Statistica Sinica Preprint No: SS-2021-0338	
Title	Pseudo-Bayesian Approach for Quantile Regression Inference: Adaptation to Sparsity
Manuscript ID	SS-2021-0338
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0338
Complete List of Authors	Yuanzhi Li and Xuming He
Corresponding Authors	Yuanzhi Li
E-mails	yzli@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Statistica Sinica

Pseudo-Bayesian Approach for Quantile Regression Inference: Adaptation to Sparsity

Yuanzhi Li and Xuming He

Department of Statistics, University of Michigan

Abstract: Quantile regression is a powerful data analysis tool that accommodates heterogeneous covariate-response relationships. We find that by coupling the asymmetric Laplace working likelihood with appropriate shrinkage priors, we can deliver pseudo-Bayesian inference that adapts automatically to 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. By avoiding dichotomous variable selection, the Bayesian computational framework demonstrates desirable inference stability with respect to tuning parameter selection. Our work helps to uncloak the value of Bayesian computational methods in frequentist inference for quantile regression.

Key words and phrases: Asymmetric Laplace distribution, Increasing dimension, Optimal weighting, Posterior asymptotics. Shrinkage prior, Working likelihood.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

1. Introduction

Quantile regression, formally introduced by Koenker and Bassett Jr (1978), has become a powerful tool for data analysis in a wide range of applications, ranging from economics (Fitzenberger et al., 2013) to public health (Wei et al., 2019). Quantile regression enables researchers to go beyond the modeling of conditional means: By modeling the effects of covariates at different conditional quantile levels of a response variable, we obtain more comprehensive information on the relationships between the response and the covariates. In particular, quantile regression reveals the differential effects of a covariate on the low and high ends of the response distribution.

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, either directly or indirectly. Existing methods include using plugged-in density estimates (Powell, 1991; Hendricks and Koenker, 1992), rank-score tests (Gutenbrunner et al., 1993; Koenker and Machado, 1999), resampling methods (Feng et al., 2011; Pan and Zhou, 2020), and Bayesian computational approaches (Chernozhukov and Hong, 2003; Yang et al., 2016).

The present paper employs the Bayesian computational framework to provide frequentist inference for quantile regression. We show that the pseudo-Bayesian approach based on a working likelihood and a shrinkage prior achieves automatic adaptation to sparsity, and it can provide asymptotically valid inference for quan-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

tile regression under heterogeneity. We investigate the asymptotic properties of the posterior distribution in a possibly sparse model, and then demonstrate the desirable efficiency and stability of the proposed method with empirical results. We use the term "posterior inference" loosely to refer to statistical inference based on the Bayesian computational framework, even though we pursue inference validity in the frequentist sense.

More specifically, we consider the asymmetric Laplace working likelihood (Yu and Moyeed, 2001; Yang et al., 2016), with appropriate continuous shrinkage priors in the spirit of common frequentist penalty functions (Wu and Liu, 2009). With a random sample of size n from a linear quantile regression model with $p \leq n$ covariates but only $s \leq p$ active (nonzero) coefficients, our results offer the following insights into the posterior inference.

1. The posterior distribution concentrates around the true quantile regression parameters at an adaptive rate: it achieves the $n^{-1/2}$ rate for active coefficients and a super-efficient rate of $o(n^{-1/2})$ for inactive (zero-valued) coefficients.

2. The posterior mean for the active coefficients is asymptotically normal and oracle efficient: it achieves the same asymptotic variance as that of the quantile regression estimator where we know which coefficients are active, yet without relying on explicit variable selection.

3. With an appropriate adjustment of the posterior variance, we construct auto-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

matically adaptive confidence intervals in the frequentist sense: they are asymptotically oracle for the active coefficients, while super-efficient for the inactive coefficients with coverage probabilities tending to one.

4. Even if we identify the active covariates correctly, we cannot obtain optimally weighted quantile regression estimators by focusing on these covariates only. Our proposed pseudo-Bayesian approach with continuous shrinkage priors does not rely on a binary selection of active/inactive covariates; thus, it can offer performance advantages over variable selection approaches.

It is important that unadjusted Bayesian inference is not automatically valid, because the posterior is constructed operationally from a misspecified asymmetric Laplace working likelihood. Even for finite-dimensional models without the use of shrinkage priors, the posterior distribution does not approximate the sampling distribution of the classical quantile regression estimator (Sriram, 2015; Yang et al., 2016). From the frequentist perspective, however, we find a relatively simple adjustment to the posterior variance that facilitates asymptotically valid and adaptive inference in possibly sparse quantile regression models, generalizing the work of Yang et al. (2016). The Bayesian computational framework allows us to circumvent the nonparametric estimation of the conditional density functions as nuisance parameters (Chernozhukov and Hong, 2003), thus serving as a valuable tool for frequentist inference.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Bayesian modeling with shrinkage priors has been quite well studied in terms of estimation accuracy (error rates) of the parameters and variable selection in high-dimensional problems; see, e.g., Narisetty et al. (2014), Song and Liang (2017), Jiang and Sun (2019), and Gao et al. (2020). The focus of the present paper is not the posterior contraction rate or variable selection consistency, but the understanding of what can be accomplished in inference for possibly sparse quantile regression models, about which relatively little has been available in the literature even when the number of predictors p is fixed. To the best of our knowledge, our work is also the first to provide an asymptotic analysis for the posterior mean and variance in the Bayesian quantile regression framework with a shrinkage prior. The main challenge in our setting is adjusting for the misspecification of the likelihood function under heterogeneity and model sparsity. To simplify the technicalities and focus on the main points, we begin by working with an asymptotic framework where the sample size n goes to infinity yet the covariate dimension p is kept fixed. We extend to the regime where p can diverge to infinity later in the paper.

The rest of the paper is organized as follows. In Section 2, we discuss the quantile regression problem and our pseudo-Bayesian framework. Then, we present the corrected posterior inference approach in Section 3, supported by the asymptotic properties of the posterior distribution. In Section 4, we extend our theoretical analysis to the asymptotic regime with an increasing covariate dimension. Section 5

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

shows some simulation results to demonstrate the effectiveness and stability of the proposed approach. Section 6 concludes the paper.

2. Problem setup

2.1 The quantile regression model

Let $Q_\tau(Y \mid X = \boldsymbol{x})$ be the τth conditional quantile of the response variable Y given covariates $X = \boldsymbol{x}$, where $\boldsymbol{x} = (x_0, \ldots, x_p)^{\mathrm{T}}$ includes an intercept term $x_0 = 1$ and p covariates, and $\tau \in (0, 1)$ is a pre-specified quantile level of interest. We consider the linear quantile regression model

$$Q_\tau(Y \mid X = \boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}^0(\tau), \tag{2.1}$$

where $\boldsymbol{\beta}^0(\tau) = (\beta_0^0(\tau), \ldots, \beta_p^0(\tau))^{\mathrm{T}}$ is the true quantile regression coefficient vector. The conditional median of $\tau = 0.5$ is a special case, and high or low quantile levels of τ are often of interest in, e.g., financial risk quantification (Taylor, 2019) and public health assessment (Wei et al., 2019). Because we focus on a fixed τ in Model (2.1), we suppress the index τ in $\boldsymbol{\beta}^0(\tau)$ in the rest of the paper for notational convenience.

In this paper, we consider Model (2.1) to be possibly sparse. Let $\mathcal{S} = \{0\} \cup \{j \in \{1, \ldots p\} : \beta_j^0 \neq 0\}$ be the index set of the active (nonzero) coefficients, including the intercept term; $\mathcal{S}^c = \{0, \ldots, p\} \backslash \mathcal{S}$ is the set of inactive coefficients. Let $s = |\mathcal{S}| - 1$ be the number of active covariates. A possibly sparse model refers to $0 \leq s \leq p$ for

some integer s; yet neither s nor $\mathcal{S}$ is known in advance. For now, we suppose the covariate dimension p is fixed; later in Section 4, we extend our theory to the case where p can increase with the sample size.

