $\hat{q}_m(\tau)$ at each x_m, which would lead to another estimator:

$$\min_{\theta}\sum_{m=1}^{M}\sum_{j=1}^{n/M}\left[(Y_{mj}-x_m^T\theta)^2\cdot\mathbf{1}\{Y_{mj}\geq\hat{q}_m(\tau)\}\right].$$

We can show that the above formulation is exactly the same as the Linearization estimator $\widehat{\beta}^{(L)}$. While $\widehat{\beta}^{(TS)}$ uses Model (2.7) to assist quantile estimation, it is not clear whether $\widehat{\beta}^{(TS)}$ is more efficient than $\widehat{\beta}^{(L)}$, even when (2.7) takes hold. We relegate more detailed comparisons to Section 2.3

Remark 2. One advantage of $\widehat{\beta}^{(TS)}$ is that both the estimation procedure and the proof of Theorem II.3 can be easily extended to the case with continuous covariates. The two-step M-estimation framework does not rely on the discreteness of the covari-

ates. On the other hand, both the m-Rock and the Two-Step approaches rely on the empirical SQ estimator at each covariate value, the analysis of which becomes more complicated with continuous covariates.

2.3 Connection and comparison

In this section, we compare the three proposed methods in the case of discrete covariates. We begin with their operational differences, followed by an explicit comparison of the asymptotic efficiency under two families of models. The focus here is to better understand the behaviour of the m-Rock approach.

2.3.1 Conceptual differences

Operationally, both the m-Rock and the Linearization approaches require initial SQ estimators at each covariate value. Among those two methods, the Linearization approach is conceptually simpler as it only involves the SQ at a single level τ, whereas the m-Rock approach needs the SQ process at many other levels. Note, however, the m-Rock approach does not require a parametric model beyond that for the τth SQ in (1.4).

On the other hand, the Two-Step approach is more straightforward and does not depend on any initial estimator, yet it relies critically on the additional linear quantile regression model in (2.7). With a joint quantile and SQ model, the Two-Step method becomes remarkably easy to implement by fitting two standard regression models. The method also applies seamlessly to cases with general covariate distributions and/or higher dimensions. Such a joint model is often used in recent works on SQ regression (*Dimitriadis and Bayer*, 2019; *Barendse*, 2020).

We highlight several important differences between the three methods in Table 2.1. To summarize, the Two-Step method requires the most stringent assumption; and the m-Rock approach requires the most computational effort.

Table 2.1: Requirements of the three SQ regression methods

	Parametric quantile regression model?	Initial estimator for the τth SQ?	Initial SQ estimator beyond τth level?
m-Rock	$\times$	✓	✓
Linearization	$\times$	✓	$\times$
Two-step	✓	$\times$	$\times$

2.3.2 Efficiency comparison I: homoscedastic models

In the following, we use two examples to compare the asymptotic variances for the three estimators. In the first example, we consider a homoscedastic linear model:

$$Y_{mj} = x_m^T \eta + \varepsilon_{mj} \quad (m = 1, \ldots, M;\, j = 1, \ldots, n_0), \tag{2.9}$$

where each x_m is a vector that includes the intercept term, and ε_{mj}'s are $i.i.d.$ from the same distribution as ε. Let $q_0(\tau)$ and $v_0(\tau)$ be the τth quantile and SQ of ε, respectively. Under Model (2.9), both the τth quantile and SQ regression model are linear, and can be written as

$$q_m(\tau) = x_m^T \eta + q_0(\tau), \qquad v_m(\tau) = x_m^T \eta + v_0(\tau).$$

In the following, we consider a fixed τ and hence omit the index τ in $q_0(\tau)$ and $v_0(\tau)$ to simplify the notations.

We fix some notations before the comparison. Let $V_0 = \mathrm{var}\,(\varepsilon \mid \varepsilon \geq q_0)$, and let $f_0(\cdot)$ be the density function for ε. Recalling f_m from Condition R-Y1 and σ_m from

Lemma 1, we can simplify the following conditional quantities under Model (2.9):

$$\begin{aligned} f_m(q_m(\tau)) &= f_0(q_0), \\ \text{var}\left[Y_{mj} \mid Y_{mj} \geq q_m(\tau)\right] &= V_0, \\ v_m(\tau) - q_m(\tau) &= v_0 - q_0, \\ \sigma_m^2 = \sigma_0^2 &\triangleq \frac{V_0 + \tau(v_0 - q_0)^2}{1-\tau}. \end{aligned} \tag{2.10}$$

With a homoscedastic model, none of the above quantities depend on x_m. Furthermore, let $\text{AVar}^{(mR)}$, $\text{AVar}^{(L)}$ and $\text{AVar}^{(TS)}$ be the re-scaled (by the sample size n) asymptotic variance-covariance matrices for the m-Rock, Linearization, and Two-Step methods, respectively.

Plugging in the quantities in (2.10) into Theorems II.1, II.2 and II.3, straightforward calculations show that

$$\text{AVar}^{(mR)} = \text{AVar}^{(L)} = \text{AVar}^{(TS)} = \sigma_0^2 D_0^{-1},$$

where σ_0^2 is also the sampling variance for the one-sample SQ (*Chen*, 2007). Under Model (2.9), all three methods are asymptotically the same in efficiency. Without any heteroscedasticity, the sandwich-form variance formulae collapse to a common one. Moreover, even though the quantile function is linear-in-covariates, the Two-Step method does not offer any efficiency improvement by exploiting this linearity.

2.3.3 Efficiency comparison II: location-scale shift models

Next we consider a heteroscedastic location-scale shift model. For simplicity, we restrict to the case with only one (discrete) scalar covariate and even omit the intercept term in the regression. With M ($M > 2$) covariate values $0 < x_1 < \ldots < x_M$

and a fixed $\gamma_1 > 0$, we consider the model

$$Y_{mj} = \eta_1 x_m + (\gamma_1 x_m) \cdot \varepsilon_{mj} \quad (m = 1, \ldots, M;\, j = 1, \ldots, n_0), \tag{2.11}$$

where the error terms ε_{mj} are *i.i.d.* across both m and j. We consider the comparison for a fixed τ, and we adopt the same notations that follow (2.9). The τth quantile and SQ regression under Model (2.11) are

$$q_m(\tau) = (\eta_1 + \gamma_1 q_0) x_m, \qquad v_m(\tau) = (\eta_1 + \gamma_1 v_0) x_m,$$

and the true SQ regression coefficient is $\beta_1 + \gamma_1 v_0$ since we have no intercept term.

Now we calculate the asymptotic variance for each estimator under Model (2.11). Parallel to the calculations in (2.10), we can show that the quantity σ_m is proportional to $(\gamma_1 x_m)$, yet $f_m(q_m(\tau))$ is inversely-proportional to $(\gamma_1 x_m)$ under the location-scale shift model (2.11). Moreover, let $\mu_j = \sum_{m=1}^{M} x_m^j / M$ for $j = 1, \ldots, 4$, and we define $R_1 = \mu_2/(\mu_1^2)$ and $R_2 = \mu_4/(\mu_2^2)$. For the m-Rock estimator, we have from Theorem II.1 that

$$\begin{aligned}
\text{AVar}^{(mR)} &= D_1^{-1} \Omega_1 D_1^{-1} \\
&= \left[\frac{1}{\gamma_1 (v_0 - q_0)} \mu_1 \right]^{-1} \left[\frac{\sigma_0^2}{(v_0 - q_0)^2} \mu_2 \right] \left[\frac{1}{\gamma_1 (v_0 - q_0)} \mu_1 \right]^{-1} \\
&= \frac{\gamma_1^2}{1 - \tau} [V_0 + \tau (v_0 - q_0)^2] \cdot R_1,
\end{aligned}$$

since $\sigma_0^2 = (1 - \tau)^{-1} [V_0 + \tau (v_0 - q_0)^2]$. Similarly, for the Linerization estimator we

have

$$\begin{aligned}\text{AVar}^{(L)} &= D_0^{-1}\Omega_0 D_0^{-1} \\ &= (\mu_2)^{-1}(\gamma_1^2\sigma_0^2\mu_4)(\mu_2)^{-1} \\ &= \frac{\gamma_1^2}{1-\tau}[V_0 + \tau(v_0 - q_0)^2] \cdot R_2.\end{aligned}$$

And for the Two-Step estimator,

$$\begin{aligned}\text{AVar}^{(TS)} &= D_0^{-1}\left(V + G\,\Sigma\,G\right)D_0^{-1} \\ &= \frac{1}{1-\tau}(\mu_2)^{-1}\left[\gamma_1^2 V_0\,\mu_4 + \tau\gamma_1^2(v_0 - q_0)^2\mu_1^{-2}\mu_2^3\right](\mu_2)^{-1} \\ &= \frac{\gamma_1^2}{1-\tau}\left[V_0 \cdot R_2 + \tau(v_0 - q_0)^2 \cdot R_1\right].\end{aligned}$$

Thus, all the asymptotic variances depend on two quantities under Model (2.11): V_0 and $\tau(v_0 - q_0)^2$; And the difference between three methods originates from how those two quantities are weighted, either by R_1 or R_2. We show in Section 2.5.4 that

$$R_1 < R_2, \tag{2.12}$$

and hence the comparison of the asymptotic efficiency follows as

$$\text{AVar}^{(mR)} < \text{AVar}^{(TS)} < \text{AVar}^{(L)},$$

which suggests that the m-Rock method is the most efficient under Model (2.11).

We provide some heuristic explanations for the asymptotic relative efficiency. The variance for $\widehat{\beta}^{(L)}$ uses the weight R_2 for both quantities V_0 and $\tau(v_0 - q_0)^2$. Next, $\widehat{\beta}^{(TS)}$ improves upon $\widehat{\beta}^{(L)}$ by utilizing the linear quantile regression model: With more accurate conditional quantile estimation, the variance of $\widehat{\beta}^{(TS)}$ uses a better weight R_1, but only for one component $\tau(v_0 - q_0)^2$.

Furthermore, the variance for the m-Rock estimator uses the better weight R_1 for both components V_0 and $\tau(v_0 - q_0)^2$. Surprisingly, the m-Rock approach achieves better efficiency than the Two-Step method, even though the latter requires more modeling assumptions. The key to this improvement is the implicit weighting induced by the m-Rock loss function. In Theorem II.1, the middle part of the sandwich form variance-covariance matrix involves a weight[5]:

$$w_m \propto \frac{1}{v_m(\tau) - q_m(\tau)} = \frac{1}{\gamma_1 x_m}, \quad m = 1, \ldots, M,$$

under Model (2.11). Although these weights w_m are not optimal as those in (2.6), they are still beneficial for efficiency because w_m reflects the scale of Y_m on the right tail: the data are down-weighted if the conditional variance of the response is larger. With heterogeneity, such weighting is beneficial in general (*Leamer*, 2010), and we conjecture that the m-Rock approach is competitive beyond the relatively simple model (2.11).

To conclude the efficiency comparisons, the m-Rock approach is superior to other two methods we considered in this section: it does the best in heteroscedastic models, while it remains equally competitive in homoscedastic models. Therefore, the m-Rock approach is partially *adaptive* to the underlying heterogeneity in data.

2.4 Discussion

In this chapter, we study new approaches for superquantile regression in the setting with discrete covariates. Under this relatively simple but illustrative setting, we are able to focus on the nature of the problem, and to think outside of the traditional M-estimation framework. We consider three new approaches that complement the literature on SQ regression, among which the m-Rock approach demonstrates the

[5]The weighting is implicit because we never have to estimate those weights in the implementation of the m-Rock approach.

most desirable statistical efficiency. Via two examples, we show that the m-Rock approach is at least as efficient, if not more efficient, than the other two approaches, yet it does not require any additional assumption beyond the τ-th SQ regression model (1.4).

The m-Rock approach originates from a modified loss function in Chapter 1. When the covariates are discrete, we compute the SQ regression using not the raw data, but an array of empirical SQ estimators at each covariate value and at a range of quantile levels. Operationally, the m-Rock approach fits a linear quantile regression to the array of initial SQ estimators. The approach is intuitive following Proposition 1 in Chapter 1, where we show the τth SQ can be interpreted as the τth quantile of the SQ process.

Admittedly, the setting with discrete covariates is relatively restrictive in this chapter; And our comparison does not include other SQ regression approaches in the recent literature. The focus here is to understand and demonstrate the potential of the novel m-Rock approach. We shall give more discussion and comparisons in the next chapter.

2.5 Technical details

In this section, we give technical details that supplement the discussion in this chapter, which include the proofs of all results.

2.5.1 Auxiliary results for the one-sample SQ process

We first present asymptotic results in the one-sample case without any covariate. These results also apply to the empirical SQ estimators at each covariate value in our regression setting of this chapter.

We fix some notations for the discussion of the one-sample problem. Suppose the data $Y_1, \ldots, Y_n$ are *i.i.d.* observations with a common distribution function $F(y)$. For

any $0 < s < 1$, let $\hat{q}(s)$ be the sample quantile from the n observations, we define the empirical SQ estimator as:

$$\hat{v}(s) = \frac{\sum_{i=1}^{n} Y_i \cdot \mathbf{1}\{Y_i \geq \hat{q}(s)\}}{\sum_{i=1}^{n} \mathbf{1}\{Y_i \geq \hat{q}(s)\}}. \tag{2.13}$$

While the parameter of interest is the τth SQ, here we consider the empirical SQ process, which is the stochastic process given by $\{\hat{v}(s) : s \in [\tau_L, \tau_U]\}$, where $0 < \tau_L < \tau < \tau_U < 1$. Let $\ell^\infty[a, b]$ be the set of all uniformly bounded functions on the interval $[a, b]$. To further simplify notations, in the remainder of this subsection, we shall write $\hat{v}_s = \hat{v}(s)$ and $\hat{q}_s = \hat{q}(s)$, and we define q_L and q_U as the τ_L-th and τ_U-th quantile, respectively. In the one-sample case, the notations here may be different than those in the regression setting.

We need the following technical condition, which is the one-sample counterpart for Conditions R-Y1 and R-Y2.

Condition U. The distribution function $F(y)$ is continuously differentiable on the interval $[q_L - \varepsilon, q_U + \varepsilon]$ for some $\varepsilon > 0$; the density function $f(y)$ is bounded away from zero on the same interval. Furthermore, we have $\mathrm{E}[Y^2 \cdot \mathbf{1}\{Y \geq 0\}] < +\infty$.

Now we present the first main result in the one-sample case, which concerns the weak convergence of the empirical SQ as a stochastic process indexed by the quantile level. Not only is the result an important technical tool for subsequent analysis, but it also is of interest on its own.

Theorem II.4. *Suppose Condition U holds, then we have*

$$\hat{v}_s - v_s = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(Y_i - q_s) \cdot \mathbf{1}\{Y_i \geq q_s\}}{1 - s} - (v_s - q_s) \right] + o_P\left(n^{-1/2}\right),$$

uniformly in $s \in [\tau_L, \tau_U]$. *Furthermore, the centered empirical SQ process converges*

weakly:

$$\sqrt{n}\,[\hat{v}(\cdot) - v(\cdot)] \rightsquigarrow \mathbb{G}(\cdot) \quad \textit{in } \ell^{\infty}[\tau_L,\, \tau_U],$$

where $\mathbb{G}(\cdot)$ *is a mean zero Gaussian Process.*

Theorem II.4 gives the uniform (weak) Bahadur representation for the empirical SQ process. To the best of our knowledge, the uniformity of the result is new. Restricting to a single quantile level τ, *Chen* (2007) and *Zwingmann and Holzmann* (2016) study the asymptotic properties of the SQ estimator $\hat{v}(\tau)$ under more general conditions; on the other hand we discuss process convergence. Practically, Theorem II.4 is a technical tool for simultaneous statistical inference for a range of superquantiles.

As a simple corollary of Theorem II.4, we can obtain the asymptotic distribution for the τth empirical SQ, which is known from, e.g., *Chen* (2007) and *Zwingmann and Holzmann* (2016). We omit the proof since it simply combines the Central Limit Theorem with the Bahadur representation in Theorem II.4.

Corollary 2. *Under Condition U, we have*

$$\sqrt{n}(\hat{v}_\tau - v_\tau) \xrightarrow{d} \mathrm{N}(0, \sigma_\tau^2),$$

with $(1-\tau)\sigma_\tau^2 = var(Y \mid Y \geq q_\tau) + \tau(v_\tau - q_\tau)^2$.

The asymptotic variance σ_τ^2 consists of two parts. The first part is the variance in estimating v_τ when q_τ is known, whereas the second part is attributable to quantile estimation (*Zwingmann and Holzmann*, 2016).

Next, we proceed to the study of the inverse SQ function, which we define below:

$$h(z) = \{s : v_s = z\} \quad \text{and} \quad \hat{h}(z) = \inf\{s \in [0,1] : \hat{v}_s \geq z\},$$

for any $z \in [v_{\tau_L}, v_{\tau_U}]$. Note v_s is strictly increasing in $s \in [\tau_L, \tau_U]$, and we show in

Lemma 3 that $\hat{v}(z)$ is also non-decreasing in $s \in [\tau_L, \tau_U]$; therefore the definitions above are well-defined. The following Lemma shows that $\hat{h}(z)$, the empirical inverse SQ, is also asymptotically Gaussian.

Lemma 2. *Under Condition U, the inverse SQ process satisfies:*

$$\hat{h}(z) - h(z) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i - q_{h(z)}) \cdot \mathbf{1}\{Y_i \geq q_{h(z)}\}}{z - q_{h(z)}} - (1 - h(z))\right] + o_P\left(n^{-1/2}\right),$$

uniformly in $z \in [v_\tau - \varepsilon', v_\tau + \varepsilon']$ *for some* $\varepsilon' > 0$. *In particular, we have:*

$$\sqrt{n}\left(\hat{h}(v_\tau) - \tau\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{(1-\tau)^2\sigma_\tau^2}{(v_\tau - q_\tau)^2}\right),$$

with σ_τ^2 *defined in Corollary 2. Furthermore, the process* $n^{1/2}[\hat{h}(z) - h(z)]$ *is asymptotically equi-continuous over* $z \in [v_\tau - \varepsilon', v_\tau + \varepsilon']$ *with respect to the Euclidean distance.*

The asymptotic property of $\hat{h}(z)$ is an essential for the analysis of the m-Rock regression approach. The proof of Lemma 2 builds upon the uniform representation in Theorem II.4, as well as the functional Delta method; See, e.g., Theorem 20.8 in *Van der Vaart* (2000). Note, the asymptotic normality of $\hat{v}_\tau$ at a single level (*Chen*, 2007; *Zwingmann and Holzmann*, 2016), is not sufficient to establish the result in Lemma 2.

2.5.2 Proof for the one-sample case

The proofs in this subsection rely on standard empirical process tools in, e.g., *Van Der Vaart and Wellner* (1996), and we adopt the same notations therein. Let $Y_1, \ldots, Y_n$ be *i.i.d.* observations from the same population. For a class of function $y \mapsto f(y;\theta)$ indexed by $\theta \in \mathbb{R}^q$, let $\mathbb{E}_n[f(Y^*;\theta)] = \sum_{i=1}^n f(Y_i;\theta)/n$, $\mathrm{E}[f(Y^*;\theta)] = \mathrm{E}[f(Y_i;\theta)]$ and $\mathbb{G}_n[f(Y^*;\theta)] = n^{1/2}\{\mathbb{E}_n[f(Y^*;\theta)] - \mathrm{E}[f(Y^*;\theta)]\}$. We sometimes use

the subscript and write $\mathbb{E}_n[f_\theta]$ instead of $\mathbb{E}_n[f(Y^*;\theta)]$ for further simplicity. For a semi-metric space $\mathbb{T}$, we use $\ell^\infty(\mathbb{T})$ to denote the functional space that consists all bounded functions of $\mathbb{T} \mapsto \mathbb{R}$. Moreover, we use the same notations in the previous subsection.

We need the following technical lemmas, the proofs of which are at the end of this subsection.

Lemma 3. *Under Condition U, for any s,t such that $\tau_L \le s < t \le \tau_U$, we have $\hat{v}_s \le \hat{v}_t$ and $v_s < v_t$. Furthermore, as a function of s, $\hat{v}_s$ is left continuous whose right limit exists everywhere.*

Lemma 4. *Let $\psi(y;\theta,s) = (y - v_s)\mathbf{1}\{y \ge \theta\}$. Under the conditions of Theorem II.4, and suppose that $|\tilde{q}_s - q_s| = o_P(1)$ uniformly over $s \in [\tau_L, \tau_U]$, then we have*

$$\sup_{s\in[\tau_L,\tau_U]} \left|\mathbb{G}_n[\psi_{(\tilde{q}_s,s)}] - \mathbb{G}_n[\psi_{(q_s,s)}]\right| = o_P(1).$$

Furthermore, the function class $\mathcal{F} = \{y \mapsto \psi(y,\theta,s) : \theta \in [q_L, q_U], s \in [\tau_L,\tau_U]\}$ is Donsker.

Proof of Theorem II.4

Proof. We first prove the Bahadur representation for a broader class of SQ estimator. Consider any estimator $\tilde{v}_s$ that solves the following estimating equation:

$$0 = \sum_{i=1}^{n} (y_i - \tilde{v}_s)\mathbf{1}\{Y_i \ge \tilde{q}_s\}, \tag{2.14}$$

where $\tilde{q}_s$ is any estimator for the q_s that satisfies (i) $\sup_{s\in[\tau_L,\tau_U]} |\tilde{q}_s - q_s| = O_{\mathrm{P}}(n^{-1/2})$; and (ii) $\tilde{q}_s \in \ell^\infty([\tau_L,\tau_U])$ as a stochastic process indexed by s. Choosing $\tilde{q}_s$ as the sample quantile in (2.14) recovers the empirical SQ estimator.

Let $\psi(y;\theta,s) = (y - v_s)\mathbf{1}\{y \ge \theta\}$. Given the quantile estimators $\tilde{q}_s$ ($s \in [\tau_L,\tau_U]$),

the estimating equation (2.14) for $\tilde{v}_s$ solves $\mathbb{E}_n[(Y^* - \tilde{v}_s)\mathbf{1}\{Y^* \geq \tilde{q}_s\}] = \mathbb{E}_n[\psi(Y^*; \tilde{q}_s, s)] + (v_s - \tilde{v}_s)\mathbb{E}_n[\mathbf{1}\{Y^* \geq \tilde{q}_s\}] = 0$. Hence the estimator $\tilde{v}_s$ satisfies:

$$
\begin{aligned}
& \sqrt{n}(\tilde{v}_s - v_s)\mathbb{E}_n[\mathbf{1}\{Y^* \geq \tilde{q}_s\}] \\
= \; & \sqrt{n}\mathbb{E}_n[\psi(Y^*; \tilde{q}_s, s)] && (2.15) \\
= \; & \sqrt{n}\left\{\mathrm{E}[\psi(Y^*; \tilde{q}_s, s)] - \mathrm{E}[\psi(Y^*; q_s, s)]\right\} && (2.16) \\
& + \underbrace{\mathbb{G}_n[\psi(Y^*; \tilde{q}_s, s)] - \mathbb{G}_n[\psi(Y^*; q_s, s)]}_{R_1(s)} \\
& + \mathbb{G}_n[\psi(Y^*; q_s, s)] \\
= \; & \frac{\partial \mathrm{E}[\psi(Y^*; q_s, s)]}{\partial q_s}[\sqrt{n}\,(\tilde{q}_s - q_s)] + \mathbb{G}_n[\psi(Y^*; q_s, s)] + R_1(s) \\
& + \underbrace{\sqrt{n}\left\{\mathrm{E}[\psi(Y^*; \tilde{q}_s, s)] - \mathrm{E}[\psi(Y^*; q_s, s)] - \frac{\partial \mathrm{E}[\psi(Y^*; q_s, s)]}{\partial q_s}\,(\tilde{q}_s - q_s)\right\}}_{R_2(s)} \\
= \; & (v_s - q_s) f_Y(q_s)[\sqrt{n}\,(\tilde{q}_s - q_s)] + \mathbb{G}_n[\psi(Y^*; q_s, s)] + R_1(s) + R_2(s),
\end{aligned}
$$

where (2.16) holds since $\mathrm{E}[\psi(Y^*; q_s, s)] = 0$ for all $s \in [\tau_L, \tau_U]$, and the last inequality follows since $\partial \mathrm{E}[\psi(Y^*; \theta, s)]/\partial\theta = (v_s - \theta) f_Y(\theta)$.

Now we show that both $R_1(s)$ and $R_2(s)$ are negligible uniformly in s. By Lemma 4, we immediately obtain $R_1(s) = o_{\mathrm{P}}(1)$ uniformly over $s \in [\tau_L, \tau_U]$. For R_2, we first re-write

$$
\mathrm{E}[\psi(Y^*; \theta, s)] = \int_{\theta}^{+\infty} y f_Y(y)\,\mathrm{d}y - v_s[1 - F_Y(\theta)] \triangleq I_1(\theta) + v_s \times I_2(\theta),
$$

and hence by Taylor expansion with respect to θ we have:

$$
\begin{aligned}
& \sup_{s \in [\tau_L, \tau_U]} |R_2(s)| \\
\leq \; & \sup_{s \in [\tau_L, \tau_U]} |\sqrt{n}\Delta_s| \times \sup_{\substack{\theta \in [q_{\tau_L}, q_{\tau_U}] \\ |\theta - \theta'| \leq \Delta_s}} |[I_1'(\theta') - I_1'(\theta)] - v_s \times [I_2'(\theta') - I_2'(\theta)]| \\
= \; & o_{\mathrm{P}}(1),
\end{aligned}
$$

where $\Delta_s = \tilde{q}_s - q_s$, the last equality follows since: (i) $n^{1/2}(\tilde{q}_s - q_s)$ is asymptotically tight, (ii) v_s is uniformly bounded over $s \in [\tau_L, \tau_U]$, and (iii) both $I_1(\theta)$ and $I_2(\theta)$ are continuously differentiable on $\theta \in [q_{\tau_L} - \varepsilon_0, q_{\tau_U} + \varepsilon_0]$ under Condition U.

Combining the results for $R_1(s)$ and $R_2(s)$ with Equation (2.15), we have

$$\sqrt{n}(\tilde{v}_s - v_s)\mathbb{E}_n[\mathbf{1}\{Y^* \geq \tilde{q}_s\}] = (v_s - q_s)f_Y(q_s)[\sqrt{n}\,(\tilde{q}_s - q_s)] + \mathbb{G}_n[\psi(Y^*; q_s, s)] + o_{\mathrm{P}}(1), \tag{2.17}$$

where the $o_{\mathrm{P}}(1)$ term is uniform in $s \in [\tau_L, \tau_U]$. From here, we can deduce the $n^{1/2}$-uniform consistency of $\tilde{v}_s$ as follows. From the Lemma 4, the function class $\{y \mapsto \psi(y, \theta, s); s \in [\tau_L, \tau_U], \theta \in [q_L, q_U]\}$ is Donsker, therefore

$$\sup_{s\in[\tau_L,\tau_U]} |\sqrt{n}\mathbb{G}_n[\psi(Y^*; q_s, s)]| \leq \sup_{\substack{s\in[\tau_L,\tau_U]\\ \theta\in[q_L,q_U]}} |\sqrt{n}\mathbb{G}_n[\psi(Y^*; \theta, s)]| = O_{\mathrm{P}}(1);$$

Furthermore, the assumptions on $\tilde{q}_s$ at the beginning of the proof implies

$$\sup_{s\in[\tau_L,\tau_U]} \sqrt{n}|\tilde{q}_s - \tilde{q}_s| = O_{\mathrm{P}}(1), \quad \text{and} \quad \mathbb{E}_n[\mathbf{1}\{y \geq \tilde{q}_s\}] = 1 - s + o_{\mathrm{P}}(1).$$

Hence, it follows from (2.17) that

$$\sup_{s\in[\tau_L,\tau_U]} |\sqrt{n}(\tilde{v}_s - v_s)| = O_{\mathrm{P}}(1).$$

From here, we can obtain the uniform Bahadur representation of $\tilde{v}_s - v_s$. Dividing both sides of (2.17) by $(1 - s)$, we obtain, since $\sqrt{n}(\tilde{v}_s - v_s)$ is asymptotically tight in $\ell^\infty([\tau_L, \tau_U])$, that

$$\sqrt{n}(\tilde{v}_s - v_s) = \frac{1}{1-s}\left\{\sqrt{n}\,(\tilde{q}_s - q_s)\,(v_s - q_s)f_Y(q_s) + \mathbb{G}_n[(Y^* - v_s)\mathbf{1}\{Y^* \geq q_s\}]\right\} + o_{\mathrm{P}}(1), \tag{2.18}$$

uniformly over $s \in [\tau_L, \tau_U]$.

In particular, if we choose $\tilde{q}_s$ to be the sample quantile that satisfies $\mathbb{E}_n[\mathbf{1}\{y \leq \tilde{q}_s\}] = s$, then for sufficiently large n, the estimator obtained from (2.14) is asymptotically equivalent to the empirical SQ estimator $\hat{v}_s$ defined in (2.13). Since the sample quantile satisfies

$$\sqrt{n}(\hat{q}_s - q_s) = \sum_{i=1}^{n} \frac{s - \mathbf{1}\{Y_i \leq q_s\}}{f_Y(q_s)} + o_{\mathrm{P}}(1),$$

uniformly in s (see, e.g., Corollary 21.5 of *Van der Vaart* (2000)). Combining the above displayed equation with (2.18), we have

$$\sqrt{n}(\hat{v}_s - v_s) = \frac{1}{1-s}\mathbb{G}_n[(y - q_s)\mathbf{1}\{y \geq q_s\}] + o_{\mathrm{P}}(1),$$

uniformly over $s \in [\tau_L, \tau_U]$.

Finally, we show that the empirical SQ process $\sqrt{n}(\hat{v}_s - v_s)$ converges towards a Gaussian Process in $\ell^\infty[\tau_L, \tau_U]$. In view of the uniform Bahadur representation, it suffices to consider the process $\mathbb{G}_n[(Y^* - q_s)\mathbf{1}\{Y^* \geq q_s\}]$. Since q_s is uniformly Lipschitz continuous in $s \in [\tau_L, \tau_U]$, it follows from Example 19.19 of *Van der Vaart* (2000) that the function class $\{y \mapsto (y - q_s)\mathbf{1}[y \geq q_s] : s \in [\tau_L, \tau_U]\}$ is Donsker. Therefore $\mathbb{G}_n[(Y^* - q_s)\mathbf{1}\{Y^* \geq q_s\}] \xrightarrow{d} \mathbb{G}_\infty(s)$ as a function of s on the space $\ell^\infty([\tau_L, \tau_U])$; here $\mathbb{G}_\infty(s)$ is a zero-mean Gaussian process with continuous sample path with respect to the semi-metric

$$\rho(s,t) = \left(\mathrm{E}\{(Y^* - q_s)\mathbf{1}[Y^* \geq q_s] - (Y^* - q_t)\mathbf{1}[Y^* \geq q_t]\}^2\right)^{1/2}, \quad s,t \in [\tau_L, \tau_U].$$

Since $\rho(s,t) \leq |q_s - q_t| \lesssim |s - t|$, the sample path of $\mathbb{G}_\infty(\cdot)$ is also continuous with respect to the Euclidean distance. This concludes the proof. □

Proof of Lemma 2

Proof. Let $a = v_\tau - \varepsilon_0$ and $b = v_\tau + \varepsilon_0$ for some constant ε_0 such that $v_{\tau_U} - 2\varepsilon_0 \geq$

$v_\tau \geq v_{\tau_L} + 2\varepsilon_0$. Define the function space $\mathbb{D}_1$ as the space of all non-decreasing, continuous function on $[\tau_L, \tau_U]$. For any function $F \in \mathbb{D}_1$, we define the inverse map $\phi(\cdot) : \mathbb{D}_1 \mapsto \ell^\infty([a,b])$ such that $\phi(F)(z) = \inf\{s \in [\tau_L, \tau_U] : F(s) \geq z\}$ for $z \in [a,b]$ [6]. Note that v_s as a function of $s \in [\tau_L, \tau_U]$ is continuously differentiable with $\partial v_s/\partial s = (v_s - q_s)/(1-s) > 0$. Following Lemma 21.4 in *Van der Vaart* (2000), the map $\phi(\cdot)$ is Hadamard-differentiable at $v_s \in \mathbb{D}_1$, tangentially to the set of all continuous (with respect to the Euclidean distance) functions on $[\tau_L, \tau_U]$. The Hadamard-derivative of the inverse map ϕ at v_s is $\phi'_v(h) = -h(v^{-1})/v'(v^{-1})$, for any continuous function h.

Next we apply the functional Delta method. Note $v_s, \hat{v}_s \in \mathbb{D}_1$, and $h(\cdot) = \phi \circ v_s(\cdot) \in \ell^\infty([a,b])$; since $\hat{v}_{\tau_L} \xrightarrow{P^*} v_{\tau_L} < a$, $\hat{v}_{\tau_U} \xrightarrow{P^*} v_{\tau_U} > b$, the inverse SQ process $\hat{h}(\cdot) = \phi \circ \hat{v}_s(\cdot)$ with probability going to 1 [7]. Therefore, applying the functional Delta method (Theorem 20.8 in *Van der Vaart* (2000)) towards the inverse map ϕ gives

$$\begin{aligned}
\sqrt{n}[\hat{h}(z) - h(z)] &= -\left[\frac{\sqrt{n}(\hat{v}_s - v_s)}{v'(s)}\right]\Bigg|_{s=v^{-1}(z)} + o_{\mathrm{P}}(1) \\
&= -\left[\frac{\mathbb{G}_n[(Y^* - q_s)\mathbf{1}\{Y^* \geq q_s\}]}{v_s - q_s}\right]\Bigg|_{s=v^{-1}(z)} + o_{\mathrm{P}}(1),
\end{aligned}$$

in $\ell^\infty[a,b]$, which shows the first part of the Lemma. Since $n^{1/2}(\hat{v}_s - v_s) \xrightarrow{d} \mathbb{G}_\infty(s)$, it follows that $n^{1/2}[\hat{h}(z) - h(z)]$ also converges towards a Gaussian process with continuous sample path (with respect to the Euclidean distance), since v_s is continuously differentiable with respect to s. Asymptotic equi-continuity of $n^{1/2}[\hat{h}(z) - h(z)]$ is then a consequence of its convergence towards a continuous stochastic process.

For the second part of the Lemma, taking $z = v_\tau$ in the above displayed equation, and recalling $\sqrt{n}(\hat{v}_\tau - v_\tau) \xrightarrow{d} \mathrm{N}(0, \sigma_\tau^2)$ from Theorem II.4 concludes the proof. □

[6] We define $\phi(F)(z) = \tau_U$ if $\sup_{s \in [\tau_L, \tau_U]} F(s) < z$, so that $\phi(F) \in \ell^\infty([a,b])$.

[7] Since the infimum in the definition of ϕ is taken only within $[\tau_L, \tau_U]$.

Proof for other auxiliary lemmas

Proof of Lemma 3. Under Condition U, the SQ process $v_s = \mathrm{E}[Y \mid Y \geq q_s]$ is continuous in s, and in particular

$$\frac{\partial v_s}{\partial s} = \frac{v_s - q_s}{1-s} > 0, \quad s \in [\tau_L, \tau_U],$$

which indicates v_s is strictly increasing in s.

Next we show that the sample SQ $\hat{v}_s \leq \hat{v}_t$ for $\tau_L \leq s < t \leq \tau_U$. Without loss of generality, we can assume $\hat{q}_s < \hat{q}_t$, where $\hat{q}$ is the sample quantile; otherwise $\hat{v}_s = \hat{v}_t$. Let $m_1 = \sum_{i=1}^n \mathbf{1}\{Y_i \geq \hat{q}_t\}$ and $m_2 = \sum_{i=1}^n \mathbf{1}\{Y_i \geq \hat{q}_s\}$; by the choice of sample quantiles $\hat{q}_s$, we have $m_2 \geq m_1 > 0$. Hence

$$\begin{aligned}
\hat{v}_t - \hat{v}_s &= \frac{\sum_{i=1}^n Y_i \cdot \mathbf{1}\{Y_i \geq \hat{q}_t\}}{m_1} - \frac{\sum_{i=1}^n Y_i \cdot \mathbf{1}\{Y_i \geq \hat{q}_s\}}{m_2} \\
&= \frac{(m_2 - m_1)\sum_{i=1}^n Y_i \cdot \mathbf{1}\{Y_i \geq \hat{q}_t\} - m_1 \sum_{i=1}^n Y_i \cdot \mathbf{1}\{\hat{q}_t > Y_i \geq \hat{q}_s\}}{m_1 m_2} \\
&\geq \frac{\hat{q}_t(m_2 - m_1)\sum_{i=1}^n \mathbf{1}\{Y_i \geq \hat{q}_t\} - m_1 \hat{q}_t \sum_{i=1}^n \mathbf{1}\{\hat{q}_t > Y_i \geq \hat{q}_s\}}{m_1 m_2} \\
&\geq 0,
\end{aligned}$$

where the equality in the penultimate inequality holds if and only if $m_1 = m_2$. Therefore, $\hat{v}_s$ is non-decreasing with respect to s.

From its monotonicity, the one-sided limit of $\hat{v}_s$ from either the left or right exists. To show the continuity from the left, note that the quantile function $\hat{q}_s$ is left-continuous over $s \in (0,1)$, thus for any $s \in (\tau_L, \tau_U)$,

$$\lim_{\varepsilon \to 0+} \sum_{i=1}^n \mathbf{1}\{Y_i \geq \hat{q}_{s-\varepsilon}\} = \lim_{\varepsilon \to 0+} \sum_{i=1}^n \mathbf{1}\{Y_i \geq \hat{q}_s - \varepsilon\} = \sum_{i=1}^n \mathbf{1}\{Y_i \geq \hat{q}_s\} > 0,$$

$$\lim_{\varepsilon \to 0+} \sum_{i=1}^n Y_i \mathbf{1}\{Y_i \geq \hat{q}_{s-\varepsilon}\} = \sum_{i=1}^n Y_i \mathbf{1}\{Y_i \geq \hat{q}_s\}.$$

Since $\hat{v}_s$ is the ratio of the above displayed equations, we conclude that $\hat{v}_s$ is also

continuous from the left. □

Proof of Lemma 4. Define a class of functions $\mathcal{F} = \{y \mapsto \psi(y, \theta, s) : \theta \in [q_L, q_U], s \in [\tau_L, \tau_U]\}$. We shall show that $\mathcal{F}$ is a Donsker class of functions. First, note that $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \triangleq \{fg : f \in \mathcal{F}_1, g \in \mathcal{F}_2\}$, where

$$\mathcal{F}_1 = \{z \mapsto z - v_s : s \in [\tau_L, \tau_U]\}, \quad \mathcal{F}_2 = \{z \to \mathbf{1}[z \geq \theta] : \theta \in [q_L, q_U]\}.$$

Since $\mathcal{F}_1$ contains only linear functions and $\mathcal{F}_2$ contains only indicator functions of half lines, it is clear that both $\mathcal{F}_1$ and $\mathcal{F}_2$ are VC classes of functions, and therefore $\mathcal{F}$ also satisfy the uniform entropy condition. (See e.g. Example 19.19 of *Van der Vaart* (2000).) Next, let $F(y) = [|y| + |v_{\tau_U}| + |v_{\tau_L}|]\mathbf{1}\{y \geq q_L\}$, we can easily verify that $\sup_{f \in \mathcal{F}} |f(z)| \leq F(z)$ and $\mathrm{E}[F^2] < +\infty$ under Condition U, i.e., F is a suqare-integrable envelope function for $\mathcal{F}$. Therefore, we conclude that $\mathcal{F}$ is Donsker, which follows from Lemma 19.14 of *Van der Vaart* (2000).

Let $\mathbb{T} = [q_{\tau_L}, q_{\tau_U}] \times [\tau_L, \tau_U]$ be the product space equipped with the semimetric $\rho((\theta, s), (\theta', s')) = \{\mathrm{E}[\psi(Y^*; \theta, s) - \psi(Y^*; \theta', s')]^2\}^{1/2}$. As a consequence of Donskerness, the stochastic process $\mathbb{G}_n[\psi(Y^*; \theta, s)]$ indexed by (θ, s) is stochastically equicontinuous on $(\mathbb{T}, \rho)$, and that $(\mathbb{T}, \rho)$ is totally bounded.

Similar to Lemma 19.24 in *Van der Vaart* (2000), define the map

$$\begin{array}{rccccc} g: & \ell^\infty(\mathbb{T}) & \times & \ell^\infty([\tau_L, \tau_U]) & \mapsto & \mathbb{R} \\ & z(\cdot, \cdot) & \times & v(\cdot) & \mapsto & \sup_{s \in [\tau_L, \tau_U]} |z(v(s), s) - z(q_s, s)|. \end{array}$$

First, it is easy to verify that $g(\cdot, \cdot)$ is continuous (with respect to the product metric on $\ell^\infty(\mathbb{T}) \times \ell^\infty([\tau_L, \tau_U])$) at (z_0, v_0), as long as $z_0(\cdot, \cdot)$ is uniformly continuous over $(\mathbb{T}, \rho)$. Second, by its Donskerness, $\mathbb{G}_n(\psi(Y^*; \theta, s)) \xrightarrow{d} \mathbb{G}_\infty(\theta, s)$ in $\ell^\infty(\mathbb{T})$, where almost all sample paths of the limit $\mathbb{G}_\infty(\theta, s)$ is uniformly continuous on $(\mathbb{T}, \rho)$. Third, by assumption we have $\tilde{q}_s \xrightarrow{P^*} q_s$ on $\ell^\infty([\tau_U, \tau_L])$, which implies that the bivariate

process $[\mathbb{G}_n(\psi(Y^*;\theta,s)),\tilde{q}_s]$ also converges weakly. Hence by the continuous mapping theorem,

$$\begin{aligned}\sup_{s\in[\tau_L,\tau_U]}\left|\mathbb{G}_n[\psi_{(\tilde{q}_s,s)}]-\mathbb{G}_n[\psi_{(q_s,s)}]\right| &= g\circ\{\mathbb{G}_n[\psi(Y^*;\theta,s)],\tilde{q}_s\}\\ &= g\circ\{\mathbb{G}_n[\psi(Y^*;\theta,s)],\tilde{q}_s\}-g\circ\{\mathbb{G}_\infty[\psi(\theta,s)],q_s\}\\ &\xrightarrow{d} 0,\end{aligned}$$

since $g\{\mathbb{G}_\infty[\psi(\theta,s)],q_s\}=0$. Weak convergence to a constant then implies convergence in probability, which concludes the proof. □

2.5.3 Proof for the m-Rock approach

Here we give the proofs for the results in Section 2.1, as well as those in Chapter 1. For a matrix A, let $\lambda_{min}(A)$ be the minimal eigenvalue of A.

For the results in Section 2.1

Proof of Proposition 2. Part 1 of the Proposition follows from the properties of the check loss $\rho_\tau(\cdot)$ function. Note that

$$\begin{aligned}L_n(u_1)+L_n(u_2) &= \sum_{m=1}^{M}\left[\int_0^1\rho_\tau\left(\hat{v}_m(\alpha)-x_m^Tu_1\right)+\rho_\tau\left(\hat{v}_m(\alpha)-x_m^Tu_2\right)\,\mathrm{d}\alpha\right]\\ &\geq \sum_{m=1}^{M}\int_0^1\rho_\tau\left(\hat{v}_m(\alpha)-x_m^T(u_1+u_2)/2\right)\,\mathrm{d}\alpha\\ &\geq L_n\left(\frac{u_1+u_2}{2}\right).\end{aligned}\tag{2.19}$$

Moreover,

$$\begin{aligned}
|L_n(u_1) - L_n(u_2)| &= \left| \sum_{m=1}^{M} \int_0^1 \rho_\tau \left(\hat{v}_m(\alpha) - x_m^T u_1 \right) - \rho_\tau \left(\hat{v}_m(\alpha) - x_m^T u_2 \right) \, \mathrm{d}\alpha \right| \\
&\leq \sum_{m=1}^{M} \int_0^1 \left| \rho_\tau \left(\hat{v}_m(\alpha) - x_m^T u_1 \right) - \rho_\tau \left(\hat{v}_m(\alpha) - x_m^T u_2 \right) \right| \, \mathrm{d}\alpha \\
&\leq \sum_{m=1}^{M} \int_0^1 \left| x_m^T (u_1 - u_2) \right| \, \mathrm{d}\alpha \\
&\leq \sum_{m=1}^{M} \|x_m\| \cdot \|u_1 - u_2\|.
\end{aligned}$$

Thus, the convexity and Lipschitz continuity of $L_n(\theta)$ follows.

For Part 2 of the Proposition, the previous Lipschitz continuity implies we can exchange the order of integration and differentiability. Therefore

$$\begin{aligned}
\nabla_w L_n(u) &= \lim_{t \to 0+} \frac{L_n(u + tw) - L_n(u)}{t} \\
&= \sum_{m=1}^{M} \int_0^1 \nabla_w \rho_\tau (\hat{v}_m(\alpha) - x_m^T u) \, \mathrm{d}\alpha \\
&= -\sum_{m=1}^{M} x_m^T w \int_0^1 \psi_\tau^* \left(\hat{v}_m(\alpha) - x_m^T u, -x_m^T w \right) \, \mathrm{d}\alpha, \qquad (2.20)
\end{aligned}$$

where ψ_τ^* originates from the gradient condition of the check loss function, as in *Koenker* (2005, page 33):

$$\begin{aligned}
\psi_\tau^*(u, v) &= \begin{cases} \tau - \mathbf{1}\{u < 0\}, & \text{if } u \neq 0, \\ \tau - \mathbf{1}\{v < 0\}, & \text{if } u = 0. \end{cases} \\
&= \tau - \mathbf{1}\{u < 0\} - \mathbf{1}\{u = 0, v < 0\}.
\end{aligned}$$

We now prove Part 3, i.e., the optimality condition for the m-Rock estimator. By the convexity of L_n, any minimizer $\hat{\beta}$ of L_n must satisfy: $\nabla_w L_n(\hat{\beta}) \geq 0$, for all $w \in \mathbb{R}^p$, $\|w\| = 1$. Using the expression in (2.20), we can re-write the optimality

condition as

$$\begin{aligned} 0 &\geq \sum_{m=1}^{M} x_m^T w \int_0^1 \psi_\tau^* \left(\hat{v}_m(\alpha) - x_m^T \hat{\beta}, -x_m^T w \right) \mathrm{d}\alpha \\ &= \sum_{m=1}^{M} x_m^T w \left(\tau - \int_0^1 \mathbf{1}\{\hat{v}_m(\alpha) < x_m^T \hat{\beta}\} \, \mathrm{d}\alpha - \mathbf{1}\{x_m^T w > 0\} \int_0^1 \mathbf{1}\{\hat{v}_m(\alpha) = x_m^T \hat{\beta}\} \, \mathrm{d}\alpha \right). \end{aligned} \tag{2.21}$$

By the monotonicity of $\hat{v}_m(\alpha)$, each of the set $\{\alpha : \hat{v}_m(\alpha) < x_m^T \hat{\beta}\}$ is an interval on $[0, 1]$. By relating the integration to Lebesgue measure, we have

$$\begin{aligned} \int_0^1 \mathbf{1}\{\hat{v}_m(\alpha) < x_m^T \hat{\beta}\} \, \mathrm{d}\alpha &= 1 - \int_0^1 \mathbf{1}\{\hat{v}_m(\alpha) \geq x_m^T \hat{\beta}\} \, \mathrm{d}\alpha \\ &= 1 - Leb\left(\{\alpha \in (0,1) : \hat{v}_m(\alpha) \geq x_m^T \hat{\theta}\} \right) \\ &= \hat{h}_m(x_m^T \hat{\beta}), \end{aligned}$$

where $Leb(\cdot)$ is the Lebesgue measure on $\mathbb{R}$, and the last inequality follows from the definition of $\hat{h}_m(\cdot)$. Therefore, (2.21) implies that

$$\begin{aligned} \left\| \sum_{m=1}^{M} x_m \left[\tau - \hat{h}_m(x_m^T \hat{\beta}) \right] \right\|_2 &= \sup_{\|w\|=1} \left[\sum_{m=1}^{M} x_m^T w \left(\tau - \hat{h}_m(x_m^T \hat{\beta}) \right) \right] \\ &\leq \sup_{\|w\|=1} \left[\sum_{m=1}^{M} x_m^T w \mathbf{1}\{x_m^T w > 0\} \cdot \int_0^1 \mathbf{1}\{\hat{v}_m(\alpha) = x_m^T \hat{\beta}\} \, \mathrm{d}\alpha \right] \\ &\leq \sum_{m=1}^{M} \|x_m\| \cdot Leb\{\alpha \in [0,1] : \hat{v}_\alpha(x_m) = x_m^T \hat{\beta}\} \\ &\lesssim \frac{M^2}{n}, \end{aligned} \tag{2.22}$$

almost surely since the covariates are bounded; the last inequality follows since there are no ties among $Y_1, \ldots, Y_n$, and hence $Leb\{\alpha \in [0,1] : \hat{v}_m(\alpha) = x_m^T \hat{\theta}\} \leq M/n$. The proof is now complete.

□

Proof of Theorem II.1. First we prove the consistency part. For any $\varepsilon_0 > 0$, $\|\widehat{\beta} - \beta\| \geq 2\varepsilon_0$ implies that $\nabla_w L_n(\beta + \varepsilon_0 w) \leq 0$ for all $\|w\| = 1$, which follows from the convexity of the loss function. Using (2.20) and (2.22) in the proof of Proposition 2, the negativity of the directional derivative further implies

$$\inf_{\|w\|=1} \left[\sum_{m=1}^{M} x_m^T w \left\{ \hat{h}_m[x_m^T(\beta + \varepsilon_0 w)] - \tau \right\} \right] \leq \frac{C_0}{n},$$

with probability one for some universal constant C_0; Note that there are no ties in the data under Condition R-Y1 with probability 1.

For small enough ε_0, let

$$R_1(w) = \hat{h}_m[x_m^T(\beta + \varepsilon_0 w)] - h_m[x_m^T(\beta + \varepsilon_0 w)].$$

Lemma 2 shows that $R_1(w) = o_{\mathrm{P}}(1)$ uniformly over $\|w\| = 1$. Furthermore, since $h_m(z)$ is continuously differentiable and $h_m(x_m^T\beta) = \tau$, we have

$$h_m(z) - \tau = (z - x_m^T\beta) h_m'(x_m^T\beta) + o(|z - x_m^T\beta|).$$

Therefore, for sufficiently small $\varepsilon_0 > 0$,

$$\begin{aligned}
& \mathrm{P}\left(\|\widehat{\beta} - \beta\| \geq 2\varepsilon_0 \right) \\
\leq\ & \mathrm{P}\left(\inf_{\|w\|=1} \left\{ \sum_{m=1}^{M} x_m^T w \left[\hat{h}_m(x_m^T\beta + \varepsilon_0 x_m^T w) - \tau \right] \right\} \leq \frac{C_0}{n} \right) \\
\leq\ & \mathrm{P}\left(\varepsilon_0 \inf_{\|w\|=1} w^T \left[\sum_{m=1}^{M} x_m x_m^T h_m'(x_m^T\beta) \right] w - o(1) \leq \sup_w |R_1(w)| + \frac{C_0}{n} \right) \\
=\ & \mathrm{P}\left(\varepsilon_0 \left[M(1-\tau)\lambda_{\min}(\Omega_1) - o(1) \right] \leq \sup_w |R_1(w)| + o(1) \right) \\
\to\ & 0,
\end{aligned}$$

since $h_m'(x_m^T\beta) = (v_m - q_m)^{-1}(1 - \tau)$ and Ω_1 is positive definite; this concludes the

consistency of $\widehat{\beta}$.

Next we derive the asymptotic distribution of $\widehat{\beta}$. From Proposition 2, we have

$$\begin{aligned} O_{\mathrm{P}}\left(\frac{1}{n}\right) &= \sum_{m=1}^{M} x_m \left[\tau - \hat{h}_m(x_m^T\widehat{\beta})\right] \\ &= \underbrace{\sum_{m=1}^{M} x_m \left[\tau - \hat{h}_m(x_m^T\beta)\right]}_{R_1} + \underbrace{\sum_{m=1}^{M} x_m \left[h_m(x_m^T\beta) - h_m(x_m^T\widehat{\beta})\right]}_{R_2} \\ &\quad + \underbrace{\sum_{m=1}^{M} x_m \left\{[\hat{h}_m(x_m^T\beta) - h_m(x_m^T\beta)] - [\hat{h}_m(x_m^T\widehat{\beta}) - h_m(x_m^T\widehat{\beta})]\right\}}_{R_3} \end{aligned} \quad (2.23)$$

We consider the three terms R_1 through R_3 separately in the following.

By Lemma 2, $\hat{h}_m(z)$ is asymptotically Gaussian at $z = v_m(\tau) = x_m^T\beta$ for each covariate value x_m, therefore:

$$\sqrt{\frac{n}{M}}\left[\hat{h}_m(x_m^T\beta) - h_m(x_m^T\beta)\right] \xrightarrow{d} \mathrm{N}\left(0,\ \frac{(1-\tau)^2\sigma_m^2}{(v_m - q_m)^2}\right).$$

Therefore, summing the above equation over m gives

$$\frac{\sqrt{n}}{M}R_1 \xrightarrow{d} \mathrm{N}\left[0,\ (1-\tau)^2\Omega_1\right],$$

where Ω_1 is defined in Theorem II.1. For the term R_2, Taylor expansion of h_m gives

$$\begin{aligned} \frac{R_2}{M} &= -\frac{1}{M}\sum_{m=1}^{M} x_m \left[x_m^T(\widehat{\beta} - \beta)h_m'(x_m^T\beta) + o_{\mathrm{P}}(\|\widehat{\beta} - \beta\|)\right] \\ &= -\left[(1-\tau)D_1 + o_{\mathrm{P}}(1)\right](\widehat{\beta} - \beta), \end{aligned}$$

since $h_m(\cdot)$ is continuously differentiable. For R_3, the asymptotic equi-continuity in

Lemma 2 shows that

$$\frac{R_3}{M} = \frac{1}{M}\sum_{m=1}^{M}\left[x_m o_{\mathrm{P}}\left(\sqrt{\frac{M}{n}}\right)\right] = o_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right),$$

since $\widehat{\beta}$ is consistent for β.

Therefore, substituting R_1, R_2 and R_3 back into (2.23) gives

$$\sqrt{n}\left[(1-\tau)D_1 + o(1)\right](\widehat{\beta}-\beta) \xrightarrow{d} \mathrm{N}\left[0,\ (1-\tau)^2\Omega_1\right],$$

which implies

$$\sqrt{n}(\widehat{\beta}-\beta) \xrightarrow{d} \mathrm{N}\left(0,\ D_1^{-1}\Omega_1 D_1^{-1}\right).$$

The proof is now complete.

□

For the results in Chapter 1 Here we prove the results in Chapter 1, where we use the same notations therein. For two matrices A and B, we write $A \succeq B$ if $A - B$ is positive semi-definite.

Proof of Proposition 1. The second equality in Equation (1.11) of the Proposition is straightforward; we only prove the first one. Note that $v_{[Y-C]}(\alpha) = v_Y(\alpha) - C$. Hence the loss function in (1.5) can be written as

$$\begin{aligned} L_\tau(C) - \int_0^1 v_{[Y]}(\alpha)\,\mathrm{d}\alpha &= \frac{1}{1-\tau}\int_0^1 (1-\tau)[C - v_{[Y]}(\alpha)] + \max\{0, v_{[Y]}(\alpha) - C\}\,\mathrm{d}\alpha \\ &= \frac{1}{1-\tau}\int_0^1 \rho_\tau(v_{[Y]}(\alpha) - C)\,\mathrm{d}\alpha, \end{aligned}$$

which follows from standard algebra. □

Proof of Theorem I.1. Letting

$$C_0 = \mathrm{E}_X \left[\int_0^1 v_{[Y|X]}(\alpha, x)\, \mathrm{d}\alpha \right],$$

Equation (1.13) then follows directly from Proposition 1 and the linearity of the operator $\mathrm{E}_X(\cdot)$. We prove the identification of β and the uniqueness of the minimizer using (1.13).

We first show that the function $g(x, u)$ is strictly increasing in all $u \in (0, 1)$ for each possible value of x. Note $g(x, u)$ is continuous and (weakly) monotonic by construction (1.3). Since the response distribution is continuous, it follows that $q_{[Y|X]}(u, x)$ is strictly increasing in u and almost everywhere continuous; Therefore

$$\begin{aligned} \frac{\partial g(x,u)}{\partial u} &= \frac{v_{[Y|X]}(u,x) - q_{[Y|X]}(u,x)}{1-u} \\ &= \frac{1}{(1-u)^2} \int_u^1 \left[q_{[Y|X]}(s,x) - q_{[Y|X]}(u,x) \right] \mathrm{d}u \\ &> 0, \end{aligned} \tag{2.24}$$

almost everywhere in u. Hence the strict monotonicity of $g(x, u)$ takes hold.

Now we prove the optimality of β for the function $L(\theta)$. From the property of the check-loss function, it follows that the function $L(\theta)$ is convex and differentiable in θ; See, e.g., *Koenker* (2005, Chapter 1.3). In fact, the derivative is

$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \tau \mathrm{E}_X \left[\Pr\left(g(X,\xi) \geq X^T\theta \mid X \right) \cdot X \right] - (1-\tau) E_X \left[\Pr\left(g(X,\xi) < X^T\theta \mid X \right) \cdot X \right] \\ &= E_X \left\{ \left[\tau - \Pr\left(g(X,\xi) < X^T\theta \mid X \right) \right] \cdot X \right\}. \end{aligned}$$

Note $x^T\beta = g(x, \tau) = v_{[Y|X]}(\tau, x)$, hence the optimality of β follows from the first-order condition

$$\Pr\left(g(x,\xi) < x^T\beta \mid X = x \right) = \Pr\left(g(x,\xi) < g(x,\tau) \mid X = x \right) = \Pr(\xi < \tau) = \tau,$$

for all x, as $g(x,u)$ is strictly increasing in all $u \in (0,1)$.

Next we show the minimizer of L is unique. Let $g^{-1}(x,z)$ be the inverse of g, such that $g \circ (x, g^{-1}(x,z)) = z$. By the conditions in Theorem I.1, $g(x,u)$ is differentiable in u for all $|u-\tau| \le c_0$, with derivative given by (2.24); Hence we have

$$\left.\frac{\partial \Pr(g(x,\xi) \le z)}{\partial z}\right|_{z=g(x,\tau)} = \left[\left.\frac{\partial g(x,s)}{\partial s}\right|_{s=\tau}\right]^{-1} \ge \delta_0 > 0,$$

for some constant δ_0 uniformly in all x. Therefore, it follows that $L(\theta)$ is twice differentiable, and the derivative satisfies

$$\left.\frac{\partial^2 L}{\partial\theta\partial\theta^T}\right|_{\theta=\beta} = \mathrm{E}_X\left[\left.\frac{\partial \Pr(g(X,\xi) \le z)}{\partial z}\right|_{z=X^T\beta} \cdot XX^T\right] \succeq \delta_0 \cdot \mathrm{E}_X\left[XX^T\right].$$

Therefore, the Hessian matrix of $L(\cdot)$ evaluated at β is positive definite, establishing the uniqueness of the minimizer β.

□

2.5.4 Proof of other results

2.5.4.1 For the Linearization method

Proof of Theorem II.2. By the SQ regression model (1.4), we have $v_m(\tau) = x_m^T\beta$ for all $m = 1, \ldots, M$. Hence we can rewrite

$$\sqrt{n}\left(\widehat{\beta}^{(L)} - \beta\right) = \left(\frac{1}{M}\sum_{m=1}^{M} x_m x_m^T\right)^{-1}\left[\frac{\sqrt{n}}{M}\sum_{m=1}^{M} x_m\left\{\hat{v}_m(\tau) - v_m(\tau)\right\}\right]. \tag{2.25}$$

Theorem II.4 implies that

$$\sqrt{\frac{n}{M}}\left\{\hat{v}_m(\tau) - v_m(\tau)\right\} \xrightarrow{d} \mathrm{N}\left(0, \sigma_m^2\right),$$

and therefore

$$\frac{\sqrt{n}}{M}\sum_{m=1}^{M} x_m \left\{\hat{v}_m(\tau) - v_m(\tau)\right\} \xrightarrow{d} \mathrm{N}\left\{0,\ \frac{1}{M}\sum_{m=1}^{M} x_m x_m^T \sigma_m^2\right\},$$

since $\hat{v}_m(\tau)$ only involves the data at x_m, they are independent across $m = 1, \ldots, M$. The proof is complete by substituting the above displayed equation into Equation (2.25). □

2.5.4.2 For the Two-Step method

Proof of Theorem II.3. The proof follows from standard M-estimation framework; See, e.g., *Van der Vaart* (2000, Section 5). Here we give a more direct proof. Let $\tilde{q}_m = x_m^T \widehat{\beta}_q$ be the linear quantile regression estimator for q_m in (2.8). At each covariate value x_m, we define an estimator for the SQ as

$$\tilde{v}_m = \frac{\sum_{j=1}^{n_m} Y_{mj} \mathbf{1}\{Y_{mj} \geq \tilde{q}_m\}}{\hat{w}_m},$$

wher $\hat{w}_m = \sum_{j=1}^{n_m} \mathbf{1}\{Y_{mj} \geq \tilde{q}_m\}/n_m$ and $n_m = n/M$.

From the estimating equation (2.8), $\widehat{\beta}^{(TS)}$ is the solution to a weighted least sqaures equation, and hence we can express it in close form:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta}^{(TS)} - \beta\right) &= \left(\frac{1}{M}\sum_{m=1}^{M} \hat{w}_m x_m x_m^T\right)^{-1} \left(\frac{\sqrt{n}}{M}\sum_{m=1}^{M} \hat{w}_m x_m [\tilde{v}_m - v_m]\right) \\
&= \left(\frac{1-\tau}{M}\sum_{m=1}^{M} x_m x_m^T + o_{\mathrm{P}}(1)\right)^{-1} \qquad (2.26) \\
&\quad \cdot \left(\frac{\sqrt{n}(1-\tau)}{M}\sum_{m=1}^{M} x_m [\tilde{v}_m - v_m] + o_{\mathrm{P}}(1)\right), \qquad (2.27)
\end{aligned}$$

the last equality follows since $\tilde{q}_m = q_m + o_{\mathrm{P}}(1)$ and therefore $\hat{w}_m = (1-\tau) + o_{\mathrm{P}}(1)$

for each m.

By (2.18) in the proof of Theorem II.4, each $\tilde{v}_m$ has the following representation

$$\begin{aligned}\tilde{v}_m - v_m &= \frac{1}{1-\tau}\left\{(\tilde{q}_m - q_m)(v_m - q_m)f_m(q_m) + \frac{1}{n_m}\sum_{j=1}^{n_m}[Y_{mj} - v_m]\mathbf{1}\{Y_{mj} \geq q_m\}\right\} \\ &\quad + o_{\mathrm{P}}\left(\frac{1}{\sqrt{n_m}}\right).\end{aligned}$$

Substituting the above expansion into the second factor of (2.27), we have

$$\begin{aligned}\frac{\sqrt{n}(1-\tau)}{M}\sum_{m=1}^{M} x_m[\tilde{v}_m - v_m] &= \left[\frac{\sqrt{n}}{M}\sum_{m=1}^{M} x_m x_m^T (v_m - q_m)f_m(q_m)\right](\widehat{\beta}_q - \beta_q) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{m=1}^{M} x_m \sum_{j=1}^{n_m}[Y_{mj} - v_m]\mathbf{1}\{Y_{mj} \geq q_m\} + o_{\mathrm{P}}(1) \\ &= (1-\tau)G\left[\frac{1}{\sqrt{n}}D_2^{-1}\sum_{m=1}^{M} x_m \sum_{j=1}^{n_m}(\tau - \mathbf{1}\{Y_{mj} < q_m\})\right] \\ &\quad + \frac{1}{\sqrt{n}}\sum_{m=1}^{M} x_m \sum_{j=1}^{n_m}[Y_{mj} - v_m]\mathbf{1}\{Y_{mj} \geq q_m\} + o_{\mathrm{P}}(1) \\ &\xrightarrow{d} \mathrm{N}\left\{0, (1-\tau)^2 G\Sigma G + (1-\tau)^2 V\right\},\end{aligned}$$

where the first equality holds since $\tilde{q}_m - q_m = x_m^T(\widehat{\beta}_3 - \beta_q)$, and the second equation follows from the classic Bahadur representation of the quantile regression estimator (see e.g., Chapter 4 of *Koenker* (2005)), and the weak convergence follows from the Central Limit Theorem; Refer to the statement of Theorem II.3 for the definition of G, Σ and V.

Combining the above displayed equation with (2.27), we obtain

$$\sqrt{n}\left(\widehat{\beta}^{(TS)} - \beta\right) \xrightarrow{d} \mathrm{N}\left\{0, \frac{1}{1-\tau}D_0^{-1}\left(G\Sigma G + V\right)D_0^{-1}\right\},$$

which finishes the proof. □

2.5.4.3 Proof for Equation (2.12)

Proof. Recall that $0 < x_1 < \ldots < x_M$ and $\mu_j = \sum_{m=1}^{M} x_m^j / M$. The result is a direct application of the Hölder's inequality. Note

$$\mu_2 = M^{-1} \sum_{m=1}^{M} \left| x_m^{4/3} \cdot x_m^{2/3} \right| \leq \left(M^{-1} \sum_{m=1}^{M} |x_m|^{4p/3} \right)^{1/p} \cdot \left(M^{-1} \sum_{m=1}^{M} |x_m|^{2q/3} \right)^{1/q}$$

for any $1/p + 1/q = 1$. Taking $p = 3$ and $q = 3/2$ in the above shows that

$$\mu_2^3 \leq \mu_4 \times \mu_1^2,$$

therefore

$$R_1 = \frac{\mu_2}{\mu_1^2} \ \leq \ \frac{\mu_4}{\mu_2^2} = R_2.$$

For our setting under Model (2.11), the above inequality is strict because $x_m^{4/3}$ and $x_m^{2/3}$ cannot be proportional to each other across $m = 1, \ldots, M$.

□

CHAPTER III

The m-Rock Approach with General Covariates

In this chapter, we extend the m-Rock approach to the case with general covariate distributions. While the same principles of Chapter 2 apply, there are several significant challenges, both practically and theoretically, for applying the m-Rock approach with continuous covariates. We study the asymptotic properties of the m-Rock estimator based on binning of the covariate space. The main focus of this chapter is the theoretical investigation of the m-Rock approach, and we demonstrate its benefits over other common approaches in the literature via asymptotic efficiency comparisons.

Following from the general formula (1.13) in Chapter 1, the key challenge for the m-Rock approach is to obtain an initial estimator for the unknown conditional SQ process. When the covariates are discrete, taking the sample SQ at each covariate value suffices in Chapter 2. With continuous covariates, we start with a general analysis that can incorporate a broad class of non-parametric initial SQ estimators; Under appropriate technical conditions, we show that the m-Rock approach is asymptotically equivalent to a weighted linearization of those initial SQ estimators. Next, we show that the local-linear estimator in the spirit of *Olma* (2021) can be used as an example of the initial estimator, and we characterize the precise asymptotic distribution for the resulting m-Rock estimator. To deal with continuous covariates, the theoretical analysis in this chapter is much more involved than those in Chapter 2.

For simplicity, we shall assume that all the covariates have a continuous distribution in our theoretical analysis. Our discussion in this chapter can be easily extended to the setting where we have both discrete and continuous covariates.

3.1 The binning method

For theoretical analysis with continuous covariates, we rely on the idea of *binning* to partition the continuous sample space into local sub-spaces, which we call *bins*. Within each bin, we can obtain an initial estimator for the conditional SQ, based on which we can implement the m-Rock approach in a way similar to Chapter 2. Overall, the binning method effectively discretizes the sample space, which facilitates our theoretical analysis.

In general, binning regression has been a popular practical tool to summarize the data (*Starr and Goldfarb*, 2020), and recently *Cattaneo et al.* (2019) gives a comprehensive theoretical analysis of its properties. However, the analysis in *Cattaneo et al.* (2019) does not apply to our setting because (i) it is restricted to one-dimensional covariates, and (ii) it does not cover the SQ process convergence. In the following, we develop new asymptotic theories for the m-Rock approach under the binning method, where we allow covariates of multiple dimensions.

We formally define the binning procedure as follows. Suppose the data $\{(X_i, Y_i) : i = 1, \ldots, n\}$ is a random sample from the distribution $(X, Y) \sim \Pr$, where $X \in \mathbb{R}^{p+1}$ includes p covariates and an intercept term. Let $\mathcal{X} \subset \mathbb{R}^{p+1}$ be the sample space of the covariate, and we partition

$$\mathcal{X} = \bigcup_{m=1}^{M} A_m,$$

where $A_1, \ldots, A_M$ are non-stochastic, disjoint bins, and the number $M = M_n$ may depend on the sample size n. For each $m = 1, \ldots, M$, let $\bar{x}_m \in \mathbb{R}^{p+1}$ be the geometric center of A_m, i.e., $\bar{x}_m = \int x \mathbf{1}\{x \in A_m\} \, \mathrm{d}x$; Note $\bar{x}_m$ is also non-stochastic. With

$M_n \to +\infty$, the conditional SQ function $v(\tau, x)$ can be approximated by the SQ in each bin. In the following, we shall omit the index n in M_n for simplicity.

The m-Rock approach can be implemented with an initial SQ estimator in each bin; in the following, we first give a general analysis that does not depend on a specific choice of initial estimator. Within each bin A_m, let $\hat{v}(\alpha, \bar{x}_m)$ be a binning estimator of $v(\alpha, \bar{x}_m)$, for a range of $\alpha \in (0, 1)$; the estimator should only use the data within the bin A_m. Furthermore, let $\hat{\gamma}_m$ be a suitable weight for each bin A_m that only depends on the covariates; we shall discuss the choice of $\hat{\gamma}_m$ later. Parallel to the implementation in Chapter 2, the m-Rock estimator can be obtained by:

$$\widehat{\beta} = \arg\min_{u \in \mathbb{R}^{p+1}} \sum_{m=1}^{M} \hat{\gamma}_m \int_0^1 \rho_\tau \left(\hat{v}(\alpha, \bar{x}_m) - \bar{x}_m^T u \right) \mathrm{d}\alpha. \tag{3.1}$$

In terms of computation, we can approximate the integration in (3.1) by a fine grid of quantile levels over $\alpha \in (0, 1)$, then solve the optimization problem via quantile regression, similar to (2.2) in Chapter 2. We relegate more computational details in Chapter 4.

There are several options when defining the m-Rock estimator (3.1) that we have not yet specified. First, we do not focus on a specific estimator $\hat{v}(\alpha, \bar{x}_m)$. Our analysis here works for a class of binning estimators that satisfy certain technical conditions given in the next section. Second, our analysis do depend on the exact shape or construction of the bins A_m; later we discuss some necessary conditions for the bin size. Third, we do not specify the choice of $\hat{\gamma}_m$. Intuitively, the purpose of those weights is to adjust for the difference in sample sizes across the bins; In practice, $\hat{\gamma}_m$ may depend on the construction of initial estimators, and we give one example later in Section 3.4.

3.2 High-level technical conditions

Since the m-Rock approach builds upon a set of initial SQ estimators, its properties would depend on that for those SQ estimators. Here we give some technical conditions on both the data generating process and the initial SQ estimators. We set τ to be the fixed quantile level of interest. Let $f_{Y|X}(y;x)$ be the conditional density function of $Y \mid X = x$, and let $q(s,x)$ and $v(s,x)$ be the conditional quantile and SQ function of $Y \mid X = x$, respectively.

We first give the following regularity conditions regarding the covariate and response distributions, which are similar to those in Section 2.1.2 for the case with discrete covariates.

Condition G-X. The covariates have bounded support $\mathcal{X} \subset \mathbb{R}^{p+1}$, and have a density function $f_X(x)$ that is uniformly bounded away from 0 and $+\infty$. Furthermore, the matrix D_1 is positive definite, where

$$D_1 = \mathrm{E}\left[\frac{XX^T}{v(\tau,X) - q(\tau,X)}\right].$$

Condition G-Y1. At each x, $f_{Y|X}(y;x)$ is continuous over y. Furthermore, there exist constants $\underline{f}$, $\overline{f}$, and $\varepsilon_0 > 0$, such that

$$0 < \underline{f} \leq \inf_{\substack{(x,y):x\in\mathcal{X} \\ |y-q(\tau,x)|\leq\varepsilon_0}} f_{Y|X}(y;x) \leq \sup_{\substack{(x,y):x\in\mathcal{X} \\ |y-q(\tau,x)|\leq\varepsilon_0}} f_{Y|X}(y;x) \leq \overline{f}.$$

Condition G-Y2. For each x, both $q(s,x)$ and $v(s,x)$ are strictly increasing and continuous over $s \in (0,1)$. Furthermore, both $q(\tau,x)$ and $v(\tau,x)$ are Lipschitz continuous over $x \in \mathcal{X}$.

We briefly discuss the conditions above. First, we require the covariates to be bounded in Condition G-X to simplify the technical derivations; the condition may be relaxed at the cost of more complicated proofs. Second, Conditions G-Y1 and G-

Y2 have several further implications on the data generating process; E.g., they imply the differentiability of $v(s,x)$ with respect to s, and that $v(s,x)-q(s,x)$ is uniformly bounded over both s and x. We give more details later in Lemma 6 of Section 3.7.1.

Next, we require the following high-level technical conditions for the binning SQ estimators used in the m-Rock approach.

Condition G-V1. For any $m = 1,\ldots,M$, the initial SQ estimator $\hat{v}(s,\bar{x}_m)$ is left-continuous and non-decreasing in $s \in (0,1)$. Furthermore, for any constant $B > 0$ and some sequence $r_n = o(n^{-1/4})$, those estimators satisfy:

1. $$\sup_{\substack{m=1,\ldots,M \\ s:|s-\tau|\leq B\cdot(r_n+n^{-1/2})}} |\hat{v}(s,\bar{x}_m) - v(s,\bar{x}_m)| = O_{\mathrm{P}}(r_n),$$

2. $$\sup_{\substack{m=1,\ldots,M \\ s:|s-\tau|\leq B\cdot(r_n+n^{-1/2})}} |[\hat{v}(s,\bar{x}_m) - v(s,\bar{x}_m)] - [\hat{v}(\tau,\bar{x}_m) - v(\tau,\bar{x}_m)]| = o_{\mathrm{P}}\left(n^{-1/2}\right).$$

Condition G-V2. The weighted aggregation of the initial SQ estimators satisfies:

$$\sum_{m=1}^{M}\left[\frac{\hat{\gamma}_m\bar{x}_m}{v(\tau,\bar{x}_m)-q(\tau,\bar{x}_m)}\{\hat{v}(\tau,\bar{x}_m)-v(\tau,\bar{x}_m)\}\right] = O_{\mathrm{P}}\left(n^{-1/2}\right).$$

Condition G-V2 requires that the τth initial SQ estimators can be aggregated over the bins, and that the resulting statistic enjoys a $\sqrt{n}$-rate of convergence. Condition G-V1 is more technical, and is about the uniform consistency and asymptotic equi-continuity of the estimated SQ process. Similar conditions are commonly used in the empirical process literature to ensure process convergence (*Van Der Vaart and Wellner*, 1996, Section 3.3). Common to those assumptions is that they only concern $\hat{v}(s,x)$ when s is in a local neighbourhood of τ. In fact, our analysis for the m-Rock approach does not depend on the properties of $\hat{v}(s,\bar{x}_m)$ in the tails when s is close to 0 or 1, as long as $\hat{v}(s,x)$ is monotonic in s. In particular, $\hat{v}(s,\bar{x}_m)$ does not need to converge uniformly over $s \in (0,1)$. Later in Section 3.4, we show those conditions can be satisfied with a practical estimator.

Outside of the technical parts in Conditions G-V1 and G-V2, the other relatively restrictive assumption is the monotonicity of $\hat{v}(s, \bar{x}_m)$ over $s \in (0, 1)$. For many typical estimators, $\hat{v}(s, \bar{x}_m)$ may not be smooth and monotonic in their usual finite-sample constructions. Here we show how monotonicity can be reached in practice. In the quantile regression literature, there are various approaches to enforce monotonicity in the estimation of monotone functions, see, e.g., *He* (1997); *Chernozhukov et al.* (2010); *Dette and Volgushev* (2008). In particular, we can use the re-arrangement approach in *Chernozhukov et al.* (2010) to monotonize a given initial SQ estimator without jeopardizing our theoretical analysis. Therefore, monotonicity of $\hat{v}(s, \bar{x}_m)$ can be safely assumed, and we do not explicitly discuss this condition in our subsequent analysis.

Remark 3. In fact, we believe our main result does not rely on the monotonicity of $\hat{v}(s, \bar{x}_m)$; this assumption is more of a proof artifact to simplify the technical derivations. Some further discussions on what can be done without monotonicity are relegated to Section 3.7.6.1. In the following of our thesis, we keep the monotonicity requirement to make the presentation concise.

Remark 4. Since Conditions G-V1 and G-V2 only concern the behaviour of $\hat{v}(s, \bar{x}_m)$ when s is near τ, there are two possible simplifications in the m-Rock estimation procedure. First, we can use Winsorization (*Wilcox*, 2005) to construct the initial estimator $\{\hat{v}(s, x) : s \in (0, 1)\}$. Specifically, we calculate $\hat{v}(s, x)$ only for s near τ, and then extrapolate with a constant into the upper and/or lower tails. Second, parallel to Corollary 1 of Chapter 1, we can use a truncated range of integration in the m-Rock loss function (3.1). These simplifications eliminate the need for initial SQ estimation at extreme tails. We give more discussions on the practical implementation in Chapter 4.

3.3 Main result

Here we present the main theoretical result of this chapter, which characterizes how $\widehat{\beta}$ links to the initial SQ estimators. Recall from (3.1) that $\hat{\gamma}_m$ is a weight for each bin in the m-Rock approach, and that the number of bins M may depend on the sample size. Let $\text{diam}(\cdot)$ be the diameter of a set in $\mathbb{R}^{p+1}$, and let $\hat{\pi}_m = n^{-1}\sum_{i=1}^{n} \mathbf{1}[X_i \in A_m]$ be the proportion of data that fall into the bin A_m.

Theorem III.1. *Suppose Conditions G-X, G-Y1 and G-Y2 hold. In addition, let the binning mechanism satisfy*

$$\sup_{m=1,\ldots,M} diam(A_m) = o(1), \qquad \sup_{m=1,\ldots,M} \left| \frac{\hat{\gamma}_m}{\hat{\pi}_m} - 1 \right| = o_P(1).$$

Given any initial SQ estimator that satisfies Conditions G-V1 and G-V2, the m-Rock estimator in (3.1) would satisfy:

$$\left(\widehat{\beta} - \beta\right) = D_1^{-1} \sum_{m=1}^{M} \left[\frac{\hat{\gamma}_m \bar{x}_m}{v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)} \{\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)\} \right] + o_P(n^{-1/2}),$$

where D_1 is given in Condition G-X. In particular, $\widehat{\beta}$ is $\sqrt{n}$-consistent for β.

Theoretically, Theorem III.1 shows that $\widehat{\beta}$ is asymptotically equivalent to a weighted linearization of $\hat{v}(\tau, \bar{x}_m)$ over each bin: it turns a set of non-parametric initial estimators to a parametric estimator using the m-Rock loss function (3.1). To better understand our asymptotic result, consider the following alternative way of linearization:

$$\widetilde{\beta} = \min_{u \in \mathbb{R}^{p+1}} \sum_{m=1}^{M} \hat{\gamma}_m w_m \left(\hat{v}(\tau, \bar{x}_m) - \bar{x}_m^T u \right)^2, \tag{3.2}$$

where w_m is a set of known weights. Simply put, (3.2) is a weighted least-squares

(WLS) of the initial τth SQ estimator on the covariates, and $\widetilde{\beta}$ satisfies

$$\sqrt{n}\left(\widetilde{\beta}-\beta\right)=\left(\sum_{m=1}^{M}\hat{\gamma}_m w_m \bar{x}_m \bar{x}_m^T\right)^{-1}\sum_{m=1}^{M}\left[\hat{\gamma}_m w_m \bar{x}_m\{\hat{v}(\tau,\bar{x}_m)-v(\tau,\bar{x}_m)\}\right].$$

Theorem III.1 shows that the m-Rock estimator is asymptotically equivalent to (3.2) with $w_m=[v(\tau,\bar{x}_m)-q(\tau,\bar{x}_m)]^{-1}$.

Remark 5. It is important that the m-Rock approach is only effective with non-linear initial estimators. To see this, consider the case when the initial SQ estimators are linear-in-covariates, i.e., $\hat{v}(\tau,\bar{x}_m)=\bar{x}_m^T\hat{\xi}$ for some $\hat{\xi}$. Noting that $v(\tau,\bar{x}_m)=\bar{x}_m^T\beta$, it then follows from Theorem III.1 that:

$$\begin{aligned}\widehat{\beta}-\beta &= D_1^{-1}\left[\sum_{m=1}^{M}\frac{\hat{\gamma}_m}{v(\tau,\bar{x}_m)-q(\tau,\bar{x}_m)}\bar{x}_m\bar{x}_m^T\right](\hat{\xi}-\beta)+o_{\mathrm{P}}\left(n^{-1/2}\right)\\ &= \hat{\xi}-\beta+o_{\mathrm{P}}\left(n^{-1/2}\right),\end{aligned}$$

since the summation in the bracket converges towards D_1. Asymptotically, the m-Rock approach would make no change to the initial linear SQ estimator $\hat{\xi}$.

3.3.1 Benefits of m-Rock: semi-efficient weight

A simpler way to linearize the initial SQ estimators is to use (3.2) with $w_m\equiv 1$; this is the Linearization method of Chapter 2 when the covariates are discrete. Specifically, the Linearization method only involves the τth SQ, while the m-Rock approach relies on the initial SQ estimator $\hat{v}(s,\bar{x}_m)$ for a range of s. In the following, we explain how the m-Rock approach uses the initial SQ process to achieve better efficiency over the Linearization method.

For the sake of theoretical illustration, it suffices to compare two different weights in (3.2): using $w_m=[v(\tau,\bar{x}_m)-q(\tau,\bar{x}_m)]^{-1}$ corresponds to the m-Rock approach, and $w_m\equiv 1$ for the Linearization method. Among the class of estimators (3.2),

the optimal weight should reflect the heterogeneity of the initial estimators, i.e., $w_m^* \propto \text{var}^{-1}[\hat{v}(\tau, \bar{x}_m)]$; See, e.g., (*Wooldridge*, 2010, Section 7). Even though the effective weights for the m-Rock approach are not optimal in general, we show that they tend to be closely related to the optimal weights w_m^*.

For general non-parametric SQ estimation, *Olma* (2021) shows that many SQ estimators has asymptotic variance in the form of:

$$\begin{aligned} a_n \text{var}[\hat{v}(\tau, \bar{x}_m) \mid X = \bar{x}_m] =& \rho_1 \text{var}[Y \mid X = \bar{x}_m, Y \geq q(\tau, \bar{x}_m)] \\ &+ \rho_2 [v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)]^2 + o_{\text{P}}(1), \end{aligned} \tag{3.3}$$

where a_n is the scaling factor, and ρ_1, ρ_2 are two constants depending on the construction of $\hat{v}(\tau, x)$. Therefore, the m-Rock weight $w_m = [v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)]^{-1}$ captures part of the variance in (3.3), and hence can be similar to the optimal weight w_m^*. In fact, we shall demonstrate in Section 3.5 that the two additive components in (3.3) are often proportional to each other across x_m; In those situations, the m-Rock weights satisfy $w_m \propto (w_m^*)^{1/2}$, and hence are partially adaptive to heterogeneity. Therefore, the m-Rock approach can often be more efficient than the simple Linearization approach, since the latter ignores any heterogeneity in the data. We relegate more detailed asymptotic efficiency comparisons to Section 3.5.

3.3.2 Benefits of m-Rock: automatic weighting

Motivated from the m-Rock weights, one may also consider a direct weighted Linearization in (3.2) using $w_m = [v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)]^{-1}$. However, the weights w_m are unknown and therefore the WLS approach is infeasible. Using some estimated weights $\hat{w}_m$ for (3.2), the resulting feasible WLS estimator may be unstable if the weights are not estimated well; See, e.g., Section 3.4.1 of *Angrist and Pischke* (2010). We also provide numerical evidence in Section 4.2.2 of Chapter 4. On the contrary, the m-Rock approach does not require estimating any weight, yet its asymptotic property

is the same as if the weight was known. The theoretical weighting is implicit and achieved automatically, which is another important feature of our method.

Central to the understanding of our approach is how those weights come into play. Recall from (3.1) that we require $\hat{v}(s, \bar{x}_m)$ for a range of s in the m-Rock approach. The key is that we borrow information from nearby quantile levels. The m-Rock weights can be written as:

$$w_m^{-1} \propto \frac{v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)}{1-\tau} = \left.\frac{\partial v(s, \bar{x}_m)}{\partial s}\right|_{s=\tau}, \quad m = 1, \ldots, M.$$

Heuristically, by invoking the initial SQ estimators at levels near τ, the m-Rock approach can implicitly approximate the derivative of $\hat{v}(s, \bar{x}_m)$, which leads to the automatic weighting in the m-Rock approach.

3.4 An example of initial estimator

In this section, we provide one concrete example of constructing the initial SQ estimator, and we show that it satisfies the technical conditions of Theorem III.1. Using this initial estimator, we can characterize the asymptotic normality of the resulting m-Rock estimator.

3.4.1 Neyman-orthogonalized local-linear estimation

Our construction is a bin-wise linear SQ estimator that uses a Neyman-orthogonalized score function. Operationally, we fit a linear SQ regression using the data within each bin, and those estimator can be viewed as an example of local-linear estimation (*Fan*, 1992; *Fan and Gijbels*, 2018) with rectangular kernels. For the SQ regression in each bin, we use the Neyman-orthogonalized least-squares regression as in *Barendse* (2020) and *Olma* (2021).

