Composite versus model-averaged quantile regression

Bloznelis D, Claeskens G, Zhou J.

Composite versus model-averaged quantile regression

Daumantas Bloznelis[1,2], Gerda Claeskens[1] and Jing Zhou[1]

[1] ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

[2] Department of Business Administration, Inland Norway University of Applied Sciences, Telthusveien 12, 2450 Rena, Norway

Daumantas.Bloznelis@inn.no; Gerda.Claeskens@kuleuven.be; Jing.Zhou@kuleuven.be

Abstract

The composite quantile estimator is a robust and efficient alternative to the least-squares estimator in linear models. However, it is computationally demanding when the number of quantiles is large. We consider a model-averaged quantile estimator as a computationally cheaper alternative. We derive its asymptotic properties in high-dimensional linear models and compare its performance to the composite quantile estimator in both low- and high-dimensional settings. We also assess the effect on efficiency of using equal weights, theoretically optimal weights, and estimated optimal weights for combining the different quantiles. None of the estimators dominates in all settings under consideration, thus leaving room for both model-averaged and composite estimators, both with equal and estimated optimal weights in practice.

Keywords: Quantile regression; model averaging; composite estimation; penalized estimation; weight choice.

1 Introduction

For low-dimensional linear regression models, ordinary least squares (OLS) estimation is the common approach. Under standard assumptions, OLS provides the minimum-variance (a.k.a. best) estimator in the class of linear unbiased estimators. However, it may misbehave when the error distribution has heavy tails. This motivated the seminal Koenker and Bassett (1978) paper that introduced quantile regression as a robust alternative to OLS. Unsurprisingly, robustness does not come at zero cost; the quantile estimator is relatively less efficient than its OLS counterpart for certain light-tailed distributions such as the Gaussian. Efforts to find robust yet efficient estimators have persisted. Koenker (1984) considered weighted composite quantile regression (weighted CQR) and weighted model-averaged quantile regression (weighted MAQR) as more efficient alternatives to the regular single-quantile estimator. He showed that both estimators are more efficient than the single-quantile one, and that both achieve the same lower bound of asymptotic variance, given a suitable choice of weights that depend on the error distribution. The origins of MAQR can be found already in Koenker and Bassett (1978), while the idea of CQR was proposed by R.V. Hogg in 1979; see (Koenker, 1984, 2005). The literature has continued expanding on the composite quantile regression (see Zou and Yuan, 2008; Bradic et al., 2011; Jiang et al., 2012, 2014). Meanwhile, the model-averaged quantile regression has garnered little attention (with a recent exception of Zhao and Xiao, 2014), although it is computationally

cheaper than CQR, and the difference in the computational cost becomes prohibitive when the number of quantiles employed is larger than about ten.

For high-dimensional models, only the composite estimator has been considered (Bradic et al., 2011). We introduce its model-averaged counterpart, obtain optimal weights for the different quantiles under a given error distribution, and compare CQR and MAQR in terms of asymptotic relative efficiency and finite-sample performance. We also draw attention to the fact that when the error distribution is unknown, theoretically optimal weights are unavailable. They need to be estimated from the data and thus become random variables. Therefore, optimality results for plug-in versions of the theoretically optimal weights may change. This is similar in spirit to the forecast combination puzzle (e.g. Claeskens et al., 2016, and references therein) where estimated optimal weights in forecast combinations may yield poorer results than equal weights. Moreover, the asymptotic distributions of the CQR and MAQR estimators under estimated optimal weights are less straightforward to obtain than under fixed weights. We examine in simulations whether estimated optimal weights or equal weights perform better in practice.

In Section 2 we first review the composite and model-averaged linear quantile estimators in low-dimensional regression models. We contribute with a theoretical comparison of equal weights and optimal weights for both types of estimators. Section 3 proceeds with composite and model-averaged estimation in high-dimensional regression models under a sparsity assumption. We obtain the limiting asymptotic distribution of a high-dimensional model-averaged quantile estimator and use the distribution to propose a vector of optimal weights. A simulation study in Section 4 and a data example in Section 5 show the estimators' performance in practice.

2 Low-dimensional linear quantile regression

Consider a linear model $Y = X\beta + \varepsilon$, with a $n \times 1$-vector Y of dependent variables and a $n \times p$ matrix X of fixed regressors. For the errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ we denote the cumulative distribution of ε_j by F, and its density function by f. We assume the errors are independent and identically distributed.

Given a single quantile level $0 < \tau < 1$, an estimator of the $100\tau\%$ quantile of the response Y in the linear model is defined as

$$(\hat{b}_\tau, \hat{\beta}_\tau^\top) = \arg \min_{b_\tau, \beta} \sum_{i=1}^n \rho_\tau(Y_i - b_\tau - X_i^\top \beta), \tag{1}$$

where $\hat{b}_\tau$ is the estimator of the quantile intercept b_τ, $\hat{\beta}_\tau$ is the estimator of the slopes β, and $\rho_\tau(z) = \tau I(z \geq 0)z + (\tau - 1)I(z < 0)z$. The true $100\tau\%$ quantile of Y given X_i is $X_i^\top \beta + b_\tau$, where $b_\tau = F^{-1}(\tau)$ is the $100\tau\%$ quantile of the distribution of the error term ε. Hence, the only regression parameter that depends on the quantile level is the quantile intercept b_τ, while the true slopes β are the same for all quantiles.