We briefly review the classical quantile regression analysis. Let $\mathbb{D}_n = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ be a random sample of size n that satisfies Model (2.1). The quantile regression estimator (Koenker and Bassett Jr, 1978) is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{u} \in \mathbb{R}^{(p+1)}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u}), \tag{2.2}$$

where $\rho_\tau(v) = v\{\tau - 1(v < 0)\}$ and $1(\cdot)$ is the indicator function. With $p \ll n$, one can perform statistical inference for Model (2.1) based on the asymptotic properties of the estimator $\widehat{\boldsymbol{\beta}}$; refer to Koenker (2005) and Koenker et al. (2017) for further discussion of quantile regression. Here, we highlight two aspects of the estimator $\widehat{\boldsymbol{\beta}}$: (i) it does not account for the possible model sparsity, and therefore does not achieve optimal efficiency when Model (2.1) is sparse; (ii) its asymptotic variance-covariance matrix involves the conditional density function of Y given X, which requires non-parametric estimation that can be unstable in practice.

2.2 A pseudo-Bayesian framework

In this section, we present the pseudo-Bayesian framework for modeling the quantile regression coefficient $\boldsymbol{\beta}$ in Model (2.1). We adopt the asymmetric Laplace working

likelihood:

$$\mathcal{L}(\mathbb{D}_n \mid \boldsymbol{\beta}) \propto \exp\left\{-\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right\}, \tag{2.3}$$

where $\propto$ means equality up to a multiplicative factor that does not depend on $\boldsymbol{\beta}$. We call $\mathcal{L}(\mathbb{D}_n|\boldsymbol{\beta})$ a working likelihood because it does not correspond to the true data-generating mechanism of $\mathbb{D}_n$ under parameter value $\boldsymbol{\beta}$; in fact, there is no "true" likelihood function, because Model (2.1) does not fully specify a conditional distribution of Y given X. Choosing a working likelihood in the form of (2.3) enjoys two benefits: (i) it allows the maximum working likelihood estimator to coincide with the classical quantile regression estimator $\widehat{\boldsymbol{\beta}}$ in (2.2); and (ii) its Fisher information matrix shares a critical component with the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ (Yang et al., 2016).

To incorporate possible model sparsity, we consider two examples of shrinkage priors in the spirit of common penalty functions:

$$\pi_{AL}(\boldsymbol{\beta}) \propto \exp\left\{-n^{1/2}\lambda_n \sum_{j=1}^{p} w_j|\beta_j|\right\}, \tag{2.4}$$

$$\pi_{CA}(\boldsymbol{\beta}) \propto \exp\left\{-n \sum_{j=1}^{p} p_{\lambda_n}(\beta_j)\right\}, \tag{2.5}$$

where w_j and the function $p_{\lambda_n}(\cdot)$ are given below; the tuning parameter λ_n depends on the sample size, but we sometimes omit the subscript n when there is no confusion. The prior (2.4) corresponds to the adaptive lasso (AL) penalty (Zou, 2006), where $w_j = 1/|\hat{\beta}_j|$ for $j \in \{1, \ldots, p\}$ as in Wu and Liu (2009), and $\hat{\beta}_j$ is the jth component

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

of $\widehat{\boldsymbol{\beta}}$ defined in (2.2). In the clipped absolute (CA) prior (2.5) we define $p_\lambda(u) = \lambda(|u| \wedge \lambda)$, which is motivated by 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 1 for a visual comparison. For either (2.4) or (2.5), the prior on β_0 is flat, that is, $\pi(\beta_0) \propto 1$; therefore, β_0 is not penalized. We discuss the prior choice further in the next subsection.

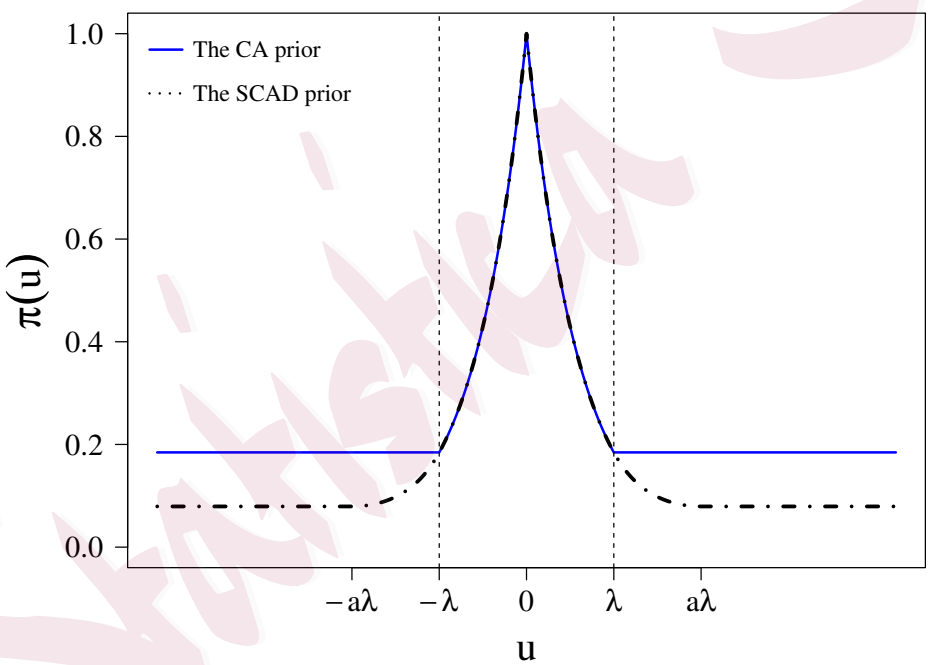

Figure 1: Comparison between the prior $\pi_{CA}(u)$ and the prior induced by the SCAD penalty of Fan and Li (2001); a is a tuning parameter in the SCAD penalty, and we set $a = 2$ in the plot. Both priors are flat when $|u| > a\lambda$.

Given the working likelihood (2.3) and any prior $\pi(\boldsymbol{\beta})$, we have the formal posterior density

$$p\left(\boldsymbol{\beta} \mid \mathbb{D}_n\right) \propto \mathcal{L}(\mathbb{D}_n \mid \boldsymbol{\beta}) \times \pi(\boldsymbol{\beta}). \tag{2.6}$$

Under either the AL (2.4) or the CA (2.5) prior, existing Markov chain Monte Carlo (MCMC) algorithms enable efficient sampling from the posterior; see Li et al. (2010) and Alhamzawi et al. (2012) for the prior (2.4); and Li (2011) and Adlouni et al. (2018) for priors similar to (2.5). Note that although the CA prior (2.5) is improper, that is, integration of $\pi_{CA}(\boldsymbol{\beta})$ over $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ diverges, the posterior (2.6) under the CA prior is still proper; see Proposition S1 in the Supplementary Material. In the rest of this paper, we examine the asymptotic properties of the posterior distribution, from which we derive valid and adaptive confidence intervals in the frequentist sense.

2.3 Discussion on the choice of prior

Priors (2.4) and (2.5) are both examples from a general family of continuous shrinkage priors; see e.g., Song and Liang (2017), Bhadra et al. (2019) and Zhang et al. (2022) for discussions in the mean regression context. Common to these priors is that the shrinkage is meant to be *adaptive* in a possibly sparse model, and such adaptivity is central to our main results. The priors (2.4) and (2.5), though relatively simple, are sufficient to demonstrate such adaptivity in our setting, and therefore we focus on these two choices.