We fix some notations first. In this section, we separate the intercept term from

the covariates and write $X = (1, \tilde{X}^T)^T$ and $\bar{x}_m = (1, \tilde{x}_m^T)^T$. Let $\hat{q}(s, x)$ be an estimator for the conditional quantile function of $Y \mid X = x$, and we define

$$Z_i(s, \theta) = \frac{(Y_i - \theta)\mathbf{1}[Y_i \geq \theta]}{1 - s} + \theta. \tag{3.4}$$

For notational simplicity, in the following we shall write $Z_i(s) = Z_i(s, q(s, X_i))$ and $\hat{Z}_i(s) = Z_i(s, \hat{q}(s, X_i))$. For each bin A_m, we define

$$\boldsymbol{X}_m = \begin{bmatrix} & (\tilde{X}_1 - \tilde{x}_m)^T \\ \mathbf{1}_n, & \vdots \\ & (\tilde{X}_n - \tilde{x}_m)^T \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}, \quad \boldsymbol{W}_m = \mathrm{diag}\{w_{1m}, \ldots, w_{nm}\} \in \mathbb{R}^{n \times n},$$

where $w_{im} = \mathbf{1}[X_i \in A_m]$ and the first column of $\boldsymbol{X}_m$ is a vector of ones. We further define the following partition:

$$n^{-1}\left[\boldsymbol{X}_m^T \boldsymbol{W}_m \boldsymbol{X}_m\right] = \begin{bmatrix} S_{0m} & S_{1m}^T \\ S_{1m} & \boldsymbol{S}_{2m} \end{bmatrix}, \tag{3.5}$$

where $S_{0m} = n^{-1} \sum_{i=1}^n \mathbf{1}\{X_i \in A_m\} = \hat{\pi}_m$, and S_{2m} is a p-by-p sample Gram matrix for the covariates without the intercept.

Now we give the explicit construction of our initial SQ estimator. For each $\bar{x}_m$ and each quantile level $s \in (0, 1)$, we fit the following bin-wise linear regression:

$$\min_{\substack{c_0 \in \mathbb{R} \\ c_1 \in \mathbb{R}^p}} \sum_{\substack{i=1 \\ X_i \in A_m}}^n \left[\hat{Z}_i(s) - c_0 - c_1^T(\tilde{X}_i - \tilde{x}_m)\right]^2,$$

and we use the estimated intercept term $\hat{c}_0$ as the initial SQ estimator at $\bar{x}_m$. The

solution can be given in closed form as:[1]:

$$\hat{v}(s,\bar{x}_m) = e_1^T \left[\boldsymbol{X}_m^T \boldsymbol{W}_m \boldsymbol{X}_m\right]^{-1} \begin{bmatrix} \sum_{i=1}^n w_{im}\hat{Z}_i(s) \\ \sum_{i=1}^n (\tilde{X}_i - \tilde{x}_m) w_{im}\hat{Z}_i(s) \end{bmatrix}, \tag{3.6}$$

where $e_1 = (1, 0, \ldots, 0)$ is a unit vector in $\mathbb{R}^{p+1}$.

Our binning SQ estimator is similar to the the non-parametric estimator in *Olma* (2021), with a key difference being that we do not require a specific form of $\hat{q}(s,x)$. Given a quantile level $s \in (0,1)$, we create an auxiliary variable $Z_i(s)$, which is approximated by $\hat{Z}_i(s)$ using estimated quantile functions. Then we obtain the initial SQ estimator by fitting an OLS using $\hat{Z}_i(s)$ as the response variable. The validity of the approach can be seen from standard local-linear regression theory (*Fan and Gijbels*, 2018) since $\mathrm{E}[Z_i(s) \mid X_i = x] = v(s,x)$.

The motivation to consider our construction is to better control the bias in $\hat{v}(s,x)$ in two ways. First, we use local-linear estimation to alleviate the binning bias since $v(\tau,x)$ is linear-in-covariates; See, e.g., (*Fan and Gijbels*, 2018) for discussions in general non-parametric regression settings. Second, recall in (3.6) that we use $\hat{Z}_i(s)$ as a proxy for $Z_i(s)$, which leads to another source of bias since $\mathrm{E}[\hat{Z}_i(s) \mid X = x] \neq v(s,x)$. In our construction (3.6), we use Neyman-orthogonalization to alleviate the bias attributable to quantile estimation. We relegate more discussions to Section 3.7.6.2.

Remark 6. The initial SQ estimator we provide here is only one possible example that fits into Theorem III.1. While the estimators need to satisfy Conditions G-V1 and G-V2, we believe there are many other possibilities and we do not claim that our construction is optimal. The focus here is to demonstrate that the technical conditions of Theorem III.1 can be satisfied under general conditions.

[1]If a square matrix A is not invertible, A^{-1} is defined as the Moore–Penrose pseudo-inverse.

3.4.2 Theoretical properties

Here we investigate the properties of our initial estimator constructed in (3.6). To this end, we need to have a few more detailed technical conditions, as well as to strengthen Condition G-Y1 in Section 3.2. To fix notations, recall $f_{Y|X}(y;x)$ is the conditional density of Y given $X = x$. For each bin A_m, let

$$\bar{h}_m = \sup_{x \in A_m} \|x - \bar{x}_m\|_2, \quad \underline{h}_m = \inf_{x \notin A_m} \|x - \bar{x}_m\|_2,$$

where $\bar{h}_m$ is the radius of the bin, and $\underline{h}_m$ is the separation between bins. We further define $\bar{h} = \max_m\{\bar{h}_m\}$ and $\underline{h} = \min_m\{\underline{h}_m\}$; both of these quantities depend on the sample size n. For a matrix G, let $\|G\|_2$ be its operator norm.

Condition G-Y1'. All requirements in Condition G-Y1 hold; in addition, we have:

1. For some constant $L_1 > 0$,

$$\sup_{x \in \mathcal{X}} |f_{Y|X}(y_1;x) - f_{Y|X}(y_2;x)| \le L_1 |y_1 - y_2|.$$

2. For some $\delta_0 > 0$,

$$\sup_{x \in \mathcal{X}} \mathrm{E}\left[(Y^+)^{2+\delta_0} \middle| X = x\right] < +\infty,$$

   where $Y^+ = \max\{Y, 0\}$ is the positive part of Y.

Condition G-A1. There exists a constant $\varepsilon_1 > 0$, such that

$$\bar{h} \to 0, \qquad \frac{n^{\min\{1/2,\, 1-2/(2+\delta_0)-\varepsilon_1\}} \underline{h}^p}{\log n} \to \infty,$$

where δ_0 is in Condition G-Y1'. Furthermore, for some constants $0 < m_h < M_h < +\infty$,

$$m_h \le \liminf_{n\to\infty} (\bar{h}^{-1}\underline{h}) \le \limsup_{n\to\infty} (\bar{h}^{-1}\underline{h}) \le M_h.$$

Condition G-A2. At least one of the following takes hold:

1. The covariate-dimension $p < 4$ and $\bar{h}^{4-p} = o(\log^{-1} n)$; furthermore, $v(s, x)$ is twice continuously differentiable with respect to x, and for some $\varepsilon_1 > 0$ and $L_2 > 0$,
$$\sup_{x \in \mathcal{X}} \left\| \frac{\partial^2 v(s,x)}{\partial x \partial x^T} - \frac{\partial^2 v(\tau,x)}{\partial x \partial x^T} \right\|_2 \leq L_2 |s - \tau|,$$
for all $|s - \tau| \leq \varepsilon_1$.

2. For all sufficiently large n, there exists $\beta_n^{(m)}(s)$ such that
$$v_n(s,x) = x^T \beta_n^{(m)}(s), \quad |s - \tau| \leq \varepsilon_2 n^{-1/4}; \, x \in A_m,$$
for each $m = 1, \ldots, M$, where $\varepsilon_2 > 0$ is a universal constant.

We comment on these conditions. First, Condition G-Y1' implies that $v(s, x)$ is finite for each s and x, and that
$$\sup_{x \in \mathcal{X}} \operatorname{var}[Y \mid X = x, \, Y \geq q(\tau, x)] < +\infty.$$
Second, Condition G-A1 ensures each bin is of appropriate size, which is in line to the general bandwidth conditions for kernel SQ estimation (*Olma*, 2021); it also implies that the number of bins $M = M_n$ is upper bounded by $\sqrt{n}$. Moreover, Condition G-A1 ensures that $\bar{h}_m$ and $\underline{h}_m$ are at the same order, which holds if e.g., all the bins are hyperspheres or hypercubes. Third, Condition G-A2 is more technical; it requires either a low-dimensional model with smooth SQ functions over $x \in \mathcal{X}$, or a piece-wise linear SQ regression model. Note the bandwidth condition in item 1 of Condition G-A2 is compatible with Condition G-A1 since $p < 4$. The motivation behind Condition G-A2 is to control the non-parametric binning bias in the initial SQ estimator $\hat{v}(s, x)$.

In addition to these conditions, we also require the following technical conditions on $\hat{q}(s, x)$ used in our initial SQ estimator (3.6). Condition G-Q is relatively weak, and

therefore we can use a wide range of conditional quantile estimators. In particular, $\hat{q}(s,x)$ does not necessarily have to be based (i) a parametric quantile regression model, or (ii) the same binning mechanism as $\hat{v}(s,x)$.

Condition G-Q. For some sequence g_{1n} and g_{2n} with $n^{1/4}g_{1n} \to 0$ and $n^{1/2}g_{2n} \to 0$, the conditional quantile estimator $\hat{q}(s,x)$ satisfies:

1. $$\sup_{\substack{x\in\mathcal{X} \\ s:|s-\tau|\leq n^{-1/4}}} |\hat{q}(s,x) - q(s,x)| = O_{\mathrm{P}}(g_{1n}).$$
2. For each $j = 0, 1$,

$$\sup_{\substack{m=1,\ldots,M \\ s:|s-\tau|\leq n^{-1/4}}} \left\| \frac{\displaystyle\sum_{\substack{i=1 \\ X_i\in A_m}}^{n} \left[\frac{\tilde{X}_i - \tilde{x}_m}{\bar{h}_m}\right]^j [\hat{q}(s,X_i) - q(s,X_i)][s - \mathbf{1}\{Y_i \leq q(s,X_i)\}]}{\displaystyle\sum_{i=1}^{n} \mathbf{1}\{X_i \in A_m\}} \right| = O_{\mathrm{P}}(g_{2n}).$$

Under these new conditions, now we are ready to give the main result of this section, which is tailored for our specific implementation as follows. We consider the m-Rock estimator $\widehat{\beta}$ in (3.1), where we use the initial estimator $\hat{v}(s,x)$ given in (3.6), and use the associated weights

$$\hat{\gamma}_m = (S_{0m} - S_{1m}^T \boldsymbol{S}_{2m}^{-1} S_{1m}), \tag{3.7}$$

where the quantities S_{jm} are defined in (3.5).

Theorem III.2. *Suppose Conditions G-X, G-Y1' and G-Y2 hold under Model (1.4); Furthermore, suppose the binning mechanism satisfies Condition G-A1 and G-A2, and that $\hat{q}(s,x)$ satisfies Condition G-Q. Then the conclusion of Theorem III.1 takes hold under our implementation; In particular, the resulting m-Rock estimator satisfies:*

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \mathrm{N}(0,\ D_1^{-1}\Omega_1 D_1^{-1}),$$

where

$$\Omega_1 = E\left[\frac{\sigma_\tau^2(X)}{[v(\tau,X)-q(\tau,X)]^2}XX^T\right],$$

and $(1-\tau)\sigma_\tau^2(x) = var(Y \mid X = x,\, Y \geq q(\tau,x)) + \tau[v(\tau,x)-q(\tau,x)]^2$.

Theorem III.2 shows that our Neyman-orthogonalized local-linear estimator (3.6) satisfies Conditions G-V1 and G-V2 required by Theorem III.1, and therefore can be used as an initial estimator for the m-Rock approach. Furthermore, the resulting asymptotic variance-covariance matrix in Theorem III.2 is the same as that in Theorem II.1 for discrete covariates. In general, however, the asymptotic distribution may depend on the construction of initial SQ estimators.

Note the m-Rock approach based on Theorem III.2 does not require a parametric quantile regression model, hence is more flexible than many other approaches in the literature (*Dimitriadis and Bayer*, 2019; *Barendse*, 2020; *Patton et al.*, 2019). Moreover, the first-order asymptotic property of $\hat{q}(s,x)$ does not affect the asymptotic variance of $\widehat{\beta}$, thanks to the Neyman-orthogonality in our initial estimator in (3.6). Asymptotically, there is no additional benefit for the m-Rock approach even if a linear quantile regression model is correct.

Compared to the case with discrete covariates, the main technical challenges behind Theorem III.2 can be summarized as follows. First, the number of bins $M = M_n$ increases with the sample size. Therefore the uniform convergence rate of the initial estimators over the bins needs to be carefully investigated. Second, we need to establish the process convergence of $\hat{v}(s,\bar{x}_m)$ for a continuum of s. Standard empirical process tools do not directly apply to the binned data. Moreover, we also need to explicitly analyze the bias in $\hat{v}(s,\bar{x}_m)$ attributable to binning and quantile estimation. In our proof of Theorem III.2, we develop new asymptotic results for the estimated superquantile process.

3.5 Theoretical comparison of SQ regression approaches

In this section, we give a selective comparison between several SQ regression methods, where we compute the asymptotic relative efficiency under different scenarios. For two estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$, the asymptotic relative efficiency (ARE) of $\widehat{\beta}_1$ relative to $\widehat{\beta}_2$ is defined as

$$ARE = \frac{\|\text{var}(\widehat{\beta}_2)\|}{\|\text{var}(\widehat{\beta}_1)\|},$$

where $\|\cdot\|$ is certain matrix norm given in the comparison later; a larger ARE indicates that $\widehat{\beta}_1$ is more efficient. Note the discussions in this section are not based on numerical simulations, but on the asymptotic analysis in Section 3.4; our discussions are more general and more extensive than those in Chapter 2.

All of our comparisons operate under the following joint linear model for quantile and SQ regression:

$$q(\tau, x) = x^T \eta_0(\tau), \quad v(\tau, x) = x^T \beta_0(\tau), \tag{3.8}$$

where only $\beta_0(\tau)$ is the parameter of interest and the index τ is often omitted. From the modeling perspective, the m-Rock approach does not rely on the linear quantile regression model. However, most other competing approaches require such a joint model. Hence we consider the stronger model (3.8) in this section for the sake of theoretical comparison.

3.5.1 Competing approaches

We consider three other approaches for estimating the superquantile regression. These approaches are by no means exhaustive, but we find them to be the most similar to our proposed approach in terms of underlying assumptions and applicability.

The first approach is the 'oracle' approach given in Remark 2.9 of *Dimitriadis and Bayer* (2019), where we assume the true conditional quantile function is known.

Note that this approach is infeasible in practice, and we include this approach only as a benchmark. Given the true conditional τth quantile function of Y given $X = x$ as $q(\tau, x)$, the approach solves

$$\hat{\beta} \leftarrow \min \sum_{i=1}^{n} \left([y_i - x_i^T \beta]^2 \mathbf{1}[y_i \geq q(\tau, x)] \right).$$

This approach estimates the SQ regression as a truncated conditional mean.

Second, we consider the Neyman-orthogonalized truncated Least-Squares (No-LS) in *Barendse* (2020). They consider the following two-step estimation procedure where we fit a quantile regression followed least-squares:

$$\hat{\eta} \leftarrow \min_{\eta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \eta),$$
$$\hat{\beta} \leftarrow \min_{\beta} \sum_{i=1}^{n} \left[Z_i(x_i^T \hat{\eta}) - x_i^T \beta \right]^2,$$

where Z_i is defined in (3.4). The validity of the second stage least-squares can be seen from $\mathrm{E}[Z_i(x_i^T \eta_0) \mid X = x_i] = x_i^T \beta_0$. Therefore we achieve estimation of superquantile regression via least-squares.

The third approach is the joint quantile and expected-shortfall approach (*Dimitriadis and Bayer*, 2019; *Patton et al.*, 2019). Given a non-decreasing function $G_1(\cdot)$ and a concave increasing function $G_2(\cdot)$, the approach minimizes a joint loss function [2]

$$(\hat{\eta}, \hat{\beta}) \leftarrow \min_{\eta, \beta} \sum_{i=1}^{n} \ell_i(\eta, \beta; G_1, G_2),$$

[2]We reverse the direction of the loss function from lower tail to upper tail to match with our context

where

$$\ell_i(\eta, \beta; G_1, G_2) = \rho_\tau \left(G_1(y_i) - G_1(x_i^T\eta)\right) + G_2'(x_i^T\beta)\left[Z_i(x_i^T\eta) - x_i^T\beta\right] + G_2(x_i^T\beta), \tag{3.9}$$

and Z_i is defined in (3.4). See *Fissler and Ziegel* (2016) and *Dimitriadis and Bayer* (2019) for more discussion on the bivariate loss function.

The joint approach depends on the so-called 'specification functions' G_1 and G_2, both numerically and theoretically. *Dimitriadis and Bayer* (2019) demonstrated that the choice of G_2 is central to the estimation efficiency of $\hat{\beta}$. However, there is no universal recommendation of which G_2 would be the most beneficial. To make a solid argument, we focus on the following two options advocated by *Dimitriadis and Bayer* (2019) and *Patton et al.* (2019):

$$G_2(u) = \log(u), \quad \text{and} \quad G_2(u) = \sqrt{u}.$$

We shall name the two versions of the Joint approaches Joint-1 (J1) and Joint-2 (J2), respectively.

Here we compute the asymptotic variance-covariance matrices for the competing approaches under Model (3.8). Let

$$m_1(x) = \text{var}(Y \mid Y \geq q(\tau, x), X = x), \quad m_2(x) = [v(\tau, x) - q(\tau, x)].$$

We write $A \prec B$ if the matrix $B - A$ is positive definite, and $A \preceq (B \wedge C)$ means both $B - A$ and $C - A$ are positive semi-definite. Next we collect some results from the literature. From Remark 2.9 in *Dimitriadis and Bayer* (2019), the Oracle approach has an asymptotic variance-covariance matrix:

$$V_{ORCL} = \frac{1}{1-\tau}\text{E}[XX^T]^{-1}\text{E}\left[XX^T m_1(X)\right]\text{E}[XX^T]^{-1}.$$

The result for NO-LS approach is given by Theorem 1 of *Barendse* (2020):

$$V_{NOLS} = \frac{1}{1-\tau}\mathrm{E}[XX^T]^{-1}\mathrm{E}\left[XX^T\{m_1(X)+\tau m_2^2(X)\}\right]\mathrm{E}[XX^T]^{-1}.$$

From Theorem 2.4 of *Dimitriadis and Bayer* (2019), we have for the Joint approach

$$V_{J1} = \frac{1}{1-\tau}\mathrm{E}\left[\frac{XX^T}{v^2(\tau,X)}\right]^{-1}\mathrm{E}\left[XX^T\left\{\frac{m_1(X)+\tau m_2^2(X)}{v^4(\tau,X)}\right\}\right]\mathrm{E}\left[\frac{XX^T}{v^2(\tau,X)}\right]^{-1},$$
$$V_{J2} = \frac{1}{1-\tau}\mathrm{E}\left[\frac{XX^T}{v^{3/2}(\tau,X)}\right]^{-1}\mathrm{E}\left[XX^T\left\{\frac{m_1(X)+\tau m_2^2(X)}{v^3(\tau,X)}\right\}\right]\mathrm{E}\left[\frac{XX^T}{v^{3/2}(\tau,X)}\right]^{-1},$$

for $G_2(u) = \log u$ and $G_2(u) = \sqrt{u}$, respectively. Finally, recalling from Theorem III.2 that the m-Rock approach has an asymptotic variance-covariance matrix:

$$V_{ROCK} = \frac{1}{1-\tau}\mathrm{E}\left[\frac{XX^T}{m_2(X)}\right]^{-1}\mathrm{E}\left[XX^T\left\{\frac{m_1(X)}{m_2^2(X)}+\tau\right\}\right]\mathrm{E}\left[\frac{XX^T}{m_2(X)}\right]^{-1}.$$

The matrices V_{NOLS}, V_{J1}, V_{J2}, and V_{ROCK} all involve the weighted expectations of $m_1(X)+\tau m_2^2(X)$. The key difference between these approaches lies in the weighting scheme. For V_{NOLS}, V_{J1}, V_{J2} and V_{ROCK} the weights are proportional to 1, $[v(\tau,x)]^{-2}$, $[v(\tau,x)]^{-3/2}$, and $[m_2(x)]^{-1}$ respectively. We shall examine their effect in the following examples.

3.5.2 Efficiency comparison I: homoscedastic models

First, we consider the following homoscedastic linear model with covariate $X \in \mathbb{R}^p$:

$$Y = \gamma_0 + \gamma_1^T X + \varepsilon,$$

where ε is independent of X and has a density function $f_0(\cdot)$. For a fixed τ, let $q_0(\tau)$ and $v_0(\tau)$ be the τth quantile and superquantile of ε, and we further define $V_0(\tau) = \mathrm{var}[\varepsilon \mid \varepsilon > q_0(\tau)]$; in the following we shall omit the index τ in these

quantities. The τth quantile and superquantile regression are:

$$q(\tau, x) = (\gamma_0 + q_0) + x^T\gamma_1, \quad v(\tau, x) = (\gamma_0 + v_0) + x^T\gamma_1.$$

In addition, we have

$$m_1(x) = V_0, \quad m_2(x) = v_0 - q_0,$$

and neither of which depend on x.

With simple algebra, we can show that

$$V_{ORCL} \prec V_{NOLS} = V_{ROCK} \preceq (V_{J1} \wedge V_{J2}),$$

where the equality takes hold if and only if $x^T\gamma_1$ is constant almost surely. We see both the NO-LS and the m-Rock approach are more efficient than the two Joint approaches. Specifically, the Joint-1 and Joint-2 approaches *loses* efficiency by incorporating non-constant weighting in homoscedastic models. On the other hand, the m-Rock approach remains efficient as the weight $m_2(x)$ is constant in homoscedastic models.

To better visualize the difference in asymptotic efficiency, we numerically compute the asymptotic variance-covariance matrix for each method below. For concreteness, we focus on $\tau = 0.9$ and consider the following setting: let $p = 3$ and X be uniformly distributed on the cube $[0, 3]^3$; furthermore, let ε follow the standard normal distribution. We fix $\gamma_0 = 1$ and we randomly sample 200 values of γ_1 from the cube $[0, 5]^3$. Figure 3.1 summarizes the results under the sampled γ_1 values; it shows the determinant and Frobenius norm of the asymptotic variance-covariance matrix for the Joint-1 and Joint-2 approaches, relative to the m-Rock (or NO-LS) approach. We see the Joint approaches are always less efficient than the m-Rock (or NOLS) approach. Remarkably, the asymptotic variance for the J1 approach can be $2 - 3$

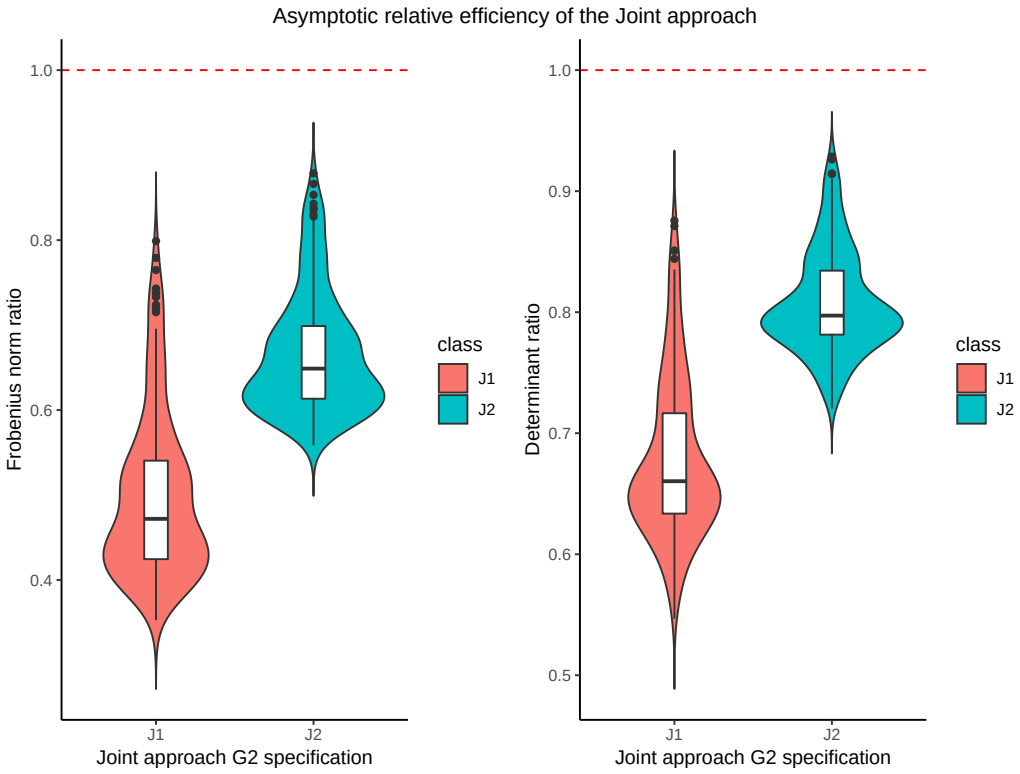


Figure 3.1: The violin plot of ARE relative to the m-Rock approach in the homoscedastic model; the plot is under 200 random values of γ_1. The left panel compares the efficiency by the Frobenius norm of the variance-covariance matrix Σ; the right panel compares by the normalized determinant $|\Sigma|^{1/p}$.

times as large as the m-Rock approach under a simple linear model. Heuristically, the Joint approaches perform worse under larger values of $\|\gamma_1\|$.

3.5.3 Efficiency comparison II: heteroscedastic models

Next we consider the heteroscedastic model

$$Y = (\gamma_1^T X) + (\gamma_2^T X)\varepsilon.$$

where $X = (1, X_1, \ldots, X_p)$ contains an intercept term and ε is independent of X. Let the covariates (excluding the intercept) be uniformly distributed on the unit cube $[0,1]^p$. We shall write $\gamma_1^T = (\gamma_{10}, \gamma_{11}^T)$ where $\gamma_{11} \in \mathbb{R}^p$; similar notation applies to γ_2. Let q_0, v_0 and V_0 be the same quantity as in the last subsection. For identifiability, we consider the case with $v_0 = 0$ and $V_0 = 1$, and that the true parameters satisfy $\gamma_1^T X > 0$ and $\gamma_2^T X > 0$ almost surely. This linear location-scale shift model satisfies the joint model (3.8), with

$$q(\tau, x) = \gamma_1^T x + q_0(\gamma_2^T x), \quad v(\tau, x) = \gamma_1^T x + v_0(\gamma_2^T x) = \gamma_1^T x.$$

Under this more complicated model, it is difficult to give an analytical comparison between the asymptotic variance-covariance matrices. Instead, we shall calculate the asymptotic relative efficiency (ARE) under different values of model parameters, similar to how we obtain Figure 3.1. We focus on $\tau = 0.9$ and suppose ε follows a (scaled) normal distribution with $v_0 = 0$ and $V_0 = 1$. In the following, we give the comparisons under three different model specifications.

First, we examine the ARE relative to the NO-LS approach when $p = 3$; we fix $\gamma_{10} = \gamma_{20} = 3$ and sample 200 different values of γ_{11} and γ_{21} independently and uniformly in $[-1, 3]^3$. Figure 3.2 summarizes the ARE under the 200 sampled parameter values. The m-Rock approach is consistently more efficient than the NO-LS approach

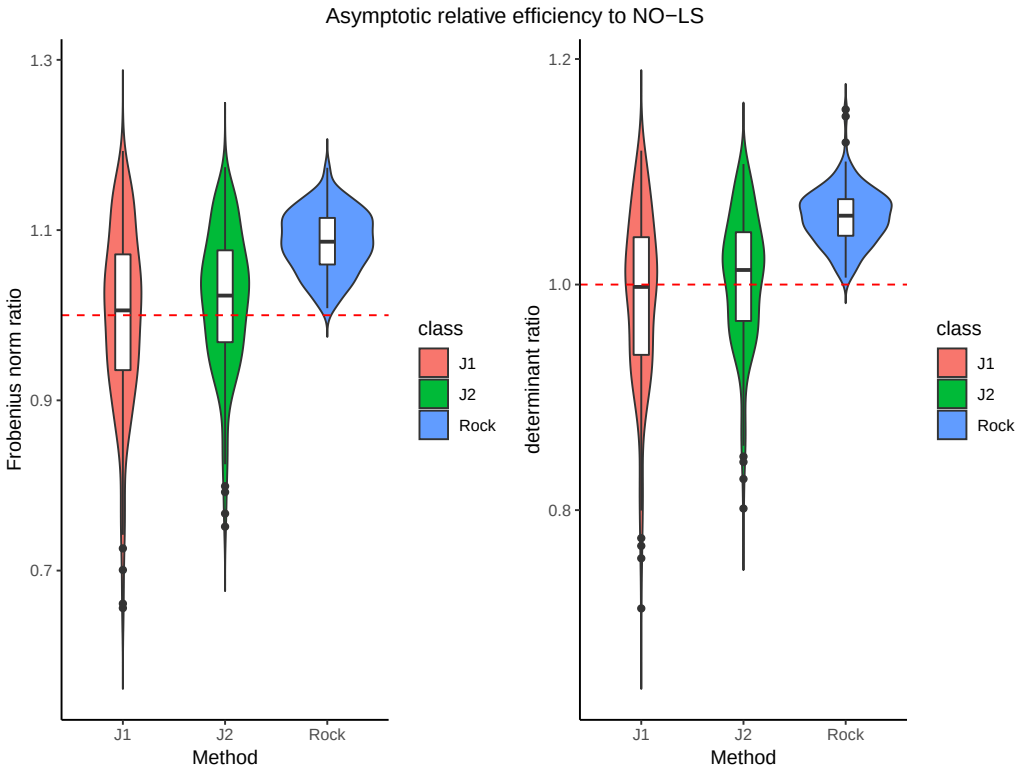


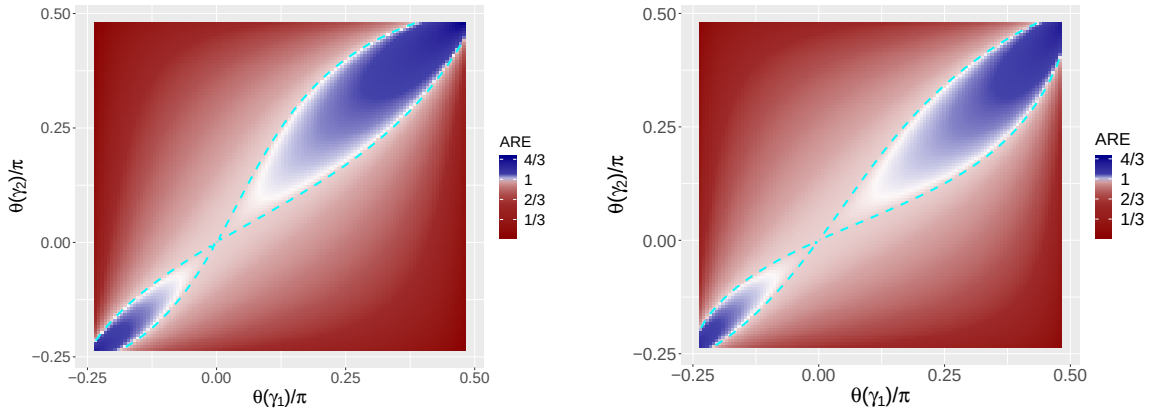
Figure 3.2: The violin plot of ARE relative to the NO-LS approach in the linear location-scale shift model with $p = 3$; the plot is under 200 random values of γ_1 and γ_2. We omit the result for the Oracle approach in the plot, whose ARE is about 2.16. Other attributes of the plots are the same as Figure 3.1.

because the weight $m_2(x)$ is always proportional to the direction of heteroscedasticity under the location-scale model. On the other hand, the performance of the Joint approach varies heavily with γ_1 and γ_2. Depending on the model coefficients, both J1 and J2 can be up to 20% more or less efficient than the NO-LS approach.

Next, we further examine the ARE of the Joint approaches relative to the m-Rock approach. To this end, we set $p = 1$ to simplify the setting, and let $\|\gamma_1\|_2 = \|\gamma_2\|_2 = 1$; note γ_1 and γ_2 are 2-dimensional vectors with an intercept term. For a two-dimensional vector γ, we define $\theta(\gamma)$ to be its angular coordinate. Figure 3.3 shows the ARE when the angular coordinate of γ_1 and γ_2 varies between $-\pi/4$ and $\pi/2$ on the Polar system. We can see that the contour is roughly characterized by the angular difference $\theta(\gamma_1) - \theta(\gamma_2)$. Only when γ_1 is approximately parallel to γ_2, the Joint approaches can be more efficient than the m-Rock approach. In general, however, either J1 or J2 can be up to 30% less efficient than the m-Rock approach.

There is a clear intuition behind the comparison. Note the Joint approaches has a weight proportional to $v(\tau, x) = x^T\gamma_1$ in the sandwich variance-covariance matrix, yet the m-Rock approach involves the weight of $m_2(x) = v(\tau, x) - q(\tau, x) = (v_0 - q_0)(x^T\gamma_2)$. Under the linear location-scale model, it is γ_2 that governs the degree of heteroscedasticity and hence should be used as weights. Therefore, the Joint approach is only competitive when γ_1 is similar to γ_2, i.e., the location shift is in the same direction as the scale shift.

Finally, we point out that even the Oracle approach may not be the most efficient. To this end, we consider the following example with $p = 1$; we set $\gamma_1 = (1, 4)$, $\gamma_2 = (0.5, \gamma_{21})$ and we vary γ_{21}. Figure 3.4 shows the ARE relative to the Oracle approach when γ_{21} grows. Both J1 and m-Rock can be more efficient than the Oracle, and when $\gamma_{21} > 12$, i.e., with strong heteroscedasticity, the m-Rock approach is the most efficient. While the Oracle approach uses the true conditional quantile, it does not apply any weight to the data. With strong heteroscedasticity, the implicit



(a) Method J1

(b) Method J2

Figure 3.3: The heatmap of ARE of the Joint approaches relative to the m-Rock approach in the linear location-scale shift model with $p = 1$, and for each value of γ_1 and γ_2 on the unit circle; the x and y axis represent the angular coordinate of γ_1 and γ_2 in the Polar coordinate system, respectively. The ARE is measured by the Frobenius norm of the asymptotic variance-covariance matrix.

weighting in the m-Rock approach can lead to better efficiency.

To conclude the asymptotic efficiency comparisons, we find that none of the approaches can be universally the most efficient. Though not optimal, the implicit weighting of the m-Rock approach is more adaptive to data heterogeneity. Hence it can often be beneficial for statistical efficiency.

3.6 Discussion

In this chapter, we study the theoretical properties of the m-Rock approach with general covariate distributions. Our analysis specializes to binning of the covariate space with certain initial SQ estimates. We show that the m-Rock approach is asymptotically equivalent to a weighted linearization of those initial SQ estimator. Via theoretical efficiency comparisons, we demonstrate that those weights are often adaptive to the heterogeneity in data; Hence, the m-Rock approach achieves desirable, if not superior, statistical efficiency compared to other common approaches in the literature.

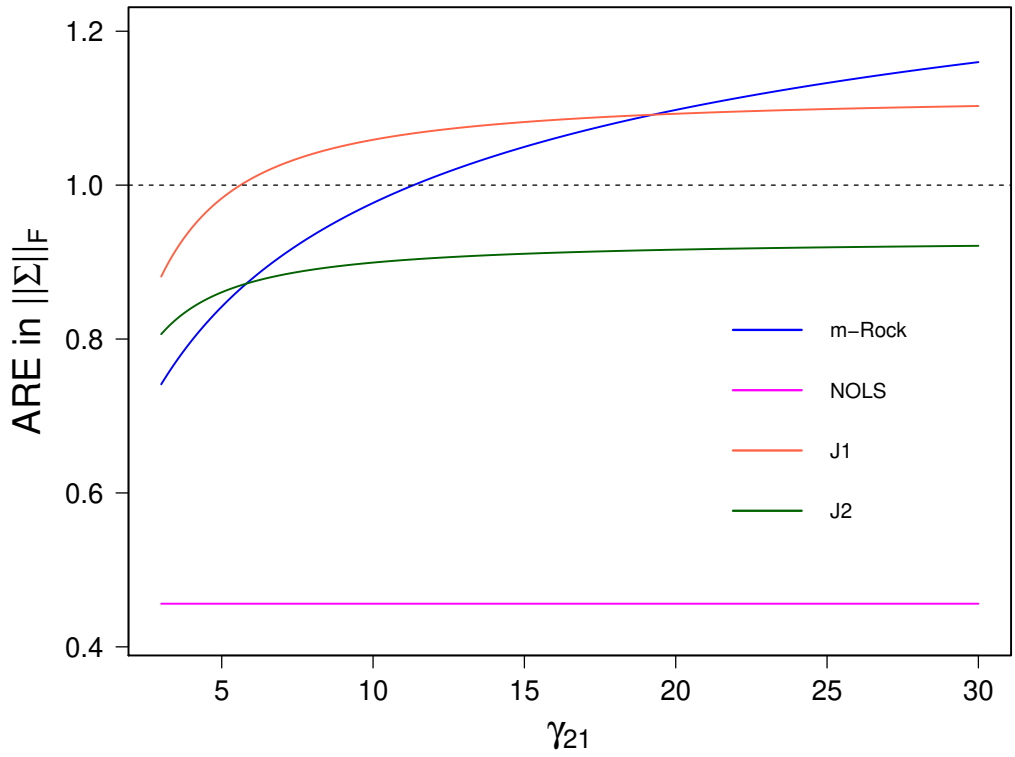


Figure 3.4: The ARE relative to the Oracle approach in the linear location-scale shift model with $p = 1$, where we fix $\gamma_1 = (1, 4)$, $\gamma_2 = (0.5, \gamma_{21})$ and vary γ_{21} from 0 to 30. The ARE measured by the Frobenius norm of the asymptotic variance-covariance matrix.

As one concrete example, we show that a Neyman-orthogonalized local-linear SQ estimator can be used for the m-Rock approach. The resulting m-Rock estimator is asymptotically normal under relatively weak conditions, and the asymptotic variance does not depend on the choice of conditional quantile estimators.

The m-Rock approach is significantly different from other existing approaches for SQ regression. Operationally, it uses a set of non-parametric SQ estimators for a more efficient estimation of the linear SQ regression model. From the modeling perspective, the m-Rock approach is one of the most flexible approaches since it does not explicitly require a joint quantile and SQ regression model (*Dimitriadis and Bayer*, 2019; *Barendse*, 2020), nor does it require a parametric model for the SQ process (*Peracchi and Tanase*, 2008; *Leorato et al.*, 2012).

There are several limitations to our theoretical analysis of the m-Rock approach. First, we rely on binning of the covariate space, where the bins have to be disjoint and cannot depend on the data. Second, our analysis builds on a set of local non-parametric SQ estimators, hence our theoretical results may not generalize to settings with high dimensions.

3.7 Technical details

3.7.1 Some technical lemmas

Here we collect some technical lemmas that are useful for our proofs later. These lemmas do not depend on the specific construction of an initial SQ estimator, and hence are applicable for te results in both Sections 3.3 and 3.4. The proof of these lemmas can be found in Sections 3.7.5.1 and 3.7.5.2.

We fix some notations here. Recall $A_1, \ldots, A_M$ are the bins. For each bin, $\bar{x}_m$ is its geometric center, and $\hat{\gamma}_m$ is a weight (that only depends on the covariates) in the m-Rock estimation procedure (3.1). Let $\hat{v}(s, x)$ and $\hat{q}(s, x)$ be the initial binning

SQ and quantile estimators, respectively. The lemmas below do not depend on any particular choice of $\hat{\gamma}_m$, $\hat{v}$ or $\hat{q}$. The total number of bins M is allowed to increase with the sample size. For the binning mechanism, let $w_{im} = \mathbf{1}\{X_i \in A_m\}$, and let

$$\hat{\pi}_m = n^{-1} \sum_{i=1}^{n} w_{im},$$

be the proportion of data that falls into bin A_m. For each bin, we write $diam(A_m) = \bar{h}_m = \sup_{x \in A_m} \|x - \bar{x}_m\|$ and $\underline{h}_m = \inf_{x \notin A_m} \|x - \bar{x}_m\|$. We further define the inverse SQ function as[3]:

$$\begin{aligned} h(z,x) &:= \int_0^1 \mathbf{1}\{v(s,x) \le z\}\, \mathrm{d}s = \sup\{s \in [0,1] : v(s,x) \le z\}, \\ \hat{h}(z,x) &:= \int_0^1 \mathbf{1}\{\hat{v}(s,x) \le z\}\, \mathrm{d}s = \sup\{s \in [0,1] : \hat{v}(s,x) \le z\}. \end{aligned} \tag{3.10}$$

In the Operations Research literature, these functions are called the 'superdistribution' functions, in duality to the superquantile functions (*Rockafellar and Royset*, 2013; *Rockafellar and Uryasev*, 2013).

We also use the following set of notations. For a vector v, let $\|v\|$ be its ℓ_2 norm; for a matrix A, let $\|A\|$ be its operator norm. For two deterministic sequences a_n and b_n, we write $a_n \ll b_n$ if $a_n = o(b_n)$ and $a_n \lesssim b_n$ if there exists a universal constant $C^* > 0$ such that $a_n \le C^* b_n$; we define $a_n \asymp b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$ hold. For stochastic sequences A_n and B_n, we use the notations $A_n \ll_{\mathrm{P}} B_n$ and $A_n \lesssim_{\mathrm{P}} B_n$ to denote $A_n = o_{\mathrm{P}}(B_n)$ and $A_n = O_{\mathrm{P}}(B_n)$, respectively.

Lemma 5. *Suppose the bins A_m and the associated weights $\hat{\gamma}_m$ satisfy:*

$$\sup_{m=1,\dots,M} diam(A_m) \xrightarrow{P^*} 0, \qquad \sup_{m=1,\dots,M} \left| \frac{\hat{\gamma}_m}{\hat{\pi}_m} - 1 \right| \xrightarrow{P^*} 0.$$

Let $g(\cdot) : \mathcal{X} \mapsto \mathbb{R}^m$ be a bounded and Lipschitz continuous function over $\mathcal{X}$, then we

[3]Without loss of generality, we assume $\hat{v}(s,x)$ and $v(s,x)$ are (weakly) increasing in s.

have

$$\sum_{m=1}^{M} \hat{\gamma}_m g(\bar{x}_m) \xrightarrow{P^*} E[g(X)].$$

In addition, if $h(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ *is a function such that* $E[|h(X)|] < \infty$*, then we have*

$$E\left[\sum_{m=1}^{M} \mathbf{1}_{\{X \in A_m\}} g(\bar{x}_m) h(X)\right] \to E[h(X)g(X)],$$

as $n \to \infty$*.*

Lemma 6. *Under Conditions G-X, G-Y1 and G-Y2, there is a constant* $c_1 > 0$ *such that the following results hold:*

1. *For some constants* $0 < \underline{m}_1 < \overline{m}_1 < +\infty$*, we have*

$$\underline{m}_1 \leq \inf_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq c_1}} |v(s,x) - q(s,x)| \leq \sup_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq c_1}} |v(s,x) - q(s,x)| \leq \overline{m}_1.$$

2. *Both* $q(s,x)$ *and* $v(s,x)$ *are differentiable with respect to* s *when* $|s - \tau| \leq c_1$*, and there exist constants* $0 < \underline{m}_2 < \overline{m}_2 < +\infty$ *such that*

$$\begin{aligned}
\underline{m}_2 \leq \inf_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq c_1}} \left|\frac{\partial q}{\partial s}(s,x)\right| &\leq \sup_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq c_1}} \left|\frac{\partial q}{\partial s}(s,x)\right| \leq \overline{m}_2, \\
\underline{m}_2 \leq \inf_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq c_1}} \left|\frac{\partial v}{\partial s}(s,x)\right| &\leq \sup_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq c_1}} \left|\frac{\partial v}{\partial s}(s,x)\right| \leq \overline{m}_2.
\end{aligned}$$

3. *There exists a constant* $L > 0$ *such that*

$$\begin{aligned}
\sup_{x \in \mathcal{X}} \left|\frac{\partial v}{\partial s}(s_1,x) - \frac{\partial v}{\partial s}(s_2,x)\right| &\leq L|s_1 - s_2|, \\
\sup_{x \in \mathcal{X}} \left|\left[\frac{\partial v}{\partial s}(s_1,x)\right]^{-1} - \left[\frac{\partial v}{\partial s}(s_2,x)\right]^{-1}\right| &\leq L|s_1 - s_2|,
\end{aligned}$$

for all $s_1, s_2 \in [\tau - c_1, \tau + c_1]$*.*

Lemma 7. *Suppose the initial estimators $\hat{v}(s, \bar{x}_m)$ satisfy Condition G-V1, and the binning weights $\hat{\gamma}_m$ satisfy*

$$\sup_{m=1,\ldots,M} \left| \frac{\hat{\gamma}_m}{\hat{\pi}_m} - 1 \right| \xrightarrow{P^*} 0.$$

The following results hold, where r_n is the same as in Condition G-V1.

1. $\sum_{m=1}^{M} \hat{\gamma}_m \sup_{\substack{(z,z'):z=v(\tau,\bar{x}_m) \\ |z'-z| \lesssim (r_n \vee n^{-1/2})}} \left| \{\hat{h}(z, \bar{x}_m) - h(z, \bar{x}_m)\} - \{\hat{h}(z', \bar{x}_m) - h(z', \bar{x}_m)\} \right| = o_P\left(\frac{1}{\sqrt{n}}\right).$

2. $\sum_{m=1}^{M} \hat{\gamma}_m \left[\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)\right]^2 = o_P\left(\frac{1}{\sqrt{n}}\right).$

3. $\sum_{m=1}^{M} \hat{\gamma}_m \left| \tau - \hat{h} \circ (\hat{v}(\tau, \bar{x}_m), \bar{x}_m) \right| = o_P\left(\frac{1}{\sqrt{n}}\right).$

Lemma 8. *Recall S_{0m}, S_{1m}, and $\boldsymbol{S}_{2m}$ given in (3.5). Under Conditions G-A1 and G-X, the following results hold:*

1. $\sup_{m=1,\ldots,M} \left| S_{0m}^{-1} \left(S_{0m} - S_{1m}^T \boldsymbol{S}_{2m}^{-1} S_{1m}\right) - 1 \right| = o_P(1).$

2. *For any fixed $c_2 > 0$,*

$$Pr\left(\sup_{m=1,\ldots,M} \left\| \bar{h}_m S_{1m}^T \boldsymbol{S}_{2m}^{-1} \right\| \geq c_2 \right) \leq \frac{C_2}{n^3},$$

   where C_2 is a constant that may depend on c_2.

3. *For some $\varepsilon_2 > 0$,*

$$Pr\left(\inf_{m=1,\ldots,M} |\bar{h}_m^{-p} S_{0m}| \leq \varepsilon_2 \right) \leq \frac{1}{n^3}$$

Lemma 9. *Suppose Condition G-Y1' holds. For any two sequences $a_n, b_n \to 0$, if the quantile estimator $\hat{q}(s, x)$ satisfies*

$$\sup_{\substack{x \in \mathcal{X} \\ s:|s-\tau| \leq a_n}} |\hat{q}(s, x) - q(s, x)| = O_P(b_n),$$

then we have

1. $$\sup_{\substack{m=1,\dots,M \\ |s-\tau|\le a_n}} \left| \frac{\sum_{i=1}^{n} w_{im}\kappa_{im}[Y_i - q(s,X_i)]\left[\mathbf{1}\{Y_i \ge \hat{q}(s,X_i)]\} - \mathbf{1}\{Y_i \ge q(s,X_i)\}\right]}{\sum_{i=1}^{n} w_{im}} \right| = O_P(a_n^2 + b_n^2);$$

2. $$\sup_{\substack{m=1,\dots,M \\ |s-\tau|\le a_n}} \left| \frac{\sum_{i=1}^{n} w_{im}\kappa_{im}\left(q(s,X_i) - \hat{q}(s,X_i)\right)\left[\mathbf{1}\{Y_i \ge \hat{q}(s,X_i)\} - \mathbf{1}\{Y_i \ge q(s,X_i)\}\right]}{\sum_{i=1}^{n} w_{im}} \right| = O_P(a_n^2 + b_n^2);$$

where $w_{im} = \mathbf{1}\{X_i \in A_m\}$ *and* $\kappa_{im} = [1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \bar{x}_m)]$*;* S_{1m}*,* $\boldsymbol{S}_{2m}$ *are given in (3.5).*

We comment on these lemmas. Lemmas 5 and 8 are about the covariate properties with the binning mechanism. Lemma 6 gives some more technical implications on the data generating process derived from the conditions in Section 3.2. In addition, Lemma 7 and 9 are more technical. In particular, the results in 7 are similar to, but stronger than the examples given by standard functional delta method [Chapter 20](*Van der Vaart*, 2000). For a fixed quantile level τ, Lemma 9 is the same as Lemma A.4 in *Olma* (2021) and Lemma A.3 in *Kato* (2012). Our result is stronger in the uniformity over s.

3.7.2 Proof of Theorem III.1

Now we give the proof to our first main result Theorem III.1. To simplify the notations, in the following proof we define $v_m(\alpha) = v(\alpha, \bar{x}_m)$, and $\hat{v}_m(\alpha) = \hat{v}(\alpha, \bar{x}_m)$; correspondingly we write $h_m(z) = h(z, \bar{x}_m)$ and $\hat{h}_m(z) = \hat{h}(z, \bar{x}_m)$ for the inverse SQ function. When there is no confusion, we shall write $v_m = v(\tau, \bar{x}_m)$ without the index

to refer to the targeting τth SQ. In our proof, the number of bins $M = M_n$ increases with the sample size, though we often omit the subscript.

Proof of Theorem III.1. In this proof, we work with the following shifted m-Rock objective function:

$$L_n(\delta) = \sum_{m=1}^{M} \hat{\gamma}_m \int_0^1 \left[\rho_\tau\left(\hat{v}_m(\alpha) - v_m(\tau) - \bar{x}_m\delta/\sqrt{n}\right) - \rho_\tau\left(\hat{v}_m(\alpha) - v_m(\tau)\right)\right] \mathrm{d}\alpha. \tag{3.11}$$

It follows that $\hat{\delta} = n^{1/2}(\widehat{\beta} - \beta)$ minimizes $L_n(\delta)$, where $\widehat{\beta}$ is the m-Rock estimator in (3.1). Therefore, it suffices to study the asymptotic properties of $\hat{\delta}$. To this end, we first show that the function $L_n(\delta)$ in (3.11) converges (pointwise) in probability to a quadratic function of δ. Then we apply the convexity argument in *Pollard* (1991) to derive the asymptotic properties of $\hat{\delta}$. We define $\Delta_m(\delta) = \hat{v}_m(\tau) - v_m(\tau) - n^{-1/2}\bar{x}_m^T\delta$.

By Knight's identity (*Knight*, 1998),

$$\rho_\tau(w - v) - \rho_\tau(w) = -v(\tau - \mathbf{1}\{w \le 0\}) + \int_0^v (\mathbf{1}\{w \le t\} - \mathbf{1}\{w \le 0\})\,\mathrm{d}t,$$

for any w and v, therefore

$$\begin{aligned}\rho_\tau(w - v_1) - \rho_\tau(w - v_2) &= [\rho_\tau(w - v_1) - \rho_\tau(w)] - [\rho_\tau(w - v_2) - \rho_\tau(w)] \\ &= (v_2 - v_1)(\tau - \mathbf{1}\{w \le 0\}) + \int_{v_2}^{v_1} (\mathbf{1}\{w \le t\} - \mathbf{1}\{w \le 0\})\,\mathrm{d}t.\end{aligned}$$

Taking $w = \hat{v}_m(\alpha) - \hat{v}_m(\tau)$, $v_1 = -\Delta_m(\delta)$, and $v_2 = v_m(\tau) - \hat{v}_m(\tau)$ in the above

displayed equation, we obtain:

$$\begin{aligned}
&\int_0^1 \rho_\tau\left(\hat{v}_m(\alpha) - v_m(\tau) - \bar{x}_m^T\delta/\sqrt{n}\right)\,\mathrm{d}\alpha - \int_0^1 \rho_\tau\left(\hat{v}_m(\alpha) - v_m(\tau)\right)\,\mathrm{d}\alpha \\
= \;& -n^{-1/2}\tilde{x}^T\delta\int_0^1 (\tau - \mathbf{1}\{\hat{v}_m(\alpha) \le \hat{v}_m(\tau)\})\mathrm{d}\alpha \\
&+ \int_0^1 \int_{v_m-\hat{v}_m}^{-\Delta_m(\delta)} (\mathbf{1}\{\hat{v}_m(\alpha) \le \hat{v}_m(\tau)+t\} - \mathbf{1}\{\hat{v}_m(\alpha) \le \hat{v}_m(\tau)\})\,\mathrm{d}t\,\mathrm{d}\alpha \\
= \;& -n^{-1/2}\bar{x}_m^T\delta[\tau - \hat{h}_m(\hat{v}_m)] + \int_{v_m-\hat{v}_m}^{-\Delta_m(\delta)} [\hat{h}_m(\hat{v}_m+t) - \hat{h}_m(\hat{v}_m)]\,\mathrm{d}t,
\end{aligned}$$

where the last equality follows from the definition of $\hat{h}$ in (3.10), and by exchanging the order of integration. Therefore, summing over $m = 1, \ldots, M$ in the above equation gives the following decomposition for $L_n(\delta)$ (defined in (3.11)):

$$\begin{aligned}
L_n(\delta) \;=\; & \underbrace{-n^{-1/2}\sum_{m=1}^M \hat{\gamma}_m\bar{x}_m^T\delta[\tau - \hat{h}_m(\hat{v}_m)]}_{A_n(\delta)} + \sum_{m=1}^M \hat{\gamma}_m \int_{v_m-\hat{v}_m}^{-\Delta_m(\delta)} [\hat{h}_m(\hat{v}_m+t) - \hat{h}_m(\hat{v}_m)]\,\mathrm{d}t \\
=\; & A_n(\delta) + \underbrace{\sum_{m=1}^M \hat{\gamma}_m \int_{v_m-\hat{v}_m}^{-\Delta_m(\delta)} [h_m(\hat{v}_m+t) - h_m(\hat{v}_m)]\,\mathrm{d}t}_{B_n(\delta)} \\
&+ \underbrace{\sum_{m=1}^M \hat{\gamma}_m \int_{v_m-\hat{v}_m}^{-\Delta_m(\delta)} \left[\{\hat{h}_m(\hat{v}_m+t) - h_m(\hat{v}_m+t)\} - \{\hat{h}_m(\hat{v}_m) - h_m(\hat{v}_m)\}\right]\,\mathrm{d}t}_{C_n(\delta)} \\
\triangleq\; & A_n(\delta) + B_n(\delta) + C_n(\delta). \qquad (3.12)
\end{aligned}$$

For any fixed $\delta = O(1)$, we shall show that both $A_n(\delta)$ and $C_n(\delta)$ are $o_{\mathrm{P}}(n^{-1})$. For $A_n(\delta)$, note $\bar{x}_m^T\delta$ is uniformly bounded over m, hence

$$|n\,A_n(\delta)| \lesssim \sqrt{n}\left(\sum_{m=1}^M \hat{\gamma}_m|\tau - \hat{h}_m(\hat{v}_m)|\right) = o_{\mathrm{P}}(1),$$

from Lemma 7. For $C_n(\delta)$, we first define

$$R_n = \sup_{m=1,\ldots,M} \max\left\{ \frac{|\bar{x}_m^T \delta|}{\sqrt{n}},\ |\hat{v}_m(\tau) - v_m(\tau)| \right\},$$

and it follows from Condition G-V1 that $R_n = O_{\mathrm{P}}(r_n \vee n^{-1/2})$. By taking the supremum within each the integration in $C_n(\delta)$, we have

$$\begin{aligned}
& |n\, C_n(\delta)| \\
\leq\ & \sqrt{n} \left(\sum_{m=1}^{M} \hat{\gamma}_m |\bar{x}_m^T \delta| \times\ 2 \sup_{|s| \leq R_n} \left| \{\hat{h}_m(v_m + s) - h(v_m + s)\} - \{\hat{h}_m(v_m) - h_m(v_m)\} \right| \right) \\
\lesssim_{\mathrm{P}}\ & \sqrt{n} \left(\sum_{m=1}^{M} \hat{\gamma}_m \sup_{|s| \lesssim r_n \vee n^{-1/2}} \left| \{\hat{h}_m(v_m + s) - h_m(v_m + s)\} - \{\hat{h}_m(v_m) - h_m(v_m)\} \right| \right) \\
=\ & o_{\mathrm{P}}(1),
\end{aligned}$$

where the last inequality follows from Lemma 7.

Next we turn to the convergence of $B_n(\delta)$, where we first give a linear approximation for $h_m(\hat{v}_m + t) - h_m(\hat{v}_m)$ in (3.12). Note the derivative for the inverse function $h_m(z) = h(z, \bar{x}_m)$ in (3.10) is:

$$h_m'(z) = \left[\left. \frac{\partial v_m(s)}{\partial s} \right|_{s = h_m(z)} \right]^{-1} = \frac{1 - h_m(z)}{v_m(h_m(z)) - q_m(h_m(z))}. \tag{3.13}$$

By using the first order Taylor-expansion and the mean value theorem, there exists a ξ_m between $\hat{v}_m$ and $\hat{v}_m + t$ such that

$$\begin{aligned}
|h_m(\hat{v}_m + t) - h_m(\hat{v}_m) - t h_m'(v_m)| &= |t[h_m'(\xi_m) - h_m'(v_m)]| \\
&\leq |t| \times (L|\xi_m - \hat{v}_m| + |\hat{v}_m - v_m|) \\
&\leq L\, t^2 + |\hat{v}_m - v_m| \times |t|, \qquad (3.14)
\end{aligned}$$

since $|\xi_m - \hat{v}_m| \leq |t|$, where L is the Lipschitz constant in Lemma 6. Therefore, $B_n(\delta)$

can be approximated as follows:

$$\begin{aligned}
& \left| B_n(\delta) - \frac{1}{2}\sum_{m=1}^{M} \hat{\gamma}_m h'_m(v_m) \left\{ \Delta_m^2(\delta) - [\hat{v}_m - v_m]^2 \right\} \right| \\
= & \left| \sum_{m=1}^{M} \hat{\gamma}_m \int_{v_m - \hat{v}_m}^{-\Delta_m(\delta)} [h_m(\hat{v}_m + t) - h_m(\hat{v}_m) - t h'_m(v_m)] \, \mathrm{d}t \right| \\
\lesssim & \left| \sum_{m=1}^{M} \hat{\gamma}_m \int_{v_m - \hat{v}_m}^{-\Delta_m(\delta)} (t^2 + |\hat{v}_m - v_m| \times |t|) \, \mathrm{d}t \right| \\
\lesssim & \frac{1}{\sqrt{n}} \sum_{m=1}^{M} \hat{\gamma}_m |\hat{v}_m - v_m|^2 + \frac{1}{n} \sum_{m=1}^{M} \hat{\gamma}_m |\hat{v}_m - v_m| + o_{\mathrm{P}}\left(\frac{1}{n}\right) \\
= & o_{\mathrm{P}}\left(\frac{1}{n}\right),
\end{aligned}$$

where the second inequality follows from (3.14), and the last equality holds from Lemma 7. Therefore, $B_n(\delta)$ can be approximated by a function of $\Delta_m^2(\delta)$.

We now show that the loss function L_n is approximately a quadratic function in δ. Let

$$D_{1n} = \frac{1}{1-\tau} \left[\sum_{m=1}^{M} \hat{\gamma}_m h'_m(v_m) \bar{x}_m \bar{x}_m^T \right], \quad \text{and} \quad u_n = \frac{\sqrt{n}}{1-\tau} \sum_{m=1}^{M} \hat{\gamma}_m h'_m(v_m) \bar{x}_m (\hat{v}_m - v_m).$$

Collecting the results for $A_n(\delta)$, $B_n(\delta)$ and $C_n(\delta)$ into (3.12), we have shown that for any fixed $\delta \in \mathbb{R}^{p+1}$,

$$\begin{aligned}
n \cdot L_n(\delta) &= \frac{1}{2} \sum_{m=1}^{M} \hat{\gamma}_m h'_m(v_m) \{ \Delta_m^2(\delta) - [\hat{v}_m - v_m]^2 \} + o_{\mathrm{P}}(1) \\
&= \frac{1}{2} \delta^T D_{1n} \delta - \delta^T u_n + o_{\mathrm{P}}(1), \qquad (3.15)
\end{aligned}$$

where the last equality follows by expanding $\Delta_m(\delta) = \hat{v}_m(\tau) - v_m(\tau) - n^{-1/2} \bar{x}_m^T \delta$. Since the m-Rock loss function $L_n(\delta)$ is convex, standard convexity argument (see e.g., *Hjort and Pollard* (2011) and *Pollard* (1991)) shows that the convergence in (3.15) is uniform in δ over any compact subset of $\mathbb{R}^{p+1}$. Furthermore, the calculation

of h'_m in (3.13) and Lemma 5 shows

$$D_{1n} \xrightarrow{P^*} D_1 = \mathrm{E}\left[\frac{XX^T}{v(\tau, X) - q(\tau, X)}\right],$$

since $v(\tau, x) - q(\tau, x)$ is bounded by Lemma 6. Therefore, (3.15) implies that for any compact set $\mathcal{B} \subset \mathbb{R}^{p+1}$,

$$n \cdot \sup_{\delta \in \mathcal{B}} | L_n(\delta) - Q_n(\delta)| = o_{\mathrm{P}}(1), \tag{3.16}$$

where $Q_n(\delta) = \frac{1}{2}\delta^T D_1 \delta - \delta^T u_n$. This shows that $L_n(\delta)$ can be uniformly approximated by a quadratic function in δ.

Finally, we show the convergence of $\hat{\delta}$, which establishes the asymptotic properties of the m-Rock estimator. As a function of δ, $Q_n(\cdot)$ in (3.16) has a unique minimizer

$$\tilde{\delta} = D_1^{-1} u_n,$$

since D_1 is positive definite. Given Condition G-V2 and (3.16), we apply the Basic Corollary in *Hjort and Pollard* (2011) to conclude that the minimizers of $L_n(\delta)$ and $Q_n(\delta)$ are asymptotically equivalent, i.e.,

$$\hat{\delta} = \tilde{\delta} + o_{\mathrm{P}}(1) = D_1^{-1}\left[\sqrt{n}\sum_{m=1}^{M} \frac{\hat{\gamma}_m \bar{x}_m}{v_m(\tau) - q_m(\tau)}[\hat{v}_m(\tau) - v_m(\tau)]\right] + o_{\mathrm{P}}(1).$$

The proof is now complete by noting that $\hat{\delta} = n^{1/2}(\widehat{\beta} - \beta)$.

□

3.7.3 Proof of Theorem III.2

To prove Theorem III.2, it entails to show that all conditions of Theorem III.1 apply to our specific construction of the initial estimator in (3.6). We break the main

technical requirements into three Propositions below, the proof of which can be found later in this subsection. In our proofs here, $\hat{v}(s, \bar{x}_m)$ refers specifically to the estimator constructed in (3.6), and the weight $\hat{\gamma}_m$ refers to the one in (3.7). Furthermore, we fix the sequence r_n to be the one defined in Proposition 4 below; We shall verify later in the proof of Theorem III.2 that r_n indeed satisfies the requirements in Theorem III.2.

Proposition 3. *Under the conditions of Theorem III.2, we have*

$$\sqrt{n}\sum_{m=1}^{M}\left[\frac{\hat{\gamma}_m \bar{x}_m}{v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)}\{\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)\}\right] \xrightarrow{d} \mathrm{N}\left(0,\ \Omega_1\right),$$

where Ω_1 is defined in Theorem III.2.

Proposition 4. *Let*

$$r_n = \sqrt{\frac{\log n}{n\underline{h}^p}}.$$

Under the condition of Theorem III.2, we have

$$\sup_{m=1,\ldots,M} |\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)| = O_P\left(r_n\right).$$

Proposition 5. *Under the condition of Theorem III.2, we have for any fixed $B > 0$,*

$$\sup_{\substack{m=1,\ldots,M \\ |t| \le B\cdot(r_n + n^{-1/2})}} |[\hat{v}(\tau + t, \bar{x}_m) - v(\tau + t, \bar{x}_m)] - [\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)]| = o_P\left(n^{-1/2}\right),$$

where r_n is given in Proposition 4.

The proof of Theorem III.2 is relatively straightforward with these Propositions, and we now give the details.

Proof of Theorem III.2. Under the conditions of Theorem III.2, the binning mecha-

nism satisfies:

$$\sup_{m=1,\ldots,M} \operatorname{diam}(A_m) \lesssim \bar{h} = o(1), \quad \text{and} \quad \sup_{m=1,\ldots,M} \left| \frac{\hat{\gamma}_m}{\hat{\pi}_m} - 1 \right| = o_{\mathrm{P}}(1),$$

which follows from Lemma 8. It then suffices to check Conditions G-V2 and G-V1.

Proposition 3 directly implies Condition G-V2. From Condition G-A1, we have

$$n^{-1/2} \ll r_n = \sqrt{\frac{\log n}{n\underline{h}^p}} \ll n^{-1/4},$$

therefore the sequence r_n constructed in Proposition 4 can be used in Condition G-V1. Next we check Condition G-V1.

The second requirement in Condition G-V1 follows from Proposition 5. Moreover, from Proposition 4 and 5 we have

$$\begin{aligned}
& \sup_{\substack{m=1,\ldots,M \\ |s-\tau| \le B \cdot r_n}} |\hat{v}(s, \bar{x}_m) - v(s, \bar{x}_m)| \\
\le \; & \sup_{m=1,\ldots,M} |\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)| \\
& + \sup_{\substack{m=1,\ldots,M \\ |t| \le B \cdot r_n}} |[\hat{v}(\tau + t, \bar{x}_m) - v(\tau + t, \bar{x}_m)] - [\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)]| \\
= \; & O_{\mathrm{P}}(r_n) + o_{\mathrm{P}}(n^{-1/2}).
\end{aligned}$$

Hence the first requirement in Condition G-V1 also holds. Since the monotonicity of $\hat{v}(s, x)$ (with respect to s) is assumed, we have checked all requirements of Theorem III.1. The proof is now complete.

□

3.7.4 Proof of Propositions 3, 4 and 5

Here we prove the three Propositions used in the proof Theorem III.2. We fix some notations used in the proof. Recall S_{0m}, S_{1m} and $\boldsymbol{S}_{2m}$ from (3.5); and note

$S_{0m} = n^{-1}\sum_{i=1}^{n} w_{im} = \hat{\pi}_m$. For the weight of each bin in (3.1), we set

$$\hat{\gamma}_m = S_{0m} - S_{1m}^T \boldsymbol{S}_{2m}^{-1} S_{1m},$$

as in (3.7). Using the block matrix inverse, our estimator $\hat{v}(s, \bar{x}_m)$ in (3.6) can be further simplified as:

$$\hat{v}(s, \bar{x}_m) = (n\hat{\gamma}_m)^{-1} \left[\sum_{i=1}^{n} w_{im} \hat{Z}_i(s) - S_{1m}^T \boldsymbol{S}_{2m}^{-1} \sum_{i=1}^{n} (X_i - \bar{x}_m) w_{im} \hat{Z}_i(s) \right], \tag{3.17}$$

where $\hat{Z}_i(s)$ is defined in (3.4). Furthermore, let $\tilde{v}(s, \bar{x}_m)$ be the oracle estimator where we know $q(s, x)$ and $Z_i(s)$, i.e.,

$$\tilde{v}(s, \bar{x}_m) = (n\hat{\gamma}_m)^{-1} \left[\sum_{i=1}^{n} w_{im} Z_i(s) - S_{1m}^T \boldsymbol{S}_{2m}^{-1} \sum_{i=1}^{n} (X_i - \bar{x}_m) w_{im} Z_i(s) \right]. \tag{3.18}$$

In the proofs of this section, we write X as the covariate vector that does not contain the intercept for simplicity.

3.7.4.1 Proof of Proposition 3

Proof. We rely on the decomposition that

$$\begin{aligned}
[\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)] &= [\hat{v}(\tau, \bar{x}_m) - \tilde{v}(\tau, \bar{x}_m)] + [\tilde{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m)] \\
&= [\hat{v}(\tau, \bar{x}_m) - \tilde{v}(\tau, \bar{x}_m)] \\
&\quad - (n\hat{\gamma}_m)^{-1} \left[S_{1m}^T \boldsymbol{S}_{2m}^{-1} \sum_{i=1}^{n} (X_i - \bar{x}_m) w_{im} [Z_i(\tau) - v(\tau, X_i)] \right] \\
&\quad + (n\hat{\gamma}_m)^{-1} \left[\sum_{i=1}^{n} w_{im} [Z_i(\tau) - v(\tau, X_i)] \right],
\end{aligned} \tag{3.19}$$

where the last equality follows from standard local-linear calculation (*Fan and Gijbels*, 2018) since $v(\tau, x)$ is linear in x.

It suffices to consider the aggregation of the three terms in the decomposition above. First, we give two claims below; and we verify them one by one at the end of this proof. In what follows, we define $\zeta_m = v(\tau, \bar{x}_m) - q(\tau, \bar{x}_m)$.

Claim 1:

$$\sqrt{n}\sum_{m=1}^{M}\left\{\frac{\hat{\gamma}_m}{\zeta_m}\bar{x}_m\left[\sum_{i=1}^{n}\frac{w_{im}}{n\hat{\gamma}_m}S_{1m}^T\boldsymbol{S}_{2m}^{-1}(X_i-\bar{x}_m)[Z_i(\tau)-v(\tau,X_i)]\right]\right\} = o_{\mathrm{P}}(1). \quad (3.20)$$

Claim 2:

$$\sqrt{n}\sum_{m=1}^{M}\left\{\frac{\hat{\gamma}_m}{\zeta_m}\bar{x}_m\left[\hat{v}(\tau,\bar{x}_m)-\tilde{v}(\tau,\bar{x}_m)\right]\right\} = O_{\mathrm{P}}\left(\sqrt{n}g_{1n}^2+\sqrt{n}g_{2n}\right) = o_{\mathrm{P}}(1), \quad (3.21)$$

where g_{1n} and g_{2n} are given in Condition G-Q.

Claims 1 and 2 together show the first two terms in Equation (3.19) are asymptotically negligible when aggregated over the bins. In particular, they show that using our initial estimator is asymptotically equivalent to using the oracle estimators. In what follows, the proof is given in three steps. In the first step, we give our main argument, which establishes a Central Limit Theorem type result; This step shows the desired asymptotic normality in Proposition 3. In the next steps, we verify Claims 1 and 2 separately.

Step 1: A CLT-type result We give the asymptotic analysis for the aggregation of the last term in (3.19) over the bins, given by

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{m=1}^{M}\left\{\frac{1}{\zeta_m}\bar{x}_m\left[\sum_{i=1}^{n}w_{im}[Z_i(\tau)-v(\tau,X_i)]\right]\right\}\\
=\;&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underbrace{[Z_i(\tau)-v(\tau,X_i)]\left[\sum_{m=1}^{M}\frac{w_{im}}{\zeta_m}\bar{x}_m\right]}_{O_{in}},
\end{aligned}$$

which holds by exchanging the order of summation. Since ζ_m and $\bar{x}_m$ are deterministic, the term $\sum_{m=1}^{M}(\zeta_m)^{-1}w_{im}\bar{x}_m$ in each O_{in} only depends on the bin that X_i falls into. For fixed n and M, the random vectors O_{in} are *i.i.d.* with mean 0 across $i = 1, \ldots, n$; and we apply the multivariate Lindeberg-Feller Central Limit Theorem for triangular arrays (E.g., Theorem 2.27 of *Van der Vaart* (2000)) in our proof below.

We check the first Lindeberg conditions here. In our setting it suffices to show:

$$\mathrm{E}\|O_{in}\|^2\mathbf{1}\{\|O_{in}\| \geq \varepsilon\sqrt{n}\} \to 0, \tag{3.22}$$

for all fixed $\varepsilon > 0$ as $n \to \infty$. Since $\bar{x}_m$ is uniformly bounded, we have that

$$\mathrm{E}\|O_{in}\|^{2+\delta_0} \lesssim \mathrm{E}|Z_i(\tau) - v(\tau, X_i)|^{2+\delta_0} \lesssim \mathrm{E}[|q(\tau, X)|^{2+\delta_0}] + \mathrm{E}\left[|Y^+|^{2+\delta_0}\right] < \infty,$$

which follows from Condition G-Y1'. Furthermore note $|x|^2\mathbf{1}\{|x| \geq a\} \leq a^{-\delta}|x|^{2+\delta}$, the Lindeberg condition (3.22) then follows from the Markov inequality.

Next we calculate the variance of each O_{in}. Parallel to the one-sample case in Corollary 2 of Chapter 2, we have that

$$\begin{aligned} \mathrm{var}[Z_i(\tau) - v(\tau, X_i) \mid X_i = x] &= \frac{\mathrm{var}(Y \mid X = x, Y \geq q(\tau, x)) + \tau[v(\tau, x) - q(\tau, x)]^2}{1 - \tau} \\ &\triangleq \sigma_\tau^2(x). \end{aligned} \tag{3.23}$$

Therefore from O_{in} in the beginning of Step 1, we have

$$\begin{aligned} \mathrm{var}(O_{in}) &= \mathrm{E}_X\left\{\left[\sum_{m=1}^{M}\frac{w_{im}}{\zeta_m^2}\bar{x}_m\bar{x}_m^T\right]\mathrm{E}_{Y|X}\left[Z_i(\tau) - v(\tau, X_i)\right]^2\right\} \\ &= \mathrm{E}_X\left\{\sigma_\tau^2(X)\left[\sum_{m=1}^{M}\frac{w_{im}}{\zeta_m^2}\bar{x}_m\bar{x}_m^T\right]\right\} \\ &\to \mathrm{E}_X\left\{\frac{\sigma_\tau^2(X)}{[v(\tau, X) - q(\tau, X)]^2}XX^T\right\}, \end{aligned}$$

as $n \to \infty$, which follows from Lemma 5. Hence, it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} O_{in} \xrightarrow{d} \mathrm{N}\left(0,\ \Omega_1\right),$$

by the Lindeberg CLT, where Ω_1 is given in Theorem III.2.

Together with Claims 1 and 2, we have proved that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{M} \left\{ \frac{\hat{\gamma}_m}{\zeta_m} \bar{x}_m \left[\hat{v}(\tau, \bar{x}_m) - v(\tau, \bar{x}_m) \right] \right\} \xrightarrow{d} \mathrm{N}\left(0,\ \Omega_1\right),$$

from the decomposition in (3.19). Therefore Proposition 3 holds.

Step 2: Verification of Claim 1 Now we check Claim 1. The left hand side of (3.20) can be written as:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ [Z_i(\tau) - v(\tau, X_i)] \underbrace{\left[\sum_{m=1}^{M} \frac{w_{im}}{\zeta_m} \bar{x}_m S_{1m}^T \boldsymbol{S}_{2m}^{-1} (X_i - \bar{x}_m) \right]}_{V_{in}} \right\}, \tag{3.24}$$

by re-arranging the summation.

We use Markov inequality to bound (3.24); To this end, we calculate the variance for each term of (3.24). Note that V_{in} depends on the covariates but not the response, by conditioning on X first we have:

$$\mathrm{E} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Z_i(\tau) - v(\tau, X_i)] V_{in} \right\|^2 = \frac{1}{n} \mathrm{E} \left(\sum_{i=1}^{n} \sigma_\tau^2(X_i) \|V_{in}\|^2 \right) \lesssim \frac{1}{n} \sum_{i=1}^{n} \mathrm{E} \|V_{in}\|^2,$$

since $\sigma_\tau^2(x)$ in (3.23) is bounded. Following the above displayed equation, we can

further expand the variance of V_{in} as:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\|V_{in}\|^2 &= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{m=1}^{M}\frac{w_{im}\|\bar{x}_m\|^2}{\zeta_m^2}\|S_{1m}^T\boldsymbol{S}_{2m}^{-1}(X_i-\bar{x}_m)\|^2\right] \\
&\lesssim \mathrm{E}\left\{\sum_{m=1}^{M}S_{1m}^T\boldsymbol{S}_{2m}^{-1}\underbrace{\left[n^{-1}\sum_{i=1}^{n}w_{im}(X_i-\bar{x}_m)(X_i-\bar{x}_m)^T\right]}_{\boldsymbol{S}_{2m}}\boldsymbol{S}_{2m}^{-1}S_{1m}\right\} \\
&= \mathrm{E}\left[\sum_{m=1}^{M}S_{1m}^T\boldsymbol{S}_{2m}^{-1}S_{1m}\right] \qquad (3.25) \\
&= o(1),
\end{aligned}$$

where the definition of $\boldsymbol{S}_{2m}$ is in (3.5), and the convergence to the $o(1)$ term in the end follows from the Dominated Convergence theorem as outlined below. First, the term inside the expectation of (3.25) is bounded by

$$\sum_{m=1}^{M}S_{1m}^T\boldsymbol{S}_{2m}^{-1}S_{1m}\le\sum_{m=1}^{M}S_{0m}=n^{-1}\sum_{m=1}^{M}\sum_{i=1}^{n}\mathbf{1}\{X_i\in A_m\}=1,$$

since $\hat{\gamma}_m=S_{0m}-S_{1m}^T\boldsymbol{S}_{2m}^{-1}S_{1m}\ge 0$. Second, we have from Lemma 8 that

$$\left|\sum_{m=1}^{M}S_{1m}^T\boldsymbol{S}_{2m}^{-1}S_{1m}\right|\le\sum_{m=1}^{M}S_{0m}\left|1-\frac{\hat{\gamma}_m}{S_{0m}}\right|=o_{\mathrm{P}}(1).$$

The convergence in expectation of (3.25) then follows.

Therefore, Claim 1 holds by applying Markov inequality for (3.24).

Step 3: Verification of Claim 2 By the construction of $\hat{v}$ and $\tilde{v}$ in (3.17) and (3.18), we have

$$\hat{v}(\tau,\bar{x}_m)-\tilde{v}(\tau,\bar{x}_m)=(n\hat{\gamma}_m)^{-1}\left[\sum_{i=1}^{n}w_{im}[1-S_{1m}^T\boldsymbol{S}_{2m}^{-1}(X_i-\bar{x}_m)][\hat{Z}_i(\tau)-Z_i(\tau)]\right].$$

Similar to the proof of Lemma 1 in *Olma* (2021), we consider the following decomposition of $\hat{Z}_i(\tau) - Z_i(\tau)$:

$$\begin{aligned}(1-\tau)[\hat{Z}_i(\tau) - Z_i(\tau)] &= [Y_i - q(\tau, X_i)]\left\{\mathbf{1}[Y_i \geq \hat{q}(\tau, X_i)] - \mathbf{1}[Y_i \geq q(\tau, X_i)]\right\} \\ &\quad + (q(\tau, X_i) - \hat{q}(\tau, X_i)) \cdot (\tau - \mathbf{1}[Y_i < q(\tau, X_i)]) \\ &\quad + (q(\tau, X_i) - \hat{q}(\tau, X_i)) \cdot \{\mathbf{1}[Y_i \geq \hat{q}(\tau, X_i)] - \mathbf{1}[Y_i \geq q(\tau, X_i)]\} \\ &\triangleq u_{1i}(\tau) + u_{2i}(\tau) + u_{3i}(\tau), . \end{aligned} \tag{3.26}$$

where we sometimes omit the index τ in this proof. Using the above two displayed equations, and by re-arranging the order of summation in (3.21) of Claim 2, we have

$$\begin{aligned} & \sqrt{n}\sum_{m=1}^{M}\left\{\frac{\hat{\gamma}_m}{\zeta_m}\bar{x}_m\left[\hat{v}(\tau, \bar{x}_m) - \tilde{v}(\tau, \bar{x}_m)\right]\right\} \\ = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{[\hat{Z}_i(\tau) - Z_i(\tau)]\underbrace{\left[\sum_{m=1}^{M}\frac{w_{im}}{\zeta_m}[1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \bar{x}_m)]\bar{x}_m\right]}_{\kappa_i}\right\} \\ = & \frac{1}{\sqrt{n}(1-\tau)}\left(\sum_{i=1}^{n}u_{1i}\kappa_i + \sum_{i=1}^{n}u_{2i}\kappa_i + \sum_{i=1}^{n}u_{3i}\kappa_i\right) \\ \triangleq & U_{1n} + U_{2n} + U_{3n}, \end{aligned} \tag{3.27}$$

where $U_{jn} = [\sqrt{n}(1-\tau)]^{-1}\sum_{i=1}^{n}u_{ji}\kappa_i$ and u_{ji} is defined in (3.26). To check Claim 2, it suffices to consider the three terms in (3.27) separately.

We consider U_{2n} first. Separating κ_i into two sums for the terms 1 and $S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i -$

$\bar{x}_m$), we have

$$
\begin{aligned}
& \|(1-\tau)\sqrt{n}U_{2n}\| \\
\lesssim\ & \|\sum_{i=1}^{n} u_{2i}\kappa_i\| \\
\leq\ & \left\|\sum_{m=1}^{M}\sum_{i=1}^{n}\left(\frac{w_{im}}{\zeta_m}u_{2i}\bar{x}_m\right)\right\| + \left\|\sum_{m=1}^{M}\sum_{i=1}^{n}\left[\frac{w_{im}}{\zeta_m}S_{1m}^T\boldsymbol{S}_{2m}^{-1}(X_i-\bar{x}_m)\cdot u_{2i}\bar{x}_m\right]\right\| \\
\lesssim\ & \sup_{m=1,\ldots,M}\left|\frac{\sum_{i=1}^{n} w_{im}u_{2i}}{n\hat{\pi}_m}\right| \cdot \sum_{m=1}^{M}\frac{n\hat{\pi}_m\|\bar{x}_m\|}{|\zeta_m|} \\
& + \sup_{m=1,\ldots,M}\left\|\frac{\sum_{i=1}^{n} w_{im}\left[\frac{X_i-\bar{x}_m}{\bar{h}_m}\right]u_{2i}}{n\hat{\pi}_m}\right\| \cdot \sum_{m=1}^{M}\frac{n\hat{\pi}_m\cdot\|\bar{x}_m\|\cdot\|\bar{h}_m\cdot S_{1m}^T\boldsymbol{S}_{2m}^{-1}\|}{|\zeta_m|} \\
\lesssim\ & Op(g_{2n})\sum_{m=1}^{M}(n\hat{\pi}_m) + O_{\mathrm{P}}(g_{2n})\sum_{m=1}^{M}\left[n\hat{\pi}_m o_{\mathrm{P}}(1)\right],
\end{aligned}
$$

where we have used the fact that $\|\bar{x}_m\|/|\zeta_m|$ is bounded; in the last inequality, the $O_{\mathrm{P}}(g_{2n})$ terms follow from Condition G-Q, and the $o_{\mathrm{P}}(1)$ term uses the bound of $\|\bar{h}_m\cdot S_{1m}^T\boldsymbol{S}_{2m}^{-1}\|$ in Lemma 8. Noting that $\sum_{m=1}^{M}\hat{\pi}_m = 1$, we conclude that $U_{2n} = O_{\mathrm{P}}(n^{1/2}g_{2n})$.

Next we consider U_{1n} and U_{3n}. Noting that $w_{im}|1-S_{1m}^T\boldsymbol{S}_{2m}^{-1}(X_i-\bar{x}_m)| = 1+o_{\mathrm{P}}(1)$ uniformly as in Lemma 8, we have from (3.27) that

$$
\begin{aligned}
\|(1-\tau)\sqrt{n}U_{1n}\| &= \left\|\sum_{i=1}^{n} u_{1i}\kappa_i\right\| \\
&\leq \sum_{i=1}^{n}\sum_{m=1}^{M}\frac{w_{im}|u_{1i}|\cdot\|\bar{x}_m\|}{|\zeta_m|}[1+o_{\mathrm{P}}(1)] \\
&\lesssim_{\mathrm{P}} \sup_{m=1,\ldots,M}\left[\frac{\sum_{i=1}^{n} w_{im}|u_{1i}|}{n\hat{\pi}_m}\right]\cdot\sum_{m=1}^{M} n\hat{\pi}_m \\
&= n\cdot\sup_{m=1,\ldots,M}\left[\frac{\sum_{i=1}^{n} w_{im}|u_{1i}|}{n\hat{\pi}_m}\right],
\end{aligned}
$$

since $\sum_{m=1}^{M} \hat{\pi}_n = 1$. Similarly

$$\|(1-\tau)\sqrt{n}U_{3n}\| \lesssim_{\mathrm{P}} n \cdot \sup_{m=1,\ldots,M} \left[\frac{\sum_{i=1}^{n} w_{im}|u_{3i}|}{n\hat{\pi}_m} \right]$$

Therefore, it follows directly from Lemma 9 that $U_{1n} = O_{\mathrm{P}}(n^{1/2}g_{1n}^2)$ and $U_{3n} = O_{\mathrm{P}}(n^{1/2}g_{1n}^2)$, hence Claim 3 holds. The proof of Proposition 3 is now complete.

□

3.7.4.2 Proof of Proposition 4

We define some additional notations. Let

$$\kappa_{im} = [1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \bar{x}_m)], \tag{3.28}$$

and let

$$A_{0m} = \sup_{i=1,\ldots,n} |w_{im}\kappa_{im}|, \quad A_{1m} = \sum_{i=1}^{n} w_{im}|\kappa_{im}|, \quad A_{2m} = \sum_{i=1}^{n} w_{im}\kappa_{im}^2. \tag{3.29}$$

Proof. Following the same calculation in (3.19), for each bin A_m we have:

$$\begin{aligned}
[\hat{v}(\tau, \tilde{x}_m) - v(\tau, \tilde{x}_m)] &= [\hat{v}(\tau, \tilde{x}_m) - \tilde{v}(\tau, \tilde{x}_m)] \\
&\quad + \hat{\gamma}_m^{-1} S_{0m} \left[\frac{\sum_{i=1}^{n} w_{im}[Z_i(\tau) - v(\tau, X_i)][1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \tilde{x}_m)]}{\sum_{i=1}^{n} w_{im}} \right] \\
&\triangleq \frac{S_{0m}}{\hat{\gamma}_m} \left[B_q(\tau, m) + C(\tau, m) \right],
\end{aligned} \tag{3.30}$$

where $B_q(\tau, m) = S_{0m}^{-1}\hat{\gamma}_m[\hat{v}(\tau, \tilde{x}_m) - \tilde{v}(\tau, \tilde{x}_m)]$ corresponds to the bias term that originates from using the estimated quantile in $\hat{Z}_i(\tau)$, and $C(\tau, m)$ has mean zero. Under the conditions of Theorem III.2, we give the following claims, which we verify at the end of the proof.

Claim 1:

$$\sup_{m=1,\ldots,M} |B_q(\tau, m)| = O_{\mathrm{P}}(g_{1n}^2 + g_{2n}),$$

where g_{1n} and g_{2n} are in Condition G-Q.