2

2.1 Composite and model-averaged estimators

For multiple quantile levels $0 < \tau_1 < \ldots < \tau_k < 1$, the equally-weighted CQR estimator is defined as

$$(\hat{b}_{\tau_1,\text{comp}}, \ldots, \hat{b}_{\tau_k,\text{comp}}, \hat{\beta}_{\text{comp}}^\top)(1_k/k) = \arg\min_{b_{\tau_1}, \ldots, b_{\tau_k}, \beta} \sum_{l=1}^{k} \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - b_{\tau_l} - X_i^\top \beta), \qquad (2)$$

the vector $(\hat{b}_{\tau_1,\text{comp}}, \ldots, \hat{b}_{\tau_k,\text{comp}}^\top, \hat{\beta}_{\text{comp}})$ depending on the weights $1_k/k$, where 1_k is a vector of length k consisting of ones. In (2), the estimating functions from the single quantile regression models as in equation (1) are simply summed, or equivalently, all given the same weight. A more general result, see Koenker (1984) and Koenker (2005, Sec. 5.5), is to allow for different weights $\nu = (\nu_1, \ldots, \nu_k)$, resulting in the weighted CQR estimator

$$(\hat{b}_{\tau_1,\text{comp}}, \ldots, \hat{b}_{\tau_k,\text{comp}}, \hat{\beta}_{\text{comp}}^\top)(\nu) = \arg\min_{b_{\tau_1}, \ldots, b_{\tau_k}, \beta} \sum_{l=1}^{k} \nu_l \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - b_{\tau_l} - X_i^\top \beta).$$

This approach of getting *composite* estimators by linearly combining the estimating functions has become quite popular, especially in quantile regression. Zou and Yuan (2008) consider equally-weighted penalized CQR in a high-dimensional setting as an alternative to penalized least squares estimation. They develop an oracle estimator that is at worst 30% less efficient than the penalized least squares estimator but in other cases can be arbitrarily more efficient. For example, it works well when the variance of the error distribution is infinite, where the penalized least squares estimator fails. Bradic et al. (2011) consider a weighted penalized CQR estimator and its oracle properties when the error distribution is unknown. They build on an idea that the loss function of the CQR with data-driven adaptive weights can approximate the true likelihood of the error distribution well and as such can lead to efficient estimation. Jiang et al. (2012) extend the research on robust yet efficient estimation and model selection in high dimensions to nonlinear models. They use weighted CQR and achieve consistent model selection together with estimation efficiency that is near that of the maximum likelihood estimator. Jiang et al. (2014) consider weighted CQR estimator for autoregressive conditionally heteroskedastic models and demonstrate its robustness and efficiency.

Model averaging is an alternative to composite estimation. Focusing only on the slopes β, one may obtain different quantile estimators $\hat{\beta}_{\tau_l}$, $l = 1, \ldots, k$, from equation (1) and take their weighted average, to arrive at a weighted MAQR estimator

$$\hat{\beta}_{\text{mod.avg}}(\omega) = \sum_{l=1}^{k} \omega_l \hat{\beta}_{\tau_l},$$

where $\omega = (\omega_1, \ldots, \omega_k)^\top$ is a vector of weights. In model averaging, one usually restricts the weights to sum to one, $\sum_{l=1}^{k} \omega_l = 1$. Another restriction that one might or might not impose is that the weights lie within $[0, 1]$. This estimator has recently been considered by Zhao and Xiao (2014) in a low-dimensional setting.

The main question we wish to investigate is, which approach is preferred: (1) separate estimation using different quantile levels τ_l, $l = 1, \ldots, k$, which is a simple procedure, followed by a weighted average of the estimators to arrive at $\hat{\beta}_{\text{mod.avg}}$, or (2) a single, though more complicated, estimation with a weighted loss function that immediately results in an estimator $\hat{\beta}_{\text{comp}}$?

2.2 Asymptotic relative efficiency

For the low-dimensional case, part of the answer to this question has been given by Koenker (1984), see also Koenker (2005, Th. 5.2). Under the assumption that $\sum_{l=1}^{k} \nu_l = 1$ to guarantee consistency, and some additional assumptions to ensure the distribution of errors and the regressors are well behaved, he obtains that for $n \to \infty$ there is a limiting mean-zero normal distribution for $\sqrt{n}(\hat{\beta}_{\text{comp}}(\nu) - \beta)$ with asymptotic covariance matrix

$$Q^{-1} \frac{\sum_{l,l'=1}^{k} \nu_l \nu_{l'} \min(\tau_l, \tau_{l'})\{1 - \max(\tau_l, \tau_{l'})\}}{\{\sum_{l,l'=1}^{k} \nu_l \nu_{l'} f(b_{\tau_l}) f(b_{\tau_{l'}})\}^2} = Q^{-1} \nu^\top A \nu / (\nu^\top \boldsymbol{f})^2,$$

where f is the density function corresponding to F, $Q = \lim_{n \to \infty} \{\frac{1}{n} X^\top X\}$, $\boldsymbol{f} = (f(b_{\tau_1}), \ldots, f(b_{\tau_k}))^\top$, $\nu = (\nu_1, \ldots, \nu_k)^\top$ and the $k \times k$ matrix A is formed by the entries $a_{ll'} = \min(\tau_l, \tau_{l'})\{1 - \max(\tau_l, \tau_{l'})\}$ for $l, l' = 1, \ldots, k$.