Here, we highlight the adaptivity of the priors (2.4) and (2.5), which is not shared by all shrinkage priors; see also Remark 1. In the AL prior, each β_j is subject to a different scaling factor w_j that is chosen adaptively from the data, hence enabling adaptive shrinkage. On the other hand, the CA prior is not data dependent, but has several desirable features (Song and Liang, 2017): (i) it has a sharp peak near zero, which shrinks the smaller coefficients towards zero; and (ii) it has a fat tail, which allows large coefficients to be unpenalized. Therefore, both priors are adaptive, owning to their distinct features.

There are many other adaptive shrinkage priors designed specifically for the Gaussian mean regression setting as in Carvalho et al. (2010), Bhattacharya et al. (2015); and Zhang et al. (2022), with the latter motivating their choice from a prior on the coefficient of determination R^2. However, few works have focused on Bayesian quantile regression, because of the absence of a true likelihood. While some existing priors can be adapted operationally for quantile regression (Alhamzawi et al., 2012; Chen et al., 2013; Kohns and Szendrei, 2020), we show in the online Supplementary Material that not all such priors are appropriate for posterior inference. In this paper, we do not promote any particular prior choice; instead, we use the relatively simple AL and CA priors to illustrate the properties of posterior inference in possibly sparse quantile regression models.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

3. Adaptive posterior inference

While posterior inference seems straightforward from the Bayesian perspective, its validity is not warranted for our pseudo-Bayesian approach because the working likelihood is misspecified (Yang et al., 2016). In this section, we begin by investigating the asymptotic properties of the posterior distribution from the frequentist perspective. Next, we propose an adjustment to the posterior variance-covariance matrix, and show that it can lead to valid confidence intervals that adapt to model sparsity. In the last subsection, we discuss an extension in which we use a weighted working likelihood to obtain optimal efficiency. Throughout this section, the covariate dimension p is fixed in Model (2.1).

3.1 Notation

Recall that we have $\boldsymbol{\beta}^0 = (\beta_0^0, \ldots, \beta_p^0)^{\mathrm{T}}$ as the true regression coefficient in Model (2.1), and that $\mathcal{S}$ is the index set of the active (nonzero) coefficients, including the intercept term. Without loss of generality, we assume $\mathcal{S} = \{0, 1, \ldots, s\}$. Recall $\widehat{\boldsymbol{\beta}}$ is the classical quantile regression estimator in (2.2); let $\tilde{\boldsymbol{\beta}}_{\mathcal{S}} \in \mathbb{R}^{s+1}$ be the oracle quantile regression estimator, which solves (2.2) using only the active covariates. For any vector $\boldsymbol{v} = (v_0, \ldots, v_p)^{\mathrm{T}}$, let $\boldsymbol{v}_{\mathcal{S}} = \{v_j : \ j \in \mathcal{S}\}$ and $\boldsymbol{v}_{\mathcal{S}^c} = \{v_j : \ j \notin \mathcal{S}\}$. For any

matrix $A \in \mathbb{R}^{(p+1)\times(p+1)}$, we partition

$$A = \begin{pmatrix} A_{\mathcal{S}} & A_{\mathcal{S},\mathcal{S}^c} \\ A_{\mathcal{S}^c,\mathcal{S}} & A_{\mathcal{S}^c} \end{pmatrix},$$

where $A_{\mathcal{S}} \in \mathbb{R}^{(s+1)\times(s+1)}$; for $i, j \in \{0, \ldots, p\}$, we write $A(i, j)$ as the $(i+1, j+1)$th entry of A.

Recall that $\mathbb{D}_n$ is a random sample of size n from the distribution $(X, Y) \sim \mathrm{pr}^*$, where the τth conditional quantile of Y satisfies Model (2.1). We also use $\mathrm{E}^*(\cdot)$ as the expectation operator under pr^*. Let $\epsilon = Y - X^{\mathrm{T}}\boldsymbol{\beta}^0$, and let $f_{\epsilon|X}$ (or $f_{\epsilon|X_{\mathcal{S}}}$) be the conditional density function of ϵ given X (or $X_{\mathcal{S}}$). Furthermore, let $D = \mathrm{E}^*(XX^{\mathrm{T}})$ and $G = \mathrm{E}^*\{XX^{\mathrm{T}} f_{\epsilon|X_{\mathcal{S}}}(0)\}$. Given the data $\mathbb{D}_n$ and the prior π, we consider the posterior probability measure as

$$\Pi\left(\mathcal{A} \mid \mathbb{D}_n\right) = \int_A p(\boldsymbol{\beta} \mid \mathbb{D}_n)\, \mathrm{d}\boldsymbol{\beta},$$

for any measurable set $\mathcal{A} \subset \mathbb{R}^{(p+1)}$, where $p(\boldsymbol{\beta} \mid \mathbb{D}_n)$ is the posterior density in (2.6).

We also use the following notation. For a vector $\boldsymbol{v}$, let $\|\boldsymbol{v}\|$ and $\|\boldsymbol{v}\|_\infty$ be its ℓ_2-norm and its maximum norm, respectively. For a matrix A, we denote its maximal/minimal eigenvalue by $\theta_{\max}(A)$ and $\theta_{\min}(A)$, respectively. For probability density functions $h(x)$ and $g(x)$, we denote their total variation distance by $\|h - g\|_{TV} = \int |h - g| \mathrm{d}x$. For covariance matrices A and B, we write $A \preceq B$ if $B - A$ is positive semi-definite. 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 $C_1 > 0$ such that $a_n \leq C_1 b_n$. For any two stochastic sequences $\hat{a}_n$ and $\hat{b}_n$, we use $\hat{a}_n \ll_{\mathrm{pr}^*} \hat{b}_n$ and $\hat{a}_n \lesssim_{\mathrm{pr}^*} \hat{b}_n$ to denote $\hat{a}_n = o_{\mathrm{pr}^*}(\hat{b}_n)$ and $\hat{a}_n = O_{\mathrm{pr}^*}(\hat{b}_n)$, respectively; we define $\hat{a}_n \asymp_{\mathrm{pr}^*} \hat{b}_n$ if both $\hat{a}_n = O_{\mathrm{pr}^*}(\hat{b}_n)$ and $\hat{b}_n = O_{\mathrm{pr}^*}(\hat{a}_n)$ hold.

3.2 Posterior asymptotics

In this subsection, we present the large-sample properties of the posterior distribution defined in (2.6). To this end, we need the following technical assumptions.

Assumption 1 (Identification)**.** For any $\delta > 0$, there exists $\varepsilon > 0$, such that

$$\lim_{n\to\infty} \mathrm{pr}^* \left[\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|\geq\delta} \left\{ \frac{L_n(\boldsymbol{\beta}^0) - L_n(\boldsymbol{\beta})}{n} \right\} \leq -\varepsilon \right] = 1,$$

where $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau(y_i - \boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\beta})$.

Assumption 2 (Covariates)**.** The covariate vector X has bounded support on $\mathcal{X} \subset \mathbb{R}^{p+1}$. Furthermore, the eigenvalues of $D = \mathrm{E}^*(XX^{\mathrm{T}})$ are all bounded away from 0 and $+\infty$.