Claim 2:

$$\Pr\left(\sup_{m=1,\ldots,M} A_{0m} \geq 2\right) \leq \frac{1}{n^3}, \quad A_{1m} \leq nS_{0m}, \text{ and } \quad A_{2m} \leq nS_{0m},$$

where the quantities A_{jm} $(j = 1, 2, 3)$ are defined in (3.29).

Claim 1 shows that the bias in the initial SQ estimator is asymptotically negligible, Claim 2 is also useful but more technical. Following Claims 1 and 2, the proof proceeds in 5 steps. In steps 1 through 3, we establish the main argument that:

$$\sup_{m=1,\ldots,M} |C(\tau, m)| = O_{\mathrm{P}}(r_n);$$

In step 1, we give a truncation argument similar to *Mack and Silverman* (1982); In step 2, we derive exponential inequalities for the truncated process; Step 3 gives some auxiliary calculations that completes the proof. In the final two steps we verify Claims 1 and 2.

Step 1: Truncation For any sequence $b_n > 0$ that satisfies: (i) $\sum_{n=1}^{\infty} b_n^{-2-\delta_0} < +\infty$ for δ_0 in Condition G-Y1'; (ii) b_n is monotonically increasing; and (iii) $b_n \to +\infty$, we define the truncated variable:

$$Z_i^{(B)}(\tau) = \frac{[Y_i - q(\tau, X_i)]\mathbf{1}\{0 \leq Y_i - q(\tau, X_i) \leq b_n\}]}{1 - \tau} + q(\tau, X_i),$$

and the corresponding truncated process:

$$C^{(B)}(\tau, m) = (nS_{0m})^{-1} \sum_{i=1}^{n} \left\{ w_{im} \left[Z_i^{(B)}(\tau) - v(\tau, X_i) \right] \left[1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1} (X_i - \tilde{x}_m) \right] \right\}.$$

We shall give the precise choice of b_n in step 2 below. For sufficiently large n, in the following we show that $C^{(B)}(\tau, m)$ is equivalent to $C(\tau, m)$ with probability one.

Comparing $C^{(B)}(\tau, m)$ with $C(\tau, m)$ in (3.30), we see that $C^{(B)}(\tau, m) \neq C(\tau, m)$ only when $Y_i - q(\tau, X_i) \geq b_n$ for some $i = 1, \ldots, n$; and we can calculate this probability:

$$\Pr\left([Y_i - q(\tau, X_i)] \geq b_n \right) \leq \frac{\mathrm{E}[Y_i - q(\tau, X_i)]^{2+\delta_0} \mathbf{1}[Y_i \geq q(\tau, X_i)]}{b_n^{2+\delta_0}} \lesssim b_n^{-2-\delta_0},$$

from Chebyshev's inequality and Condition G-Y1'. Following Proposition 1 of *Mack and Silverman* (1982), under our choice of b_n we have

$$\Pr\left(\liminf_{n \to \infty} \left\{ \sup_{m=1,\ldots,M} |C^{(B)}(\tau, m) - C(\tau, m)| = 0 \right\} \right) = 1,$$

where $\liminf_{n\to\infty}$ denotes the limit infimum for a sequence of events.

Step 2: Exponential inequality Here we derive exponential tail bounds for the centered truncated sequence $C^{(B)}(\tau, m) - \mathrm{E}[C^{(B)}(\tau, m)]$. Note $C^{(B)}(\tau, m)$ does not have mean zero after truncation. With κ_{im} in (3.28), we write

$$C^{(B)}(\tau, m) = (nS_{0m})^{-1} \sum_{i=1}^{n} \left\{ w_{im} \kappa_{im} \left[Z_i^{(B)}(\tau) - v(\tau, X_i) \right] \right\}.$$

For a small enough $\varepsilon_3 > 0$, let the truncation threshold satisfy

$$b_n \asymp n^{\frac{1}{2+\delta_0} + \varepsilon_3}, \tag{3.31}$$

where δ_0 is in Condition G-Y1'; it is easy to check that this choice of b_n satisfies the requirements in Step 1 of the proof.

We apply Bernstein inequality on the truncated and centered process; to this end, we compute some key quantities below. Conditional on the covariates X, we have

$$\sup_{x\in\mathcal{X}} \operatorname{var}\left[Z_i^{(B)}(\tau) - v(\tau, X_i) \,\middle|\, X_i = x \right] \le C_1 < +\infty,$$

for some constant C_1, which follows from Condition G-Y1'; hence

$$\sum_{i=1}^{n} \operatorname{var}\left[w_{im}\kappa_{im}[Z_i^{(B)}(\tau) - v(\tau, X_i)] \,\middle|\, X \right] \le C_1 A_{2m} \le nC_1 S_{0m}.$$

Furthermore, each of summands in $C^{(B)}(\tau, m)$ can be bounded by

$$\left| w_{im}\kappa_{im}[Z_i^{(B)}(\tau) - v(\tau, X_i)] \right| \lesssim A_{0m} b_n.$$

Refer to Claim 2 for the properties of A_{0m} and A_{2m}.

Now, a direct application of the (conditional on X) Bernstein inequality (e.g., Theorem 2.8.4 of *Vershynin* (2018)) and a union bound gives

$$\begin{aligned}
& \Pr\left(\sup_{m=1,\dots,M} \left| C^{(B)}(\tau, m) - \mathrm{E}[C^{(B)}(\tau, m)] \right| \ge M_1 r_n \,\middle|\, X \right) \\
\le\ & \sum_{m=1}^{M} \Pr\left(\left| \sum_{i=1}^{n} w_{im}\kappa_{im} \left[Z_i^{(B)}(\tau) - \mathrm{E}[Z_i^{(B)}(\tau)] \right] \right| \ge M_1 \cdot nS_{0m} r_n \,\middle|\, X \right) \\
\le\ & 2\exp\left\{ \log n - \frac{(M_1^2 n r_n^2/2) \cdot \inf_m S_{0m}}{C_1 + (M_1 b_n r_n/3) \cdot \sup_m A_{0m}} \right\} \\
\lesssim\ & 2\exp\left\{ \log n - \frac{(M_1^2 n r_n^2/2) \cdot \inf_m S_{0m}}{C_1(1 + \sup_m A_{0m})} \right\},
\end{aligned} \tag{3.32}$$

for sufficiently large n, where the $\log n$ factor comes from $M \lesssim \bar{h}^{-p} \le n$ under

Condition G-A1; the last inequality follows since

$$b_n r_n = \sqrt{\frac{\log n}{n^{1-2/(2-\delta_0)-2\varepsilon_3}\,\bar{h}^p}} \to 0,$$

under Condition G-A1, with b_n in (3.31) and r_n in Proposition 4

Here we give the unconditional tail bound from the conditional one in (3.32). Let Γ denote the event that $\sup_m A_{0m} \leq 2$ and $\inf_m |\bar{h}_m^{-p} S_{0m}| \geq \varepsilon_2$ for some $\varepsilon_2 > 0$; With Lemma 8 and Claim 2, we have $\Pr(\Gamma^c) \lesssim n^{-3}$. With the law of total expectation applied to (3.32), the unconditional tail bound is:

$$\begin{aligned}
& \Pr\left(\sup_{m=1,\ldots,M} |C^B(\tau,m) - \mathrm{E}[C^B(\tau,m)]| \geq M_1 r_n\right) \\
\leq\ & \mathrm{E}_X\left[2\exp\left\{\log n - \frac{(M_1^2 n r_n^2/2)\cdot \inf_m S_{0m}}{C_1(1+\sup_m A_{0m})}\right\}\cdot \mathbf{1}\{\Gamma\}\right] + \mathrm{E}\left[\exp\{\log n\}\cdot \mathbf{1}\{\Gamma^c\}\right] \\
\lesssim\ & \mathrm{E}_X\left[2\exp\left\{\log n - \frac{\log n \cdot M_1^2\varepsilon_2/2}{3C_1}\right\}\right] + n\Pr\left(\Gamma^c\right) \\
\lesssim\ & \frac{1}{n}, \qquad (3.33)
\end{aligned}$$

for sufficiently large M_1 since C_1 and ε_2 are fixed.

Step 3: Final calculations Noting that

$$\left|C(\tau,m) - C^{(B)}(\tau,m)\right| = (nS_{0m})^{-1}\sum_{i=1}^n w_{im}|\kappa_{im}|\left\{\frac{[Y_i - q(\tau,X_i)]\mathbf{1}\{b_n \leq Y_i - q(\tau,X_i)\}}{1-\tau}\right\},$$

the expectation of $C^{(B)}$ can be bounded by

$$\begin{aligned}
& \mathrm{E}\left[|C(\tau,m) - C^{(B)}(\tau,m)| \mid X = x\right] \\
= \;& (nS_{0m})^{-1} \sum_{i=1}^{n} \frac{w_{im}|\kappa_{im}|}{1-\tau} \mathrm{E}\left\{[Y_i - q(\tau,X_i)]\mathbf{1}[Y_i - q(\tau,X_i) \geq b_n] \;\middle|\; X = x\right\} \\
\leq \;& (nS_{0m})^{-1} \sum_{i=1}^{n} \frac{w_{im}|\kappa_{im}|}{1-\tau} \mathrm{E}\left\{\frac{[Y_i - q(\tau,X_i)]^{2+\delta_0}\mathbf{1}[Y_i \geq q(\tau,X_i)]}{b_n^{1+\delta_0}} \;\middle|\; X = x\right\} \\
\lesssim \;& (nS_{0m})^{-1} A_{1m} \cdot b_n^{-1-\delta_0} \\
\leq \;& b_n^{-1-\delta_0},
\end{aligned}$$

where the second inequality follows from Chebyshev's inequality, the third inequality follows from the moment bound in Condition G-Y1' and the last inequality from Claim 2. Taking expectation again with expect to X we obtain:

$$\sup_{m=1,\ldots,M} \mathrm{E}\left[|C^{(B)}(\tau,m) - C(\tau,m)|\right] \lesssim b_n^{-1-\delta_0} \lesssim r_n,$$

from the choice of b_n in (3.31) and r_n in Proposition 4.

Combining the above expectation bounds with the results of steps 1 and 2, we have

$$\sup_{m=1,\ldots,M} |C(\tau,m)| = O_{\mathrm{P}}\left(r_n\right).$$

Together with Claim 1, we've proved that

$$\sup_{m=1,\ldots,M} |B_q(\tau,m) + C(\tau,m)| = O_{\mathrm{P}}(g_{1n}^2 + g_{2n} + r_n) = O_{\mathrm{P}}(r_n),$$

since $g_{1n}^2 + g_{2n} \ll n^{-1/2} \ll r_n$ in Proposition 4. Furthermore, note from Lemma 8 we have $\hat{\gamma}_m^{-1} S_{0m} = 1 + o_{\mathrm{P}}(1)$ uniformly over $m = 1, \ldots, M$. The conclusion of Proposition 4 then holds from the decomposition (3.30).

Step 4: Verification of Claim 1 We follow the same decomposition of $\hat{Z}_i(\tau) - Z_i(\tau)$ as in (3.26) in the proof of Proposition 3. From the definition of B_q in (3.30) we have

$$\begin{aligned} B_q(\tau, m) &= \frac{1}{(1-\tau)} \left\{ (nS_{0m})^{-1} \sum_{i=1}^{n} [u_{1i}(\tau) + u_{2i}(\tau) + u_{3i}(\tau)] w_{im} \kappa_{im} \right\} \\ &\triangleq \frac{1}{(1-\tau)} \left[U_{1n}(\tau, m) + U_{2n}(\tau, m) + U_{3n}(\tau, m) \right], \end{aligned} \tag{3.34}$$

where $u_{ji}(\tau)$ is defined in (3.26). We consider the three terms separately.

We consider $U_{2n}(\tau, m)$ first. By separating κ_{im} in (3.28) into two parts, we have:

$$\begin{aligned} \sup_{m=1,\ldots,M} |U_{2n}(\tau, m)| &\leq \sup_{m=1,\ldots,M} \left| \frac{\sum_{i=1}^{n} w_{im} [S_{1m}^T \boldsymbol{S}_{2m} (X_i - \bar{x}_m)] u_{2i}(\tau)}{nS_{0m}} \right| \\ &\quad + \sup_{m=1,\ldots,M} \left| \frac{\sum_{i=1}^{n} w_{im} u_{2i}(\tau)}{nS_{0m}} \right| \\ &\leq \sup_{m=1,\ldots,M} \left\| \frac{\sum_{i=1}^{n} w_{im} \left[\frac{X_i - \bar{x}_m}{\bar{h}_m} \right] u_{2i}(\tau)}{nS_{0m}} \right\| \cdot \sup_{m=1,\ldots,M} \| \bar{h}_m S_{1m}^T \boldsymbol{S}_{2m}^{-1} \| \\ &\quad + O_{\mathrm{P}}(g_{2n}) \\ &= O_{\mathrm{P}}(g_{2n}), \end{aligned}$$

which follows from Condition G-Q and Lemma 8.

For U_{1n}, we have

$$\begin{aligned} \sup_{m=1,\ldots,M} |U_{1n}(\tau, m)| &\leq \sup_{m=1,\ldots,M} \frac{\sum_{i=1}^{n} w_{im} \kappa_{im} |u_{1i}(\tau)|}{nS_{0m}} \\ &\leq \sup_{m=1,\ldots,M} A_{0m} \cdot \sup_{m=1,\ldots,M} \frac{\sum_{i=1}^{n} w_{im} |u_{1i}(\tau)|}{\sum_{i=1}^{n} w_{im}} \\ &= O_{\mathrm{P}}(g_{1n}^2), \end{aligned}$$

which follows from Claim 2 and Lemma 9. Similarly,

$$\sup_{m=1,\ldots,M} |U_{3n}(\tau, m)| = O_{\mathrm{P}}(g_{1n}^2).$$

Combining the results with $U_{2n}(\tau, m)$, we have verified

$$B_q(\tau, m) = O_{\mathrm{P}}(g_{1n}^2 + g_{2n}),$$

hence Claim 1 holds. The proof is now complete.

Step 5: Verification of Claim 2 We check the conditions for A_{0m}, A_{1m}, and A_{2m} separately. For A_{2m}, standard algebra gives

$$\begin{aligned} A_{2m} &= \sum_{i=1}^{n} w_{im} + S_{1m}^T \boldsymbol{S}_{2m}^{-1} \left[\sum_{i=1}^{n} w_{im} (X_i - \tilde{x}_m)(X_i - \tilde{x}_m)^T \right] \boldsymbol{S}_{2m}^{-1} S_{1m} \\ &\quad - \ 2 S_{1m}^T \boldsymbol{S}_{2m}^{-1} \sum_{i=1}^{n} w_{im} (X_i - \tilde{x}_m) \\ &= n S_{0m} - n S_{1m}^T \boldsymbol{S}_{2m}^{-1} S_{1m} \\ &\leq n S_{0m}, \end{aligned}$$

similar to how we obtain (3.25). For A_{1m}, Cauchy–Schwartz inequality gives

$$A_{1m} \leq \sqrt{A_{2m} \sum_{i=1}^{n} w_{im}} \leq n S_{0m}.$$

For A_{0m}, note

$$\begin{aligned}
\Pr\left(\sup_{m=1,\ldots,M} A_{0m} \geq 2\right) &\leq \Pr\left(\sup_{\substack{i=1,\ldots,n \\ m=1,\ldots,M}} \left|w_{im} S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \tilde{x}_m)\right| \geq 1\right) \\
&\leq \Pr\left(\sup_{m=1,\ldots,M} \left\|\bar{h}_m \cdot S_{1m}^T \boldsymbol{S}_{2m}^{-1}\right\| \geq 1\right) \\
&\leq \frac{1}{n^3},
\end{aligned}$$

which follows from Lemma 8. We have verified Claim 2, and hence the proof of Proposition 4 is complete.

□

3.7.4.3 Proof of Proposition 5

Proof. For each $s \in (0,1)$, the decomposition in (3.30) gives

$$\begin{aligned}
[\hat{v}(s,\tilde{x}_m) - v(s,\tilde{x}_m)] &= \frac{S_{0m}}{\hat{\gamma}_m}\left[B_q(s,m) + C(s,m)\right] \\
&\quad + \hat{\gamma}_m^{-1} S_{0m}\left[\frac{\sum_{i=1}^n w_{im}\kappa_{im}[v(s,X_i) - v(s,\tilde{x}_m)]}{\sum_{i=1}^n w_{im}}\right] \\
&= \frac{S_{0m}}{\hat{\gamma}_m}\left[B_q(s,m) + C(s,m) + B_{np}(s,m)\right], \qquad (3.35)
\end{aligned}$$

where the additional term $B_{np}(s,m)$ corresponds to the non-parametric binning bias; In (3.30), this bias does not exist because $v(\tau,x)$ is linear in x. Following the above three-term decomposition,

$$\begin{aligned}
&[\hat{v}(s,\tilde{x}_m) - v(s,\tilde{x}_m)] - [\hat{v}(\tau,\tilde{x}_m) - v(\tau,\tilde{x}_m)] \\
= \ &\frac{S_{0m}}{\hat{\gamma}_m}\left\{[B_q(s,m) - B_q(\tau,m)] + [B_{np}(s,m) - B_{np}(\tau,m)] + [C(s,m) - C(\tau,m)]\right\}.
\end{aligned}$$

Recall that

$$r_n = \sqrt{\frac{\log n}{n\underline{h}^p}},$$

from Proposition 4. We give two claims below, the verification of which is at the end of the proof.

Claim 1:

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |B_q(s,m)| = O_{\mathrm{P}}(g_{1n}^2 + g_{2n} + r_n^2),$$

for any fixed $B > 0$, where where g_{1n} and g_{2n} are in Condition G-Q.

Claim 2:

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |B_{np}(s,m)| = o_{\mathrm{P}}\left(n^{-1/2}\right),$$

for any fixed $B > 0$, if any one of the requirements in Condition G-A2 holds.

Claims 1 and 2 show that the bias terms are uniformly (over s) negligible; They are stronger than those in the proof of Proposition 4. Following Claims 1 and 2, the proof proceeds in 5 steps. In steps 1 to 3, we establish the main argument:

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |C(s,m) - C(\tau,m)| = o_{\mathrm{P}}\left(n^{-1/2}\right).$$

In steps 4 and 5 we verify Claims 1 and 2.

Step 1: Decomposition We use the following decomposition of $Z_i(s)$ defined in (3.4):

$$\begin{aligned}
& (1-s)[Z_i(s) - v(s, X_i)] - (1-\tau)[Z_i(\tau) - v(\tau, X_i)] \\
= \; & \underbrace{(q(s, X_i) - y_i)\mathbf{1}[q(\tau, X_i) \le y_i \le q(s, X_i)]}_{u_{4i}(s)} \\
& - \underbrace{\{(1-s)[v(s, X_i) - q(s, X_i)] - (1-\tau)[v(\tau, X_i) - q(s, X_i)]\}}_{\mathrm{E}_i[u_{4i}(s)]} \\
& + \underbrace{[q(\tau, X_i) - q(s, X_i)]\{\tau - \mathbf{1}[y_i \le q(\tau, X_i)]\}}_{u_{5i(s)}} \\
\triangleq \; & u_{4i}(s) - \mathrm{E}_i[u_{4i}(s)] + u_{5i}(s),
\end{aligned}$$

where we define $\mathrm{E}_i[\cdot]$ as the conditional expectation given $X = X_i$; and note $\mathrm{E}_i[u_{5i}(\tau, s)] = 0$. Therefore from the definition of $C(s, m)$ in (3.30) and κ_{im} in (3.28), we have

$$\begin{aligned}
(1-s)C(s,m) - (1-\tau)C(\tau, m) = & \underbrace{(nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}\kappa_{im}\{u_{4i}(s) - \mathrm{E}_i[u_{4i}(s)]\}}_{U_{4n}(s,m)} \\
& + \underbrace{(nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}\kappa_{im}u_{5i}(s)}_{U_{5n}(s,m)}.
\end{aligned} \tag{3.36}$$

Furthermore, we separate κ_{im} into:

$$\kappa_{im} = \kappa_{im}\mathbf{1}[\kappa_{im} \ge 0] + \kappa_{im}\mathbf{1}[\kappa_{im} < 0] \triangleq \kappa_{im}^{(+)} - \kappa_{im}^{(-)},$$

and correspondingly we define $U_{4n}(s, m) = U_{4n}^{(+)}(s, m) + U_{4n}^{(-)}(s, m)$

$$\begin{aligned}
U_{4n}^{(+)}(s, m) &= (nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}\kappa_{im}^{(+)}\{u_{4i}(s) - \mathrm{E}_i[u_{4i}(s)]\}, \\
U_{4n}^{(-)}(s, m) &= (nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}\kappa_{im}^{(-)}\{u_{4i}(s) - \mathrm{E}_i[u_{4i}(s)]\}.
\end{aligned}$$

In the following we consider $U_{4n}^{(+)}(s,m)$, $U_{4n}^{(-)}(s,m)$ and $U_{5n}(s,m)$ separately.

Step 2: Bound for U_{4n} Let $s_+ = \tau + Br_n$, it suffices to consider the convergence of $U_{4n}^{(+)}(s,m)$ over the over $s \in [\tau, s_+]$. The result for $s < \tau$ and/or $U_{4n}^{(-)}(s,m)$ follows analogously.

We use a monotonicity argument to show the uniformity over s. Since $u_{4i}(s)$ is monotonically increasing in s, we have the sandwich-type bound for $U_{4n}^{(+)}$:

$$\frac{\sum_{i=1}^{n} w_{im}\kappa_{im}^{(+)} \{u_{4i}(\tau) - \mathrm{E}_i[u_{4i}(s_+)]\}}{nS_{0m}} \leq U_{4n}^{(+)}(s,m) \leq \frac{\sum_{i=1}^{n} w_{im}\kappa_{im}^{(+)} \{u_{4i}(s_+) - \mathrm{E}_i[u_{4i}(\tau)]\}}{nS_{0m}},$$

which holds for all $s \in [\tau, s_+]$. Noting that $u_{4i}(\tau) = 0$, using the monotonicity argument in *Van der Vaart* (2000, Theorem 19.1) gives

$$\sup_{\substack{m=1,\ldots,M \\ s\in[\tau,s_+]}} |U_{4n}^{(+)}(s,m)| \leq \sup_{m=1,\ldots,M} |U_{4n}^{(+)}(s_+,m)| + \sup_{m=1,\ldots,M} \left| (nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}\kappa_{im}^{(+)} \mathrm{E}_i[u_{4i}(s_+)] \right|. \tag{3.37}$$

Next we bound the two terms separately.

For the first term in (3.37), note each of the summand in $U_{4n}^{(+)}(s_+,m)$ is bounded from (3.36), and

$$0 \leq \sum_{i=1}^{n} |w_{im}\kappa_{im}^{(+)} u_{4i}(s_+)|^2 \leq (\underline{f}^{-1}|\tau - s_+|)^2 \sum_{i=1}^{n} w_{im}\kappa_{im}^2 \leq A_{2m}\underline{f}^{-2}B^2 r_n^2,$$

where A_{2m} is in (3.29) and $\underline{f}$ from Condition G-Y1. For any $\varepsilon_4 > 0$, application of the (conditional on X) Hoeffding's inequality and the union bound gives

$$\Pr\left(\sup_{m=1,\ldots,M} |U_{4n}^{(+)}(s_+,m)| \geq \varepsilon_4 n^{-1/2} \middle| X \right) \leq \sum_{m=1}^{M} 2\exp\left\{ -\frac{2n\varepsilon_4^2 \cdot \inf_m S_{0m}}{\underline{f}^{-2}B^2 r_n^2 \cdot \sup_m A_{2m}} \right\}.$$

Since $r_n^{-1} \gg \log(n)$ under Condition G-A1, we can obtain the unconditional tail

bound, which implies

$$\sup_{m=1,\ldots,M} |U_{4n}^{(+)}(s_+, m)| = o_{\mathrm{P}}\left(n^{-1/2}\right),$$

similar to how we obtain (3.33).

Next we consider the conditional expectation on the right hand side of (3.37). From (3.36), each $\mathrm{E}_i[u_{4i}(s_+)]$ is bounded as

$$|\mathrm{E}_i[u_{4i}(s_+)]| \lesssim |q_i(s) - q_i(\tau)|^2 \leq \underline{f}^{-2} B^2 r_n^2;$$

Hence

$$\begin{aligned}
\left|(nS_{0m})^{-1} \sum_{i=1}^{n} w_{im} \kappa_{im}^{(+)} \mathrm{E}_i[u_{4i}(s_+)]\right| &\lesssim (nS_{0m})^{-1} \sum_{i=1}^{n} w_{im} |\kappa_{im}| B^2 r_n^2 \\
&\leq (nS_{0m})^{-1} A_{1m} r_n^2 \\
&= O_{\mathrm{P}}(r_n^2),
\end{aligned}$$

where A_{1m} and its property are in Claim 2 of Proposition 4.

We now conclude from (3.37) that

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau| \leq B \cdot r_n}} |U_{4n}^{(+)}(s, m)| = o_{\mathrm{P}}\left(n^{-1/2}\right),$$

since $r_n^2 = o(n^{-1/2})$ under Condition G-A1.

Step 3: Bound for U_{5n} For any $s, s' \in (0,1)$, from the decomposition in (3.36) we have

$$\begin{aligned} |U_{5n}(s,m) - U_{5n}(s',m)| &= (nS_{0m})^{-1} \max\{\tau, 1-\tau\} \sum_{i=1}^{n} w_{im}|\kappa_{im}||q(s,X_i) - q_i(s',X_i)| \\ &\lesssim \frac{A_{1m}}{nS_{0m}}|s-s'| \\ &\leq |s-s'|, \end{aligned}$$

since $q(s, X_i)$ is Lipschitz continuous in s, and we use the bound for A_{1m} in Claim 2 of Proposition 4.

We use a discretization argument to show the uniform convergence over s. Define

$$\tau - Br_n = s_0 < s_1 < \ldots, s_J = \tau + Br_n,$$

as an equally-spaced grid, such that $s_{j+1} - s_j \asymp n^{-1}$; therefore there are $J \lesssim n$ sub-intervals. Similar to (3.37), we have

$$\begin{aligned} \sup_{\substack{m=1,\ldots,M \\ |s-\tau| \leq B \cdot r_n}} |U_{5n}(s,m)| &\leq \sup_{\substack{m=1,\ldots,M \\ j=0,\ldots,J}} |U_{5n}(s_j,m)| + \sup_{\substack{m=1,\ldots,M \\ j=0,\ldots,J \\ s \in I_j}} |U_{5n}(s,m) - U_{5n}(s_j,m)| \\ &\leq \sup_{\substack{m=1,\ldots,M \\ j=0,\ldots,J}} |U_{5n}(s_j,m)| + \sup_{m=1,\ldots,M} \left| \frac{A_{1m}}{n^2 S_{0m}} \right|, \end{aligned}$$

where the last term is from the beginning of step 3.

Next we apply Bernstein inequality for the discretized $U_{5n}(s_j, m)$. For each $|s - \tau| \leq Br_n$, from (3.36) and the Lipschitz continuity of $q(s, X_i)$ (over s):

$$\begin{aligned} \mathrm{E}[u_{5i}(s) \mid X] = 0, \qquad |w_{im}\kappa_{im}u_{5i}(s)| \leq C_{51}A_{0m}Br_n, \\ \sum_{i=1}^{n} \mathrm{var}[w_{im}\kappa_{im}u_{5i}(s) \mid X] \leq C_{52}(Br_n)^2 A_{2m}, \end{aligned}$$

where C_{51} and C_{52} are two constants, and A_{0m} and A_{2m} are in (3.29). For small enough $\varepsilon_5 > 0$, we apply the (conditional on X) Bernstein inequality as in (3.32), which shows that:

$$\begin{aligned}
& \Pr\left(\sup_{m,j}|U_{5n}(s_j,m)| \geq \varepsilon_5 n^{-1/2}\middle| X\right) \\
\leq\ & 2\sum_{m=1}^{M}\sum_{j=0}^{J}\exp\left\{-\frac{n\varepsilon_5^2 S_{0m}^2}{C_{52}B^2r_n^2A_{2m}+n^{1/2}\varepsilon S_{0m}C_{51}A_{0m}Br_n/3}\right\} \\
\lesssim\ & 2\exp\left\{2\log n-\frac{\varepsilon_5^2\cdot\inf_m S_{0m}}{B^2r_n^2+n^{-1/2}Br_n\cdot\sup_m A_{0m}}\right\}.
\end{aligned}$$

Similar to how we obtain (3.33), we can show that the corresponding unconditional probability is $o(1)$, which implies that:

$$\sup_{\substack{m=1,\ldots,M\\ |s-\tau|\leq B\cdot r_n}} U_{5n}(s,m) = o_{\mathrm{P}}\left(n^{-1/2}\right).$$

Therefore, with the decomposition in (3.36), we have established that

$$\sup_{\substack{m=1,\ldots,M\\ |s-\tau|\leq B\cdot r_n}} |C(s,m)-C(\tau,m)| = o_{\mathrm{P}}\left(n^{-1/2}\right),$$

from steps 1 through 3. Using Claims 1, 2 and Equation (3.35), we would complete the proof of Proposition 5.

Step 4: Verification of Claim 1 We check Claim 1 under two scenarios separately. First, consider the case where the second requirement in Condition G-A2 holds. Then $v(s,x)$ is piece-wise linear in all the bins, for all $s \lesssim n^{-1/4} \ll r_n$. Therefore the same calculations in (3.30) apply, and the non-parametric bias does not exist, i.e.,

$$\sup_{\substack{m=1,\ldots,M\\ |s-\tau|\leq B\cdot r_n}} |B_{np}(s,m)| = 0,$$

which follows from the nature of local-linear estimation (*Fan and Gijbels*, 2018).

In the following, we consider the case when only the first requirement in Condition G-A2 holds. For notational simplicity, let v'_x denote the p-dimensional gradient vector with respect to covariates x, and $\boldsymbol{v}''_{xx}$ be the p by p Hessian matrix. With Taylor expansion at each $\tilde{x}_m$,

$$v(s, X_i) - v(s, \tilde{x}_m) = (X_i - \tilde{x}_m)^T v'_x(s, \tilde{x}_m) + \frac{1}{2}(X_i - \tilde{x}_m)^T \left[\frac{\partial^2 v(s, \hat{x}_{im})}{\partial x \partial x^T}\right](X_i - \tilde{x}_m),$$

for some $\hat{x}_{im}$ in between $\tilde{x}_m$ and X_i.

Note B_{np} is the linear combination of $v(s, X_i) - v(s, \tilde{x}_m)$ as in (3.35); We plug in the two terms in the above displayed equation into (3.35) separately. First, the first-order terms sum up to exactly 0:

$$\sum_{i=1}^{n} w_{im}(X_i - \tilde{x}_m)^T v'_x(s, \tilde{x}_m)[1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \tilde{x}_m)] = 0,$$

due to standard local-linear calculation (*Fan and Gijbels*, 2018).

Second, note that $\boldsymbol{v}''_{xx}(\tau, x) = 0$ for all x due to the linearity of τ-th SQ, therefore the first item in Condition G-A2 implies $\|\boldsymbol{v}''_{xx}(s, x)\|_2 \leq L_2|s - \tau|$ uniformly for all x. Hence

$$\begin{aligned}
& \frac{1}{2nS_{0m}} \sum_{i=1}^{n} w_{im}(X_i - \tilde{x}_m)^T [\boldsymbol{v}''_{xx}(s, \hat{x}_{im})](X_i - \tilde{x}_m)[1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \tilde{x}_m)] \\
\leq & \left[\frac{L_2|s - \tau|}{2nS_{0m}} \sum_{i=1}^{n} w_{im}\|X_i - \tilde{x}_m\|^2\right] \left|1 - S_{1m}^T \boldsymbol{S}_{2m}^{-1}(X_i - \tilde{x}_m)\right| \\
\lesssim & |s - \tau| \left|\frac{\sum_{i=1}^{n} w_{im}\|X_i - \tilde{x}_m\|^2}{\sum_{i=1}^{n} w_{im}}\right| (1 + o_{\mathrm{P}}(1)) \\
\lesssim & |s - \tau|\bar{h}^2(1 + o_{\mathrm{P}}(1)),
\end{aligned}$$

where the first inequality owns to the operator norm bound for $\boldsymbol{v}''_{xx}$, and the $o_{\mathrm{P}}(1)$ terms are uniform in m and independent of s due to Lemma 8.

Combining the previous two displayed equations, we obtain

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |B_{np}(s,m)| = O_{\mathrm{P}}\left(r_n\bar{h}^2\right) = o_{\mathrm{P}}\left(n^{-1/2}\right),$$

since

$$r_n\bar{h}^2 = \sqrt{\frac{\bar{h}^4\log n}{nh^p}} \ll \frac{1}{\sqrt{n}},$$

under the first requirement of Condition G-A2.

Therefore, Claim 1 holds under either requirements of Condition G-A2.

Step 5: Verification of Claim 2 With Condition G-Q and Lemma 9, the proof here is a simple extension of step 4 in the proof of Proposition 4. We only give an outline here. Using same decomposition used in (3.34), we have

$$B_q(s,m) \;=\; \frac{1}{(1-s)}[U_{1n}(s,m)+U_{2n}(s,m)+U_{3n}(s,m)].$$

For U_{2n}, similar to step 4 in the proof of Proposition 4, we have:

$$\begin{aligned}\sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |U_{2n}(s,m)| \;&\leq\; \sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} \left\| \frac{\sum_{i=1}^{n} w_{im}\left[\frac{X_i-\bar{x}_m}{\bar{h}_m}\right]u_{2i}(s)}{nS_{0m}} \right\| \cdot \sup_{m=1,\ldots,M} \|\bar{h}_m S_{1m}^T \boldsymbol{S}_{2m}^{-1}\| \\ &\quad + \sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} \left| \frac{\sum_{i=1}^{n} w_{im}u_{2i}(s)}{nS_{0m}} \right| \\ &=\; O_{\mathrm{P}}(g_{2n}),\end{aligned}$$

which follows from Condition G-Q.

For $U_{1n}(s,m)$, it follows verbatim to part 4 of Proposition 4 that:

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |U_{1n}(\tau,m)| = O_{\mathrm{P}}(g_{1n}^2), \qquad \sup_{\substack{m=1,\ldots,M \\ |s-\tau|\leq B\cdot r_n}} |U_{3n}(\tau,m)| = O_{\mathrm{P}}(g_{1n}^2),$$

from the uniform convergence (over s) in Lemma 9.

Noting that $g_{1n}^2 + g_{2n} \ll n^{-1/2}$, we have verified Claim 2.

□

3.7.5 Proof of the technical lemmas

3.7.5.1 Proof of Lemmas 5, 6 and 7

In this section, we use the same notations as in the proof of Theorem III.1 in Section 3.7.2.

Proof of Lemma 5. First we prove the second claim. By the Lipschitz continuity of $g(\cdot)$, we have

$$\begin{aligned}
& \left\| \mathrm{E}\left[\sum_{m=1}^{M} \mathbf{1}\{X \in A_m\} g(\bar{x}_m) h(X)\right] - \mathrm{E}[g(X)h(X)] \right\| \\
\leq\ & \mathrm{E}\left[\sum_{m=1}^{M} \mathbf{1}\{X \in A_m\} \cdot \|g(\bar{x}_m) - g(X)\| \cdot |h(X)|\right] \\
\lesssim\ & \sup_{m=1,\ldots,M} \mathrm{diam}(A_m) \cdot \mathrm{E}\left[|h(X)|\right] \\
=\ & o_{\mathrm{P}}(1),
\end{aligned}$$

where the last equality follows from the binning conditions in Lemma 5 as well as the absolute integrability of h.

For the first claim, we first show the convergence when $\hat{\gamma}_m$ is replaced by $\hat{\pi}_m$, where $\hat{\pi}_m$ is given in Lemma 5. By re-arranging the summation we have

$$\sum_{m=1}^{M} \hat{\pi}_m g(\bar{x}_m) = \frac{1}{n} \sum_{m=1}^{M} \sum_{i=1}^{n} \mathbf{1}\{X_i \in A_m\} g(\bar{x}_m) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sum_{m=1}^{M} \mathbf{1}\{X_i \in A_m\} g(\bar{x}_m)}_{U_i^{(n)}},$$

where $U_i^{(n)}$ depends on the sample size through binning. For each fixed n, $U_i^{(n)}$ are *i.i.d.* across $i = 1, \ldots, n$, and $\mathrm{E}[|U_i^{(n)}|] < +\infty$ since $g(\cdot)$ is bounded. The Law of Large

Numbers gives

$$\sum_{m=1}^{M} \hat{\pi}_m g(\bar{x}_m) - \mathrm{E}[U_i^{(n)}] \xrightarrow{P^*} 0.$$

Furthermore, using the second claim of the lemma, we have $\mathrm{E}[U_i^{(n)}] \to \mathrm{E}[g(X)]$, thus

$$\sum_{m=1}^{M} \hat{\pi}_m g(\bar{x}_m) \xrightarrow{P^*} \mathrm{E}[g(X)]. \tag{3.38}$$

Next, we show the first claim of the lemma holds with $\hat{\gamma}_m$. Under the conditions on $\hat{\gamma}_m$ in Lemma 5, we have

$$\left| \sum_{m=1}^{M} \hat{\gamma}_m g(\bar{x}_m) - \sum_{m=1}^{M} \hat{\pi}_m g(\bar{x}_m) \right| \leq \sum_{m=1}^{M} \left| \frac{\hat{\gamma}_m - \hat{\pi}_m}{\hat{\pi}_m} \right| \hat{\pi}_m |g(\bar{x}_m)| = o_{\mathrm{P}}(1),$$

since $|g(\cdot)|$ is bounded. The proof is now complete by combining the above displayed equation with (3.38).

□

Proof of Lemma 6. We prove the three items separately.

For claim 1, note that $|v(s,x) - q(s,x)|$ is a continuous function in s and x under Condition G-Y2; therefore the upper bound holds since continuous functions are always bounded on compact intervals. We consider the lower bound. By the uniform continuity of $q(s,x)$ on compact intervals, there is a constant $c_1 > 0$ such that $|s-\tau| \leq 2c_1$ implies $|q(s,x) - q(\tau,x)| \leq \varepsilon_0$ for all x, where ε_0 is defined in Condition G-Y1. Without loss of generality we assume $c_1 < \varepsilon_0/2$. For each $|s-\tau| < c_1$, we have

$$\begin{aligned} v(s,x) - q(s,x) &= \frac{\int_s^1 q(\alpha,x) - q(s,x)\,\mathrm{d}\alpha}{1-s} \\ &\geq \frac{\inf\limits_{x,|\alpha-\tau|\leq 2c_1} \left[\dfrac{\partial q(\alpha,x)}{\partial s}\right] \cdot \displaystyle\int_{\tau+c_1}^{\tau+2c_1} |\alpha - s|\,\mathrm{d}\alpha}{1-s} \\ &\geq \frac{1}{\sup\limits_{x,|y-q(\tau,x)|\leq\varepsilon_0} f_{Y|X}(y;x)} \cdot \frac{c_1^2}{2(1-s)}. \end{aligned}$$

The lower bound in the first claim hence follows follows from Condition G-Y1.

For claim 2, the derivative of $q(s,x)$ is bounded because $f_{Y|X}(y;x)$ is bounded in Condition G-Y1. For the derivative of $v(s,x)$, note

$$\frac{\partial v(s,x)}{\partial s} = \frac{v(s,x) - q(s,x)}{1-s}; \tag{3.39}$$

the boundedness of the above derivative then follows from item 1 of Lemma 6.

Finally we prove claim 3. Since claim 2 of Lemma 6 implies both $v(s,x)$ and $q(s,x)$ are uniformly (in x) Lipschitz continuous in $s \in [\tau - c_1, \tau + c_1]$. Therefore, the derivative $\partial v(s,x)/\partial s$ is also uniformly (over both x and $s \in [\tau - c_1, \tau + c_1]$) Lipschitz continuous from (3.39). Furthermore, the Lipschitz continuity of $[\partial v(s,x)/\partial s]^{-1}$ follows since $\partial v(s,x)/\partial s$ is uniformly bounded away from 0 and $+\infty$. □

Proof of Lemma 7. We need to check items 1, 2 and 3 in the lemma separately. As a preliminary result, note that

$$\left| \sum_{m=1}^{M} (\hat{\gamma}_m - \hat{\pi}_m) \right| \leq \sum_{m=1}^{M} \hat{\pi}_m \left| \frac{\hat{\gamma}_m - \hat{\pi}_m}{\pi_m} \right| = o_{\mathrm{P}}(1),$$

under the conditions of the lemma. Therefore $\sum_{m=1}^{M} \hat{\gamma}_m = O_{\mathrm{P}}(1)$.

To check item 2, it follows that

$$\sqrt{n} \sum_{m=1}^{M} \hat{\gamma}_m \left[\hat{v}_m(\tau) - v_m(\tau)\right]^2 \leq O_{\mathrm{P}}(\sqrt{n}\, r_n^2),$$

which is a direct consequence of Condition G-V1.

Next we check item 3 in Lemma 7. From the monotonicity and (left-)continuity of $\hat{v}(s, \bar{x}_m)$ we have $\tau \leq \hat{h}_m[\hat{v}_m(\tau)] < \tau + g_n$, for any $g_n > 0$ satisfying $\hat{v}_m(\tau + g_n) > \hat{v}_m(\tau)$. Therefore it suffices to show that there exists a sequence $0 < g_n \ll n^{-1/2}$, such that

$$\inf_{m=1,\ldots,M} [\hat{v}_m(\tau + g_n) - \hat{v}_m(\tau)] > 0,$$

with high probability; the above displayed inequality means the functions $\hat{v}_m(\cdot)$ are not flat near τ. Note for any $0 < g_n \ll n^{-1/2}$, we shall have

$$\begin{aligned}\inf_{m=1,\ldots,M}[\hat{v}_m(\tau+g_n)-\hat{v}_m(\tau)] &\geq \inf_m[v_m(\tau+g_n)-v_m(\tau)] \\ &\quad - \underbrace{\sup_{\substack{m=1,\ldots,M \\ s:|s-\tau|\lesssim n^{-1/2}}} |[\hat{v}_m(s)-v_m(s)]-[\hat{v}_m(\tau)-v_m(\tau)]|}_{O_{\mathrm{P}}(G_n)} \\ &\geq g_n \cdot \left[\inf_{\substack{m=1,\ldots,M \\ |s-\tau|\leq g_n}} v_m'(s)\right] - O_{\mathrm{P}}(G_n),\end{aligned}$$

where $G_n \ll n^{-1/2}$ as in the second requirement of Condition G-V1. By Lemma 6, $v_m'(s)$ is uniformly bounded from below; therefore by choosing any g_n such that $G_n \ll g_n \ll n^{-1/2}$, the last displayed inequality is positive with probability tending to 1. Item 3 in Lemma 7 thus takes hold.

Finally we check item 1 in Lemma 7. Our proof follows the classical treatment in *Bahadur* (1966). We first show that $\hat{h}_m(z)$ converge uniformly at a rate of r_n, which is given in Lemma 7. For each s in a shrinking neighbourhood of τ, and for any fixed $C_1 > 0$, it follows from the definition of $\hat{h}$ in (3.10) that

$$\begin{aligned}\hat{h}_m[v_m(s)] < s - C_1 r_n &\;\Rightarrow\; v_m(s) \leq \hat{v}_m(s - C_1 r_n), \\ \hat{h}_m[v_m(s)] > s + C_1 r_n &\;\Rightarrow\; v_m(s) \geq \hat{v}_m(s + C_1 r_n),\end{aligned}$$

which shows that

$$\sup_{m=1,\ldots,M} |\hat{h}_m[v_m(s)] - h_m[v_m(s)]| > C_1 r_n$$

$$\Downarrow$$

$$\sup_{\substack{m=1,\ldots,M \\ |u|<C_1 r_n}} |\hat{v}_m(s+u) - v_m(s+u)| \geq C_1 r_n \inf_{|u|\leq C_1 r_n} v_m'(s+u).$$

Note $v'_m(\cdot)$ is uniformly bounded in Lemma 6, hence for sufficiently large C_1, the probability of the right hand side of the above displayed equation converges to zero by Condition G-V1. Adding uniformity with respect to s, we have that

$$\sup_{\substack{m=1,\ldots,M \\ |z-v_m(\tau)|\leq C_2\,(r_n+n^{-1/2})}} |\hat{h}_m(z) - h_m(z)| = O_{\mathrm{P}}(r_n) = o_{\mathrm{P}}(1), \tag{3.40}$$

for some $C_2 > 0$; we can use the range $|z - v_m(\tau)| \leq C_2\,(r_n + n^{-1/2})$ since h_m is uniformly (over m) Lipschitz continuous by Lemma 6.

Next we consider the asymptotic equi-continuity of $\hat{h}_m$. Let $z_m = v_m(\tau)$, and fix a z'_m such that $|z'_m - z_m| \leq C_2(n^{-1/2} + r_n)$. Define $\hat{\xi}_m = \hat{h}_m(z_m)$, $\hat{\xi}'_m = \hat{h}_m(z'_m)$. Fixing n, from the monotonicity and (left-)continuity of $\hat{v}_m$ we have:

$$\hat{v}_m(\hat{\xi}_m) \leq z_m \leq \hat{v}_m(\hat{\xi}_m + \varepsilon_n), \quad \hat{v}_m(\hat{\xi}'_m) \leq z'_m \leq \hat{v}_m(\hat{\xi}'_m + \varepsilon_n),$$

for any $\varepsilon_n > 0$; See *Van der Vaart* (2000, Chapter 19). Letting $\Delta_m(\cdot) = \hat{v}_m(\cdot) - v_m(\cdot)$, the first set of inequalities above on the left implies

$$\Delta_m(\hat{\xi}_m) \leq [z_m - v_m(\hat{\xi}_m)] \leq \Delta_m(\hat{\xi}_m + \varepsilon_n) + v_m(\hat{\xi}_m + \varepsilon_n) - v_m(\hat{\xi}_m).$$

Re-arranging the above displayed inequalities gives

$$\begin{aligned} \Delta_m(\hat{\xi}_m) - \Delta_m(\hat{\xi}'_m + \varepsilon_n) - \eta_k(z'_m) \leq [z_m - z'_m] - [v_m(\hat{\xi}_m) - v_m(\hat{\xi}'_m)] \\ \leq \Delta_m(\hat{\xi}_m + \varepsilon_n) - \Delta_m(\hat{\xi}'_m) + \eta_k(z'_m), \end{aligned} \tag{3.41}$$

where

$$\eta_k(z'_m) = \max\left\{|v_m(\hat{\xi}_m + \varepsilon_n) - v_m(\hat{\xi}_m)|, |v_m(\hat{\xi}'_m + \varepsilon_n) - v_m(\hat{\xi}'_m)|\right\}.$$

We derive the desired asymptotic equi-continuity of $\hat{h}_m$ from (3.41). To this end,

we bound its left and right hand sides separately. An application of the results in (3.40) shows that both $\hat{\xi}_m$ and $\hat{\xi}'_m$ converges in probability towards τ uniformly over m. It then follows from the Lipschitz continuity of v_m in Lemma 6 that

$$\sup_{\substack{m=1,\ldots,M \\ |z'_m - v_m(\tau)| \leq C_2(n^{-1/2}+r_n)}} |\eta_m(z'_m)| = O_{\mathrm{P}}(\varepsilon_n).$$

In addition, from (3.40) we have

$$\sup_{m=1,\ldots,M} |\hat{\xi}_m - \tau| = O_{\mathrm{P}}(r_n), \qquad \sup_{\substack{m=1,\ldots,M \\ |z'_m - v_m(\tau)| \leq C_2(n^{-1/2}+r_n)}} |\hat{\xi}'_m - \tau| = O_{\mathrm{P}}(r_n + n^{-1/2}),$$

Then, by choosing $\varepsilon_n = o(n^{-1/2} \wedge r_n)$ we have

$$\sup_{\substack{m=1,\ldots,M \\ |z'_m - v_m(\tau)| \leq C_2(n^{-1/2}+r_n)}} |\Delta_m(\hat{\xi}_m) - \Delta_m(\hat{\xi}'_m + \varepsilon_n)| = o_{\mathrm{P}}\left(n^{-1/2}\right),$$

from the second claim of Condition G-V1. The right hand side of Equation (3.41) can be bounded with the same argument, hence we have

$$\sup_{\substack{m=1,\ldots,M \\ |z'_m - v_m(\tau)| \leq C_1(n^{-1/2}+r_n)}} \left| [z_m - z'_m] - [v_m(\hat{\xi}_m) - v_m(\hat{\xi}'_m)] \right| = o_{\mathrm{P}}\left(n^{-1/2}\right), \tag{3.42}$$

by our choice of ε_n.

Finally, we connect Equation (3.42) with the desired asymptotic equi-continuity of $\hat{h}_m$. Let $\xi_m = h_m(z_m)$, $\xi'_m = h_m(z'_m)$, and therefore $z_m - z'_m = v_m(\xi_m) - v_m(\xi'_m)$.

By the first-order Taylor expansion of $v_m(\cdot)$ we have

$$\begin{aligned}\left|[z_m - z'_m] - [v_m(\hat{\xi}_m) - v_m(\hat{\xi}'_m)]\right| &= \left|v'_m(\tilde{s}_1)[\xi_m - \hat{\xi}_m] - v'_m(\tilde{s}_2)[\xi'_m - \hat{\xi}'_m]\right| \\ &\geq v'_m(\tilde{s}_1)\left|[\xi_m - \hat{\xi}_m] - [\xi'_m - \hat{\xi}'_m]\right| \\ &\quad - \underbrace{\sup_{m=1,\ldots,M}\left|[v'_m(\tilde{s}_1) - v'_m(\tilde{s}_2)] \cdot [\hat{\xi}'_m - \xi'_m]\right|}_{O_{\mathrm{P}}(H_n)} \\ &\geq c_1\left|[h_m(z_m) - \hat{h}_m(z_m)] - [h_m(z'_m) - \hat{h}_m(z'_m)]\right| \\ &\quad - H_n,\end{aligned}$$

for some $\tilde{s}_1$ in between ξ_m and $\hat{\xi}_m$ and some $\tilde{s}_2$ in between ξ'_m and $\hat{\xi}'_m$; the last inequality follows by expanding ξ_m and $\hat{x}_m$, and that v'_m is bounded in Lemma 6. Since both h_m and v'_m is Lipschitz continuous, it follows from (3.40) that we can take $H_n = (r_n + n^{-1/2})^2$. We conclude from the above displayed equation and (3.42) that

$$\begin{aligned}&\sup_{\substack{m=1,\ldots,M \\ z_m = v_m(\tau) \\ |z'_m - v_m(\tau)| \leq C_1 \cdot (r_n + n^{1/2})}}\left|[h_m(z_m) - \hat{h}_m(z_m)] - [h_m(z'_m) - \hat{h}_m(z'_m)]\right| \\ &= o_{\mathrm{P}}\left(n^{-1/2}\right) + O_{\mathrm{P}}\left((r_n + n^{-1/2})^2\right),\end{aligned}$$

which proves item 1 of Lemma 7. The proof is now complete.

□

3.7.5.2 Proof of Lemmas 8 and 9

In this section, we use the same notations as in the proof of Theorem III.2 in Section 3.7.3.

Proof of Lemma 8. We first prove the first claim; By definition

$$\left\|S_{0m}^{-1} S_{1m}\right\| \leq \frac{\sum_{i=1}^n \|X_i - \tilde{x}_m\| \mathbf{1}\{X_i \in A_m\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in A_m\}} \leq \bar{h}_m,$$

for all $m = 1, \ldots, M$. Therefore

$$\left|1 - \frac{\hat{\gamma}_m}{S_{0m}}\right| = \left|S_{0m}^{-1} S_{1m}^T \boldsymbol{S}_{2m}^{-1} S_{1m}\right| \leq \|\bar{h}_m \boldsymbol{S}_{2m}^{-1} S_{1m}\|,$$

and hence the first claim follows from the second claim. It suffices to show

$$\sup_m \|\bar{h}_m \boldsymbol{S}_{2m}^{-1} S_{1m}\| = o_{\mathrm{P}}(1).$$

For Claim 2, we first give a uniform probability order bound for $\|S_{1m}\|_2$, where

$$n \cdot S_{1m} = \sum_{i=1}^{n} (X_i - \tilde{x}_m) \mathbf{1}\{X_i \in A_m\} \in \mathbb{R}^p.$$

We apply the covering argument to show the convergence of the ℓ_2 norm. For any $\alpha \in \mathbb{R}^p$, with $\|\alpha\| = 1$, we have

$$\begin{aligned} \left|\mathrm{E}\left(\alpha^T S_{1m}\right)\right| &= \left|\alpha^T \int_{z \in A_m} (z - \tilde{x}_m) f_X(z) \mathrm{d}z\right| \\ &\leq f_X(\tilde{x}_m) \cdot \left|\alpha^T \int_{z \in A_m} (z - \tilde{x}_m) \mathrm{d}z\right| \\ &\quad + \alpha^T \int_{z \in A_m} |f_X(z) - f_X(\tilde{x}_m)| \cdot \|z - \tilde{x}_m\| \, \mathrm{d}z \\ &= 0 + O\left(\bar{h}_m^{p+2}\right), \end{aligned} \tag{3.43}$$

uniformly over m, where the last inequality owns to $\tilde{x}_m$ being the geometric center of A_m, as well as the Lipschitz continuity of f_X. Similarly, we have the following uniform bound for variance

$$\begin{aligned} \mathrm{var}\left(\alpha^T S_{1m}\right) &\leq \mathrm{E}[(\alpha^T (X_i - \tilde{x}_m) \mathbf{1}\{X_i \in A_m\})^2] \\ &= O\left(\bar{h}_m^{p+2}\right). \end{aligned}$$

Furthermore, note the boundedness of $\|X_i - \tilde{x}_m\| \leq \bar{h}_m$ when $X_i \in A_m$. Application

of the Bernstein's inequality (*Vershynin*, 2018, Theorem 2.8.4) gives for any $\varepsilon > 0$,

$$\begin{aligned}
& \Pr\left(n \cdot |\alpha^T S_{1m}| \geq n\bar{h}_m^{p+1}\varepsilon\right) \\
\leq\ & \Pr\left(|\mathrm{E}(\alpha^T S_{1m})| \geq \bar{h}_m^{p+1}\varepsilon\right) + \Pr\left(n \cdot |\alpha^T S_{1m} - \mathrm{E}(\alpha^T S_{1m})| \geq n\bar{h}_m^{p+1}\varepsilon\right) \\
\leq\ & 2\exp\left\{-\frac{n^2\bar{h}_m^{2p+2}\varepsilon^2/2}{nO(\bar{h}_m^{p+2}) + n\bar{h}_m^{p+2}\varepsilon/3}\right\} \\
=\ & 2\exp\left\{-C_1 n\bar{h}_m^p\varepsilon^2\right\},
\end{aligned}$$

for some constant $C_1 > 0$ whenever n is sufficiently large. With the standard covering argument, see e.g., *Vershynin* (2018, Chapter 4), we have

$$\|S_{1m}\| = \sup_{\alpha} \alpha^T S_{1m} \leq 2 \sup_{j=1,\ldots,J} \alpha_t^T S_{1m},$$

where $\{\alpha_j\}$ forms a 1/2-net in the unit p-dimensional sphere and the covering number $J \leq 2^p$. Using a union bound over m and t gives

$$\begin{aligned}
\Pr\left(\sup_{m=1,\ldots,M}\left\|\frac{n \cdot S_{1m}}{n\bar{h}_m^{p+1}}\right\| \geq 2\varepsilon\right) &= 2\sum_{m=1}^{M}\sum_{j=1}^{J}\exp\{-C_1 n\bar{h}_m^p\varepsilon^2\} \\
&\leq 2\exp\left\{\log M + \log J - C_1 n\varepsilon^2 \cdot \inf_k \bar{h}_m^p\right\} \\
&\lesssim \frac{1}{n^3},
\end{aligned}$$

for sufficiently large n under Condition G-A1, which implies

$$\sup_{m=1,\ldots,M}\left\|\frac{S_{1m}}{\bar{h}_m^{p+1}}\right\| = o_{\mathrm{P}}(1).$$

Next, we prove an analogous result for the operator norm of $\boldsymbol{S}_{2m}^{-1}$. Basic matrix algebra gives

$$\|\boldsymbol{S}_{2m}^{-1}\|_{op} = \left(\min_{\alpha\in\mathbb{R}^p} \alpha^T \boldsymbol{S}_{2m}\alpha\right)^{-1}, \tag{3.44}$$

and hence it suffices to bound the right hand side. For any fixed α in the p-dimensional

unit sphere, we have, using similar to the derivation in (3.43):

$$\begin{aligned}\mathrm{E}\left(n\cdot\alpha^T\boldsymbol{S}_{2m}\alpha\right) &\geq m_1\cdot n\underline{h}_m^{p+2},\\ \mathrm{var}\left(n\cdot\alpha^T\boldsymbol{S}_{2m}\alpha\right) &= O\left(n\bar{h}_m^{p+4}\right),\end{aligned}$$

for some constant C_2; the expectation is lower bounded since the A_m covers a ball with radius $\underline{h}_m$. With Bernstein's inequality, similar to that used for S_{1m}, we have for any $\varepsilon>0$:

$$\Pr\left(n\cdot|\alpha^T\boldsymbol{S}_{2m}\alpha-\mathrm{E}(\alpha^T\boldsymbol{S}_{2m}\alpha)|\geq n\bar{h}_m^{p+2}\varepsilon\right)\leq 2\exp\left\{-C_2n\bar{h}_m^p\varepsilon^2\right\},$$

and hence for any sufficiently large $M_2>0$,

$$\begin{aligned}&\Pr\left(\frac{n\bar{h}_m^{p+2}}{n\cdot\alpha^T\boldsymbol{S}_{2m}\alpha}\geq M_2\right)\\ =\;&\Pr\left(n\cdot\alpha^T\boldsymbol{S}_{2m}\alpha\leq\frac{1}{M_2}n\bar{h}_m^{p+2}\right)\\ \leq\;&\Pr\left(n\cdot|\alpha^T\boldsymbol{S}_{2m}\alpha-\mathrm{E}(\alpha^T\boldsymbol{S}_{2m}\alpha)|\geq n\cdot\mathrm{E}[\alpha^T\boldsymbol{S}_{2m}\alpha]-\frac{1}{M_2}n\bar{h}_m^{p+2}\right)\\ \leq\;&\Pr\left(n\cdot|\alpha^T\boldsymbol{S}_{2m}\alpha-\mathrm{E}(\alpha^T\boldsymbol{S}_{2m}\alpha)|\geq m_1n\bar{h}_m^{p+2}/2\right)\\ \leq\;&2\exp\{-C_2n\bar{h}_m^pm_1^2/4\},\end{aligned}$$

since $\bar{h}_m/\underline{h}_m$ is uniformly bounded and the expectation is bounded from below.

Next, applying the same covering argument again, and note the relationship from (3.44), we have that

$$\Pr\left(\sup_{m=1,\ldots,M}\left\|\bar{h}_m^{p+2}\boldsymbol{S}_{2m}^{-1}\right\|_{op}\geq M_2\right)\;\lesssim\;\exp\{\log M+p\log 2-C_2n\underline{h}^pm_1^2/4\}\lesssim\frac{1}{n^3},$$

implying

$$\sup_{m=1,\ldots,M}\left\|\bar{h}_m^{p+2}\boldsymbol{S}_{2m}^{-1}\right\|_{op}=O_{\mathrm{P}}(1).$$

Therefore Claim 2 follows by combining the norm bounds for S_{1m} and $\boldsymbol{S}_{2m}^{-1}$. In particular,

$$\begin{aligned} & \Pr\left(\sup_m \|\bar{h}_m \boldsymbol{S}_{2m}^{-1} S_{1m}\| \geq \frac{1}{2}\right) \\ \leq\ & \Pr\left(\sup_m \left\|(\bar{h}_m^{p+1})^{-1} S_{1m}\right\| \geq \frac{1}{2M_2}\right) + \Pr\left(\sup_m \left\|\bar{h}_m^{p+2} \boldsymbol{S}_{2m}^{-1}\right\|_{op} \geq M_2\right) \\ \lesssim\ & \frac{1}{n^3}. \end{aligned}$$

For Claim 3, we give a bound for

$$S_{0m} = n^{-1} \sum_{i=1}^{n} \mathbf{1}[X_i \in A_m],$$

similar to what we did in Claim 2. Note that

$$\mathrm{E}[S_{0m}] = \Pr(X \in A_m) \gtrsim \underline{h}_m^p, \quad \mathrm{var}[S_{0m}] \leq \Pr(X \in A_m) \lesssim \bar{h}_m^p,$$

since the density of X is bounded. Therefore for small enough $\varepsilon_0 > 0$, an application of Bernstein's inequality gives

$$\begin{aligned} \Pr\left(S_{0m} \leq \varepsilon_0 \underline{h}_m^p\right) &\leq \Pr\left(S_{0m} - \mathrm{E}[S_{0m}] \leq -\varepsilon_0 \underline{h}_m^p / 2\right) \\ &\leq \exp\left\{-C_3 n \bar{h}_m^p \varepsilon_0^2\right\}, \end{aligned}$$

for some constant $C_3 > 0$. Taking a union bound with all $m = 1, \ldots, M$ shows

$$\Pr\left(\inf_m \frac{S_{0m}}{\bar{h}_m^p} \leq \varepsilon_0\right) \leq \sum_{m=1}^{M} \exp\left\{-C_3 n \bar{h}_m^p \varepsilon_0^2\right\} \leq \frac{1}{n^3},$$

for sufficiently large n with the bandwidth in Condition G-A1. The proof is now complete.

□

Proof of Lemma 9. We only give detailed proof for item 1; the conclusion for item 2 holds similarly and we only give an outline. We prove the conclusion specifically for $a_n = r_n$ in Proposition 4 and $b_n = g_{1n}$ as in Condition G-Q. We use the same notations in (3.26) and define

$$u_{1i}(s) = [Y_i - q(s, X_i)] \left[\mathbf{1}\{Y_i \geq \hat{q}(s, X_i)] - \mathbf{1}\{Y_i \geq q(s, X_i)\}\right],$$
$$u_{3i}(s) = (q(s, X_i) - \hat{q}(s, X_i)) \cdot \left[\mathbf{1}\{Y_i \geq \hat{q}(s, X_i)\} - \mathbf{1}\{Y_i \geq q(s, X_i)\}\right];$$

Correspondingly the left hand sides in Lemma 9 can be written as

$$U_{1n}(s, m) = \frac{\sum_{i=1}^n w_{im}\kappa_{im}u_{1i}(s)}{\sum_{i=1}^n w_{im}}, \quad U_{3n}(s, m) = \frac{\sum_{i=1}^n w_{im}\kappa_{im}u_{3i}(s)}{\sum_{i=1}^n w_{im}}.$$

Moreover, let

$$R_q = \sup_{\substack{m=1,\ldots,M \\ |s-\tau| \leq B \cdot r_n}} |\hat{q}(s, X_i) - q(s, X_i)| = O_{\mathrm{P}}(g_{1n}),$$

as in Condition G-Q.

We consider the decomposition:

$$\begin{aligned}
|U_{1n}(s, m)| \lesssim_{\mathrm{P}} & (nS_{0m})^{-1} \sum_{i=1}^n w_{im}[y_i - q(s, X_i)]\mathbf{1}\{q(s, X_i) \leq y_i \leq \hat{q}(s, X_i)\} \\
& + (nS_{0m})^{-1} \sum_{i=1}^n w_{im}[q(s, X_i) - y_i]\mathbf{1}\{\hat{q}(s, X_i) \leq y_i \leq q(s, X_i)\} \\
\leq & (nS_{0m})^{-1} \sum_{i=1}^n w_{im}[y_i - q(s, X_i)]\mathbf{1}\{q(s, X_i) \leq y_i \leq q(s, X_i) + R_q\} \\
& + (nS_{0m})^{-1} \sum_{i=1}^n w_{im}[q(s, X_i) - y_i]\mathbf{1}\{q(s, X_i) - R_q \leq y_i \leq q(s, X_i)\} \\
\triangleq & U_{1n}^{(+)}(s, m) + U_{1n}^{(-)}(s, m),
\end{aligned}$$

where we use a constant to upper bound $|\kappa_{im}|$ (see Claim 2 in the proof of Proposition 4); and the second inequality holds by monotonicity of the indicator functions. By

symmetry, hereafter we focus on the term $U_{1n}^{(+)}(s,m)$.

Let $s_- = \tau - Br_n$, $s_+ = \tau + Br_n$. For any $s \in [s_-, s_+]$ we have

$$\begin{aligned} 0 \le U_{1n}^{(+)}(s,m) &\le (nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}[y_i - q(s_-, X_i)]\mathbf{1}\{q(s_-, X_i) \le y_i \le q(s_+, X_i) + R_q\} \\ &\triangleq \bar{U}_{1n}^{(+)}(k), \end{aligned}$$

Therefore we can drop the supremum over s by relying on $\bar{U}_{1n}^{(+)}(k)$, we bound its expectation and centered process separately. In what follows we bound the centered empirical process and the expectation of $\bar{U}_{1n}^{(+)}(k)$ separately.

We give a tail bound for $\mathrm{E}[\bar{U}_{1n}^{(+)}(k)] - \bar{U}_{1n}^{(+)}(k)$ using Hoeffding's inequality (conditional on X) and a union bound. By Condition G-Y1, each summand in $\bar{U}_{1n}^{(+)}(k)$ is bounded by $[y_i - q(s_-, x)]\mathbf{1}\{q(s_-, x) \le y_i \le q(s_+, x) + R_q\} \lesssim (s_+ - s_-) + R_q$ for all x. For any $\delta_1 > 0$, there exists a large enough $M_1 > 0$ that

$$\begin{aligned} &\Pr\left(\sup_{m=1,\ldots,M} \left| \bar{U}_{1n}^{(+)}(k) - \mathrm{E}[\bar{U}_{1n}^{(+)}(k)] \right| \ge M_1(g_{1n} + r_n) \middle| X \right) \\ \le\ & 2\exp\left\{ \log n - 2nM_1^2 \cdot \inf_m S_{0m} \right\} + \delta_1, \end{aligned}$$

where the δ_1 comes from the probability that $R_q \ge M_1 g_{1n}$. Similar to how we obtain (3.33), the following unconditional tail bound holds from the above displayed conditional bound:

$$\Pr\left(\sup_{m=1,\ldots,M} \left| \bar{U}_{1n}^{(+)}(k) - \mathrm{E}[\bar{U}_{1n}^{(+)}(k)] \right| \ge M_1(g_{1n} + r_n) \right) \lesssim 2\delta_1.$$

Here we bound the expectation $\mathrm{E}[\bar{U}_{1n}^{(+)}(k)]$. By Condition G-Y1' we have

$$\begin{aligned} & \mathrm{E}[Y_i - q(s_-, X_i)]\mathbf{1}\{q(s_-, X_i) \leq Y_i \leq q(s_+, X_i) + R_q\} \\ = \; & \mathrm{E}[Y_i - q(s_-, X_i)]\mathbf{1}\left\{0 \leq [Y_i - q(s_-, X_i)] \leq R_q + \underline{f}^{-1}|s_+ - s_-|\right\} \\ = \; & O\left((g_{1n} + r_n)^2\right), \end{aligned}$$

since $s_+ - s_- \lesssim r_n$ and $R_q = O_{\mathrm{P}}(g_{1n})$. Therefore

$$\sup_{m=1,\ldots,M} \mathrm{E}[\bar{U}_{1n}^{(+)}(k)] = O_{\mathrm{P}}\left((g_{1n} + r_n)^2\right).$$

Combining the bounds for the expectation and the centered empirical process, we arrive at

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau| \leq B \cdot r_n}} U_{1n}^{(+)}(s, m) \lesssim_{\mathrm{P}} \sup_{m=1,\ldots,M} \bar{U}_{1n}^{(+)}(k) = O_{\mathrm{P}}\left((g_{1n} + r_n)^2\right).$$

Repeating the same procedure for $U_{1n}^{(-)}(s, m)$ would complete the proof for the first item of the Lemma.

For $U_{3n}(s, m)$, note

$$|U_{3n}(s, m)| \lesssim_{\mathrm{P}} R_q(nS_{0m})^{-1} \sum_{i=1}^{n} w_{im}\mathbf{1}[q(s, X_i) - R_q \leq y_i \leq q(s, X_i) + R_q];$$

therefore we can follow the same line of reasoning and establish

$$\sup_{\substack{m=1,\ldots,M \\ |s-\tau| \leq B \cdot r_n}} |U_{3n}(s, m)| = O_{\mathrm{P}}\left((g_{1n} + r_n)^2\right).$$

The proof is now complete.

□

3.7.6 Auxiliary discussions

3.7.6.1 On the monotonicity of the initial SQ estimator

In Condition G-V1, we require the initial SQ estimator $\hat{v}(s, \bar{x}_m)$ to be monotonically increasing in $s \in (0, 1)$. In our discussions below Condition G-V1, we demonstrate that monotonicity can be achieved by re-arrangement of a given estimator (*Chernozhukov et al.*, 2009, 2010). Here we provide a technical perspective that suggests monotonicity may not be necessary for Theorem III.1.

Even when the initial SQ estimators are not monotone, the m-Rock approach still has a clear interpretation as finding the τth 're-arranged' SQ (*Chernozhukov et al.*, 2009). To see this, consider the univariate case with no covariate as an illustrative example. Following (2.2), the m-Rock approach solves:

$$\min_C \int_0^1 \rho_\tau \left(\hat{v}(s) - C\right) \mathrm{d}s \approx \frac{1}{J} \min_C \sum_{j=1}^{J} \rho_\tau \left(\hat{v}(s_j) - C\right), \tag{3.45}$$

where we discretize the integral above as a grid $s_1, \ldots, s_J \in (0, 1)$. The solution to (3.45) is approximately the τ-th quantile of the (possible unordered) set $\{\hat{v}(s_1), \ldots, \hat{v}(s_J)\}$. Operationally, the m-Rock approach gives exactly the τ-th *monotonically re-arranged* superquantile in *Chernozhukov et al.* (2009).

Following this insight, we now demonstrate that the proof of Theorem III.1 may adapt to situations where $\hat{v}(s, \bar{x}_m)$ is not monotonic. As we've demonstrated in the proof of Theorem III.1, central to the main result is the asymptotic properties of $\hat{h}(\cdot, \tilde{x}_m)$; Without monotonicity, $\hat{h}(\cdot, \tilde{x}_m)$ is defined by

$$\hat{h}(z, x) := \int_0^1 \mathbf{1}\{\hat{v}(s, x) \leq z\} \mathrm{d}s = \sup\{s \in [0, 1] : \hat{v}(s, x) \leq z\}.$$

The functional $\hat{h}(\cdot, \tilde{x}_m)$ is the *monotonized inverse* operator in *Chernozhukov et al.* (2010); note when $\hat{v}(\cdot, \tilde{x}_m)$ is indeed monotonic, $\hat{h}(\cdot, \tilde{x}_m)$ reduces to the classic inverse

operator defined in (3.10). Corollary 3 of *Chernozhukov et al.* (2010) establishes the Hadamard differentiability of $\hat{h}(\cdot, \tilde{x}_m)$, and shows that its asymptotic property does not rely on the (finite-sample) monotonicity of $\hat{v}(\cdot, \tilde{x}_m)$. With some technical modification, we expect that their proof can be adapted to our setting, therefore our main result, i.e., Lemma 7 and Theorem III.1, can be established without the monotonicity requirement in Condition G-V1.

Remark 7. As yet another technical solution, one may pursue the following strategy: first find J equally-spaced grid points $0 < \tau_1 < \ldots < \tau_J < 1$ that spans the interval $[0, 1]$. Then we can estimate the initial SQ on the grid with linear interpolations in between the grid points. The monotonicity follows with probability going to 1 provided that $r_n \ll (\tau_{j+1} - \tau_j) \ll n^{-1/4}$, where r_n is in Condition G-V1.

3.7.6.2 On the bias in the initial SQ estimator

Here we illustrate the importance to control the bias in the initial SQ estimator. Consider the following example with fixed design in a unit cube $[0, 1]^p$ (excluding intercept); and we define the bins as hypercubes with edge length $\bar{h}$ and therefore we have $M \leq \lceil \bar{h}^{-p} \rceil$ total bins. Suppose we use a standard Nadaraya-Watson type estimator for the initial SQ in each bin m, and the τ-th SQ estimator can be represented as $\hat{v}_m(\tau) - v_m(\tau) = B_m + U_m$, where $\mathrm{E}[U_m] = 0$ and B_m is the bias.

We consider the plausibility of Condition G-V2. From the results in (*Kato*, 2012), $B_m = O_{\mathrm{P}}(\bar{h}^2)$ and $U_m = O_{\mathrm{P}}((n\bar{h}^p)^{-1/2})$. Because each $\hat{v}_m$ are based on local observations in disjoint bins, (B_m, U_m) is independent across $m = 1, \ldots, M$. Therefore, the aggregation for U_m gives

$$\sqrt{n} \sum_{m=1}^{M} U_m = \sqrt{nM} \cdot O_{\mathrm{P}} \left(\frac{1}{\sqrt{nh^p}} \right) = O_{\mathrm{P}}(1),$$

by the Central Limit Theorem, provided that $\sqrt{n}\bar{h}^p \to 0$. On the other hand, for the

bias terms we have

$$\sqrt{n}\sum_{m=1}^{M} B_m = \sqrt{n}M \cdot O_{\mathrm{P}}(\bar{h}^2) = O_{\mathrm{P}}\left(\sqrt{n}h^{2-p}\right);$$

the order of the above sum goes to infinity if $p > 1$, hence the bias dominates in the aggregation of $\hat{v}_m$ in Condition G-V2. Therefore, it is critical to reduce the bias when constructing the initial SQ estimator.

CHAPTER IV

Numerical and Empirical Investigations

In this chapter, we demonstrate the practical applicability of the m-Rock approach via numerical and empirical examples. Inspired by our theoretical framework in Chapter 3, we first discuss the computational aspects of the m-Rock approach and give a prototype implementation of the approach. Then we demonstrate the numerical stability and statistical efficiency of our approach via simulation studies. We also apply the m-Rock approach to two empirical examples related to finance and public health.

4.1 Implementation

Our theoretical analysis in Chapter 3 relies on finding disjoint bins that partition the sample space. However, it is often more practical to consider overlapping and possibly data-dependent bins. Specifically, we use subsampling and k-Nearest Neighbours (kNN) to find those overlapping bins. Suppose the observed covariates are $X_1, \ldots, X_n$. For each X_i, we find its K nearest neighbours in the data as $B_i = \{X_{n_{ij}} : j = 1, \ldots, K\}$. We then choose a subsample of size m out of those n sets as $B_{(1)}, \ldots, B_{(m)}$; those sets form m effective bins that discretize the sample space and may overlap with each other. The purpose of subsampling is to reduce the computational cost and we set $m = \lceil 20n/K \rceil$ in our experiments. On the other hand,

the value of K can be regarded as a tuning parameter of the m-Rock approach.

Based on those selected bins, the m-Rock approach can then be implemented from the discussions in Chapter 3. We give the detailed implementation in Algorithm 1. We review some notations here. Let $X \in \mathbb{R}^{p+1}$ be the covariate vector including an intercept term. The number K is the tuning parameter we described above, and δ ($\delta \in [0, 1)$), J and m are three user-specified parameters that we explain below.

Algorithm 1 Estimation of the τth SQ regression via the m-Rock approach.

1: Form an equally-spaced grid over the interval $[\tau - \delta\tau,\, \tau + \delta(1-\tau)]$ as

$$\tau - \delta\tau = s_0 < s_1 < \ldots < s_J = \tau + \delta(1-\tau).$$

2: Subsample m out of n covariate vectors $X_{(1)}, \ldots, X_{(m)}$.
3: **for** $i = 1$ to m **do**
4: Find the K nearest neighbours of $X_{(i)}$ in the full data, collected in bin B_i.
5: **for** $j = 0$ to J **do**
6: Obtain the conditional quantile estimator at level s_j:

$$\hat{q}^{(i)}(s_j, x), \quad x \in B_i.$$

7: Obtain the initial SQ estimator at level s_j from (3.6) in Chapter 3:

$$\hat{v}_{ij} \leftarrow \hat{v}\left(s_j, X_{(i)}\right),$$

8: **end for**
9: **end for**
10: Solve the (approximate) optimization problem via quantile regression

$$\begin{aligned}\widehat{\theta} \;&\leftarrow\; \min_{\theta \in \mathbb{R}^{p+1}} \sum_{i=1}^{m} \hat{\gamma}_m \int_{\tau-\delta\tau}^{\tau+\delta(1-\tau)} \rho_\tau\left(\hat{v}\left(s, X_{(i)}\right) - X_{(i)}^T\theta\right) \mathrm{d}s. \\ &\approx\; \min_{\theta \in \mathbb{R}^{p+1}} \frac{1}{1+J} \sum_{i=1}^{m} \sum_{j=0}^{J} \hat{\gamma}_m \rho_\tau\left(\hat{v}_{ij} - X_{(i)}^T\theta\right),\end{aligned}$$

where $\hat{\gamma}_m$ is given in (3.7) in Chapter 3.

We explain several aspects of Algorithm 1. First, the choice of J in Step 1 may have to increase with more sample size, so that the approximation error is negligible.

Second, in Step 10 we use a δ-truncated interval in the m-Rock loss function, in accordance with Corollary 1; Moreover, we may use a left-winsorized estimator (for the lower quantile levels) for $\hat{v}(s, x)$ in Step 7; see also Remark 4 in Chapter 3. The truncation and winsorization can reduce the computational cost. Third, in Step 2 we can use the k-medoids algorithm (*Schubert and Rousseeuw*, 2019) for subsampling, so that the selected bins are more representative. In our experience, choosing $m = \lceil 20n/K \rceil$, $J = \lceil \sqrt{n \log(n)} \rceil$, $\delta = 0.8$, and using 50% left-winsorization gives relatively stable performances; therefore we fix those parameters in our subsequent experiments.

Another important ingredient to Algorithm 1 is the quantile estimator in Step 6. While there are many possibilities, in our experiments we restrict to two types of estimators: local or global estimation. For local estimation, we use the data within each bin to fit a bin-wise linear quantile regression; and we name this approach m-Rock with *kNN quantile*. For global estimation, we use all available data to fit either (i) linear quantile regression, or (ii) B-splines quantile regression.

Remark 8. Instead of specifying the size (volume) for each bin, in Algorithm 1 we specify the number (K) of observations within each bin. This is in parallel to using a variable bandwidth for kernel-based estimation (*Muller and Stadtmuller*, 1987; *Fan and Gijbels*, 1992). By ensuring each bin has sufficient data, our implementation based on kNN helps with the numerical stability in the initial SQ estimation.

Remark 9. In the scenario with only one covariate, we have conducted extensive numerical experiments to compare our Algorithm 1 with the implementation using non-overlapping bins. With only one covariate, it is relatively simple to find non-overlapping bins based on the sample quantiles, which are consistent with our theoretical framework in Chapter 3. We find that our Algorithm 1 is more stable with respect to the bin sizes.

4.2 Numerical experiments

4.2.1 The effect of tuning K

In this section, we use Monte Carlo simulations to investigate the effect of K in the m-Rock approach. Effectively, K is the bandwidth parameter for the initial non-parametric SQ estimation in Chapter 3; and the choice of K reflects the bias-variance trade-off in the m-Rock approach. Here we use two models to study the practical effect of selecting K when the m-Rock approach is implemented as in Algorithm 1. For each simulation setting, we generate $1,000$ Monte Carlo datasets and report the average estimation accuracy.

Theoretically, we can derive some theoretical requirements for K. For each bin, $K \asymp n\bar{h}^p$ under the notations in Section 3.4, where $\bar{h}$ is the radius of the bin and p is the dimension of covariates. Condition G-A1 in Chapter 3 then implies that K needs to satisfy:

$$\sqrt{n}\log n \ll K \ll n.$$

In the subsequent simulations, we shall examine the performance of the m-Rock approach when K varies within the above range.

4.2.1.1 A one-dimensional model

We first consider a heteroscedastic model with one continuous covariate $X \sim \mathcal{U}(0,4)$:

$$Y = -1 + 2X + (2(X-2)^2 + 1)(\varepsilon - v_0), \tag{4.1}$$

where ε is independent of X and follows the standardized skewed-t_5 distribution with skewness set to 2, see *Hansen* (1994); v_0 is chosen to be the 90% SQ of ε. Figure 4.1 shows the density function of ε, as well as a scatterplot of data generated from Model (4.1) with sample size $n = 2000$. Under Model (4.1), the 90% conditional SQ

is linear in X, yet the 90% conditional quantile is highly non-linear. Therefore many other approaches in the literature (e.g., those we compare with in Section 3.5) are not directly applicable.

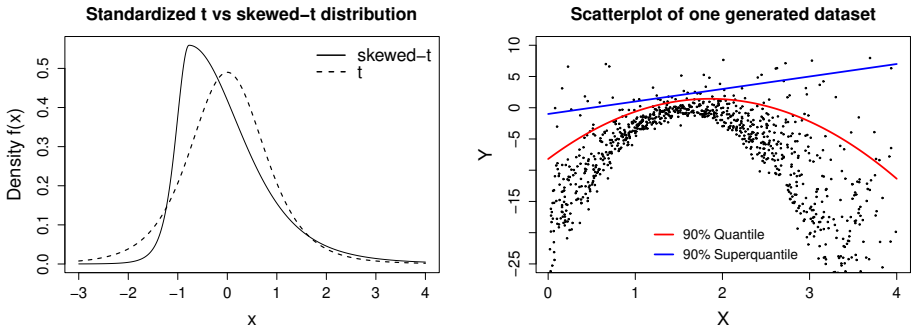
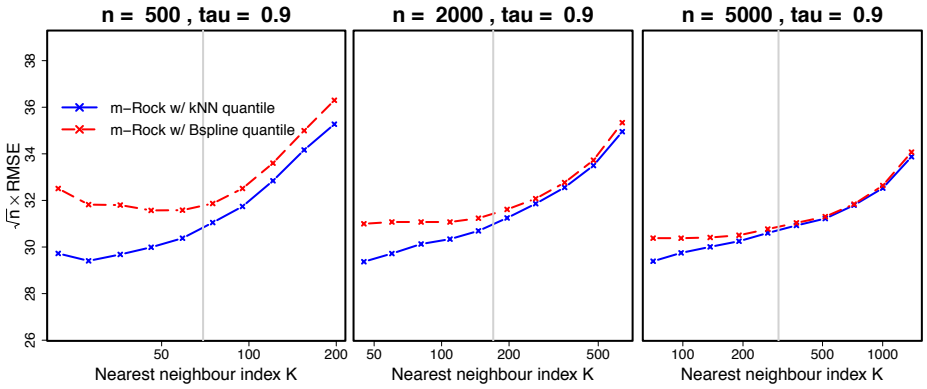


Figure 4.1: Illustration of the skewed-t error distribution. Left: The density functions of the standardized t_5-distribution and skewed t_5-distribution with the skewness parameter equals to 2. Right: The scatterplot of one dataset generated from Model (4.1).

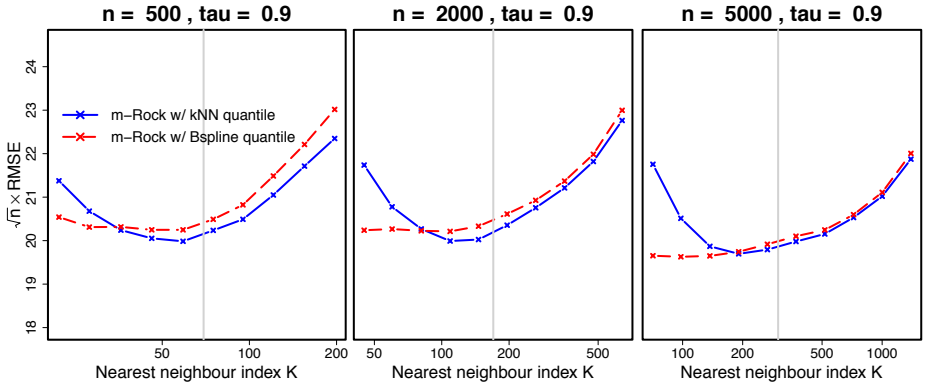
We consider three different sample sizes $n = 500$, $n = 2000$ and $n = 5000$ under Model (4.1), and we focus on $\tau = 0.9$. For initial estimation of the conditional quantile function in Step 6, we consider two options: (i) kNN quantile regression; and (ii) B-splines quantile regression with 5 degrees of freedom. In this setting, we find that the B-splines quantile regression provides an accurate approximation for the conditional quantile function, and it does not depend on the tuning of K. We consider the performance of these two m-Rock implementations when the tuning parameter K varies within $[n^{0.5}, n^{0.85}]$.

Figure 4.2 shows the scaled RMSE (Root Mean Squared Error) of the m-Rock approach for a wide range of K. Note the x-axis is on a log-scale; and the RMSE is multiplied by $\sqrt{n}$ so that the results under different sample sizes are comparable. We observe that both versions of the m-Rock approach can achieve a desirable bias-variance trade-off for a wide range of K near $\sqrt{n}\log n/2$, marked with a vertical line. When using a proper K, the two m-Rock implementations are similar.

We explain the effect of K in more details. First, choosing too large a K will



(a) For the intercept term β_0



(b) For the slope term β_1

Figure 4.2: The scaled (by $\sqrt{n}$) RMSE for each coefficients under Model (4.1). The x-axis is displayed on the log scale, and the vertical line marks our recommended value of $K = \sqrt{n}\log n/2$.

lead to a worse performance. This phenomenon is persistent in both implementations of the m-Rock approach and is due to the binning bias: the local structure of the covariate space cannot be well-approximated by the bins if they are too large in size. Second, using a smaller K in Figure 4.2 also drives up the RMSE for the approach with kNN quantile; This is particularly notable for the estimation of the slope term β_1. With kNN quantile, the tuning of K affects both the m-Rock approach itself and the initial quantile estimation. The results in Figure 4.2 suggest that the kNN quantile regression can be unstable with a small K, which impairs the performance of the m-Rock approach. On the other hand, the m-Rock approach with B-splines quantile (which does not depend on K) stays relatively stable with smaller values of

K.

To conclude, we find that when K is in a suitable range, both implementations of the m-Rock approach can achieve relatively stable performance. In our subsequent experiments, we shall fix $K = \sqrt{n}\log(n)/2$, which is marked with a vertical line in Figure 4.2. Furthermore, if we have a good conditional quantile estimator that does not depend on K, the performance of the m-Rock approach can be even more stable, and smaller values of K may be preferred.