In order to study the asymptotic distribution of the model-averaged estimator we need the joint limiting distribution of the estimators $\hat{\beta}_{\tau_l}$ for $l = 1, \ldots, k$. Koenker and Bassett (1978) obtain the limiting normal distribution of a vector of quantile regression estimators

$$\sqrt{n}\{(\hat{\beta}_{\tau_1}^\top, \ldots, \hat{\beta}_{\tau_k}^\top) - (\beta^\top, \ldots, \beta^\top)\}.$$

This limiting distribution has mean zero and a covariance matrix $\Omega \otimes Q^{-1}$ where $\otimes$ denotes the Kronecker product and where the $k \times k$ matrix Ω has l, l' entry equal to $a_{ll'}/\{f(b_{\tau_l}) f(b_{\tau_{l'}})\}$. This immediately leads to the asymptotic normality of $\hat{\beta}_{\text{mod.avg}}$. Denote the vector of weights $\omega = (\omega_1, \ldots, \omega_k)^\top$. It follows that, when $n \to \infty$, with $\tilde{\beta} = (\sum_{l=1}^{k} \omega_l)\beta$,

$$\sqrt{n}(\hat{\beta}_{\text{mod.avg}}(\omega) - \tilde{\beta}) \xrightarrow{d} N(0, \omega^\top \Omega \omega Q^{-1}). \tag{3}$$

When the weights sum to 1, $\tilde{\beta} = \beta$ and both the CQR and the MAQR estimators are asymptotically unbiased. A comparison of the asymptotic mean squared error (MSE) values of $\hat{\beta}_{\text{mod.avg}}$ and $\hat{\beta}_{\text{comp}}$ boils down to comparing the asymptotic variances. The asymptotic relative efficiency (ARE) of the model-averaged and the composite estimators is

$$\text{ARE}\{\hat{\beta}_{\text{mod.avg}}(\omega), \hat{\beta}_{\text{comp}}(\nu)\} = \frac{\text{asyVar}\,\hat{\beta}_{\text{mod.avg}}(\omega)}{\text{asyVar}\,\hat{\beta}_{\text{comp}}(\nu)} = \sum_{l,l'=1}^{k} \omega_l \omega_{l'} \frac{a_{l,l'}}{f(b_{\tau_l}) f(b_{\tau_{l'}})} \frac{\{\sum_{j=1}^{k} \nu_j f(b_{\tau_j})\}^2}{\sum_{j,j'=1}^{k} \nu_j \nu_{j'} a_{j,j'}}.$$

Using inequalities related to eigenvalues of the matrix A, Koenker (2005, Th. 5.2, Cor. 5.1) explains that there exists a choice of the weight vectors ω and ν such that both estimators,

$\hat{\beta}_{\mathrm{mod.avg}}(\omega)$ and $\hat{\beta}_{\mathrm{comp}}(\nu)$, achieve the same lower bound for the asymptotic variance, asyVar $=$ $(\boldsymbol{f}^{\top}A^{-1}\boldsymbol{f})^{-1}$. That is, both estimators can achieve the same efficiency and with optimal weights the ARE of the two estimators is 1. The optimal choices of the weights ω and ν follow expressions

$$\omega_{\mathrm{opt}} = (\boldsymbol{f}^{\top}A^{-1}\boldsymbol{f})^{-1}\mathrm{diag}(\boldsymbol{f})A^{-1}\boldsymbol{f} \text{ and } \nu_{\mathrm{opt}} = A^{-1}\boldsymbol{f}.$$

Consequently, one may conclude that both approaches are worth pursuing, each with its own optimal choice of weights depending on the choice of the quantile levels τ_l, $l = 1, \ldots, k$ and on the true error distribution f.

Note that some components of the optimal weight vectors ω_{opt} and ν_{opt} may be negative. From the perspective of estimation algorithms, this causes no difficulty for the MAQR estimator as the weighting is done after having estimated the individual regressions for each quantile. However, it is a bigger problem in the composite regression setting. There, negative weights lead to nonconvexity of the objective function and thus conventional convex optimization algorithms cannot be applied. In practice this may be prohibitive and may effectively prevent the use of the CQR estimator when some of the weights are negative.

2.3   Weights: theoretically optimal versus estimated optimal versus equal

The optimal weights ω_{opt} and ν_{opt} are often not computable due to an incompletely specified density function f, which may be either entirely unknown in a nonparametric setting, or partly unknown in a parametric setting. When estimators replace unknown quantities in the computation of optimal weights, the resulting estimated weights $\hat{\omega}$ and $\hat{\nu}$ are obviously random. While ω_{opt} and ν_{opt} minimize the asymptotic variance of, respectively, $\hat{\beta}_{\mathrm{mod.avg}}(\omega)$ and $\hat{\beta}_{\mathrm{comp}}(\nu)$, no such guarantee can be given for their estimated counterparts $\hat{\omega}$ and $\hat{\nu}$. In fact, it might well be the case that an equally-weighted estimator yields a lower mean squared error than its counterpart with estimated optimal weights. This phenomenon is known as the 'forecast combination puzzle' (e.g., Smith and Wallis, 2009). Claeskens et al. (2016) worked out first and second moments of the forecast combination with estimated weights and showed that such a phenomenon may take place when estimation uncertainty is neglected while deriving the optimal weights. Whereas explicit formulas of moments are harder to obtain for the quantile estimators, it is immediately clear that the same problem may occur. Indeed, for the model-averaged estimator with estimated optimal weights,