Assumption 3 (Conditional densities)**.** The conditional density function of $\epsilon = Y - X^{\mathrm{T}}\boldsymbol{\beta}^0$ given $X = \boldsymbol{x}$ satisfies the following: (i) there exists $L > 0$, such that for all $u, u' \in \mathbb{R}$,

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f_{\epsilon|X=\boldsymbol{x}}(u) - f_{\epsilon|X=\boldsymbol{x}}(u')| \leq L|u - u'|;$$

and (ii) there exist two constants $\underline{f}$ and $\overline{f}$, such that

$$0 < \underline{f} \leq \inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ f_{\epsilon|X=\boldsymbol{x}}(0) \right\} \leq \sup_{\substack{u\in\mathbb{R} \\ \boldsymbol{x}\in\mathcal{X}}} \left\{ f_{\epsilon|X=\boldsymbol{x}}(u) \right\} \leq \overline{f}.$$

Assumption 4 (Separation)**.** For some constant $b_0 > 0$, we have

$$\min_{j\in\mathcal{S}\setminus\{0\}} |\beta_j^0| > b_0.$$

We briefly discuss the assumptions. Assumptions 1–3 are standard in pseudo-Bayesian modeling with a working likelihood (Chernozhukov and Hong, 2003; Yang et al., 2016) and in the quantile regression literature (Knight, 1998; Pan and Zhou, 2020); see also Koenker (2005, Section 4). In particular, the two assertions in Assumption 3 hold for the conditional density $f_{\epsilon|X_{\mathcal{S}}}(u)$ as well. Furthermore, Assumption 2 implies that the eigenvalues of $G = \mathrm{E}^*\{XX^{\mathrm{T}} f_{\epsilon|X_{\mathcal{S}}}(0)\}$ are also bounded. Assumption 4 holds automatically when we posit a fixed model as (2.1), where p is a constant; Similar separation conditions are needed to achieve consistent model selection (Fan and Li, 2001; Wu and Liu, 2009; Belloni et al., 2011).

Now, we present the main theoretical result on the posterior distribution defined in (2.6).

Theorem 1. *Consider the posterior distribution under either the AL prior (2.4) or the CA prior (2.5). Suppose Assumptions 1–4 hold, and the tuning parameter λ satisfies $n^{-1/2} \ll \lambda \ll 1$. Then, we have the following results:*

1. *Adaptive rate of contraction: for any sequence* $M_n \to +\infty$,

$$\Pi\left(\left\|\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}^0_{\mathcal{S}}\right\| \leq \frac{M_n}{n^{1/2}}\,,\, \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_\infty \leq \frac{M_n}{n\lambda} \;\middle|\; \mathbb{D}_n\right) \to 1,$$

in pr**-probability.*

2. *Distributional approximation: for some density functions* $\pi_j(u) = O_{\mathrm{pr}^*}(1)$ $(u \in \mathbb{R}, j \in \mathcal{S}^c)$,

$$\left\| p\left(\boldsymbol{\beta} \mid \mathbb{D}_n\right) - \phi\left(\boldsymbol{\beta}_S; \tilde{\boldsymbol{\beta}}_S, \frac{1}{n} G_{\mathcal{S}}^{-1}\right) \times \prod_{j \notin \mathcal{S}} \{n\lambda \pi_j(n\lambda\beta_j)\} \right\|_{TV} \to 0,$$

in pr**-probability, where* $\phi(\cdot\,;\mu,\Sigma)$ *is the density function of a multivariate-Gaussian distribution. In particular,* $\pi_j(u) = (n^{-1/2}w_j/2)\exp\{-n^{-1/2}w_j|u|\}$ *if we use the AL prior (2.4), and* $\pi_j(u) = (1/2)\exp\{-|u|\}$ *if we use the CA prior (2.5).*

Theorem 1 shows that, despite the misspecification of the likelihood, the posterior under either prior can separate the active and inactive coefficients with a wide range of choices of λ. With $n\lambda \gg n^{1/2}$, part 1 of Theorem 1 shows that the posterior for the inactive coefficients concentrates toward zero at a second-order rate, which is super-efficient. Furthermore, part 2 of Theorem 1 shows that the posterior for $\boldsymbol{\beta}_{\mathcal{S}}$ and $\boldsymbol{\beta}_{\mathcal{S}^c}$ are approximately independent. In particular, the posterior for $\boldsymbol{\beta}_{\mathcal{S}}$ is "oracle", that is, the Gaussian limiting posterior for $\boldsymbol{\beta}_{\mathcal{S}}$ is the same as if we knew the true model $X_{\mathcal{S}}$ in advance (Sriram, 2015), regardless of the prior we use. Thus,

using the two shrinkage priors in Section 2.2, the posterior distribution can adapt automatically to the model sparsity.

Although slightly different in the limit, the posterior shares the same adaptation principle under both the AL and the CA priors in Section 2.2. For an active coefficient, the prior casts no asymptotic effect on the posterior distribution; For an inactive coefficient $\beta_j (j \in \mathcal{S}^c)$, the shrinkage prior dominates the working likelihood, because the limiting posterior density $n\lambda \times \pi_j(n\lambda\beta_j)$ is proportional to the corresponding prior when $|n\lambda\beta_j| = O(1)$. Therefore, the shrinkage prior can separate the inactive coefficient from the active ones, which is in line with Theorem 2.4 in Song and Liang (2017) for the Gaussian linear model setting.

Remark 1. The adaptivity of the posterior shrinkage in Theorem 1 is not shared under all popular Bayesian priors. For example, Castillo et al. (2015) shows that the traditional Bayesian lasso (Park and Casella, 2008) cannot 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.

3.3 Confidence intervals from posterior moments

Since the working likelihood (2.3) is misspecified, Theorem 1 alone does not imply correct inference for quantile regression, and the posterior needs to be properly calibrated. However, the correction on the posterior variance proposed in Yang et al.

(2016) is no longer valid when we use shrinkage priors. In light of Theorem 1, we give a modified adjustment that yields confidence intervals based on posterior moments that are automatically adaptive to model sparsity.

We construct the confidence intervals for $\boldsymbol{\beta}^0$ based on the posterior mean $\check{\boldsymbol{\beta}} = (\check{\beta}_0, \ldots, \check{\beta}_p)$ and the posterior variance-covariance matrix $\check{\Sigma}$ obtained from any posterior sampling algorithm. We start from the adjustment used in Yang et al. (2016) by letting $\widehat{D} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}/n$ and $\check{\Sigma}_{adj} = n\tau(1-\tau)\check{\Sigma}\,\widehat{D}\,\check{\Sigma}$. Our proposed level $1-\alpha$ confidence interval for each β_j^0 takes the form

$$\check{\beta}_j \pm z_{\alpha/2}\eta_j \left\{\check{\Sigma}_{adj}(j,j)\right\}^{1/2}, \quad j \in \{0, 1, \ldots, p\}, \tag{3.1}$$

where $\eta_j = \min\{n^{1/2}\lambda, \max\{1, \lambda/|\hat{\beta}_j|\}\}$ is the adjustment weight, and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. Theorem 2 gives several properties of the proposed interval (3.1).

Theorem 2. *Consider the posterior distribution under either the AL prior (2.4) or the CA prior (2.5). Under the conditions of Theorem 1, we have the following results:*

1. *Convergence of the posterior mean:*

$$\begin{aligned} n^{1/2}(\check{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^0) &\to \mathrm{N}\left\{0\,,\, \tau(1-\tau)G_{\mathcal{S}}^{-1}D_{\mathcal{S}}G_{\mathcal{S}}^{-1}\right\}, \\ n\lambda(\check{\boldsymbol{\beta}}_{\mathcal{S}^c} - \mathbf{0}) &\to 0, \end{aligned}$$

in distribution as $n \to \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

2. Properties of the adjusted variance:

$$n\,\check{\Sigma}_{adj,\mathcal{S}} = \tau(1-\tau)G_{\mathcal{S}}^{-1}D_{\mathcal{S}}G_{\mathcal{S}}^{-1} + o_{\mathrm{pr}^*}(1),$$

$$(n^{1/2}\lambda)^{-2} \lesssim_{\mathrm{pr}^*} (n\lambda)^2\,\check{\Sigma}_{adj}(j,j) \ll_{\mathrm{pr}^*} 1, \quad j \notin \mathcal{S}.$$

Theorem 2 informs us of several aspects of the proposed inferential approach. First, the posterior mean for the active coefficient is first-order equivalent to the oracle quantile regression estimator, as if we knew the set $\mathcal{S}$. Furthermore, the adjusted posterior variance-covariance matrix captures the sampling variance-covariance of the posterior mean. For those coefficients, the adjustment weight $\eta_j = 1 + o_{\mathrm{pr}^*}(1)$, because $n^{-1/2} \ll \lambda \ll 1$; hence the confidence intervals in the form of (3.1) can be viewed as standard Wald-type intervals in the oracle model.