4.2.1.2 A three-dimensional model

Here we consider a model with three continuous covariates to examine the effect of dimensionality in the selection of K. Specifically, we consider the following model:

$$Y = -3 + 2X_1 - 3X_2 + 2X_3 + (5X_1 + 2)\varepsilon \tag{4.2}$$

where (X_1, X_2, X_3) is uniformly distributed on the 3-dimensional cube $[0, 4]^3$; the error term ε follows a standard normal distribution independent of the covariates. While heterogeneity is present in Model (4.2), both the conditional quantile and SQ of Y are linear in covariates at all quantile levels.

We generate Monte Carlo datasets under Model (4.2) with four difference sample sizes $n = 500$; $2,000$; $5,000$ and $10,000$; for each generated data, we estimate the SQ regression at three quantile levels $\tau = 0.8$, 0.9 and 0.95. For the m-Rock approach, we consider two implementations based on (i) kNN quantile; or (ii) linear quantile We also include the Two-Step approach in (2.8) of Chapter 2 as a relatively simple benchmark: we first fit the τ-th linear quantile regression, followed by least-squares regression using the data above the fitted τ-th quantile; We name this approach TS-LS. The TSLS approach is similar to the m-Rock implementation (ii) as they both rely on the linearity of the conditional quantile function. On the other hand, the m-

Rock implementation (i) does not rely on a linear quantile regression model *a priori*, even though it takes hold under Model (4.2).

We first examine the average RMSE over the four estimated SQ regression coefficients under Model (4.2). Figure 4.3 gives a matrix of plots that shows the rescaled (by $\sqrt{n}$) RMSE at different (n, τ) combinations; In each plot, we vary K within the range $[n^{0.55}, n^{0.85}]$ for the m-Rock approach. We observe that the m-Rock approaches are consistently more efficient than the benchmark approach with almost all choices of K; and the performances are relatively steady when K is close to our recommended value $\sqrt{n}\log n/2$, marked by a vertical line in Figure 4.3.

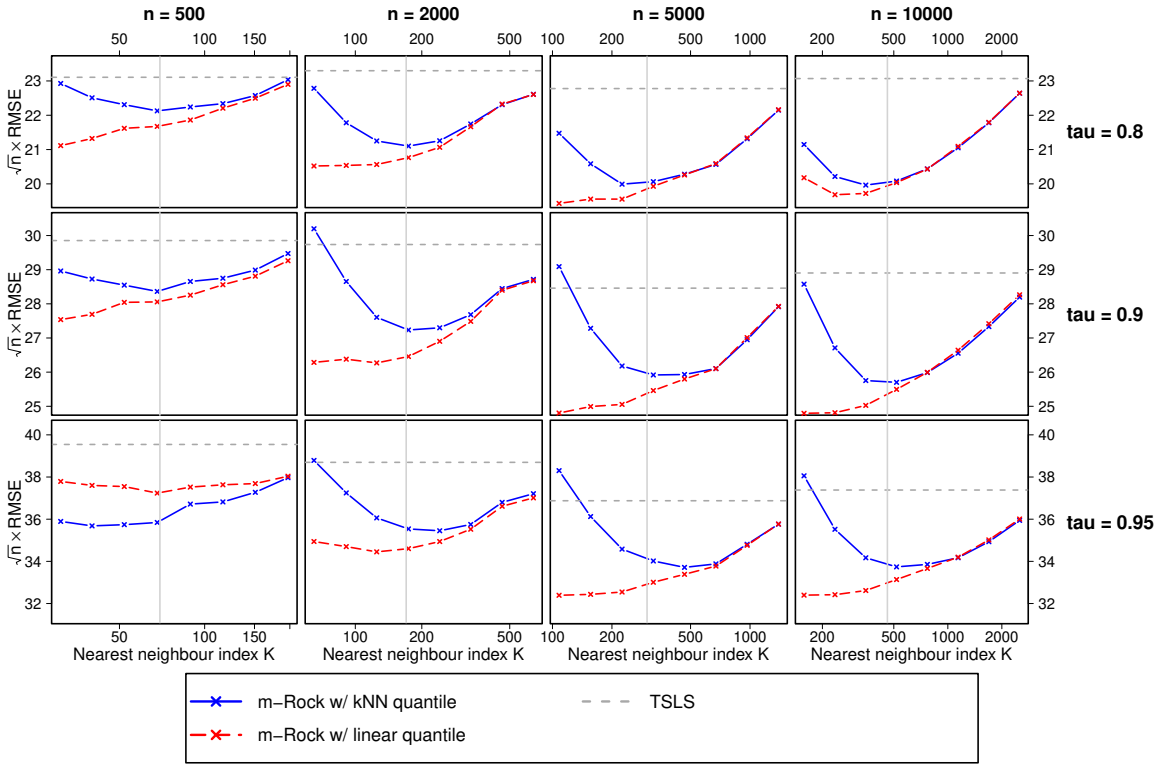
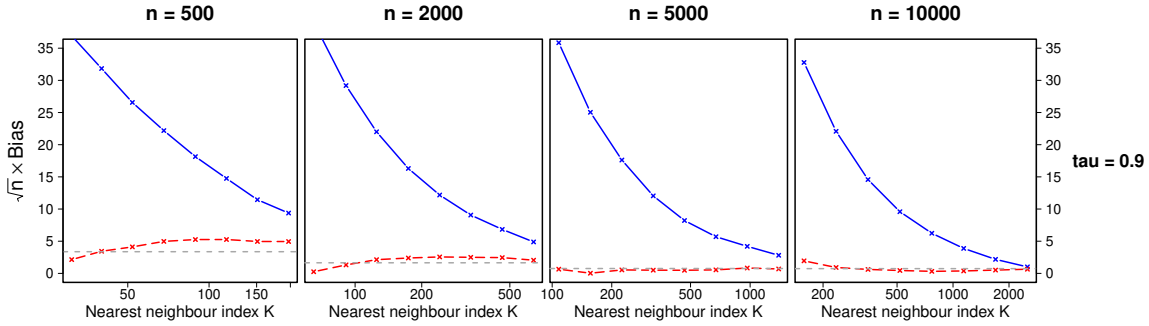


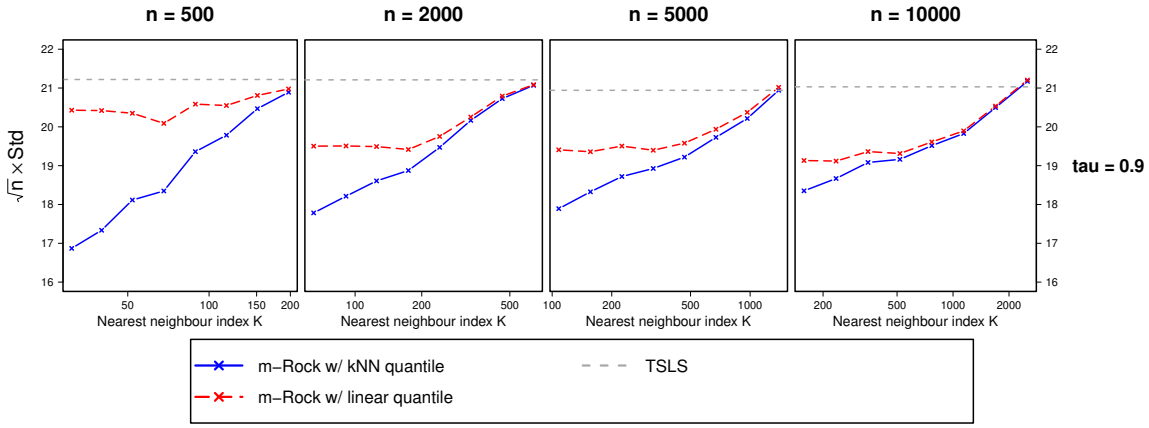
Figure 4.3: The scaled (by $\sqrt{n}$) average RMSE across four coefficients under Model (4.2); each row represents a fixed quantile level, and each column shares a fixed sample size. The x-axis is displayed on the log scale, and the vertical line marks our recommended value of $K = \sqrt{n}\log n/2$.

With multiple continuous covariates, it becomes even more challenging for the m-Rock with kNN quantile regression. This becomes more evident in the bias-variance

decomposition in Figure 4.4, where we focus on one of the coefficient β_1 and one quantile level $\tau = 0.9$. For the m-Rock Implementation with kNN quantile, there is a significant bias for most values of K; the magnitude of the bias is comparable to that of the standard deviation. From Figure 4.4, using a smaller K quickly increases the bias, while using a larger K leads to inflated variance. Therefore, it may be difficult for the m-Rock Implementation (i) to achieve desired bias-variance balance by tuning K. On the other hand, the m-Rock approach with linear quantile is much more stable when K is small, though it is less robust because it hinges on a linear quantile regression model.



(a) Scaled absolute bias



(b) Scaled standard deviation

Figure 4.4: The scaled (by $\sqrt{n}$) absolute bias and standard deviation for β_1 with $\tau = 0.9$ under Model (4.2); other attributes of the figure are the same as Figure 4.3.

In summary, the m-Rock approach is still applicable with multiple covariates; and the choice $K = \sqrt{n}\log n/2$ is still suitable under Model (4.2). Nonetheless, the kNN

quantile regression can be less stable in this setting with respect to the tuning of K. it is desirable to have a joint quantile and SQ regression model to assist the initial quantile estimation, though it is not necessary for our theoretical analysis in Chapter 3. Moreover, due to the use of non-parametric initial SQ estimators, the m-Rock approach requires a much larger sample size when the covariate dimension increases.

4.2.2 More comparisons with a fixed K

In this section, we present more numerical results to compare the m-Rock approach with other approaches in the literature. We shall fix the value of K in Algorithm 1 to be $K = \sqrt{n}\log(n)/2$. We find the comparisons would not be qualitatively different when we use a different K at the same magnitude.

We compare with a limited number of approaches in the literature, including those in Section 3.5. The approaches can be categorized into three classes. The first class involves two-stage estimation of the quantile and SQ regression, which includes the Neyman-Orthogonalized Least-Squares (NO-LS) approach in *Barendse* (2020); and the simple Two-Step (TS-LS) approach described in (2.8) of Chapter 2. The second class is based on the joint estimation of *Dimitriadis et al.* (2020), where we consider two different specification functions $G_2(z) = \log(z)$ and $G_2(z) = \sqrt{z}$ as in Section 3.5 of Chapter 3, which we name Joint-1 and Joint-2 respectively. We rely on the `R` package `esreg` (*Dimitriadis and Bayer*, 2022) for computation of the Joint approaches. For the third class of method, we consider the Original Rockafellar (O-Rock) approach in *Rockafellar et al.* (2014), which is implemented based on the duality theory in *Miranda* (2014) and *Rockafellar and Royset* (2018). We only include the third method in limited scenarios, since it is theoretically biased and hence invalid in general; see our discussions in Chapter 1.

4.2.2.1 The quadratic-scale model revisited

Here we examine the one-dimensional Model (4.1) in more detail; we focus on the same Monte Carlo settings but with more competing methods. For the m-Rock approach, we include the same two implementations used in Section 4.2.1.1, where mRock-1 refers to using kNN quantile and mRock-2 refers to B-splines quantile.

Since the 90% conditional quantile of Y given X is not linear, the first two classes of methods are not directly applicable. We present the bias of those approaches in Figure 4.5. Except for the m-Rock approaches, all other methods are highly biased, especially for the intercept term β_0; this is because the bias from quantile regression carries over to the targeted SQ regression. On the other hand, the m-Rock approaches have no visible bias when the sample size is sufficiently large.

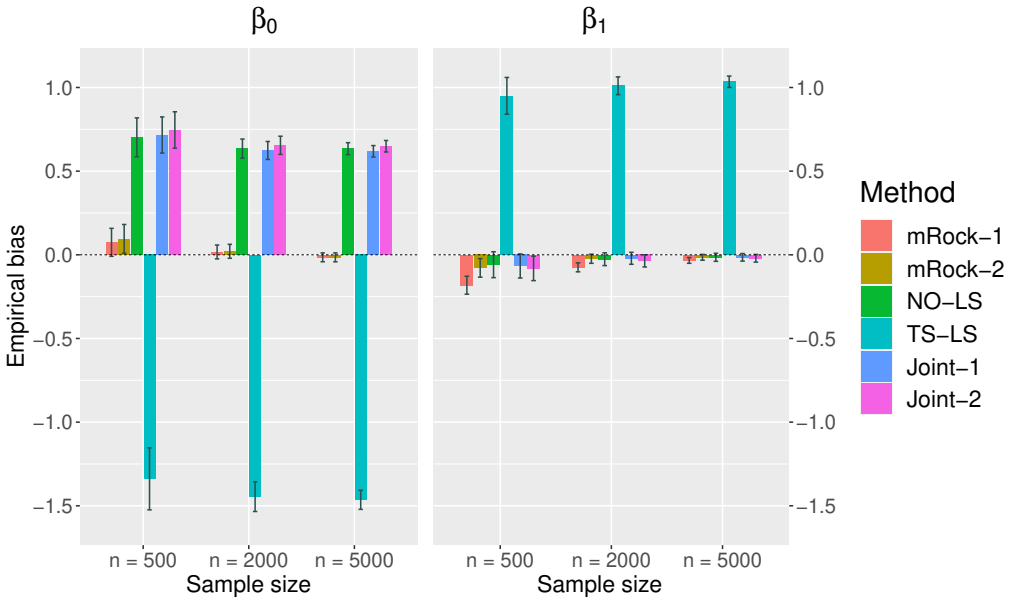


Figure 4.5: The bias for various SQ regression approaches under Model (4.1); the error bars show two times the estimated standard errors. For details on abbreviations of the methods' names; see the beginning of Section 4.2.2.

For a fair comparison, we now use the same B-splines quantile regression with 5 degrees of freedom for all the competing approaches, therefore alleviating the bias from quantile modeling. Operationally, we simply use the B-splines basis matrix

generated by the observed covariate data as the predictors for quantile regression modeling, which is straightforward to implement for all approaches in Classes 1 and 2[1]. For the m-Rock approach, we also focus on the implementation with B-splines quantile regression. The estimation accuracies are presented in Table 4.1.

Since the modeling of conditional quantile is the same, all approaches in Table 4.1 are comparable. We highlight several findings. First, the m-Rock approach is consistently more efficient than all competing methods in this example with heterogeneity, with around 30% reduction in RMSE and MAE. Second, all the other approaches share similar performance. While *Barendse* (2020) shows that choosing $G_2(z) = \log(z)$ is semi-parametric efficient under certain forms of heterogeneity, in the current Model (4.1) the Joint approaches do not offer any significant efficiency gain. Therefore, the m-Rock approach can help under broader forms of heterogeneity.

Furthermore, we include one further comparison with the weighted linearization method described in Equation (3.2). In practice, we can use an estimated weight and solve the feasible weighted least-squares (WLS) problem:

$$\min_{\theta} \sum_{j=1}^{m} \frac{[\hat{v}(\tau, X_{(j)}) - X_{(j)}^T \theta]^2}{\hat{v}(\tau, X_{(j)}) - \hat{q}(\tau, X_{(j)})}, \tag{4.3}$$

where both $\hat{v}(\tau, X_{(j)})$ and $\hat{q}(\tau, X_{(j)})$ are available from Steps 6 and 7 in Algorithm 1. Parallel to the m-Rock implementations, we consider two linearization methods, labeled by Linearize-1 and Linearize-2; the former uses kNN quantile regression and the latter uses B-splines quantile regression to estimate the weights. We also include the oracle linearization method (Linearize-OR) as if the weights $v(\tau, x) - q(\tau, x)$ in (4.3) were known.

Figure 4.6 compares the estimation accuracy of the linearization and the m-Rock approaches. We observe that the m-Rock approaches are similar to the (infeasible)

[1]Though the theoretical results for those approaches may not apply.

Linearize-OR approach, thereby agreeing with the asymptotic theory in Theorem III.1. On the other hand, the linearization approaches using estimated weights do not achieve the same estimation accuracy with limited sample sizes. Therefore, the m-Rock approach achieves implicit oracle weighting[2] without having to estimate the weights, which may not be consistently reliable in practice with limited sample size.

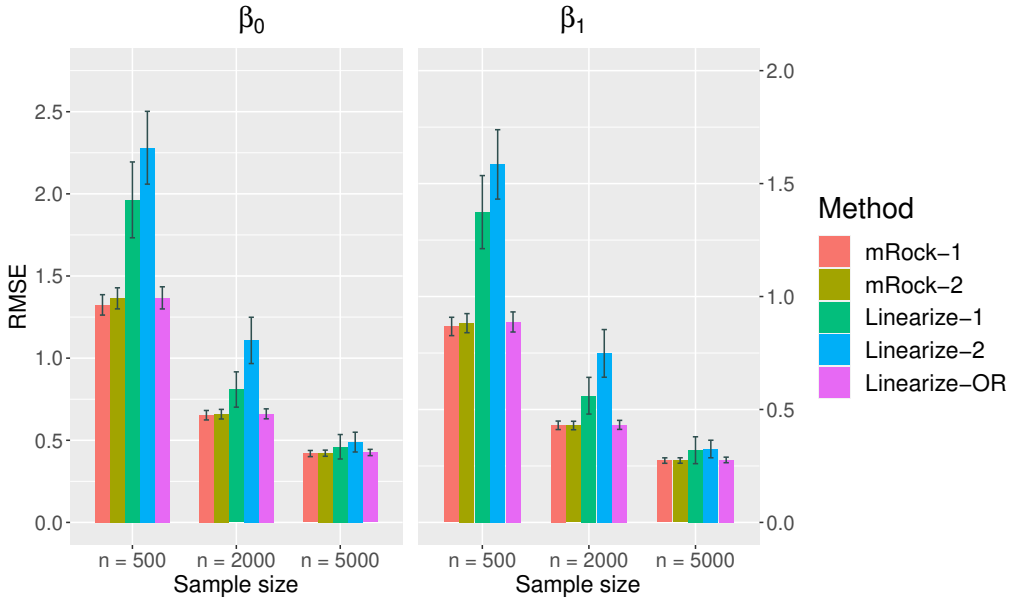


Figure 4.6: The RMSE of m-Rock and linearization approaches under Model (4.1). The error bars show two times the estimated standard errors. For details on abbreviations of the methods' names; see Section 4.2.2.

[2]Such 'oracle' weighting does not refer to the semi-parametric efficient weights, but the true weight in (4.3).

Table 4.1: The estimation accuracy for 90% SQ regression under Model (4.1); the conditional quantiles are modeled by B-splines regression with 5 degrees of freedom for all methods. RMSE is the root-mean-squared error, and MAE is the mean absolute error. The numbers in parentheses are the estimated standard errors.

Method		Bias (×10)		RMSE (×10)		MAE (×10)	
		β_0	β_1	β_0	β_1	β_0	β_1
$n = 500$							
m-Rock	$\hat{q}_\tau(x)$ by B-splines	0.95 (0.42)	-0.78 (0.27)	13.64 (0.29)	8.82 (0.19)	10.74 (0.25)	6.96 (0.16)
Two-stage	NO-LS	0.69 (0.56)	-1.31 (0.38)	19.13 (0.43)	12.88 (0.28)	14.67 (0.34)	9.99 (0.23)
	TS-LS	0.74 (0.57)	-1.52 (0.38)	19.66 (0.43)	13.23 (0.28)	15.14 (0.35)	10.33 (0.23)
Joint	$G_2(z) = \log(z)$	0.76 (0.53)	-1.36 (0.35)	17.88 (0.39)	11.97 (0.25)	13.96 (0.32)	9.39 (0.21)
	$G_2(z) = \sqrt{z}$	0.54 (0.53)	-1.26 (0.36)	18.08 (0.40)	12.13 (0.25)	14.04 (0.33)	9.46 (0.21)
$n = 2000$							
m-Rock	$\hat{q}_\tau(x)$ by B-splines	0.21 (0.22)	-0.24 (0.14)	6.59 (0.15)	4.29 (0.10)	5.21 (0.13)	3.40 (0.08)
Two-stage	NO-LS	0.06 (0.30)	-0.30 (0.20)	9.48 (0.25)	6.42 (0.18)	7.41 (0.19)	5.04 (0.13)
	TS-LS	0.02 (0.30)	-0.35 (0.20)	9.53 (0.26)	6.46 (0.19)	7.45 (0.19)	5.08 (0.13)
Joint	$G_2(z) = \log(z)$	0.05 (0.28)	-0.32 (0.19)	8.92 (0.22)	6.02 (0.16)	6.98 (0.18)	4.74 (0.12)
	$G_2(z) = \sqrt{z}$	-0.02 (0.29)	-0.30 (0.19)	9.03 (0.22)	6.11 (0.16)	7.04 (0.18)	4.79 (0.12)
$n = 5000$							
m-Rock	$\hat{q}_\tau(x)$ by B-splines	0.00 (0.14)	-0.04 (0.09)	4.21 (0.09)	2.74 (0.06)	3.35 (0.08)	2.19 (0.05)
Two-stage	NO-LS	0.06 (0.19)	-0.07 (0.13)	5.95 (0.13)	4.02 (0.09)	4.74 (0.11)	3.21 (0.08)
	TS-LS	0.03 (0.19)	-0.07 (0.13)	5.96 (0.13)	4.03 (0.09)	4.74 (0.11)	3.21 (0.08)
Joint	$G_2(z) = \log(z)$	0.01 (0.18)	-0.07 (0.12)	5.67 (0.13)	3.81 (0.08)	4.52 (0.11)	3.04 (0.07)
	$G_2(z) = \sqrt{z}$	-0.03 (0.18)	-0.06 (0.12)	5.68 (0.13)	3.84 (0.09)	4.53 (0.11)	3.06 (0.07)

4.2.2.2 The linear location-scale shift models

Here we consider the linear model:

$$Y = -2 + 2X + (1 + \gamma X)\varepsilon, \tag{4.4}$$

where X is uniformly distributed on $[0, 4]$, $\gamma > 0$ controls the heterogeneity and ε is the error term independent of X. Depending on the heterogeneity parameter γ and the error term ε, Equation (4.4) represents a class of linear location-scale shift models. Here we consider four different specifications, where ε follows either (i) a standard normal distribution, or (ii) a standardized skewed-t_5 distribution in (4.1) with skewness 2; and γ is set to be (i) $\gamma = 2$ or (ii) $\gamma = 0$. Therefore, we cover both homogeneous and heterogeneous cases with different error distributions.

For each model specification, we consider three sample sizes $n = 500$; $n = 2,000$ and $n = 5,000$. For each Monte Carlo dataset, we estimate the SQ regression coefficients at four different quantile levels $\tau = 0.8$, $\tau = 0.9$, $\tau = 0.95$ and $\tau = 0.975$. For the m-Rock approaches, we consider two implementations with (i) kNN quantile, and (ii) linear quantile. We also include all three classes of competing methods described in Section 4.2.2.

Tables 4.2 and 4.3 show the average RMSE across the two SQ regression coefficients β_0 and β_1 under Model (4.4); each one of the tables is for a fixed γ value, representing a scenario with homogeneous or heterogeneous model. For each sample size n and quantile level τ, we report the maximum of estimated standard errors among all methods, since the standard errors for different methods are comparable in this example.

With sufficiently large sample sizes and/or less extreme quantile levels, both implementations of the m-Rock approach are among the most efficient in all scenarios. With homogeneous models ($\gamma = 0$), Table 4.2 shows the m-Rock approaches and the

Table 4.2: The average RMSE (multiplied by 10) for the SQ regression coefficients under Model (4.4) in homogeneous settings with $\gamma = 0$. The numbers in the parentheses show the maximum estimated standard error across all methods for each (n, τ). For abbreviations of methods' names, see Section 4.2.2.

Method		Normal error					Skewed-t error				
		Quantile levels (%)					Quantile levels (%)				
		80	90	95	97.5	99	80	90	95	97.5	99
$n = 500$											
m-Rock	kNN $\hat{q}$	1.02	1.26	1.64	2.12	2.83	1.80	2.77	4.22	6.40	10.97
	linear $\hat{q}$	1.02	1.27	1.66	2.24	3.35	1.83	2.85	4.57	7.53	14.66
Rock	Original	1.06	1.32	1.67	2.17	3.13	1.68	2.60	4.16	6.54	14.68
Two-stage	NO-LS	1.01	1.25	1.60	2.03	2.89	1.80	2.81	4.46	7.06	12.57
	TS-LS	1.01	1.27	1.61	2.04	3.02	1.80	2.84	4.47	7.10	13.03
Joint	Joint-1	1.17	1.44	1.80	2.28	3.16	2.15	3.18	4.84	7.42	13.04
	Joint-2	1.11	1.36	1.71	2.18	3.05	1.99	3.01	4.65	7.21	12.66
	max. s.e.	(0.02)	(0.03)	(0.04)	(0.05)	(0.08)	(0.05)	(0.07)	(0.15)	(0.25)	(0.99)
$n = 2000$											
m-Rock	kNN $\hat{q}$	0.50	0.64	0.83	1.11	1.61	0.95	1.46	2.27	3.51	6.16
	linear $\hat{q}$	0.50	0.64	0.81	1.05	1.58	0.94	1.45	2.25	3.56	6.65
Rock	Original	0.52	0.65	0.83	1.06	1.54	0.88	1.33	2.06	3.20	5.80
Two-stage	NO-LS	0.50	0.63	0.80	1.02	1.45	0.94	1.44	2.24	3.52	6.35
	TS-LS	0.50	0.63	0.79	1.02	1.45	0.93	1.44	2.23	3.50	6.37
Joint	Joint-1	0.57	0.71	0.89	1.13	1.60	1.09	1.62	2.44	3.73	6.55
	Joint-2	0.54	0.67	0.85	1.09	1.54	1.03	1.54	2.35	3.62	6.41
	max.s.e.	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.05)	(0.08)	(0.16)
$n = 5000$											
m-Rock	kNN $\hat{q}$	0.31	0.40	0.52	0.70	1.07	0.58	0.90	1.41	2.25	4.11
	linear $\hat{q}$	0.31	0.40	0.51	0.65	0.99	0.58	0.90	1.40	2.20	4.03
Rock	Original	0.32	0.41	0.52	0.66	0.96	0.54	0.83	1.28	1.99	3.68
Two-stage	NO-LS	0.31	0.39	0.50	0.64	0.92	0.58	0.90	1.41	2.24	4.18
	TS-LS	0.31	0.39	0.50	0.64	0.93	0.58	0.90	1.41	2.24	4.23
Joint	Joint-1	0.35	0.44	0.55	0.71	1.01	0.66	1.00	1.53	2.39	4.36
	Joint-2	0.33	0.42	0.53	0.68	0.98	0.62	0.95	1.48	2.32	4.27
	max. s.e.	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.04)	(0.07)	(0.19)

Table 4.3: The average RMSE (multiplied by 10) for the SQ regression coefficients under Model (4.4) in heterogeneous settings with $\gamma = 2$. Other attributes of the table are the same as Table 4.2.

Method		Normal error					Skewed-t error				
		Quantile levels (%)					Quantile levels (%)				
		80	90	95	97.5	99	80	90	95	97.5	99
$n = 500$											
m-Rock	kNN $\hat{q}$	3.23	4.19	5.44	7.06	9.74	5.78	8.80	13.46	21.17	41.74
	linear $\hat{q}$	3.26	4.27	5.66	8.11	13.11	5.89	9.13	14.69	25.15	55.99
Rock	Original	3.76	4.62	5.80	7.36	12.00	9.25	11.96	16.46	23.69	49.07
Two-stage	NO-LS	4.27	5.44	7.05	9.10	12.99	7.98	12.10	18.89	29.99	55.43
	TS-LS	3.68	4.60	6.11	7.94	13.42	7.00	10.70	16.94	27.32	51.06
Joint	Joint-1	3.20	3.99	5.11	6.61	9.33	5.37	8.00	12.31	19.14	33.19
	Joint-2	3.40	4.25	5.46	7.04	10.06	5.75	8.55	13.14	20.50	36.35
	max. s.e.	(0.09)	(0.12)	(0.16)	(0.24)	(1.11)	(0.18)	(0.29)	(0.57)	(1.83)	(4.13)
$n = 2000$											
m-Rock	kNN $\hat{q}$	1.60	2.04	2.67	3.51	5.01	2.92	4.44	6.77	10.61	19.22
	linear $\hat{q}$	1.60	2.05	2.67	3.48	5.34	2.94	4.48	6.86	11.11	21.70
Rock	Original	2.41	2.84	3.32	3.92	5.24	7.94	9.65	11.78	15.16	23.30
Two-stage	NO-LS	2.23	2.82	3.60	4.60	6.47	4.04	6.15	9.43	14.77	26.94
	TS-LS	1.91	2.38	3.02	3.90	5.51	3.61	5.43	8.41	13.08	24.12
Joint	Joint-1	1.72	2.12	2.67	3.36	4.69	2.90	4.29	6.41	9.84	17.62
	Joint-2	1.82	2.26	2.84	3.58	5.02	3.10	4.60	6.87	10.55	18.88
	max. s.e.	(0.05)	(0.06)	(0.08)	(0.10)	(0.15)	(0.08)	(0.13)	(0.20)	(0.34)	(0.70)
$n = 5000$											
m-Rock	kNN $\hat{q}$	0.95	1.22	1.62	2.18	3.26	1.83	2.82	4.35	6.80	12.50
	linear $\hat{q}$	0.95	1.23	1.62	2.13	3.19	1.83	2.82	4.34	6.78	12.87
Rock	Original	1.92	2.21	2.49	2.85	3.59	7.51	9.03	10.77	13.13	18.03
Two-stage	NO-LS	1.34	1.71	2.21	2.85	4.06	2.57	3.96	6.18	9.69	17.83
	TS-LS	1.13	1.45	1.88	2.43	3.48	2.27	3.55	5.55	8.59	15.71
Joint	Joint-1	1.04	1.31	1.68	2.16	3.05	1.86	2.79	4.21	6.38	11.40
	Joint-2	1.10	1.39	1.78	2.28	3.23	1.99	2.99	4.53	6.87	12.28
	max. s.e.	(0.03)	(0.04)	(0.05)	(0.06)	(0.08)	(0.06)	(0.09)	(0.14)	(0.23)	(0.50)

two-stage approaches are more efficient than the Joint approaches. When heterogeneity is present ($\gamma = 2$), however, the Joint approach with $G_2(z) = \log(z)$ and the m-Rock approaches are similar and consistently the most efficient[3]. These efficiency comparisons are in line with our theoretical findings in Chapter 3; and the differences between methods are more evident with heavy-tailed errors. To conclude, the estimation accuracy of the m-Rock approach remains competitive in either homogeneous or heterogeneous scenarios.

We note that SQ regression can be less stable with limited sample sizes and/or at extreme quantile levels. The m-Rock approach can be especially sensitive due to its need for an initial estimator at a range of quantile levels, e.g., when targeting the 99% SQ regression, we often need to estimate up to the 99.9% initial SQ for the m-Rock approach, which can be unstable with limited sample sizes. Therefore in Tables 4.2 and 4.3, we sometimes observe the m-Rock approach can be less competitive at extreme quantile levels (e.g., $\tau = 0.975$ or $\tau = 0.99$).

We further zoom in to the case of $n = 2000$, $\tau = 0.9$ with skewed-t_5 error in Figure 4.7, which shows a more detailed bias-variance decomposition of the results in Tables 4.2 and 4.3. We highlight two findings from Figure 4.7. First, Figure 4.7 visualizes the adaptivity of the m-Rock approach: the variances are among the smallest in either homoscedastic or heteroscedastic cases. Note, however, that the m-Rock approach with kNN quantile can introduce some more bias with limited sample sizes.

Second, Figure 4.7 informs us that the Original Rock approach can be highly biased in heterogeneous scenarios, though it can be valid and even more efficient than others in homogeneous settings with heavy-tailed errors. In Figure 4.8, we further compare the bias for the Original Rock approach with the m-Rock approach. The bias for the O-Rock approach is persistent, while the bias for our m-Rock approach

[3]We note that the `R` package `esreg` for the Joint approach will shift the response variable before fitting, therefore the numerical performance may differ from the theoretical discussions in Section 3.5.

vanishes as the sample size increases; Therefore, our modification is critical for a consistent SQ regression. These results agree with our findings in Chapter 1.

Overall, we conclude that the m-Rock approach demonstrates desirable estimation accuracy that is automatically adaptive to the heterogeneity in data. The m-Rock approach can also incorporate either parametric or non-parametric quantile regression estimators; and when the sample size is large, its performance does not change much with the quantile estimation. Those empirical findings corroborate our theoretical discussions in Chapter 3.

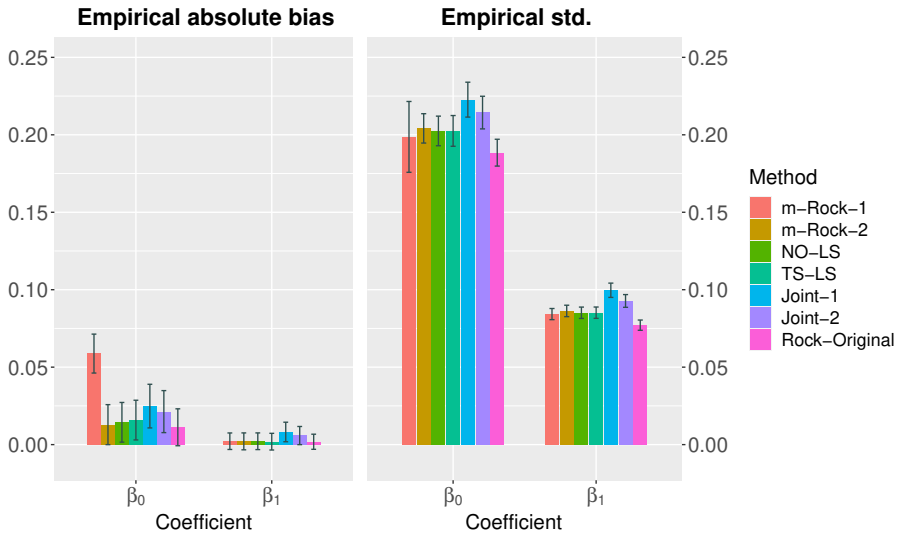
4.3 Empirical data applications

In this section, we use two empirical applications to illustrate the use of the m-Rock approach in practice. In our examples, the standard errors for all involved statistical procedures are from $B = 500$ bootstrap samples.

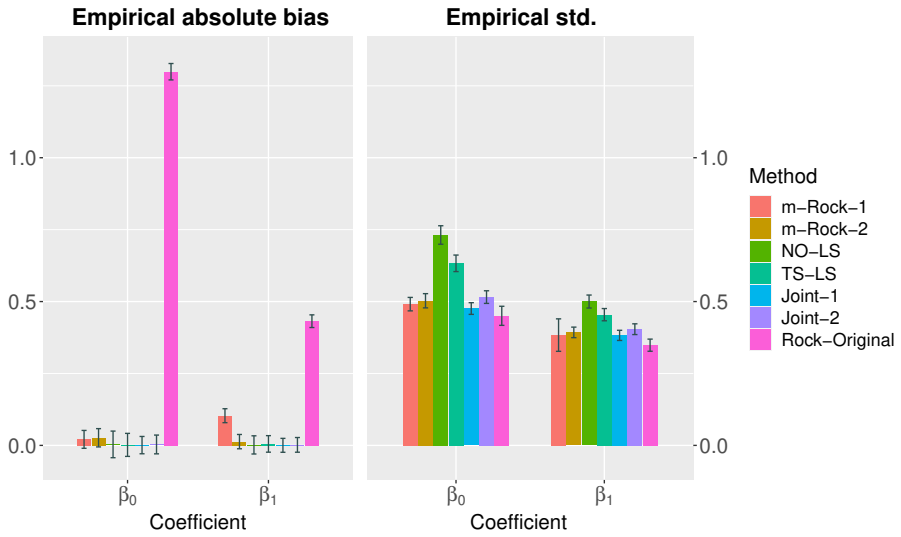
4.3.1 Financial data

We first study an example related to financial risk analysis. In financial applications, superquantile is often used to quantify the risk in an investment portfolio, which reflects the potential losses in adverse situations. In this example, we use the proposed m-Rock approach to study the risk exposures of some investment portfolios to the Fama-French (F-F) three factors model (*Fama and French*, 1993, 1995). Our settings are similar to those in *Chetverikov et al.* (2022) and *Barendse* (2020).

We investigate the superquantile risk of 6 different investment strategies based on the the company's size and operating profitability (OP) (*Novy-Marx*, 2013). The first four portfolios are the double-sorted portfolios based on small/large market capitalization (i.e., size) and low/high OP. The other two portfolios are of a long-short type that takes a long position in high OP stocks and a short position in low OP stocks; each of those two portfolios focuses only on stocks with big (or small) market

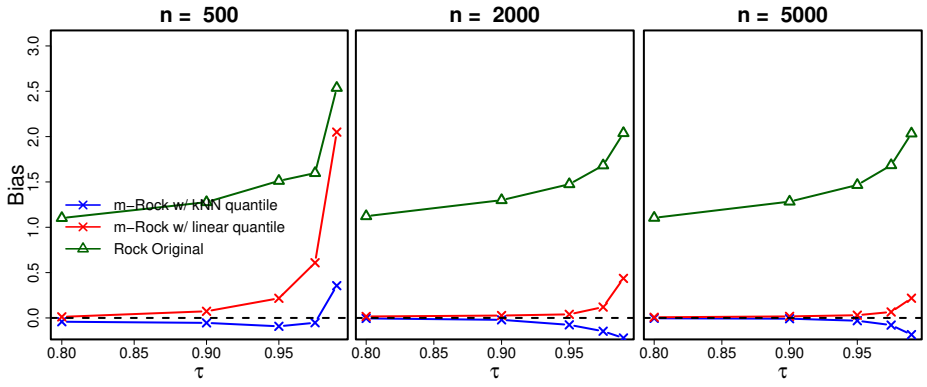


(a) For the homoscedastic case

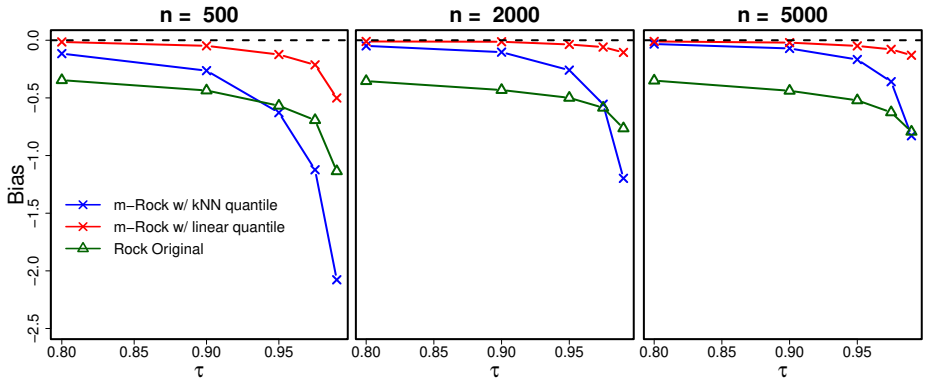


(b) For the heteroscedastic case

Figure 4.7: The absolute bias and standard deviation when $\tau = 0.9$ and $n = 2000$ under Model (4.4) with skewed-t error. The error bars show two times the estimated standard errors. For abbreviations of methods' names, see Section 4.2.2.



(a) For the intercept term β_0



(b) For the slope term β_1

Figure 4.8: The bias for each coefficient at different quantile levels τ under Model (4.4) with heterogeneity ($\gamma = 2$) and skewed-t error.

capitalization. For the covariates, we use the Fama-French three factors, which include the market factor (MktRF), the size factor (SMB), and the value factor (HML). Those factors are standard in the financial literature as systematic macro-economic risk factors. We refer to *Fama and French* (1993) for more detailed discussions of factor models; and *Fama and French* (2015) for the portfolio construction based on profitability. We focus on the U.S. stock market in this example; all data for the investment portfolios and the F-F factors are publicly available from the data library of Professor Ken French (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

In our analysis, we use the daily loss percentage (negative return) of the six portfolios as different response variables. The sampling period is from July 1963 to

April 2022, which consists of $n = 14,810$ daily observations. Table 4.4 presents the result for $\tau = 95\%$ superquantile regression using the m-Rock approach, where we fix the nearest-neighbour parameter $K = 584$ in Algorithm 1. To properly capture the serial correlation, the reported standard errors are from the block bootstrap in *Kunsch* (1989) using a fixed block length of one year. We use the name 'Small/hi' to represent the portfolio with small size and high profitability, and 'Small/LS' for the portfolio with small size and takes a long-short position based on profitability; the names of other portfolios follow the same convention. Each row in Table 4.4 represents one regression model with the corresponding portfolio as the response. In what follows, we refer to the estimated coefficient for each factor as factor loadings.

In Table 4.4, the factor loadings represent the risk exposures of the targeting portfolio on the Fama-French factor returns; In particular, a negative loading indicates that risk decreases, i.e., a smaller loss, when the corresponding factor return is high. The left and right panels in Table 4.4 display similar results for the factor loadings; yet the approach with kNN quantile consistently under-estimates the intercept term α. This observation is in line with our result in Section 4.2.2 that kNN quantile may be more biased. Therefore, we focus on the linear quantile in our subsequent discussion.

In general, the factor loadings are quite heterogeneous across the six portfolios. For the first four long-only portfolios, the risk exposures to the market factor are all approximately -1, which is consistent with the traditional CAPM model; since we focus on the loss (negative return) distribution, these exposures are negative. The portfolios of smaller market caps also have significant negative risk exposures to the size factor; this means that smaller-sized companies have lower risk when the size premium is high. This is intuitive as the size premium is driven by the excess return of smaller-sized companies. Except for the Small/hi portfolio, the long-only portfolios have relatively little risk exposure to the value factor.

Table 4.4: The 95% SQ regression using the m-Rock approach for the six investment portfolios. Left/right panels reflects two m-Rock implementations using different conditional quantile estimators. The term α is the estimated intercept. The numbers in the parentheses show the block bootstrap standard errors.

Portfolios	linear quantile				kNN quantile			
		F-F factors				F-F factors		
	α (%)	MktRF	SMB	HML	α (%)	MktRF	SMB	HML
Small/low	0.578 (0.064)	-1.069 (0.020)	-1.034 (0.047)	-0.090 (0.049)	0.509 (0.047)	-1.074 (0.016)	-1.027 (0.034)	-0.098 (0.038)
Small/hi	0.717 (0.107)	-1.015 (0.023)	-0.866 (0.051)	-0.433 (0.076)	0.603 (0.071)	-1.027 (0.021)	-0.853 (0.044)	-0.406 (0.065)
Big/low	0.914 (0.126)	-1.146 (0.039)	-0.223 (0.068)	0.009 (0.087)	0.793 (0.089)	-1.133 (0.031)	-0.206 (0.053)	-0.035 (0.072)
Big/hi	0.450 (0.066)	-0.938 (0.024)	0.186 (0.034)	0.087 (0.043)	0.404 (0.047)	-0.946 (0.021)	0.183 (0.029)	0.099 (0.039)
Small/LS	1.007 (0.126)	0.075 (0.030)	0.214 (0.077)	-0.323 (0.101)	0.879 (0.087)	0.065 (0.029)	0.202 (0.066)	-0.303 (0.091)
Big/LS	1.246 (0.167)	0.226 (0.058)	0.286 (0.078)	0.126 (0.110)	1.108 (0.123)	0.208 (0.047)	0.306 (0.069)	0.158 (0.098)

For the two long-short type portfolios, the risk exposures to all three factors are relatively small. In particular, the exposures to the market factor are now approximately zero, or even positive; By taking long-short positions, the market risk is hedged away in these portfolios. Moreover, those long-short type portfolios have the largest estimated α among the six portfolios. Note α quantifies the potential loss that is not captured by the F-F factors (*Barendse*, 2020). As in Table 4.4, both the long-short portfolio has an unexplained average daily loss of over 1% in the worst 5% situations. Following the reasoning in *Fama and French* (2015), these long-short portfolios are more focused on the profitability effect yet have less exposure to other systematic risks, comparing with the long-only portfolios. Our results corroborate the findings in *Novy-Marx* (2013) that the F-F three factors model is insufficient to explain variations in the stock market related to profitability.

We also give the results for the Original Rock approach in Table 4.5. While the Original Rock approach is not valid in general, in this example it gives similar results to the m-Rock approach. The reason is the lack of strong heterogeneity in this dataset. Nonetheless, when comparing with the m-Rock approach using linear quantile, the Original Rock approach seems to consistently under-estimate α.

Table 4.5: The 95% SQ regression from the Original Rockafellar's approach for the six investment portfolios; the setting is the same as Table 4.4.

Portfolios		F-F factors		
	α (%)	MktRF	SMB	HML
Small/low	0.508	-1.081	-1.012	-0.110
	(0.041)	(0.014)	(0.034)	(0.039)
Small/hi	0.588	-1.034	-0.864	-0.407
	(0.057)	(0.018)	(0.052)	(0.055)
Big/low	0.793	-1.128	-0.214	-0.046
	(0.078)	(0.031)	(0.042)	(0.076)
Big/hi	0.399	-0.952	0.190	0.092
	(0.043)	(0.019)	(0.024)	(0.046)
Small/LS	0.875	0.064	0.160	-0.270
	(0.070)	(0.025)	(0.075)	(0.086)
Big/LS	1.088	0.200	0.313	0.166
	(0.120)	(0.045)	(0.076)	(0.118)

Furthermore, we compare our results with those obtained using the NO-LS approach and the Joint approach with $G_2(z) = \log(z)$. For our m-Rock approach, we focus on the one with linear quantile. Figure 4.9 shows the point estimates and the standard error bars for the six investment portfolios. While all methods give qualitatively similar point estimates, the estimated standard errors of our m-Rock approach are consistently among the smallest. These results indicate that our m-Rock approach may offer improved estimation efficiency compared to other approaches in the literature.

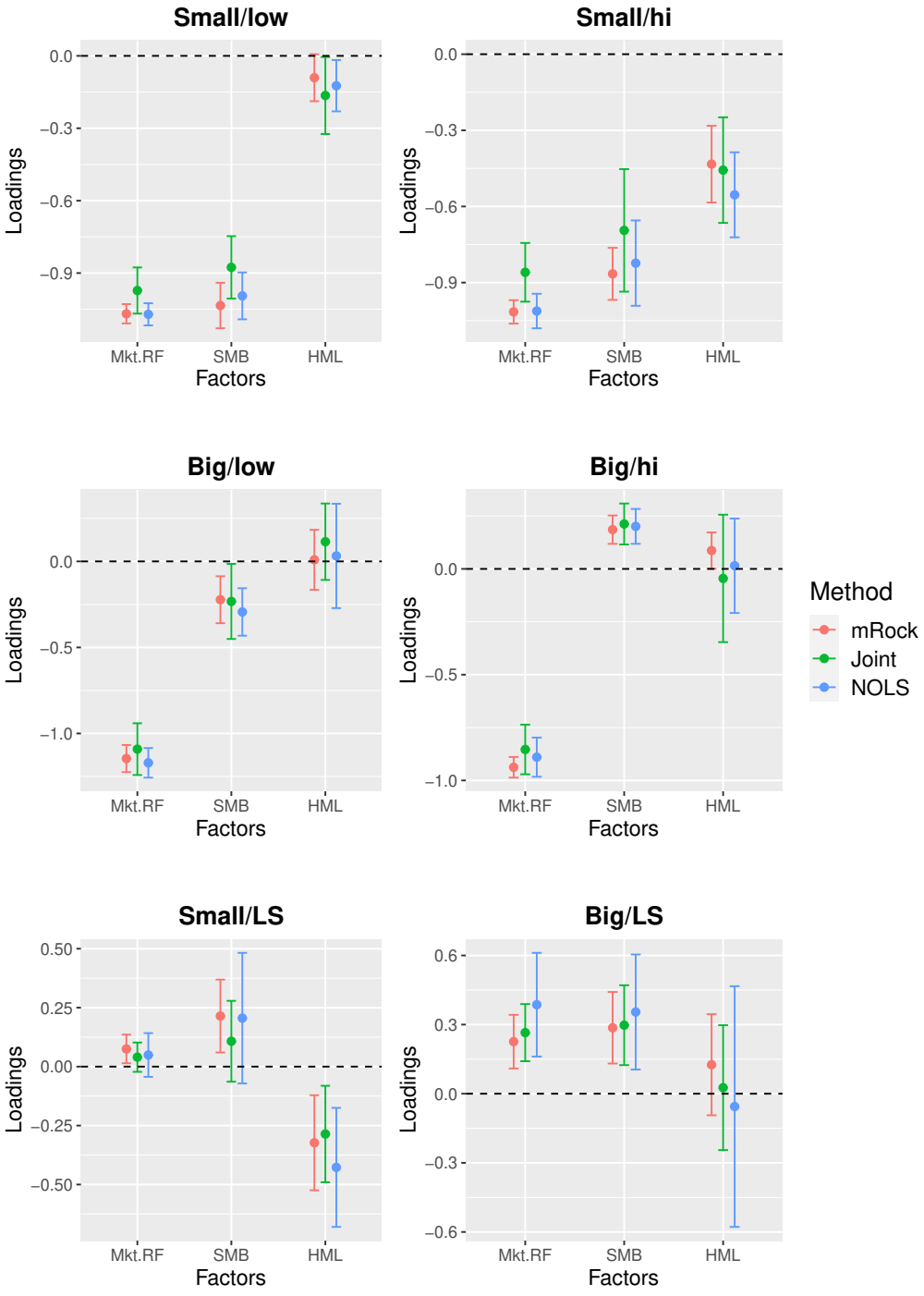


Figure 4.9: The comparison between the m-Rock approach, the Joint approach, and the NOLS approach. Each panel represents an investment portfolio, and the x-axis shows three F-F factors. We omit the intercept terms. The error bars show two times the block bootstrap standard errors.

4.3.2 Birth-weight data

Here we study another example related to birth weight. A low birth weight ($<$ 2500g) is long-known to be associated with increased infant mortality risk and long-term health issues; See *Hughes et al.* (2017) for a recent review. Since low birth weights are considered at risk, in this example we focus on the lower-superquantile of the birth weight distribution, which is the average birth weight *below* a certain quantile level. The method developed in the dissertation is still applicable for the problem by reversing the direction of both the covariate and response [4].

We focus on the effect of parity on the birth weight distribution. Parity is defined as the number of live births a mother has given, e.g., a parity of 1 indicates the first live birth of a mother. In our analysis, we consider the difference in birth weight between two groups: (i) parity $= 1$ and (ii) parity > 1; therefore, the variable of interest is a binary indicator, where parity of 1 is the reference level. This reference group is commonly referred to as *nulliparous* in the public health setting. Many previous studies (*Shah*, 2010; *Duong et al.*, 2012; *Hinkle et al.*, 2014; *Lin et al.*, 2021) have shown that nulliparous mothers are exposed to higher risk of low birth weight and/or birth defect. In this example, we use the m-Rock approach to study the superquantile effect of parity on the birth weight distribution.

The data we use is the 2020 U.S. birth-weight dataset, which is available online at the National Center for Health Statistics (`https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`). In our analysis, we focus on the cases of male singleton births only; and we restrict to the subpopulation of black or white mothers that are recorded to be married, college-educated, non-smokers, and at least 36 years old. This subpopulation focuses on the mothers that are relatively older but in good conditions otherwise; therefore, taking the subpopulation eliminates some possible

[4] The lower τth superquantile of Y given X is equivalent to the upper $(1-\tau)$th superquantile of $-Y$ given $-X$.

confounders in our analysis of parity (*Yang et al.*, 2006; *Muula et al.*, 2011). We include three other maternal factors in our SQ regression model: mother's race, age, and weight gain during pregnancy; We also include a quadratic term for the mother's weight gain after centering the variable. Any record with missing data is removed. We retain $n = 79,336$ birth records from the 2020 U.S. birth-weight dataset. Some summary statistics are presented in Table 4.6.

Table 4.6: Average values of the variables used in the birth weight example, stratified by parity groups. For continuous variables, the numbers in parentheses are the interquartile range.

	Parity = 1	Parity > 1
Birth weight (g)	3301 (3005 – 3657)	3482 (3185 – 3820)
Maternal age	38.0 (36.0 – 39.0)	38.1 (36.0 – 39.0)
Gestational weight gain (lb)	30.4 (21.0 – 38.0)	30.2 (22.0 – 38.0)
Race = Black (%)	10.2	10.7

Table 4.7 presents the result using m-Rock SQ regression at two quantile levels $\tau = 0.05$ and $\tau = 0.2$, where we fix $K = 787$ and use linear quantile in the m-Rock approach. The unit for the reported birth weight effect is in grams. We observe that the nulliparous group (i.e., parity = 1) has an adverse effect on the lower end of the birth weight distribution; Moreover, the effect is heterogeneous and not a simple location shift. For those with the 5% lowest birth weight, babies born to nulliparous mothers are, on average, over 300 grams lighter than those born to multiparous mothers. This effect may be interpreted as that mothers giving their first birth may be inexperienced in pregnancy and giving births (*Bisai et al.*, 2006; *Muula et al.*, 2011).

In addition, we observe in Table 4.7 that race is another significant risk factor. The difference in birth weight between white and black mothers is over half a kilogram at the 5% superquantile. Coupling this effect with the estimated intercept term, we obtain that the 20% SQ for the birth weight distribution for black mothers at the age

of 38 years old and had a weight gain of 30 pounds[5] is less than $2,500$ grams; such a low birth weight is considered to at risk in the public health context (*Hughes et al.*, 2017).

Table 4.7: The lower-SQ regression using the m-Rock approach for the birth weight example. White mothers with parity 1 are the baseline groups; The other continuous covariates are centered prior to the regression. The numbers in the parenthesis show the bootstrap standard errors.

Covariates	Quantile levels	
	0.05	0.2
(Intercept)	2006.83 (21.53)	2580.32 (9.62)
Parity > 1	325.93 (23.35)	241.22 (10.64)
Race = Black	-505.92 (39.09)	-297.43 (16.62)
Mom age	-34.03 (5.33)	-22.20 (2.48)
Mom weight gain	21.39 (0.92)	12.67 (0.43)
Mom weight gain^2	-0.41 (0.04)	-0.22 (0.02)

Furthermore, we compare the mean, quantile, and superquantile effects of parity and race in Figure 4.10. The effects of parity and race are differential, in the sense that the effects at lower tail of the birth weight distribution are different than the average effect. Figure 4.10 suggests that our superquantile-based approach can better capture those differential effects, whereas the quantile effect at 20% or 25% does not show a significant difference from the mean-effect. Furthermore, the SQ effect curves are smoother than the quantile regression curves, because the SQ is an average over the entire tail distribution.

As a comparison to the results of the m-Rock approach, we also give the results

[5]These are the average age and weight gain in the sample from Table 4.6.

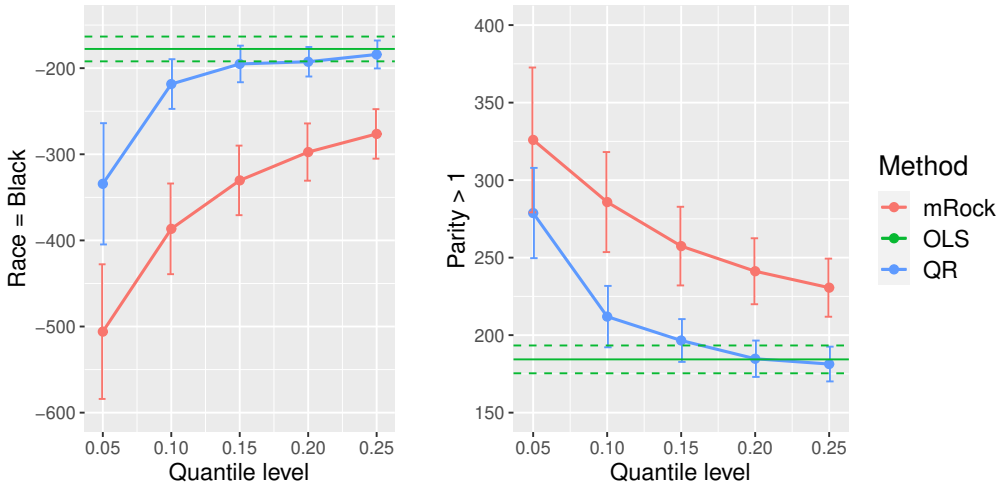


Figure 4.10: The race and parity effects estimated from the m-Rock (lower-)SQ regression, quantile regression (QR) and OLS; the reference levels are white mothers with parity > 1. The error bars show two times the bootstrap standard errors.

using the NO-LS approach and the O-Rock approach in Table 4.8[6]; see the beginning of Section 4.2.2 for abbreviations of methods' names. Comparing with Table 4.7, the NO-LS approach give similar results to our m-Rock approach; the O-Rock approach often underestimates the SQ effects by a magnitude of one standard error. This suggests that the O-Rock approach may implicitly shrink the coefficients towards 0. These findings echo our simulation findings in Section 4.2.2 that the O-Rock approach can be biased but has reduced variance.

4.4 Discussion

In this chapter, we use simulations and empirical applications to demonstrate the performance of the m-Rock approach. Our m-Rock implementation depends on a tuning parameter K, and a set of quantile regression estimators $\hat{q}(s, x)$. We demonstrate via numerical experiments that the m-Rock approach is relatively stable under a wide range of K, and several different choices of $\hat{q}(s, x)$. In practice, we

[6]The Joint approach is excluded because the algorithm in the `esreg` package has convergence issues in our example

Table 4.8: The lower-SQ regression using the O-Rock and NO-LS approaches for the birth weight data. Other attributes in the table are the same as Table 4.7

Covariates	O-Rock		NO-LS	
	Quantile levels		Quantile levels	
	0.05	0.2	0.05	0.2
(Intercept)	2002.53 (19.79)	2570.77 (9.08)	1996.39 (21.69)	2569.08 (9.62)
Parity > 1	327.33 (21.56)	239.08 (9.64)	341.10 (23.85)	247.71 (10.63)
Race = Black	-507.97 (43.49)	-286.75 (18.44)	-497.74 (40.49)	-289.97 (17.77)
Mom age	-29.04 (4.38)	-18.95 (2.07)	-28.89 (4.67)	-20.09 (2.04)
Mom weight gain	21.46 (0.82)	13.11 (0.44)	21.26 (0.82)	12.65 (0.39)
Mom weight gain^2	-0.40 (0.03)	-0.21 (0.02)	-0.37 (0.03)	-0.19 (0.02)

find that it is often the best to use a parametric quantile regression estimator for the m-Rock approach, especially with limited sample sizes and/or multiple continuous covariates; even though our theoretical analysis in Section 3.3 does not rely on a linear quantile regression model.

In our simulations, the m-Rock approach demonstrates desirable estimation efficiency in a wide range of models. Parallel to our discussions in Section 3.5, we confirm numerically that the m-Rock approach is adaptive to a broad form of heterogeneity. Our empirical studies further illustrate how the m-Rock approach can be useful for data analysis in financial and public health applications.

We hasten to add that our Algorithm 1 is only a prototype implementation for the m-Rock approach, and our empirical applications are relatively simple. Importantly, we have restricted examples with only a few continuous covariates because the kNN binning strategy can be unstable otherwise. Furthermore, it is not yet clear how to select the tuning parameter in a data-driven way. Therefore, we would need to develop

a more general implementation of better applicability of the m-Rock approach.

CHAPTER V

Posterior Inference for Quantile Regression with Shrinkage Priors

5.1 Introduction

Quantile regression, since its first debut in *Koenker and Bassett Jr* (1978), has become a popular data analysis tool in a wide range of applications, from economics (*Fitzenberger et al.*, 2013) to public health (*Wei et al.*, 2019). Importantly, quantile regression allows researchers to go beyond the conditional mean analysis: it examines the effect of the covariates at different conditional quantile levels, thus providing more comprehensive information on the relationship between the response and the covariates. Another celebrated virtue of quantile regression is its robustness. In the presence of heavy tails or extreme outliers, the median regression, also known as the Least Absolute Deviation regression, serves as an attractive alternative to the least-squares regression (*Narula and Wellington*, 1982; *Wilson*, 1978). The asymptotic theory and related inferential methods have been well explored for quantile regression. We refer the readers to *Koenker* (2005) and *Koenker et al.* (2017) for a comprehensive discussion on quantile regression.

In this chapter, we consider a pseudo-Bayesian framework for quantile regression, where the quantile level of interest is fixed at a pre-specified value. Following *Yu and*

Moyeed (2001), we adopt the asymmetric Laplace working likelihood, which permits efficient posterior computations with MCMC algorithms (*Tsionas*, 2003; *Kozumi and Kobayashi*, 2011). However, direct posterior inference is invalid since the asymmetric Laplace working likelihood is generally mis-specified (*Sriram*, 2015; *Yang et al.*, 2016). Since the quantile regression model allows a broad form of heteroscedasticity, there is little reason to believe our working likelihood is close to the true one. Therefore, posterior inference is not justified by Bayes' Theorem, not even in the asymptotic sense (*Kleijn and Van der Vaart*, 2012). Specifically, the posterior credible intervals do not provide valid coverage probabilities, either in the frequentist or Bayesian sense.

Despite the likelihood mis-specification, the pseudo-Bayesian method offers a valuable computational tool for frequentist inference. *Yang et al.* (2016) and *Sriram* (2015) recognize that we can provide valid frequentist inference after a sandwich-form adjustment on the posterior variance. Using the adjusted posterior variance, the Wald-type interval can have valid frequentist coverage asymptotically. This idea of adjusted-posterior inference dates back to *Chernozhukov and Hong* (2003). Because the sampling distributions of the quantile regression estimators involve the conditional density functions as nonparametric nuisance parameters, inferential methods have to approximate those quantities directly or indirectly; see (*Koenker*, 2005, Section 3). The Bayesian computational approach trades optimization and nuisance parameter estimation for posterior sampling, and therefore provides a convenient framework for inference. The posterior-based method further stands out in more complex settings, such as censored regression (*Powell*, 1986) or missing covariates (*Sherwood et al.*, 2013), where the computational burden worsens for frequentist inferential procedures; See also *Yang et al.* (2016).

In this chapter, we extend the pseudo-Bayesian framework by considering shrinkage priors under a possibly sparse quantile regression model. In the big data era, datasets with a large amount of variables are becoming more and more common.

Given a large number of potential covariates, it is not unreasonable to believe that only a small portion of them affects the conditional quantile function. Under this sparsity regime, it is recognized that a shrinkage approach greatly improves estimation accuracy and statistical efficiency (*Tibshirani*, 1996). In this chapter, we use shrinkage priors in the Bayesian computational framework to capture possible sparsity in the model. Our goal is to provide valid, and more importantly, efficient inference for the quantile regression coefficients under sparsity.

In the Bayesian literature, using shrinkage priors is empirically shown to give improved performance when the model is sparse. Some examples of the priors include the Bayesian Lasso (*Park and Casella*, 2008), the horseshoe (*Carvalho et al.*, 2010), and the Dirichlet-Laplace prior (*Bhattacharya et al.*, 2015). There are also computational developments that adapt the shrinkage priors to quantile regression settings (*Li et al.*, 2010; *Alhamzawi et al.*, 2012; *Adlouni et al.*, 2018; *Kohns and Szendrei*, 2020). From a theoretical perspective, however, most results in the literature focus on the Gaussian mean regression and related settings, where the likelihood specification is approximately correct (*Bai and Ghosh*, 2021; *Gao et al.*, 2020; *Zhang et al.*, 2022); see also *Bhadra et al.* (2019) for a recent review. In the context of quantile regression, there is so far no theoretical understanding of how shrinkage priors can be used for valid and efficient inference.

In this chapter, we bring together the strength of the pseudo-Bayesian framework and shrinkage priors. We first establish two contributions when the covariate-dimension is fixed. On the theoretical side, we provide an asymptotic characterization of the posterior distribution. With a suitable prior, we show the posterior is consistent at the root-n rate regardless of the likelihood mis-specification, and we show the posterior is adaptive to model sparsity. Asymptotically, the posterior factors into two independent components: One for the non-zero (active) coefficients that achieves oracle efficiency as if we knew the true model; The other component for the inactive

components will concentrate toward 0 at a second-order rate. Based on these theoretical results, we present a unified approach for adaptive posterior inference in quantile regression. With an appropriate adjustment of the posterior variance, we can construct automatically adaptive confidence intervals in the frequentist sense: For the active coefficients, the interval achieves oracle efficiency; For the inactive coefficients, the interval is super-efficient and centers at 0. The confidence interval is adaptive in the sense that it automatically distinguishes active and inactive components without an additional variable selection step.

Then we extend our theoretical results to an increasing dimensional regime. That is, the covariate-dimension can grow with, but not exceed, the sample size. We find that the adaptivity result of the posterior distribution still applies, provided that the dimension grows at a controlled rate. The regime with increasing dimension is relevant when we approximate a non-parametric conditional quantile function by series expansion, e.g., splines, wavelets or local-polynomials. The number of effective regressors is typically chosen to increase with the sample size at a certain rate (*He and Shi*, 1994; *Belloni et al.*, 2019a). In empirical studies, it is also common to incorporate a large number of dummy variables, with possible interactions among them. In econometrics, this is referred to as the 'many regressors' regime (*Cattaneo et al.*, 2018).

Our setting of possibly sparse quantile regression is different from the high-dimensional regime where the regression coefficients are not identifiable without stringent sparsity constraints (*Belloni and Chernozhukov*, 2011; *Belloni et al.*, 2019b). Direct estimation and inference are feasible in our setup with increasing dimensions, though it may be inefficient if the true model is sparse. In another related regime with increasing dimension, *Belloni et al.* (2019a); *Pan and Zhou* (2021) considers bootstrap inference for the quantile regression. However, these bootstrap methods cannot incorporate the model sparsity, and specialized bootstrap procedures are needed for penalized

quantile regression; see *Wang et al.* (2018) for such an approach when the covariate dimensions are fixed.

Key to the pseudo-Bayesian framework is the choice of likelihood and prior. Since the quantile regression model does not assume any parametric likelihood function, it is common to rely on a working likelihood to pursue pseudo-Bayesian inference. Examples of other working likelihoods include the empirical likelihood (*Yang and He*, 2012; *Xi et al.*, 2016), the score likelihood (*Wu and Narisetty*, 2021), or the approximate likelihood (*Feng et al.*, 2015). The use of different shrinkage priors is also prevalent in practice for more efficient estimation (*Li et al.*, 2010; *Chen et al.*, 2013; *Adlouni et al.*, 2018; *Kohns and Szendrei*, 2020). This paper adopts the asymmetric Laplace working likelihood and focuses on two easy-to-understand examples of continuous shrinkage priors for their interpretability and computational attractiveness.

We hasten to add that our focus is not variable selection consistency or estimation accuracy, but the understanding of what can be accomplished with inference in the pseudo-Bayesian framework. In the recent literature, many have discussed the Bayesian variable selection performances under slab-and-spike type priors (*Ishwaran and Rao*, 2005; *Narisetty and He*, 2014; *Ročková and George*, 2018), and the posterior contraction rates under continuous shrinkage priors (*Song and Liang*, 2017; *Jiang and Sun*, 2019; *Gao et al.*, 2020). However, the literature is relatively sparse for adaptive pseudo-Bayesian inference in quantile regression, where the likelihood is mis-specified.

The rest of this chapter is organized as follows. In Section 5.2, we discuss the quantile regression problem and our pseudo-Bayesian framework. In Section 5.3, we present our main theoretical results under a fixed dimension. We discuss the posterior inference procedure in Section 5.4. The extension to increasing dimensions is considered in Section 5.5. We provide the computational details of posterior sampling in Section 5.6, followed by simulation studies in Section 5.7. We conclude in Section 5.8. All proofs are relegated to Section 5.9.

5.2 Modeling framework

5.2.1 Quantile regression model and working likelihood

We consider the linear quantile regression model. Let $Q_\tau(Y \mid X = \boldsymbol{x})$ be the τ-th conditional quantile of a continuous response Y given covariates $X = \boldsymbol{x} \in \mathbb{R}^p$, which includes an intercept term. Here the dimension p is fixed; later in Section 5.5, we extend our discussion to allow p to increase with the sample size. For a pre-specified quantile level $\tau \in (0, 1)$, we consider the model

$$Q_\tau(Y \mid X = \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}^0(\tau). \tag{5.1}$$

In the rest of the paper, we shall suppress the index τ whenever there is no confusion. Furthermore, we consider the case where the model (5.1) is possibly sparse, i.e.,

$$S = \{j \in \{1, \ldots, p\} : \ \beta_j^0 \neq 0\}, \quad |S| = s \leq p,$$

for some integer $s \geq 0$. This possible sparsity implies that some of the covariates may be irrelevant for modeling the τ-th conditional quantile of Y.

Let $\mathbb{D}_n = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ be a random sample of size n from (X, Y) that satisfies model (5.1). The classical quantile regression estimator (*Koenker and Bassett Jr*, 1978) minimizes the following check-loss function

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{u}} \ L_n(\boldsymbol{u}) = \arg\min_{\boldsymbol{u}} \ \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{u}), \tag{5.2}$$

where $\rho_\tau(\cdot)$ is the check function $\rho_\tau(u) = u[\tau - I(u < 0)]$. It is well known that $L_n(\boldsymbol{u})$ is convex, and (5.2) is equivalent to a linear programming problem. Under mild conditions, $\widehat{\boldsymbol{\beta}}$ consistently estimates the true quantile regression coefficients in (5.1). See *Koenker* (2005) and *Koenker et al.* (2017) for more discussion on quantile

regression.

In the pseudo-Bayesian framework of quantile regression, we consider the asymmetric Laplace working likelihood popularized by *Yu and Moyeed* (2001):

$$\mathcal{L}(\mathbb{D}_n \mid \boldsymbol{\beta}) = \frac{\tau^n(1-\tau)^n}{\sigma^n} \exp\left\{ \frac{-\sum_{i=1}^n \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})}{\sigma} \right\}, \tag{5.3}$$

where σ is a fixed scale parameter. The asymmetric Laplace likelihood (5.3) exponentiates the check-loss objective function L_n in (5.2); equivalently, (5.3) coincides with the likelihood function of an asymmetric Laplace distribution. We call (5.3) a working likelihood because it does not correspond to the true data-generating mechanism of $\mathbb{D}_n$ given $\boldsymbol{\beta}$. The model (5.1) does not impose any distributional assumption of the data, except for the τ-th conditional quantile. Therefore, the true likelihood may well be different from (5.3). Nonetheless, the maximum likelihood estimator under (5.3) coincides with the classic quantile regression estimator in (5.2). Furthermore, we consider the scale parameter σ to be fixed at 1 throughout this chapter. Alternatively, *Choi and Hobert* (2013) and *Yu and Moyeed* (2001) consider a full Bayesian approach with a prior on σ.

5.2.2 Penalization and shrinkage priors

Since the model (5.1) is possible sparse, we use shrinkage priors in our pseudo-Bayesian framework. In this chapter, we focus on two choices that are motivated by the frequentist penalized regression procedures. When there are a relatively large number of covariates, it is common to consider the following penalized quantile regression problem, which can improve efficiency and interpretability (*Tibshirani*, 1996).

$$\min_{\boldsymbol{u}} \sum_{i=1}^n \rho_\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{u}) + Q_\lambda(\boldsymbol{u}),$$

for some non-negative penalty function $Q_\lambda(\boldsymbol{u})$. Among others, the Smoothly Clipped Absolute Deviation (SCAD) in *Fan and Li* (2001) and the adaptive Lasso in *Zou* (2006) are known to enjoy 'oracle' properties for quantile regression *Wu and Liu* (2009).

In this chapter, we consider priors that are motivated from the SCAD and adaptive Lasso penalty functions as below:

$$\pi_{AL}(\boldsymbol{\beta}) \propto \exp\left\{-\sqrt{n}\sum_{j=1}^{p}\frac{\lambda}{w_j}|\beta_j|\right\}, \tag{5.4}$$

$$\pi_{CA}(\boldsymbol{\beta}) \propto \exp\left\{-n\sum_{j=1}^{p}p_\lambda(\beta_j)\right\}, \tag{5.5}$$

where the choice of w_j and the function $p_\lambda(\cdot)$ will be given shortly. The prior (5.4) corresponds to the Adaptive Lasso (AL) penalty (*Zou*, 2006), where $w_j = |\hat{\beta}_j|$ for $j \in \{1, \ldots, p\}$ as in *Wu and Liu* (2009) and $\hat{\beta}_j$ is the j-th component of $\widehat{\boldsymbol{\beta}}$; Similar prior has been studied in *Alhamzawi et al.* (2012) and *Li et al.* (2010) via a full Bayesian approach. In the Clipped Absolute (CA) prior (5.5) we define $p_\lambda(u) = \lambda(|u| \wedge \lambda)$, which is motivated from the Smoothly Clipped Absolute Deviation (SCAD) penalty of *Fan and Li* (2001). However, we remove the smoothing component to simplify the theoretical derivation; See Figure 5.1 for a visual comparison. Our choices of priors are relatively simple examples from the broader class of shrinkage priors *Griffin and Brown* (2010); *Carvalho et al.* (2010); *Bhattacharya et al.* (2015); *Zhang et al.* (2022). We stick to the choices (5.4) and (5.5) to fix ideas and simplify the technical derivations.

Given the choice of a prior $\pi(\boldsymbol{\beta})$, the working posterior density is then

$$p(\boldsymbol{\beta} \mid D_n) \propto \mathcal{L}(\mathbb{D}_n \mid \boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta}). \tag{5.6}$$

Here $\propto$ means equal up to some constants that does not depend on β (but could

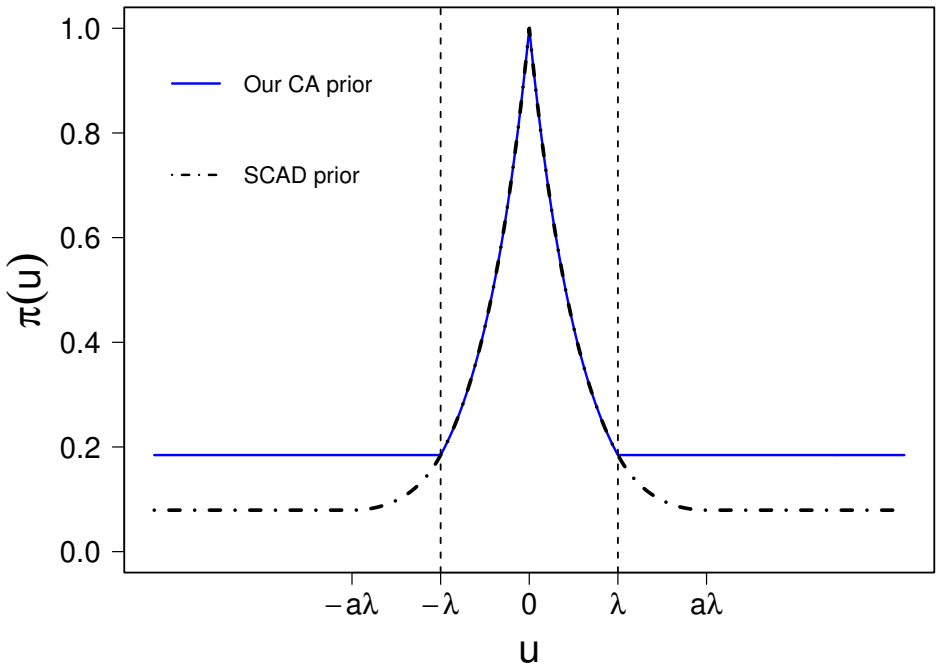


Figure 5.1: Comparison between the prior $\pi_{CA}(u)$ and the prior induced by the SCAD penalty in *Fan and Li* (2001); a is a tuning parameter in the SCAD penalty and we set $a = 2$ in the plot. Both priors are flat when $|u| > a\lambda$.

depend on the data $\mathbb{D}_n$). Note each of the prior functions (5.4) and (5.5) is uniformly upper bounded by 1. Thus they give rise to a proper posterior, in the sense that the above posterior leads to a valid probability distribution on $\boldsymbol{\beta}$ (*Yu and Moyeed*, 2001; *Tsionas*, 2003). Thus, various MCMC techniques readily applies to the model (5.6).

Below we provide some additional comments to our prior choices. Both classes of priors involve an additional scalar parameter λ. We regard this $\lambda = \lambda_n$ as a tuning parameter that depends on n, though suppressing any subscript to simplify notation. This dependency is the essence of the family of 'shrinkage priors', which aims to shrink the irrelevant coefficients to 0 in the posterior. See for example *Armagan et al.* (2013a) and *Song and Liang* (2017). For any fixed prior that does not depend on n, the impact of the prior will eventually get washed away as the sample size grows (*Van der Vaart*, 2000; *Bontemps*, 2011). In our setting, as λ grows with n, the priors (5.4) and (5.5) will show a sharper peak at the origin, flatter tail, and places

more mass in the neighbourhood of 0. these properties make our priors *adaptive* in a possibly sparse model, which are in line with other commonly used priors in the Bayesian framework (*Song and Liang*, 2017).

5.3 Asymptotic properties of the posterior distribution

In this section, we present the main theoretical results under either choices of the shrinkage prior (5.5) or (5.4). Specifically, we study the large-sample properties of the posterior distribution under the repeated sampling perspective.

5.3.1 Regularity conditions

We fix some notations first. Let $\mathbb{D}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ be *i.i.d.* samples from $(X, Y) \sim \mathrm{P}^*$ that satisfies model (5.1), where each $\boldsymbol{x}_i \in \mathbb{R}^p$ and each y_i is a scalar for $i = 1, \ldots, n$. Note P^* does not necessarily satisfy the asymmetric Laplace working likelihood (5.3). We denote convergence in P^*-probability by $\xrightarrow{P^*}$; we denote the expectation under P^* by $\mathrm{E}^*(\cdot)$. Given the data $\mathbb{D}_n$, let $p(\boldsymbol{\beta} \mid \mathbb{D}_n)$ be the working posterior density for $\boldsymbol{\beta}$ as in (5.6); we define the corresponding posterior probability measure as

$$\Pi\left(\mathcal{A} \mid \mathbb{D}_n\right) = \int_{\mathcal{A}} p(\boldsymbol{\beta} \mid \mathbb{D}_n)\, \mathrm{d}\boldsymbol{\beta},$$

for any measurable set $\mathcal{A} \subset \mathbb{R}^p$. Note both $p(\boldsymbol{\beta} \mid \mathbb{D}_n)$ and $\Pi(\mathcal{A} \mid \mathbb{D}_n)$ are random variables under P^*-probability.

Let $\boldsymbol{\beta}^0 = (\beta_1, \ldots, \beta_p)^T$ be the true values of regression coefficients in model (5.1); let $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ be the classic quantile regression estimator from (5.2). For a vector $\boldsymbol{v}$, let $\|\boldsymbol{v}\|$ be its L_2 norm and $\|\boldsymbol{v}\|_\infty$ be its L_∞ norm. For probability density functions $f(x)$ and $g(x)$, we denote their total variation distance by $\|f - g\|_{TV} = \int |f - g| \mathrm{d}x$. For non-stochastic sequences a_n and b_n, we write $a_n \ll b_n$ if $a_n / b_n = o(1)$; we write $a_n \lesssim b_n$ if there is a universal constant C_0 such that $a_n \leq C_0 \cdot b_n$. For

stochastic sequences A_n and B_n, we define $A_n \ll_{\mathrm{P}^*} B_n$ (or $A_n \lesssim_{\mathrm{P}^*} B_n$) if $A_n \ll B_n$ (or $A_n \lesssim B_n$) holds with P*-probability tending to unity. Recall $S = \{1, \ldots, s\}$ is the index set for active covariates, for any vector $\boldsymbol{v}$ we write $\boldsymbol{v}^T = (\boldsymbol{v}_1^T, \boldsymbol{v}_2^T)$ with $\boldsymbol{v}_1 \in \mathbb{R}^s$. For any matrix $A \in \mathbb{R}^{p \times p}$, we partition

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where $A_{11} \in \mathbb{R}^{s \times s}$; for $i, j \in \{0, \ldots, p\}$, we shall write $A(i, j)$ as the (i, j)th entry of A.

Now we introduce some technical assumptions, most of which are standard in quantile regression (*Koenker*, 2005; *Belloni and Chernozhukov*, 2011) and variable selection literature (*Wu and Liu*, 2009; *Zhao and Yu*, 2006).

Assumption E.1 (Identification)**.** For any $\delta > 0$, there exists $\varepsilon > 0$, such that

$$\limsup_{n \to \infty} \ \mathrm{P}^* \left\{ \sup_{\|\beta - \beta_0\| \geq \delta} \frac{1}{n} (L_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}^0)) \geq \varepsilon) \right\} = 1.$$

Assumption E.2 (Smooth and bounded conditional densities)**.** (i) The conditional distribution of Y given X has a density function $f_{Y|X=\boldsymbol{x}}(u)$. (ii) There exist $0 < \underline{f} < \overline{f}$, such that

$$\underline{f} \leq \inf_{\boldsymbol{x}} \left[f_{Y|X=\boldsymbol{x}}(\boldsymbol{x}^T \boldsymbol{\beta}^0) \right] \leq \sup_{\substack{u \in \mathbb{R} \\ \boldsymbol{x}}} \left[f_{Y|X=\boldsymbol{x}}(u) \right] \leq \overline{f}.$$

(iii) $f_{Y|X=\boldsymbol{x}}(u)$ is uniformly (over $\boldsymbol{x}$) Lipschitz continous (in u).

Assumption E.3 (Eigenvalue condition)**.** The matrix $D = \mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right]$ is positive definite; its eigenvalues are all bounded away from 0 and $+\infty$.

Assumption E.4 (Bounded covariates)**.** The covariates X has bounded support.

Assumption E.5 (Sparsity)**.** For a constant $b_0 > 0$, the true regression coefficients

satisfy:

$$\min_{j=1,\ldots,s} |\beta_j^0| > b_0,$$
$$\beta_{s+1}^0 = \ldots = \beta_p^0 = 0.$$

Here we briefly discuss the assumptions. Assumptions E.1–E.4 are standard in pseudo-Bayesian modeling with a working likelihood (*Chernozhukov and Hong*, 2003; *Yang et al.*, 2016) and the quantile regression literature (*Knight*, 1998; *Pan and Zhou*, 2021); see also *Koenker* (2005, Section 4). In particular, Assumption E.4 is to simplify the technical treatment in our theoretical development. When the dimension is fixed, it implies the boundedness of $\|X_i\|$. Assumption E.5 requires the non-zero coefficients to be well separated from others. This so called beta-min condition is necessary for a consistent model selection in either frequentist or Bayesian literature (*Wu and Liu*, 2009; *Belloni and Chernozhukov*, 2011; *Castillo et al.*, 2015). Although our target is not variable selection, these conditions are necessary for adaptive inference.

5.3.2 Main results

Before bringing in the shrinkage prior, we first present a Proposition regarding the rate of consistency for the posterior (5.6) under a flat prior and without concerning the model sparsity. Here flat means a improper uniform prior on the entire space $\mathbb{R}^p$, which induces a posterior proportional to the likelihood (5.3) itself. Nonetheless, the asymptotic result here also applies to any fixed prior that does not depend on the sample size. The Proposition below is not only a useful lemma for the remaining asymptotic results, but also of independent interest itself.

Proposition 6. *Given Assumptions E.1 - E.4 hold, and consider the flat prior* $\pi(\beta) \propto$

1. *The posterior is consistent at a $\sqrt{n}$-rate, that is,*

$$\Pi(\sqrt{n} \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \geq M_n \mid \mathbb{D}_n) \xrightarrow{P^*} 0,$$

for any sequence $M_n \to +\infty$.

The above result gives a $\sqrt{n}$-rate of posterior contraction, which is the Bayesian counterpart to the frequentist result of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 = O_p(n^{-1/2})$. This is intuitive as the posterior mode coincides exactly with the classic quantile regression estimator $\widehat{\boldsymbol{\beta}}$. Note the sequence M_n can tend to infinity arbitrarily slow, yet it can not be replaced by any fixed constant M.

Such $\sqrt{n}$-rate is necessary for the adjusted posterior inference in quantile regression. For example, the validity of inference scheme in *Yang et al.* (2016) and *Sriram* (2015) depends on a $\sqrt{n}$-rate. However, this result has not been thoroughly studied yet. It is hinted in Theorem 1 of *Chernozhukov and Hong* (2003), yet no rigorous derivation is present. Our Proposition 6 provides a general treatment specifically for quantile regression. The results in *Kleijn and Van der Vaart* (2012) does not apply with the absence of distributional assumption in Model (5.1). Specific for quantile regression, *Sriram et al.* (2013) claims the same $\sqrt{n}$ consistency result as ours, but a later correction by *Sriram and Ramamoorthi* (2017) voids their contribution. Recently, a manuscript of *Sriram and Ramamoorthi* (2018, Theorem 2) gives the same conclusion. Their result is only valid for a proper prior, whereas ours is valid for the improper flat prior. In terms of technical treatment, they rely on the piece-wise nature of the check function $\rho_\tau(\cdot)$, while we provide a more general treatment via empirical process theory.

Now we are ready to introduce shrinking priors to incorporate model sparsity. In particular, we present consistency result similar to Proposition 6 and a Bernstein-von-Mises (BvM) type theorem about the distributional approximation of the posterior,

under both priors (5.4) or (5.5). In smooth parametric models, the usual BvM theorem (*Van der Vaart*, 2000; *Kleijn and Van der Vaart*, 2012) asserts that the posterior can be approximated by a normal distribution in total variation distance. That is

$$\|p(\cdot \mid \mathbb{D}_n) - \phi(\cdot)\|_{TV} \xrightarrow{P^*} 0,$$

where $\phi(\cdot)$ is the density function for the limiting Gaussian distribution. For Bayesian quantile regression with no sparsity, *Sriram and Ramamoorthi* (2017) concludes that the posterior is indeed asymptotically normal, despite that the working likelihood is mis-specified. Here we show that shrinkage priors can make the posterior adaptive to model sparsity, in correspondence with the frequentist penalized quantile regression (*Wu and Liu*, 2009; *Li and Zhu*, 2008). We first consider the CA shrinkage prior (5.5). We define

$$G = \mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T \cdot f_{Y|X=\boldsymbol{x}_i}(\boldsymbol{x}_i^T \boldsymbol{\beta}^0)\right].$$

Further, we denote the upper-left s-by-s sub-matrix of G by G_{11}.

Theorem V.1 (CA shrinkage)**.** *Consider the improper CA prior (5.5). Suppose Assumptions E.1 through E.5 hold, and the tuning parameter λ satisfies $1/\sqrt{n} \ll \lambda \ll 1$. We have the following:*

1. *Posterior consistency:*

$$\Pi(\sqrt{n} \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \geq M_n \mid \mathbb{D}_n) \xrightarrow{P^*} 0,$$

   for any sequence $M_n \to +\infty$.