$$E[\hat{\beta}_{\mathrm{mod.avg}}(\hat{\omega})] = \sum_{l=1}^{k} E[\hat{\omega}_l \hat{\beta}_{\tau_l}],$$

$$\mathrm{Var}[\hat{\beta}_{\mathrm{mod.avg}}(\hat{\omega})] = \sum_{l=1}^{k} \mathrm{Var}[\hat{\omega}_l \hat{\beta}_{\tau_l}] + 2\sum_{l=1}^{k}\sum_{l'=1,\ l'<l}^{k} \mathrm{Cov}[\hat{\omega}_l \hat{\beta}_{\tau_l}, \hat{\omega}_{l'} \hat{\beta}_{\tau_{l'}}].$$

Both quantities depend on the joint distribution of the weight vector $\hat{\omega}$ and the vector of quantile estimators $(\hat{\beta}_{\tau_1}, \ldots, \hat{\beta}_{\tau_k})$. A similar argument holds for the composite quantile estimator with estimated weighs $\hat{\nu}$. Since ω_{opt} and ν_{opt} are the theoretical minimizers of the fixed-weight

asymptotic variances, using estimated weights $\hat{\omega}$ and $\hat{\nu}$ of course results in values of the asymptotic variance that are at least as large as the minimal variance. Moreover, the asymptotic variance of the estimator with estimated weights may exceed its counterpart with equal weights. Hence, when the joint distribution of the estimated weights and the estimated quantile slopes is not available, one might as well resort to employing the simpler equal weights. A simulated comparison of equally-weighted versus optimally-weighted MAQR estimator is offered in Section 4.

To avoid the estimation of optimal weights when the error distribution is unknown, we might use the equally-weighted MAQR estimator where $\omega_l = 1/k$ for $l \in \{1, \ldots, k\}$; or the equally-weighted CQR estimator (2) of Zou and Yuan (2008). First, we find the choice of the weights ω of the model-averaged estimator that achieves the same ARE as the equally-weighted composite estimator with $\nu = 1_k/k$. The equally-weighted composite estimator gains precisely the same asymptotic variance as a weighted model-averaged estimator with weights proportional to the density $f(b_{\tau_l})$, denoted $\omega_l^{[1]}$, that is, $\omega_l^{[1]} = f(b_{\tau_l})/\{\sum_{j=1}^{k} f(b_{\tau_j})\}$. Thus more weight is assigned to the quantile estimator $\hat{b}_{\tau_l}$ for which the density $f(b_{\tau_l})$ is larger. Indeed, it can be verified that

$$\mathrm{ARE}\{\hat{\beta}_{\mathrm{mod.avg}}(\omega^{[1]}), \hat{\beta}_{\mathrm{comp}}(1_k/k)\} = 1.$$

Meanwhile, taking equal weights for the model-averaged estimator corresponds to the same asymptotic variance as when using the perhaps less intuitive weights $\nu_l^{[1]} = 1/f(b_{\tau_l})$ for the composite estimator, since it is readily verified that

$$\mathrm{ARE}\{\hat{\beta}_{\mathrm{comp}}(\nu^{[1]}), \hat{\beta}_{\mathrm{mod.avg}}(1_k/k)\} = 1.$$

Here, for the composite estimator to get the same efficiency as the equally-weighted model-averaged estimator, one should weight inversely proportional to the density, thus giving higher weights to the low-density areas.

Replacing optimal weights by equal weights will generally lead to a less efficient estimator, and the effect will vary depending on the error distribution and the number of quantiles under consideration. Figure 1 contains the asymptotic relative efficiency of the equally-weighted MAQR estimator and the equally-weighted CQR estimator to their respective optimally-weighted counterparts. Hence, the vertical axis measures the values of the asymptotic relative efficiencies $\mathrm{ARE}(\{\hat{\beta}_{\mathrm{mod.avg}}(1_k/k), \hat{\beta}_{\mathrm{mod.avg}}(\omega_{\mathrm{opt}})\}$ and $\mathrm{ARE}(\{\hat{\beta}_{\mathrm{comp}}(1_k/k), \hat{\beta}_{\mathrm{comp}}(\nu_{\mathrm{opt}})\}$. Different panels correspond to different error distributions, and a range of equally-spaced quantiles $k = 1, \ldots, 20$ is used on the horizontal axis. Note that the optimal asymptotic variance is the same for both the composite and the model-averaged cases.