Next, we consider any inactive coefficient β_j, where $j \notin \mathcal{S}$. In this case, $\hat{\beta}_j = O_{\mathrm{pr}^*}(n^{-1/2})$, so the adjustment weight $\eta_j \asymp_{\mathrm{pr}^*} n^{1/2}\lambda \to \infty$ in (3.1) works to inflate the Wald-type interval. Theorem 2 implies

$$\frac{\check{\beta}_j - 0}{\eta_j\left\{\check{\Sigma}_{adj}(j,j)\right\}^{1/2}} \to 0, \quad n^{1/2}\,\eta_j\left\{\check{\Sigma}_{adj}(j,j)\right\}^{1/2} \to 0, \quad j \notin \mathcal{S},$$

in pr*-probability. Therefore, the confidence interval in (3.1) achieves a conservative 100% asymptotic coverage probability, but the interval length remains super-efficient at the order of $o_{\mathrm{pr}^*}(n^{-1/2})$.

In summary, the proposed procedure is valid for all coefficients, and the resulting confidence intervals (3.1) are automatically adaptive to possible sparsity in the model

without relying on a dichotomous variable selection step. In a sparse model ($s < p$), such interval estimates are more efficient than the classical quantile regression inference using all the coefficients. Empirically, we later show that the proposed intervals are less sensitive to tuning than direct quantile regression inference following model selection.

Remark 2. (Value of the Bayesian computational framework) Theorems 1 and 2 show that the posterior variance-covariance matrix approximates $G_{\mathcal{S}}^{-1}$, which is an essential quantity for oracle inference in quantile regression (Yang et al., 2016). Because $G_{\mathcal{S}}$ involves the conditional density function of Y given X, common frequentist approaches require non-parametric estimation, even if we know the true model. Refer to Chernozhukov and Hong (2003) for an in-depth discussion of frequentist inference via MCMC.

Remark 3. (Statistical efficiency for inactive coefficients) For an inactive coefficient β_j, Theorem 2 suggests the width of its confidence interval (3.1), denoted by ℓ_n, satisfies

$$\frac{1}{n\lambda_n} \ll \ell_n \ll \frac{1}{n^{1/2}}.$$

However, Theorem 1 suggests that the unadjusted posterior distribution for β_j is at the scale of $1/(n\lambda_n)$, which is of higher order than ℓ_n. Therefore, statistical efficiency of the pseudo-Bayesian inference involves more than just the convergence rate of the posterior distribution. Under a misspecified likelihood, we need to investigate the

distributional properties of the posterior mean, together with any necessary adjustment, in order to achieve valid inference. This is where our work differs from others in the literature that focus on the concentration of the posterior distributions.

3.4 Optimally weighted posterior inference

In the presence of heteroscedasticity, it is well-known since Newey (1990) that the following optimally weighted quantile regression estimator is semi-parametric efficient for estimating $\boldsymbol{\beta}^0$ when there is no sparsity:

$$\hat{\boldsymbol{\beta}}^{(w)} = \arg\min_{\boldsymbol{u}\in\mathbb{R}^{(p+1)}} \sum_{i=1}^{n} \zeta_i \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{u}),$$

where $\zeta_i = f_{\epsilon|X=\boldsymbol{x}_i}(0)$. In a possibly sparse quantile regression model, a natural question is whether we can achieve the optimal semi-parametric efficiency by using only the data on $(X_{\mathcal{S}}, y)$, i.e., after "oracle" model selection is attained. The answer is, somewhat surprisingly, negative, because the "optimal" weights $f_{\epsilon|X_{\mathcal{S}}}(0)$ under the "oracle model" do not capture the full heteroscedasticity in the data. Instead, we show that the statistical efficiency can be further improved in our pseudo-Bayesian framework with the following optimally weighted asymmetric Laplace working likelihood,

$$\mathcal{L}^{(w)}(\mathbb{D}_n \mid \boldsymbol{\beta}) \propto \exp\left\{-\sum_{i=1}^{n} \zeta_i \rho_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right\}. \tag{3.2}$$

Coupling (3.2) with the shrinkage priors in Section 2.2, we obtain the posterior density $p^{(w)}(\boldsymbol{\beta} \mid \mathbb{D}_n)$. Let $\check{\boldsymbol{\beta}}^{(w)}$ be the posterior mean under the weighted likelihood.

The following result gives the sampling distribution of the posterior mean for the active coefficients.

Proposition 1. *Consider the weighted working likelihood (3.2) and either of the prior (2.4) or (2.5). Under the same conditions as those in Theorem 1, the posterior mean satisfies*

$$n^{1/2}(\check{\boldsymbol{\beta}}_{\mathcal{S}}^{(w)} - \boldsymbol{\beta}_{\mathcal{S}}^{0}) \quad \to \quad \mathrm{N}\left\{0\,,\, \tau(1-\tau)Q_{\mathcal{S}}^{-1}\right\},$$

in distribution, where $Q_{\mathcal{S}} = \mathrm{E}^{*}\{X_{\mathcal{S}}X_{\mathcal{S}}^{\mathrm{T}} f_{\epsilon|X}^{2}(0)\}$.

On the other hand, if classical quantile regression is applied to $(X_{\mathcal{S}}, Y)$ with the inactive covariates left out, the "optimally" weighted quantile regression has an asymptotic variance of $\tau(1-\tau)V_{\mathcal{S}}^{-1}$ (Newey and Powell, 1990), where $V_{\mathcal{S}} = \mathrm{E}^{*}\{X_{\mathcal{S}}X_{\mathcal{S}}^{\mathrm{T}} f_{\epsilon|X_{\mathcal{S}}}^{2}(0)\}$ relies only on the active covariates. We show in the Supplementary Material that

$$Q_{\mathcal{S}}^{-1} \preceq V_{\mathcal{S}}^{-1}. \tag{3.3}$$

Thus, focusing only on the oracle quantile regression model (even when it is available) does not lead to optimal efficiency for the active coefficients.

There is a simple reason why the inactive set of covariates should not be abandoned. Even though $X_{\mathcal{S}^c}$ does not affect the conditional τth quantile of Y given X, it may still affect other aspects of the conditional distribution of Y given X; in particular, the density function $f_{\epsilon|X}(0)$ may depend on $X_{\mathcal{S}^c}$. Unless $f_{\epsilon|X}(0) = f_{\epsilon|X_{\mathcal{S}}}(0)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

optimal efficiency for quantile regression analysis cannot be achieved if we focus only on the active covariates. In general, a true "oracle" model should identify covariates that affect the conditional density function $f_{\epsilon|X}(0)$, alongside with $X_{\mathcal{S}}$.

Remark 4. To focus on the main idea, we suppose that the optimal weight $\zeta_i = f_{\epsilon|X=\boldsymbol{x}_i}(0)$ in (3.2) is known. In practice, it is possible to use estimated weights and achieve the same asymptotic efficiency as if we knew ζ_i; see, e.g., Newey and Powell (1990), Koenker and Zhao (1994), and Zhao (2001) for some theoretical investigations. See also Section S1.6 of the online Supplementary Material for empirical results using estimated weights.