2. *Distributional approximation:*

$$\left\| p(\boldsymbol{\beta} \mid \mathbb{D}_n) - \phi\left(\boldsymbol{\beta}_1; \tilde{\boldsymbol{\beta}}_1, \frac{G_{11}^{-1}}{n}\right) \otimes \prod_{j=s+1}^{p} \frac{n\lambda}{2} \exp\{-n\lambda|\beta_j|\} \right\|_{TV} \xrightarrow{P^*} 0.$$

Here $\phi(\cdot; \mu, \Sigma)$ represents a Gaussian density function with mean vector μ and covariance matrix Σ, and $\tilde{\beta}_1$ is the oracle QR estimator using the true model.

Part 1 of the above Theorem states that the rate of contraction is the same as in the un-penalized case. By choosing a proper value of λ, the CA shrinkage prior does not cast any bias or performance loss on the first order. Further, part 2 of Theorem V.1 shows that there is actually some efficiency gain. The limiting distribution breaks down into two independent components, corresponding to the active coefficients $\boldsymbol{\beta}_1$ and those inactive ones $\boldsymbol{\beta}_2$, respectively. For non-zero coefficients $\boldsymbol{\beta}_1$, the posterior achieves oracle efficiency. In particular, the limiting distribution is the same as if we know the true model (*Sriram*, 2015). On the other hand, the posterior for inactive coefficients $\boldsymbol{\beta}_2$ is highly concentrated toward 0 at a faster-than-first-order rate, since $1/(n\lambda) \ll 1/\sqrt{n}$. Consequently, the posterior is automatically adaptive to the model sparsity, and the efficiency for both components $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be improved under sparsity.

Theorem V.1 shares the same spirit of Theorem 2.1 in *Song and Liang* (2017), where the name 'near-oracle property' is adopted. The main difference between them is the model mis-specification in our setting, and that the working likelihood is non-smooth, therefore our results require more delicate technical treatment. On the contrary, Theorem V.1 is in sharp contrast with those under discrete spike-and-slab priors (*Ishwaran and Rao*, 2005; *Castillo et al.*, 2015), where the posterior gives sparse solution automatically. Our shrinkage priors are not tailored for variable selection, therefore the posterior distribution is continuous. As we illustrate later, such smoothness is beneficial for more stable posterior inference.

Remark 10. The posterior variance for the active component, G_{11}^{-1}, does not match the sampling variance of the (oracle) QR estimator, as recognized by *Yang et al.* (2016). This is a consequence of using the mis-specified asymmetric Laplace likelihood (5.3). Nonetheless, Theorem V.1 leads to an convenient tool for constructing the Wald

interval, particularly so in quantile regression. As the MCMC samples from the posterior provide an efficient estimation of G_{11}^{-1} (*Chernozhukov and Hong*, 2003), which is an essential piece for quantile regression inference. In the next section we develop an easy-to-implement inferential procedure that targets sparse model.

Remark 11. The convergence in total variation in Theorem V.1 does not immediately imply convergence of moment. Later in Proposition 7, we show that the posterior moments converges toward the moments in the limiting distribution.

Next we present the result for Adaptive Lasso shrinkage. The results are similar to Theorem V.1. However the different natures of the two kinds of priors requires different treatment in the proof.

Theorem V.2 (Adaptive Lasso shrinkage)**.** *Consider the Adaptive Lasso prior (5.4). Suppose assumptions E.1 through E.5 hold and the tuning parameter satisfies* $1/\sqrt{n} \ll \lambda \ll 1$. *Then we have*

1. *Posterior consistency:*

$$\Pi(\sqrt{n} \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \geq M_n \mid \mathbb{D}_n) \xrightarrow{P^*} 0,$$

   for any sequence $M_n \to +\infty$.

2. *Distributional approximation:*

$$\left\| p(\boldsymbol{\beta} \mid \mathbb{D}_n) - \phi\left(\boldsymbol{\beta}_1; \tilde{\boldsymbol{\beta}}_1, \frac{G_{11}^{-1}}{n}\right) \otimes \prod_{j=s+1}^{p} \frac{\sqrt{n}\lambda}{2w_j} \exp\left\{-\frac{\sqrt{n}\lambda}{w_j}|\beta_j|\right\} \right\|_{TV} \xrightarrow{P^*} 0.$$

   Here $\phi(\cdot; \mu, \Sigma)$ *represents a Gaussian density function with mean vector* μ *and covariance matrix* Σ, $\tilde{\beta}_1$ *is the oracle QR estimator under the true model, and* w_j *is defined in (5.4).*

Theorem V.2 is in the same spirit as Theorem V.1, though the limiting distribution is in slightly different forms. For the inactive coefficients $\boldsymbol{\beta}_2$, the posterior share the same order at $O_{p^*}(n\lambda)$ since $w_j = |\hat{\beta}_j| = O_{p^*}(1/\sqrt{n})$ for $j = s+1, \ldots, p$. Such an order matches with Theorem V.1. In fact, the posterior limit is the same as the Adaptive Lasso prior (5.4). This coincidence is in line with *Song and Liang* (2017), in the sense that the posterior for the inactive coefficients is asymptotically driven by the prior.

We further provide an heuristic explanation of the technical results in Theorem V.1 and V.2, through which we hope to shed some light on the subtle difference between the two kinds of priors. The CA prior is the same for each coordinate of $\boldsymbol{\beta}$, and its adaptivity is driven by its flat tail and sharp peak. For the zero coefficients, the prior dominates the likelihood using the sharp peak around zero; while for the non-zero coefficients the prior is washed out by the likelihood. On the other hand, the Adaptive Lasso prior achieves the adaptation via the choice of weights $w_j = |\widehat{\beta}_j|$: For those active coefficients, the scale for prior shrinkage in (5.4) is of order $O_{\mathrm{P}^*}(\sqrt{n}\lambda) = o_{\mathrm{P}^*}(\sqrt{n})$. For an inactive coefficient it would be in an order of $O_{\mathrm{P}^*}(n\lambda) \gg O_{\mathrm{P}^*}(\sqrt{n})$. Since the likelihood itself has a $\sqrt{n}$-scale by Proposition 6, it can dominate the prior for the active coefficients, yet not the the inactive ones. From the above reasoning, it should also be clear that any $\sqrt{n}$-consistent estimator can be used in w_j as in (*Zou*, 2006).

Remark 12. We emphasize that the adaptivity of the posterior shrinkage in Theorems V.1 and V.2 is not shared under all popular Bayesian priors. For example, *Castillo et al.* (2015) shows that the traditional Bayesian-lasso (*Park and Casella*, 2008) can not achieve the adaptation in the Gaussian mean regression setting, in the sense that the posterior either over-shrinks the active coefficients or under-shrinks the inactive coefficients.

5.4 Adaptive posterior inference

In this section we develop an asymptotically valid inferential procedure based on the posterior moments. It is known since *Yang et al.* (2016) that the posterior variance matches the sampling variance of $\widehat{\boldsymbol{\beta}}$ after a feasible correction. We extend their approach to target possibly sparse models, which we show to be automatically adaptive to sparsity. It is important that our inferential procedures are valid in the frequentist sense. Though adopting a pseudo-Bayesian framework, the posterior only serves as an computational tool to construct the confidence intervals. See *Chernozhukov and Hong* (2003) for a in-depth discussion of this idea.

5.4.1 Inferential procedure using posterior moments

We construct interval estimates that achieves frequentist validity based on the posterior moments. We focus on the CA prior (5.5) for simplicity. We introduce some additional notations. Denote by $\check{\boldsymbol{\beta}}$ the (finite sample) posterior mean and $\check{\Sigma}$ the posterior variance-covariance matrix in (5.6). Let $\widehat{D} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T/n$ and $D = \mathrm{E}[XX^T]$. In Theorem V.1, the limiting posterior has mean $(\widetilde{\boldsymbol{\beta}}_1, \mathbf{0})$ and variance Σ. Recall $\widetilde{\boldsymbol{\beta}}_1$ is the oracle QR estimator, and

$$\Sigma = \begin{bmatrix} \frac{1}{n}\boldsymbol{G}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2}{n^2\lambda^2}\boldsymbol{I}_{p-s} \end{bmatrix}, \tag{5.7}$$

where $\boldsymbol{I}_{p-s}$ is the identity matrix of dimension $p-s$.

In accordance with *Yang et al.* (2016), we define the adjusted variance-covariance matrix

$$\check{\Sigma}_{adj} = n\tau(1-\tau)\check{\Sigma}\,\widehat{D}\,\check{\Sigma}, \tag{5.8}$$

using all p covariates. We propose an weighted Wald-interval using $\check{\boldsymbol{\beta}}$ as the center, and $\check{\Sigma}_{adj}$ as the standard error. In particular, for any one-dimensional component β_j,

we construct the $1-\alpha$ confidence interval as

$$\check{\beta}_j \pm z_{\alpha/2} \cdot \eta_j \cdot \sqrt{\check{\Sigma}_{adj}(j,j)}, \tag{5.9}$$

with weights $\eta_j = \max\{1, \lambda/|\hat{\beta}_j|\}$, and $z_{\alpha/2}$ Above we have used the upper $\alpha/2$ quantile for the standard normal distribution, and the j-th diagonal element of the adjusted variance-covariance matrix. Note that $\hat{\beta}_j$ in the weight is the un-penalized quantile regression estimator.

To study the property of our interval estimate in (5.9), we need the following result regarding the convergence of posterior moments. The key is the adaptive rate of convergence: For active component, the convergence occur at $\sqrt{n}$-rate, while for inactive components, the moments converge at a faster $n\lambda$-rate. Such an adaptive rate is necessary for technical derivations hereafter.

Proposition 7 (Posterior moments)**.** *Under the conditions of Theorem V.1, we have*

$$\begin{bmatrix} \sqrt{n}(\check{\boldsymbol{\beta}}_1 - \widetilde{\boldsymbol{\beta}}_1) \\ n\lambda(\check{\boldsymbol{\beta}}_2 - \mathbf{0}) \end{bmatrix} \xrightarrow{P^*} 0$$

$$\begin{bmatrix} n\check{\Sigma}_{11} & n^{1.5}\lambda\check{\Sigma}_{12} \\ n^{1.5}\lambda\check{\Sigma}_{21} & n^2\lambda^2\check{\Sigma}_{22} \end{bmatrix} - \begin{bmatrix} \boldsymbol{G}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & 2\boldsymbol{I}_{p-s} \end{bmatrix} \xrightarrow{P^*} 0$$

Based on Proposition 7, we now show the interval estimate (5.9) is valid in the frequentist sense, and adaptive to sparsity. We first consider an active component $j \in S$, and we show that the interval achieves oracle efficiency. Proposition 7 implies for an active component $j \in S$, the posterior mean $\check{\beta}_j$ enjoys the same first-order asymptotic behavior with the oracle QR estimator $\tilde{\boldsymbol{\beta}}_1$. The results in *Wu and Liu* (2009) then applies to the posterior mean $\check{\boldsymbol{\beta}}_1$,

$$\sqrt{n}\left(\check{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0\right) \xrightarrow{d} N\left(\mathbf{0}, \tau(1-\tau)G_{11}^{-1}D_{11}G_{11}^{-1}\right).$$

For the standard error in (5.9), Proposition 7 implies that the posterior variance $\check{\Sigma}$ converges to Σ defined in (5.7). We can rewrite the adjusted variance-covariance matrix as

$$\begin{aligned} n \cdot \check{\Sigma}_{adj} &= n^2\tau(1-\tau)[\Sigma + o_{P^*}(1/n)] \cdot [D + o_{P^*}(1)] \cdot [\Sigma + o_{P^*}(1/n)] \\ &= \tau(1-\tau)(n\Sigma) \cdot D \cdot (n\Sigma) + o_{P^*}(1) \\ &= \begin{bmatrix} \tau(1-\tau)G_{11}^{-1}D_{11}G_{11}^{-1} & \frac{2\tau(1-\tau)}{n\lambda^2}G_{11}^{-1}D_{12} \\ \frac{2\tau(1-\tau)}{n\lambda^2}D_{21}G_{11}^{-1} & \frac{4\tau(1-\tau)}{n^3\lambda^4}D_{22} \end{bmatrix} + o_{P^*}(1) \\ &= \begin{bmatrix} \tau(1-\tau)G_{11}^{-1}D_{11}G_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + o_{P^*}(1). \end{aligned} \tag{5.10}$$

The last identity is due to $1/\sqrt{n} \ll \lambda \ll 1$ in Theorem V.1. For an active $1 \le j \le s$, this implies our adjusted variance (5.8) consistently estimates the sampling variance of $\check{\beta}_j$. For $j \in S$ we have also the weight $\eta_j \xrightarrow{P^*} 1$, concluding (5.9) attains $1-\alpha$ coverage asymptotically.

Next we consider the case when $j \notin S$. We show the interval is super-efficient, in the sense that the coverage achieves 100%, while the length vanishes faster than a rate of $1/\sqrt{n}$. Denote by $\check{\boldsymbol{\sigma}}_j$ the j-th row of $\check{\Sigma}$. Proposition 2 shows the posterior mean $\check{\boldsymbol{\beta}}_2 = o_{P^*}(1/n\lambda)$, and

$$\check{\boldsymbol{\sigma}}_j = \frac{\boldsymbol{e}_j}{n^2\lambda^2} + \left[\frac{\boldsymbol{r_1}}{n^{1.5}\lambda}\,,\ \frac{\boldsymbol{r}_2}{n^2\lambda^2}\right],$$

where $\boldsymbol{e}_j$ is the unit vector with 1 on its j-th entry and $\boldsymbol{r}_k = o_{P^*}(1)$. From (5.8), the adjusted variance for β_j becomes

$$\begin{aligned} \check{\Sigma}_{adj}(j,j) &\asymp n\tau(1-\tau)\check{\sigma}_j^T\check{\sigma}_j \\ &\asymp \frac{\|\boldsymbol{r}_1\|^2}{n^2\lambda^2} + \frac{\|\boldsymbol{r}_2 + \boldsymbol{e}_j\|^2}{n^3\lambda^4}, \end{aligned}$$

where $A \asymp B$ means that $A = O_{P^*}(B)$ and $B = O_{P^*}(A)$. Together with the weight $\eta_j = \lambda/|\hat{\beta}_j| \asymp \sqrt{n}\lambda$ for $j \notin S$, we have

$$\frac{\check{\beta}_j - 0}{z_{\alpha/2} \cdot \eta_j \cdot \sqrt{\check{\Sigma}_{adj}(j,j)}} \xrightarrow{P^*} 0,$$
$$\sqrt{n} \cdot \eta_j \cdot \sqrt{\check{\Sigma}_{adj}(j,j)} \xrightarrow{P^*} 0.$$

The first line shows the width of the interval (5.9) dominates the magnitude of the center, implying the interval will cover $\beta_j^0 = 0$ with probability attaining one; The second line shows the length of the interval is of order $o_{P^*}(1/\sqrt{n})$, achieving super-efficiency.

We summarize the behaviour of the interval estimate (5.9), denoted as $CI_j(\alpha)$. Combining the above arguments, we have:

$$\begin{aligned} \mathrm{P}^*\left(\beta_j^0 \in CI_j(\alpha)\right) &\to 1-\alpha, \quad \text{if } j \in S, \\ \mathrm{P}^*\left(\beta_j^0 \in CI_j(\alpha)\right) &\to 1, \quad \text{if } j \notin S. \end{aligned}$$

For $j \in S$, $CI_j(\alpha)$ achieves 'oracle efficiency' as in (*Wu and Liu*, 2009; *Zou*, 2006). For $j \notin S$, $CI_j(\alpha)$ is super-efficient, and its width is narrower than the usual $1/\sqrt{n}$ order. This reveals an important feature of our inferential procedure: The constructed confidence interval (5.9) can automatically distinguish the active and inactive coefficients. Given the MCMC samples from the posterior, a unified variance-adjustment step serves any coefficient β_j. Our discussion also applies to constructing confidence intervals for any linear combination of the coefficients.

Remark 13. The weight η_j in (5.9) is necessary to achieve desired coverage for inactive components. For $j \notin S$, denote by ℓ_n as the width of (5.9). Our weighting scheme induced by η_j inflates the standard error such that the width of the interval is at the correct order $1/(n\lambda) \lesssim \ell_n \ll 1/\sqrt{n}$. Such an order ensures that the interval is super-

efficient, yet has guaranteed coverage probability in theory. Without the weighting adjustment, the interval length would be too narrow to have valid coverage.

Remark 14. Under the same conditions, the SCAD quantile regression estimator enjoys the oracle property asymptotically (*Wu and Liu*, 2009). One may use the SCAD-penalized estimator as the center in our interval (5.9). Our interval based on the posterior mean, however, has the following two advantages. First, all pieces used to form (5.9) are readily available from MCMC computation. Second, numerical evidences in Section 5.7 show the inferential procedure is less sensitive to the choice of λ when using the posterior mean.

5.4.2 Comparison with the frequentist approach

Another common approach for inference in sparse models is the following two-step procedure: we first conduct variable selection, then apply frequentist inferential methods on the selected model. Here we compare this two-step procedure with our method. First, the two-step approach is valid only when the variable selection is correct. In finite samples, however, variable selection procedures often fail to give oracle selection 100% of the time (*Wang et al.*, 2020). As we show in the simulation studies of Section 5.7, our pseudo-Bayesian approach does not depend on the binary variable selection, and therefore is often more stable in practice.

Second, the classical Wald-type interval in quantile regression requires estimating G_{11}, which involves a weighted average of the conditional densities of Y given $X = x_i$. While many non-parametric approaches are available for estimating those density functions (*Koenker*, 2005), those methods rely critically on other additional assumption and/or the proper selection of a bandwidth parameter. E.g., *Powell* (1991) propose a kernel-based weighted density estimator; *Hendricks and Koenker* (1992) develop a method based on a differentiation formula, assuming the conditional quantile remains linear in x for a range of quantile level close to τ. In practice, the

performance of the resulting Wald interval are often unstable, even when the sample size is moderately large (*Yang et al.*, 2016, Section 4). Our posterior-based method provides a versatile inferential method that can approximate G_{11} using MCMC (*Chernozhukov and Hong*, 2003).

5.5 Theoretical investigation with increasing dimensions

In this section, we present some theoretical results when the dimension p_n diverges with, while is still of smaller order than, the sample size n; We also allow the size of the true model, s_n, to depend on the sample size. Sometimes we shall omit the index n if there is no confusion. We show similar asymptotic behaviours to that in Section 5.3 hold true under some additional conditions. We first investigate the asymptotic behaviour of the posterior distribution under the flat prior, where no model sparsity is assumed. Then, we incorporate shrinkage priors to deal with sparse models. For technical simplicity, we shall only focus on the CA prior (5.5) in this section.

We discuss some extensions of the regularity conditions in Section 5.3. When $p = p_n \to +\infty$, some of those conditions may not be practical anymore. We still need Assumptions E.1 and E.2, which are standard in the quantile regression literature (*Belloni et al.*, 2019a; *Pan and Zhou*, 2021). As for Assumption E.3 through E.5, we replace them by the following:

Assumption E.3′ (Relaxed eigenvalue condition)**.** The maximal/minimal eigenvalues of the p-by-p matrix D satisfy $p^{-1} \lesssim \theta_{\min}(D) \leq \theta_{\max}(D) \lesssim p$. Further, the minimal eigenvalue of the s-by-s matrix D_{11} satisfy $\theta_{\min}(D_{11}) \geq \theta_{11} > 0$ for some constant θ_{11}.

Assumption E.4′ (Regular covariates)**.** $\mathrm{E}^*[x_{ij}] = 0$ and $\mathrm{E}^*[x_{ij}^2] = 1$. Furthermore,

the standardized covariate $D^{-1/2}x_i$ is sub-exponential, i.e., for some constant $\sigma_1 > 0$,

$$\mathrm{P}^*\left(|u^T D^{-1/2} x_i| \geq \sigma_1 t\right) \leq 2\mathrm{e}^{-t}, \tag{5.11}$$

for all $\|u\| = 1$ and $t > 0$.

Assumption E.5′ (Sparsity)**.** We assume $\beta^0_{s+1} = \ldots = \beta^0_p = 0$, and there exists a sequence $\underline{b}_n > 0$ such that $\min_{j=1,\ldots,s}|\beta^0_j| > \underline{b}_n > 0$.

Here we discuss the above generalizations. Assumption E.3′ relaxes the bounded-eigenvalue condition in Assumption E.3 by allowing the matrix D to have vanishing/diverging eigenvalues. That is, our analysis can tolerate some degree of co-linearity among the covariates. Nonetheless, we assume that the eigenvalues of D_{11}, the active covariates, are bounded from below; so that the asymptotic results for quantile regression (*He and Shao*, 2000; *Belloni et al.*, 2019a) apply to the s-dimensional oracle estimator. In our setting, the same eigenvalue conditions in Assumption E.3′ also applies to the matrix G defined above Assumption E.3. For Assumption E.4′, it is implied by the original Assumption E.4 when the dimensions are fixed. In high-dimensions, however, Assumption E.4′ is stronger than the boundedness of x_{ij}. We refer the readers to the recent book *Vershynin* (2018, Section 3.3) for examples of sub-exponential distributions in high-dimensions. Finally, Assumption E.5′ requires all non-zero coefficients to be sufficiently-separated from zero by a margin of $\underline{b}_n$. So far we do not specify any requirement for s and $\underline{b}_n$; We discuss them in our main results below.

5.5.1 Posterior consistency with a dense model

We first consider the case without model sparsity, and we focus on the flat prior, i.e., $p(\boldsymbol{\beta}) \propto 1$. The following Proposition is a natural extension of Proposition 6 that allows a diverging number of predictors. The proof relies on two auxiliary Lemmas

IB.1 and IB.3 in the Appendix. The remaining proof is identical to that of Proposition 6 and *Sriram* (2015, Theorem 1), hence we omit it in this dissertation.

Proposition 8. *Suppose that Assumptions E.1, E.2, E.3′ and E.4′ hold. Suppose in addition we have* $\theta_{\min}(D) \geq c_0 > 0$ *for some universal constant* c_0*. Then under the flat prior, if the dimension* p *satisfy* $p^6 = o(n)$*, we have*

- *Posterior Consistency:*

$$\Pi\left(\sqrt{\frac{n}{p}} \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \geq M_n \mid \mathbb{D}_n\right) \xrightarrow{P^*} 0,$$

  for any non-random sequence $M_n \to \infty$*.*

- *Berstein-von-Mises Theorem:*

$$\left\| p(\boldsymbol{\beta} \mid \mathbb{D}_n) - \phi\left(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \frac{G^{-1}}{n}\right) \right\|_{TV} \xrightarrow{P^*} 0,$$

where $\hat{\boldsymbol{\beta}}$ *is the classic quantile regression estimator,* G *and* ϕ *are the same in Theorem V.1.*

Part 1 of the above Proposition characterizes the $\sqrt{n/p}$-rate of convergence, which coincides with the convergence rate for the frequentist estimator in increasing dimensions (*He and Shao*, 2000). Our result strengthens *Sriram and Ramamoorthi* (2018) by providing the exact rate of contraction. Along with most literature for Bayesian estimation, the convergence is measured in terms of the ℓ_2 error for the entire p-dimensional vector. Part 2 of Proposition 8 asserts the joint posterior distribution approaches a multivariate Gaussian distribution. Since the convergence is in Total Variation norm, the proposition guarantees that the posterior distribution for the entire p-dimensional vector is uniformly close to its limit, even though the dimension of the distribution may increase with the sample size. Without model sparsity and

shrinkage prior, our result suggests that the adjusted posterior inference for quantile regression in *Yang et al.* (2016) applies to the case with $p \to +\infty$.

The growth rate of p in Proposition 8 is much stronger than that in frequentist analysis of quantile regression (*He and Shao*, 2000). This is certainly not the best possible rate. However, note the Bayesian and frequentist results are different in nature: Our result considers the distributional approximation for the p-dimensional regression vector as a whole, whereas the frequentist results consider the asymptotic normality for a low-dimensional linear combination of all parameters. Even though *He and Shao* (2000) establishes a Bahadur-type representation for the regression vector, the joint Gaussian approximation does not necessarily apply in diverging dimensions. This is because the conditions for CLT in high dimensions is not known until recently (*Chernozhukov et al.*, 2013, 2017). Using the idea of coupling, *Belloni et al.* (2019a) establishes the distributional approximation for the entire quantile regression vector under comparable conditions to our Proposition 8.

Here we make some connections with the literature on Bernstein-von-Mises (BvM) theorems for general regression models, which corresponds to part 2 of Proposition 8. With increasing dimensions, the literature is relatively sparse. For general smooth parametric models, *Ghosal* (1999) and *Panov and Spokoiny* (2015) derive the BvM theorem for the posterior distribution under the condition $p^4 = o(n)$. Here smooth means the likelihood is second-order differentiable with respect to the parameter. Further, *Spokoiny* (2013) shows that the condition $p^3 = o(n)$ is necessary for the BvM theorem to hold in an one-sample Poisson model. Even for those smooth models, this necessary condition is stronger than the known rate for the frequentist analysis of general M-estimators, which is $p^2 \log p = o(n)$ (*He and Shao*, 2000). We refer the readers to *Panov and Spokoiny* (2015) for a detailed discussion on the differences. Turning to Bayesian quantile regression, less is known about its asymptotic behaviour since our working likelihood (5.3) is neither smooth nor correctly specified. Therefore

in our asymptotic analysis we require a more stringent condition of $p^6 = o(n)$.

5.5.2 Posterior asymptotics under the CA prior

Now we incorporate shrinkage priors and model sparsity into the pseudo-Bayesian framework. Our theoretical results apply to both sparse ($s_n \ll p_n$) and dense ($s_n = p_n$) models. For technical simplicity, we only consider the CA prior (5.5).

Let $\theta_{\min}(\cdot)$ be the minimal eigenvalue of a matrix, and recall $\underline{b}_n$ from Assumption E.5′, as well as the matrix $D = \mathrm{E}^*[x_i x_i^T]$. The following result generalizes Theorem V.1 to the case with increasing dimensions, under a few more technical conditions.

Theorem V.3 (CA shrinkage for increasing dimension)**.** *Consider the improper CA prior (5.5). Suppose Assumptions E.1, E.2, and E.3′ through E.5′ hold. If $s^4 p^2 \log^2 n = o(n)$, and the tuning parameter λ_n is chosen such that*

$$\frac{\sqrt{s} p \log p}{\sqrt{n}} \ll \lambda_n \ll \min\left\{ \underline{b}_n,\ \frac{1}{\sqrt{s}},\ \underline{b}_n \sqrt{\theta_{\min}(D)} \right\},$$

then we have the following results:

1. *Posterior consistency:*

$$\Pi\left(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^0\|_2 \geq M_n \sqrt{\frac{s}{n}}\ ;\ \text{or}\ \ \|\boldsymbol{\beta}_2\|_\infty \geq M_n \frac{s \log p}{n \lambda_n} \,\middle|\, \mathbb{D}_n \right) \xrightarrow{P^*} 0,$$

   for any sequence $M_n \to +\infty$.

2. *Distributional approximation:*

$$\left\| p(\boldsymbol{\beta} \mid D_n) - \phi\left(\boldsymbol{\beta}_1; \tilde{\boldsymbol{\beta}}_1, \frac{G_{11}^{-1}}{n} \right) \otimes \prod_{j=s+1}^{p} \frac{n\lambda_n}{2} \exp\left\{ -n\lambda_n |\beta_j| \right\} \right\|_{TV} \xrightarrow{P^*} 0.$$

   Here $\phi(\cdot)$ and G are the same as in Theorem V.1, and $\tilde{\boldsymbol{\beta}}_1$ is the oracle estimator using the true model.

To cover a wide range of scenarios, the range for the tuning parameter is entangled with many other factors; We shall provide comments on the conditions later via a few examples. Part 1 of the above Theorem states that with the CA prior, the pseudo-Bayesian quantile regression achieves adaptive and near-oracle performance in terms of estimation accuracy. For those active coefficients, the posterior achieves the oracle $\sqrt{n/s}$-rate of contraction around the true parameter. The rate is the same as if we knew the true model in advance. On the other hand, the convergence rate for inactive coefficients is $(n\lambda_n)/(s \log p) \gg \sqrt{p^2 n/s}$, per the conditions on λ_n. Therefore, the posterior for inactive coefficients will be highly concentrated around 0, explaining the name 'adaptive'. Since the measure is the ℓ_∞ norm, this rate of contraction holds simultaneously for every single component of the inactive coefficients, even when there is a large number of inactive coefficients. Such an adaptive contraction rate is comparable with Theorem 2.3 of *Song and Liang* (2017), where they focus on the Gaussian linear model. Our result is more general, as we work with the mis-specified likelihood (5.3).

Part 2 of Theorem V.3 provides the distributional approximation for the posterior, similar to Theorem V.1. In the diverging dimension case, nonetheless, it is sometimes more realistic to consider a linear combination of the p-dimensional parameters $\boldsymbol{\alpha}^T\boldsymbol{\beta}$, where $\|\boldsymbol{\alpha}\| = 1$ (*Fan and Peng*, 2004; *He and Shao*, 2000). Proposition V.3 then implies that the marginal posterior distribution for $\boldsymbol{\alpha}^T\boldsymbol{\beta}$ is asymptotically normal, provided that $\boldsymbol{\alpha}_1 \neq 0$.

Here we comment on the condition in Theorem V.3. The involved conditions depend on: (i) the dimension, (ii) the minimal eigenvalue of $D = \mathrm{E}^*[x_i x_i^T]$, and (iii) the minimal signal strength $\underline{b}_n$. We shall provide a few examples later to make the requirement for the theorem explicit and visible. The difficulty behind such complicated requirements is twofold. First, we allow both the dimension p_n and s_n to diverge with sample size at a reasonable rate, and we do not explicitly require the

model to be sparse; see Example 2 below. Second, our Assumption E.3′ does not restrict the gram matrix D to have uniformly bounded eigenvalues as n and p grows; see Example 3 below Therefore, our setting is more general than many other existing results on shrinkage estimation and variable selection, either in the frequentist (*Kim et al.*, 2008; *Huang et al.*, 2008) or the Bayesian framework (*Armagan et al.*, 2013b; *Song and Liang*, 2017).

We give a few representative examples to explain Theorem V.3 under different scenarios. For each of those specific cases, we are able to derive more explicit and intuitive conditions under which Theorem V.3 holds.

Example 1. Consider a sparse model where s_n stays a constant yet $p_n \to \infty$. In addition to Assumption E.3′, suppose the minimal eigenvalue for D is uniformly bounded from below $\theta_{\min}(D) \geq c_0 > 0$. Suppose all other assumptions required by Theorem V.3 hold.

In this simple case, our conclusions in Theorem V.3 hold as long as $p^2 \log^2 p = o(n)$ and $\underline{b}_n \gg \frac{p \log p}{\sqrt{n}}$. The growth rate condition for the dimension p is considerably relaxed compared with Proposition 8. In addition, the range for the tuning parameter in Theorem V.3 reduces to

$$\frac{p \log p}{\sqrt{n}} \ll \lambda_n \ll \underline{b}_n.$$

This rate for the tuning parameter is more intuitive and matches with that in the literature (*Fan and Peng*, 2004).

Example 2. Consider a dense model where $s_n = p_n \to \infty$. Suppose in this case $\|\boldsymbol{\beta}^0\| = c_0$ stays fixed, and $\underline{b}_n \asymp 1/\sqrt{p}$. This example holds if the magnitudes of all regression coefficient are the same. Again, suppose all other assumptions required by Theorem V.3 hold. Note Assumption E.3′ implies $\theta_{\min}(D) \geq \theta_{11} > 0$ uniformly in a dense model.

Under this type of dense models, Theorem V.3 still holds as long as $p^6 \log^2 n =$

$o(n)$. This dimensionality constraint is roughly equivalent to that of Proposition 8. Using a shrinkage prior, the conclusions in Theorem V.3 reduce to those in Proposition 8 under a flat prior. This suggest that a properly-tuned shrinkage prior does not hurt the performance, even when the model is not sparse.

Now, the proper range for the tuning parameter in Theorem V.3 becomes

$$\frac{p^{1.5}\log p}{\sqrt{n}} \ll \lambda_n \ll \frac{1}{\sqrt{p}}. \tag{5.12}$$

However, this range for the tuning parameter (5.12) is not the best possible. If we know the model is dense, we can simply choose $\lambda_n \equiv 0$, i.e., use the flat prior. The asymptotic results in Theorem V.3 would still hold.

Example 3. Consider a quantile regression model with one active predictor Z:

$$Q_\tau(Y \mid Z = z) = \beta_0 + \beta_1 z,$$

with a large enough β_1. Beyond Z, suppose we have a sequence of other predictors $X_1, X_2, \ldots$ that is irrelevant to Y, yet predictive for Z. To be more precise, we assume

$$Z = \sum_{k=1}^{\infty} \alpha_k X_k, \quad \text{and} \quad X_k \overset{i.i.d.}{\sim} \mathrm{N}(0,1), \;\; k = 1, 2, \ldots,$$

where $\alpha_k \asymp 1/k$. Note the infinite sum on the right hand side converges in the L_2 sense.

For each p, there is collinearity among the covariates when fitting the model using all of Z and $X_1, \ldots, X_p$. The population covariance matrix of $(Z, X_1, \ldots, X_p)$ can be

written as:

$$D_p = \left(\begin{array}{c|ccc} \sum_{k=1}^{\infty} \alpha_k^2 & \alpha_1 & \cdots & \alpha_p \\ \hline \alpha_1 & & & \\ \vdots & & I_p & \\ \alpha_p & & & \end{array}\right).$$

We show in Lemma IA.4 in Section 5.9.6 that $\theta_{\min}(D_p) \asymp 1/p$, which satisfies Assumption E.3′. Therefore Theorem V.3 holds when $p^3 \log^2 p = o(n)$, and

$$\frac{p \log p}{\sqrt{n}} \ll \lambda_n \ll \frac{1}{\sqrt{p}},$$

given the other assumptions on the data generating process. Note the condition on p is more stringent than the explicit requirement in Theorem V.3, i.e., $p^2 \log^2 p = o(n)$. This is due to the range of the tuning parameter has to be non-empty. Hence, we show that Theorem V.3 covers certain situations with collinearity.

5.5.3 Practical posterior inference in higher dimensions

The results in this section do not apply to the high-dimensional regime where $p \gg n$. Under such a scenario, the asymmetric Laplace working likelihood (5.3) degenerates, and Bayesian inference for quantile regression is much less understood in the literature. Furthermore, the variance adjustment in Section 5.4 is not applicable when $n < p$, since it relies on estimation of the full covariance matrix $\mathrm{E}^*[XX^T]$. Therefore, direct application of the pseudo-Bayesian approach becomes problematic.

Nonetheless, the pseudo-Bayesian approach can be useful when combined with the idea of marginal screening (*Fan and Lv*, 2008). For high-dimensional sparse problems with $s \ll n < p$, it is often practically useful to employ a fast screening step to reduce the dimension to a manageable scale, prior to further statistical analysis (*Fan and Lv*, 2010; *Liu et al.*, 2015; *Barut et al.*, 2016). Such screening is routinely applied in

many real-world applications (*Bermingham et al.*, 2015; *Tamba et al.*, 2017).

For inference in high dimensional quantile regression, we suggest using our pseudo-Bayesian framework after applying a quantile sure screening procedure such as those proposed by *He et al.* (2013), *Wu and Yin* (2015), *Shao and Zhang* (2014) and *Ma et al.* (2017). Under appropriate conditions, those screening procedures keep all relevant covariates with probability approaching one, while at the same time the total number of retained covariates is $d_n = O(n^r)$ for some $r < 1$. Our Theorem V.3 then applies to the d_n-dimensional posterior distribution post-screening.

5.6 Computational details

In this section, we present the Bayesian hierarchical model for quantile regression under shrinkage prior, and derive the induced posterior sampling techniques. Though theoretical properties under CA (5.5) and AL (5.4) priors are similar, the AL prior is computational more efficient (*Alhamzawi et al.*, 2012; *Benoit and Van den Poel*, 2017). For the CA prior, we also propose a feasible Gibbs sampling device based on the piece-wise interpretation of the Bayesian Lasso (*Hans*, 2009).

5.6.1 Bayesian hierarchy under the AL prior

We first provide a review of the Bayesian Adaptive Lasso quantile regression. See *Alhamzawi et al.* (2012) and *Li et al.* (2010) for a more detailed discussion. Following *Kozumi and Kobayashi* (2011), denote

$$\theta = \frac{1-2\tau}{\tau(1-\tau)} \quad \text{and} \quad \rho = \sqrt{\frac{2}{\tau(1-\tau)}}.$$

The Asymmetric Laplace distribution with skewness τ in (5.3) can be succinctly formulated as $\theta v + \rho\sqrt{v}z$, where $v_i \sim Exp(1)$ and $z_i \sim N(0,1)$. By *Andrews and*

Mallows (1974), the double-exponential prior can be represented as follows:

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{z^2}{2s}\right) \cdot \exp\left(-\frac{a^2 s}{2}\right) \, \mathrm{d}s.$$

This is a scale mixture of Normal distribution, where the mixing distribution is Exponential. When λ and the scale parameter $\sigma = 1$ in (5.3) are fixed, the Bayesian hierarchy of the Adaptive Lasso quantile regression is

$$\begin{aligned}
\text{Likelihood:} \quad p(y_i \mid v_i, \boldsymbol{\beta}) &\propto \sqrt{\frac{1}{2\pi\rho^2 v_i}} \cdot \exp\left(-\frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}_{-0} - \theta v_i)^2}{2\rho^2 v_i}\right), \\
p(v_i \mid \beta) &\propto \exp\left(-v_i\right), \\
\text{Prior:} \quad p(\beta_0) &\propto 1, \\
p(\beta_j \mid s_j) &\propto \sqrt{\frac{1}{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j}\right), \quad j = 1, \ldots, p, \\
p(s_j) &\propto \exp\left(-\frac{n\lambda^2 s_j}{2|\hat{\beta}_j|^2}\right).
\end{aligned}$$

The full posterior distribution of $\beta, \boldsymbol{v}, \boldsymbol{s}$ is

$$\begin{aligned}
p(\beta_0, \boldsymbol{\beta}_{-0}, \boldsymbol{v}, \boldsymbol{s} \mid D_n) \propto & \prod_{i=1}^n \sqrt{\frac{1}{2\pi\rho^2 v_i}} \cdot \exp\left(-\frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}_{-0} - \theta v_i)^2}{2\rho^2 v_i} - v_i\right) \\
& \times \prod_{j=1}^p \sqrt{\frac{1}{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j} - \frac{n\lambda^2 s_j}{2|\hat{\beta}_j|^2}\right).
\end{aligned}$$

The above expression yields the following list of conditional distributions.

1. The conditional distribution of $s_j \mid \cdot$ is $(j = 1, \ldots, p)$

$$p(s_j \mid \boldsymbol{\beta}, \boldsymbol{v}, D_n) \propto \sqrt{\frac{1}{s_j}} \exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2}{s_j} + \frac{n\lambda^2 s_j}{|\hat{\beta}_j|^2}\right)\right\}.$$

   This is a generalized Inverse Gaussian distribution $GIG(1/2, n\lambda^2/|\hat{\beta}_j|^2, \beta_j^2)$ (*Jorgensen*, 2012). A generalized Inverse Gaussian distribution $GIG(p, a, b)$ is

parametrized as

$$f(x; p, a, b) \propto x^{p-1} e^{-(ax+b/x)/2}.$$

2. The conditional distribution of $v_i \mid \cdot$ is $(i = 1, \ldots, n)$

$$\begin{aligned} p(v_i \mid \boldsymbol{\beta}, \boldsymbol{s}, D_n) &\propto \sqrt{\frac{1}{v_i}} \exp\left\{ -\frac{1}{2\rho^2} \left(\frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \beta_{-0})^2}{v_i} + \left(\theta^2 + 2\rho^2\right) v_i \right) \right\} \\ &\sim GIG\left(\frac{1}{2}, \ \frac{\theta^2}{\rho^2} + 2, \ \frac{[y - \beta_0 - x_i^T \beta_{-0}]^2}{\rho^2} \right) \end{aligned}$$

This is again a generalized Inverse Gaussian distribution.

3. The conditional distribution of $\beta_0 \mid \cdot$ is

$$\begin{aligned} p(\beta_0 \mid \boldsymbol{\beta}_{-0}, \boldsymbol{v}, \boldsymbol{s}, \boldsymbol{D}_n) &\propto \exp\left\{ -\sum_{i=1}^{n} \frac{(\beta_0 - y_i + x_i^T \boldsymbol{\beta}_{-0} + \theta v_i)^2}{2\rho^2 v_i} \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left(\sum_{i=1}^{n} \left[\frac{1}{\rho^2 v_i} \right] \beta_0^2 - 2 \sum_{i=1}^{n} \left[\frac{y_i - x_i \beta_{-0} - \theta v_i}{\rho^2 v_i} \right] \beta_0 \right) \right\} \\ &\sim N(\bar{\mu}_0, \ \bar{\sigma}_0^2), \end{aligned}$$

where $1/\bar{\sigma}_0^2 = \rho^{-2} \sum_{i=1}^{n} v_i^{-1}$ and

$$\bar{\mu}_0 = \bar{\sigma}_0^2 \cdot \sum_{i=1}^{n} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{-0} - \theta v_i}{\rho^2 v_i}.$$

4. The conditional distribution of the vector $\boldsymbol{\beta}_{-0} \mid \cdot$ follows from a similar strategy.

Let $\tilde{y}_i = y_i - \beta_0 - \theta v_i$, $\boldsymbol{W} = diag\{\rho^2 v_i\}$ and $\boldsymbol{S} = diag\{s_i\}$

$$\begin{aligned} p(\boldsymbol{\beta}_{-0} \mid \beta_0, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{D}_n) &\propto \exp\left\{ -\sum_{i=1}^{n} \frac{(\tilde{y}_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_{-0})^2}{2\rho^2 v_i} - \sum_{j=1}^{p} \frac{\beta_j^2}{2 s_j} \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}_{-0}^T (X^T W^{-1} X + S^{-1}) \boldsymbol{\beta}_{-0} + \tilde{\boldsymbol{Y}} W^{-1} X \boldsymbol{\beta}_{-0} \right\} \\ &\propto \exp\left\{ -\frac{1}{2} (\beta_{-0} - \bar{\boldsymbol{\mu}}) \bar{B} (\beta_{-0} - \bar{\boldsymbol{\mu}}) \right\} \\ &\sim N\left(\bar{\boldsymbol{\mu}}, \bar{B}^{-1} \right), \end{aligned} \tag{5.13}$$

where

$$\bar{\boldsymbol{B}} = X^T W^{-1} X + S^{-1}, \qquad \bar{\boldsymbol{\mu}} = \bar{\boldsymbol{B}}^{-1} X^T W^{-1} \tilde{\boldsymbol{Y}}.$$

In the above algorithm, step (5.13) can be implemented in a sequential way as in *Li et al.* (2010). That is, sample one component β_j from its full conditional distribution $p(\beta_j \mid \boldsymbol{\beta}_{-j}, \boldsymbol{v}, \boldsymbol{s}, \boldsymbol{D}_n)$ at a time for $j = 1, \ldots, p$. The block update (5.13) offers a faster mixing behaviour in the Gibbs sampler, at a cost of computing the inverse of a p by p matrix in each iteration.

5.6.2 Bayesian hierarchy under the CA prior

While Metropolis-Hastings algorithms for Bayesian quantile regression under the smooth SCAD prior are available (*Adlouni et al.*, 2018), here we derive a direct Gibbs sampling device based on our modified CA prior. Our method adapts from the Gibbs sampler for Bayesian Lasso (*Hans*, 2009). With the same characterization of Asymmetric Laplace likelihood, the Bayesian hierarchy follows similarly from the previous section.

$$\begin{aligned}
\text{Likelihood:} \quad p(y_i \mid v_i, \boldsymbol{\beta}) &\propto \sqrt{\frac{1}{2\pi\rho^2 v_i}} \cdot \exp\left(-\frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}_{-0} - \theta v_i)^2}{2\rho^2 v_i} \right), \\
p(v_i \mid \beta) &\propto \exp\left(-v_i\right), \\
\text{Prior:} \quad p(\beta_0) &\propto 1, \\
p(\beta_j) &\propto \exp\left(-n\lambda \min\{|\beta_j|, \lambda\}\right), \quad j = 1, \ldots, p.
\end{aligned}$$

Note the prior on (β_j) is improper, in the sense that the integration over β_j diverges. Nonetheless, the following full posterior is a proper distribution that can be normalized. Thus the validity of using the posterior distribution (5.6) is still well justified

(*Gelman et al.*, 2013).

$$p(\beta_0, \boldsymbol{\beta}_{-0}, \boldsymbol{v}, \boldsymbol{s} \mid D_n) \propto \prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\rho^2 v_i}} \cdot \exp\left(-\frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}_{-0} - \theta v_i)^2}{2\rho^2 v_i} - v_i\right) \times \prod_{j=1}^{p} \exp\left(-n\lambda \min\{|\beta_j|, \lambda\}\right).$$

The above posterior yields the following full conditional distributions.

1. The conditional distribution of $v_i \mid \cdot$ is the same Generalized Inverse Gaussian as the previous section $(i = 1, \ldots, n)$

$$p(v_i \mid \boldsymbol{\beta}, D_n) \sim GIG\left(\frac{1}{2}, \frac{\theta^2}{\rho^2} + 2, \frac{[y - \beta_0 - x_i^T \beta_{-0}]^2}{\rho^2}\right)$$

2. The conditional distribution of $\beta_0 \mid \cdot$ is the same Normal distribution as the previous section

$$p(\beta_0 \mid \boldsymbol{\beta}_{-0}, \boldsymbol{v}, \boldsymbol{D}_n) \sim N(\bar{\mu}_0, \bar{\sigma}_0^2).$$

3. Now we derive the conditional distribution of $\beta_j \mid \cdot$ for each $j = 1, \ldots, p$. We start by noting the prior $p(\beta_j)$ can be written in a piece-wise fashion.

$$p(\beta_j) = \begin{cases} n\lambda\beta_j, & \text{if } 0 < \beta_j < \lambda, \\ -n\lambda\beta_j, & \text{if } -\lambda < \beta_j < 0, \\ n\lambda^2, & \text{if } |\beta_j| \geq \lambda. \end{cases}$$

Following *Hans* (2009), we write the conditional distribution of $\beta_j \mid \cdot$ as a mixture of truncated Normal. Let $\tilde{y}_i = y_i - \theta v_i - \beta_0 - \sum_{k \neq j} x_{ik}\beta_k$. Denote by $N^+(\mu, \sigma^2)$ as a generic truncated Normal distribution on $0 < x \leq \lambda$, $N^-(\mu, \sigma^2)$ as the truncated Normal on $-\lambda < x < 0$, and $N^0(\mu, \sigma^2)$ as the one truncated

on $|x| \geq \lambda$.

$$
\begin{aligned}
& p(\beta_j \mid \beta_{-j}, \boldsymbol{v}, \boldsymbol{D}_n) && (5.14)\\
\propto\ & \exp\left\{ -\sum_{i=1}^{n} \frac{(\tilde{y}_i - x_{ij}\beta_j)^2}{2\rho^2 v_i} - n\lambda \min\{|\beta_j|, \lambda\} \right\} \\
\propto\ & \exp\left\{ -\frac{1}{2\rho^2} \left(\sum_{i=1}^{n} \left[\frac{x_{ij}^2}{v_i} \right] \beta_j^2 - 2 \sum_{i=1}^{n} \left[\frac{x_{ij}\tilde{y}_i}{v_i} \right] \beta_j \right) - n\lambda \min\{|\beta_j|, \lambda\} \right\} \\
\propto\ & \begin{cases} \dfrac{\exp(-n\lambda^2)}{\phi\left(0; \mu_j^O, \sigma_j^2\right)} \cdot \phi\left(\beta_j; \mu_j^O, \sigma_j^2\right), & \text{if } |\beta_j| \geq \lambda \\ \dfrac{1}{\phi\left(0; \mu_j^+, \sigma_j^2\right)} \cdot \phi\left(\beta_j; \mu_j^+, \sigma_j^2\right), & \text{if } 0 < \beta_j < \lambda \\ \dfrac{1}{\phi\left(0; \mu_j^-, \sigma_j^2\right)} \phi\left(\beta_j; \mu_j^-, \sigma_j^2\right), & \text{if } -\lambda < \beta_j < 0 \end{cases} \\
\sim\ & \kappa_j^+ N^+\left(\mu_j^+, \sigma_j^2\right) + \kappa_j^- N^-\left(\mu_j^-, \sigma_j^2\right) + \kappa_j^0 N^0\left(\mu_j^O, \sigma_j^2\right), && (5.15)
\end{aligned}
$$

where $\phi(x; \mu, \sigma^2)$ is the density function for a generic Normal distribution $N(\mu, \sigma^2)$. Let $W = diag\{\rho^2 v_i\}$ The parameters in (5.15) are given by

$$
\begin{aligned}
\sigma_j^{-2} &= \frac{\sum_{i=1}^{n} x_{ij}^2 / v_i}{\rho^2} = (X^T W^{-1} X)_{j,j}, \\
\mu_j^0 &= \frac{\sigma_j^2 \sum_{i=1}^{n} x_{ij}\tilde{y}_i / v_i}{\rho^2}, \\
\mu_j^+ &= \mu_j^0 - n\lambda\sigma_j^2, \\
\mu_j^- &= \mu_j^0 + n\lambda\sigma_j^2, \\
\kappa_j^+ &= \left[\frac{P_j^+}{\phi(0; \mu_j^+, \sigma_j^2)} \right] \Bigg/ \left[\frac{\exp(-n\lambda^2) P_j^0}{\phi(0; \mu_j^0, \sigma_j^2)} + \frac{P_j^-}{\phi(0; \mu_j^-, \sigma_j^2)} + \frac{P_j^+}{\phi(0; \mu_j^+, \sigma_j^2)} \right], \\
\kappa_j^- &= \left[\frac{P_j^-}{\phi(0; \mu_j^-, \sigma_j^2)} \right] \Bigg/ \left[\frac{\exp(-n\lambda^2) P_j^0}{\phi(0; \mu_j^0, \sigma_j^2)} + \frac{P_j^-}{\phi(0; \mu_j^-, \sigma_j^2)} + \frac{P_j^+}{\phi(0; \mu_j^+, \sigma_j^2)} \right], \\
\kappa_j^0 &= 1 - \kappa_j^+ - \kappa_j^-.
\end{aligned}
$$

Note the variances for the three mixing components are the same, and are equal to the j-th diagonal element of $X^T W^{-1} X$. The corresponding constants P_j^+, P_j^- and P_j^0 normalizes the truncated distributions N^+, N^- and N^0 respectively. Let $\Phi(\cdot)$ denote the cdf for a standard normal distribution, the constants can be calculated explicitly as

$$\begin{aligned} P_j^+ &= \Phi\left(\frac{\lambda - \mu_j^+}{\sigma_j}\right) - \Phi\left(-\frac{\mu_j^+}{\sigma_j}\right), \\ P_j^- &= \Phi\left(-\frac{\mu_j^-}{\sigma_j}\right) - \Phi\left(\frac{-\lambda - \mu_j^-}{\sigma_j}\right), \\ P_j^0 &= 1 - \Phi\left(\frac{\lambda - \mu_j^0}{\sigma_j}\right) + \Phi\left(\frac{-\lambda - \mu_j^0}{\sigma_j}\right). \end{aligned}$$

The Gibbs Sampling under CA prior requires a component-wise update for each β_j in step (5.15). In one Gibbs iteration, we sample each β_j from a different mixture representation (5.15). The simultaneous block update of the entire p-dimensional vector $\boldsymbol{\beta}$ is not tractable as in the Adaptive Lasso case (5.13). This is due to the piece-wise nature of the prior $p_\lambda(\beta_j)$, leading to a posterior of mixture-form. Though the mixture weights κ_j are explicitly tractable in the univariate case, the multivariate mixture weights are not.

Though the posterior sampling under the CA prior are easy to implement, its mixing properties hinges on the correlation structure of the predictors. Compared with the block-update (5.13), *Hans* (2009) points out that a component-wise update as in (5.15) may increase the auto-correlation in the MCMC chain. The phenomena exacerbate when the features X_j are highly correlated. An orthogonalized Gibbs sampling step may replace (5.15) to improve the mixing behaviour of the Gibbs sampler.

5.6.3 The role of the tuning parameter

The remaining practical issue is selecting a hyper-parameter λ. We first comment on the role of λ. A large λ helps improve the efficiency for those zero coefficients, as their posterior variances are of the order $1/(n\lambda)^2$. A small λ ensures that the posterior is unbiased for those non-zero coefficients. A reasonable choice of λ should strike a balance in between. In theory, as Theorems V.1 V.2 show, the posterior inference remains valid and improves over no shrinkage as long as λ falls in a certain range. In our numerical experience, we have confirmed that the performance of the interval estimates is not much affected by λ within a proper range. The choice of λ does, however, affect the amount of efficiency gain of the posterior inference.

Unfortunately, there has not been a systematic approach for automatic tuning parameter selection in our context. For a Bayesian treatment, one practical method is the Empirical Bayes approach in *Casella* (2001), which finds the marginal maximum-likelihood estimator of λ in the joint likelihood function of λ and $\boldsymbol{\beta}$. However, since our working-likelihood may be seriously mis-specified, such a model-based approach is not well-justified. From a frequentist perspective, cross-validation (CV) is prevalent for tuning parameter selection (*Wu and Liu*, 2009; *Zou*, 2006). Since CV aims to minimize the prediction error, it may not work for posterior inference. Through numerical evidence, we find that CV sometimes chooses a λ that is too large. We suggest using a smaller λ than that chosen from CV, to ensure the posterior inference is valid under the shrinkage prior.

5.7 Simulation

In this section, we present our findings from simulation studies. We verify that the adjusted posterior inference provides adaptive confidence intervals for quantile regressions, under a wide range of settings. By adaptive, we mean the method (i)

achieves oracle efficiency for active coefficients and (ii) achieves super-efficiency for inactive coefficients, as discussed in Section 5.4. We also compare the results with some standard frequentist approaches. In what follows, we introduce the methods and the simulation scenarios that we consider in this section.

Throughout this section, we may use four different methods to construct confidence intervals for quantile regression coefficients. We first introduce some useful abbreviations for those methods, which are highlighted in **bold** below.

- **Full** – Fit the classic quantile regression using all available covariates, then apply the rank-score approach (*Gutenbrunner and Jurecková*, 1992) to construct confidence intervals.

- **Refit** – First apply the Adaptive Lasso (*Wu and Liu*, 2009) for variable selection, and refit the quantile regression with the selected covariates. Then apply the rank-score approach on the refitted model. For a coefficient that is not selected by the Adaptive Lasso, we report its confidence interval as a single point $\{0\}$.

- **BayesM** – The adjusted posterior inference in Section 5.4, under the Adaptive Lasso prior (5.4). The confidence intervals are centered at the posterior means.

- **BayesF** – The same adjusted posterior inference as **BayesM**, but we use the Adaptive Lasso estimators as the center. As in Remark 14, we do not implement the weighting adjustment.

A few comments are in place for those methods. First, we comment on the choice of the frequentist inferential procedure. Among the variety of inferential methods for quantile regression (See e.g., *Koenker* (2005, Chapter 3)), we only consider the rank-score method in this section. In the quantile regression literature, it is known that the rank-score method enjoys robust and competitive performances. We do not

consider those resampling methods as they usually have similar performances to the rank-score method, yet with an increased computational cost. Second, we comment on the choice of the prior in the posterior-based methods. We do not consider the CA prior (5.5) due to the long mixing time of the Gibbs sampler, as mentioned in Section 5.6.2. Instead, the Adaptive Lasso prior (5.4) offers easy computation and satisfactory performance. Third, the methods **Refit**, **BayesM** and **BayesF** all need a tuning parameter λ. To make a fair comparison, we shall always use the same λ when comparing the performances of those shrinkage methods.

Now we provide a summary of different data generating processes in this section. Let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be the covariates, y be the response, and u be the error terms independent of $\boldsymbol{x}$. Define $Q_\tau(u)$ as the marginal τ-th quantile of u. In what follows, we shall fix $p = 6$ unless stated otherwise. Note the values of $\boldsymbol{\beta}$ and sample sizes will be given in each specific example later. In this section, we consider the following four data generating models.

(A). An *i.i.d.* error model:

$$y = 1 + \boldsymbol{x}^T \boldsymbol{\beta}_A + u,$$

where $\boldsymbol{x} \sim \mathrm{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$; We consider two possible error distributions: (i) $u \sim \mathrm{N}(0, 1)$ and (ii) $u \sim \mathrm{Exp}(1)$. In this model, the conditional τ-th quantile of y given $\boldsymbol{x}$ is $[1 + Q_\tau(u)] + \boldsymbol{x}^T \boldsymbol{\beta}_A$, which is linear in $\boldsymbol{x}$ at any τ.

(B). A linear location-scale model:

$$y = 1 + \boldsymbol{x}^T \boldsymbol{\beta}_B + [3.5 + 2x_6]\, u,$$

where x_6 is generated from the uniform distribution on the interval $[-1.5, 1.5]$; Conditioning on x_6, we generate $x_1, \ldots, x_5 \sim \mathrm{N}(x_5/3,\, 1)$ independently; The error term $u \sim (\Gamma(3, 2) - c_0)$, where the constant c_0 is chosen such that the first quartile of u is 0. In this model, the conditional quantile of y given $\boldsymbol{x}$ is

$[1 + 3.5Q_\tau(u)] + \boldsymbol{x}^T\boldsymbol{\beta}_B + [2Q_\tau(u)]x_6$, which is linear in $\boldsymbol{x}$ at any τ.

(C). A global conditional quantile model:

$$Q_\tau(y \mid \boldsymbol{x}) = 3\Phi^{-1}(\tau) + \boldsymbol{x}^T\boldsymbol{\beta}_C(\tau), \quad \forall \tau \in (0,1),$$

where $x_j \sim \text{Unif}(-1,1)$, $j = 1,\ldots,6$; $\Phi^{-1}(\tau)$ is the quantile function of the standard Normal distribution. Note the displayed conditional quantile function uniquely determines the conditional distribution of y given $\boldsymbol{x}$. We can simulate the observation y from

$$y = 3\Phi^{-1}(u) + \boldsymbol{x}^T\boldsymbol{\beta}_C(u),$$

with $u \sim \text{Unif}(0,1)$.

(D). An *i.i.d.* error model when the dimension p may increase with the sample size n:

$$y = 1 + \boldsymbol{x}^T\boldsymbol{\beta}_D + u,$$

where $\boldsymbol{x} \sim \text{N}(\boldsymbol{0}, \Sigma)$ with $\Sigma_{ij} = 0.85^{|i-j|}$; $u \sim t(2)$. In this model, the conditional τ-th quantile of y given $\boldsymbol{x}$ is $[1 + Q_\tau(u)] + \boldsymbol{x}^T\boldsymbol{\beta}_D$, which is linear in $\boldsymbol{x}$ at any τ.

Here we give some details about our upcoming simulation results. For each simulation setting, we generate 1000 Monte Carlo data-sets independently, and we evaluate the methods based on their performances on those 1000 realizations. We use the R package '*quantreg*' (*Koenker*, 2018) to compute the frequentist approaches. For the posterior-based method, we run the Gibbs sampler in Section 5.6.1 of length 20000 with a burn-in period of 3000 iterations. We find the posterior chain achieves sufficient mixing for our simulations.

5.7.1 The effect of the tuning parameter

While Theorem V.2 informs us that a wide range of λ provides asymptotically valid inference, the finite-sample effect of λ remains unclear. In this first part, we explore how the methods **BayesF**, **BayesM** and **Refit** depend on λ. To this end, we evaluate all candidate methods' performances at a range of different λ values. To keep the presentation concise, we only include two simulation scenarios here: model (A) with a small coefficient, and model (B).

5.7.1.1 In models with small coefficients

Our first example examines the effect of λ when the true model has a small coefficient. We consider model (A) with

$$\boldsymbol{\beta}_A = (1/10,\, 3,\, 0,\, -5,\, 0,\, 0),$$

and the error term $u \sim \mathrm{N}(0, 1)$; We fix the sample size at $n = 80$. Here we focus on the case with $\tau = 0.5$, i.e., the median regression. The active covariates include x_1, x_2 and x_4, despite that the regression coefficient for x_1 is relatively small. Such a small coefficient makes it difficult for shrinkage methods to deliver valid inference (*Leeb and Pötscher*, 2005). Figure 5.2 presents how different methods perform under a range of different λ; We shall only focus on β_1, β_2 and β_3 for simplicity. Note in this particular example, the interval lengths for **BayesF** and **BayesM** are identical. The weighting adjustment in Remark 13 does not affect the performance with limited sample sizes.

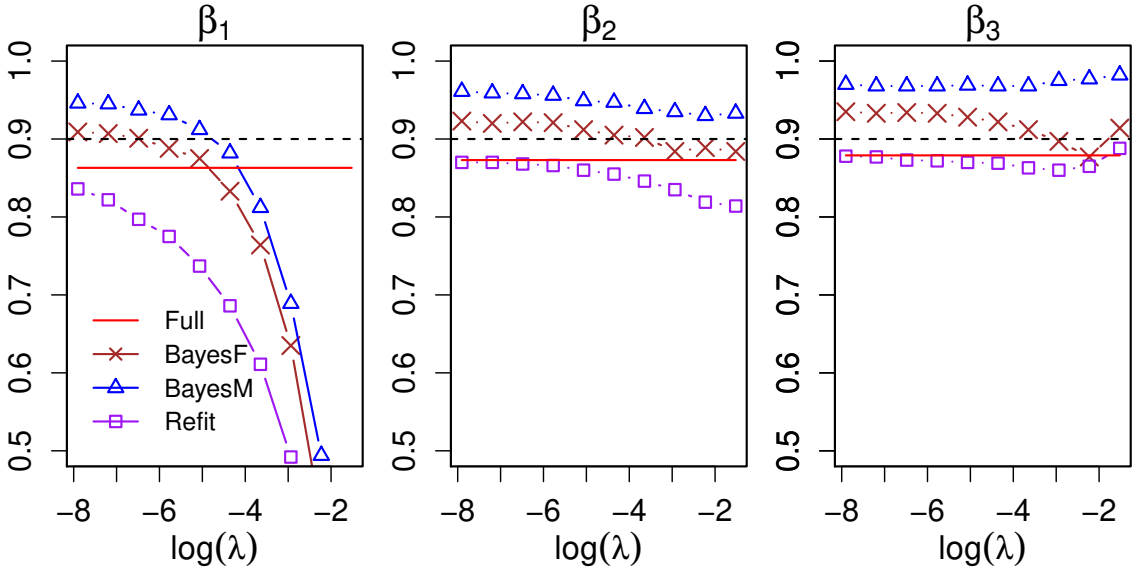
As we can see in Figure 5.2, while **Full** delivers valid inference for all coefficients, the performance of other shrinkage methods depend heavily on λ. With a small enough λ, all the shrinkage methods deliver satisfactory performances: Their empirical coverage probabilities are all around 90% for the presented coefficients. With a

limited sample size, there seem to be some size distortions for all methods: In general, **Refit** is more liberal while **BayesF** and **BayesM** are more conservative.

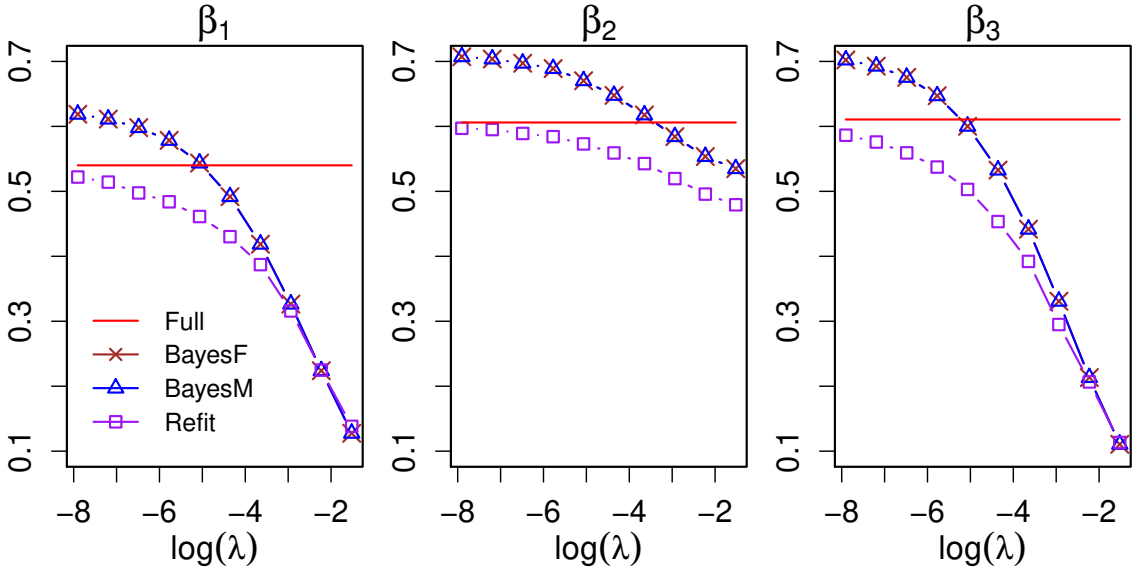
When λ increases, the inference for β_1, β_2 and β_3 shows different trends. For the inactive coefficient β_3, the shrinkage methods keep to provide valid inference: Increased λ only reduces the length of the interval, without jeopardizing the coverage probability too much. For the active coefficient β_2, the performance remains acceptable, as long as λ is not too large; This is because the true coefficient $\beta_2^0 = 3$ is relatively large. However, the inference for the small coefficient, β_1, breaks down rapidly as λ grows, which is a well-recognized limitation of shrinkage methods (*Leeb and Pötscher*, 2005). To be more specific, the coverage probabilities for all shrinkage methods are below 50% when $\lambda \approx$ 2e–01 = 0.2.

Among the three shrinkage methods, the posterior-based methods, **BayesM** and **BayesF**, are more robust to the change of λ, provided that λ is not too large. In Figure 5.2a for β_1, at λ = 1e-02 = 0.01, the empirical coverage probabilities of **BayesM** and **BayesF** are still close to 90%, while the coverage for **Refit** is even below 70%. This is not simply because the method **BayesM** is conservative. When looking at the shape of the coverage curve, the one for **Refit** drops much more sharply than **BayesM** and **BayesF**. We can observe the same behaviour in the plot for β_2 as well.

The reason behind this robustness is that, **Refit** relies on a dichotomous variable selection step. If the true coefficient is small yet non-zero, the variable selection method will often select that coefficient as exactly 0, therefore providing no uncertainty estimation (*Pötscher and Leeb*, 2009); The corresponding interval will then be a singleton $\{0\}$. This issue is known as the *non-uniformity* for shrinkage methods, as pointed out in *Leeb and Pötscher* (2005). On the other hand, the posterior-based methods will always provide a non-empty confidence interval for each coefficient, even when the true coefficient is close to 0. Thus, **BayesM** and **BayesF** mitigates the



(a) Empirical coverage



(b) Average length

Figure 5.2: Inference using different λ under model (A) at $\tau = 0.5$ and with normal error. The x-axis is on the log scale. The true regression coefficients are $\beta_1^0 = 0.10$, $\beta_2^0 = 3$, and $\beta_3^0 = 0$. Nominal level is 90%, marked with a black dashed line.

non-uniformity issue by providing a small margin of error for those small yet non-zero coefficients.

Here we conclude our findings in this example. When the true regression coefficients are relatively small, it is difficult for shrinkage methods to achieve adaptive inference. To guarantee the coverage probability for the coefficient β_1, we have to choose a small λ; In turn, this choice limits the efficiency gain for the inactive coefficients. Comparing with the classic **Refit** method, the posterior-based methods are more robust to poor choices of λ.

5.7.1.2 In models with heteroscedasticity

Our second example demonstrates the effect of λ in heteroscedastic models. We consider model (B) with

$$\boldsymbol{\beta}_B = (0, 3, 0, -5, 0, 0),$$

and we fix the sample size at $n = 500$. We focus on two different quantile levels $\tau = 0.25$ and $\tau = 0.75$. At those quantile levels, the true quantile regression coefficients are

$$\boldsymbol{\beta}^0_{0.25} = (0,\ 3,\ 0,\ -5,\ 0,\ 0), \quad \text{and} \quad \boldsymbol{\beta}^0_{0.75} = (0,\ 3,\ 0,\ -5,\ 0,\ 2.19),$$

respectively. At level $\tau = 0.25$, only x_2 and x_4 are active; Whereas x_6 becomes active at $\tau = 0.75$. In this example, we also include the results where we choose λ adaptively by 10-fold cross validation (CV). Note the CV approach is data-dependent, that is, the value of λ will be different from each simulated dataset.

Since the heteroscedasticity issue is severe in this example, we shall use the modified frequentist approaches, **Full**-*nid* and **Refit**-*nid*, in place of **Full** and **Refit**.

- **Full**-*nid* – The same as **Full**, but we do not use the original rank-score method for inference. Instead, we use the modified rank-score approach in *Koenker and Machado* (1999) that is robust to heteroscedasticity.

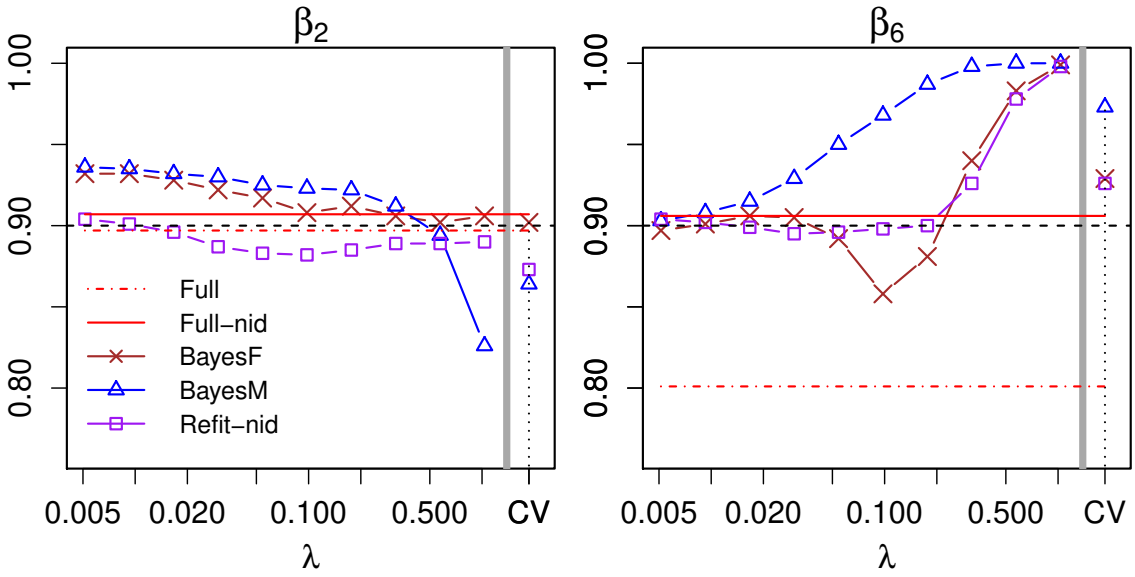
- **Refit**-*nid* – The same as **Refit**, the only difference is that we use the modified rank-score approach as above, after refitting the model.

The modified rank-score approch can adapt to heteroscedasiticity with large enough sample sizes; See *Koenker* (2005, Chapter 3). Correspondingly, its performance is less stable than the original rank-score method when the sample size is small. The 'nid' approaches are also implemented in the R package '*quantreg*' (*Koenker*, 2018).

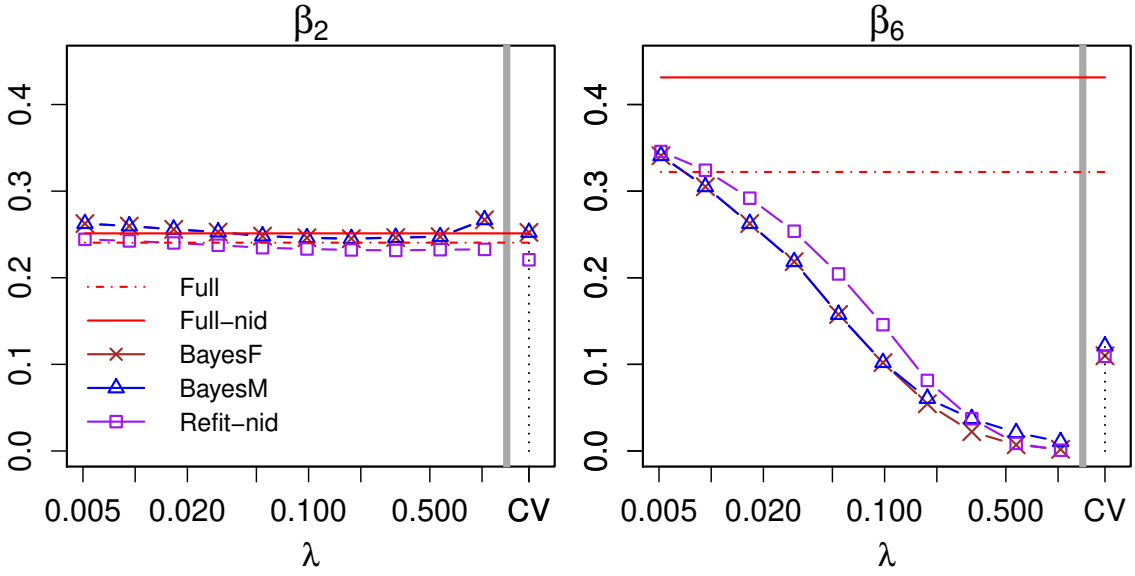
Figure 5.3 and 5.4 show the results for β_2 and β_6, respectively for $\tau = 0.25$ and $\tau = 0.75$. As we can see, **Full**-*nid* provides valid inference for all presented coefficients; Hence, we shall use **Full**-*nid* as a benchmark to compare other methods. The original **Full** method, however, fails drastically for the inference of β_6.

First we focus on the performance of **Refit**-*nid*, **BayesM** and **BayesF** for inactive coefficients. To this end, we examine the inference for β_6 at $\tau = 0.25$ in Figure 5.3. With smaller values of λ, all shrinkage methods provide valid inference. As λ grows, we can see that the coverage probability for **BayesM** approaches 100% much faster than all other methods. The coverage probabilities for **BayesF** and **Refit**-*nid* are similar, which stay below 95% for most values of λ in our range. When λ is large enough, nonetheless, all methods achieve near 100% coverage probability. For the lengths of the intervals, all shrinkage methods are similar, and they provide narrower interval than **Full**-*nid* for a wide range of λ. As λ grows, the efficiency gain is even more significant. In this example, we observe the same behaviour for other inactive coefficients as well. Thus, we conclude for inactive coefficients: (i) The posterior-based method provides efficient inference, where a larger λ leads to better performance; (ii) **BayesM** gives higher coverage for a wide range of λ, compared among the shrinkage methods.

For active coefficients, the posterior-based methods are sometimes sensitive to the choice of λ. For example, at $\tau = 0.75$, the coefficient β_6 is active in the quantile regression model. As in Figure 5.4, we can see that **BayesM** and **BayesF** are both

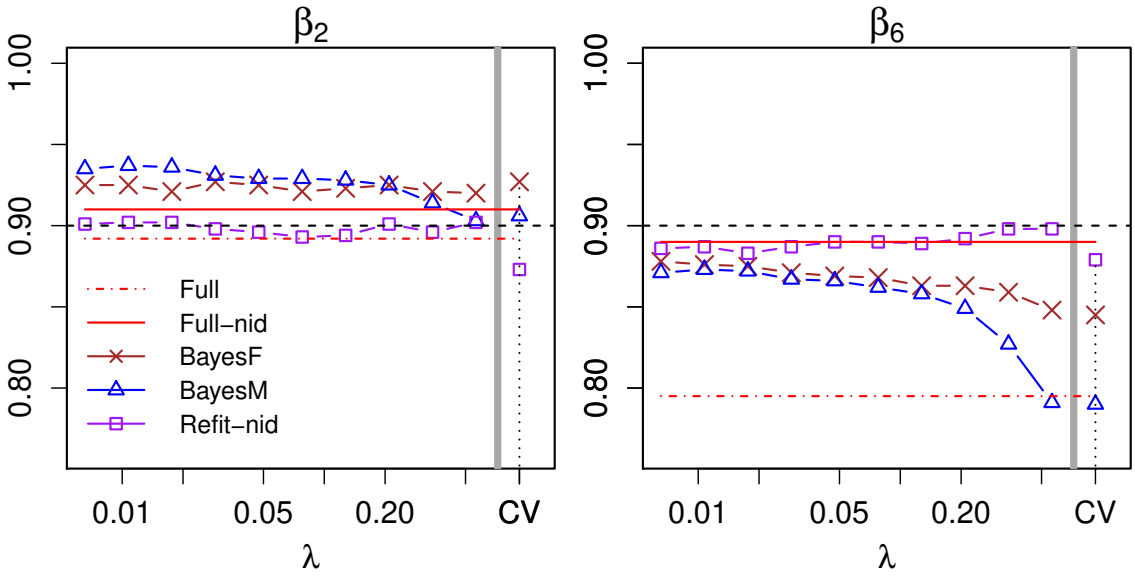


(a) Empirical coverage

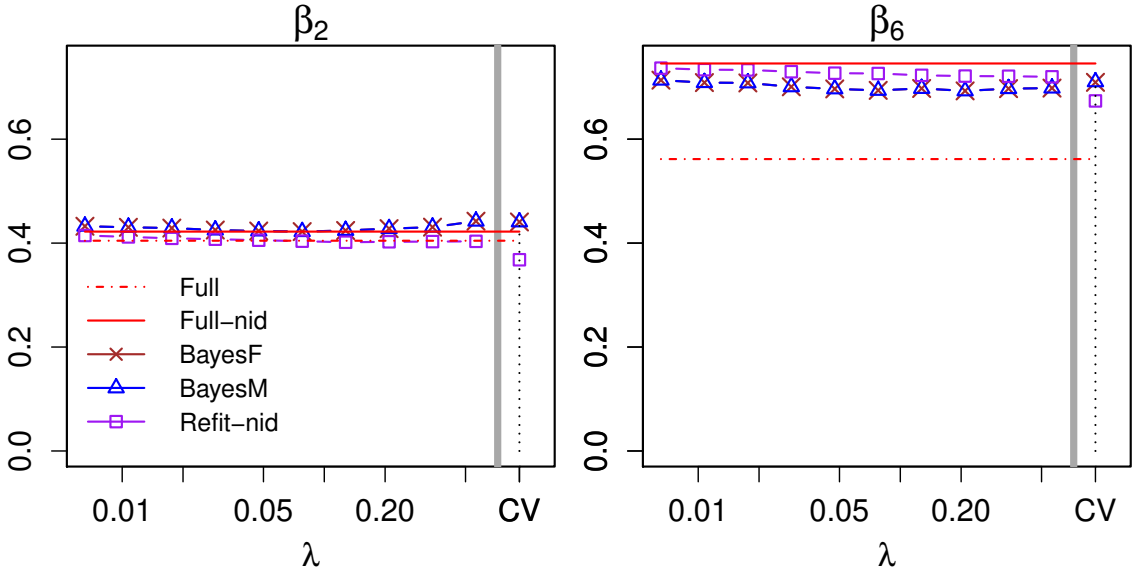


(b) Average length

Figure 5.3: Inference using different λ under model (B) at $\tau = 0.25$. The x-axis is on the log scale, with the largest tick at 1. The true regression coefficients are $\beta_2^0 = 3$ and $\beta_6^0 = 0$. Nominal level is 90%, marked with a black dashed line.



(a) Empirical coverage



(b) Average length

Figure 5.4: Inference using different λ under model (B) at $\tau = 0.75$. The x-axis is on the log scale, with the largest tick at 0.50. The true regression coefficients are $\beta_2^0 = 3$ and $\beta_6^0 = 2.19$. Nominal level is 90%, marked with a black dashed line.

sensitive to the choice of λ: They have satisfactory coverage probability for smaller λ, yet the inference for β_6 breaks down for larger λ. Among the posterior-based methods, **BayesM** seems to be more affected by large λ. For some other active coefficients, however, the posterior-based methods are less sensitive to λ: The inference for β_2 is valid for a wide range of λ, for both quantile levels $\tau = 0.25$ and $\tau = 0.75$. We think the reason may be the following: The covariate x_6 is associated with the heteroscedasticity in model (B), whereas x_2 is not. Through more extensive simulations, we observe the same phenomenon that the heteroscedasticity-related coefficients are more sensitive to the choice of λ.

To get a better picture of why the posterior-based inference deteriorates for β_6 at $\tau = 0.75$, we compare the centers of the intervals from **BayesM** and **BayesF**. Note **BayesM** uses the posterior mean, while **BayesF** uses the Adaptive Lasso estimator. As Figure 5.5 shows, the centers from both methods suffer from non-negligible bias as λ increases; Yet the refitted estimator of **Refit** does not suffer from any variable selection nor penalization bias. Furthermore, we observe the posterior mean is more susceptible to the penalization bias, compared with the Adaptive Lasso estimator. Those penalization bias explains why **BayesM** and **BayesF** are sensitive to λ.

Here we comment on the CV approach for choosing λ. When $\tau = 0.25$, the cross-validated λ seems to deliver satisfactory performances for **BayesM** and **BayesF**, but the chosen λ may not the best. From Figure 5.3, we can see that the CV should have chosen a larger λ: The length of the interval for β_6 can be further reduced; At the same time the inference for β_2 remains valid. When $\tau = 0.75$, the CV fails to select a proper λ: The coverage probabilities for the cross-validated **BayesM** and **BayesF** are both below 85%. This coverage is somewhat disappointing, considering that we have a relatively large sample size $n = 500$. We find the failure of CV persists even when (i) we have larger sample sizes, or (ii) we use more refined CV schemes like the leave-one-out cross validation. Thus, we conclude that CV may not be the best

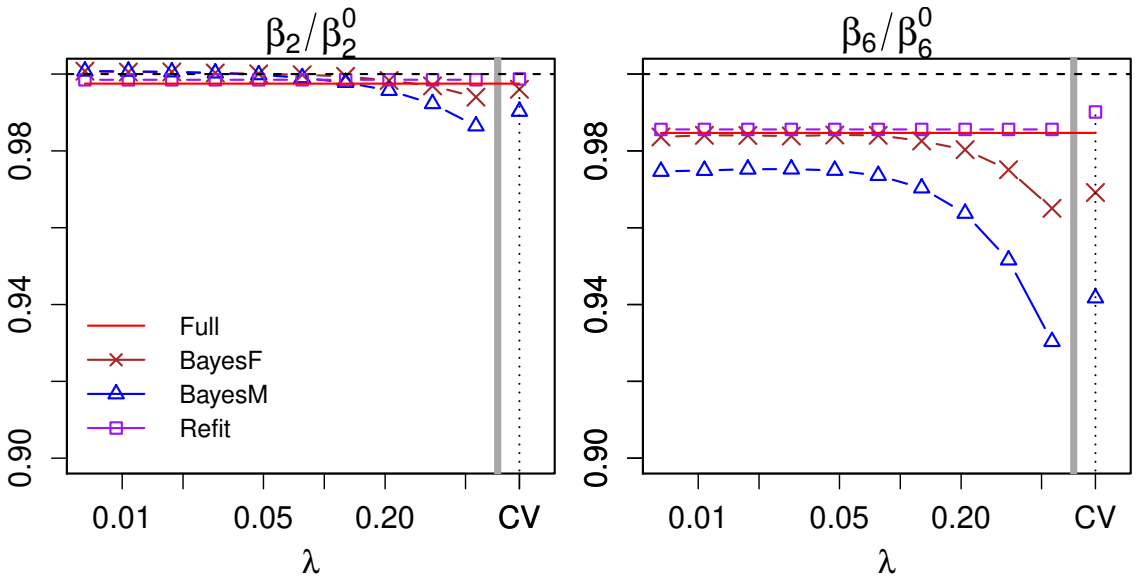


Figure 5.5: Comparison of the relative bias (estimated divided by true values) of different point estimators. The x-axis is on the log scale, with the largest tick at 0.50. The true regression coefficients are $\beta_2^0 = 3$ and $\beta_6^0 = 2.19$. The horizontal dashed line is at 1.

approach for tuning λ, if we want to achieve valid and efficient inference.

As a summary, we find that **BayesM** provides better coverage for inactive coefficients; Whereas **BayesF** is slightly more robust to λ for active coefficients. With suitable values of λ, both **BayesM** and **BayesF** provide adaptive inference under sparse heteroscedastic models, confirming their asymptotic properties in Section 5.4. In the meantime, choosing a proper λ for inference is difficult; The CV is not capable of finding the best λ. For λ's that are too small, the inference is generally valid but inefficient; Larger λ's provide more efficient confidence intervals, but some of those intervals may be invalid. The desired value of λ should strike a balance between efficiency and validity.

5.7.2 When the tuning parameter is fixed

In this section, we further examine the performance of the posterior-based methods in a wide range of settings. For the presentation to be more concise, we shall fix a

single value of λ in each simulation setting below. The particular value of λ is chosen as follows: We run the CV on 100 preliminary datasets generated under the same scenario, which gives 100 cross-validated λ's; Then we fix λ at the 40% quantile of those 100 values. Admittedly, CV is not designed to select the best λ for posterior inference; But here we use that value for simplicity. In general, we find that the selected λ delivers valid inference for shrinkage methods.

5.7.2.1 A sparse homoscedastic model

Here we consider model (A) with

$$\boldsymbol{\beta}_A = (0\,, 3\,, 0\,, -5\,, 0\,, 0)^T.$$

Besides the intercept, only β_2 and β_4 are active in the quantile regression model. Unlike the example in Section 5.7.1.1, all non-zero coefficients are sufficiently separated from 0 here. Here we target three different quantile levels $\tau = 0.3, 0.5, 0.9$; At each τ, we consider two different sample sizes $n = 80$ and $n = 500$. We also present the results with both error distributions as described in model (A): N$(0, 1)$ and Exp(1). Tables 5.1 and 5.2 summarize the results. Here we omit the results of **BayesF** to keep the table concise.

We can see that **Full** delivers consistent coverage probability across all scenarios. For the shrinkage methods, we discuss the inference for active and inactive coefficients separately. For the active coefficients, we observe that **BayesM** is always on the conservative side, especially when the sample size is small. The **BayesM** confidence intervals are sometimes even wider than the **Full** intervals, e.g., in Table 5.2 with $n = 80$ and $\tau = 0.3$. Nonetheless, **BayesM** still provides robust and valid coverage probability. When the sample size increases, **BayesM** delivers more satisfactory performance: The coverage probabilities are closer to 90%, while the average lengths are

Table 5.1: Empirical coverage and average length for 90% confidence intervals under model (A) with N$(0, 1)$ error. The numbers in the parentheses are the empirical standard errors. The row named $\beta_{inactive}$ shows the average over all inactive coefficients β_1, β_3, β_5 and β_6.