Comparing the equally-weighted MAQR estimator to its optimally-weighted counterpart, we find that the loss in efficiency generally grows with the number of quantiles k, but the growth rate differs considerably across the different distributions. The loss is negligible for the normal and the logistic distribution, e.g. at $k = 15$ the variance ratio equals 1.001 for the normal distribution and 1.037 for the logistic distribution. Meanwhile, for distributions with heavier tails, the losses in efficiency are larger, e.g. the variance ratio is 6.017 for the $t(1)$ distribution and 13.461 for the exponential distribution at $k = 15$.
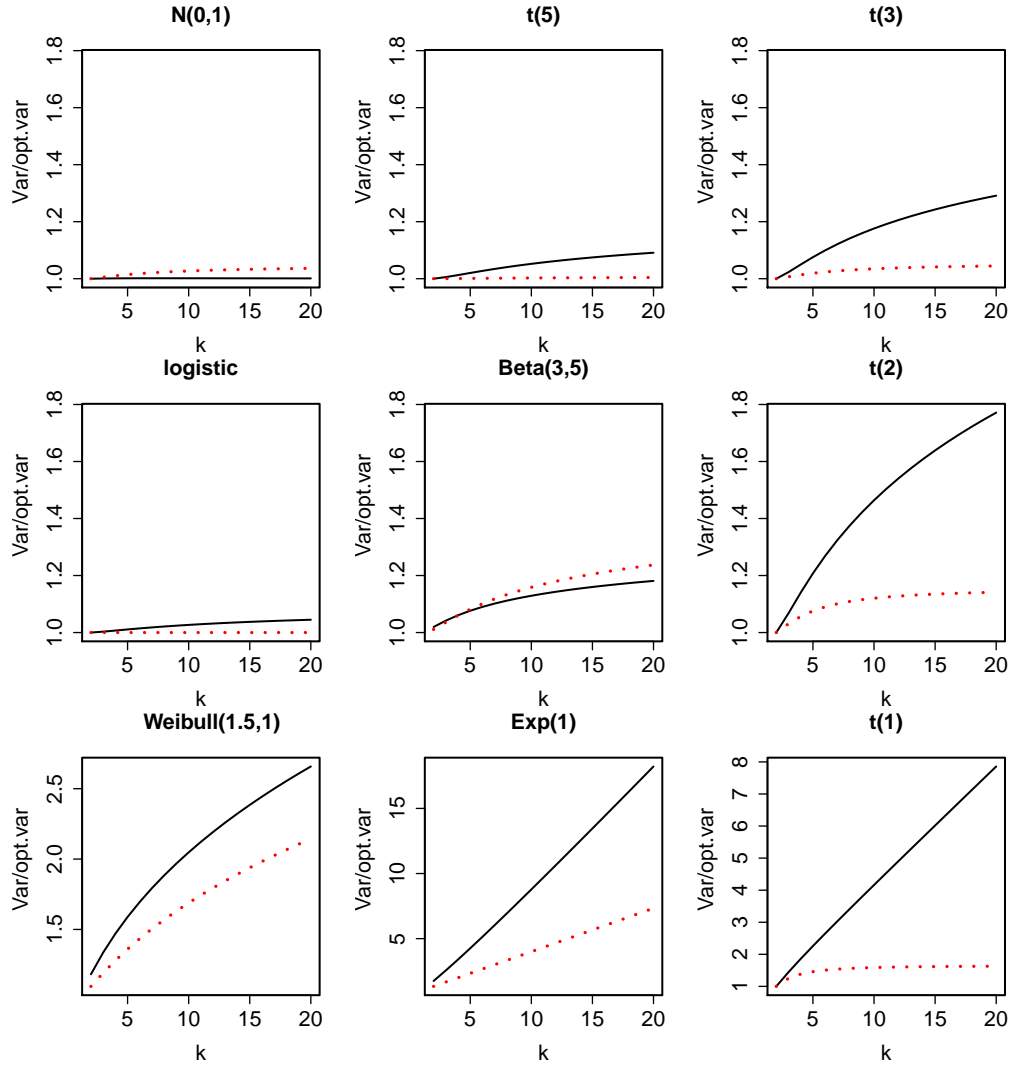
Figure 1: Asymptotic variance of the equally-weighted model-averaged estimator (solid line) and the equally-weighted composite estimator (dotted line) over the optimal variance for various distributions and different numbers k of equally-spaced quantiles.

The loss in efficiency of the equally-weighted CQR estimator in relation to its optimally-weighted counterpart grows with the number of quantiles as well. For the light-tailed normal distribution the loss is small, e.g. some 3%. This is more than for the MAQR estimator but not by much. The estimator is fully efficient for the logistic distribution. Unlike its model-averaged counterpart, the equally-weighted composite estimator is quite efficient for the heavy-tailed distributions; the variance ratio is only 1.618 for $t(1)$ at $k = 15$.

For the skewed distributions in the example, exponential and Weibull, there is a significant loss in efficiency for both equally-weighted methods relative to the optimally-weighted cases.

In general, algebraic calculations reveal the following relationship between the equally-

7

weighted composite and model-averaged estimators:

$$\text{ARE}\{\hat{\beta}_{\text{mod.avg}}(1_k/k), \hat{\beta}_{\text{comp}}(1_k/k)\} < 1 \Leftrightarrow \left(\frac{\overline{f}}{f}\right)^{\top} \cdot A \cdot \frac{\overline{f}}{f} < 1_k^{\top} \cdot A \cdot 1_k, \tag{4}$$

where $\overline{f}/f = (\overline{f}/f(b_{\tau_1}), \ldots, \overline{f}/f(b_{\tau_k}))^{\top}$ and $\overline{f} = \sum_{l=1}^{k} f(b_{\tau_l})/k$. If this condition holds, the equally-weighted model-averaged estimator is more efficient than its equally-weighted composite counterpart. Condition (4) can be verified for different error distributions. For mean-zero normal distributions, regardless of the variance, and for Student t-distributions with a large-enough degrees of freedom, the condition is satisfied, implying that model-averaged estimators with equal weights are more efficient than equally-weighted composite estimators. For example, for the $t(10)$ distribution when $k \leq 4$, model-averaged estimation is better than composite estimation when both methods use equal weights. For t distributions with 5, 3 and 1 degree of freedom, model-averaged estimation is equally efficient as composite estimation only for $k = 2$, while for values of $k \geq 3$ the composite estimation method is better. For the skewed distributions (exponential and Weibull), the condition fails under all values $k \geq 1$, implying that the equally-weighted composite estimator is the better choice.