4. Posterior inference with diverging dimensions

In this section, we extend the results in Section 3.2 to the case with a large number of covariates. Here we focus on the case where the covariate dimension $p = p_n$ diverges with, while still at a smaller order of, the sample size n. In Section 4.2, we discuss some practical recommendations when the dimension is even higher, i.e., p_n may grow faster than the sample size. Under the asymptotic regimes of this section, the true model parameter $\boldsymbol{\beta}^0 = \boldsymbol{\beta}_n^0$ may depend on n; nevertheless we sometimes suppress the index n for the ease of presentation.

4.1 Posterior asymptotics under moderately increasing dimensions

Here, we consider the asymptotic regime of $p = p_n \ll n$. Under this setting, we also allow the size of the active covariates, $|\mathcal{S}| = s_n$, to grow with the sample size. For illustration purposes, we only focus on the CA prior (2.5), and show that the posterior distribution still achieves adaptation to sparsity, even in the regime of moderately increasing dimensions.

The asymptotic regime with a moderately increasing dimension is often of practical interest. In conditional quantile modeling, it is common that the complexity of Model (2.1) may depend on the available sample size. A common example is when we approximate the unknown conditional quantile function by a linear combination of series/basis expansions, e.g., B-splines, polynomials, and wavelets (Chao et al., 2017; Belloni et al., 2019). To control the approximation error, the number of basis functions typically increases with the sample size at a certain rate (He and Shi, 1994). The regime also covers the so-called "many regressors" model in econometrics, where a large number of variables is often necessary to model economic theories (Cattaneo et al., 2018).

We first discuss some generalizations of the conditions in Section 3.2 when the dimension $p_n \to \infty$. With $p_n = o(n)$, Assumptions 1 and 3 are standard in the quantile regression literature (Belloni et al., 2019; Pan and Zhou, 2020). On the other hand, Assumptions 2 and 4 may not be suitable for the increasing dimensional

regime, therefore we adjust them as follows.

Assumption 2′ (Covariates)**.** There exists a constant $\sigma_0 > 0$, such that for all $\|u\| = 1$ and $t > 0$,

$$\mathrm{pr}^*\left(|u^T D^{-1/2} X| \geq \sigma_0 t\right) \leq 2\mathrm{e}^{-t}. \tag{4.1}$$

Furthermore, the eigenvalues of the matrix $D = \mathrm{E}^*[XX^T]$ satisfy

$$p_n^{-1} \lesssim \theta_{\min}(D) \leq \theta_{\max}(D) \lesssim p_n \quad \text{and} \quad \theta_{\min}(D_{\mathcal{S}}) \geq \theta_1 > 0, \tag{4.2}$$

for some constant $\theta_1 > 0$.

Assumption 4′ (Sparsity)**.** There exists a sequence $\underline{b}_n > 0$ such that for each n,

$$\min_{j \in \mathcal{S} \setminus \{0\}} |\beta_j^0| > \underline{b}_n.$$

Assumption 2′ consists of two parts. First, (4.1) states that the standardized covariate $D^{-1/2}X$ is sub-exponential, which strengthens the boundedness of X in Assumption 2; see Vershynin (2018, Section 3.3) for examples of sub-exponential distributions in high dimensions. Second, (4.2) relaxes Assumption 2 by allowing certain co-linearity among the $p = p_n$ covariates. Finally, Assumption 4′ requires that all nonzero coefficients be sufficiently separated from zero, yet the threshold $\underline{b}_n$ is allowed to shrink toward zero as the sample size grows.

Our next result generalizes Theorem 1 to an increasing dimensional regime, where we drop the subscript n in s and p for simplicity.

Theorem 3. *Consider the posterior distribution under the CA prior (2.5) and $p \to \infty$. Suppose Assumptions 1, 2′, 3, and 4′ hold. If $s^4 p^2 \log^2 n = o(n)$, and the tuning parameter λ is chosen such that*

$$\frac{s^{1/2} p \log^{3/2} p}{n^{1/2}} \ll \lambda \ll \min\left\{s^{-1/2},\ \underline{b}_n,\ \underline{b}_n[\theta_{\min}(D)]^{1/2}\right\}. \tag{4.3}$$

Then, we have the following results:

1. *Adaptive rate of contraction: for any sequence $M_n \to +\infty$,*

$$\Pi\left(\left\|\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}^0_{\mathcal{S}}\right\| \le M_n\sqrt{\frac{s}{n}},\ \left\|\boldsymbol{\beta}_{\mathcal{S}^c}\right\|_\infty \le M_n \frac{s\log p}{n\lambda} \,\middle|\, \mathbb{D}_n\right) \to 1,$$

   in pr**-probability.*

2. *Distributional approximation: for $\pi_j(u) = (1/2)\exp\{-|u|\}$, $(\forall j \in \mathcal{S}^c)$,*

$$\left\| p\left(\boldsymbol{\beta} \mid \mathbb{D}_n\right) - \phi\left(\boldsymbol{\beta}_S; \tilde{\boldsymbol{\beta}}_S, \frac{1}{n}G_{\mathcal{S}}^{-1}\right) \times \prod_{j\notin\mathcal{S}} \{n\lambda\pi_j(n\lambda\beta_j)\} \right\|_{TV} \to 0,$$

   in pr**-probability, where $\phi(\cdot\,;\mu,\Sigma)$ is the density function of a multivariate-Gaussian distribution; $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ and G are defined in Section 3.1.*

Theorem 3 explicitly characterizes the effect of increasing model dimensions. Since $(n\lambda)/(s\log p) \gg (np)^{1/2}$, part 1 of Theorem 3 shows that the posterior distribution for all inactive coefficients concentrates toward zero at a second-order rate, even if the number of such coefficients diverges. For part 2 of Theorem 3, it is sometimes more informative to consider a one-dimensional linear combination of parameters $\boldsymbol{\alpha}^T\boldsymbol{\beta}$ for $\|\boldsymbol{\alpha}\| = 1$ in the regime of increasing dimension (Fan et al., 2004).

If $\boldsymbol{\alpha}_{\mathcal{S}} \neq \mathbf{0}$, then the posterior for $\boldsymbol{\alpha}^T\boldsymbol{\beta}$ is asymptotically "oracle"; otherwise, the scale of the posterior is at the order of $p^{1/2}/(n\lambda) \ll n^{-1/2}$, which is super-efficient.

The range (4.3) for tuning parameter depends explicitly on design and spasity conditions. As one example, we consider a sparse model where $s_n = s_0$ stays fixed and $p_n \to \infty$. In addition to Assumptions 1, 2′, 3, and 4′, we suppose the design matrix satisfies $\theta_{\min}(D) \geq \theta_0 > 0$, which aligns with the setting in Belloni et al. (2019). Under this model setting, the conclusions in Theorem 3 hold if

$$p^2 \log^2 n = o(n), \quad \underline{b}_n \gg \frac{p \log^{3/2} p}{n^{1/2}},$$
$$\text{and} \qquad \frac{p \log^{3/2} p}{n^{1/2}} \ll \lambda_n \ll \underline{b}_n,$$

where $\underline{b}_n$ is defined in Assumption 4′. With a sparse model, the above conditions are comparable with those in the literature on Gaussian mean regression (Fan et al., 2004; Huang et al., 2008; Armagan et al., 2013).