		Empirical Coverage			Average Length (s.e.)		
			$n = 80$				
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.3$	β_2	0.88	0.83	0.93	0.65 (0.21)	0.51 (0.18)	0.58 (0.12)
$\lambda =$	β_4	0.89	0.84	0.94	0.65 (0.20)	0.51 (0.17)	0.58 (0.12)
0.140	$\beta_{inactive}$	0.88	0.87	0.97	0.62 (0.20)	0.16 (0.26)	0.17 (0.15)
$\tau = 0.5$	β_2	0.89	0.83	0.94	0.61 (0.19)	0.48 (0.15)	0.54 (0.10)
$\lambda =$	β_4	0.89	0.84	0.94	0.61 (0.19)	0.48 (0.15)	0.54 (0.11)
0.140	$\beta_{inactive}$	0.90	0.88	0.98	0.58 (0.17)	0.16 (0.25)	0.16 (0.14)
$\tau = 0.9$	β_2	0.90	0.82	0.97	0.90 (0.38)	0.67 (0.30)	0.88 (0.22)
$\lambda =$	β_4	0.89	0.81	0.97	0.91 (0.37)	0.67 (0.30)	0.88 (0.22)
0.089	$\beta_{inactive}$	0.90	0.87	0.99	0.85 (0.35)	0.24 (0.37)	0.31 (0.24)
			$n = 500$				
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.3$	β_2	0.90	0.86	0.92	0.25 (0.05)	0.20 (0.05)	0.21 (0.03)
$\lambda =$	β_4	0.90	0.84	0.90	0.25 (0.05)	0.20 (0.05)	0.21 (0.03)
0.067	$\beta_{inactive}$	0.90	0.89	0.97	0.23 (0.05)	0.06 (0.10)	0.05 (0.06)
$\tau = 0.5$	β_2	0.90	0.87	0.92	0.24 (0.05)	0.20 (0.04)	0.20 (0.03)
$\lambda =$	β_4	0.90	0.84	0.90	0.24 (0.05)	0.20 (0.04)	0.20 (0.03)
0.071	$\beta_{inactive}$	0.90	0.89	0.96	0.23 (0.04)	0.05 (0.09)	0.05 (0.05)
$\tau = 0.9$	β_2	0.89	0.84	0.93	0.33 (0.09)	0.26 (0.08)	0.28 (0.05)
$\lambda =$	β_4	0.90	0.84	0.94	0.32 (0.08)	0.27 (0.07)	0.28 (0.05)
0.051	$\beta_{inactive}$	0.90	0.88	0.97	0.31 (0.08)	0.08 (0.13)	0.08 (0.07)

Table 5.2: Empirical coverage and average length for 90% confidence intervals under model (A) with Exp(1) error. Other attributes in the table are the same as Table 5.1.

		Empirical Coverage			Average Length		
			$n = 80$				
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.3$	β_2	0.90	0.85	0.99	0.32 (0.12)	0.25 (0.09)	0.35 (0.07)
$\lambda =$	β_4	0.89	0.83	0.98	0.32 (0.12)	0.25 (0.09)	0.35 (0.07)
0.062	$\beta_{inactive}$	0.89	0.88	0.99	0.30 (0.11)	0.08 (0.14)	0.12 (0.09)
$\tau = 0.5$	β_2	0.89	0.84	0.95	0.46 (0.14)	0.36 (0.13)	0.42 (0.09)
$\lambda =$	β_4	0.89	0.85	0.96	0.46 (0.14)	0.36 (0.12)	0.42 (0.08)
0.102	$\beta_{inactive}$	0.89	0.88	0.98	0.43 (0.14)	0.12 (0.19)	0.14 (0.11)
$\tau = 0.9$	β_2	0.90	0.83	0.95	1.34 (0.54)	1.01 (0.46)	1.19 (0.34)
$\lambda =$	β_4	0.90	0.82	0.95	1.36 (0.55)	1.06 (0.47)	1.23 (0.37)
0.132	$\beta_{inactive}$	0.89	0.87	0.99	1.26 (0.52)	0.39 (0.58)	0.41 (0.35)
			$n = 500$				
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.3$	β_2	0.91	0.87	0.94	0.12 (0.03)	0.10 (0.02)	0.10 (0.01)
$\lambda =$	β_4	0.88	0.85	0.92	0.12 (0.03)	0.10 (0.02)	0.10 (0.01)
0.039	$\beta_{inactive}$	0.90	0.89	0.95	0.11 (0.02)	0.02 (0.04)	0.03 (0.02)
$\tau = 0.5$	β_2	0.89	0.85	0.91	0.19 (0.04)	0.15 (0.03)	0.16 (0.02)
$\lambda =$	β_4	0.90	0.85	0.91	0.19 (0.04)	0.15 (0.03)	0.16 (0.02)
0.058	$\beta_{inactive}$	0.90	0.89	0.96	0.18 (0.04)	0.04 (0.07)	0.04 (0.04)
$\tau = 0.9$	β_2	0.90	0.85	0.91	0.56 (0.15)	0.44 (0.12)	0.47 (0.10)
$\lambda =$	β_4	0.90	0.86	0.91	0.56 (0.14)	0.45 (0.13)	0.47 (0.10)
0.093	$\beta_{inactive}$	0.90	0.89	0.98	0.53 (0.14)	0.12 (0.21)	0.11 (0.11)

more similar to **Refit**. Note, however, **Refit** is usually more liberal with insufficient coverage probability.

For the inactive coefficients, **BayesM** seems to stand out. While the method **Full** is valid, the lengths of the confidence intervals are much wider, compared with **Refit** and **BayesM**. On average, the two shrinkage methods give intervals of similar length, yet **BayesM** has a much higher coverage probability. Note **Refit** never exceeds the nominal 90% coverage for inactive coefficients, although it is claimed to be asymptotically oracle (*Wu and Liu*, 2009). In the meantime, the performance of **BayesM** is also more stable; Because the lengths of **BayesM** intervals have smaller standard errors.

To get a clearer picture of how **Refit** and **BayesM** differ, we zoom in on β_1 and β_3 in Table 5.3; For each coefficient, we compute the average lengths of the interval in two different cases: Depending on whether the Adaptive Lasso selects that coefficient as 0 or not. We only focus on the scenario with $n = 500$ and $\tau = 0.5$. When the Adaptive Lasso identifies β_1 as 0, **Refit** gives perfect inference: Its 'confidence interval' is a singleton for β_1, which coincides with the true coefficient. In the other case when the Adaptive Lasso fails to select β_1 as 0, **Refit** will refit the quantile regression model that includes x_1, with no shrinkage imposed on β_1. Therefore, the intervals will be wider in the latter case, similar to that in **Full**. On the other hand, **BayesM** is more stable in terms of the interval lengths: we get much narrower intervals than **Full** in either of the cases.

To conclude, **BayesM** achieves adaptive inference in this example. By paying the price of being a little conservative, **BayesM** provides robust coverage for the active coefficients. Furthermore, **BayesM** achieves super-efficiency for the inactive coefficients, providing consistently narrower interval than **Full**.

Table 5.3: The average interval lengths for β_1 and β_3, separately for two cases: (i) the Adaptive Lasso (AL) selection is correct for that coefficient; and (ii) the AL selection is incorrect. The column 'Prop. zeros' shows the empirical probability that the AL is correct. The results are for $n = 500$ and $\tau = 0.5$.

		Prop. zeros	AL correct		AL incorrect	
			Refit	BayesM	Refit	BayesM
$u \sim$	β_1	73.8%	0	0.026	0.200	0.110
N(0, 1)	β_3	76.7%	0	0.031	0.228	0.123
$u \sim$	β_1	74.6%	0	0.023	0.157	0.088
Exp(1)	β_3	74.4%	0	0.026	0.173	0.095

5.7.2.2 A dense homoscedastic model

In this section, we consider model (A) with

$$\boldsymbol{\beta}_A = (+3, +3, +3, -3, -3, -3)^T.$$

In this example, all coefficients are active and sufficiently-separated from 0. For simplicity, we only present the results under N(0, 1) errors; We focus on the median regression $\tau = 0.5$ at three different sample sizes $n = 80$, $n = 200$ and $n = 500$. Table 5.4 compares the results for **Full**, **BayesM** and **BayesF**. Here we omit the results for **Refit**, as the Adaptive Lasso always selects the full model. We only compare the results for β_0, β_1, β_2 and β_3.

Even when the model is dense, the shrinkage methods **BayesM** and **BayesF** still provide valid coverage probabilities for the λ values we choose, though they may be more conservative. The conservative nature of posterior-based methods is consistent with our findings in the previous example. Notably, we can see that **BayesM** is always more conservative than **BayesF**. With $n = 500$, **BayesF** provides coverage very close to 90%, whereas **BayesM** is still conservative.

Since the two methods **BayesM** and **BayesF** provide intervals of identical lengths, the difference between their coverage lies in the choice of centers: Recall that **BayesM**

Table 5.4: Empirical coverage and average length for 90% confidence intervals under model (A) with dense coefficients and normal errors. The numbers in the parenthesis are the empirical standard errors. These results are for $\tau = 0.5$.

Sample Size		Empirical Coverage Full	BayesF	BayesM	Average Length (s.e.) Full	BayesF	BayesM
$n = 80$ $\lambda =$ 0.266	β_0	0.89	0.93	0.96	0.46 (0.12)	0.54 (0.07)	0.54 (0.07)
	β_1	0.88	0.93	0.96	0.55 (0.17)	0.64 (0.12)	0.64 (0.12)
	β_3	0.89	0.93	0.94	0.61 (0.18)	0.72 (0.13)	0.72 (0.13)
	β_5	0.90	0.92	0.96	0.61 (0.18)	0.71 (0.13)	0.71 (0.13)
$n = 200$ $\lambda =$ 0.237	β_0	0.88	0.91	0.94	0.29 (0.06)	0.31 (0.03)	0.31 (0.03)
	β_1	0.89	0.92	0.94	0.33 (0.08)	0.36 (0.05)	0.36 (0.05)
	β_3	0.89	0.90	0.92	0.38 (0.09)	0.41 (0.06)	0.41 (0.06)
	β_5	0.88	0.91	0.94	0.37 (0.09)	0.41 (0.06)	0.41 (0.06)
$n = 500$ $\lambda =$ 0.174	β_0	0.88	0.90	0.93	0.18 (0.03)	0.19 (0.02)	0.19 (0.02)
	β_1	0.91	0.91	0.93	0.21 (0.04)	0.22 (0.03)	0.22 (0.03)
	β_3	0.91	0.92	0.93	0.24 (0.05)	0.25 (0.03)	0.25 (0.03)
	β_5	0.90	0.90	0.92	0.23 (0.05)	0.25 (0.03)	0.25 (0.03)

uses the posterior mean, while **BayesF** uses the Adaptive Lasso estimator. Those point estimators have different finite sample properties. Figure 5.6 shows the bias and standard error for the two standardized point estimators. We can see that, the posterior mean and the Adaptive Lasso estimator are both unbiased, but the former has a smaller standard error. The difference in standard error seems to be more obvious when $n = 80$. Therefore, **BayesM** gives more conservative confidence intervals, especially with limited sample sizes.

5.7.2.3 A global conditional quantile model

In this section, we consider model (C) with

$$\boldsymbol{\beta}_C(\tau) = \left(5 \cdot \min\left\{\tau - \frac{3}{4}, 0\right\}, 3 + \tau, 0, -5, 0, 0\right)^T, \quad \tau \in (0, 1),$$

which is similar to the that in *Reich and Smith* (2013). With this choice of $\boldsymbol{\beta}_C(\tau)$, the conditional distribution of y given $\boldsymbol{x}$ is complicated and heteroscedastic; See

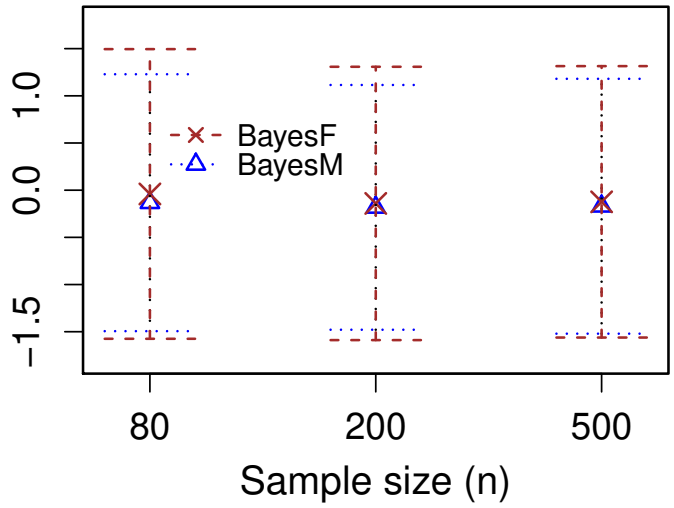


Figure 5.6: The scaled (by $\sqrt{n}$) empirical bias for the two point estimators $\widehat{\beta}_2^{\text{BayesF}}$ and $\widehat{\beta}_2^{\text{BayesM}}$; The error bars show $\pm$ 1 estimated standard error for the scaled bias. The true coefficient is $\beta_2^0 = 3$.

Rodrigues et al. (2019, Figure 4) for an illustration. We target two different quantile levels $\tau = 0.25$, and $\tau = 0.75$; At those quantile levels, the true quantile regression coefficients are

$$\boldsymbol{\beta}_C(0.25) = (-2.5,\, 3.25,\, 0,\, -5,\, 0,\, 0), \quad \text{and} \quad \boldsymbol{\beta}_C(0.75) = (0,\, 3.75,\, 0,\, -5,\, 0,\, 0),$$

respectively. Note β_1 is active when $\tau = 0.25$, yet it is inactive when $\tau = 0.75$. We consider two different sample sizes $n = 200$ and $n = 500$ for each of the quantile level. Table 5.5 presents the results.

In this heteroscedastic example, **BayesM** continues to provide adaptive inference: It gives valid coverage for active coefficients, though the intervals are sometimes wider than the **Full** intervals; It gives near-100% coverage for inactive coefficients, with much shorter intervals than **Full**. On the other hand, **Refit** often fails to provide sufficient coverage probability, especially for the coefficient β_1. The performance of **Refit** does not seem to improve when the sample size grows.

Notably, the **BayesM** intervals for inactive coefficients are even shorter than the

Table 5.5: Empirical coverage and average length for 90% confidence intervals under model (C). The numbers in the parentheses are the empirical standard errors. The row named $\beta_{inactive}$ shows the average over all inactive coefficients β_3, β_5 and β_6.

		Empirical Coverage			Average Length		
			$n = 200$				
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.25$ $\lambda =$ 0.162	β_1	0.88	0.86	0.88	1.53 (0.39)	1.45 (0.39)	1.67 (0.39)
	β_2	0.90	0.87	0.90	1.51 (0.38)	1.42 (0.38)	1.58 (0.35)
	β_4	0.90	0.88	0.92	1.52 (0.39)	1.44 (0.37)	1.59 (0.34)
	$\beta_{inactive}$	0.89	0.88	0.98	1.50 (0.38)	0.54 (0.73)	0.45 (0.46)
$\tau = 0.75$ $\lambda =$ 0.192	β_1	0.87	0.85	0.98	1.57 (0.43)	0.59 (0.79)	0.47 (0.51)
	β_2	0.89	0.87	0.90	1.59 (0.41)	1.46 (0.39)	1.59 (0.35)
	β_4	0.90	0.88	0.91	1.55 (0.39)	1.44 (0.37)	1.58 (0.35)
	$\beta_{inactive}$	0.90	0.89	0.98	1.57 (0.41)	0.50 (0.72)	0.40 (0.43)
			$n = 500$				
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.25$ $\lambda =$ 0.125	β_1	0.86	0.85	0.89	0.95 (0.20)	0.92 (0.19)	1.03 (0.20)
	β_2	0.90	0.88	0.91	0.92 (0.19)	0.89 (0.19)	0.96 (0.19)
	β_4	0.90	0.89	0.91	0.93 (0.18)	0.89 (0.18)	0.96 (0.17)
	$\beta_{inactive}$	0.88	0.87	0.98	0.92 (0.18)	0.32 (0.44)	0.24 (0.25)
$\tau = 0.75$ $\lambda =$ 0.170	β_1	0.87	0.85	0.99	0.99 (0.21)	0.29 (0.46)	0.22 (0.28)
	β_2	0.90	0.88	0.89	0.98 (0.20)	0.92 (0.19)	0.98 (0.17)
	β_4	0.91	0.89	0.91	0.99 (0.20)	0.93 (0.19)	0.99 (0.18)
	$\beta_{inactive}$	0.89	0.88	0.99	0.98 (0.20)	0.26 (0.42)	0.19 (0.23)

Refit intervals on average. Figure 5.7 shows the average interval length for β_1 at $\tau = 0.75$, separately in two cases. When the Adaptive Lasso correctly identifies β_1 as 0, **Refit** gives perfect inference with the 'confidence interval' as a singleton. In the meantime, for more than one-third of the simulated datasets, the Adaptive Lasso does not select β_1 as 0. In those cases, **Refit** provides intervals that are of similar length to that from **Full**; While **BayesM** intervals are still much narrower than the **Full** intervals, on average. In practice, perfect variable selection is rare for the Adaptive Lasso; See *Wang et al.* (2020), *Bühlmann and Van De Geer* (2011, Chapter 7). Therefore, **BayesM** can often help to improve the efficiency for inactive coefficients, even when **Refit** cannot.

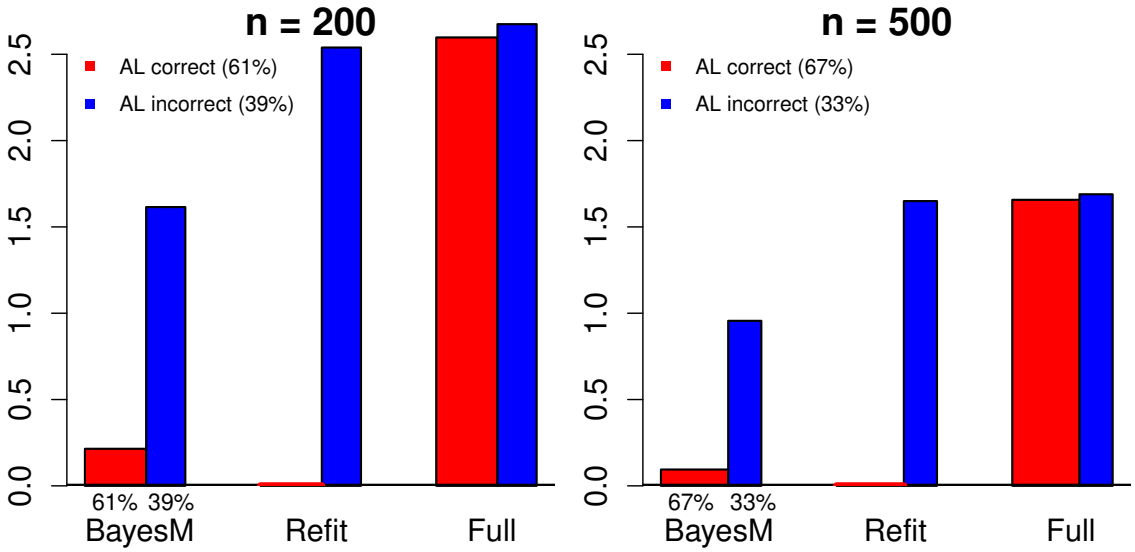


Figure 5.7: The average interval lengths for β_1 separately in two cases: (i) the Adaptive Lasso (AL) selection is correct for β_1, shown in red; and (ii) the AL selection is incorrect, shown in blue. The results are for $\tau = 0.75$, where β_1 is inactive.

5.7.2.4 A sparse model with increasing dimensions

This section considers model (D) in a series of experiments where the dimension p grows with the sample size n. We choose three different combinations of the pair (n, p): $(140, 10)$, $(350, 20)$ and $(700, 30)$, following the growth rate $n = 0.7p^2 + 70$.

For a given dimension p, we set

$$\beta_D = (1.5,\, 0\underbrace{,\ldots,\, 0}_{\lfloor p/2\rfloor-2},\, 3,\, -5,\, 2,\, 0\underbrace{,\ldots,\, 0}_{\lceil p/2\rceil-2})^T.$$

Besides the intercept, there are 4 active coefficients in the quantile regression model, with all others being inactive. Note in this example, there is a relatively high correlation among all covariates, with $\text{corr}(x_i, x_j) = 0.85^{|i-j|}$. Table 5.6 summarizes the results. We only present: (i) the result for the intercept, (ii) the average result over the 4 active coefficients, and (iii) the average result over all inactive coefficients.

We can see that **BayesM** continues to provide valid and adaptive inference, similar to what we see in previous examples. Two comments are in place to explain some new observations in this example. First, the performance of **Refit** is consistently poor throughout Table 5.6. When there are many covariates, the Adaptive Lasso rarely achieve consistent variable selection in finite sample, which leads to invalid inference (*Leeb and Pötscher*, 2005). **BayesM** seems to be a safer alternative than **Refit**: For active coefficients, **BayesM** achieves nominal coverage, at the cost of wider intervals; For inactive coefficients, **BayesM** provides near-100% coverage with narrower intervals on average, compared with **Refit**.

Second, **BayesM** offers narrower intervals for the active coefficients, compared with **Full**. We do not observe such an obvious improvement in previous examples, where there is a small number of covariates. In this example, there are many highly correlated covariates; Therefore, **Full** becomes inefficient when it uses all the covariates. **BayesM**, on the other hand, implicitly uses a smaller number of covariates, as **BayesM** shrinks some inactive coefficients toward zero. Thus, **BayesM** is close to the 'oracle efficiency' (*Zou*, 2006) for the active coefficients, as if we knew the true model in advance.

In summary, if we have enough samples, **BayesM** can deliver valid inference for

Table 5.6: Empirical coverage and average length for 90% confidence intervals under model (D). The numbers in the parenthesis are the empirical standard errors. The row named β_{active} and $\beta_{inactive}$ shows the average over all active or inactice coefficients in the slope, respectively.

		Empirical Coverage			Average Length (s.e.)		
$n = 140$, $p = 10$							
τ		Full	Refit	BayesM	Full	Refit	BayesM
$\tau = 0.3$	β_0	0.89	0.85	0.92	0.50 (0.13)	0.44 (0.12)	0.53 (0.09)
$\lambda =$	β_{active}	0.89	0.83	0.92	1.14 (0.33)	0.94 (0.30)	1.06 (0.22)
0.100	$\beta_{inactive}$	0.89	0.85	0.97	1.17 (0.34)	0.36 (0.48)	0.35 (0.28)
$\tau = 0.5$	β_0	0.91	0.87	0.94	0.41 (0.10)	0.37 (0.08)	0.43 (0.05)
$\lambda =$	β_{active}	0.89	0.84	0.92	0.98 (0.26)	0.81 (0.24)	0.90 (0.16)
0.101	$\beta_{inactive}$	0.90	0.86	0.97	1.00 (0.27)	0.29 (0.40)	0.28 (0.23)
$\tau = 0.9$	β_0	0.86	0.79	0.89	1.32 (0.57)	1.15 (0.52)	1.40 (0.43)
$\lambda =$	β_{active}	0.90	0.83	0.89	2.46 (0.93)	1.96 (0.86)	2.21 (0.68)
0.090	$\beta_{inactive}$	0.90	0.85	0.97	2.50 (0.94)	0.96 (1.15)	0.93 (0.72)
$n = 350$, $p = 20$							
$\tau = 0.3$	β_0	0.87	0.84	0.92	0.31 (0.07)	0.27 (0.06)	0.32 (0.04)
$\lambda =$	β_{active}	0.90	0.83	0.92	0.73 (0.17)	0.56 (0.15)	0.62 (0.10)
0.106	$\beta_{inactive}$	0.89	0.86	0.97	0.76 (0.18)	0.14 (0.24)	0.14 (0.12)
$\tau = 0.5$	β_0	0.89	0.84	0.93	0.26 (0.05)	0.23 (0.04)	0.27 (0.03)
$\lambda =$	β_{active}	0.89	0.82	0.91	0.62 (0.13)	0.48 (0.12)	0.53 (0.08)
0.102	$\beta_{inactive}$	0.89	0.86	0.97	0.65 (0.14)	0.12 (0.20)	0.11 (0.10)
$\tau = 0.9$	β_0	0.85	0.79	0.88	0.80 (0.25)	0.66 (0.21)	0.83 (0.18)
$\lambda =$	β_{active}	0.90	0.80	0.89	1.58 (0.46)	1.18 (0.40)	1.40 (0.32)
0.089	$\beta_{inactive}$	0.90	0.83	0.98	1.65 (0.49)	0.43 (0.58)	0.42 (0.34)
$n = 700$, $p = 30$							
$\tau = 0.3$	β_0	0.90	0.87	0.93	0.22 (0.04)	0.19 (0.03)	0.22 (0.02)
$\lambda =$	β_{active}	0.90	0.83	0.91	0.51 (0.10)	0.39 (0.09)	0.42 (0.06)
0.090	$\beta_{inactive}$	0.90	0.87	0.97	0.54 (0.11)	0.08 (0.15)	0.08 (0.07)
$\tau = 0.5$	β_0	0.88	0.84	0.92	0.18 (0.03)	0.16 (0.02)	0.18 (0.01)
$\lambda =$	β_{active}	0.90	0.83	0.91	0.43 (0.08)	0.34 (0.07)	0.36 (0.04)
0.084	$\beta_{inactive}$	0.90	0.88	0.97	0.45 (0.09)	0.07 (0.13)	0.07 (0.06)
$\tau = 0.9$	β_0	0.84	0.79	0.88	0.56 (0.14)	0.45 (0.11)	0.54 (0.09)
$\lambda =$	β_{active}	0.91	0.81	0.90	1.12 (0.28)	0.82 (0.23)	0.96 (0.18)
0.091	$\beta_{inactive}$	0.91	0.85	0.98	1.18 (0.30)	0.23 (0.36)	0.21 (0.19)

models with many covariates, whereas **Refit** oftentimes can not; With many highly-correlated covariates, **BayesM** offers efficiency gain for both the active and inactive coefficients, compared with **Full**.

5.7.3 A summary of the simulation studies

Here we summarize our findings from all those simulation studies. Overall, we confirm that **BayesM** provides valid and adaptive inference in a wide range of settings. With a suitable λ, **BayesM** achieves nominal coverage probability for active coefficients, though it is sometimes conservative; **BayesM** also gives near-100% coverage for inactive coefficients, with much shorter intervals than the analysis using the full model.

First, we compare two posterior-based methods **BayesM** and **BayesF**. For inactive coefficients, **BayesM** has much higher coverage for a wide range of λ, whereas the coverage for **BayesF** is often insufficient. For active coefficients, **BayesF** is slightly more robust to the penalization bias when λ is too large.

Second, **BayesM** is more stable than **Refit**, especially for the inactive coefficients. Furthermore, **Refit** often fails to provide sufficient coverage.

Third, we comment on the role of the tuning parameter λ. Often, there is a wide range of λ that can achieve a balance between (i) valid coverage probability for active coefficients, and (ii) better efficiency than **Full** for inactive coefficients. When λ is too large, however, the inference may be incorrect.

Finally, for tuning λ, CV sometimes chooses a λ that is too large to deliver valid inference for active coefficients. We suggest using a smaller λ than that chosen from CV.

5.8 Discussion

In this Chapter, we show that the Bayesian computational framework can be useful for constructing frequentist confidence intervals in possibly sparse quantile regression analysis. By employing appropriate shrinkage priors, we show the posterior inference can adapt automatically to model sparsity. Asymptotically, the proposed confidence intervals are oracle efficient for the active coefficients, and are super-efficient for the inactive coefficients. Our work helps to uncloak the value of Bayesian computational methods in frequentist inference with a mis-specified likelihood.

The proposed pseudo-Bayesian inference enjoys two distinct advantages over other commonly-used frequentist approaches based on variable selection: (i) it avoids the need to pursue dichotomous variable selection which is often non-oracle in finite-sample problems; (ii) it avoids direct (non-parametric) estimation of the nuisance-parameter needed for frequentist inference. These two properties often lead to more stable results for quantile regression inference. In addition, the Bayesian computational framework can be especially valuable in other complex settings, e.g., censored quantile regression problems (*Yang et al.*, 2016; *Wu and Narisetty*, 2021) where the objective function can be highly non-convex (*Powell*, 1986). Our pseudo-Bayesian approach can be used to produce statistical inference without direct optimization of the objective function while incorporating possible model sparsity.

There are several limitations of our work on the pseudo-Bayesian framework. First, we focus on problems with fixed or moderately increasing dimensions. Second, we use two relatively simple shrinkage priors as examples, which do not easily generalize to high-dimensional settings. It remains an interesting problem, however, to study what the pseudo-Bayesian approach can offer in higher dimensions when coupled with other hierarchical shrinkage priors.

5.9 Technical details

We first review some common notations, which shall appear throughout the appendix. Let P^* be the true data generating probability measure, and let E^* be the expectation under P^*. The posterior probability is $\Pi(\cdot|\mathbb{D}_n)$, where $\mathbb{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Conversely, we shall use Pr for generic probability calculations.

For a vector a, let $\|a\|$ be its L_2 norm, and $\|a\|_q$ be its L_q norm for $1 \leq q \leq \infty$. For any symmetric matrix A, we define $\theta_{\max}(A)$ and $\theta_{\min}(A)$ as the maximal/minimal eigenvalue of A; and let $|A|$ be its determinant. For p-by-p symmetric matrices A and B, we write $A \preceq B$ if $a^T A a \leq a^T B a$ for all $a \in \mathbb{R}^p$. For two probability density functions f and g, we define $\|f - g\|_{TV}$ as their total variation distance. For two real numbers a and b, let $a \wedge b = \min\{a, b\}$, and $a \vee b = \max\{a, b\}$. For two sequences a_n and b_n, we define $a_n \ll b_n$ if $a_n/b_n \to 0$; and $a_n \lesssim b_n$ if there is a universal constant $C_0 > 0$, such that $a_n \leq C_0 \cdot b_n$. We define $a_n \leq_{\mathrm{P}^*} b_n$ if the inequality holds with P^*-probability tending to 1.

5.9.1 Some preliminary lemmas

Let $\chi^2_d(\nu)$ represent the chi-square distribution with d degrees of freedom and non-centrality parameter ν; let $Laplace(b)$ represent the Laplace distribution with density function

$$f_b(x) = \frac{1}{2b} \exp\left\{-\frac{|x|}{b}\right\}, \quad x \in \mathbb{R};$$

and let $\boldsymbol{N}(\boldsymbol{\mu}, \Sigma)$ represent the multivariate normal distribution. We first present Lemma IA.1 – Lemma IA.3 regarding the properties for those distributions.

Lemma IA.1. *Let $X \sim \chi^2_d(\nu)$, then for all $x \geq 4(d + 2\nu)$, we have*

$$P(X \geq x) \leq \exp(-x/4).$$

Furthermore, let $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^d$*, then if* $x^2 \geq 4\theta_{\max}(\Sigma) \cdot (d + 2\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})$*, we have*

$$P(\|\boldsymbol{Z}\| \geq x) \leq \exp\left(-\frac{x^2}{4\theta_{\max}(\Sigma)}\right).$$

Proof. The first inequality follows from Lemma 8.1 of *Birgé* (2001). To show the second inequality, note that $\boldsymbol{Z}^T\Sigma^{-1}\boldsymbol{Z} \sim \chi^2_d(v)$, where $v = \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}$; therefore if $x^2\theta_{\min}(\Sigma^{-1}) \geq 4(d + 2v)$, we have

$$\Pr(\|\boldsymbol{Z}\| \geq x) \leq \Pr\left(\boldsymbol{Z}^T\Sigma^{-1}\boldsymbol{Z} \geq x^2\theta_{\min}(\Sigma^{-1})\right) \leq \exp\{-x^2/(4\theta_{\max}(\Sigma))\}.$$

□

Lemma IA.2. *Given* $a_1, \ldots, a_m \geq a_{\min} > 0$*, let* $X_1, \ldots, X_m$ *be independent random variables with*

$$X_j \sim \textit{Laplace}\left(\frac{1}{a_j}\right), \quad j = 1, \ldots m,$$

Then, for all $x > 0$*, we have*

$$\Pr\left(\max_{j=1,\ldots m} |X_j| \geq x\right) \leq m \cdot \exp\{-a_{\min} \cdot x\};$$

furthermore, if $a_{\min}/b > 1/2$*, then*

$$\Pr\left(\max_{j=1,\ldots m} |X_j| \geq \frac{x}{b}\right) \leq m \cdot \exp\{-x/2\}.$$

Proof. The first inequality follows from a standard union bound, since

$$\Pr(|X_j| \geq x) = 2\int_x^\infty \frac{a_j}{2}\exp\{-a_j|u|\}\,\mathrm{d}u \leq \exp\{-a_{\min} \cdot x\}.$$

The second inequality follows similarly, since

$$\Pr(|X_j| \geq x/b) = \exp\{-a_{\min}x/b\}.$$

□

Lemma IA.3. *Let* $\boldsymbol{w} = (w_1, \ldots, w_s)^T$*, and let* $\boldsymbol{X} \in \mathbb{R}^s$ *be distributed as*

$$\boldsymbol{X} \sim N\left(\boldsymbol{\mu}, \frac{\sigma_0^2}{n}\boldsymbol{I}_s\right).$$

For any positive integer k *and any* $0 < \varepsilon < 1/2$*, if*

$$\|\boldsymbol{w}\| \leq \varepsilon \cdot \min\left\{\frac{\sqrt{2n}}{k\sigma_0}, \frac{1}{k\|\boldsymbol{\mu}\|}\right\},$$

then we have

$$\left|\mathrm{E}_X\left(\exp\{-k \cdot \boldsymbol{w}^T X\}\right) - 1\right| \leq 4\varepsilon.$$

Furthermore, if a constant K *satisfy* $K \geq 3\sigma_0$ *and* $K^2 s \geq 16n\|\boldsymbol{\mu}\|^2$*, we have*

$$\mathrm{E}_X\left(\exp\{-k \cdot \boldsymbol{w}^T X\} \cdot \mathbf{1}\left[\|X\| \geq K\sqrt{\frac{s}{n}}\right]\right) \lesssim \exp\left(-\frac{K^2 s}{8\sigma_0^2}\right).$$

Proof. By leveraging the moment generating function of the normal distribution, and using the upper bound for $\|\boldsymbol{w}\|$,

$$\begin{aligned}
\mathrm{E}\left(\exp\{-k \cdot \boldsymbol{w}^T \boldsymbol{X}\}\right) &= \exp\left\{k\boldsymbol{\mu}^T\boldsymbol{w} + \frac{k\sigma_0^2}{2n}\|\boldsymbol{w}\|^2\right\} \\
&\leq \exp\{2\varepsilon\} \\
&\leq 1 + 4\varepsilon,
\end{aligned}$$

for $0 < \varepsilon < 1/2$. In a similar manner, we can establish the lower bound for

$\mathrm{E}\left(\exp\{-k\cdot\boldsymbol{w}^T\boldsymbol{X}\}\right)$, which shows the first result.

For the second inequality, using Cauchy-Schwartz inequality gives

$$\begin{aligned}
& \mathrm{E}^2\left(\exp\{-k\cdot\boldsymbol{w}^T\boldsymbol{X}\}\cdot\mathbf{1}\left[\|\boldsymbol{X}\|\geq K\sqrt{\frac{s}{n}}\right]\right)\\
\leq\ & \mathrm{E}\left(\exp\{-2k\cdot\boldsymbol{w}^T\boldsymbol{X}\}\right)\cdot\Pr\left(\|\boldsymbol{X}\|\geq K\sqrt{\frac{s}{n}}\right)\\
\leq\ & \exp\{4\varepsilon\}\cdot\exp\left\{-\frac{K^2 s}{4\sigma_0^2}\right\},
\end{aligned}$$

where the tail probability is bounded by Lemma IA.1. Taking the square root of the above inequality gives the desired result. □

Furthermore, the following result is needed for Example 3 in Section 5.5.

Lemma IA.4. *Let $\alpha_k \asymp 1/k$, and*

$$D_p = \left(\begin{array}{c|ccc} \sum_{k=1}^{\infty}\alpha_k^2 & \alpha_1 & \cdots & \alpha_p \\ \hline \alpha_1 & & & \\ \vdots & & I_p & \\ \alpha_p & & & \end{array}\right).$$

Then the eigenvalues of D_p satisfy:

$$p^{-1}\lesssim\theta_{\min}(D_p)\leq\theta_{\max}(D_p)\lesssim p,$$

as $p\to\infty$.

Proof of Lemma IA.4. Let $A_p=\sum_{k=1}^{p}\alpha_k^2$, $A=\sum_{k=1}^{\infty}\alpha_k^2$ and $\boldsymbol{b}=(\alpha_1,\ldots,\alpha_p)^T$. Any eigenpair of D_p, denoted by $(\lambda,\boldsymbol{u})$, satisfies:

$$\begin{aligned}
Au_0+\boldsymbol{b}^T\boldsymbol{u}_1&=\lambda u_0\\
u_0\boldsymbol{b}+\boldsymbol{u}_1&=\lambda\boldsymbol{u}_1.
\end{aligned}$$

From simple linear algebra, we have either $\lambda = u_0 = 1$, or $\boldsymbol{u}_1 = [u_0/(\lambda - 1)] \cdot \boldsymbol{b}$.

It suffices to consider the case where $\lambda \neq 1$. From the first line of the above displayed equations we have

$$A + \frac{A_p}{\lambda - 1} = \lambda,$$

since $\boldsymbol{u}_1 = [u_0/(\lambda - 1)] \cdot \boldsymbol{b}$ and $u_0 \neq 0$. Therefore it follows from simple algebra that

$$-\sqrt{(A+1)^2 - 4(A - A_p)} \lesssim \lambda - (A+1) \lesssim \sqrt{(A+1)^2 - 4(A - A_p)}.$$

Noting that $A - A_p \asymp 1/p$, the above displayed inequality shows that all eigenvalues of D_p are upper bounded by a constant, and lower bounded by a multiple of p^{-1}. Hence the proof is now complete □

The following lemma is simple but useful; we will implicitly use the lemma in the upcoming proofs.

Lemma IA.5. *Let $f(z;\theta)$ be a probability density function indexed by $\theta \in \Theta \subset \mathbb{R}^k$. We write $Z \sim f(z;\theta)$ and define $\Pr_\theta(Z \geq x) = \int_{z \geq x} f(z;\theta) \mathrm{d}z$, where Z is independent of the data. Let $g(\cdot,\cdot)$ be a bivariate function of $\Theta \times \mathbb{R} \to \mathbb{R}$; suppose we have*

$$\sup_{\theta:\, g(\theta,x) \leq B} \Pr_\theta(X \geq x) \leq a,$$

for some real numbers a, x, and B. For any statistic θ_n that satisfies $g(\theta_n, x) \leq_{\mathrm{P}^*} B$, it holds that

$$\Pr_{\theta_n}(X \geq x) \leq_{\mathrm{P}^*} a.$$

We need the following variants of the Bernstein inequality. They are Theorem 2.10 of *Boucheron et al.* (2013) and Theorem 2.8.2 of *Vershynin* (2018), respectively.

Lemma IA.6. *Let $X_1, \ldots, X_n$ be independent random variables. Suppose there exist*

positive constants $c, v > 0$ *such that* $\sum_{i=1}^{n} \mathrm{E}(X_i^2) \leq v$, *and*

$$\sum_{i=1}^{n} \mathrm{E}\left[(X_i^q)_+\right] \leq \frac{q!}{2} v c^{q-2} \quad \textit{for all integers } q > 3.$$

Then

$$\mathrm{P}\left(\sum_{i=1}^{n} \left[X_i - E(X_i)\right] \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(v+ct)}\right).$$

Lemma IA.7. *Let* $X_1, \ldots, X_n$ *be independent, mean zero random variables that satisfy*

$$\sup_{i=1,\ldots,n} \Pr\left(|X_i| \geq x\right) \leq \exp(-x/\sigma_0),$$

for some constant σ_0. *Then there is a universal constant* C_2, *such that for every* $t \geq 0$ *and* $\boldsymbol{a} = (a_1, \ldots, a_n)$,

$$\Pr\left(\left|\sum_{i=1}^{n} a_i X_i\right| \geq t\right) \leq 2\exp\left(-C_2 \cdot \min\left\{\frac{t^2}{\sigma_0^2 \|\boldsymbol{a}\|_2^2}, \frac{t}{\sigma_0 \|\boldsymbol{a}\|_\infty}\right\}\right).$$

We also need the following lemma, which is from Lemma 4 of *Belloni and Chernozhukov* (2011).

Lemma IA.8 (Expectation of log-likelihood)**.** *Let* $L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \boldsymbol{\beta})$ *and* $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$ *Suppose Assumptions E.2 and E.3′ hold, then there exists a constant* q_0 *such that*

$$\frac{1}{n}\mathrm{E}^*\left[L_n(\boldsymbol{\beta}^0 + \boldsymbol{\delta}) - L_n(\boldsymbol{\delta})\right] \geq \min\left\{\frac{\|G^{1/2}\boldsymbol{\delta}\|^2}{4}, q_0 \|G^{1/2}\boldsymbol{\delta}\|\right\}.$$

In particular, when $\|G^{1/2}\boldsymbol{\delta}\| \geq 4q_0$, *the left-hand-side of the displayed equation is lower bounded by* $q_0 \|G^{1/2}\boldsymbol{\delta}\|$.

5.9.2 Technical lemmas with increasing dimensions

In this subsection, we state and prove two key results, Lemma IB.1 and IB.2, which controls the uniform variation of the empirical quantile-loss function. When the dimension p grows with the sample size n, the results are new; they are not implied by generic combinatoric arguments (*Belloni et al.*, 2019a) under our current conditions. When p is fixed, those lemmas are standard from the empirical process literature; see e.g., *Knight* (1998), *Pollard* (1985) and *Andrews* (1994).

Before stating the lemmas, we first review some notations. We shall continue to use the notations in the beginning of the appendix. In addition, recall x_i, y_i and $\boldsymbol{\beta}^0$ from the quantile regression model (5.1); and recall the quantile-loss function $\rho_\tau(\cdot)$ and $L_n(\cdot)$ from (5.2). Let $\boldsymbol{X} = [x_1, \ldots, x_n]^T$ be the design matrix. We define $\phi_\tau(u) = \tau - \mathbf{1}[u \leq 0]$, and we shall write $\boldsymbol{\phi} = [\phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0)]_{i=1}^n$ as a vector. Recall from Assumption E.3 that $G = \mathrm{E}^*[x_i x_i^T f_{y|x}(x_i^T\boldsymbol{\beta}^0)]$, $D = \mathrm{E}^*[x_i x_i^T]$, as well as the block-partition

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where G_{11} is an s-by-s matrix corresponding to the active coefficients under Assumption E.5′. Furthermore, for any vector $a \in \mathbb{R}^p$, we shall write $a = (a_1^T, a_2^T)^T$, where $a_1 \in \mathbb{R}^s$ corresponds to the active components.

Now we give the key lemmas and their proofs.

Lemma IB.1 (Stochastic Differentiability)**.** *Let* $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$ *and*

$$r_n(\boldsymbol{\delta}) = L_n(\boldsymbol{\beta}^0 + \boldsymbol{\delta}) - L_n(\boldsymbol{\beta}^0) + \sum_{i=1}^n \phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0)x_i^T\boldsymbol{\delta}.$$

Suppose Assumptions E.1, E.2, E.3′ and E.4′ hold and $p^2 \log^2 p = o(n)$*. Then we have that*

$$\sup_{\boldsymbol{\delta}\in\mathbb{R}^p} \left| \frac{r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]}{n\|D^{1/2}\boldsymbol{\delta}\| + 1} \right| \xrightarrow{P^*} 0.$$

Proof. First, it is easy to see $|r_n(\boldsymbol{\delta})| \le \sum_{i=1}^n |x_i^T\boldsymbol{\delta}|$. Therefore, for large enough $a > 0$, we have that

$$\begin{aligned}\sup_{\|D^{1/2}\boldsymbol{\delta}\|\ge n^a}\left|\frac{r_n(\boldsymbol{\delta})-\mathrm{E}^*[r_n(\boldsymbol{\delta})]}{n\|D^{1/2}\boldsymbol{\delta}\|+1}\right| &\le \sup_{\|D^{1/2}\boldsymbol{\delta}\|\ge n^a}\frac{\sum_{i=1}^n|x_i^T\boldsymbol{\delta}|+\sum_{i=1}^n\mathrm{E}^*\left[|x_i^T\boldsymbol{\delta}|\right]}{n\cdot\|D^{1/2}\boldsymbol{\delta}\|^2}\\ &\le \frac{\sum_{i=1}^n\|D^{-1/2}x_i\|}{n^{1+a}}+\frac{\mathrm{E}^*\left[\|D^{-1/2}x_i\|\right]}{n^a}\\ &\xrightarrow{P^*} 0,\end{aligned}$$

where the last inequality follows since $\mathrm{Cov}^*(D^{-1/2}x_i) = \boldsymbol{I}_p$. Therefore, it suffices to show

$$\sup_{\boldsymbol{\delta}:\|D^{1/2}\boldsymbol{\delta}\|\le n^a}\left|\frac{r_n(\boldsymbol{\delta})-\mathrm{E}^*[r_n(\boldsymbol{\delta})]}{n\|D^{1/2}\boldsymbol{\delta}\|+1}\right|\xrightarrow{P^*}0,$$

for any constant $a > 0$.

Let $\gamma_n^4 = (p^2\log^2 n)/n \to 0$. In the following steps, we apply a generic chaining argument to show that the above display is of order $O_{P^*}(\gamma_n)$. To simplify notations, we define $f_n(\boldsymbol{\delta}) = r_n(\boldsymbol{\delta})/(n\|D^{1/2}\boldsymbol{\delta}\|^2+1)$.

Step I: Main chaining First, we define the following concentric 'cubes':

$$\mathcal{C}_k = \{\boldsymbol{\delta}\in\mathbb{R}^p : \|D^{1/2}\boldsymbol{\delta}\|_\infty \le d_k\},\quad k = 0,\dots,K_n,$$

where d_k is the edge length of each cube. For $0 < \gamma_n < 1$, we take the lengths to be

$$d_k = (k+1)\varepsilon_n,\quad \text{with}\quad \varepsilon_n = \frac{\gamma_n}{np},\quad k = 0,\dots,K_n.$$

Letting $K_n = \lceil\frac{pn^{a+1}}{\gamma_n}\rceil - 1$, it is easy to check $\{\|D^{1/2}\boldsymbol{\delta}\|\le n^a\}\subset\mathcal{C}_{K_n}$. It then suffices to show the uniform convergence in $\mathcal{C}_{K_n}$.

For each of $\mathcal{C}_k\backslash\mathcal{C}_{k-1}$, we further partition it into smaller cubes of length at most ε_n. That is, $\mathcal{C}_k$ builds upon $\mathcal{C}_{k-1}$ by one layer of such small cubes with edge ε_n. For

each $k \geq 1$, there are $B_k = (2k)^p - (2(k-1))^p$ such smaller cubes, which are denoted as $\mathcal{C}_k^j$, $j = 1, \ldots, B_k$. Letting $\boldsymbol{\delta}_k^j$ be the center of $\mathcal{C}_k^j$, we have

$$\begin{aligned}
& \sup_{\boldsymbol{\delta}:\|D^{1/2}\boldsymbol{\delta}\| \leq n^a} |f_n(\boldsymbol{\delta}) - \mathrm{E}^*(f_n(\boldsymbol{\delta}))| \\
\leq & \sup_{\boldsymbol{\delta} \in \mathcal{C}_0} |f_n(\boldsymbol{\delta}) - \mathrm{E}^*(f_n(\boldsymbol{\delta}))| + \sup_{\substack{k=1,\ldots,K_n \\ j=1,\ldots,B_k}} \sup_{\boldsymbol{\delta} \in \mathcal{C}_k^j} |f_n(\boldsymbol{\delta}) - \mathrm{E}^*(f_n(\boldsymbol{\delta}))| \\
\leq & \left(\sup_{\boldsymbol{\delta} \in \mathcal{C}_0} |f_n(\boldsymbol{\delta}) - \mathrm{E}^*(f_n(\boldsymbol{\delta}))| \right) \\
& + \left(\sup_{\substack{k=1,\ldots,K_n \\ j=1,\ldots,B_k}} \sup_{\boldsymbol{\delta} \in \mathcal{C}_k^j} \left[|f_n(\boldsymbol{\delta}) - f_n(\boldsymbol{\delta}_k^j)| + \mathrm{E}^*|f_n(\boldsymbol{\delta}) - f_n(\boldsymbol{\delta}_k^j)| \right] \right) \\
& + \left(\sup_{\substack{k=1,\ldots,K_n \\ j=1,\ldots,B_k}} |f_n(\boldsymbol{\delta}_k^j) - \mathrm{E}^*[f_n(\boldsymbol{\delta}_k^j)]| \right) \\
\triangleq & \; R_1 + R_2 + R_3.
\end{aligned}$$

In the next step, we shall compute the stochastic order of R_1, R_2, and R_3 separately.

Step II: Auxiliary chaining Let $v(\boldsymbol{\delta}) = \|D^{1/2}\boldsymbol{\delta}\|^2$; for any $\boldsymbol{\delta}, \boldsymbol{\delta}' \neq \mathbf{0}$, define

$$\begin{aligned}
\Delta_0(\boldsymbol{\delta}) &= |f_n(\boldsymbol{\delta})| = \frac{|r_n(\boldsymbol{\delta})|}{n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1}, \\
\Delta_1(\boldsymbol{\delta}, \boldsymbol{\delta}') &= \frac{|r_n(\boldsymbol{\delta}') - r_n(\boldsymbol{\delta})|}{n\|D^{1/2}\boldsymbol{\delta}'\|^2 + 1}, \\
\Delta_2(\boldsymbol{\delta}, \boldsymbol{\delta}') &= \left| r_n(\boldsymbol{\delta}) \cdot \frac{1}{n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1} - \frac{1}{n\|D^{1/2}\boldsymbol{\delta}'\|^2 + 1} \right|.
\end{aligned} \tag{IB.1}$$

In step IV below, we show separately that for any $d < 1/\sqrt{n}$,

$$\mathrm{E}^* \left[\sup_{\|D^{1/2}\boldsymbol{\delta}\| \leq d} \Delta_0(\boldsymbol{\delta}) \right] \lesssim npd^2,$$

and

$$\mathrm{E}^* \left[\sup_{\substack{\boldsymbol{\delta},\boldsymbol{\delta}' \subset \mathcal{C}_{K_n} \\ \|D^{1/2}(\boldsymbol{\delta}'-\boldsymbol{\delta})\| \le d}} \Delta_u(\boldsymbol{\delta},\boldsymbol{\delta}') \right] \lesssim n\sqrt{p}d,$$

for $u = 1, 2$.

Now, we show that we can control R_1 and R_2 using Δ_0 through Δ_2 defined in (IB.1). Recall $d_0 = \varepsilon_0 = \gamma_n/(np)$. For R_1, note when $\boldsymbol{\delta} \in \mathcal{C}_0$ we have $\|D^{1/2}\boldsymbol{\delta}\| \le \sqrt{p}\|D^{1/2}\boldsymbol{\delta}\|_\infty \le \sqrt{p}d_0 \le 1/\sqrt{n}$, therefore

$$\mathrm{E}^*[R_1] \le 2\mathrm{E}^* \left[\sup_{\|D^{1/2}\boldsymbol{\delta}\| \le \sqrt{p}\varepsilon_n} \Delta_0(\boldsymbol{\delta}) \right] \lesssim np^2\varepsilon_n^2 \le \gamma_n.$$

For R_2, note for any $\boldsymbol{\delta}, \boldsymbol{\delta}' \neq 0$, we have

$$\begin{aligned} |f_n(\boldsymbol{\delta}') - f_n(\boldsymbol{\delta})| &\le \frac{|r_n(\boldsymbol{\delta}') - r_n(\boldsymbol{\delta})|}{n\|D^{1/2}\boldsymbol{\delta}'\|^2 + 1} + \left| r_n(\boldsymbol{\delta}) \cdot \frac{1}{n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1} - \frac{1}{n\|D^{1/2}\boldsymbol{\delta}'\|^2 + 1} \right| \\ &= \Delta_1(\boldsymbol{\delta},\boldsymbol{\delta}') + \Delta_2(\boldsymbol{\delta},\boldsymbol{\delta}'); \end{aligned}$$

furthermore, in each of the small cubes $\mathcal{C}_k^j$, we have $\|D^{1/2}(\boldsymbol{\delta} - \boldsymbol{\delta}_k^j)\| \le \sqrt{p}\varepsilon_n \le 1/\sqrt{n}$. Therefore

$$\mathrm{E}^*[R_2] \le 2\sum_{u=1}^{2} \left(\mathrm{E}^* \left[\sup_{\substack{\boldsymbol{\delta},\boldsymbol{\delta}' \subset \mathcal{C}_{K_n} \\ \|D^{1/2}(\boldsymbol{\delta}'-\boldsymbol{\delta})\| \le \sqrt{p}\varepsilon_n}} \Delta_u(\boldsymbol{\delta},\boldsymbol{\delta}') \right] \right) \lesssim np\varepsilon_n \le \gamma_n.$$

Hence, Chebyshev's inequality implies

$$R_1 = O_{P^*}(\gamma_n), \qquad R_2 = O_{P^*}(\gamma_n).$$

Now we bound R_3. In step III below, we show that for any fixed $\boldsymbol{\delta}$, the following

inequality holds for all $t_n > 0$:

$$\mathrm{P}^*\left(|r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]| \geq t_n\right) \;\leq\; 2\exp\left\{-\frac{t_n^2}{2\left(c_1 n[v(\boldsymbol{\delta})]^{3/2} + c_2\sqrt{v(\boldsymbol{\delta})}t_n\right)}\right\}, \quad \text{(IB.2)}$$

where $v(\boldsymbol{\delta}) = \|D^{1/2}\boldsymbol{\delta}\|^2$ and $c_1, c_2 > 0$ are two constants. Recall there are at most $(2K_n)^p \leq (4pn^{a+1}/\gamma_n)^p$ small cubes with edge-length ε_n; therefore, for large enough $M > 0$,

$$\begin{aligned}
& \mathrm{P}^*(R_3 \geq M\gamma_n) \\
\leq\; & \sum_{k=1}^{K_n}\sum_{j=1}^{B_k} \mathrm{P}^*\left(\frac{r_n(\boldsymbol{\delta}_k^j) - \mathrm{E}^*\left[r_n(\boldsymbol{\delta}_k^j)\right]}{n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1} \geq M\gamma_n\right) \\
\leq\; & \left(\frac{4pn^{a+1}}{\gamma_n}\right)^p \cdot \exp\left\{-\inf_{\boldsymbol{\delta}\in\mathcal{C}_{K_n}} \frac{M^2\gamma_n^2(nv(\boldsymbol{\delta})+1)^2}{2\left(c_1 n[v(\boldsymbol{\delta})]^{3/2} + c_2\sqrt{v(\boldsymbol{\delta})}\cdot M\gamma_n(nv(\boldsymbol{\delta})+1)\right)}\right\} \\
\leq\; & \exp\left\{(a+2)p\log n - \frac{M^2\sqrt{n}\gamma_n^2}{c_1}\right\} \\
\to\; & 0,
\end{aligned}$$

when M is large enough, since $\sqrt{n}\gamma_n^2 = p\log n$; to compute the infimum in the penultimate inequality, we define $z = (n\sqrt{v(\boldsymbol{\delta})} + 1/\sqrt{v(\boldsymbol{\delta})}$, which gives

$$\frac{(nv(\boldsymbol{\delta})+1)^2}{c_1 n[v(\boldsymbol{\delta})]^{3/2} + c_2\sqrt{v(\boldsymbol{\delta})}\cdot M\gamma_n(nv(\boldsymbol{\delta})+1)} \geq \frac{z^2}{(c_1 + M\gamma_n\cdot c_2)z} \geq \frac{\sqrt{n}}{c_1}.$$

Collecting the results for R_1, R_2 and R_3 and recalling that $\gamma_n^4 = (p^2\log^2 n)/n$, we have

$$\sup_{\boldsymbol{\delta}:\|D^{1/2}\boldsymbol{\delta}\|\leq n^a} |f_n(\boldsymbol{\delta}) - \mathrm{E}^*(f_n(\boldsymbol{\delta}))| = O_{P^*}\left(\sqrt[4]{\frac{p^2\log^2 n}{n}}\right) = o_{P^*}(1),$$

since $p^2\log^2 n \ll n$. Thus, we have shown the asserted claim of the Lemma.

Step III: Exponential inequality Here we show the exponential inequality (IB.2) holds. Without loss of generality, we assume the scale-parameter $\sigma_0 = 1$ in Assumption E.4′; therefore, standard calculation leads to

$$\mathrm{E}^*\left[|x_i^T D^{-1/2} u|^q\right] \lesssim q! \cdot \|u\|^q \tag{IB.3}$$

by Assumption E.4′.

Note for fixed $\boldsymbol{\delta}$, $r_n(\boldsymbol{\delta})$ can be written as

$$r_n(\boldsymbol{\delta}) = \sum_{i=1}^n \int_0^{x_i^T\boldsymbol{\delta}} \left(1[y_i - x_i^T\boldsymbol{\beta}^0 \le s] - 1[y_i - x_i^T\boldsymbol{\beta}^0 \le 0]\right) \mathrm{d}s \triangleq \sum_{i=1}^n \int_0^{x_i^T\boldsymbol{\delta}} h_i(s)\,\mathrm{d}s, \tag{IB.4}$$

which follows directly from Knight's identity (*Knight*, 1998); note the above summands $\int h_i(s)\mathrm{d}s$ are non-negative. To apply Lemma IA.6, we check the condition in the next paragraph.

Let $F_i(y)$ and $f_i(y)$ denote the conditional cumulative distribution function and conditional density function of $(y - x_i^T\boldsymbol{\beta}^0) \mid x = x_i$, respectively. Letting $A_n = n\|D^{1/2}\boldsymbol{\delta}\|^3$, we first have

$$\begin{aligned}
& \sum_{i=1}^n \mathrm{E}^*\left[\int_0^{x_i^T\boldsymbol{\delta}} h_i(s)\,\mathrm{d}s\right]^2 \\
= & \sum_{i=1}^n \mathrm{E}_X^*\left[\int_0^{x_i^T\boldsymbol{\delta}}\int_0^{x_i^T\boldsymbol{\delta}} \mathrm{E}_{Y|X=x}^*\left\{h_i(u)h_i(s)\right\}\,\mathrm{d}s\,\mathrm{d}u\right] \\
= & n\cdot\mathrm{E}_X^*\left[\int_0^{x_i^T\boldsymbol{\delta}}\int_0^{x_i^T\boldsymbol{\delta}} F_i(u\wedge s) + F_i(0) - F_i(u\wedge 0) - F_i(s\wedge 0)\,\mathrm{d}s\,\mathrm{d}u\right] \\
\le & 2n\cdot\mathrm{E}_X^*\left[\int_0^{x_i^T\boldsymbol{\delta}} \mathrm{d}s\int_0^s |u| f_i(\tilde{u})\,\mathrm{d}u\right] \\
\lesssim & n\cdot\mathrm{E}^*\left[|x_i^T\boldsymbol{\delta}|^3\right] \\
\lesssim & A_n,
\end{aligned}$$

where the first inequality owns to the mean value theorem; the penultimate inequality follows since f_i is bounded from above in Assumption E.2; the last inequality follows from (IB.3). Next, it is easy to see from induction that for all integers $q \geq 3$,

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{E}^* \left[\int_0^{x_i^T \boldsymbol{\delta}} h_i(s)\, \mathrm{d}s \right]^q &\leq \sum_{i=1}^{n} \mathrm{E}^* \left(|x_i^T \boldsymbol{\delta}|^q \cdot \mathbf{1}_{[0 \leq |y_i - x_i^T \boldsymbol{\beta}^0| \leq |x_i^T \boldsymbol{\delta}|]} \right) \\
&\lesssim n \cdot \mathrm{E}_X^* \left(|x_i^T \boldsymbol{\delta}|^{q+1} \right) \\
&\lesssim n \cdot (q+1)! \cdot \|D^{1/2} \boldsymbol{\delta}\|^{q+1} \\
&\leq q! \cdot A_n \cdot B_n^{q-2},
\end{aligned}$$

where $B_n = 2\|D^{1/2}\boldsymbol{\delta}\|$, and the last inequality follows from (IB.3).

Now we can readily apply Lemma IA.6, which gives

$$\begin{aligned}
\mathrm{P}^* \left(\sum_{i=1}^{n} \int_0^{x_i^T \boldsymbol{\delta}} (h_i(s) - \mathrm{E}^*[h_i(s)])\, \mathrm{d}s \geq t_n \right) &\leq \exp\left\{ -\frac{t_n^2}{2(c_1 A_n + c_2 B_n t_n)} \right\} \\
&\leq \exp\left\{ -\frac{t_n^2}{2\left(n c_1 [v(\boldsymbol{\delta})]^{3/2} + 2c_2 \sqrt{v(\boldsymbol{\delta})} t_n \right)} \right\},
\end{aligned}$$

which is precisely the one-sided version of (IB.2). The inequality for the opposite direction follows in a similar manner since $\int h_i(s)\mathrm{d}s \geq 0$.

Step IV: Control of the supremum Here we compute the expectation of the supremum of Δ_0, Δ_1 and Δ_2 defined in (IB.1).

For Δ_1, from the proof of Theorem 1 in *Pollard* (1991), we deduce,

$$\begin{aligned}
\mathrm{E}^* \sup_{\substack{\boldsymbol{\delta}, \boldsymbol{\delta}' \in \mathcal{C}_{K_n} \\ \|D^{1/2}(\boldsymbol{\delta}' - \boldsymbol{\delta})\| \leq d}} |r_n(\boldsymbol{\delta}) - r_n(\boldsymbol{\delta}')| &\leq \mathrm{E}^* \sup_{\boldsymbol{\delta}, \boldsymbol{\delta}'} \left(\sum_n |x_i^T (\boldsymbol{\delta}' - \boldsymbol{\delta})| \cdot \mathbf{1}_{[|y_i - x_i^T \boldsymbol{\beta}^0| \leq |x_i^T \boldsymbol{\delta}| \vee |x_i^T \boldsymbol{\delta}'|]} \right) \\
&\leq \mathrm{E}^* \sup_{\boldsymbol{\delta}, \boldsymbol{\delta}'} \left(\sum_{i=1}^{n} \|x_i^T D^{-1/2}\| \cdot \|D^{1/2}(\boldsymbol{\delta}' - \boldsymbol{\delta})\| \right) \\
&\leq nd\sqrt{p},
\end{aligned}$$

since $\mathrm{E}^*[\|x_i^T D^{-1/2}\|] \le \sqrt{\mathrm{E}^*[\|x_i^T D^{-1/2}\|^2]} = \sqrt{p}$. Observing the denominator of $f_n(\boldsymbol{\delta})$ is no less than 1, we obtain for $d < 1$,

$$\mathrm{E}^* \sup_{\substack{\boldsymbol{\delta},\boldsymbol{\delta}' \in \mathcal{C}_{K_n} \\ \|D^{1/2}(\boldsymbol{\delta}'-\boldsymbol{\delta})\| \le d}} \Delta_1(\boldsymbol{\delta}', \boldsymbol{\delta}) \lesssim nd\sqrt{p}.$$

For Δ_2, observe that for any $\|D^{1/2}(\boldsymbol{\delta} - \boldsymbol{\delta}')\| \le d$, we have $|v(\boldsymbol{\delta}') - v(\boldsymbol{\delta})| \le d^2 + 2d\sqrt{v(\boldsymbol{\delta})}$, which further implies

$$\left| \frac{1}{n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1} - \frac{1}{n\|D^{1/2}\boldsymbol{\delta}'\|^2 + 1} \right| \le \frac{n|v(\boldsymbol{\delta}) - v(\boldsymbol{\delta}')|}{(nv(\boldsymbol{\delta}) + 1)(nv(\boldsymbol{\delta}') + 1)} \le \frac{nd^2 + 2nd\sqrt{v(\boldsymbol{\delta})}}{(nv(\boldsymbol{\delta}) + 1)}.$$

Therefore, we obtain for Δ_2:

$$\begin{aligned}
\mathrm{E}^* \sup_{\substack{\boldsymbol{\delta},\boldsymbol{\delta}' \in \mathcal{C}_{K_n} \\ \|D^{1/2}(\boldsymbol{\delta}'-\boldsymbol{\delta})\| \le d}} \Delta_2(\boldsymbol{\delta}, \boldsymbol{\delta}') &\le \mathrm{E}^* \sup_{\boldsymbol{\delta} \in \mathcal{C}_{K_n}} \left[r_n(\boldsymbol{\delta}) \cdot \frac{nd^2 + 2nd\sqrt{v(\boldsymbol{\delta})}}{(nv(\boldsymbol{\delta}) + 1)} \right] \\
&\le \mathrm{E}^* \sup_{\boldsymbol{\delta} \in \mathcal{C}_{K_n}} \left[\sum_{i=1}^n |x_i^T \boldsymbol{\delta}| \cdot \frac{nd^2 + 2nd\sqrt{v(\boldsymbol{\delta})}}{(nv(\boldsymbol{\delta}) + 1)} \right] \\
&\le \mathrm{E}^* \left[\sum_{i=1}^n \|D^{-1/2} x_i\| \right] \cdot \sup_{\boldsymbol{\delta} \in \mathcal{C}_{K_n}} \left(\sqrt{v(\boldsymbol{\delta})} \cdot \frac{nd^2 + 2nd\sqrt{v(\boldsymbol{\delta})}}{(nv(\boldsymbol{\delta}) + 1)} \right) \\
&\lesssim n\sqrt{p}(\sqrt{n}d^2 + 2d).
\end{aligned}$$

with the second inequality owns to (IB.4), and the last inequality owns to $\mathrm{E}^*[\|x_i^T D^{-1/2}\|] \le \sqrt{p}$. Therefore, with $d \le 1/\sqrt{n}$, the above display is bounded by $n\sqrt{p}d$.

For Δ_0, we have from (IB.4)

$$\begin{aligned}
\mathrm{E}^* \sup_{\boldsymbol{\delta}:\|D^{1/2}\boldsymbol{\delta}\|\le d} f_n(\boldsymbol{\delta}') &\le \mathrm{E}^* \sup_{\boldsymbol{\delta}:\|D^{1/2}\boldsymbol{\delta}\|\le d} |r_n(\boldsymbol{\delta})| \\
&\le \mathrm{E}^* \sup_{\boldsymbol{\delta}} \left(\sum_n |x_i^T\boldsymbol{\delta}| \cdot \mathbf{1}\big[y_i - x_i^T\boldsymbol{\beta}^0| \le |x_i^T\boldsymbol{\delta}|\big] \right) \\
&\le \mathrm{E}^* \left(\sum_{i=1}^n d \cdot \|x_i^T D^{-1/2}\| \cdot \mathbf{1}\big[|y_i - x_i^T\boldsymbol{\beta}^0| \le d \cdot \|x_i^T D^{-1/2}\|\big] \right) \\
&\lesssim nd^2p.
\end{aligned}$$

where the last inequality follows by first taking conditional expectation over $y \mid x$. □

Lemma IB.2 (Restricted Quadratic Expansion)**.** *Suppose Assumptions E.1, E.2, and E.3′ through E.5′ hold and $s^4p^2\log^2 n = o(n)$. Furthermore, if the tuning parameter satisfies*

$$\lambda \gg \frac{\sqrt{s}p\log p}{\sqrt{n}},$$

then we have

$$L_n(\boldsymbol{\beta}^0 + \boldsymbol{\delta}) - L_n(\boldsymbol{\beta}^0) = \frac{n}{2}\boldsymbol{\delta}^T G\boldsymbol{\delta} - \sum_{i=1}^n x_i^T\boldsymbol{\delta}\phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0) + o_{P^*}(1),$$

uniformly on $\boldsymbol{\delta} \in \mathcal{B}_n(K)$ for any constant K, where

$$\mathcal{B}_n(K) = \left\{ \|G_{11}^{1/2}\boldsymbol{\delta}_1\|_2 \le K\sqrt{\frac{s}{n}};\ \|\boldsymbol{\delta}_2\|_\infty \le K\frac{s\log p}{n\lambda} \right\},$$

with $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \boldsymbol{\delta}_2^T)^T$, where $\boldsymbol{\delta}_1 \in \mathbb{R}^s$ corresponds to the active coefficients.

Proof. Recall the definition of $r_n(\boldsymbol{\delta})$ in Lemma IB.1, it suffices to show

$$\sup_{\boldsymbol{\delta}\in\mathcal{B}_n} \left| r_n(\boldsymbol{\delta}) - \frac{n}{2}\boldsymbol{\delta}^T G\boldsymbol{\delta} \right| = o_{P^*}(1).$$

Due to assumption E.2, we have $\boldsymbol{\delta}_1^T D_{11}\boldsymbol{\delta}_1 \le \boldsymbol{\delta}_1^T G_{11}\boldsymbol{\delta}_1/\overline{f}$. Therefore, when $\boldsymbol{\delta} \in$

$\mathcal{B}_n(K)$

$$\begin{aligned}
\|D^{1/2}\boldsymbol{\delta}\|^2 &\leq 2\left(\boldsymbol{\delta}_1 D_{11}\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2 D_{22}\boldsymbol{\delta}_2\right) \\
&\leq 2K^2\frac{s}{n\overline{f}} + 2\theta_{\max}(D_{22})\cdot p\|\boldsymbol{\delta}_2\|_\infty^2 \\
&\lesssim 2K^2\frac{s}{n} + 2p\left(K\frac{s\sqrt{p}\log p}{n\lambda}\right)^2 \\
&\lesssim \frac{s}{n}, \qquad \text{(IB.5)}
\end{aligned}$$

since $\theta_{max}(D_{22}) \leq p$ from Assumption E.3′, and that $\lambda \gg \frac{\sqrt{s}p\log p}{\sqrt{n}}$. Lemma IB.1 then implies that

$$\sup_{\boldsymbol{\delta}\in\mathcal{B}_n(K)}\left|\frac{r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]}{K^2 s + 1}\right| \lesssim \sup_{\boldsymbol{\delta}\in\mathcal{B}_n(K)}\left|\frac{r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]}{n\mathrm{E}^*[|x^T\boldsymbol{\delta}|^2] + 1}\right| = O_{P^*}\left(\sqrt[4]{\frac{p^2\log^2 p}{n}}\right).$$

Therefore, it follows that for any fixed K

$$\sup_{\boldsymbol{\delta}\in\mathcal{B}_n(K)}|r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]| = o_{P^*}(1),$$

if $s^4p^2\log^2 p = o(n)$.

Next we compute $\mathrm{E}^*[r_n(\boldsymbol{\delta})]$. Denote by F_i as the conditional distribution function for $(y_i - x_i^T\boldsymbol{\beta}^0) \mid X = x_i$, and f_i as the corresponding conditional density function. By Knight's identity (IB.4) we have

$$\begin{aligned}
\mathrm{E}^*[r_n(\boldsymbol{\delta})] &= \sum_{i=1}^n \mathrm{E}_X^*\left(\int_0^{x_i^T\boldsymbol{\delta}} [F_i(s) - F_i(0)]\,\mathrm{d}s\right) \\
&= \sum_{i=1}^n \mathrm{E}_X^*\left(\int_0^{x_i^T\boldsymbol{\delta}} \left[sf_i(0) + \frac{s^2}{2}f_i'(\tilde{s}_i)\right]\mathrm{d}s\right) \qquad \text{(by the mean value-theorem)} \\
&= \frac{1}{2}\boldsymbol{\delta}^T\mathrm{E}^*\left[\sum_{i=1}^n x_i x_i^T f_i(0)\right]\boldsymbol{\delta} + O\left(\overline{f_1}\,\mathrm{E}^*\left[\sum_{i=1}^n |x_i^T\boldsymbol{\delta}|^3\right]\right) \\
&= \frac{n}{2}\boldsymbol{\delta}^T G\boldsymbol{\delta} + O\left(n\cdot\|D^{1/2}\boldsymbol{\delta}\|^3\cdot\sup_{\|u\|=1}\mathrm{E}^*[|u^TD^{-1/2}x_i|^3]\right),
\end{aligned}$$

with $\overline{f}_1$ is the uniform upper bound for $|f_i'|$ in Assumption E.2. Note when $\boldsymbol{\delta} \in \mathcal{B}_n(K)$ we have $\|D^{1/2}\boldsymbol{\delta}\|^3 \leq K^3(s/n)^{3/2}$ as in (IB.5); and also $\sup_{\|u\|=1} \mathrm{E}^*[|u^T D^{-1/2} x_i|^3]$ is uniformly bounded from (IB.3). Therefore, when $s^3 \ll n$ we have

$$\mathrm{E}^*[r_n(\boldsymbol{\delta})] = \frac{n}{2}\boldsymbol{\delta}^T G \boldsymbol{\delta},$$

which completes the proof. □

The following results are simple corollaries from Lemma IB.1 and IB.2.

Lemma IB.3 (Unrestricted Quadratic Expansion)**.** *Suppose Assumptions E.1, E.2, E.3′ and E.4′ hold and $p^6 \log^2 n = o(n)$. In addition, suppose $\theta_{\min}(G) \geq c_0 > 0$, then for $r_n(\boldsymbol{\delta})$ defined in Lemma IB.1 we have*

$$\sup_{\boldsymbol{\delta} \in \mathcal{B}_n(K)} |r_n(\boldsymbol{\delta}) - \frac{n}{2}\boldsymbol{\delta}^T G \boldsymbol{\delta}| = o_{P^*}(1),$$

where $\mathcal{B}_n(K) = \{\|G^{1/2}\boldsymbol{\delta}\|_2 \leq K\sqrt{p/n}\}$.

Proof. The proof is similar to that of Lemma IB.2 and is therefore omitted. □

Corollary IB.1. *Define*

$$\mathcal{B}_n(K_n) = \left\{ \|G_{11}^{1/2}\boldsymbol{\delta}_1\|_2 \leq K_n\sqrt{\frac{s}{n}}\,;\ \|\boldsymbol{\delta}_2\|_\infty \leq K_n \frac{s\log p}{n\lambda} \right\}.$$

Under the condition of Lemma IB.2, there exists a sequence $K_n \to \infty$ such that the conclusion therein holds uniformly when $\boldsymbol{\delta} \in \mathcal{B}_n(K_n)$. That is, for $r_n(\boldsymbol{\delta})$ defined in Lemma IB.1,

$$\sup_{\boldsymbol{\delta} \in \mathcal{B}_n(K_n)} \left| r_n(\boldsymbol{\delta}) - \frac{n}{2}\boldsymbol{\delta}^T G \boldsymbol{\delta} \right| = o_{P^*}(1).$$

Proof. The desired result of the Corollary follows from a generic diagonalization argument. □

Corollary IB.2. *Define*

$$\mathcal{E}_n(K_n) = \left\{ \|G_{11}^{1/2}\boldsymbol{\delta}_1\|_2 \le K_n\sqrt{\frac{s}{n}}\,;\, \|\boldsymbol{\delta}_2\|_\infty \le K_n\frac{\log p}{n\lambda} \right\}.$$

Under the condition of Lemma IB.2, there exists a sequence $K_n \to \infty$ such that

$$\sup_{\boldsymbol{\delta}\in\mathcal{E}_n(K_n)} |r_n(\boldsymbol{\delta}) - \frac{n}{2}\boldsymbol{\delta}_1^T G_{11}\boldsymbol{\delta}_1| = o_{P^*}(1),$$

where $r_n(\boldsymbol{\delta})$ is define in Lemma IB.1.

Proof. We only need to verify $n\boldsymbol{\delta}^T G\boldsymbol{\delta} - n\boldsymbol{\delta}_1^T G_{11}\boldsymbol{\delta}_1 = o_{P^*}(1)$ on $\mathcal{E}_n$. Observe that

$$\begin{aligned}
\left|n\boldsymbol{\delta}^T G\boldsymbol{\delta} - n\boldsymbol{\delta}_1^T G_{11}\boldsymbol{\delta}_1\right| &\le n\boldsymbol{\delta}_2^T G_{22}\boldsymbol{\delta}_2 + 2n\boldsymbol{\delta}_1^T G_{12}\boldsymbol{\delta}_2 \\
&\le n\theta_{\max}(G_{22})p\|\boldsymbol{\delta}_2\|_\infty^2 + 2n\|\boldsymbol{\delta}_1^T G_{11}^{1/2}\| \cdot \|G_{11}^{-1/2}G_{12}\boldsymbol{\delta}_2\| \\
&\lesssim np^2\left(\frac{K_n\log p}{n\lambda}\right)^2 + 2nK_n^2\sqrt{\frac{s}{n}}\cdot\sqrt{\theta_{\max}(G_{22})p}\cdot\frac{\log p}{n\lambda} \\
&\lesssim K_n^2\frac{p^2\log^2 p}{n\lambda^2} + K_n^2\frac{\sqrt{s}p\log p}{\sqrt{n}\lambda} \\
&\to 0,
\end{aligned}$$

provided that K_n diverges slow enough; the second to the last inequality holds due to $\|G_{11}^{-1/2}G_{12}\|^2 \le \theta_{\max}(G_{22}) \le p$ as in Assumption E.3′. □

5.9.3 Proof under the flat prior

Proof of Proposition 6. Denote $\mathcal{A}_n = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \geq M_n/\sqrt{n}\}$. Let the empirical check-loss function be

$$L_n(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \boldsymbol{\beta}).$$

We prove that

$$\frac{\displaystyle\int_{\mathcal{A}_n} \exp\left\{-\sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\boldsymbol{\beta})\right\} \mathrm{d}\boldsymbol{\beta}}{\displaystyle\int_{\mathbb{R}^p} \exp\left\{-\sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\boldsymbol{\beta})\right\} \mathrm{d}\boldsymbol{\beta}} = \frac{\displaystyle\int_{\mathcal{A}_n} \exp\left\{L_n(\beta_0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta}}{\displaystyle\int_{\mathbb{R}^p} \exp\left\{L_n(\beta_0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta}} \xrightarrow{P^*} 0.$$

Without loss of generality, we assume that all X_i are bounded such that $\|X_i\|_2 \leq 1$. Also, we consider the sequence M_n such that $M_n/\sqrt{n} \to 0$. Letting $\delta = \beta - \beta^0$ and $\hat{\delta} = \hat{\beta} - \beta^0$. The proof goes by three parts. We first lower bound the denominator. Then we upper bound the numerator on two disjoint regions $\mathcal{C}_n$ and $\mathcal{D}_n$, where the constant k will be specified later.

$$\mathcal{C}_n = \left\{\beta : \frac{M_n}{\sqrt{n}} \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \leq k\right\},$$
$$\mathcal{D}_n = \left\{\beta : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \geq k\right\}.$$

The proof completes by showing the upper bound converges to 0.

Lower bound the denominator. For any fixed constant K, we consider the integral on $\mathcal{B}_n = \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\| \leq K/\sqrt{n}\}$. Letting the constant

$$\log S_n = n(\beta^0 - \hat{\beta})^T D_1 (\beta^0 - \hat{\beta})/2, \tag{IB.6}$$

which does not depend on β. Recall $\theta_{\max}$ ($\theta_{\min}$) is the maximal (minimal) eigenvalue of the matrix D_1. We proceed by translating into a non-central chi-square distribution. By the uniform expansion in Lemma IB.3, we have

$$\begin{aligned}
& \int_{\mathcal{B}_n} \exp\left\{L_n(\beta_0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta} \\
= \; & S_n \int_{\mathcal{B}_n} \exp\left\{-\frac{n}{2}(\beta-\hat{\beta})^T D_1 (\beta-\hat{\beta}) + R_n(\boldsymbol{\beta})\right\} \mathrm{d}\boldsymbol{\beta} \\
\geq \; & S_n(1+R_n(\boldsymbol{\beta})) \int_{\mathcal{B}_n} \exp\left\{-\frac{n\theta_{\max}}{2}(\beta-\hat{\beta})^T(\beta-\hat{\beta})\right\} \mathrm{d}\boldsymbol{\beta} \\
\geq \; & S_n(1-o_P(1)) \int_{\|\boldsymbol{\delta}\|\leq K/\sqrt{n}} \exp\left\{-\frac{n\theta_{\max}}{2}(\delta-\hat{\delta})^T(\delta-\hat{\delta})\right\} \mathrm{d}\boldsymbol{\delta} \\
= \; & S_n(1-o_P(1)) \left(\frac{2\pi}{n\theta_{\max}}\right)^{p/2} \mathrm{P}(\chi_p^2(g_n) \leq K^2\theta_{\max}),
\end{aligned}$$

where the non-centrality parameter $g_n = n\theta_{\max}\|\hat{\beta}-\beta^0\|^2 = O_p(1)$. $R_n(\boldsymbol{\beta}) = o_P(1)$ uniformly on $\mathcal{B}_n$ from Lemma IB.3.

For any K large enough, on the event $E_n(K) = \{K^2\theta_{\max} \geq 4(p+2g_n)\} \cap \{R_n(\boldsymbol{\beta}) \leq 1/2\}$, Lemma IA.1 gives the following lower bound for the preceding display:

$$\begin{aligned}
\int_{\mathbb{R}^p} \exp\left\{L_n(\beta^0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta} \; & \geq \; \frac{1}{2} S_n \left(\frac{2\pi}{n\theta_{\max}}\right)^{p/2} \left[1 - \mathrm{P}(\chi_p^2(g_n) \geq K^2\theta_{\max})\right] \\
& \geq \; \frac{1}{2} S_n \left(\frac{2\pi}{n\theta_{\max}}\right)^{p/2} \left[1 - \exp(-K^2\theta_{\max}/4)\right] \\
& \geq \; \frac{S_n}{3} \cdot \left(\frac{2\pi}{n\theta_{max}}\right)^{p/2}. \qquad \text{(IB.7)}
\end{aligned}$$

Note for every $\varepsilon > 0$, we can choose a K such that $\overline{\lim}_n P(E_n(K)) \geq 1-\varepsilon$ by the tightness of g_n. Letting $\varepsilon \to 0$ proves the above lower bound holds with probability tending to 1.

Bound the numerator on area $\mathcal{C}_n$ Let $\phi_\tau(u) = \tau - 1[u \leq 0]$. By Knight's identity (*Knight*, 1998), we have

$$\begin{aligned} L_n(\beta^0) - L_n(\beta) &= \sum_{i=1}^n x_i^T \phi_\tau(y_i - x_i^T \boldsymbol{\beta}^0) \cdot \boldsymbol{\delta} \\ &\quad - \sum_{i=1}^n \int_0^{x_i^T \boldsymbol{\delta}} \left(1[y_i - x_i^T \boldsymbol{\beta}^0 \leq s] - 1[y_i - x_i^T \boldsymbol{\beta}^0 \leq 0]\right) \mathrm{d}s \\ &\triangleq \sum_{i=1}^n x_i^T \phi_\tau(y_i - x_i^T \boldsymbol{\beta}^0)\boldsymbol{\delta} - B_n \\ &\leq \sum_{i=1}^n x_i^T \phi_\tau(y_i - x_i^T \boldsymbol{\beta}^0)\boldsymbol{\delta} - E(B_n) + |E(B_n) - B_n|. \qquad \text{(IB.8)} \end{aligned}$$

To proceed further, we first provide a lower bound on $E(B_n)$, followed by a uniform upper bound on $E(B_n) - B_n$.

1. Bound $E(B_n)$ Denote by F_i as the cdf for $(Y - x_i^T \beta^0) \mid X = x_i$, where f_i is the conditional density.

$$\begin{aligned} E(B_n \mid X) &= \sum_{i=1}^n \int_0^{x_i^T \delta} [F_i(s) - F_i(0)] \, \mathrm{d}s \\ &= \sum_{i=1}^n \int_0^{x_i^T \delta} \left[s f_i(0) + \frac{s^2}{2} f_i'(\tilde{s}_i) \right] \mathrm{d}s \qquad \text{(by mean value-theorem)} \\ &\geq \frac{1}{2} \boldsymbol{\delta}^T \left[\sum_{i=1}^n x_i x_i^T f_i(0) \right] \boldsymbol{\delta} - \frac{1}{6} \overline{f_1} \sum_{i=1}^n |x_i^T \boldsymbol{\delta}|^3. \end{aligned}$$

The last inequality is due the bound of the $f'(s)$ in Assumption E.2. Taking expectation again yields

$$\begin{aligned} E(B_n) &\geq \frac{n}{2}\boldsymbol{\delta}^T D_1 \boldsymbol{\delta} - \frac{n}{6}\overline{f_1}\, E|x_i^T\boldsymbol{\delta}|^3 \\ &\geq \frac{n}{4}\boldsymbol{\delta}^T D_1 \boldsymbol{\delta} + \frac{n}{4}\underline{f}\,\boldsymbol{\delta}^T E[x_i x_i^T]\boldsymbol{\delta} - \frac{n}{6}\overline{f_1}\,(E\sup|x_i^T\boldsymbol{\delta}|)\cdot E|x_i^T\boldsymbol{\delta}|^2 \\ &\geq \frac{n}{4}\boldsymbol{\delta}^T D_1 \boldsymbol{\delta} + \left(\frac{1}{4}\underline{f} - \frac{1}{6}\overline{f_1}\cdot \sup_{\boldsymbol{\beta}\in\mathcal{C}_n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_2\right) n\boldsymbol{\delta}^T E[x_i x_i^T]\boldsymbol{\delta} \\ &\geq \frac{n}{4}\boldsymbol{\delta}^T D_1 \boldsymbol{\delta}. \end{aligned}$$

The second identity uses the fact that $|u|^3 \leq \sup|u|\cdot|u|^2$. In the last identity resides the choice of k, such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\| \leq k \leq 6\underline{f}/4\overline{f_1}$ on $\mathcal{C}_n$.

2. Bound $|E(B_n) - B_n|$ For any fixed $0 < \varepsilon_0 < \lambda_{min}/8$, Lemma IB.1 implies that there's a constant k such that

$$\sup_{\boldsymbol{\beta}\in\mathcal{C}_n}\left|\frac{B_n - E(B_n)}{n\boldsymbol{\delta}^T D_1\boldsymbol{\delta}}\right| \leq \sup_{\boldsymbol{\beta}\in\mathcal{C}_n}\left|\frac{B_n - E(B_n)}{n\theta_{min}\|\boldsymbol{\delta}\|^2}\right| \leq \frac{\varepsilon_0}{\theta_{min}} \leq 1/8,$$

with probability approaching unity.