3 Weighted quantile estimators in high dimensions

Consider now a sparse *high-dimensional* linear model as in Bradic et al. (2011),

$$Y = X\beta + \varepsilon \tag{5}$$

with independent and identically distributed mean-zero errors ε and with p, the number of columns of X large relative to the sample size n, allowing for an exponential order such that $\log(p) = O(n^{\delta})$ with $\delta \in (0, 1)$. The number of nonzero components of β, or sparsity, is assumed to be $s = O(n^{\alpha_0})$ with $\alpha_0 \in (0, 1)$. Under such circumstances, penalized estimation is employed.

3.1 Penalized composite quantile estimator

Bradic et al. (2011) consider a penalized composite quantile estimator

$$(\hat{b}_{\tau_1,\text{comp,pen}}, \ldots, \hat{b}_{\tau_k,\text{comp,pen}}, \hat{\beta}^{\top}_{\text{comp,pen}})(\nu) \tag{6}$$

$$= \arg\min_{b_{\tau_1}, \ldots, b_{\tau_k}, \beta} \left\{ \sum_{l=1}^{k} \sum_{i=1}^{n} \rho_l(Y_i - b_{\tau_l} - X_i^{\top}\beta) + n \sum_{j=1}^{p} \gamma_{\lambda}(|\beta_j^{(0)}|)|\beta_j| \right\}$$

where $\rho_l(z) = \tau_l I(z \geq 0)z + (\tau_l - 1)I(z < 0)z$. For the penalty term, $\beta_j^{(0)}$ is an initial slope estimator and γ_{λ} is some function, e.g. the derivative of some penalty function, allowing for (adaptive) lasso (Tibshirani, 1996; Zou, 2006) where $\gamma_{\lambda}(u) = \lambda|u|^{-a}$ for some constant $a \geq 0$ and SCAD (Fan and Li, 2001) where $\gamma_{\lambda}(u) = \lambda[I(u \leq \lambda) + \max(a\lambda - x, 0)I(x > \lambda)/\{(a - 1)\lambda\}]$ for $a > 0$.

A related work of Jiang et al. (2012) addresses estimation of a high-dimensional nonlinear regression model of the form $Y = h(X; \beta) + \varepsilon$ with a known function h under the assumption that $p = p_n$ is such that $p^3/n \to 0$ for $n \to \infty$. They consider penalized estimation using lasso and SCAD. From Jiang et al. (2012), the weight vector that minimizes the asymptotic variance of the weighted composite quantile estimator is given by, using the same notation as in Section 2.2,

$$\nu = (\boldsymbol{f}^\top A^{-2} \boldsymbol{f})^{-1/2} A^{-1} \boldsymbol{f}.$$

These weights may be negative, as the authors explicitly mention.

In Bradic et al. (2011), the optimal value of the weights is given by $\nu = A^{-1} \boldsymbol{f}$ to achieve the lower bound for the variance, $(\boldsymbol{f}^\top A^{-1} \boldsymbol{f})^{-1}$. While such weights may be negative and thus lead to a nonconvex objective function that is hard to optimize, an alternative weight vector ν^+ is obtained by minimizing $\nu^\top A \nu$ subject to having all weights nonnegative and $\boldsymbol{f}^\top \nu = 1$. There is no explicit expression for the nonnegative optimal weights ν^+. The authors show by simulations that both types of optimal weights outperform the equally-weighted estimator.

Noteworthy, Bradic et al. (2011) comment upon the computational complexity of the composite quantile estimation method with a large number of quantiles, but report that usually $k \leq 10$ suffices. Jiang et al. (2012) also suggest that $k = 10$ is large enough to get close-to-optimal efficiency.

3.2 Model-averaged penalized quantile estimator

To our knowledge, a model-averaged quantile regression estimator has not yet been investigated in the high-dimensional setting. We define the estimator as follows,

$$\hat{\beta}_{\text{mod.avg,pen}}(\omega) = \sum_{l=1}^{k} \omega_l \hat{\beta}_{\tau_l,\text{pen}},$$

where

$$(\hat{b}_{\tau_l}, \hat{\beta}_{\tau_l,\text{pen}}^\top) = \arg\min_{b_{\tau_l},\beta} \left\{ \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - b_{\tau_l} - X_i^\top \beta) + n \sum_{j=1}^{p} \gamma_{\lambda_l}(|\beta_j^{(0)}|)|\beta_j| \right\}. \tag{7}$$

Note that different penalization constants can be used for the separate quantile estimators, allowing for high flexibility. A major advantage of using the model-averaged penalized quantile estimator is that optimization is carried out for a single quantile at a time, which makes the estimator simple and fast to compute.