4.2 Practical posterior inference in higher dimensions

For problems with even higher dimensions, where the number of covariates may exceed the number of observations, Bayesian inference for quantile regression is much less understood. Furthermore, the variance adjustment in Section 3.3 is not applicable when $n < p_n$, since it relies on estimating the full covariance matrix $\mathrm{E}^*[XX^T]$. Therefore, a direct application of the pseudo-Bayesian approach becomes problematic.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Nonetheless, the pseudo-Bayesian approach becomes useful when combined with the idea of marginal screening (Fan and Lv, 2008). For high-dimensional sparse problems with $s_n \ll n < p_n$, 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, these 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 3 then applies to the d_n-dimensional posterior distribution post-screening.

5. Simulation

We use a set of Monte Carlo simulation to demonstrate that the asymptotic properties established in this paper are present in finite-sample problems. We include a limited comparison with some other inferential methods for quantile regression. We highlight several key findings here; the implementations and more detailed results are relegated to the online Supplementary Material. The Supplementary Material also contain

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

more discussions on variable selection approaches and the use of other priors, as well as an additional simulation setting.

We generate random samples of size n from the following regression model:

$$Y = 1 + 3X_2 - 5X_5 + \left\{ \frac{1 + (X_6 - 1)^2}{3} \right\} e,$$

where $e \sim \mathrm{N}(0, 1)$ is independent of the covariate vector $X = (X_1, \ldots, X_6)^T \sim \mathrm{N}(0, \Sigma)$, with the (i, j)th entry of Σ being $0.8^{|i-j|}$, for $i, j \in \{1, \cdots, 6\}$. The data-generating process satisfies Model (2.1) at $\tau = 0.5$, where X_2 and X_5 are active, but X_6 is inactive for the conditional median of Y given X. We consider two sample sizes, $n = 200$ and $n = 500$, and use $2,000$ Monte Carlo data sets in each simulation.

For the proposed pseudo-Bayesian approach, we use the AL prior (2.4) in the simulation study because of its computational attractiveness. We compare the proposed approach with four other approaches for constructing 90% confidence intervals of the median regression coefficients. The first three approaches are the robust rank-score method of Koenker and Machado (1999) applied to: (i) the full model with $(X_1, \ldots, X_6)$ included, (ii) the oracle model with (X_2, X_5) included, and (iii) the selected model from adaptive lasso variable selection, respectively. The fourth approach is the wild bootstrap for the adaptive lasso quantile regression proposed recently by Wang et al. (2018). Because an exhaustive comparison with all other methods is infeasible, we focus on those four approaches that are known to exhibit good performance under heteroscedastic models.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

We first compare the performance of the approaches under a fixed tuning parameter in Table 1. To ensure a fair comparison, we keep the tuning parameter value λ the same across all Monte Carlo data sets at a given sample size, for both the pseudo-Bayesian approach and the adaptive lasso model selection procedure. We relegate further implementation details, including the tuning parameter specification, to Section S1.1 of the online Supplementary Material.

Table 1 suggests that the adjusted posterior inference indeed achieves adaptive performance. For the active coefficients, the adjusted posterior inference gives much shorter intervals than those from the full model, and the results are reasonably competitive with those from the oracle model. For the inactive coefficients, the adjusted posterior inference gives much shorter intervals than those under the full model with higher-than-nominal coverage probabilities. On the other hand, the wild bootstrap approach and the rank-score method after variable selection both fall short in terms of coverage. These approaches are all asymptotically equivalent if oracle model selection is achieved. However, we find that common variable selection approaches, including the adaptive lasso, do not achieve oracle selection often enough in this case with limited sample sizes; we refer to Sections S1.2 and S1.3 of the online Supplementary Material for detailed results. Therefore, approaches based on variable selection may not be consistently reliable for inference (Leeb and Pötscher, 2005; Wang et al., 2020).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Table 1: Empirical coverage probabilities and average lengths ($\times 100$) for 90% confidence intervals.

	Empirical coverage			Average length (s.e.)		
			$n = 200$			
	β_2	β_5	β_{zeros}	β_2	β_5	β_{zeros}
Full	92	91	90	43.7 (0.25)	43.9 (0.25)	41.7 (0.10)
Oracle	89	93	100	22.5 (0.13)	32.6 (0.21)	0.0 (0.00)
Refit	84	86	89	28.1 (0.21)	34.7 (0.22)	11.4 (0.06)
WildPen	85	84	89	27.0 (0.13)	30.2 (0.14)	20.1 (0.08)
BayesAdj	93	93	96	28.1 (0.11)	32.7 (0.12)	11.7 (0.04)
			$n = 500$			
	β_2	β_5	β_{zeros}	β_2	β_5	β_{zeros}
Full	91	91	90	26.7 (0.12)	26.6 (0.12)	25.6 (0.06)
Oracle	90	93	100	14.0 (0.06)	20.5 (0.11)	0.0 (0.00)
Refit	81	86	89	17.2 (0.11)	21.5 (0.11)	6.3 (0.05)
WildPen	84	85	91	16.7 (0.07)	18.9 (0.07)	11.3 (0.06)
BayesAdj	89	91	95	16.3 (0.06)	19.3 (0.06)	6.1 (0.03)

"Full" refers to the rank-score method applied to all the covariates, "Oracle" uses only the active covariates for the conditional median, and "Refit" is the rank-score method applied to a model selected by the adaptive lasso. "WildPen" is the wild bootstrap approach of Wang et al. (2018). "BayesAdj" refers to the adjusted posterior inference in Section 3.3. For the "Refit" and "Oracle" methods, if a covariate is not included in the model, we report its confidence interval as a singleton $\{0\}$. The column β_{zeros} averages over all inactive coefficients β_1, β_3, β_4, and β_6. The numbers shown in parentheses are the estimated standard errors. For the coverage estimates, their standard errors are all below 0.9. For penalization/shrinkage, we used $\lambda = 0.066$ when $n = 200$, and $\lambda = 0.051$ when $n = 500$.

In addition, the adjusted posterior inference gives more stable confidence intervals, as the standard errors for interval lengths are among the smallest of all methods in Table 1. This finite-sample stability of our psuedo-Bayesian approach is because it does not require dichotomous variable selection; wee Figure S1 in the Supplementary

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Material for more details.

Next, we examine the effect of the tuning parameter in the comparisons of shrinkage-based methods. To this end, we vary λ through a wide range of values, and compare the performance in Figure 2 when the sample size $n = 500$; see Figure S3 in the Supplementary Material for the results when $n = 200$. For the active coefficients, the coverage probabilities of the pseudo-Bayesian approach are more stable around the nominal levels than the other methods for a wide range of λ values. For the inactive coefficient β_6, the coverage probability for the proposed method remains high, without any sacrifice in the lengths of the intervals relative to other non-oracle methods. More extensive empirical experiments show that the proposed approach tends to lose coverage if the shrinkage parameter λ is too large. As a practical guide, we suggest choosing λ to be slightly smaller than the value one would obtain from cross-validation for the adaptive lasso.

In the online Supplementary Material, we show the proposed pseudo-Bayesian approach continues to have desirable inferential performance with higher covariate dimensions, and at other quantile levels. Furthermore, we demonstrate that not all common shrinkage priors are directly suitable for the pseudo-Bayesian framework. Refer to Sections S1.4, S1.5, and S2 of the online Supplementary Material for detailed numerical results.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

6. Conclusion

In this paper, 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 that the posterior inference adapts 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 results reveal the value of Bayesian computational methods in frequentist inference even with a misspecified likelihood.

Under appropriate assumptions to ensure oracle model selection asymptotically, the adjusted posterior inference is first-order equivalent to the following two-step procedure: variable selection followed by quantile regression inference on the selected model. With the goal being inference rather than variable selection, the proposed pseudo-Bayesian approach enjoys two distinct advantages: (i) it avoids the need to pursue dichotomous variable selection, which is often non-oracle in finite-sample problems; and (ii) it avoids direct non-parametric estimation of the nuisance parameter needed for frequentist inference. Therefore, the proposed approach often leads to more stable inference for quantile regression. Additional numerical results in the online Supplementary Material further demonstrate the stability of our pseudo-Bayesian approach.