Combining the two steps above, (IB.8) shows with probability going to 1,

$$L_n(\beta^0) - L_n(\beta) \leq \sum_{i=1}^n \phi_\tau(y_i - x_i^T\boldsymbol{\beta}_0)x_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0) - \frac{n}{8}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T D_1(\boldsymbol{\beta}-\boldsymbol{\beta}^0), \quad \text{(IB.9)}$$

uniformly on $\beta \in \mathcal{C}_n$. Define

$$\log(C_n) = 2[\sum_{i=1}^n x_i\phi_\tau(e_i)]^T D_1^{-1}[\sum_{i=1}^n x_i\phi_\tau(e_i)]/n \quad \text{and} \quad \tilde{\mu} = \frac{4}{n}D_1^{-1}\sum_{i=1}^n x_i\phi_\tau(e_i). \quad \text{(IB.10)}$$

And $e_i = y_i - x_i^T\beta^0$. We arrive at the following bound on the integral on $\mathcal{C}_n$

$$\begin{aligned}
\int_{\mathcal{C}_n} \exp\left\{L_n(\beta^0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta} &\leq \int_{\mathcal{C}_n} \exp\left\{-\frac{n}{8}\boldsymbol{\delta}^T D_1 \boldsymbol{\delta} + \sum_{i=1}^{n} x_i^T \phi(e_i)\boldsymbol{\delta}\right\} \mathrm{d}\boldsymbol{\beta} \\
&= C_n \int_{\|\boldsymbol{\delta}\| \geq M_n/\sqrt{n}} \exp\left\{-\frac{n}{8}(\boldsymbol{\delta} - \tilde{\mu})^T D_1 (\boldsymbol{\delta} - \tilde{\mu})\right\} \mathrm{d}\boldsymbol{\delta} \\
&\leq C_n \int_{\|\boldsymbol{\delta}\| \geq M_n/\sqrt{n}} \exp\left\{-\frac{n\theta_{min}}{8}(\boldsymbol{\delta} - \tilde{\mu})^T (\boldsymbol{\delta} - \tilde{\mu})\right\} \mathrm{d}\boldsymbol{\delta} \\
&\leq C_n \cdot \left(\frac{8\pi}{n\theta_{min}}\right)^{p/2} \cdot \mathrm{P}\left[\chi_p^2\left(\frac{n\theta_{min}}{4}\|\tilde{\mu}\|^2\right) \geq \theta_{min} M_n^2/4\right].
\end{aligned}$$

Classic Central Limit Theorem shows the non-centrality parameter $\nu_n = \frac{n\theta_{min}}{4}\|\tilde{\mu}\|^2 = O_p(1)$. Such tightness implies $\theta_{\min} M_n^2/4 \geq \nu_n$ with probability going to 1. Applying Lemma IA.1 yields the following bound

$$\int_{\mathcal{C}_n} \exp\left\{L_n(\beta^0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta} \leq C_n \cdot \left(\frac{8\pi}{n\theta_{min}}\right)^{p/2} \cdot \exp(-\theta_{min} M_n^2/16), \qquad \text{(IB.11)}$$

with probability approaching unity.

Bound the numerator on area $\mathcal{D}_n$. From Assumption E.1 and the convexity of the objective function (5.2), there exists a constant ε_0 such that on $\mathcal{D}_n$

$$L_n(\beta) - L_n(\beta^0) \geq \frac{n\varepsilon_0}{k}\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geq \frac{n\varepsilon_0}{k\sqrt{p}}\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_1, \qquad \text{(IB.12)}$$

with probability going to unity. Note the inequality holds uniformly in $\beta \in \mathcal{D}_n$. Thus,

$$\begin{aligned}
\int_{\mathcal{D}_n} \exp\left\{L_n(\beta^0) - L_n(\beta)\right\} \mathrm{d}\boldsymbol{\beta} &\leq \int_{\mathcal{D}_n} \exp\left\{-\frac{n\varepsilon_0}{k\sqrt{p}}\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_1\right\} \mathrm{d}\boldsymbol{\beta} \\
&\leq \int_{\|\boldsymbol{\delta}\|_1 \geq k} \exp\left\{-\frac{n\varepsilon_0}{k\sqrt{p}} \sum_{j=1}^{p} |\delta_j|\right\} \mathrm{d}\boldsymbol{\beta} \\
&= \left(\frac{2k\sqrt{p}}{n\varepsilon_0}\right)^p \cdot P\left(\Gamma\left(p, \frac{n\varepsilon_0}{k\sqrt{p}}\right) \geq k\right) \\
&\leq \left(\frac{2k\sqrt{p}}{n\varepsilon_0}\right)^p \cdot \exp\left\{-\frac{n\varepsilon_0}{4\sqrt{p}}\right\}, \qquad \text{(IB.13)}
\end{aligned}$$

where we have used the fact that sum of independent exponential distribution forms a Gamma distribution. The last inequality follows from the tail bound for Gamma distribution (*Boucheron et al.*, 2013, Section 2.4)

From the Bahadur representation of the QR estimator (*Koenker*, 2005, Section 4.1), we have that $\log C_n = 4 \log S_n + o_P(1)$, who are defined separately in (IB.6) and (IB.10). Collecting the results in (IB.7), (IB.11) and (IB.13), the proof is now complete. □

5.9.4 Proof under the CA prior

In this subsection, we prove Theorem V.3 in Section 5.5, where the dimension p grows with the sample size. The result when p is fixed, i.e., Theorem V.1, follows as a simple corollary.

Here we review some of the notations, most of which are defined in the beginning of the previous subsection. Recall $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2^0)$ as the true quantile regression coefficients, with $\boldsymbol{\beta}_2^0 = \mathbf{0}$ under Assumption E.5′. The quantile-loss function $\rho_\tau(\cdot)$ and $L_n(\cdot)$ are defined in (5.2). We shall write $\boldsymbol{\phi} = [\phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0)]_{i=1}^n$ as a vector. Recall from Assumption E.3 that $G = \mathrm{E}^*[x_i x_i^T f_{y|x}(x_i^T\boldsymbol{\beta}^0)]$, $D = \mathrm{E}^*[x_i x_i^T]$, as well as the sub-matrices $G_{k\ell}$ for $k = 1, 2$. By Assumption E.3′, define the constant θ_{11} such that $0 < \theta_{11} < \theta_{\min}(G_{11})$. Furthermore, for any vector $a \in \mathbb{R}^p$, we shall write $a = (a_1^T, a_2^T)^T$, where $a_1 \in \mathbb{R}^s$ corresponds to the active components. In particular, we shall write $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$. Finally, we define $\Delta_p = G^{-1}\boldsymbol{X}^T\boldsymbol{\phi}$ and $\Delta_s = G_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{\phi}$.

Under the CA prior (5.5), the posterior density with respect to $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$ is (up to a normalization constant)

$$p_n(\boldsymbol{\delta}) = \pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \cdot \exp\left\{L_n(\boldsymbol{\beta}^0) - L_n\left(\boldsymbol{\delta} + \boldsymbol{\beta}^0\right)\right\}.$$

The posterior probability that $\boldsymbol{\delta} \in \mathcal{A}$ under p_n is then

$$\Pi(\boldsymbol{\delta} \in \mathcal{A}|\mathbb{D}_n) = \frac{\displaystyle\int_{\mathcal{A}^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\displaystyle\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}.$$

To better organize the proofs, we prove the two parts of Theorem V.3 separately in the following subsections.

5.9.4.1 Part 1 of Theorem V.3: Adaptive rate of posterior consistency

Here we prove Part 1 of Theorem V.3, which we re-phrase as the following theorem.

Theorem IC.1. *Consider the CA prior (5.5). Suppose Assumptions E.1, E.2, and E.3′ through E.5′ hold. In addition, suppose* $s^4p^2\log^2 n = o(n)$*, and the tuning parameter* λ_n *satisfies*

$$\frac{\sqrt{s}p\log p}{\sqrt{n}} \ll \lambda_n \ll \min\left\{\underline{b}_n,\ \frac{1}{\sqrt{s}},\ \underline{b}_n\sqrt{\theta_{\min}(D)}\right\}.$$

For any sequence $M_n \to \infty$*, we define*

$$\mathcal{B}_n = \left\{\boldsymbol{\delta} : \|\boldsymbol{\delta}_1\|_2 \le M_n\sqrt{\frac{s}{n}}\ ,\ \|\boldsymbol{\delta}_2\|_\infty \le M_n\frac{s\log p}{n\lambda}\right\}. \tag{IC.1}$$

Then, we have

$$\Pi\left(\boldsymbol{\delta} \in \mathcal{B}_n^C \,\middle|\, \mathbb{D}_n\right) \xrightarrow{P^*} 0,$$

To facilitate the proof of Theorem IC.1, we need the following four lemmas, the proof of which are deferred to Section 5.9.6.

Lemma IC.1 (Lower bounding the integration – CA prior)**.** *Under the conditions of Theorem IC.1, we have*

$$\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \gtrsim_{\mathrm{P}^*} \left(\frac{2\pi}{n}\right)^{s/2}\left(\frac{2}{n\lambda}\right)^{p-s}\cdot\frac{\exp\left(-sn\lambda^2 + n\Delta_s^T G_{11}\Delta_s/2\right)}{\sqrt{|G_{11}|}}.$$

Lemma IC.2 (Preliminary contraction region of the CA prior)**.** *Suppose the conditions of Theorem IC.1 hold. Let*

$$\mathcal{A}_n = \left\{\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0 : \min_{j=1,\ldots,s}|\beta_j| < \lambda,\ \textit{or}\ \max_{j=s+1,\ldots,p}|\beta_s| > \lambda\right\}, \tag{IC.2}$$

then we have

$$\Pi\left(\boldsymbol{\delta} \in \mathcal{A}_n \,\middle|\, \mathbb{D}_n\right) \xrightarrow{P^*} 0.$$

Lemma IC.3. *Recall* $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ *from the beginning of this subsection. Suppose*

Assumptions E.2, E.3′, and E.4′ hold. For any matrix $A \in \mathbb{R}^{q\times p}$ such that all diagonal elements of ADA^T are bounded from above by $C_0 > 0$, we have

$$\left\| \sum_{i=1}^{n} \phi_i A x_i \right\|_\infty = O_{P^*}(\sqrt{n \log q}).$$

Lemma IC.4. *Recall from the beginning of this subsection that*

$$\Delta_s = G_{11}^{-1} \sum_{i=1}^{n} x_{1i} \phi_\tau(y_i - x_i^T \boldsymbol{\beta}^0)/n, \quad \textit{and} \quad \Delta_p = G^{-1} \sum_{i=1}^{n} x_i \phi_\tau(y_i - x_i^T \boldsymbol{\beta}^0)/n.$$

Suppose Assumption E.2 holds, then we have

$$\Delta_s^T G_{11} \Delta_s = O_{P^*}(s/n),$$
$$\Delta_p^T G \Delta_p = O_{P^*}(p/n).$$

Now we are ready to prove Theorem IC.1.

Proof of Theorem IC.1. For the constant q_0 in Lemma IA.8, we define

$$\mathcal{C}_n = \left\{ \boldsymbol{\delta} : \, \|G^{1/2}\boldsymbol{\delta}\| \le 4q_0 \right\}.$$

Recalling that $\mathcal{A}_n$ from (IC.2), Lemma IC.2 shows the posterior probability of $\mathcal{A}_n$ converges to zero. Therefore, to show that $\Pi(\mathcal{B}_n^C \mid \mathbb{D}_n)$ converges to zero, it suffices to show the posterior probabilities of the following areas are all $o_{P^*}(1)$:

1. $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$
2. $\mathcal{C}_n^C \cap \mathcal{A}_n^C$,

where B_n is defined in (IC.1).

We divide our proof into five parts. In step I, we first give upper bounds for the posterior density $p_n(\boldsymbol{\delta})$; in steps II - III below, we obtain upper bounds for the posterior integral $\int p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta}$ on the two areas $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$ and $\mathcal{C}_n^C \cap \mathcal{A}_n^C$ separately;

then in step IV, we show the posterior probabilities of those two areas are both $o_{P^*}(1)$; step V contains some auxiliary calculations to supplement the proof.

Since both areas $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$ and $\mathcal{C}_n^C \cap \mathcal{A}_n^C$ are in $\mathcal{A}_n^C$, the CA prior for $\boldsymbol{\delta}$ becomes

$$\pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\left\{-sn\lambda^2 - n\lambda \cdot \|\boldsymbol{\delta}_2\|_1\right\},$$

throughout the proof, as defined in (5.5).

Step I: Bounding the posterior density $p_n(\boldsymbol{\delta})$ We give two different upper bounds for $p_n(\boldsymbol{\delta})$, one for $\boldsymbol{\delta} \in \mathcal{C}_n \cap \mathcal{A}_n^C$ and the other for $\boldsymbol{\delta} \in \mathcal{C}_n^C \cap \mathcal{A}_n^C$.

Here we consider the case on $\mathcal{C}_n^C \cap \mathcal{A}_n^C$; and we first provide upper bounds for $L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta} + \boldsymbol{\beta}^0)$. By the convexity of L_n, we have

$$\begin{aligned} L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \quad &\leq \quad \frac{n\|G^{1/2}\boldsymbol{\delta}\|}{4q_0} \cdot \sup_{\|G^{1/2}\boldsymbol{\delta}\| \geq 4q_0} \left\{ \frac{L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta} + \boldsymbol{\beta}^0)}{n} \right\} \\ &\leq_{\mathrm{P}^*} \quad -\frac{n\varepsilon_0\|G^{1/2}\boldsymbol{\delta}\|}{4q_0}, \end{aligned} \tag{IC.3}$$

uniformly in $\boldsymbol{\delta} \in \mathcal{C}_n^C$; here ε_0 is the constant due to Assumption E.1. Therefore, combining with the formula of the CA prior, we have

$$\begin{aligned} p_n(\boldsymbol{\delta}) \quad &\leq_{\mathrm{P}^*} \quad \exp\left\{-\frac{n\varepsilon_0\|G^{1/2}\boldsymbol{\delta}\|_2}{q_0} - sn\lambda^2 - n\lambda\|\boldsymbol{\delta}_2\|_1\right\} \\ &\leq \quad \exp\left\{-\frac{n\varepsilon_0\|G^{1/2}\boldsymbol{\delta}\|_2}{q_0} - sn\lambda^2\right\} \\ &\triangleq \quad \exp\left\{-sn\lambda^2\right\} \cdot \overline{p}_{1n}(\boldsymbol{\delta}), \end{aligned}$$

uniformly in $\boldsymbol{\delta} \in \mathcal{C}_n^C \cap \mathcal{A}_n^C$.

Next we consider the case on $\mathcal{C}_n \cap \mathcal{A}_n$. First we bound the function $r_n(\boldsymbol{\delta}) =$

$L_n(\boldsymbol{\delta}+\boldsymbol{\beta}^0) - L_n(\boldsymbol{\beta}^0) + \boldsymbol{\phi}^T\boldsymbol{X}\boldsymbol{\delta}$ using Lemma IB.1:

$$\begin{aligned} -r_n(\boldsymbol{\delta}) \quad &\leq \quad -\mathrm{E}^*[r_n(\boldsymbol{\delta})] + \left(\sup_{\boldsymbol{\delta}\in\mathcal{C}_n} \frac{|r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]|}{n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1}\right) \cdot (n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1) \\ &\leq_{\mathrm{P}^*} \quad -\frac{n\|G^{1/2}\boldsymbol{\delta}\|^2}{4} + \frac{1}{8}\underline{f}(n\|D^{1/2}\boldsymbol{\delta}\|^2 + 1) \\ &\leq \quad -\frac{n}{8}\boldsymbol{\delta}^T G\boldsymbol{\delta} + \frac{1}{8}\underline{f}, \end{aligned} \tag{IC.4}$$

uniformly on $\mathcal{C}_n$, where $\underline{f}$ is a constant introduced in Assumption E.2; the second inequality uses Lemma IB.1 to bound the centered empirical process $r_n(\boldsymbol{\delta}) - \mathrm{E}^*[r_n(\boldsymbol{\delta})]$, and Lemma IA.8 to bound $\mathrm{E}^*[r_n(\boldsymbol{\delta})]$; (IC.4) follows since $\underline{f}\cdot D \preceq G$. Using (IC.4) and the relationship between r_n and L_n, we can upper bound the quantile-loss function as

$$\begin{aligned} L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta}+\boldsymbol{\beta}^0) \quad &\leq_{\mathrm{P}^*} \quad \boldsymbol{\phi}^T\boldsymbol{X}\boldsymbol{\delta} - \frac{n}{8}\boldsymbol{\delta}^T G\boldsymbol{\delta} + \frac{1}{8}\underline{f} \\ &\leq \quad \boldsymbol{\phi}^T\boldsymbol{X}\boldsymbol{\delta} - \frac{n}{8}(\boldsymbol{\delta}_1 + A_2\boldsymbol{\delta}_2)^T G_{11}(\boldsymbol{\delta}_1 + A_2\boldsymbol{\delta}_2) + \frac{1}{8}\underline{f} \\ &\leq \quad -\frac{n}{8}(\boldsymbol{\delta}_1 + A_2\boldsymbol{\delta}_2 - 4\Delta_s)^T G_{11}(\boldsymbol{\delta}_1 + A_2\boldsymbol{\delta}_2 - 4\Delta_s) \\ &\quad + \|\boldsymbol{\phi}^T\boldsymbol{X}_2 - \boldsymbol{\phi}^T\boldsymbol{X}_1 A_2\|_\infty \cdot \|\boldsymbol{\delta}_2\|_1 \\ &\quad + 2n\Delta_s^T G_{11}\Delta_2 + \frac{1}{8}\underline{f}, \end{aligned} \tag{IC.5}$$

where $A_2 = G_{11}^{-1}G_{12}$; the second inequality relies on $\boldsymbol{\delta}^T G\boldsymbol{\delta} \geq (\boldsymbol{\delta}_1 + A_2\boldsymbol{\delta}_2)^T G_{11}(\boldsymbol{\delta}_1 + A_2\boldsymbol{\delta}_2)$ by using the Schur-decomposition of G; and the last inequality follows by completing the squares with respect to $\boldsymbol{\delta}_1$ and Hölder's inequality.

Let $\alpha_n = \|\boldsymbol{\phi}^T\boldsymbol{X}_2 - \boldsymbol{\phi}^T\boldsymbol{X}_1 A_2\|_\infty$ and $\widetilde{\mu}(\boldsymbol{\delta}_2) = 4\Delta_s - A_2\boldsymbol{\delta}_2$. Combining (IC.5) with the CA prior, we have

$$\begin{aligned} p_n(\boldsymbol{\delta}) \quad &\lesssim_{\mathrm{P}^*} \quad \exp\left\{-sn\lambda^2 + 2n\Delta_s^T G_{11}\Delta_s\right\} \cdot \exp\left\{-\frac{n}{8}(\boldsymbol{\delta}_1 - \widetilde{\mu}(\boldsymbol{\delta}_2))^T G_{11}(\boldsymbol{\delta}_1 - \widetilde{\mu}(\boldsymbol{\delta}_2))\right\} \\ &\qquad \cdot \exp\left\{-(n\lambda - \alpha_n)\cdot\|\boldsymbol{\delta}_2\|_1\right\} \\ &\triangleq \quad \exp\left\{-sn\lambda^2 + 2n\Delta_s^T G_{11}\Delta_s\right\} \cdot \overline{p}_{2n}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2), \end{aligned}$$

uniformly on $\mathcal{C}_n \cap \mathcal{A}_n^C$.

Step II: Bounding the posterior integral on $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$ Here we bound the posterior integral of $p_n(\boldsymbol{\delta})$ on $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$. Using the upper bound of $\overline{p}_{2n}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ in step I, we relate its integration to probabilistic tail bounds. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{p-s})$ be distributed as

$$\begin{aligned} \xi_1, \ldots, \xi_{p-s} &\overset{i.i.d.}{\sim} \text{Laplace}\left(\frac{1}{n\lambda - \alpha_n}\right), \\ \boldsymbol{Z} \mid \boldsymbol{\xi} &\sim \text{N}\left(\widetilde{\mu}(\boldsymbol{\xi}),\ \frac{4}{n} G_{11}^{-1}\right), \end{aligned}$$

where α_n and $\widetilde{\mu}$ are defined at the end of Step I. In what follows, we shall write $\Pr(\cdot)$ as the probability with respect to $(\boldsymbol{Z}, \boldsymbol{\xi})$, while holding Δ_s and α_n fixed. Given any fixed Δ_s and $\alpha_n < n\lambda$, the function $\overline{p}_{2n}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ is proportional to the joint density function of the vector $(\boldsymbol{Z}, \boldsymbol{\xi})$. Therefore, the integration of $\overline{p}_{2n}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ can be related to the probabilistic statements about $(\boldsymbol{Z}, \boldsymbol{\xi})$, which gives

$$\begin{aligned} & \int_{\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta} \\ \lesssim_{\mathrm{P}^*} \ & \exp\{-sn\lambda^2 + 2n\Delta_s^T G_{11} \Delta_s\} \cdot \int_{\mathcal{B}_n^C \cap \mathcal{A}_n^C} \overline{p}_{2n}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)\, \mathrm{d}\boldsymbol{\delta}_1\, \mathrm{d}\boldsymbol{\delta}_2 \\ = \ & \exp\{-sn\lambda^2 + 2n\Delta_s^T G_{11} \Delta_s\} \cdot \left(\frac{2}{n\lambda - \alpha_n}\right)^{p-s} \cdot \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{8\pi}{n}\right)^{s/2} \\ & \cdot \Pr\left(\|\boldsymbol{Z}\|_2 \geq M_n \sqrt{\frac{s}{n}} \ \text{ or } \ \|\boldsymbol{\xi}\|_\infty \geq \frac{M_n s \log p}{n\lambda}\right), \end{aligned} \tag{IC.6}$$

where we insert the normalizing constants of Laplace and Gaussian distributions in the second equality; note the displayed equation holds on the event $E_1(\gamma) = \{p\alpha_n < \gamma \cdot n\lambda\}$ for small enough constant $\gamma > 0$.

To compute the tail probability in (IC.6), we break it into two parts. First, we

have

$$\Pr\left(\|\boldsymbol{\xi}\|_\infty \geq \frac{M_n s \log p}{n\lambda}\right) \leq p \cdot \exp\left\{-M_n s \log p/2\right\} \leq \exp\left\{-M_n s \log p/4\right\},$$

on the event $E_1(\gamma)$, which follows from Lemma IA.2. Second, using standard conditional probability formula, we can show

$$\begin{aligned} & \Pr\left(\|\boldsymbol{Z}\|_2 \geq M_n\sqrt{\frac{s}{n}}\,,\ \|\boldsymbol{\xi}\|_\infty \leq \frac{M_n s \log p}{n\lambda}\right) \\ \leq & \sup_{\|\boldsymbol{\delta}_2\|_\infty \leq M_n s \log p/(n\lambda)} \Pr\left(\|\boldsymbol{Z}\|_2 \geq M_n\sqrt{\frac{s}{n}} \;\middle|\; \boldsymbol{\xi} = \boldsymbol{\delta}_2\right) \\ \leq & \exp\left\{-\frac{\theta_{11} M_n^2 s}{16}\right\}, \end{aligned}$$

where the last inequality holds on the event

$$E_2 = \left\{8n \cdot \sup_{\|\boldsymbol{\delta}_2\|_\infty \leq M_n s \log p/(n\lambda)} \left[\widetilde{\mu}(\boldsymbol{\delta}_2)^T G_{11} \widetilde{\mu}(\boldsymbol{\delta}_2)\right] \leq M_n^2 \theta_{11} s\right\},$$

by Lemma IA.1; since $\boldsymbol{Z} \mid \boldsymbol{\xi} = \boldsymbol{\delta}_2$ follows a Gaussian distribution with mean $\widetilde{\mu}(\boldsymbol{\delta}_2)$. Thus, combining the two tail bounds above, we have

$$\begin{aligned} & \Pr\left(\|\boldsymbol{Z}\|_2 \geq M_n\sqrt{\frac{s}{n}} \ \text{ or } \ \|\boldsymbol{\xi}\|_\infty \geq \frac{M_n s \log p}{n\lambda}\right) \\ \leq & \exp\left\{-\frac{\theta_{11} M_n^2 s}{16}\right\} + \exp\left\{-M_n s \log p/4\right\} \\ \leq & 2\exp\left\{-c_0 \cdot M_n s\right\}, \end{aligned} \tag{IC.7}$$

for some constant $c_0 > 0$.

To further simplify (IC.6), note on the event $E_1(\gamma) = \{p\alpha_n < \gamma \cdot n\lambda\}$, we have

$$\begin{aligned} \left(\frac{2}{n\lambda - \alpha_n}\right)^{p-s} &= \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \left(\frac{1}{1 - \alpha_n/(n\lambda)}\right)^{p-s} \\ &\leq \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \frac{1}{1-\gamma}\ , \end{aligned} \tag{IC.8}$$

which follows since $(1-x)^p \geq 1-px$ for all $0 < x < 1$. Therefore, substituting (IC.7) and (IC.8) into (IC.6), we have

$$\begin{aligned}
&\int_{\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \\
\lesssim_{\mathrm{P}^*} \quad & \exp\{-sn\lambda^2 + 2n\Delta_s^T G_{11}\Delta_s\} \cdot \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{8\pi}{n}\right)^{s/2} \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \exp\left\{-c_0 \cdot M_n s\right\},
\end{aligned}$$

on the events $E_1(\gamma)$ and E_2.

Step III: Bounding the posterior integral on $\mathcal{C}_n^C \cap \mathcal{A}_n^C$ When $\boldsymbol{\delta} \in \mathcal{C}_n^C \cap \mathcal{A}_n^C$, we first bound the posterior density $p_n(\boldsymbol{\delta})$ with the upper bound $\overline{p}_{1n}(\boldsymbol{\delta})$ in step I. Then, by letting $\boldsymbol{u} = G^{1/2}\boldsymbol{\delta}$, the posterior integral is bounded by

$$\begin{aligned}
&\int_{\mathcal{C}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \\
\lesssim_{\mathrm{P}^*} \quad & \exp\left\{-sn\lambda^2\right\} \cdot \int_{\mathcal{C}_n^C \cap \mathcal{A}_n^C} \overline{p}_{1n}(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \\
\leq \quad & \exp\{-sn\lambda^2\} \cdot \frac{1}{\sqrt{|G|}} \cdot \int_{\|\boldsymbol{u}\|_2 \geq 4q_0} \exp\left\{-\frac{n\varepsilon_0\|\boldsymbol{u}\|_2}{4q_0}\right\} \mathrm{d}\boldsymbol{u} \\
\leq \quad & \exp\{-sn\lambda^2\} \cdot \frac{1}{\sqrt{|G|}} \cdot \int_{\mathbb{R}^p} \exp\left\{\theta_n \cdot (\|\boldsymbol{u}\|_2 - 4q_0) - \frac{n\varepsilon_0\|\boldsymbol{u}\|_2}{4q_0}\right\} \mathrm{d}\boldsymbol{u} \\
\leq \quad & \exp\{-sn\lambda^2 - 4\theta_n q_0\} \cdot \frac{1}{\sqrt{|G|}} \cdot \int_{\mathbb{R}^p} \exp\left\{-\left(\frac{n\varepsilon_0}{4q_0} - \theta_n\right) \cdot \frac{\|\boldsymbol{u}\|_1}{\sqrt{p}}\right\} \mathrm{d}\boldsymbol{u} \\
= \quad & \exp\{-sn\lambda^2 + p - n\varepsilon_0\} \cdot \frac{1}{\sqrt{|G|}} \cdot \left(\frac{4q_0}{\sqrt{p}}\right)^p, \qquad \text{(IC.9)}
\end{aligned}$$

where $\theta_n = (n\varepsilon_0 - p)/(4q_0) > 0$; the second inequality uses the Cramér-Chernoff method (*Boucheron et al.*, 2013, Section 2.2); the penultimate inequality follows from $\sqrt{p}\|x\|_2 \geq \|x\|_1$ for $x \in \mathbb{R}^p$; the last inequality follows by the normalizing constant of the Laplace distribution.

Step IV: Final bounds on the posterior probability Recall the events $E_1(\gamma)$ and E_2 in Step II; and we further define a event $E_3(\gamma) = \left\{n\Delta_s^T G_{11}\Delta_s \leq \gamma \cdot M_n s\right\}$.

In step V later, we shall verify that all three events have P*-probability going to 1, which, by Lemma IA.5, implies that the upper bounds for posterior integrals in steps II and III hold with P*-probability going to 1.

Finally, we close the proof by showing the posterior probability of $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$ and $\mathcal{C}_n^C \cap \mathcal{A}_n^C$ are both $o_{P^*}(1)$. Since Lemma IC.1 implies that

$$\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \gtrsim_{\mathrm{P}^*} \left(\frac{2\pi}{n}\right)^{s/2} \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \frac{\exp\left(n \cdot \Delta_s^T G_{11} \Delta_s/2 - sn\lambda^2\right)}{\sqrt{|G_{11}|}} \triangleq \widetilde{P}_n,$$

it suffices to verify that

$$\int_{\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} + \int_{\mathcal{C}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} = o_{P^*}\left(\widetilde{P}_n\right).$$

For the first area $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C$, we compare $\widetilde{P}_n$ with the bound displayed at the end of step II. After cancellation, we have

$$\begin{aligned} \frac{\displaystyle\int_{\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} 2^s \cdot \exp\left\{\frac{3n}{2}\Delta_s^T G_{11}\Delta_s - c_0 \cdot M_n s\right\} \\ &= o_{P^*}(1), \end{aligned}$$

which follows since the event $E_3(\gamma) = \left\{n\Delta_s^T G_{11}\Delta_s \le \gamma M_n s\right\}$ has P*-probability tending to 1.

For the second area $\mathcal{C}_n^C \cap \mathcal{A}_n^C$, we compare $\widetilde{P}_n$ with the bound in step III. To facilitate the comparison, note that

$$\widetilde{P}_n \ge \exp\left(-sn\lambda^2\right) \cdot \left(\frac{1}{n}\right)^p \cdot \frac{1}{\sqrt{|G_{11}|}},$$

for all sufficiently large n since $\lambda \ll 1$. Therefore, we obtain:

$$\begin{aligned}
\frac{\int_{\mathcal{C}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} \left(\frac{4nq_0}{\sqrt{p}}\right)^p \cdot \frac{1}{\sqrt{|\widetilde{G}_{22}|}} \cdot \exp\{p - n\varepsilon_0\} \\
&\leq \exp\left\{p \log\left(\frac{4nq_0}{\sqrt{p}}\right) + \frac{p}{2}\log[\theta_{\min}^{-1}(\widetilde{G}_{22})] + p - n\varepsilon_0\right\} \\
&\leq \exp\left\{c_2 \cdot p\log n + c_3 \cdot p \log p - n\varepsilon_0\right\} \\
&= o_{P^*}(1), \qquad \text{(IC.10)}
\end{aligned}$$

for some constants $c_2, c_3 > 0$, where we use $|G| = |G_{11}| \cdot |\widetilde{G}_{22}|$ in the first inequality, with $\widetilde{G}_{22} = G_{22} - G_{21}G_{11}^{-1}G_{12}$ being the Schur-complement of G_{11}; the second inequality bounds the determinant with eigenvalues; the penultimate equality follows since $\theta_{\min}(\widetilde{G}_{22}) \geq \theta_{\min}(G) \gtrsim p^{-1}$ in Assumption E.3′; and the last equation owns to $p\log(n \vee p) \ll n$.

Therefore, the proof is now complete.

Step V: Auxiliary calculations Now we show that each of the events

$$E_1(\gamma) = \{p\alpha_n \leq \gamma \cdot n\lambda\}, \quad E_3(\gamma) = \left\{n\Delta_s^T G_{11}\Delta_s \leq \gamma M_n s\right\},$$
$$\text{and} \quad E_2 = \left\{8n \cdot \sup_{\|\boldsymbol{\delta}_2\|_\infty \leq M_n s\log p/(n\lambda)} \left[\widetilde{\mu}(\boldsymbol{\delta}_2)^T G_{11}\widetilde{\mu}(\boldsymbol{\delta}_2)\right] \leq M_n^2\theta_{11}s\right\},$$

holds with P*-probability tending to 1, for all small enough $\gamma > 0$.

We consider the event $E_1(\gamma)$ first. Let $A = [-A_2^T, \boldsymbol{I}_{p-s}]$ and $v_i = Ax_i \in \mathbb{R}^{p-s}$, then $\alpha_n = \|\sum_{i=1}^n \phi_i v_i^T\|_\infty$. We first show that $e_j^T ADA^T e_j$ is uniformly bounded for all $j = 1, \ldots, p-s$, where e_j be the j-unit vector in $\mathbb{R}^{p-s}$; and then apply Lemma

IC.3. Note $G/\overline{f} \preceq D \preceq G/\underline{f}$ in Assumption E.2, multiplying by block gives

$$\begin{aligned} e_j^T ADA^T e_j &\leq \frac{1}{\underline{f}} \cdot e_j^T AGA^T e_j \\ &= \frac{1}{\underline{f}} \cdot e_j^T (G_{22} - G_{21}G_{11}^{-1}G_{12}) e_j \\ &\leq c_0, \end{aligned}$$

uniformly for all $j = 1, \ldots, p-s$, where $c_0 = \overline{f}/\underline{f}$; we have used the fact that $G_{22} - G_{21}G_{11}^{-1}G_{12} \preceq G_{22} \preceq \overline{f} D_{22}$, and that $D_{jj} = 1$ for all $j = 1, \ldots, p$, as in Assumption E.3′. Then, Lemma IC.3 implies that

$$\mathrm{P}^*\left(E_1^C(\gamma)\right) = \mathrm{P}^*\left(\left\|\sum_{i=1}^n \phi_i A x_i\right\|_\infty \geq \gamma \cdot \frac{n\lambda}{p}\right) \to 0.$$

since $\lambda \gg p \log p/\sqrt{n}$.

For the event $E_3(\gamma)$, Lemma IC.4 directly implies that $\mathrm{P}^*(E_3(\gamma)) \to 1$ as $n \to \infty$.

Finally, we consider the event E_2. Note that $\|x-y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$, we have

$$n \cdot \widetilde{\mu}(\boldsymbol{\delta}_2)^T G_{11} \widetilde{\mu}(\boldsymbol{\delta}_2) \lesssim n \cdot \Delta_s^T G_{11} \Delta_s + n \cdot \boldsymbol{\delta}_2^T A_2^T G_{11} A_2 \boldsymbol{\delta}_2, \tag{IC.11}$$

with $A_2 = G_{11}^{-1} G_{12}$. We consider the two terms in (IC.11) separately.

For the first term in (IC.11), Lemma IC.4 implies that $n \cdot \Delta_s^T G_{11} \Delta_s = O_{P^*}(s)$. For the second term, note $A_2^T G_{11} A_2 G_{21} \preceq G_{11}^{-1} G_{12} \preceq G_{22}$ by the properties of Schur-complements; therefore, uniformly when $\|\boldsymbol{\delta}_2\|_\infty \leq M_n s \log p/(n\lambda)$

$$\begin{aligned} n \cdot \boldsymbol{\delta}_2^T A_2^T G_{11} A_2 \boldsymbol{\delta}_2 &\leq n \cdot \boldsymbol{\delta}_2^T G_{22} \boldsymbol{\delta}_2 \\ &\leq n \cdot \theta_{\max}(G_{22}) p \|\boldsymbol{\delta}_2\|_\infty^2 \\ &\lesssim M_n^2 n p^2 \left(\frac{s \log p}{n\lambda}\right)^2 \\ &= M_n^2 s \cdot o(1), \end{aligned}$$

where the penultimate inequality holds as $\theta_{\max}(G_{22}) \leq \theta_{\max}(G) \lesssim p$ in Assumption E.3′; the last equation holds since $\lambda \gg \sqrt{s}p \log p/\sqrt{n}$.

Thus, plugging the bounds for $\Delta_s^T G_{11} \Delta_s$ and $\boldsymbol{\delta}_2^T A_2^T G_{11} A_2 \boldsymbol{\delta}_2$ into (IC.11), we have

$$\sup_{\|\boldsymbol{\delta}_2\|_\infty \leq M_n s \log p/(n\lambda)} \left[n \cdot \widetilde{\mu}(\boldsymbol{\delta}_2)^T G_{11} \widetilde{\mu}(\boldsymbol{\delta}_2) \right] = o_{P^*}(M_n^2 s),$$

which conclude that $\mathrm{P}^*(E_2) \to 1$. □

5.9.4.2 Part 2 of Theorem V.3: Distributional Approximation

We first introduce some additional notations. Let $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ be the classic quantile regression estimator from minimizing (5.2); and let $\widetilde{\boldsymbol{\beta}}_1$ be the oracle estimator using the true model. Furthermore, let $\log T_n = n\Delta_s^T G_{11} \Delta_s/2$ and $\log S_n = n\Delta_p^T G \Delta_p/2$; we define the following functions with respect to $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$:

$$\begin{aligned} h_n(\boldsymbol{\delta}) &= S_n \cdot \exp\left\{ -\frac{n}{2}(\boldsymbol{\delta} - \Delta_p)^T G (\boldsymbol{\delta} - \Delta_p) - sn\lambda^2 - n\lambda \|\boldsymbol{\delta}_2\|_1 \right\} \\ f_n(\boldsymbol{\delta}) &= T_n \cdot \exp\left\{ -\frac{n}{2}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1)^T G_{11} (\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1) - sn\lambda^2 - n\lambda \|\boldsymbol{\delta}_2\|_1 \right\}, \end{aligned} \tag{IC.12}$$

where $\widetilde{\boldsymbol{\delta}}_1 = \widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0$.

Using the notations above, we re-phrase part 2 of Theorem V.3 below.

Theorem IC.2. *Suppose the conditions of Theorem IC.1 hold, and in addition*

$$\lambda_n \gg \frac{\sqrt{s}p \log^{1.5} p}{\sqrt{n}}.$$

Recalling that $p_n(\boldsymbol{\delta})$ is the posterior density function for $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$; we have

$$\left\| \frac{p_n}{\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}} - \frac{f_n}{\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}} \right\|_{TV} \xrightarrow{P^*} 0.$$

where f_n is defined in (IC.12).

To prove Theorem IC.2, we need the following lemma, the proof of which is deferred to Section 5.9.6.

Lemma IC.5 (Normal likelihood with CA prior)**.** *Under the conditions of Theorem IC.2, we have the following:*

$$\left\| \frac{f_n(\boldsymbol{\delta})}{\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{h_n(\boldsymbol{\delta})}{\int_{\mathbb{R}^p} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right\|_{TV} \xrightarrow{P^*} 0,$$

where f_n and h_n are defined in (IC.12).

Now we are ready to prove Theorem IC.2.

Proof of Theorem IC.2. In what follows, we shall write $\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$ for integrating a function f_n on $\mathbb{R}^p$. In view of Lemma IC.5, we only need to show that p_n converges to h_n,

$$\left\| \frac{p_n(\boldsymbol{\delta})}{\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{h_n(\boldsymbol{\delta})}{\int h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right\|_{TV} \xrightarrow{P^*} 0.$$

Note Lemma IC.1 implies that

$$\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \gtrsim_{\mathrm{P}^*} \left(\frac{2\pi}{n}\right)^{s/2} \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \frac{T_n \cdot \exp\left(-sn\lambda^2\right)}{\sqrt{|G_{11}|}} \triangleq \widetilde{P}_n.$$

Therefore, following the proof of Theorem 1 of *Chernozhukov and Hong* (2003), it suffices to show

$$\frac{\int_{\mathbb{R}^p} |p_n(\boldsymbol{\delta}) - h_n(\boldsymbol{\delta})|\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n \vee \int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \xrightarrow{P^*} 0.$$

Fix a diverging sequence $K_n \to +\infty$ that satisfies the condition in Corollary IB.1,

we define

$$\mathcal{B}_n = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_1\|_2 \leq K_n \sqrt{\frac{s}{n}} \text{ , and } \|\boldsymbol{\delta}_2\|_\infty \leq K_n \frac{s \log p}{n\lambda} \right\}.$$

In the following, we upper bound the integral of $|p_n - h_n|$ on $\mathcal{B}_n$ and its complement, separately in steps I and II.

Step I: Bounding $\int |p_n - h_n| \, \mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n$ First, we analyze the CA prior when $\boldsymbol{\delta} \in \mathcal{B}_n$. For the active coefficients we have $|\delta_j + \beta_j^0| > \lambda$ for $1 \leq j \leq s$, since $|\beta_j^0| \geq \underline{b} \gg \sqrt{s/n}$ from Assumption E.5′; and for the inactive coefficients $|\delta_j + \beta_j^0| < \lambda$ for $s < j \leq p$, since $\log p/(n\lambda) \ll \lambda$. Therefore, the CA prior for $\boldsymbol{\delta}$ then becomes

$$\pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\left\{ -sn\lambda^2 - n\lambda \|\boldsymbol{\delta}_2\|_1 \right\}, \tag{IC.13}$$

when $\boldsymbol{\delta} \in \mathbb{B}_n$.

Since h_n contains the same factor as the above CA prior, Corollary IB.1 shows that

$$\begin{aligned} \sup_{\boldsymbol{\delta} \in \mathcal{B}_n} \left| \log \left(\frac{h_n(\boldsymbol{\delta})}{p_n(\boldsymbol{\delta})} \right) \right| &= \sup_{\boldsymbol{\delta} \in \mathcal{B}_n} \left| L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta} + \boldsymbol{\beta}^0) + \frac{n}{2} \boldsymbol{\delta}^T G \boldsymbol{\delta} - n \cdot \Delta_p^T G \boldsymbol{\delta} \right| \\ &= o_{P^*}(1), \end{aligned}$$

which further implies $|h_n(\boldsymbol{\delta})/p_n(\boldsymbol{\delta}) - 1| = o_{P^*}(1)$ uniformly when $\boldsymbol{\delta} \in \mathcal{B}_n$. Therefore, we have:

$$\begin{aligned} \int_{\mathcal{B}_n} |p_n(\boldsymbol{\delta}) - h_n(\boldsymbol{\delta})| \, \mathrm{d}\boldsymbol{\delta} &= \int_{\mathcal{B}_n} p_n(\boldsymbol{\delta}) \left| 1 - \left(\frac{h_n(\boldsymbol{\delta})}{p_n(\boldsymbol{\delta})} \right) \right| \mathrm{d}\boldsymbol{\delta} \\ &= o_{P^*} \left(\int p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} \right). \end{aligned}$$

Step II: Bounding $\int |p_n - h_n| \, \mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n^C$ Here we control the integration of $|p_n - h_n|$ on $\mathcal{B}_n^C$ by showing that that both $\int_{\mathcal{B}_n^C} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}$, and $\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}$ are $o_{P^*}(\widetilde{P}_n)$.

For p_n, Theorem IC.1 directly implies that

$$\int_{\mathcal{B}_n^C} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} = o_{P^*}\left(\widetilde{P}_n\right).$$

Let $A_2 = G_{11}^{-1} G_{12}$, and we further define $\tilde{\mu}(\boldsymbol{\delta}_2) = \Delta_s - A_2 \boldsymbol{\delta}_2$ and $\alpha_n = \|\boldsymbol{\phi}^T \boldsymbol{X}_2 - \boldsymbol{\phi}^T \boldsymbol{X}_1 A_2\|_\infty$. For $h_n(\boldsymbol{\delta})$, we first upper bound it by

$$\begin{aligned} h_n(\boldsymbol{\delta}) &= \exp\left\{-\frac{n}{2}\boldsymbol{\delta}^T G \boldsymbol{\delta} + n\Delta_s^T G \boldsymbol{\delta} - sn\lambda^2 - n\lambda\|\boldsymbol{\delta}_2\|_1\right\} \\ &\leq T_n \cdot \exp\left\{-sn\lambda^2 - \frac{n}{2}(\boldsymbol{\delta}_1 - \tilde{\mu}_1(\boldsymbol{\delta}_2))^T G_{11}(\boldsymbol{\delta}_1 - \tilde{\mu}_1(\boldsymbol{\delta}_2)) - (n\lambda - \alpha_n) \cdot \|\boldsymbol{\delta}_2\|_1\right\} \\ &\triangleq T_n \cdot \exp\{-sn\lambda^2\} \cdot \overline{h}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2), \end{aligned}$$

where T_n is defined before (IC.12); we use the decomposition of $\boldsymbol{\delta}^T G \boldsymbol{\delta}$ that is similar to the one in (IC.5).

Similar to step II in the proof of Theorem IC.1, we can relate the integration of $\overline{h}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ to Gaussian and Laplace tail bounds. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{p-s})$ be distributed as

$$\begin{aligned} \xi_1, \ldots, \xi_{p-s} &\overset{i.i.d.}{\sim} \text{Laplace}\left(\frac{1}{n\lambda - \alpha_n}\right), \\ \boldsymbol{Z} \mid \boldsymbol{\xi} &\sim \text{N}\left(\widetilde{\mu}(\boldsymbol{\xi}), \frac{1}{n} G_{11}^{-1}\right). \end{aligned}$$

Following (IC.6), the function $\overline{h}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ is proportional to the joint density of $(\boldsymbol{Z}, \boldsymbol{\xi})$. Recalling the definition of $\widetilde{P}_n$ in the beginning of the proof, the integral of h_n can

then be bounded by:

$$\begin{aligned}\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} &\leq T_n \cdot \exp\{-sn\lambda^2\} \cdot \left(\frac{2}{n\lambda-\alpha_n}\right)^{p-s} \cdot \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{2\pi}{n}\right)^{s/2} \\ &\quad \cdot \Pr\left(\|\boldsymbol{Z}\|_2 \geq K_n\sqrt{\frac{s}{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \geq \frac{K_n s\log p}{n\lambda}\right) \\ &\lesssim \left(\frac{n\lambda}{n\lambda-\alpha_n}\right)^{p-s} \cdot \widetilde{P}_n \cdot \exp\left\{-c_0 K_n s\right\} \\ &\lesssim \widetilde{P}_n \cdot \exp\left\{-c_0 K_n s\right\}, \end{aligned} \tag{IC.14}$$

for some constant $c_0 > 0$; in the second inequality we bound the tail probability with (IC.7), which holds on the events

$$\begin{aligned} E_1(\gamma) &= \{p\cdot\alpha_n \leq \gamma\cdot n\lambda\}, \\ E_2 &= \left\{8n \cdot \sup_{\|\boldsymbol{\delta}_2\|_\infty \leq K_n^2 s\log p/(n\lambda)} \left[\tilde{\mu}(\boldsymbol{\delta}_2)^T G_{11}\tilde{\mu}(\boldsymbol{\delta}_2)\right] \leq K_n s\theta_{\min}(G_{11})\right\}, \end{aligned}$$

for a small enough constant $\gamma > 0$; the last inequality follows from (IC.8).

Similar to the proof of Theorem IC.1 (Step V), we can show that both the events E_1 and $E_2(\gamma)$ have P*-probability tending to 1; therefore, we have proved

$$\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} = o_{P^*}\left(\widetilde{P}_n\right),$$

as $K_n \to \infty$.

Combining steps I and II, we obtain

$$\frac{\int_{\mathbb{R}^p} |p_n(\boldsymbol{\delta}) - h_n(\boldsymbol{\delta})|\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n \vee \int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \xrightarrow{P^*} 0,$$

which concludes the proof by Theorem 1 of *Chernozhukov and Hong* (2003).

□

5.9.4.3 Proposition 7: Posterior Moments

We focus on the scenario where the dimension p is fixed. We show the posterior moments converge to the moments of the limiting distribution in Theorem V.1 at a $(\sqrt{n}, n\lambda)$-rate. Besides the convergence in total variation in Theorem V.1, it remains to verify a uniform integrability condition (*Van der Vaart*, 2000). Under the CA prior (5.5), recall that $p_n(\boldsymbol{\delta})$ is the posterior density, and recall from (IC.12)

$$f_n(\boldsymbol{\delta}) = T_n \cdot \exp\left\{-\frac{n}{2}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1)^T G_{11}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1) - sn\lambda^2 - n\lambda\|\boldsymbol{\delta}_2\|_1\right\},$$

Proof of Proposition 7. Define $A \cdot \boldsymbol{\delta} = (\sqrt{n}\boldsymbol{\delta}_1^T, n\lambda\boldsymbol{\delta}_2^T)^T$, it suffices to show

$$\int_{\mathbb{R}^p} \|A\boldsymbol{\delta}\|^\alpha \left| \frac{f_n(\boldsymbol{\delta})}{\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{p_n(\boldsymbol{\delta})}{\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right| \mathrm{d}\boldsymbol{\delta} \xrightarrow{P^*} 0$$

for $\alpha = 1, 2$. For a fixed M that will be given below, define the partition $\mathcal{A}_n = \{\boldsymbol{\delta} : \sqrt{n}\|\boldsymbol{\delta}_1\| \le M,\ n\lambda\|\boldsymbol{\delta}_2\| \le M\}$. On $\mathcal{A}_n$, Theorem V.1 implies

$$\begin{aligned} & \int_{\mathbb{R}^p} \|A\boldsymbol{\delta}\|^\alpha \left| \frac{f_n(\boldsymbol{\delta})}{\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{p_n(\boldsymbol{\delta})}{\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right| \mathrm{d}\boldsymbol{\delta} \\ \le\ & 2M^\alpha \int_{\mathcal{A}_n} \|A\boldsymbol{\delta}\|^\alpha \left| \frac{f_n(\boldsymbol{\delta})}{\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{p_n(\boldsymbol{\delta})}{\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right| \mathrm{d}\boldsymbol{\delta} \\ \xrightarrow{P^*}\ & 0 \end{aligned} \tag{IC.15}$$

On $\mathcal{A}_n^c$, we show that the moments of f_n and p_n are both negligible. First, the result for f_n follows from (ID.13) in the proof of Lemma IC.5. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and

$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{p-s})$ be distributed as

$$\begin{aligned} \xi_1, \ldots, \xi_{p-s} &\overset{i.i.d.}{\sim} \text{Laplace}\left(\frac{1}{n\lambda}\right), \\ \boldsymbol{Z} &\sim \text{N}\left(\widetilde{\boldsymbol{\delta}}_1,\ \frac{1}{n} G_{11}^{-1}\right), \end{aligned}$$

and $\boldsymbol{\xi}$ is independent of $\boldsymbol{Z}$. Since f_n proportional to the joint density function of $(\boldsymbol{Z}, \boldsymbol{\xi})$, we have

$$\begin{aligned} \int_{\mathcal{A}_n^c} \|A\boldsymbol{\delta}\|^\alpha f_n(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} &\leq \left(\frac{2\pi}{n\theta_{11}}\right)^{s/2} \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \left(\sqrt{E\|n\lambda\boldsymbol{\xi}\|^{2\alpha}} + \sqrt{E\|\sqrt{n}\boldsymbol{Z}\|^{2\alpha}}\right) \\ &\quad \cdot \sqrt{P\left(\|\boldsymbol{\xi}\|_1 \geq M/(n\lambda)\right) + P\left(\|\boldsymbol{Z}\|_2 \geq M/\sqrt{n}\right)} \\ &\leq 2Q_n \left(\frac{2\pi}{n\theta_{11}}\right)^{s/2} \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \exp(-M^2\theta_{11}/8), \end{aligned}$$

with probability at least $1 - \gamma$, where the last inequality follows from computing the moments of Normal and Laplace distributions, and

$$Q_n = [8(p-s)]^{\alpha/2} + \left(n\|\widetilde{\boldsymbol{\delta}}_1\| + \frac{2s}{\theta_{\min}}\right)^{\alpha/2} = O_P(1).$$

Note also

$$\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} = \left(\frac{2\pi}{n}\right)^{s/2} \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{2}{n\lambda}\right)^{p-s}.$$

Comparing the above two displayed inequalities gives

$$\frac{\int_{\mathcal{A}_n^c} \|A\boldsymbol{\delta}\|^\alpha f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \leq \varepsilon, \tag{IC.16}$$

with probability at least $1 - \gamma$.

For $p_n(\boldsymbol{\delta})$, using an argument similar to (IC.16) above, it follows from Lemma

IC.2 that for any ε, γ, there is a constant M such that

$$\frac{\int_{\mathcal{A}_n^c} \|A\boldsymbol{\delta}\|^\alpha p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \quad \leq \quad \varepsilon, \tag{IC.17}$$

with probability at least $1-\gamma$.

The proof is complete by combining (IC.15) through (IC.17).

□

5.9.5 Proof under the AL prior

In this subsection, we prove Theorem V.2 in Section 5.3, where the dimension p is kept fixed.

Here we review some of the notations, most of which are carried from the proof of Theorem V.3 in the previous subsection. Recall $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2^0)$ as the true quantile regression coefficients, with $\boldsymbol{\beta}_2^0 = \mathbf{0}$ under Assumption E.5. We shall write $\boldsymbol{\phi} = [\phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0)]_{i=1}^n$ as a vector. Recall from Assumption E.3 that $G = \mathrm{E}^*[x_i x_i^T f_{y|x}(x_i^T\boldsymbol{\beta}^0)]$, $D = \mathrm{E}^*[x_i x_i^T]$, as well as the sub-matrices $G_{k\ell}$ for $k = 1, 2$. By Assumption E.3, we define two constants $\theta_m = \theta_{\min}(G)$ and $\theta_M = \theta_{\max}(G)$. Finally, we define $\Delta_p = G^{-1}\boldsymbol{X}^T\boldsymbol{\phi}$ and $\Delta_s = G_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{\phi}$, where $\boldsymbol{X}_1$ is the design matrix of the active covariates.

In addition, let $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ be the classic quantile regression estimator, and let $\boldsymbol{w} = (w_1, \ldots, w_p)$ with $w_j = \sqrt{n}\lambda/|\hat{\beta}_j|$. The adaptive Lasso prior (5.4) is then

$$\pi_{AL}(\boldsymbol{\beta}) = \exp\left\{-\sum_{j=1}^{p} w_j|\beta_j|\right\}.$$

We shall partition $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$, where $\boldsymbol{w}_1 \in \mathbb{R}^s$ corresponds to the active coefficients. Furthermore, let $w_{\min} = \min\{w_j : j = s+1, \ldots, p\}$ and $w_{\max} = \max\{w_j : j = s+1, \ldots, p\} = \|\boldsymbol{w}_2\|_\infty$. Under the adaptive Lasso prior (5.4), the posterior density (up to a normalization constant) with respect to $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$ is

$$p_n(\boldsymbol{\delta}) = \pi_{AL}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \cdot \exp\left\{L_n(\boldsymbol{\beta}^0) - L_n\left(\boldsymbol{\delta} + \boldsymbol{\beta}^0\right)\right\}.$$

Similar to the proof of Theorem V.3, we prove the two parts of Theorem V.2 separately in the following subsections.

5.9.5.1 Part 1 of Theorem V.2: Adaptive rate of posterior consistency

Here we prove Part 1 of Theorem V.2, which we re-phrase as the following theorem.

Theorem ID.1. *Consider the adaptive Lasso prior (5.4). Suppose Assumptions E.1 through E.5 hold, and the tuning parameter λ_n satisfies*

$$\frac{1}{\sqrt{n}} \ll \lambda_n \ll 1.$$

For any sequence $M_n \to \infty$, we define

$$\mathcal{B}_n = \left\{ \boldsymbol{\delta} : \|\boldsymbol{\delta}_1\|_2 \leq \frac{M_n}{\sqrt{n}} \,,\ \|\boldsymbol{\delta}_2\|_\infty \leq \frac{M_n}{n\lambda} \right\}.$$

Then, we have

$$\Pi\left(\boldsymbol{\delta} \in \mathcal{B}_n^C \,\middle|\, \mathbb{D}_n \right) \xrightarrow{P^*} 0,$$

To facilitate the proof of Theorem ID.1, we need the following two lemmas, the proof of which are deferred to Section 5.9.6.

Lemma ID.1 (Lower bounding the denominator – adaptive Lasso prior)**.** *Under the conditions of Theorem ID.1, and suppose $\boldsymbol{\beta}_1^0 > \mathbf{0}$ holds element-wise. Then we have*

$$\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta}) \,\mathrm{d}\boldsymbol{\delta} \gtrsim_{\mathrm{P}^*} \prod_{j=s+1}^{p} \left(\frac{2}{w_j} \right) \cdot \left(\frac{2\pi}{n\theta_M} \right)^{s/2} \cdot \exp\left\{ -\boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 + \frac{n}{2} \Delta_s^T G_{11} \Delta_s \right\},$$

where θ_M is the maximum eigenvalue of G.

Lemma ID.2 (Sign consistency of the adaptive Lasso prior)**.** *Under the conditions of Theorem ID.1, we have*

$$\Pi\left(sgn(\boldsymbol{\beta}_1) = sgn(\boldsymbol{\beta}_1^0) \,\middle|\, \mathbb{D}_n \right) \xrightarrow{P^*} 1,$$

where the equality of sign functions holds element-wise.

Now we are ready to prove Theorem ID.1.

Proof of Theorem ID.1. The proof is similar to, but simpler than, the proof of The-

orem IC.1; because we only consider the regime where p is fixed. Without loss of generality, we assume that the true values of the active coefficients are all positive, i.e., $\boldsymbol{\beta}_1^0 > \mathbf{0}$. The results under other scenarios holds by symmetry.

For the constant q_0 in Lemma IA.8, we define the following regions for $\boldsymbol{\delta}$:

$$\mathcal{A}_n = \{\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0 : sgn(\boldsymbol{\beta}_1) = sgn(\boldsymbol{\beta}_1^0)\ \},$$
$$\mathcal{C}_n = \{\boldsymbol{\delta} : \|G^{1/2}(\beta - \beta^0)\| \leq 4q_0\},$$

where $sgn(\boldsymbol{\beta}_1^0)$ is the signs for true values of active coefficients. In view of Lemma ID.2, the posterior probability of $\mathcal{A}_n^C$ converges to 0 in P*-probability; therefore, it suffices to show the posterior probability are $o_{P^*}(1)$ for each of the areas below:

1. $\mathcal{C}_n \cap \mathcal{B}_n \cap \mathcal{A}_n$.
2. $\mathcal{C}_n^C \cap \mathcal{A}_n$.

In step I below, we give upper bounds for the posterior density; in step II and III, we bound the posterior integrals of p_n on each of the two areas above; then we compute the posterior probabilities in step IV; step V contains some auxiliary calculations that supplement the proof.

Since we only consider $\boldsymbol{\delta} \in \mathcal{A}_n$ throughout the proof, the adaptive Lasso prior becomes

$$\pi_{AL}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\left\{-\boldsymbol{w}_1^T(\boldsymbol{\beta}_1^0 + \boldsymbol{\delta}_1) - \sum_{j=s+1}^{p} w_j|\delta_j|\right\},$$

where $\boldsymbol{w}_1 = (w_1, \ldots, w_s)$; since we assume the true values for active coefficients are all positive.

Step I: Bounding the posterior density $p_n(\boldsymbol{\delta})$ We give two different upper bounds for $p_n(\boldsymbol{\delta})$, one for $\boldsymbol{\delta} \in \mathcal{C}_n \cap \mathcal{A}_n$, and the other for $\boldsymbol{\delta} \in \mathcal{C}_n^C \cap \mathcal{A}_n$.

When $\boldsymbol{\delta} \in \mathcal{C}_n^C \cap \mathcal{A}_n$, we rely on (IC.3) to upper bound the working likelihood, as

in the proof of Theorem V.3; since the adaptive Lasso prior is upper bounded by 1, we have

$$\begin{aligned} p_n(\boldsymbol{\delta}) &\leq_{\mathrm{P}^*} \exp\left\{-\frac{n\varepsilon_0\|G^{1/2}\boldsymbol{\delta}\|_2}{q_0}\right\} \\ &\triangleq \overline{p}_{1n}(\boldsymbol{\delta}), \end{aligned}$$

uniformly when $\boldsymbol{\delta} \in \mathcal{A}_n$.

When $\boldsymbol{\delta} \in \mathcal{C}_n \cap \mathcal{A}_n$, we rely on (IC.4) to bound the working likelihood; since $\boldsymbol{\delta}^T G \boldsymbol{\delta} \geq \theta_m \boldsymbol{\delta}_1^T \boldsymbol{\delta}_1$, we have

$$\begin{aligned} L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\beta}^0 + \boldsymbol{\delta}) &\leq_{\mathrm{P}^*} -\frac{n\theta_m}{8}\boldsymbol{\delta}_1^T\boldsymbol{\delta}_1 + \boldsymbol{\phi}^T\boldsymbol{X}\boldsymbol{\delta} + \frac{1}{8}\underline{f} \\ &\leq -\frac{n\theta_m}{8}\left\|\boldsymbol{\delta}_1 - \frac{4}{n\theta_m}\boldsymbol{X}_1^T\boldsymbol{\phi}\right\|^2 + \frac{2}{n\theta_m}\|\boldsymbol{X}_1^T\boldsymbol{\phi}\|^2 \\ &\quad + \|\boldsymbol{X}_2^T\boldsymbol{\phi}\|_\infty \cdot \|\boldsymbol{\delta}_2\|_1 + \frac{1}{8}\underline{f}, \end{aligned} \tag{ID.1}$$

which followed by completing the squares for $\boldsymbol{\delta}_1$. Combining the above equation with the adaptive Lasso prior on $\mathcal{A}_n$, we have

$$\begin{aligned} p_n(\boldsymbol{\delta}) &\lesssim_{\mathrm{P}^*} \exp\left\{\frac{2}{n\theta_m}\|\boldsymbol{X}_1^T\boldsymbol{\phi}\|^2 - \boldsymbol{w}_1^T\boldsymbol{\beta}_1^0\right\} \cdot \exp\left\{-\frac{n\theta_m}{8}\|\boldsymbol{\delta}_1 - \boldsymbol{\mu}_1\|^2 - \boldsymbol{w}_1^T\boldsymbol{\delta}_1\right\} \\ &\quad \cdot \exp\left\{-\sum_{j=s+1}^{p}(w_j - \alpha_n)|\delta_j|\right\} \\ &\triangleq \exp\left\{\frac{2}{n\theta_m}\|\boldsymbol{X}_1^T\boldsymbol{\phi}\|^2 - \boldsymbol{w}_1^T\boldsymbol{\beta}_1^0\right\} \cdot \overline{p}_{2n}(\boldsymbol{\delta}), \end{aligned} \tag{ID.2}$$

uniformly on $\boldsymbol{\delta} \in \mathcal{C}_n \cap \mathcal{A}_n$, where $\alpha_n = \|\boldsymbol{\phi}^T\boldsymbol{X}_k\|_\infty$ and $\boldsymbol{\mu}_1 = 4\boldsymbol{X}_1^T\boldsymbol{\phi}/(n\theta_m)$.

Step II: Bound on $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n$ Here we bound the posterior integral of $p_n(\boldsymbol{\delta})$ by using its upper bound $\overline{p}_{2n}(\boldsymbol{\delta})$. Let $\gamma > 0$ be a small enough constant, we define the event $E_1(\gamma) = \{\alpha_n \leq \gamma \cdot w_{\min}\}$, where α_n is defined in the end of step I.

Similar to the proof of Theorem IC.1, we relate the integration to probabilistic

calculations. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and $\boldsymbol{\xi} = (\xi_{s+1}, \ldots, \xi_p)$ be distributed as

$$\begin{aligned} \xi_j &\overset{ind.}{\sim} \text{Laplace}\left(\frac{1}{w_j - \alpha_n}\right), \quad j = s+1, \ldots, p, \\ \boldsymbol{Z} &\sim \text{N}\left(\boldsymbol{\mu}_1, \frac{4}{n\theta_m}\boldsymbol{I}_s\right), \end{aligned}$$

where $\boldsymbol{Z}$ and $\boldsymbol{\xi}$ are independent. On the event E_1, the function $\overline{p}_{2n}(\boldsymbol{\delta})$ can be related with the moments of $(\boldsymbol{Z}, \boldsymbol{\xi})$; using a similar argument to (IC.6), we have:

$$\begin{aligned} & \int_{\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} \\ \lesssim_{\mathrm{P}^*} \quad & \exp\left\{\frac{2}{n\theta_m}\|\boldsymbol{X}_1^T\boldsymbol{\phi}\|^2 - \boldsymbol{w}_1^T\boldsymbol{\beta}_1^0\right\} \cdot \left(\frac{8\pi}{n\theta_m}\right)^{s/2} \cdot \prod_{j=s+1}^{p}\left(\frac{2}{w_j - \alpha_n}\right) \\ & \cdot \mathrm{E}\left(\exp\left\{-\boldsymbol{w}_1^T\boldsymbol{Z}\right\} \cdot \mathbf{1}\left[\|\boldsymbol{Z}\| \geq \frac{M_n}{\sqrt{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \geq \frac{M_n}{n\lambda}\right]\right) \\ \leq \quad & \exp\left\{\frac{2}{n\theta_m}\|\boldsymbol{X}_1^T\boldsymbol{\phi}\|^2 - \boldsymbol{w}_1^T\boldsymbol{\beta}_1^0\right\} \cdot \left(\frac{8\pi}{n\theta_m}\right)^{s/2} \cdot \prod_{j=s+1}^{p}\left(\frac{2}{w_j - \alpha_n}\right) \\ & \cdot \sqrt{\mathrm{E}\left(\exp\left\{-2\boldsymbol{w}_1^T\boldsymbol{Z}\right\}\right) \cdot \Pr\left(\|\boldsymbol{Z}\| \geq \frac{M_n}{\sqrt{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \geq \frac{M_n}{n\lambda}\right)}, \end{aligned} \tag{ID.3}$$

where the last inequality follows from Cauchy-Schwarz inequality. Next we bound the expectation and probability terms in (ID.3) separately.

For the expectation term with respect to $\boldsymbol{Z}$, we have by Lemma IA.3 that

$$\mathrm{E}\left(\exp\left\{-2\boldsymbol{w}_1^T\boldsymbol{Z}\right\}\right) \leq (1 + 4\gamma),$$

which holds on the event $E_2(\gamma) = \left\{\|\boldsymbol{w}_1\| \leq \gamma \cdot \left(\sqrt{2n\theta_m} \wedge \|\boldsymbol{\mu}_1\|^{-1}\right)\right\}$.

For the probability term in (ID.3), we break it into two parts. First, Lemma IA.1 gives

$$\Pr\left(\|\boldsymbol{Z}\| \geq \frac{M_n}{\sqrt{n}}\right) \leq \exp\left\{-\frac{M_n^2\theta_m}{16}\right\},$$

which holds on the event $E_3(\gamma) = \left\{ n \cdot \boldsymbol{\mu}_1^T \boldsymbol{\mu}_1 \leq \gamma \cdot M_n^2 \right\}$; second, Lemma IA.2 gives

$$\begin{aligned} \Pr\left(\|\boldsymbol{\xi}\|_\infty \geq \frac{M_n}{n\lambda} \right) &\leq \Pr\left(\|\boldsymbol{\xi}\|_\infty \geq \sqrt{M_n} \cdot \frac{1}{n\lambda/\sqrt{M_n}} \right) \\ &\leq p \cdot \exp\left\{ -\frac{\sqrt{M_n}}{2} \right\}, \end{aligned}$$

which holds on the event $E_4(\gamma) = \left\{ \sqrt{M}_n \cdot (w_{\min} - \alpha_n) \geq \gamma \cdot n\lambda \right\}$. Therefore, the tail probability in (ID.3) is bounded by

$$\Pr\left(\|\boldsymbol{Z}\| \geq \frac{M_n}{\sqrt{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \geq \frac{M_n}{n\lambda} \right) \lesssim \exp\left\{ -\frac{\sqrt{M_n}}{2} \right\}, \tag{ID.4}$$

since p is constant and $M_n \to \infty$; the equation holds on the events $E_3(\gamma)$ and $E_4(\gamma)$.

Similar to (IC.8), we can further simplify (ID.3) by showing that

$$\prod_{j=s+1}^{p} \left(\frac{2}{w_j - \alpha_n} \right) \lesssim \prod_{j=s+1}^{p} \left(\frac{2}{w_j} \right), \tag{ID.5}$$

on the event $E_1(\gamma)$. Therefore, from (ID.3), the posterior integral on $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n$ is bounded from above by

$$\begin{aligned} \int_{\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} \quad \lesssim_{\mathrm{P}^*} \quad & \exp\left\{ \frac{n\theta_m}{8} \|\boldsymbol{\mu}_1\|^2 - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 \right\} \\ & \cdot \left(\frac{8\pi}{n\theta_m} \right)^{s/2} \cdot \prod_{j=s+1}^{p} \left(\frac{2}{w_j} \right) \cdot \exp\left\{ -\frac{\sqrt{M_n}}{4} \right\}, \end{aligned}$$

on the events $E_1(\gamma)$ through $E_4(\gamma)$, where $\boldsymbol{\mu}_1$ is defined in step I.

Step III: Bound on $\mathcal{C}_n^C \cap \mathcal{A}_n$ On $\mathcal{C}_n^C$, we use the upper bound $\overline{p}_{1n}$ in step I. Similar to (IC.9), the posterior integral on $\mathcal{C}_n^C \cap \mathcal{A}_n$ is bounded from above by

$$\begin{aligned}\int_{\mathcal{C}_n^C \cap \mathcal{A}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} &\lesssim_{\mathrm{P}^*} \int_{\mathcal{C}_n^C} \overline{p}_{1n}(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \\ &\leq \exp\{p - n\varepsilon_0\} \cdot \frac{1}{\sqrt{|G|}} \cdot \left(\frac{4q_0}{\sqrt{p}}\right)^p \\ &\lesssim \exp\{-n\varepsilon_0/4\}, \end{aligned} \tag{ID.6}$$

where the last inequality follows since p is constant, and that $|G| \geq \theta_{\min}(G)^p$ is bounded away from 0.

Step IV: Bounding posterior probabilities Here we close the proof by showing the posterior probability of both $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n$ and $\mathcal{C}_n^C \cap \mathcal{A}_n$ are $o_{P^*}(1)$. From Lemma ID.1, we have

$$\begin{aligned}\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} &\gtrsim_{\mathrm{P}^*} \prod_{j=s+1}^{p} \left(\frac{2}{w_j}\right) \cdot \left(\frac{2\pi}{n\theta_M}\right)^{s/2} \cdot \exp\left(-\boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 + \frac{n}{2}\Delta_s^T G_{11} \Delta_s\right) \\ &\triangleq \widetilde{P}_n. \end{aligned}$$

Similar to the proof of Theorem IC.1, it suffices to show that the integral of $p_n(\boldsymbol{\delta})$ on both $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n$ and $\mathcal{C}_n^C \cap \mathcal{A}_n$ are are $o_{P^*}(\widetilde{P}_n)$

Let $E_5(\gamma) = \{w_{\max} \leq \gamma \cdot n\}$. In step V later, we shall show that the events $E_1(\gamma)$ through $E_5(\gamma)$ holds with P*-probability tending to 1, for small enough γ. Therefore the bounds derived in steps II and III holds with P*-probability tending to 1.

For the area $\mathcal{C}_n \cap \mathcal{B}_n^C \cap \mathcal{A}_n$, we use the upper bound displayed at the end of step

II; comparing it with $\widetilde{P}_n$ gives

$$\begin{aligned}\frac{\int_{\mathcal{C}_n\cap\mathcal{B}_n^C\cap\mathcal{A}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} \left(\frac{4\theta_M}{\theta_m}\right)^{s/2}\cdot\exp\left\{\frac{n\theta_m}{8}\|\boldsymbol{\mu}_1\|^2-\frac{\sqrt{M_n}}{4}\right\}\\ &\lesssim_{\mathrm{P}^*} \exp\left\{-\frac{\sqrt{M}_n}{8}\right\},\end{aligned}$$

since s and θ_m are both bounded; the last inequality holds on the event $E_3(\gamma)$ for small enough γ.

To bound the posterior probability of the area $\mathcal{C}_n^C\cap\mathcal{A}_n$, we first simplify $\widetilde{P}_n$. Since the events $E_5(\gamma)$ and $E_2(\gamma)$ both have P^*-probability going to 1, we have $\|\boldsymbol{w}_1\|\lesssim\sqrt{n}$ and $w_j\le\gamma\cdot n$ for all $j=s+1,\ldots,p$; therefore,

$$\begin{aligned}\widetilde{P}_n &\gtrsim_{\mathrm{P}^*} \left(\frac{2}{\gamma n}\right)^{p-s}\cdot\left(\frac{2\pi}{n\theta_M}\right)^{s/2}\cdot\exp\left\{-\|\boldsymbol{w}_1\|\cdot\|\boldsymbol{\beta}_1^0\|\right\}\\ &\ge \left(\frac{C_1}{n}\right)^p\exp\{-C_2\sqrt{n}\},\end{aligned}$$

for some constant $C_1,C_2>0$, since $\|\boldsymbol{\beta}_1^0\|=O(1)$. Comparing $\widetilde{P}_n$ with the posterior integral in step III, we have

$$\begin{aligned}\frac{\int_{\mathcal{C}_n^C\cap\mathcal{A}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} \exp\left\{p\log n+C_2\sqrt{n}-\frac{n\varepsilon_0}{4}\right\}\\ &\lesssim \exp\left\{-\frac{n\varepsilon_0}{8}\right\},\end{aligned}\tag{ID.7}$$

for large enough n.

Therefore, the proof is now complete.

Step V: Auxiliary calculations Now we show that each of the events

$$E_1(\gamma) = \{\alpha_n \le \gamma \cdot w_{\min}\}, \quad E_2(\gamma) = \left\{\|\boldsymbol{w}_1\| \le \gamma \cdot \left(\sqrt{2n\theta_m} \wedge \frac{1}{\|\boldsymbol{\mu}_1\|}\right)\right\},$$
$$E_3(\gamma) = \left\{n \cdot \boldsymbol{\mu}_1^T \boldsymbol{\mu}_1 \le \gamma \cdot M_n^2\right\}, \quad E_4(\gamma) = \left\{\sqrt{M}_n \cdot (w_{\min} - \alpha_n) \ge \gamma \cdot n\lambda\right\},$$
$$\text{and} \quad E_5(\gamma) = \{w_{\max} \le \gamma \cdot n\},$$

holds with P*-probability going to 1, where $\boldsymbol{\mu}_1 = 4\boldsymbol{X}_1^T\boldsymbol{\phi}/(n\theta_m)$ and $\alpha_n = \|\boldsymbol{X}_2^T\boldsymbol{\phi}\|_\infty$ are defined in step I. From standard asymptotic results for quantile regression (*Koenker*, 2005, Section 4.2), we have

$$\begin{aligned}
(\sqrt{n})^{-1} \cdot \max_{s+1 \le j \le p} \left\{|\hat{\beta}_j|^{-1}\right\} &= O_{P^*}(1), \\
\sqrt{n}\|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^0\|_\infty &= O_{P^*}(1), \\
\sqrt{n} \cdot \left(\sum_{j=1}^{s} |\hat{\beta}_j|^{-2} - \sum_{j=1}^{s} |\beta_j^0|^{-2}\right) &= O_{P^*}(1),
\end{aligned} \tag{ID.8}$$

since s and p are both fixed.

First we consider the event $E_1(\gamma)$. For $w_{\min}$ we have

$$\frac{1}{w_{\min}} = \frac{\sqrt{n}\|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^0\|_\infty}{n\lambda} = O_{P^*}\left(\frac{1}{n\lambda}\right),$$

from (ID.8). Furthermore, Lemma IC.3 implies that $\alpha_n = O_{P^*}(\sqrt{n})$. Since $\sqrt{n}\lambda \to \infty$, it follows that $\alpha_n/w_{\min} = o_{P^*}(1)$, which further implies $\mathrm{P}^*(E_1(\gamma)) \to 1$.

Next we consider the event $E_3(\gamma)$. Since $\boldsymbol{\mu}_1 = 4\sum_{i=1}^n x_{1i}\phi_i/(n\theta_m)$, the Central Limit Theorem gives $\boldsymbol{\mu}_1 = O_{P^*}(1/\sqrt{n})$. Therefore $\mathrm{P}^*(E_3(\gamma)) \to 1$ follows.

For the event $E_2(\gamma)$, again from (ID.8), we have

$$\|\boldsymbol{w}_1\|^2 = n\lambda^2 \sum_{j=1}^{s} |\hat{\beta}_j|^{-2} = O_{P^*}(n\lambda^2).$$

Furthermore, since $\|\boldsymbol{\mu}_1\| = O_{P^*}(1/\sqrt{n})$, it follows that $\mathrm{P}^*(E_2(\gamma)) \to 1$, as $\lambda \to 0$.

Now we consider the event $E_4(\gamma)$. Given that the event $E_1(\gamma)$ holds with P^*-probability tending to 1, it suffices to show

$$\mathrm{P}^*\left(\sqrt{M}_n \cdot w_{\min} \leq \gamma \cdot n\lambda\right) = \mathrm{P}^*\left(\frac{1}{w_{\min}} \geq \frac{\sqrt{M}_n}{\gamma \cdot n\lambda}\right) \to 0,$$

for any small enough $\gamma > 0$; the desired result is then implied by the event $E_1(\gamma)$.

Finally, for the event $E_5(\gamma)$, we have from (ID.8) that

$$w_{\max} = \sqrt{n}\lambda \cdot O_{P^*}(\sqrt{n}) = o_{P^*}(n),$$

which implies $\mathrm{P}^*(E_5(\gamma)) \to 1$, as $\lambda \to 0$.

□

5.9.5.2 Part 2 of Theorem V.2: Distributional Approximation

We first introduce some additional notations. We shall continue to use the notations in the previous subsection, i.e., the proof of Theorem ID.1. Let $\widetilde{\boldsymbol{\beta}}_1$ be the oracle estimator using the true model, and let $\widetilde{\boldsymbol{\delta}}_1 = \widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0$. Furthermore, let $\log T_n = n\widetilde{\boldsymbol{\delta}}_1^T G_{11}\widetilde{\boldsymbol{\delta}}_1/2 - \sum_{j=1}^s w_j|\beta_j^0|$ and $\log S_n = n\Delta_p^T G\Delta_p/2 - \sum_{j=1}^s w_j|\beta_j^0|$. We define the following functions

$$\begin{aligned} h_n(\boldsymbol{\delta}) &= S_n \cdot \exp\left\{-\frac{n}{2}(\boldsymbol{\delta} - \Delta_p)^T G(\boldsymbol{\delta} - \Delta_p) - \sum_{j=s+1}^p w_j|\delta_j|\right\} \\ f_n(\boldsymbol{\delta}) &= T_n \cdot \exp\left\{-\frac{n}{2}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1)^T G_{11}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1) - \sum_{j=s+1}^p w_j|\delta_j|\right\}. \end{aligned} \tag{ID.9}$$

Using the notations above, we re-phrase part 2 of Theorem V.2 below.

Theorem ID.2. *Suppose the conditions of Theorem ID.1 hold. Recalling that* $p_n(\boldsymbol{\delta})$

is the posterior density function for $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$*; we have*

$$\left\| \frac{p_n}{\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{f_n}{\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right\|_{TV} \xrightarrow{P^*} 0.$$

where f_n *is defined in (ID.9).*

To prove Theorem ID.2, we need the following lemma, the proof of which is deferred to Section 5.9.6.

Lemma ID.3 (Normal likelihood with adaptive Lasso shrinkage)**.** *Under the conditions of Theorem ID.2, we have the following:*

$$\left\| \frac{f_n(\boldsymbol{\delta})}{\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{h_n(\boldsymbol{\delta})}{\int h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right\|_{TV} \xrightarrow{P^*} 0,$$

where f_n *and* h_n *are defined in (ID.9).*

Now we are ready to prove Theorem ID.2.

Proof of Theorem ID.2. Without loss of generality, we assume $\boldsymbol{\beta}_1^0 > \mathbf{0}$, i.e., the true values of the active coefficients are all positive. In what follows, we shall write $\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$ for integrating a function f_n on $\mathbb{R}^p$. In view of Lemma ID.3, we only need to show that p_n converges to h_n,

$$\left\| \frac{p_n(\boldsymbol{\delta})}{\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{h_n(\boldsymbol{\delta})}{\int h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right\|_{TV} \xrightarrow{P^*} 0.$$

Similar to the proof of Theorem IC.2, it suffices to show

$$\frac{\int_{\mathbb{R}^p} |p_n(\boldsymbol{\delta}) - h_n(\boldsymbol{\delta})|\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n \vee \int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \xrightarrow{P^*} 0,$$

where

$$\widetilde{P}_n = \prod_{j=s+1}^{p} \left(\frac{2}{w_j}\right) \cdot \left(\frac{2\pi}{n\theta_M}\right)^{s/2} \cdot \exp\left\{-\boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 + \frac{n}{2}\Delta_s^T G_{11} \Delta_s\right\},$$

as in Lemma ID.1.

Fix a diverging sequence $K_n \to +\infty$, we define

$$\mathcal{B}_n = \left\{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_1\|_2 \leq \frac{K_n}{\sqrt{n}} \text{ , and } \|\boldsymbol{\delta}_2\|_\infty \leq \frac{K_n}{n\lambda}\right\};$$

we shall specify K_n later. In the following, we upper bound the integral of $|p_n - h_n|$ on $\mathcal{B}_n$ and its complement, separately in steps I and II.

Step I: Bounding $\int |p_n - h_n| \,\mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n$ When $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0 \in \mathcal{B}_n$, the adaptive Lasso prior becomes

$$\pi_{SCAD}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\left\{-\boldsymbol{w}_1^T(\boldsymbol{\delta}_1 + \boldsymbol{\beta}_1^0) - \sum_{j=s+1}^{p} w_j|\delta_j|\right\},$$

which holds since $\boldsymbol{\delta}_1 + \boldsymbol{\beta}_1^0 > \mathbf{0}$ on $\mathcal{B}_n$. Similar to the proof of Theorem IC.2, we have

$$\begin{aligned}
\sup_{\boldsymbol{\delta}\in\mathcal{B}_n} \left|\log\left(\frac{h_n(\boldsymbol{\delta})}{p_n(\boldsymbol{\delta})}\right)\right| &= \sup_{\boldsymbol{\delta}\in\mathcal{B}_n} \left|L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta}+\boldsymbol{\beta}^0) + \frac{n}{2}\boldsymbol{\delta}^T G \boldsymbol{\delta} - n\cdot\Delta_p^T G\boldsymbol{\delta}\right| \\
&\quad + \sup_{\boldsymbol{\delta}\in\mathcal{B}_n} \left|\boldsymbol{w}_1^T\boldsymbol{\delta}_1\right| \\
&= o_{P^*}(1);
\end{aligned}$$

the first supremum is $o_{P^*}(1)$ if we choose K_n that satisfies Corollary IB.1; and the second supremum is $o_{P^*}(1)$ if $K_n \ll 1/\lambda$, as $\boldsymbol{w}_1 = O_{P^*}(\sqrt{n}\lambda)$ in (ID.8). Thus, we have $|h_n(\boldsymbol{\delta})/p_n(\boldsymbol{\delta}) - 1| = o_{P^*}(1)$ uniformly when $\boldsymbol{\delta} \in \mathcal{B}_n$, which further implies

$$\begin{aligned}
\int_{\mathcal{B}_n} |p_n(\boldsymbol{\delta}) - h_n(\boldsymbol{\delta})| \,\mathrm{d}\boldsymbol{\delta} &= \int_{\mathcal{B}_n} p_n(\boldsymbol{\delta}) \left|1 - \left(\frac{h_n(\boldsymbol{\delta})}{p_n(\boldsymbol{\delta})}\right)\right| \mathrm{d}\boldsymbol{\delta} \\
&= o_{P^*}\left(\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}\right).
\end{aligned}$$

Step II: Bounding $\int |p_n - h_n| \, \mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n^C$ Here we control the integration of $|p_n - h_n|$ on $\mathcal{B}_n^C$ by showing that that both $\int_{\mathcal{B}_n^C} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}$, and $\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}$ are $o_{P^*}(\widetilde{P}_n)$.

For p_n, Theorem ID.1 directly implies that

$$\int_{\mathcal{B}_n^C} p_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} = o_{P^*}\left(\widetilde{P}_n\right).$$

For h_n, we first provide an upper bound before computing the integral. Note that $\boldsymbol{\delta}^T G \boldsymbol{\delta} \geq \theta_m \boldsymbol{\delta}_1^T \boldsymbol{\delta}_1$, and $G \cdot \Delta_p = \boldsymbol{X}^T \boldsymbol{\phi}/n$; similar to (ID.2), we have

$$\begin{aligned} h_n(\boldsymbol{\delta}) &\leq \exp\left\{ \frac{n\theta_m}{2} \|\boldsymbol{\mu}_1\|^2 - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 - \frac{n\theta_m}{2} (\boldsymbol{\delta}_1 - \boldsymbol{\mu}_1)^T (\boldsymbol{\delta}_1 - \boldsymbol{\mu}_1) - \sum_{j=s+1}^{p} (w_j - \alpha_n) |\delta_j| \right\} \\ &\triangleq \exp\left\{ \frac{n\theta_m}{2} \|\boldsymbol{\mu}_1\|^2 - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 \right\} \cdot \overline{h}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2), \end{aligned}$$

where $\alpha_n = \|\boldsymbol{X}_2^T \boldsymbol{\phi}\|_\infty$ and $\boldsymbol{\mu}_1 = \boldsymbol{X}_1^T \boldsymbol{\phi}/(n\theta_m)$.

Now we can relate the integration of $\overline{h}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ to Gaussian and Laplace tail bounds. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and $\boldsymbol{\xi} = (\xi_{s+1}, \ldots, \xi_p)$ be distributed as

$$\begin{aligned} \xi_j &\overset{ind.}{\sim} \text{Laplace}\left(\frac{1}{w_j - \alpha_n}\right), \quad j = s+1, \ldots, p \\ \boldsymbol{Z} &\sim \text{N}\left(\boldsymbol{\mu}_1, \frac{1}{n\theta_m} \boldsymbol{I}_s\right), \end{aligned}$$

and $\boldsymbol{Z}$ is independent of $\boldsymbol{\xi}$; the function $\overline{h}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ is then proportional to the joint density of $(\boldsymbol{Z}, \boldsymbol{\xi})$. Following the arguments in (ID.3), we have

$$\begin{aligned} \frac{\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\leq \frac{1}{\widetilde{P}_n} \left[\exp\left\{ \frac{n\theta_m}{2} \|\boldsymbol{\mu}_1\|^2 - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 \right\} \cdot \prod_{j=s+1}^{p} \left(\frac{2}{w_j - \alpha_n} \right)^{p-s} \cdot \left(\frac{2\pi}{n\theta_m} \right)^{s/2} \right. \\ &\qquad \left. \cdot \Pr\left(\|\boldsymbol{Z}\|_2 \geq \frac{K_n}{\sqrt{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \geq \frac{K_n}{n\lambda} \right) \right] \\ &\lesssim \exp\left\{ \frac{n\theta_m}{2} \|\boldsymbol{\mu}_1\|^2 - \sqrt{K_n} \right\}, \end{aligned} \tag{ID.10}$$

where we bound the tail probability by (ID.4) and bound the product of $(w_j - \alpha_n)$ by (ID.5); the displayed equations hold on the events $E_1(\gamma) = \{\alpha_n \leq \gamma \cdot w_{\min}\}$, $E_2(\gamma) = \left\{n \cdot \boldsymbol{\mu}_1^T \boldsymbol{\mu}_1 \leq \gamma \cdot K_n^2\right\}$, and $E_3(\gamma) = \left\{\sqrt{K}_n \cdot (w_{\min} - \alpha_n) \geq \gamma \cdot n\lambda\right\}$, as required by (ID.4) and (ID.5).