We now derive, under the same assumptions as in Bradic et al. (2011) the asymptotic distribution of the penalized model-averaged quantile estimator. We divide the design matrix X into two parts, $X = (X_a, X_b)$ where the columns of X_a are the columns of X for which the corresponding components of the coefficient vector β are nonzero. Hence, X_a is the 'active' part of the design matrix, with accompanying vector β_a. Likewise, X_b is the non-active part, concomitant to β_b, the latter vector consisting of zero components only. Due to the sparsity

assumption, the dimension of X_a is $n \times s$. When performing model averaging, the estimators $\hat{\beta}_{\tau_l,\text{pen}}$, $l \in \{1, \ldots, k\}$ may contain different components that are estimated nonzero for different quantiles τ_l. Under such a scenario, it is not possible to average the estimated active components of β. Therefore, instead of the estimators $\hat{\beta}_{\tau_l,\text{pen}}$ we consider predictions of a linear combination $u^\top \beta$ constructed with each of the estimated vectors $\hat{\beta}_{\tau_l,\text{pen}}$ for the different values of l, where u is a known vector.

Given the assumptions of Bradic et al. (2011, Th. 1), since a single estimation is a special case of composite estimation, we obtain (i) the existence, (ii) model selection consistency, and (iii) sign consistency of the estimators $\hat{\beta}_{\tau_l,\text{pen}}$ for $l \in \{1, \ldots, k\}$.

Likewise, with some adaptations to the prediction setting, we arrive at the asymptotic normality of the model-averaged predictions under the assumptions of Bradic et al. (2011, Th. 2), including their assumption of perfect asymptotic model selection where only the active variables are estimated as nonzero. Care is required since the matrix corresponding to the active set, $Q_a = \lim_{n \to \infty} \frac{1}{n} X_a^\top X_a$, has growing dimension when $n \to \infty$. As the dimension of β also grows with the sample size n, a correct limiting statement for the distribution of the estimators can be obtained by considering the limiting distribution of their linear combination. For this purpose, let U_a be any design matrix of dimension $r \times s$, containing the information about which $r \geq 1$ predictions we wish to make.

Proposition 1. *Under the above-mentioned assumptions and the assumptions of Theorem 2 of Bradic et al. (2011), for the model-averaged penalized quantile predictions it holds that*

$$\sqrt{n}(\frac{1}{n}U_a X_a^\top X_a U_a^\top)^{-1/2}\{\omega_1 U_a(\hat{\beta}_{a,\tau_1,\text{pen}} - \beta_a) + \ldots + \omega_k U_a(\hat{\beta}_{a,\tau_k,\text{pen}} - \beta_a)\}$$
$$\to_d N_r(0, (\omega^\top \Omega \omega)I_r)$$

where $\hat{\beta}_{a,\tau_l,\text{pen}}$ is the τ_l-quantile estimator of the active parameters β_a, for $l = 1, \ldots, k$.

Proof. We make use of Kronecker products to rewrite the model-averaged estimator as

$$\sum_{l=1}^{k} \omega_l\{\hat{\beta}_{a,\tau_l,\text{pen}} - \beta_a\} = (\omega^\top \otimes I_s)(\hat{\beta}_{a,\tau,\text{pen}} - 1_k \otimes \beta_a),$$

where $\hat{\beta}_{a,\tau,\text{pen}} = (\hat{\beta}_{a,\tau_1,\text{pen}}^\top, \ldots, \hat{\beta}_{a,\tau_k,\text{pen}}^\top)^\top$. For vectors of fixed length, $\hat{\beta}_{a,\tau,\text{pen}}$ is jointly asymptotically normal (Ruppert and Carroll, 1980, Cor. 1). For a growing length, we consider either linear combinations with a unit vector e in $\mathbb{R}^s$ to state the limiting result as

$$e^t \sqrt{n}(\frac{1}{n}X_a^\top X_a)^{-1/2}(\omega \otimes I_s)(\hat{\beta}_{a,\tau,\text{pen}} - 1_k \otimes \beta_a) \to_d N(0, \omega^\top \Omega \omega),$$

or predictions using a prediction matrix U of fixed dimension $r \times s$ to obtain the asymptotic normality of the model-averaged predictions

$$\sum_{l=1}^{k} \omega_l U_a\{\hat{\beta}_{a,\tau_l,\text{pen}} - \beta_a\} = (\omega^\top \otimes U)(\hat{\beta}_{a,\tau,\text{pen}} - 1_k \otimes \beta_a),$$

by the convergence in distribution of $\sqrt{n}(\frac{1}{n}U_aX_a^\top X_aU_a^\top)^{-1/2}U_a(\omega \otimes I_s)(\hat{\beta}_{a,\tau,\text{pen}} - 1_k \otimes \beta_a)$ to $N_s(0, (\omega^\top \Omega \omega) \otimes I_r)$. Under the assumption that $\alpha_0 \in [0, 2/3]$, the estimation bias asymptotically disappears (Bradic et al., 2011, Th. 2). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Note that here we make the same assumptions as in Bradic et al. (2011), namely, the effect of the penalty disappears, there is no shrinkage bias, and the asymptotic results are for the part of the coefficient vector corresponding to the true active set. Hence, there is no selection uncertainty.

The optimal weights for this scenario can be found by noting the resemblance with the low-dimensional case (3), from which we readily arrive at the following expression:

$$\omega_{\text{opt}} = (\boldsymbol{f}^\top A^{-1}\boldsymbol{f})^{-1}\text{diag}(\boldsymbol{f})A^{-1}\boldsymbol{f}.$$

We might restrict the weights to be nonnegative to mimic the case of the CQR estimator with nonnegative optimal weights.