This paper focuses on problems with fixed or moderately increasing dimensions.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Even in fixed-dimensional problems, our asymptotic analysis for posterior inference is new, under a misspecified likelihood and shrinkage priors. An interesting avenue for future research would be to investigate what the pseudo-Bayesian approach can offer in even higher dimensions.

We use two relatively simple shrinkage priors to demonstrate the properties of the posterior inference. Nonetheless, it would be of interest to study the appropriate use of more sophisticated priors, and to identify priors that lead to optimal posterior inference for quantile regression.

Finally, we note 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 nonconvex (Powell, 1984, 1986). Our pseudo-Bayesian approach can be used to produce statistical inference without direct optimization of the objective function, while still incorporating possible model sparsity.

Supplementary Material

The online supplementary material contains some additional simulation results, as well as the proofs of all the results in this paper.

Acknowledgments

The work was partially supported by the NSF Awards DMS-1914496 and DMS-1951980.

References

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.

Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized double pareto shrinkage. *Statistica Sinica 23*(1), 119.

Barut, E., J. Fan, and A. Verhasselt (2016). Conditional sure independence screening. *Journal of the American Statistical Association 111*(515), 1266–1277.

Belloni, A., V. Chernozhukov, et al. (2011). $\ell 1$-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 (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(1), 4–29.

Bermingham, M. L., R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, 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.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Castillo, I., J. Schmidt-Hieber, A. Van der Vaart, et al. (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.

Chao, S.-K., S. Volgushev, G. Cheng, et al. (2017). Quantile processes for semi and nonparametric regression. *Electronic Journal of Statistics 11*(2), 3272–3331.

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.

Chernozhukov, V. and H. Hong (2003). An mcmc approach to classical estimation. *Journal of Econometrics 115*(2), 293–346.

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., H. Peng, et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The annals of statistics 32*(3), 928–961.

Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. *Biometrika 98*(4), 995–999.

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, H. H. Zhou, et al. (2020). A general framework for bayes structured linear models. *Annals of Statistics 48*(5), 2848–2878.

Gutenbrunner, C., J. Jurečková, R. Koenker, and S. Portnoy (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics 2*(4), 307–331.

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., L. Wang, H. G. Hong, et al. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *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.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica 18*(4), 1603–1618.

Jiang, B. and Q. Sun (2019). Bayesian high-dimensional linear regression with generic spike-and-slab priors. *arXiv preprint arXiv:1912.08993v2*.

Knight, K. (1998). Limiting distributions for l1 regression estimators under general conditions. *Annals of Statistics 26*(2), 755–770.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of Quantile*

Eegression. CRC press.

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. and Q. Zhao (1994). L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics 3*(3-4), 223–235.

Kohns, D. and T. Szendrei (2020). Horseshoe prior bayesian quantile regression. *arXiv preprint arXiv:2006.07655*.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Li, J. (2011). *The bayesian lasso, Bayesian SCAD and Bayesian group lasso with applications to genome-wide association studies*. Ph. D. thesis, Pennsylvania State University.

Li, Q., R. Xi, N. Lin, et al. (2010). Bayesian regularized quantile regression. *Bayesian Analysis 5*(3), 533–556.

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.

Narisetty, N. N., X. He, et al. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics 42*(2), 789–817.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics 5*(2), 99–135.

Newey, W. K. and J. L. Powell (1990). Efficient estimation of linear and type i censored regression models under conditional quantile restrictions. *Econometric Theory 6*(3), 295–317.

Pan, X. and W.-X. Zhou (2020). Multiplier bootstrap for quantile regression: non-

asymptotic theory under random design. *Information and Inference: A Journal of the IMA 10*(3), 813–861.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics 25*(3), 303–325.

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*. Cambridge University Press, Cambridge, UK.

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.

Song, Q. and F. Liang (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*.

Sriram, K. (2015). A sandwich likelihood correction for bayesian quantile regression based on the misspecified asymmetric laplace density. *Statistics & Probability Letters 107*, 18–26.

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), e1005357.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics 37*(1), 121–133.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 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.

Wu, T. and N. N. Narisetty (2021). Bayesian multiple quantile regression for linear models using a score likelihood. *Bayesian Analysis 16*(3), 875–903.

Wu, Y. and Y. Liu (2009). Variable selection in quantile regression. *Statistica Sinica 19*(2), 801–817.

Wu, Y. and G. Yin (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika 102*(1), 65–76.

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, Q. (2001). Asymptotically efficient median regression in the presence of heteroskedasticity of unknown form. *Econometric Theory 17*(4), 765–784.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Yuanzhi Li

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

Department of Statistics, University of Michigan, Ann Arbor, MI, USA

E-mail: yzli@umich.edu

Xuming He

Department of Statistics, University of Michigan, Ann Arbor, MI, USA

E-mail: xmhe@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0338

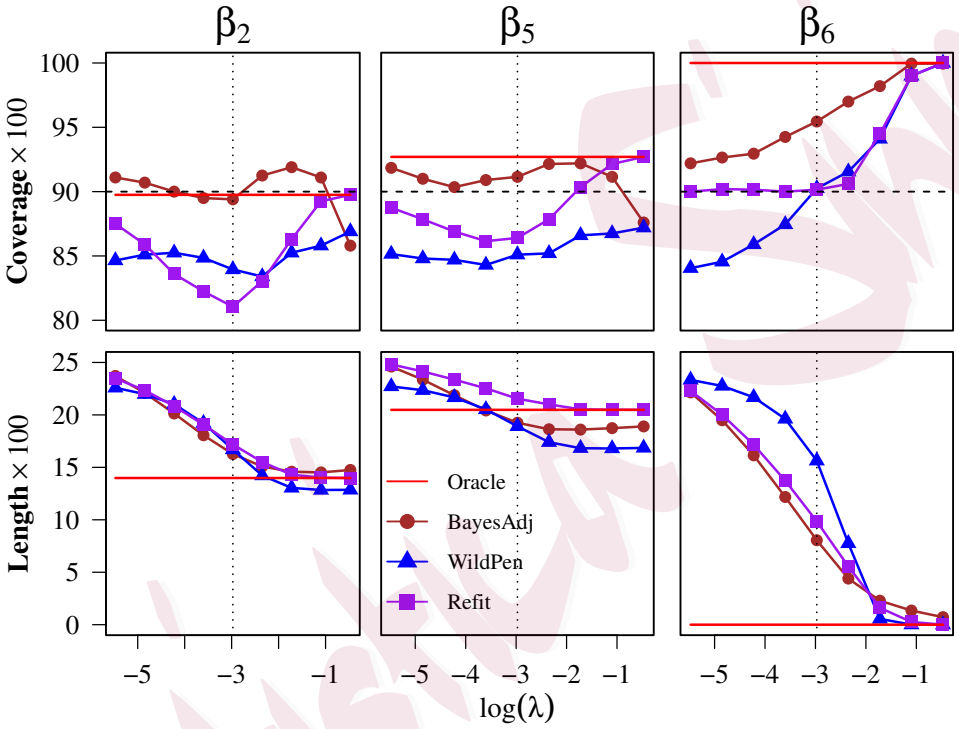

Figure 2: Empirical coverage probabilities and average lengths for 90% confidence intervals with different λ when $n = 500$. The true regression coefficients are $\beta_2^0 = 3$, $\beta_5^0 = -5$, and $\beta_6^0 = 0$. The value of λ marked by a vertical broken line is used to produce Table 1, and the abbreviated method names are the same as those in Table 1.