As in the proof of Theorem ID.1 (Step V), we can show that all the events $E_1(\gamma)$ through $E_3(\gamma)$ have P*-probability tending to 1; therefore, the bounds in this step holds with P*-probability going to 1, which implies

$$\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} = o_{P^*}\left(\widetilde{P}_n\right),$$

as $K_n \to \infty$.

Combining steps I and II, we obtain

$$\frac{\displaystyle\int_{\mathbb{R}^p} |p_n(\boldsymbol{\delta}) - h_n(\boldsymbol{\delta})|\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n \vee \displaystyle\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \xrightarrow{P^*} 0,$$

which concludes the proof by Theorem 1 of *Chernozhukov and Hong* (2003).

□

5.9.6 Proof of some auxiliary results

5.9.6.1 Proof of Lemmas IC.3 and IC.4

Proof of Lemma IC.3. First note that ϕ_i and x_i are independent, as the conditional distribution is

$$\mathrm{P}^*\left(\phi_i = \tau \mid x_i\right) = \mathrm{P}^*\left(y_i < x_i^T \boldsymbol{\beta}^0 \mid x_i\right) = 1 - \tau,$$

which does not depend on x_i.

For each $k = 1, \ldots, q$, let e_k be the k-th unit vector in $\mathbb{R}^q$ and $v_k = D^{1/2} A^T e_k$. Under the conditions of the Lemma, it holds that $\|v_k\|^2 = e_k^T A D A^T e_k \leq C_0$. Therefore

for each $i = 1, \ldots, n$, we have

$$\begin{aligned} \Pr\left(|(Ax_i)_k| \geq z\right) &\leq \Pr\left(|v_k^T D^{-1/2} x_i| \geq z\right) \\ &\leq 2\exp\left\{-\frac{z}{C_0\sigma_0}\right\}, \end{aligned}$$

by Assumption E.4′. Therefore, we have verified that each component of Ax_i is sub-exponential.

Finally, by applying a union bound and conditioning on ϕ_i's,

$$\begin{aligned} \mathrm{P}^*\left(\left\|\sum_{i=1}^n \phi_i A x_i\right\|_\infty \geq M\sqrt{n\log q}\right) &\leq \sum_{k=1}^q \mathrm{E}^*_{\boldsymbol{\phi}}\left[\mathrm{P}^*\left(\sum_{i=1}^n |\phi_i(Ax_i)_k| \geq M\sqrt{n\log q} \;\middle|\; \boldsymbol{\phi}\right)\right] \\ &\leq q\cdot \mathrm{E}^*_{\boldsymbol{\phi}}\left[\exp\left\{-C_2\frac{M^2 n\log q}{\|\boldsymbol{\phi}\|^2}\right\}\right] \\ &\leq \exp\left\{\log(q) - C_2 M^2 \log q\right\} \\ &\leq \exp\left\{-(C_2/2)\cdot M^2\log(q)\right\}. \end{aligned}$$

where the second inequality holds for some constant $C_2 > 0$ by Lemma IA.7, as $\|\boldsymbol{\phi}\|_\infty \leq 1$; and the penultimate inequality holds since $\|\boldsymbol{\phi}\|^2 \leq n$. The displayed probability is then arbitrarily small by making M large. The proof is now complete.

□

Proof of Lemma IC.4. We only prove the result for $\Delta_s^T G_{11}\Delta_s$, the conclusion for $\Delta_p^T G\Delta_p$ follows in a similar fashion. Let $\phi_i = \phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0)$, and define $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]$. First note that x_i and ϕ_i are independent, as proved in Lemma IC.3. Therefore,

$$\mathrm{E}^*\left[\boldsymbol{\phi}\boldsymbol{\phi}^T \mid x_1, \ldots, x_n\right] = \tau(1-\tau)I_n,$$

where $\boldsymbol{I}_n$ is the n by n identity matrix.

Let $\boldsymbol{X}_1 = [x_{11}^T, \ldots, x_{1n}^T]^T$ and $D_{11} = \mathrm{E}^*[x_{1i}x_{1i}^T] = \mathrm{E}^*[\boldsymbol{X}_1^T\boldsymbol{X}_1]/n$. We can re-write $\Delta_s^T G_{11}\Delta_s = \boldsymbol{\phi}^T\boldsymbol{X}_1 G_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{\phi}/n^2 \geq 0$ since G_{11} is positive definite. By Chebyshev's

inequality and switching the expectation with trace, we have

$$\begin{aligned}
\mathrm{P}^*\left(\Delta_s^T G_{11}\Delta_s \geq M\frac{s}{n}\right) &\leq \frac{\mathrm{E}^*[\boldsymbol{\phi}^T\boldsymbol{X}_1 G_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{\phi}]/n^2}{M\cdot s/n} \\
&= \frac{\mathrm{tr}\left(\mathrm{E}^*[G_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{\phi}\boldsymbol{\phi}^T\boldsymbol{X}_1]\right)}{M\cdot sn} \\
&= \frac{n\tau(1-\tau)\cdot \mathrm{tr}\left(G_{11}^{-1}\cdot D_{11}\right)}{M\cdot sn} \\
&\lesssim \frac{n\tau(1-\tau)\cdot s/\underline{f}}{M\cdot sn} \\
&\lesssim \frac{1}{M},
\end{aligned}$$

where the third equality holds by conditioning on $\boldsymbol{X}_1$ first; and the penultimate inequality holds as $D_{11} \preceq G_{11}/\underline{f}$ as in Assumption E.2. The proof is now complete.

□

5.9.6.2 Proof of Lemmas IC.1, IC.2 and IC.5 under the CA prior

Proof of Lemma IC.1. We provide a lower bound of the integral by restricting to the area

$$\mathcal{B}_n = \{\boldsymbol{\delta}\in\mathbb{R}^p : \|\boldsymbol{\delta}_1\| \leq K_n\sqrt{s/n};\ \ \|\boldsymbol{\delta}_2\|_\infty \leq K_n \log p/(n\lambda)\},$$

where the sequence $K_n \to \infty$ satisfies the requirement in Corollary IB.2. We define

$$\widetilde{P}_n = \left(\frac{2\pi}{n}\right)^{s/2}\left(\frac{2}{n\lambda}\right)^{p-s}\cdot\frac{\exp\left(-sn\lambda^2 + n\Delta_s^T G_{11}\Delta_s/2\right)}{\sqrt{|G_{11}|}}.$$

In step I, we provide a lower bound for the posterior density

$$p_n(\boldsymbol{\delta}) = \pi_{CA}(\boldsymbol{\delta}+\boldsymbol{\beta}^0)\cdot\exp\left\{L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\beta}^0+\boldsymbol{\delta})\right\},$$

on the area $\boldsymbol{\delta}\in\mathcal{B}_n$, which is denoted by $\underline{p}_n$; and in step II we integrate $\underline{p}_n$ on $\mathcal{B}_n$ to conclude the proof.

Step I: Lower bounding the posterior density First we analyze the quantile loss function. By Corollary IB.2, we have

$$\begin{aligned} L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta} + \boldsymbol{\beta}^0) &= \boldsymbol{\phi}^T \boldsymbol{X} \boldsymbol{\delta} - \frac{n}{2} \boldsymbol{\delta}_1^T G_{11} \boldsymbol{\delta}_1 + o_{P^*}(1) \\ &\geq -\frac{n}{2} (\boldsymbol{\delta}_1 - \Delta_s)^T G_{11} (\boldsymbol{\delta}_1 - \Delta_s) \\ &\quad + \frac{n}{2} \Delta_s^T G_{11} \Delta_s - \left\| \boldsymbol{\phi}^T \boldsymbol{X}_2 \right\|_\infty \cdot \|\boldsymbol{\delta}_2\|_1 + o_{P^*}(1) \quad \text{(ID.11)} \end{aligned}$$

uniformly on $\mathcal{B}_n$, which follows by completing the squares for $\boldsymbol{\delta}_1$ and Holder's inequality. Next we analyze the CA prior. As in (IC.13), the CA prior for $\boldsymbol{\delta}$ is

$$\pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\{-sn\lambda^2 - n\lambda \|\boldsymbol{\delta}_2\|_1\},$$

when $\boldsymbol{\delta} \in \mathcal{B}_n$.

Combining (ID.11) with the CA prior above, the posterior density on $\mathcal{B}_n$ is bounded from below by

$$\begin{aligned} p_n(\boldsymbol{\delta}) &\gtrsim_{\mathrm{P}^*} \exp\left\{-sn\lambda^2 + \frac{n}{2} \Delta_s^T G_{11} \Delta_s\right\} \cdot \exp\left\{-\frac{n}{2} (\boldsymbol{\delta}_1 - \Delta_s)^T G_{11} (\boldsymbol{\delta}_1 - \Delta_s)\right\} \\ &\quad \cdot \exp\left\{-\left(n\lambda + \alpha_n\right) \|\boldsymbol{\delta}_2\|_1\right\} \\ &\triangleq \exp\left\{-sn\lambda^2 + \frac{n}{2} \Delta_s^T G_{11} \Delta_s\right\} \cdot \underline{p}_n(\boldsymbol{\delta}), \end{aligned}$$

where $\alpha_n = \|\boldsymbol{\phi}^T \boldsymbol{X}_2\|_\infty$ and $\Delta_s = G_{11}^{-1} \boldsymbol{X}_1^T \boldsymbol{\phi}$.

Step II: Bounding the posterior integration Now, we relate the integration of the lower bound $\underline{p}_n(\boldsymbol{\delta})$ to probabilistic calculations. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{p-s})$

be distributed as

$$\begin{aligned} \xi_1, \ldots, \xi_{p-s} &\overset{i.i.d.}{\sim} \text{Laplace}\left(\frac{1}{n\lambda + \alpha_n}\right), \\ \boldsymbol{Z} &\sim \text{N}\left(\Delta_s,\ \frac{1}{n} G_{11}^{-1}\right), \end{aligned}$$

and $\boldsymbol{Z}$ is independent of $\boldsymbol{\xi}$. Similar to the arguments that leads to (IC.6), we have

$$\begin{aligned} \int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta} \quad &\gtrsim_{\mathrm{P}^*} \quad \exp\left\{-sn\lambda^2 + \frac{n}{2}\Delta_s^T G_{11}\Delta_s\right\} \cdot \left(\frac{2}{n\lambda + \alpha_n}\right)^{p-s} \cdot \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{2\pi}{n}\right)^{s/2} \\ &\qquad \cdot \Pr\left(\|\boldsymbol{Z}\|_2 \le K_n\sqrt{\frac{s}{n}}\right) \cdot \Pr\left(\|\boldsymbol{\xi}\|_\infty \le \frac{K_n \log p}{n\lambda}\right) \\ &= \quad \left(\frac{1}{1 + \alpha_n/(n\lambda)}\right)^{p-s} \cdot \widetilde{P}_n \\ &\qquad \cdot \left[1 - \Pr\left(\|\boldsymbol{Z}\|_2 \ge K_n\sqrt{\frac{s}{n}} \ \text{ or } \ \|\boldsymbol{\xi}\|_\infty \ge \frac{K_n \log p}{n\lambda}\right)\right] \\ &\gtrsim \quad \widetilde{P}_n, \end{aligned}$$

where we rely on two techniques: (i) we bound the tail probabilities as in (IC.14); (ii) we bound the leading factor by $(1+x)^{-1} \ge \exp(-x)$; therefore the last inequality holds on the events

$$E_1(\gamma) = \{p\alpha_n \le \gamma \cdot n\lambda\}, \quad E_2(\gamma) = \{n\Delta_s^T G_{11}\Delta_s \le \gamma \cdot K_n^2 s\}.$$

for small enough $\gamma > 0$, as required in (IC.14).

Similar to the proof of Theorem IC.1 (Step V), we can show that the events $E_1(\gamma)$ and $E_2(\gamma)$ have P*-probability going to 1. Therefore, the proof is now complete that

$$\int_{\mathbb{R}^p} \pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \cdot \exp\left\{L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\beta}^0 + \boldsymbol{\delta})\right\}\, \mathrm{d}\boldsymbol{\delta} \gtrsim_{\mathrm{P}^*} \widetilde{P}_n.$$

□

Proof of Lemma IC.2. Let $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$; recall $p_n(\boldsymbol{\delta})$ is the posterior density under the CA prior (5.5), and

$$\mathcal{A}_n = \left\{ \boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0 : \min_{j=1,\ldots,s} |\beta_j| < \lambda, \text{ or } \max_{j=s+1,\ldots,p} |\beta_s| > \lambda \right\}.$$

In the following steps I – III, we shall upper bound the posterior integral of $p_n(\boldsymbol{\delta})$ on $\mathcal{A}_n \cap \mathcal{C}_n$ and $\mathcal{C}_n^C$ separately, where $\mathcal{C}_n = \{\|G^{1/2}\boldsymbol{\delta}\| \le 4q_0\}$, as defined in the proof of Theorem IC.1; and in step IV we verify their posterior probabilities are $o_{P^*}(1)$.

First we provide a decomposition of $\mathcal{A}_n$. Let $Q(\boldsymbol{\delta}) = \{j = 1, \ldots, s : |\delta_j + \beta_j^0| < \lambda\}$, and $R(\boldsymbol{\delta}) = \{j = s+1, \ldots, p : |\delta_j| > \lambda\}$ be two index sets; then $\mathcal{A}_n$ can be decomposed into

$$\begin{aligned} \mathcal{A}_n &\subseteq \bigcup_{\substack{0\le q\le s;\ 0\le r\le p-s \\ q+r>0}} \{\boldsymbol{\delta} : |Q(\boldsymbol{\delta})| = q, \ \ |R(\boldsymbol{\delta})| = r\} \\ &\triangleq \bigcup_{\substack{0\le q\le s;\ 0\le r\le p-s \\ q+r>0}} \mathcal{E}_{q,r}. \end{aligned} \tag{ID.12}$$

Therefore, to provide an upper bound for integrating p_n on $\mathcal{A}_n \cap \mathcal{C}_n$, we first bound the integral on each of $\mathcal{E}_{q,r} \cap \mathcal{C}_n$, which is in step I below.

Step I: Bounding the posterior integral on $\mathcal{E}_{q,r} \cap \mathcal{C}_n$. We first give upper bounds for $p_n(\boldsymbol{\delta})$ on each of $\mathcal{E}_{q,r} \cap \mathcal{C}_n$. For the CA prior, on each of $\boldsymbol{\delta} \in \mathcal{E}_{q,r}$, we have

$$\pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\left\{ -n \sum_{i=1}^{n} p_\lambda(\beta_j^0 + \delta_j) \right\} \le \exp\{-(s - q + r)n\lambda^2\},$$

as $p_\lambda(x) = \exp(-sn\lambda^2)$ for $|x| > \lambda$ in (5.5). To upper bound the working likelihood, we rely on the same bound in (IC.4), which, together with the CA prior, implies

$$\begin{aligned} p_n(\boldsymbol{\delta}) &\lesssim_{\mathrm{P}^*} \exp\left\{-\frac{n}{8}\boldsymbol{\delta}^T G\boldsymbol{\delta} + \boldsymbol{\phi}^T X\delta - (s-q+r)n\lambda^2\right\} \\ &= \exp\left\{2n\Delta_p^T G\Delta_p - \frac{n}{8}(\boldsymbol{\delta}-4\Delta_p)^T G(\boldsymbol{\delta}-4\Delta_p) - (s-q+r)n\lambda^2\right\} \\ &\triangleq \exp\left\{2n\Delta_p^T G\Delta_p - (s-q+r)n\lambda^2\right\}\cdot \overline{p}_n(\boldsymbol{\delta}), \end{aligned}$$

uniformly on $\boldsymbol{\delta} \in \mathcal{E}_{q,r} \cap \mathcal{C}_n$, which followed by completing the squares for $\boldsymbol{\delta}$; recall that $\Delta_p = G^{-1}\sum_{i=1}^n x_i\phi_\tau(y_i - x_i^T\boldsymbol{\beta}^0)$.

Next, note when $\boldsymbol{\delta} \in \mathcal{E}_{q,r}$, we have

$$\|\boldsymbol{\delta}_1\|^2 \geq \sum_{j\in Q(\boldsymbol{\delta})}(|\beta_j^0| - \lambda)^2 \geq \frac{q\cdot \underline{\mathrm{b}}^2}{4} \geq \frac{32q\cdot\lambda^2}{\theta_{\min}(G)},$$
$$\|\boldsymbol{\delta}_2\|^2 \geq \sum_{j\in R(\boldsymbol{\delta})}(\lambda)^2 \geq r\cdot\lambda^2,$$

since $\lambda \ll \underline{b}\cdot(1 \wedge \theta_{\min}(G))$ as stated in the lemma; hence, we have $\|\boldsymbol{\delta}\|^2 \geq \lambda^2(r + 32q/\theta_{\min}(G))$.

Finally, we upper bound the integral of $p_n(\boldsymbol{\delta})$ by relying on Gaussian tail bounds, similar to (IC.6). Let $\widetilde{\boldsymbol{Z}} \in \mathbb{R}^p$ follow

$$\widetilde{\boldsymbol{Z}} \sim \mathrm{N}\left(4\Delta_p, \frac{4}{n}G^{-1}\right);$$

as in (IC.6), we have for $r+q>0$:

$$\begin{aligned}\int_{\mathcal{E}_{q,r}\cap\mathcal{C}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \;&\lesssim_{\mathrm{P}^*}\; \exp\{2n\Delta_p^T G\Delta_p - sn\lambda^2\}\cdot\left(\frac{8\pi}{n}\right)^{p/2}\cdot\frac{1}{\sqrt{|G|}}\cdot \\ &\qquad\cdot\exp\{(q-r)n\lambda^2\}\cdot\Pr\left(\left\|\widetilde{\boldsymbol{Z}}\right\|^2\geq\lambda^2\cdot\left[r+\frac{32q}{\theta_{\min}(G)}\right]\right) \\ &\leq\; \exp\{2n\Delta_p^T G\Delta_p - sn\lambda^2\}\cdot\left(\frac{8\pi}{n}\right)^{p/2}\cdot\frac{1}{\sqrt{|G|}}\cdot \\ &\qquad\cdot\exp\left\{-(r+q)n\lambda^2\right\},\end{aligned}$$

where the last inequality holds on the events

$$\begin{aligned}E_1(\gamma) &= \left\{n\cdot\Delta_p^T G\Delta_p\leq\gamma\cdot n\lambda^2\cdot(\theta_{\min}(G)\wedge 1)\right\},\\ E_2 &= \left\{n\lambda^2(\theta_{\min}(G)\wedge 1)\gg p\right\}\end{aligned}$$

for small enough $\gamma>0$ by Lemma IA.1.

Step II: Bounding the posterior integral on $\mathcal{A}_n\cap\mathcal{C}_n$ Motivated the decomposition (ID.12), the posterior integral on $\mathcal{A}_n\cap\mathcal{C}_n$ is bounded by the following:

$$\begin{aligned}\int_{\mathcal{A}_n\cap\mathcal{C}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \;&\leq\; \sum_{q=0}^{s}\sum_{\substack{r=0\\ r+q>0}}^{p-s}\int_{\mathcal{E}_{q,r}\cap\mathcal{C}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \\ &\lesssim_{\mathrm{P}^*}\; \exp\{2n\Delta_p^T G\Delta_p - sn\lambda^2\}\cdot\left(\frac{8\pi}{n}\right)^{p/2}\cdot\frac{1}{\sqrt{|G|}}\cdot \\ &\qquad\cdot\left[\sum_{q=0}^{s}\sum_{r=0}^{p-s}\exp\left\{-(r+q)n\lambda^2\right\}-1\right] \\ &=\; \exp\{2n\Delta_p^T G\Delta_p - sn\lambda^2\}\cdot\left(\frac{8\pi}{n}\right)^{p/2}\cdot\frac{1}{\sqrt{|G|}} \\ &\qquad\cdot\left[\left(\frac{1}{1-\exp\{-n\lambda^2\}}\right)^2-1\right],\end{aligned}$$

where the second inequality holds on the event $E_1(\gamma)$ and E_2, given by step I; and the last inequality uses the property of geometric series.

Since $\theta_{\min}(G) \gtrsim 1/p$ and $\lambda \gg p/\sqrt{n}$, the deterministic event E_2 always holds; furthermore, Lemma IC.4 implies that $E_1(\gamma)$ has P*-probability tending to 1. Thus, the previous displayed equation holds with P*-probability tending to 1.

Step III: Bounding the posterior integral on $\mathcal{C}_n^C$. On $\boldsymbol{\delta} \in \mathcal{C}_n^C$, note that $\pi_{CA}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \leq 1$; using a similar argument that leads to (IC.9), we have

$$\int_{\mathcal{C}_n^C} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta} \quad \lesssim_{\mathrm{P}^*} \quad \frac{1}{\sqrt{|G|}} \cdot \left(\frac{4q_0}{\sqrt{p}}\right)^p \exp\{p - n\varepsilon_0\},$$

by invoking a Cramér-Chernoff device.

Step IV: Final bound for posterior probability Finally we verify the posterior probability for $\mathcal{A}_n \cap \mathcal{C}_n$ and $\mathcal{C}_n^C$ are both $o_{P^*}(1)$.

First, Lemma IC.1 gives

$$\begin{aligned}
\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta} \quad &\gtrsim_{\mathrm{P}^*} \quad \left(\frac{2\pi}{n}\right)^{s/2} \left(\frac{2}{n\lambda}\right)^{p-s} \cdot \frac{\exp\left(\Delta_s^T G_{11} \Delta_s - sn\lambda^2\right)}{\sqrt{|G_{11}|}} \\
&\gtrsim \quad \left(\frac{1}{n}\right)^p \cdot \exp\left(-sn\lambda^2\right) \cdot \frac{1}{\sqrt{|G_{11}|}} \\
&\triangleq \quad \widetilde{P}_n,
\end{aligned}$$

since $\lambda \ll 1$. Therefore, following the proof of Theorem IC.1 (step IV), it suffices to show that both

$$\int_{\mathcal{A}_n \cap \mathcal{C}_n} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta}, \quad \text{and} \quad \int_{\mathcal{C}_n^C} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta},$$

are $o_{P^*}(\widetilde{P}_n)$.

On $\mathcal{A}_n \cap \mathcal{C}_n$, we bound the integral of $p_n(\boldsymbol{\delta})$ as in step II; comparing it with $\widetilde{P}_n$

gives

$$\begin{aligned}\frac{\int_{\mathcal{A}_n\cap\mathcal{C}_n} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} (\sqrt{8\pi n})^p \cdot \frac{1}{\sqrt{|\widetilde{G}_{22}|}} \cdot \exp\left\{2n\Delta_p^T G\Delta_p\right\} \cdot \left[\left(1-\exp\{-n\lambda^2\}\right)^{-2}-1\right] \\ &\leq \exp\left\{c_1\cdot p\log n + c_2 p\log p + 2n\Delta_p^T G\Delta_p\right\}\cdot 8\exp\left\{-n\lambda^2\right\} \\ &\lesssim_{\mathrm{P}^*} \exp\left\{-n\lambda^2/4\right\},\end{aligned}$$

where c_1, c_2 are some positive constants, and $\widetilde{G}_{22}$ is the Schur complement of G_{22}; the second inequality follows from (IC.10) and that $(1-x)^{-2}-1\leq 8x$ for all $0<x<1/2$; and the last inequality holds by Lemma IC.4, since $p\log(p\vee n)\ll n\lambda^2$.

On $\mathcal{C}_n^C$, we bound the integral of $p_n(\boldsymbol{\delta})$ as in step III; comparaing it with $\widetilde{P}_n$ gives

$$\begin{aligned}\frac{\int_{\mathcal{C}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} \exp\left\{c_3\cdot p\log n + c_4\cdot p\log p + sn\lambda^2 - n\varepsilon_0\right\} \\ &\leq \exp\left\{-n\varepsilon_0/4\right\},\end{aligned}$$

which followed exactly as (IC.10) for some constant c_3, $c_4>0$; since $p\log(p\vee n)\ll n$ and $\lambda^2\ll 1/s$.

Finally, note that both of the above two displayed equations are $o_{P^*}(1)$, the proof is now complete. □

Proof of Lemma IC.5. We introduce some notations first. Recall $h_n(\boldsymbol{\delta})$ and $f_n(\boldsymbol{\delta})$ from (IC.12), we further define

$$\widetilde{f}_n(\boldsymbol{\delta}) = T_n\cdot\exp\left\{-\frac{n}{2}(\boldsymbol{\delta}_1-\Delta_s)^T G_{11}(\boldsymbol{\delta}_1-\Delta_s) - sn\lambda^2 - n\lambda\|\boldsymbol{\delta}_2\|_1\right\},$$

where Δ_s and T_n are defined before (IC.12). Let $\widetilde{F}_n=\int\widetilde{f}_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$. Fix a diverging

sequence $K_n \to +\infty$ to be specified later, we define

$$\mathcal{B}_n = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_1\|_2 \leq K_n \sqrt{\frac{s}{n}} \text{ , and } \|\boldsymbol{\delta}_2\|_\infty \leq K_n \frac{\log p}{n\lambda} \right\}.$$

Following the proof of Theorem IC.2, we first show that

$$\frac{\int_{\mathbb{R}^p} \left| h_n(\boldsymbol{\delta}) - \widetilde{f}_n(\boldsymbol{\delta}) \right| \mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} \xrightarrow{P^*} 0,$$

which implies h_n converges to $\widetilde{f}_n$ in total variation; to achieve this, in the following steps I - II we bound the integral of $|\widetilde{f}_n - h_n|$ on $\mathcal{B}_n$ and its complement separately; finally in step III, we show that $\widetilde{f}_n$ converges to f_n in total variation, which concludes the proof.

First, the normalizing constant of $\widetilde{f}_n(\boldsymbol{\delta})$, i.e., $\widetilde{F}_n$, can be explicitly computed; using the normalizing constant of Gaussian and Laplace distributions:

$$\begin{aligned} \widetilde{F}_n &= \int_{\mathbb{R}^p} \widetilde{f}_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} \\ &= T_n \cdot \exp\{-sn\lambda^2\} \cdot \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{2\pi}{n}\right)^{s/2} \cdot \left(\frac{2}{n\lambda}\right)^{p-s}. \end{aligned}$$

Step I: Bounding $\int |\widetilde{f}_n - h_n| \, \mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n$. Recall $\Delta_p = G^{-1}\boldsymbol{X}^T\boldsymbol{\phi}$, $\Delta_s = G_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{\phi}$. First we observe that $G \cdot \Delta_p = [(G_{11} \cdot \Delta_s)^T, \ \boldsymbol{\phi}^T\boldsymbol{X}_2/n]^T$, which implies

$$\begin{aligned} \left| \log\left(\frac{h_n(\boldsymbol{\delta})}{\widetilde{f}_n(\boldsymbol{\delta})}\right) \right| &= \frac{n}{2} \left| \boldsymbol{\delta}_1^T G_{11} \boldsymbol{\delta}_1 - 2\boldsymbol{\delta}_1^T G_{11} \Delta_s - \delta^T G \delta + 2\delta^T G \Delta_p \right| \\ &\leq \frac{n}{2} \left| \delta^T G \delta - \delta_1^T G_{11} \delta_1 \right| + |\boldsymbol{\phi}^T \boldsymbol{X}_2 \delta_2| \\ &\triangleq R_1(\boldsymbol{\delta}) + R_2(\boldsymbol{\delta}). \end{aligned}$$

In what follows, we shall prove that $R_1(\boldsymbol{\delta})$ and $R_2(\boldsymbol{\delta})$ are all $o_{P^*}(1)$, uniformly when $\delta \in \mathcal{B}_n$. For $R_1(\boldsymbol{\delta})$, Corollary IB.2 directly implies that

$$\sup_{\boldsymbol{\delta}\in\mathcal{B}_n} R_1(\boldsymbol{\delta}) = o(1).$$

For $R_2(\boldsymbol{\delta})$, by Holder's inequality and Lemma IC.3,

$$\begin{aligned}
\sup_{\boldsymbol{\delta}\in\mathcal{B}_n} R_2(\boldsymbol{\delta}) &\leq p \cdot \left\| \sum_{i=1}^{n} \phi_i x_{2i} \right\|_\infty \cdot \sup_{\boldsymbol{\delta}\in\mathcal{B}_n} \|\boldsymbol{\delta}_2\|_\infty \\
&= p \cdot O_{P^*}(\sqrt{n \log p}) \cdot K_n \frac{\log p}{n\lambda} \\
&= o_{P^*}(1),
\end{aligned}$$

if we choose $K_n \ll (\sqrt{n}\lambda)/(p \log^{3/2} p)$, provided that $\lambda \gg (p \log^{3/2} p)/\sqrt{n}$ as required in the theorem.

Combining the results for R_1 and R_2, we have shown

$$\sup_{\boldsymbol{\delta}\in\mathcal{B}_n} \left| \log \left(\frac{h_n(\boldsymbol{\delta})}{\widetilde{f}_n(\boldsymbol{\delta})} \right) \right| = o_{P^*}(1),$$

which further implies $|h_n(\boldsymbol{\delta})/\widetilde{f}_n(\boldsymbol{\delta}) - 1| = o_{P^*}(1)$ uniformly when $\boldsymbol{\delta} \in \mathcal{B}_n$. Therefore, we have:

$$\begin{aligned}
\int_{\mathcal{B}_n} \left| h_n(\boldsymbol{\delta}) - \widetilde{f}_n(\boldsymbol{\delta}) \right| \mathrm{d}\boldsymbol{\delta} &= \int_{\mathcal{B}_n} \widetilde{f}_n(\boldsymbol{\delta}) \left| 1 - \left(\frac{h_n(\boldsymbol{\delta})}{\widetilde{f}_n(\boldsymbol{\delta})} \right) \right| \mathrm{d}\boldsymbol{\delta} \\
&= o_{P^*} \left(\int_{\mathbb{R}^p} \widetilde{f}_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} \right).
\end{aligned}$$

Step II: Bounding $\int |\widetilde{f}_n - h_n| \, \mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n^C$. Here we show that both $\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}$, and $\int_{\mathcal{B}_n^C} \widetilde{f}_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta}$ are $o_{P^*}(\widetilde{F}_n)$.

For $\widetilde{f}_n$, an argument similar to (IC.6) applies. Note that $\widetilde{F}_n$ normalizes $\widetilde{f}_n(\boldsymbol{\delta})$ as

a probability density function, we have,

$$\begin{aligned}\frac{\int_{\mathcal{B}_n^C} \widetilde{f}_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} &= \Pr\left(\|\boldsymbol{Z}\|_2 \geq K_n\sqrt{\frac{s}{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \geq \frac{K_n \log p}{n\lambda}\right) \\ &\lesssim \left(p\cdot\exp\{-K_n\log p\} + \exp\left\{-\frac{\theta_{11}K_n^2 s}{4}\right\}\right), \qquad \text{(ID.13)}\end{aligned}$$

by Lemma IA.1 and IA.2, similar to (IC.6); θ_{11} is the minimal eigenvalue of G_{11}; the above equations hold on the event $E_1 = \{8n\Delta_s^T G_{11}\Delta_s \leq \theta_{11}K_n^2 s\}$.

For $h_n(\boldsymbol{\delta})$, by relating the integral with probabilistic tail bounds as in (IC.14), we have

$$\begin{aligned}\frac{\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} &\lesssim \left(\frac{n\lambda}{n\lambda-\alpha_n}\right)^{p-s}\cdot\exp\{-c_0K_n(s\wedge\log p)\} \\ &\lesssim \exp\{-c_0K_n\},\end{aligned}$$

for some constant $c_0 > 0$, where the last inequality follows from (IC.8). The displayed equations hold provided that both of the events

$$\begin{aligned}E_2(\gamma) &= \{p\cdot\alpha_n \leq \gamma\cdot n\lambda\}, \\ E_3 &= \left\{8n\cdot\sup_{\|\boldsymbol{\delta}_2\|_\infty\leq K_n\log p/(n\lambda)}\left[\tilde{\mu}(\boldsymbol{\delta}_2)^T G_{11}\tilde{\mu}(\boldsymbol{\delta}_2)\right] \leq K_n s\theta_{\min}(G_{11})\right\},\end{aligned}$$

are true for a small enough constant $\gamma > 0$, where $\tilde{\mu}(\boldsymbol{\delta}_2) = \Delta_s - A_2\boldsymbol{\delta}_2$ and $\alpha_n = \|\boldsymbol{\phi}^T X_2 - \boldsymbol{\phi}^T\boldsymbol{X}_1 A_2\|_\infty$.

Similar to proof of Theorem IC.1, we can show the events E_1 through E_3 happens with P*-probability tending to 1. Therefore, we have

$$\int_{\mathcal{B}_n^C} |h_n(\boldsymbol{\delta}) - \widetilde{f}_n(\boldsymbol{\delta})|\,\mathrm{d}\boldsymbol{\delta} = o_{P^*}(\widetilde{F}_n).$$

Combining steps I and II, we obtain

$$\frac{\int_{\mathbb{R}^p} \left| h_n(\boldsymbol{\delta}) - \widetilde{f}_n(\boldsymbol{\delta}) \right| \, \mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} = o_{P^*}(1),$$

which further implies

$$\left\| \frac{\widetilde{f}_n(\boldsymbol{\beta})}{\int \widetilde{f}_n(\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\beta}} - \frac{h_n(\boldsymbol{\beta})}{\int h_n(\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\beta}} \right\|_{TV} \xrightarrow{P^*} 0,$$

by Theorem 1 of *Chernozhukov and Hong* (2003).

Step III: Convergence of $\widetilde{f}_n$ to f_n. Here we show that $\widetilde{f}_n$ converges to f_n in total variation by bounding their KL divergence.

First note that

$$\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} = \int_{\mathbb{R}^p} \widetilde{f}_n(\boldsymbol{\delta}) \, \mathrm{d}\boldsymbol{\delta} = \widetilde{F}_n,$$

by using the normalizing constants for Gaussian and Laplace distributions. Furthermore,

$$\begin{aligned} \log\left(\frac{\widetilde{f}_n(\boldsymbol{\delta})}{f_n(\boldsymbol{\delta})}\right) &= \frac{n}{2}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1)^T G_{11}(\boldsymbol{\delta}_1 - \widetilde{\boldsymbol{\delta}}_1) - \frac{n}{2}(\boldsymbol{\delta}_1 - \Delta_s)^T G_{11}(\boldsymbol{\delta}_1 - \Delta_s) \\ &= \frac{n}{2}(\Delta_s - \widetilde{\boldsymbol{\delta}}_1)^T G_{11}(\Delta_s - \widetilde{\boldsymbol{\delta}}_1) - n(\Delta_s - \widetilde{\boldsymbol{\delta}}_1)^T G_{11}(\boldsymbol{\delta}_1 - \Delta_s). \end{aligned}$$

Recall that $\widetilde{f}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)/F_n$ coincides with the joint density of $(\boldsymbol{Z}, \boldsymbol{\xi})$, as defined in step II. Now, by Pinsker's inequality (*Tsybakov*, 2008, Lemma 2.5), we can bound

the total variation distance by their KL divergence,

$$\begin{aligned}\left\| \frac{\widetilde{f}_n(\boldsymbol{\delta})}{\int \widetilde{f}_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} - \frac{f_n(\boldsymbol{\delta})}{\int f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}} \right\|_{TV} &\lesssim \left(\mathrm{E}_{(\boldsymbol{Z},\boldsymbol{\xi})} \left[\log \frac{f_n(\boldsymbol{Z},\boldsymbol{\xi})}{\widetilde{f}_n(\boldsymbol{Z},\boldsymbol{\xi})} \right] \right)^{1/2} \\ &= \left(\frac{n}{2} (\Delta_s - \widetilde{\boldsymbol{\delta}}_1)^T G_{11} (\Delta_s - \widetilde{\boldsymbol{\delta}}_1) \right)^{1/2} \\ &\lesssim \sqrt{n} \|\Delta_s - \widetilde{\boldsymbol{\delta}}_1\| \\ &= o_{P^*}(1),\end{aligned}$$

where the first equality holds since $\mathrm{E}_{\boldsymbol{Z}}(\boldsymbol{Z} - \Delta_s) = \mathbf{0}$; the penultimate inequality holds by Assumption E.3′; and the last inequality follows from the Bahadur representation of the quantile regression estimators (*He and Shao*, 2000).

Thus, the proof is now complete. □

5.9.6.3 Proof of Lemmas ID.1, ID.2 and ID.3 under the AL prior

Proof of Lemma ID.1. Recall that the posterior density for $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$ is

$$p_n(\boldsymbol{\delta}) = \pi_{AL}(\boldsymbol{\beta}^0 + \boldsymbol{\delta}) \cdot \exp\left\{ L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \right\},$$

under the Adaptive Lasso prior (5.4). We provide a lower bound of the integral $\int p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$ by restricting to the area

$$\mathcal{B}_n = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_1\| \le K_n/\sqrt{n};\ \ \|\boldsymbol{\delta}_2\|_\infty \le K_n/(n\lambda)\},$$

for a sequence K_n such that Corollary IB.1 holds. We define

$$\widetilde{P}_n = \exp\left\{ \frac{n}{2} \Delta_s^T G_{11} \Delta_s - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 \right\} \cdot \left(\frac{2\pi}{n\theta_M} \right)^{s/2} \cdot \prod_{j=s+1}^{p} \left(\frac{2}{w_j} \right).$$

In step I, we first lower bound the posterior density by $\underline{p}_n(\boldsymbol{\delta})$ on $\mathcal{B}_n$, and in step II we compute the integration of $\underline{p}_n(\boldsymbol{\delta})$ to conclude the proof.

Step I: Lower bounding the posterior density $p_n(\boldsymbol{\delta})$ We first analyze the adaptive Lasso prior when $\boldsymbol{\delta} \in \mathcal{B}_n$. By Assumption E.5, we have $\boldsymbol{\delta}_1 + \boldsymbol{\beta}_1^0 > \mathbf{0}$ for any $\boldsymbol{\delta} \in \mathcal{B}_n$. Recalling $\boldsymbol{\delta}_2 = (\delta_{s+1}, \ldots, \delta_p)$, the adaptive Lasso prior becomes

$$\pi_{AL}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) = \exp\left\{ -\boldsymbol{w}_1^T(\boldsymbol{\delta}_1 + \boldsymbol{\beta}_1^0) - \sum_{j=s+1}^{p} w_j |\delta_j| \right\},$$

for all $\boldsymbol{\delta} \in \mathcal{B}_n$. Second, the likelihood is lower bounded by (ID.11) as in the proof of Lemma IC.1. Therefore, we have the following lower bound of the posterior density:

$$\begin{aligned} p_n(\boldsymbol{\delta}) \;\; \gtrsim_{\mathrm{P}^*} \;\; & \exp\left\{ \frac{n}{2} \Delta_s^T G_{11} \Delta_s - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 \right\} \cdot \exp\left\{ -\frac{n\theta_M}{2} \|\boldsymbol{\delta}_1 - \Delta_s\|^2 - \boldsymbol{w}_1^T \boldsymbol{\delta}_1 \right\} \\ & \cdot \exp\left\{ -\sum_{j=s+1}^{p} (\alpha_n + w_j) |\delta_j| \right\} \\ \triangleq \;\; & \exp\left\{ \frac{n}{2} \Delta_s^T G_{11} \Delta_s - \boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 \right\} \cdot \underline{p}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2), \end{aligned}$$

where $\theta_M = \theta_{\max}(G_{11})$, $\alpha_n = \|\boldsymbol{\phi}^T \boldsymbol{X}_2\|_\infty$ and $\Delta_s = G_{11}^{-1} \boldsymbol{X}_1^T \boldsymbol{\phi}$.

Step II: Lower bounding the integral Similar to the proof of Lemma IC.1, we relate the integral of $\underline{p}_n(\boldsymbol{\delta})$ to probabilistic calculations. Let $\boldsymbol{Z} \in \mathbb{R}^s$ and $\boldsymbol{\xi} = (\xi_{s+1}, \ldots, \xi_p)$ be distributed as

$$\begin{aligned} \xi_j \;\; & \overset{ind.}{\sim} \;\; \text{Laplace}\left(\frac{1}{w_j + \alpha_n} \right), \quad j = s+1, \ldots, p, \\ \boldsymbol{Z} \;\; & \sim \;\; \mathrm{N}\left(\Delta_s, \; \frac{1}{n\theta_M} \boldsymbol{I}_s \right), \end{aligned}$$

where $\boldsymbol{Z}$ and $\boldsymbol{\xi}$ are independent; the function $\underline{p}_n(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ can be related to the moment of $(\boldsymbol{Z}, \boldsymbol{\xi})$. Similar to (ID.3) in the proof of Theorem ID.1, we have

$$\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \;\gtrsim_{\mathrm{P}^*}\; \exp\left\{\frac{n}{2}\Delta_s^T G_{11}\Delta_s - \boldsymbol{w}_1^T\boldsymbol{\beta}_1^0\right\}\cdot\prod_{j=s+1}^{p}\left(\frac{2}{w_j+\alpha_n}\right)\cdot\left(\frac{2\pi}{n\theta_M}\right)^{s/2}$$
$$\cdot\mathrm{E}\left(\exp\left\{-\boldsymbol{w}_1^T\boldsymbol{Z}\right\}\cdot\mathbf{1}[\|\boldsymbol{Z}\|\le K_n/\sqrt{n}]\right)\cdot\Pr\left(\|\boldsymbol{\xi}\|_\infty\le\frac{K_n}{n\lambda}\right), \tag{ID.14}$$

by inserting the normalizing constants for Gaussian and Laplace distributions.

Next we show that both the expectation and probability terms in (ID.14) are bounded from below by a constant. First, for the expectation with respect to $\boldsymbol{Z}$, we have

$$\begin{aligned}
&\mathrm{E}\left(\exp\left\{-\sum_{j=1}^{s}w_jZ_j\right\}\cdot\mathbf{1}[\|\boldsymbol{Z}\|\le K_n/\sqrt{n}]\right)\\
\gtrsim\;& 1-4\gamma-5\exp\left\{-\frac{K_n^2\theta_M}{8}\right\}\\
\gtrsim\;& 1,
\end{aligned}$$

which holds for large enough n on the events

$$E_1(\gamma)=\left\{\|\boldsymbol{w}_1\|\le\gamma\cdot\left(\sqrt{2n\theta_M}\wedge\frac{1}{\|\Delta_s\|}\right)\right\},\quad\text{and}\quad E_2(\gamma)=\left\{n\|\Delta_s\|^2\le\gamma\cdot K_n^2\right\},$$

by Lemma IA.3. Second, for the probability with respect to $\boldsymbol{\xi}$, using an argument similar to the proof of Theorem ID.1 (Step II) gives

$$\Pr\left(\|\boldsymbol{\xi}\|_\infty\ge\frac{K_n}{n\lambda}\right)\ge 1-p\cdot\exp\left\{-\frac{\sqrt{M_n}}{2}\right\}\gtrsim 1,$$

which holds on the event $E_3(\gamma)=\left\{\sqrt{M}_n\cdot(w_{\min}+\alpha_n)\ge\gamma\cdot n\lambda\right\}$.

Combining the above lower bounds, the integration (ID.14) becomes

$$\begin{aligned}\int_{\mathbb{R}^p} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \;&\gtrsim_{\mathrm{P}^*}\; \widetilde{P}_n \cdot \prod_{j=s+1}^{p}\left(\frac{1}{1+\alpha_n/w_j}\right)\\ &\geq\; \widetilde{P}_n \cdot \exp\left\{-\frac{p\cdot\alpha_n}{w_{\min}}\right\}\\ &\gtrsim\; \widetilde{P}_n,\end{aligned}$$

where the second inequality owns to $(1+x)^{-1} \geq \exp\{-x\}$; and the last inequality holds on the event $E_4(\gamma) = \{p\alpha_n \leq \gamma \cdot w_{\min}\}$.

As in the proof of Theorem ID.1, the events $E_1(\gamma)$ through $E_4(\gamma)$ holds with P^*-probability going to 1; Therefore, the proof is now complete. □

Proof of Lemma ID.2. Let $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$, the posterior density is

$$p_n(\boldsymbol{\delta}) \propto \pi_{AL}(\boldsymbol{\delta}+\boldsymbol{\beta}^0)\cdot\exp\left\{L_n(\boldsymbol{\beta}^0) - L_n\left(\boldsymbol{\delta}+\boldsymbol{\beta}^0\right)\right\},$$

Let

$$\mathcal{A}_n = \{\boldsymbol{\delta} = \boldsymbol{\beta}-\boldsymbol{\beta}^0 : sgn(\boldsymbol{\beta}_1) = sgn(\boldsymbol{\beta}_1^0)\ \}.$$

In the following steps I - II, we shall upper bound the posterior integral of $p_n(\boldsymbol{\delta})$ on $\mathcal{A}_n^C \cap \mathcal{C}_n$ and $\mathcal{A}_n^C \cap \mathcal{C}_n^C$ separately, where $\mathcal{C}_n = \{\|G^{1/2}\boldsymbol{\delta}\| \leq 4q_0\}$ as defined in the proof of Theorem V.2; in step III we show the desired posterior probabilities are $o_{P^*}(1)$.

Step I: Bounding the posterior integral on $\mathcal{A}_n^C \cap \mathcal{C}_n$ We first provide an upper bound of the posterior density $p_n(\boldsymbol{\delta})$. The working likelihood can be bounded by (ID.1) when $\boldsymbol{\delta} \in \mathcal{C}_n$. Furthermore, the adaptive Lasso prior is

$$\pi_{AL}(\boldsymbol{\delta}+\boldsymbol{\beta}^0) \leq \exp\left\{-\sum_{j=s+1}^{p} w_j|\delta_j|\right\},$$

as $\boldsymbol{\beta}_2^0 = \mathbf{0}$. Therefore, similar to (ID.1), we have

$$\begin{aligned} p_n(\boldsymbol{\delta}) &\lesssim_{\mathrm{P}^*} \exp\left\{\frac{2}{n\theta_m}\|\boldsymbol{X}_1^T\boldsymbol{\phi}\|^2\right\} \cdot \exp\left\{-\frac{n\theta_m}{8}\|\boldsymbol{\delta}_1 - \boldsymbol{\mu}_1\|^2\right\} \\ &\quad \cdot \exp\left\{-\sum_{j=s+1}^{p}(w_j - \alpha_n)|\delta_j|\right\} \\ &\triangleq \exp\left\{\frac{n\theta_m}{8}\|\boldsymbol{\mu}_1\|^2\right\} \cdot \overline{p}_n(\boldsymbol{\delta}), \end{aligned}$$

when $\boldsymbol{\delta} \in \mathcal{C}_n$, where $\alpha_n = \|\boldsymbol{\phi}^T\boldsymbol{X}_k\|_\infty$ and $\boldsymbol{\mu}_1 = 4\boldsymbol{X}_1^T\boldsymbol{\phi}/(n\theta_m)$.

Now we bound the integral of $p_n(\boldsymbol{\delta})$ by using its upper bound $\overline{p}_n(\boldsymbol{\delta})$. Similar to (ID.3), we can relate the integration to probabilistic tail bounds Note when $\boldsymbol{\delta} \in \mathcal{A}_n^C$, we have $\|\boldsymbol{\delta}_1\| \geq \underline{b}_0$, as in Assumption E.5. Therefore, the posterior integral on $\mathcal{A}_n^C \cap \mathcal{C}_n$ is bounded by

$$\begin{aligned} \int_{\mathcal{C}_n \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} &\lesssim_{\mathrm{P}^*} \exp\left\{\frac{n\theta_m}{8}\|\boldsymbol{\mu}_1\|^2\right\} \cdot \left(\frac{8\pi}{n\theta_m}\right)^{s/2} \cdot \prod_{j=s+1}^{p}\left(\frac{2}{w_j - \alpha_n}\right) \\ &\quad \cdot \Pr\left(\|\boldsymbol{Z}\| \geq \underline{b}_0\right) \\ &\lesssim_{\mathrm{P}^*} \exp\left\{\frac{n\theta_m}{8}\|\boldsymbol{\mu}_1\|^2 - \frac{n\theta_m\underline{b}_0^2}{16}\right\} \cdot \left(\frac{8\pi}{n\theta_m}\right)^{s/2} \cdot \prod_{j=s+1}^{p}\left(\frac{2}{w_j}\right), \end{aligned}$$

where we rely on two techniques: (i) we bound the tail probability with Lemma IA.1, which holds on the event $E_2(\gamma) = \left\{n \cdot \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 \leq \gamma \cdot M_n^2\right\}$; (ii) we bound the normalizing constant $2/(w_j - \alpha_n)$ similar to (IC.8), which holds on the event $E_1(\gamma)$. Since both events $E_1(\gamma)$ and $E_2(\gamma)$ have P*-probability tending to 1 as in the proof of Theorem V.2 (Step V), the last inequality takes hold.

Step II: Bounding the posterior integral on $\mathcal{A}_n^C \cap \mathcal{C}_n^C$ In this step, we use the exact same argument as in (ID.6), since $\pi_{AL}(\boldsymbol{\delta} + \boldsymbol{\beta}^0) \leq 1$ on $\mathcal{A}_n^C$, which gives

$$\int_{\mathcal{C}_n^C \cap \mathcal{A}_n^C} p_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \lesssim_{\mathrm{P}^*} \exp\{-n\varepsilon_0/4\}.$$

Step III: Bounding the posterior probabilities Here we show that the posterior probabilities of both $\mathcal{A}_n^C \cap \mathcal{C}_n$ and $\mathcal{A}_n^C \cap \mathcal{C}_n^C$ are $o_{P^*}(1)$. Let

$$\widetilde{P}_n = \prod_{j=s+1}^{p} \left(\frac{2}{w_j}\right) \cdot \left(\frac{2\pi}{n\theta_M}\right)^{s/2} \cdot \exp\left(-\boldsymbol{w}_1^T \boldsymbol{\beta}_1^0 + \frac{n}{2}\Delta_s^T G_{11} \Delta_s\right);$$

following the proof of Theorem V.2, it then suffices to show that both

$$\int_{\mathcal{A}_n^C \cap \mathcal{C}_n} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta} \quad \text{and} \quad \int_{\mathcal{A}_n^C \cap \mathcal{C}_n^C} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta},$$

are $o_{P^*}(\widetilde{P}_n)$.

For the first integral on $\mathcal{A}_n^C \cap \mathcal{C}_n$, we use its upper bound derived in step I; comparing it with $\widetilde{P}_n$ gives

$$\begin{aligned} \frac{\displaystyle\int_{\mathcal{A}_n^C \cap \mathcal{C}_n} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} &\lesssim_{\mathrm{P}^*} \exp\left\{\boldsymbol{w}_1^T \boldsymbol{\beta}^0 + \frac{n\theta_m}{8}\|\boldsymbol{\mu}_1\|^2 - \frac{n\theta_m \underline{b}_0^2}{16}\right\} \cdot \left(\frac{4\theta_M}{\theta_m}\right)^{s/2} \\ &= o_{P^*}(1), \end{aligned}$$

which holds since $\|\boldsymbol{w}_1\| = O_{P^*}(\sqrt{n})$ and $\boldsymbol{\mu}_1 = O_{P^*}(1/\sqrt{n})$, as in the proof of Theorem V.2 (Step V).

For the second integral on $\mathcal{A}_n^C \cap \mathcal{C}_n^C$, we have the exact same result as in (ID.7), implying

$$\frac{\displaystyle\int_{\mathcal{A}_n^C \cap \mathcal{C}_n^C} p_n(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta}}{\widetilde{P}_n} = o_{P^*}(1).$$

Therefore, the proof is now complete. □

Proof of Lemma ID.3. Fix a diverging sequence $K_n \to +\infty$ to be specified later, we define

$$\mathcal{B}_n = \left\{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_1\|_2 \leq \frac{K_n}{\sqrt{n}}\ , \text{ and } \ \|\boldsymbol{\delta}_2\|_\infty \leq \frac{K_n}{n\lambda}\right\}.$$

Let $\widetilde{F}_n = \int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$. Following the proof of Theorem V.2, we ought to show that

$$\frac{\int_{\mathbb{R}^p} \left| h_n(\boldsymbol{\delta}) - \widetilde{f}_n(\boldsymbol{\delta}) \right| \mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} \xrightarrow{P^*} 0.$$

In the following steps I - II, we bound the integral of $|f_n - h_n|$ on $\mathcal{B}_n$ and its complement separately.

First, similar to the proof of Lemma IC.5, $\widetilde{F}_n$ can be explicitly computed as

$$\begin{aligned} \widetilde{F}_n &= \int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta} \\ &= T_n \cdot \frac{1}{\sqrt{|G_{11}|}} \cdot \left(\frac{2\pi}{n}\right)^{s/2} \cdot \prod_{j=s+1}^{p} \left(\frac{2}{w_j}\right). \end{aligned}$$

Step I: Bounding $\int |f_n - h_n|\,\mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n$. First we observe that $G \cdot \Delta_p = [(G_{11} \cdot \Delta_s)^T,\ \boldsymbol{\phi}^T \boldsymbol{X}_2/n]^T$, which implies

$$\begin{aligned} \left| \log\left(\frac{h_n(\boldsymbol{\delta})}{f_n(\boldsymbol{\delta})}\right) \right| &= \frac{n}{2} \left| \boldsymbol{\delta}_1^T G_{11} \boldsymbol{\delta}_1 - 2\boldsymbol{\delta}_1^T G_{11} \widetilde{\boldsymbol{\delta}}_1 - \boldsymbol{\delta}^T G \boldsymbol{\delta} + 2\boldsymbol{\delta}_1^T G_{11} \Delta_s + 2\boldsymbol{\delta}_2^T \boldsymbol{X}_2^T \boldsymbol{\phi}/n \right| \\ &\leq \frac{n}{2} \left| \boldsymbol{\delta}^T G \boldsymbol{\delta} - \boldsymbol{\delta}_1^T G_{11} \delta_1 \right| + \left| \boldsymbol{\delta}_1^T G_{11} (\widetilde{\boldsymbol{\delta}}_1 - \Delta_s) \right| + |\boldsymbol{\delta}_2^T \boldsymbol{X}_2^T \boldsymbol{\phi}| \\ &\triangleq R_1(\boldsymbol{\delta}) + R_2(\boldsymbol{\delta}) + R_3(\boldsymbol{\delta}). \end{aligned}$$

As in the proof of Lemma IC.5, $R_1(\boldsymbol{\delta})$ and $R_3(\boldsymbol{\delta})$ are both $o_{P^*}(1)$, uniformly when $\delta \in \mathcal{B}_n$. For $R_2(\boldsymbol{\delta})$, it follows from the Bahadur representation of the quantile regression estimators (*Koenker*, 2005, Section 4.2) that

$$\begin{aligned} \sup_{\boldsymbol{\delta} \in \mathcal{B}_n} R_2(\boldsymbol{\delta}) &\leq \sup_{\boldsymbol{\delta} \in \mathcal{B}_n} \|\boldsymbol{\delta}_1\| \cdot \left\| G_{11} (\widetilde{\boldsymbol{\delta}}_1 - \Delta_s) \right\| \\ &\leq \frac{K_n}{\sqrt{n}} \cdot o_{P^*}\left(\frac{1}{\sqrt{n}}\right) \\ &= o_{P^*}(1), \end{aligned}$$

if we choose a sequence K_n that grows slow enough. Therefore, we have shown

$$\sup_{\boldsymbol{\delta}\in\mathcal{B}_n}\left|\log\left(\frac{h_n(\boldsymbol{\delta})}{f_n(\boldsymbol{\delta})}\right)\right| = o_{P^*}(1),$$

which further implies $|h_n(\boldsymbol{\delta})/\widetilde{f}_n(\boldsymbol{\delta}) - 1| = o_{P^*}(1)$ uniformly when $\boldsymbol{\delta}\in\mathcal{B}_n$. Thus,

$$\begin{aligned}\int_{\mathcal{B}_n}|h_n(\boldsymbol{\delta}) - f_n(\boldsymbol{\delta})|\,\mathrm{d}\boldsymbol{\delta} &= \int_{\mathcal{B}_n} f_n(\boldsymbol{\delta})\left|1-\left(\frac{h_n(\boldsymbol{\delta})}{f_n(\boldsymbol{\delta})}\right)\right|\mathrm{d}\boldsymbol{\delta}\\ &= o_{P^*}\left(\int_{\mathbb{R}^p} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}\right).\end{aligned}$$

Step II: Bounding $\int|f_n - h_n|\,\mathrm{d}\boldsymbol{\delta}$ on $\mathcal{B}_n^C$. Here we show that both $\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$, and $\int_{\mathcal{B}_n^C} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}$ are $o_{P^*}(\widetilde{F}_n)$.

Let $\boldsymbol{\mu}_1 = \boldsymbol{X}_1^T\boldsymbol{\phi}/n$ and $\alpha_n = \|\boldsymbol{X}_2^T\boldsymbol{\phi}\|_\infty$. For $h_n(\boldsymbol{\delta})$, similar to (ID.10), we can show that

$$\frac{\int_{\mathcal{B}_n^C} h_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} \lesssim \exp\left\{\frac{n\theta_m}{2}\|\boldsymbol{\mu}_1\|^2 - \sqrt{K_n}\right\},$$

by relating the integral with probabilistic tail bounds; provided that all of the events $E_1(\gamma) = \{\alpha_n \le \gamma\cdot w_{\min}\}$, $E_2(\gamma) = \left\{n\cdot\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 \le \gamma\cdot K_n^2\right\}$, and

$$E_3(\gamma) = \left\{\sqrt{K}_n\cdot(w_{\min} - \alpha_n) \ge \gamma\cdot n\lambda\right\}$$

holds

For f_n, note that $\widetilde{F}_n$ normalizes $f_n(\boldsymbol{\delta})$; therefore, using a similar argument to (ID.4), we have

$$\begin{aligned}\frac{\int_{\mathcal{B}_n^C} f_n(\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} &= \Pr\left(\|\boldsymbol{Z}\|_2 \ge \frac{K_n}{\sqrt{n}} \text{ or } \|\boldsymbol{\xi}\|_\infty \ge \frac{K_n}{n\lambda}\right)\\ &\lesssim \exp\left\{-\sqrt{K_n}\right\},\end{aligned}$$

which holds on the events $E_3(\gamma)$ and $E_4(\gamma) = \left\{ n \cdot \widetilde{\boldsymbol{\delta}}_1^T G_{11} \widetilde{\boldsymbol{\delta}}_1 \leq \gamma \cdot K_n^2 \right\}$, by Lemma IA.1 and IA.2.

As in the proof of Theorem ID.1 (Step V), the events $E_1(\gamma)$ through $E_3(\gamma)$ happens with P*-probability tending to 1. For $E_4(\gamma)$, standard asymptotic results on quantile regression (*Koenker*, 2005, Section 4.2) shows that

$$n\widetilde{\boldsymbol{\delta}}_1^T G_{11} \widetilde{\boldsymbol{\delta}}_1 \leq \theta_M \left\| \sqrt{n} \widetilde{\boldsymbol{\delta}}_1 \right\| = O_{P^*}(1),$$

which implies $\mathrm{P}^*(E_4(\gamma)) \to 1$. Therefore, we have shown

$$\int_{\mathcal{B}_n^C} |h_n(\boldsymbol{\delta}) - f_n(\boldsymbol{\delta})| \, \mathrm{d}\boldsymbol{\delta} = o_{P^*}(\widetilde{F}_n).$$

Combining steps I and II, we obtain

$$\frac{\displaystyle\int_{\mathbb{R}^p} |h_n(\boldsymbol{\delta}) - f_n(\boldsymbol{\delta})| \, \mathrm{d}\boldsymbol{\delta}}{\widetilde{F}_n} = o_{P^*}(1),$$

which completes the proof by Theorem 1 of *Chernozhukov and Hong* (2003). □

BIBLIOGRAPHY

BIBLIOGRAPHY

Acerbi, C., and D. Tasche (2002), On the coherence of expected shortfall, *Journal of Banking & Finance*, *26*(7), 1487–1503.

Adlouni, S. E., G. Salaou, and A. St-Hilaire (2018), Regularized bayesian quantile regression, *Communications in Statistics–Simulation and Computation*, *47*(1), 277–293.

Alhamzawi, R., K. Yu, and D. F. Benoit (2012), Bayesian adaptive lasso quantile regression, *Statistical Modelling*, *12*(3), 279–297.

Andrews, D. F., and C. L. Mallows (1974), Scale mixtures of normal distributions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *36*(1), 99–102.

Andrews, D. W. (1994), Empirical process methods in econometrics, in *Handbook of Econometrics*, vol. 4, pp. 2247–2294, Elsevier.

Angrist, J. D., and J.-S. Pischke (2010), The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of Economic Perspectives*, *24*(2), 3–30.

Armagan, A., D. B. Dunson, and J. Lee (2013a), Generalized double pareto shrinkage, *Statistica Sinica*, *23*(1), 119.

Armagan, A., D. B. Dunson, J. Lee, W. U. Bajwa, and N. Strawn (2013b), Posterior consistency in linear models under shrinkage priors, *Biometrika*, *100*(4), 1011–1018.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999), Coherent measures of risk, *Mathematical Finance*, *9*(3), 203–228.

Bahadur, R. R. (1966), A note on quantiles in large samples, *The Annals of Mathematical Statistics*, *37*(3), 577–580.

Bai, R., and M. Ghosh (2021), On the beta prime prior for scale parameters in high-dimensional bayesian regression models, *Statistica Sinica*, *31*(2), 843.

Barendse, S. (2020), Efficiently weighted estimation of tail and interquantile expectations, *Tinbergen Institute Discussion Paper 2017-034/III*, doi:10.2139/ssrn.2937665.

Barut, E., J. Fan, and A. Verhasselt (2016), Conditional sure independence screening, *Journal of the American Statistical Association*, *111*(515), 1266–1277.

Basel Committee on Banking Supervision (2013), Fundamental review of the trading book: A revised market risk framework, *Consultative Document, October*.

Belloni, A., and V. Chernozhukov (2011), ℓ1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics*, *39*(1), 82–130.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019a), Conditional quantile processes based on series or many regressors, *Journal of Econometrics*, *213*(1), 4–29.

Belloni, A., V. Chernozhukov, and K. Kato (2019b), Valid post-selection inference in high-dimensional approximately sparse quantile regression models, *Journal of the American Statistical Association*, *114*(526), 749–758.

Benoit, D. F., and D. Van den Poel (2017), Bayesqr: A bayesian approach to quantile regression, *Journal of Statistical Software*, *76*(7), 1–32.

Bercu, B., M. Costa, and S. Gadat (2021), Stochastic approximation algorithms for superquantiles estimation, *Electronic Journal of Probability*, *26*, 1–29.

Bermingham, M. L., et al. (2015), Application of high-dimensional feature selection: Evaluation for genomic prediction in man, *Scientific Reports*, *5*(1), 1–12.

Bhadra, A., J. Datta, N. G. Polson, and B. Willard (2019), Lasso meets horseshoe: A survey, *Statistical Science*, *34*(3), 405–427.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015), Dirichlet–laplace priors for optimal shrinkage, *Journal of the American Statistical Association*, *110*(512), 1479–1490.

Birgé, L. (2001), An alternative point of view on lepski's method, *Lecture Notes-Monograph Series*, *36*, 113–133.

Bisai, S., A. Sen, D. Mahalanabis, N. Datta, and K. Bose (2006), The effect of maternal age and parity on birth weight among bengalees of kolkata, india, *Human Ecology*, *14*, 139–143.

Bontemps, D. (2011), Bernstein–von mises theorems for gaussian regression with increasing number of regressors, *The Annals of Statistics*, *39*(5), 2557–2584.

Boucheron, S., G. Lugosi, and P. Massart (2013), *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press.

Bühlmann, P., and S. Van De Geer (2011), *Statistics for high-dimensional data: Methods, theory and applications*, Springer Science & Business Media.

Cai, Z., and X. Wang (2008), Nonparametric estimation of conditional var and expected shortfall, *Journal of Econometrics*, *147*(1), 120–130.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010), The horseshoe estimator for sparse signals, *Biometrika*, *97*(2), 465–480.

Casella, G. (2001), Empirical bayes gibbs sampling, *Biostatistics*, *2*(4), 485–500.

Castillo, I., J. Schmidt-Hieber, and A. Van der Vaart (2015), Bayesian linear regression with sparse priors, *The Annals of Statistics*, *43*(5), 1986–2018.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018), Inference in linear regression models with many covariates and heteroscedasticity, *Journal of the American Statistical Association*, *113*(523), 1350–1361.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2019), On binscatter, *arXiv preprint arXiv:1902.09608*.

Chen, C. W., D. B. Dunson, C. Reed, and K. Yu (2013), Bayesian variable selection in quantile regression, *Statistics and its Interface*, *6*(2), 261–274.

Chen, L.-Y., and Y.-M. Yen (2021), Estimations of the conditional tail average treatment effect, *arXiv preprint arXiv:2109.08793*.

Chen, S. X. (2007), Nonparametric estimation of expected shortfall, *Journal of Financial Econometrics*, *6*(1), 87–107.

Chernozhukov, V., and H. Hong (2003), An mcmc approach to classical estimation, *Journal of Econometrics*, *115*(2), 293–346.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009), Improving point and interval estimators of monotone functions by rearrangement, *Biometrika*, *96*(3), 559–575.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010), Quantile and probability curves without crossing, *Econometrica*, *78*(3), 1093–1125.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013), Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors, *The Annals of Statistics*, *41*(6), 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017), Central limit theorems and bootstrap in high dimensions, *The Annals of Probability*, *45*(4), 2309–2352.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018), Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal*, *21*(1), C1–C68.

Chetverikov, D., Y. Liu, and A. Tsyvinski (2022), Weighted-average quantile regression, *Tech. rep.*, National Bureau of Economic Research.

Choi, H. M., and J. P. Hobert (2013), Analysis of mcmc algorithms for bayesian linear regression with laplace errors, *Journal of Multivariate Analysis*, *117*, 32–40.

Chun, S. Y., A. Shapiro, and S. Uryasev (2012), Conditional value-at-risk and average value-at-risk: Estimation and asymptotics, *Operations Research*, *60*(4), 739–756.

Deng, K., and J. Qiu (2021), Backtesting expected shortfall and beyond, *Quantitative Finance*, *21*(7), 1109–1125.

Dette, H., and S. Volgushev (2008), Non-crossing non-parametric estimates of quantile curves, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(3), 609–627.

Dimitriadis, T., and S. Bayer (2019), A joint quantile and expected shortfall regression framework, *Electronic Journal of Statistics*, *13*(1), 1823–1871.

Dimitriadis, T., and S. Bayer (2022), *Esreg: Joint quantile and expected shortfall regression*, r package version 0.6.0.

Dimitriadis, T., T. Fissler, and J. F. Ziegel (2020), The efficiency gap, *arXiv preprint arXiv:2010.14146*.

Duong, H. T., A. T. Hoyt, S. L. Carmichael, S. M. Gilboa, M. A. Canfield, A. Case, M. L. McNeese, D. K. Waller, and N. B. D. P. Study (2012), Is maternal parity an independent risk factor for birth defects?, *Birth Defects Research Part A: Clinical and Molecular Teratology*, *94*(4), 230–236.

Emmer, S., M. Kratz, and D. Tasche (2015), What is the best risk measure in practice? a comparison of standard measures, *Journal of Risk*, *18*(2), 31–60.

Fama, E. F., and K. R. French (1993), Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, *33*(1), 3–56.

Fama, E. F., and K. R. French (1995), Size and book-to-market factors in earnings and returns, *The Journal of Finance*, *50*(1), 131–155.

Fama, E. F., and K. R. French (2015), A five-factor asset pricing model, *Journal of Financial Economics*, *116*(1), 1–22.

Fan, J. (1992), Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, *87*(420), 998–1004.

Fan, J., and I. Gijbels (1992), Variable bandwidth and local linear regression smoothers, *The Annals of Statistics*, *20*(4), 2008–2036.

Fan, J., and I. Gijbels (2018), *Local polynomial modelling and its applications: Monographs on statistics and applied probability 66*, Routledge.

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

Fan, J., and J. Lv (2008), Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., and J. Lv (2010), A selective overview of variable selection in high dimensional feature space, *Statistica Sinica*, *20*(1), 101.

Fan, J., and H. Peng (2004), Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, *32*(3), 928–961.

Feng, Y., Y. Chen, and X. He (2015), Bayesian quantile regression with approximate likelihood, *Bernoulli*, *21*(2), 832–850.

Fissler, T., and J. F. Ziegel (2016), Higher order elicitability and osband's principle, *The Annals of Statistics*, *44*(4), 1680–1707.

Fitzenberger, B., R. Koenker, and J. A. Machado (2013), *Economic applications of quantile regression*, Springer Science & Business Media.

Gao, C., A. W. van der Vaart, and H. H. Zhou (2020), A general framework for bayes structured linear models, *The Annals of Statistics*, *48*(5), 2848–2878.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013), *Bayesian data analysis*, Chapman and Hall/CRC.

Ghosal, S. (1999), Asymptotic normality of posterior distributions in high-dimensional linear models, *Bernoulli*, *5*(2), 315–331.

Gneiting, T. (2011), Making and evaluating point forecasts, *Journal of the American Statistical Association*, *106*(494), 746–762.

Golodnikov, A., V. Kuzmenko, and S. Uryasev (2019), Cvar regression based on the relation between cvar and mixed-quantile quadrangles, *Journal of Risk and Financial Management*, *12*(3), 107.

Griffin, J. E., and P. J. Brown (2010), Inference with normal-gamma prior distributions in regression problems, *Bayesian Analysis*, *5*(1), 171–188.

Gutenbrunner, C., and J. Jurecková (1992), Regression rank scores and regression quantiles, *The Annals of Statistics*, *20*(1), 305–330.

Hans, C. (2009), Bayesian lasso regression, *Biometrika*, *96*(4), 835–845.

Hansen, B. E. (1994), Autoregressive conditional density estimation, *International Economic Review*, *35*(3), 705–730.

He, X. (1997), Quantile curves without crossing, *The American Statistician*, *51*(2), 186–192.

He, X., and Q.-M. Shao (2000), On parameters of increasing dimensions, *Journal of Multivariate Analysis*, *73*(1), 120–135.

He, X., and P. Shi (1994), Convergence rate of b-spline estimators of nonparametric conditional quantile functions, *Journal of Nonparametric Statistics*, *3*(3-4), 299–308.

He, X., Y.-H. Hsu, and M. Hu (2010), Detection of treatment effects by covariate-adjusted expected shortfall, *The Annals of Applied Statistics*, *4*(4), 2114–2125.

He, X., L. Wang, and H. G. Hong (2013), Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *The Annals of Statistics*, *41*(1), 342–369.

Hendricks, W., and R. Koenker (1992), Hierarchical spline models for conditional quantiles and the demand for electricity, *Journal of the American statistical Association*, *87*(417), 58–68.

Hinkle, S. N., P. S. Albert, P. Mendola, L. A. Sjaarda, E. Yeung, N. S. Boghossian, and S. K. Laughon (2014), The association between parity and birthweight in a longitudinal consecutive pregnancy cohort, *Paediatric and perinatal epidemiology*, *28*(2), 106–115.

Hjort, N. L., and D. Pollard (2011), Asymptotics for minimisers of convex processes, *arXiv preprint arXiv:1107.3806*.

Huang, J., S. Ma, and C.-H. Zhang (2008), Adaptive lasso for sparse high-dimensional regression models, *Statistica Sinica*, *18*, 1603–1618.

Hughes, M. M., R. E. Black, and J. Katz (2017), 2500-g low birth weight cutoff: History and implications for future research and policy, *Maternal and Child Health Journal*, *21*(2), 283–289.

Ishwaran, H., and J. S. Rao (2005), Spike and slab variable selection: Frequentist and bayesian strategies, *The Annals of Statistics*, *33*(2), 730–773.

Jiang, B., and Q. Sun (2019), Bayesian high-dimensional linear regression with generic spike-and-slab priors, *arXiv preprint arXiv:1912.08993*.

Jorgensen, B. (2012), *Statistical properties of the generalized inverse Gaussian distribution*, vol. 9, Springer Science & Business Media.

Kato, K. (2012), Weighted nadaraya–watson estimation of conditional expected shortfall, *Journal of Financial Econometrics*, *10*(2), 265–291.

Kim, Y., H. Choi, and H.-S. Oh (2008), Smoothly clipped absolute deviation on high dimensions, *Journal of the American Statistical Association*, *103*(484), 1665–1673.

Kleijn, B. J. K., and A. W. Van der Vaart (2012), The bernstein-von-mises theorem under misspecification, *Electronic Journal of Statistics*, *6*, 354–381.

Knight, K. (1998), Limiting distributions for l1 regression estimators under general conditions, *The Annals of Statistics*, *26*(2), 755–770.

Koenker, R. (2005), *Quantile regression*, Econometric Society Monographs, i-vi pp., Cambridge University Press.

Koenker, R. (2018), *Quantreg: Quantile regression*, r package version 5.38.

Koenker, R., and G. Bassett Jr (1978), Regression quantiles, *Econometrica*, *46*(1), 33–50.

Koenker, R., and J. A. Machado (1999), Goodness of fit and related inference processes for quantile regression, *Journal of the American Statistical Association*, *94*(448), 1296–1310.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017), *Handbook of quantile regression*, CRC press.

Kohns, D., and T. Szendrei (2020), Horseshoe prior bayesian quantile regression, *arXiv preprint arXiv:2006.07655*.

Koumou, G. B. (2020), Diversification and portfolio theory: a review, *Financial Markets and Portfolio Management*, *34*(3), 267–312.

Kozumi, H., and G. Kobayashi (2011), Gibbs sampling methods for bayesian quantile regression, *Journal of Statistical Computation and Simulation*, *81*(11), 1565–1578.

Kunsch, H. R. (1989), The jackknife and the bootstrap for general stationary observations, *The Annals of Statistics*, *17*(3), 1217–1241.

Laguel, Y., K. Pillutla, J. Malick, and Z. Harchaoui (2021a), A superquantile approach to federated learning with heterogeneous devices, in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE.

Laguel, Y., K. Pillutla, J. Malick, and Z. Harchaoui (2021b), Superquantiles at work: Machine learning applications and efficient subgradient computation, *Set-Valued and Variational Analysis*, *29*, 967–996.

Leamer, E. E. (2010), Tantalus on the road to asymptopia, *Journal of Economic Perspectives*, *24*(2), 31–46.

Leeb, H., and B. M. Pötscher (2005), Model selection and inference: Facts and fiction, *Econometric Theory*, *21*(1), 21–59.

Leorato, S., F. Peracchi, and A. V. Tanase (2012), Asymptotically efficient estimation of the conditional expected shortfall, *Computational Statistics & Data Analysis*, *56*(4), 768–784.

Li, Q., R. Xi, and N. Lin (2010), Bayesian regularized quantile regression, *Bayesian Analysis*, *5*(3), 533–556.

Li, Y., and J. Zhu (2008), L1-norm quantile regression, *Journal of Computational and Graphical Statistics*, *17*(1), 163–185.

Lin, L., C. Lu, W. Chen, C. Li, and V. Y. Guo (2021), Parity and the risks of adverse birth outcomes: A retrospective study among chinese, *BMC Pregnancy and Childbirth*, *21*(1), 1–11.

Liu, J., W. Zhong, and R. Li (2015), A selective overview of feature screening for ultrahigh-dimensional data, *Science China Mathematics*, *58*(10), 1–22.

Ma, S., R. Li, and C.-L. Tsai (2017), Variable screening via quantile partial correlation, *Journal of the American Statistical Association*, *112*(518), 650–663.

Mack, Y.-p., and B. W. Silverman (1982), Weak and strong uniform consistency of kernel regression estimates, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *61*(3), 405–415.

Martins-Filho, C., F. Yao, and M. Torero (2018), Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory, *Econometric Theory*, *34*(1), 23–67.

Miranda, S. I. (2014), Superquantile regression: theory, algorithms, and applications, Ph.D. thesis, Naval Postgraduate School.

Muller, H.-G., and U. Stadtmuller (1987), Variable bandwidth kernel estimators of regression curves, *The Annals of Statistics*, *15*(1), 182–201.

Muula, A., S. Siziya, and E. Rudatsikira (2011), Parity and maternal education are associated with low birth weight in malawi, *African Health Sciences*, *11*(1), 65–71.

Nadarajah, S., B. Zhang, and S. Chan (2014), Estimation methods for expected shortfall, *Quantitative Finance*, *14*(2), 271–291.

Narisetty, N. N., and X. He (2014), Bayesian variable selection with shrinking and diffusing priors, *The Annals of Statistics*, *42*(2), 789–817.

Narula, S. C., and J. F. Wellington (1982), The minimum sum of absolute errors regression: A state of the art survey, *International Statistical Review*, *50*(3), 317–326.

Nolde, N., and J. F. Ziegel (2017), Elicitability and backtesting: Perspectives for banking regulation, *The Annals of Applied Statistics*, *11*(4), 1833–1874.

Novy-Marx, R. (2013), The other side of value: The gross profitability premium, *Journal of Financial Economics*, *108*(1), 1–28.

Olma, T. (2021), Nonparametric estimation of truncated conditional expectation functions, *arXiv preprint arXiv:2109.06150*.

Pan, X., and W.-X. Zhou (2021), Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design, *Information and Inference: A Journal of the IMA*, *10*(3), 813–861.

Panov, M., and V. Spokoiny (2015), Finite sample bernstein–von mises theorem for semiparametric problems, *Bayesian Analysis*, *10*(3), 665–710.

Park, T., and G. Casella (2008), The bayesian lasso, *Journal of the American Statistical Association*, *103*(482), 681–686.

Patton, A. J., J. F. Ziegel, and R. Chen (2019), Dynamic semiparametric models for expected shortfall (and value-at-risk), *Journal of Econometrics*, *211*(2), 388–413.

Peng, X. (2022), Advances in subgroup identification and expected shortfall regression, Ph.D. thesis, The George Washington University.

Peracchi, F., and A. V. Tanase (2008), On estimating the conditional expected shortfall, *Applied Stochastic Models in Business and Industry*, *24*(5), 471–493.

Pollard, D. (1985), New ways to prove central limit theorems, *Econometric Theory*, *1*(3), 295–313.

Pollard, D. (1991), Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, *7*(2), 186–199.

Pötscher, B. M., and H. Leeb (2009), On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding, *Journal of Multivariate Analysis*, *100*(9), 2065–2082.

Powell, J. L. (1986), Censored regression quantiles, *Journal of Econometrics*, *32*(1), 143–155.

Powell, J. L. (1991), Estimation of monotonic regression models under quantile restrictions, in *Nonparametric and Semiparametric Methods in Econometrics*, pp. 357–384, Cambridge University Press, Cambridge, UK.

Reich, B. J., and L. B. Smith (2013), Bayesian quantile regression for censored data, *Biometrics*, *69*(3), 651–660.

Rockafellar, R. T., and J. O. Royset (2010), On buffered failure probability in design and optimization of structures, *Reliability Engineering & System Safety*, *95*(5), 499–510.

Rockafellar, R. T., and J. O. Royset (2013), Superquantiles and their applications to risk, random variables, and regression, in *INFORMS TutORials in Operations Research*, pp. 151–167, Informs.

Rockafellar, R. T., and J. O. Royset (2018), Superquantile/cvar risk measures: second-order theory, *Annals of Operations Research*, *262*(1), 3–28.

Rockafellar, R. T., and S. Uryasev (2013), The fundamental risk quadrangle in risk management, optimization and statistical estimation, *Surveys in Operations Research and Management Science*, *18*(1-2), 33–53.

Rockafellar, R. T., S. Uryasev, and M. Zabarankin (2008), Risk tuning with generalized linear regression, *Mathematics of Operations Research*, *33*(3), 712–729.

Rockafellar, R. T., J. O. Royset, and S. I. Miranda (2014), Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk, *European Journal of Operational Research*, *234*(1), 140–154.

Ročková, V., and E. I. George (2018), The spike-and-slab lasso, *Journal of the American Statistical Association*, *113*(521), 431–444.

Rodrigues, T., J.-L. Dortet-Bernadet, and Y. Fan (2019), Pyramid quantile regression, *Journal of Computational and Graphical Statistics*, *28*(3), 732–746.

Scaillet, O. (2004), Nonparametric estimation and sensitivity analysis of expected shortfall, *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, *14*(1), 115–129.

Schubert, E., and P. J. Rousseeuw (2019), Faster k-medoids clustering: improving the pam, clara, and clarans algorithms, in *International Conference on Similarity Search and Applications*, pp. 171–187, Springer.

Shah, P. S. (2010), Parity and low birth weight and preterm birth: A systematic review and meta-analyses, *Acta Obstetricia et Gynecologica Scandinavica*, *89*(7), 862–875.

Shao, X., and J. Zhang (2014), Martingale difference correlation and its use in high-dimensional variable screening, *Journal of the American Statistical Association*, *109*(507), 1302–1318.

Sherwood, B., L. Wang, and X.-H. Zhou (2013), Weighted quantile regression for analyzing health care cost data with missing covariates, *Statistics in medicine*, *32*(28), 4967–4979.

Soleimani, H., and K. Govindan (2014), Reverse logistics network design and planning utilizing conditional value at risk, *European Journal of Operational Research*, *237*(2), 487–497.

Song, Q., and F. Liang (2017), Nearly optimal bayesian shrinkage for high dimensional regression, *arXiv preprint arXiv:1712.08964*.

Spokoiny, V. (2013), Bernstein-von mises theorem for growing parameter dimension, *arXiv preprint arXiv:1302.3430*.

Sriram, K. (2015), A sandwich likelihood correction for bayesian quantile regression based on the misspecified asymmetric laplace density, *Statistics & Probability Letters*, *107*, 18–26.

Sriram, K., and R. Ramamoorthi (2017), Correction to:"posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density", *Bayesian Analysis*, *12*(4), 1217–1219.

Sriram, K., and R. Ramamoorthi (2018), On $\sqrt{n}-$ consistency for bayesian quantile regression based on the misspecified asymmetric laplace likelihood, *arXiv preprint arXiv:1812.03652*.

Sriram, K., R. Ramamoorthi, and P. Ghosh (2013), Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density, *Bayesian Analysis*, *8*(2), 479–504.

Starr, E., and B. Goldfarb (2020), Binned scatterplots: A simple tool to make research easier and better, *Strategic Management Journal*, *41*(12), 2261–2274.

Sun, S., and F. Cheng (2018), Bootstrapping the expected shortfall, *Theoretical Economics Letters*, *8*(04), 685–698.

Tamba, C. L., Y.-L. Ni, and Y.-M. Zhang (2017), Iterative sure independence screening em-bayesian lasso algorithm for multi-locus genome-wide association studies, *PLoS Computational Biology*, *13*(1).

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288.

Tsionas, E. G. (2003), Bayesian quantile inference, *Journal of Statistical Computation and Simulation*, *73*(9), 659–674.

Tsybakov, A. B. (2008), *Introduction to nonparametric estimation*, Springer Science & Business Media.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, vol. 3, Cambridge university press.

Van Der Vaart, A. W., and J. A. Wellner (1996), *Weak convergence and empirical processes*, Springer.

Vershynin, R. (2018), *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press.

Wang, J., X. He, and G. Xu (2020), Debiased inference on treatment effect in a high-dimensional model, *Journal of the American Statistical Association*, *115*(529), 442–454.

Wang, L., I. Van Keilegom, and A. Maidman (2018), Wild residual bootstrap inference for penalized quantile regression with heteroscedastic errors, *Biometrika*, *105*(4), 859–872.

Wei, Y., R. D. Kehm, M. Goldberg, and M. B. Terry (2019), Applications for quantile regression in epidemiology, *Current Epidemiology Reports*, *6*(2), 191–199.

Wilcox, R. (2005), Trimming and winsorization, in *Encyclopedia of Biostatistics*, vol. 8, Wiley Online Library.

Wilson, H. G. (1978), Least squares versus minimum absolute deviations estimation in linear models, *Decision Sciences*, *9*(2), 322–335.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT Press.

Wu, T., and N. N. Narisetty (2021), Bayesian multiple quantile regression for linear models using a score likelihood, *Bayesian Analysis*, *1*(1), 1–29.

Wu, Y., and Y. Liu (2009), Variable selection in quantile regression, *Statistica Sinica*, *19*, 801–817.

Wu, Y., and G. Yin (2015), Conditional quantile screening in ultrahigh-dimensional heterogeneous data, *Biometrika*, *102*(1), 65–76.

Xi, R., Y. Li, and Y. Hu (2016), Bayesian quantile regression based on the empirical likelihood with spike and slab priors, *Bayesian Analysis*, *11*(3), 821–855.

Xiao, Z. (2014), Right-tail information in financial markets, *Econometric Theory*, *30*(1), 94–126.

Xu, Q., Y. Zhou, C. Jiang, K. Yu, and X. Niu (2016), A large cvar-based portfolio selection model with weight constraints, *Economic Modelling*, *59*, 436–447.

Yamai, Y., and T. Yoshiba (2005), Value-at-risk versus expected shortfall: A practical perspective, *Journal of Banking & Finance*, *29*(4), 997–1015.

Yang, Q., S. Greenland, and W. D. Flanders (2006), Associations of maternal age- and parity-related factors with trends in low-birthweight rates: United states, 1980 through 2000, *American Journal of Public Health*, *96*(5), 856–861.

Yang, Y., and X. He (2012), Bayesian empirical likelihood for quantile regression, *The Annals of Statistics*, *40*(2), 1102–1131.

Yang, Y., H. J. Wang, and X. He (2016), Posterior inference in bayesian quantile regression with asymmetric laplace likelihood, *International Statistical Review*, *84*(3), 327–344.

Yu, K., and R. A. Moyeed (2001), Bayesian quantile regression, *Statistics & Probability Letters*, *54*(4), 437–447.

Zhang, Y. D., B. P. Naughton, H. D. Bondell, and B. J. Reich (2022), Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior, *Journal of the American Statistical Association*, *117*(538), 862–874.

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *Journal of Machine Learning Research*, *7*(90), 2541–2563.

Zou, H. (2006), The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zwingmann, T., and H. Holzmann (2016), Asymptotics for the expected shortfall, *arXiv preprint arXiv:1611.07222*.