To make the model-averaged prediction well defined in practical computations, we define the prediction matrix U in dimension $r \times p$ and denote by U_{a_l} the restriction of U to the estimated active set for the lth penalized quantile estimator. We end up with the model-averaged prediction $\sum_{l=1}^k \omega_l U_{a_l}\hat{\beta}_{a_l,\tau_l,\text{pen}}$ instead of $\sum_{l=1}^k \omega_l U_a\hat{\beta}_{a,\tau_l,\text{pen}}$.

4 Simulation study

4.1 Quantile estimators in low dimensions

Since the low-dimensional case has been well-studied, we restrict our attention to the lesser known model-averaging estimator and to comparing its equally- versus optimally-weighted versions. In the simulation study in low dimensions, we consider a linear model

$$Y = X\beta + \varepsilon, \text{ where } X \sim \mathcal{N}(0, \Sigma_X) \text{ and } (\Sigma_X)_{i,j} = (0.5)^{|i-j|}, i, j \in \{1, \dots, p\}. \tag{8}$$

The number of columns in X is set at $p = 3$, and the coefficient vector at $\beta = (3, 1.5, 2)^\top$. The number of observations in one simulated sample is $n = 100$. The error distribution varies from symmetric to asymmetric and from light- to heavy-tailed: normal $\mathcal{N}(0, 1)$; t distribution with degrees of freedom equal to 1, 2, 3 and 5 (t_1, t_2, t_3 and t_5); Beta(3,5); Weibull(1.5, 1); Logistic(0, 1); and Exponential(1). The simulation is repeated 1000 times for each error distribution. The design matrix X is generated once by random sampling and is fixed across the simulation runs.

We implement the equally-weighted composite and model-averaged estimators as well as the (unrestricted) optimally-weighted model-averaged estimator. The (unrestricted) optimally-weighted composite estimator could not be implemented due to the nonconvexity of the objective function in presence of negative weights. We use the R package `quantreg` (Koenker, 2017). The equally-weighted estimation is carried out in one step for the CQR estimator and in two steps for the MAQR estimator; there, first the individual models for the different quantiles are

estimated, and then the estimators from these models are averaged. The optimally-weighted MAQR estimation is done as follows:

1. Apply OLS to equation (8) to obtain the residuals. Estimate the optimal weights from the empirical distribution of the residuals.

2. Estimate individual quantile regressions for all quantiles.

3. Obtain a weighted average of the individual estimators above.

A variation to this scheme uses median regression in step 1, with similar results. We present the simulation results in Tables 1 and 2. Table 1 shows the ratios of the empirical mean squared errors of the estimated coefficient vectors $\hat{\beta}$ from the equally-weighted MAQR estimator to the equally-weighted CQR estimator. Simplifying the ratio of MSEs to the ratio of variances is not possible here, unlike in the previous section, since the two estimators may be biased in finite samples. MSE ratios below one favour the model-averaged estimator, while those above one favour the composite one. Apparently, the composite estimator dominates almost everywhere except for low numbers of quantiles for t_1 and t_2 distributions. There, the differences in MSEs are up to 18% to the advantage of the model-averaged estimator, while elsewhere the composite estimator dominates by up to 64%. The differences in performance are quite small for t_3, t_5, and logistic distributions but larger for normal, Beta, Weibull, and exponential distributions. Additional simulation results at larger sample sizes ($n = 10^3$ and 10^4, not shown) exhibit the same trend.

$f(\varepsilon) \big/ \quad k:$	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	1.22	1.13	1.09	1.07	1.06	1.05	1.04	1.03	1.02
t_5	1.10	1.05	1.03	1.03	1.03	1.03	1.03	1.03	1.03
t_3	1.02	1.00	1.01	1.02	1.02	1.02	1.03	1.05	1.06
t_2	0.96	0.97	0.98	1.01	1.03	1.05	1.08	1.10	1.12
t_1	0.82	0.90	0.99	1.11	1.21	1.32	1.43	1.54	1.64
Logistic(0,1)	1.12	1.08	1.06	1.05	1.05	1.04	1.04	1.04	1.04
Beta(3,5)	1.39	1.26	1.19	1.16	1.14	1.12	1.10	1.09	1.08
Weibull(1.5,1)	1.29	1.21	1.17	1.15	1.14	1.13	1.12	1.12	1.12
Exp(1)	1.33	1.31	1.30	1.30	1.31	1.33	1.34	1.35	1.35

Table 1: Simulated relative efficiency of the equally-weighted model-averaged estimator compared to the equally-weighted composite estimator for different distributions and numbers of quantiles. Ratios less than 1 (colored in gray) indicate that the model-averaged estimator has lower simulated MSE than the composite estimator.

Table 2 contains ratios of the MSEs of the estimated coefficient vectors $\hat{\beta}$ from the optimally-weighted MAQR estimator to the equally-weighted MAQR estimator. Ratios below one suggest that the optimally-weighted estimator is more efficient than its equally-weighted counterpart, while ratios above one signal the opposite. We see that the optimally-weighted estimator is superior for t_2, t_3, Beta, and especially Weibull and exponential distributions. For the latter distribution, the simulated MSEs of the optimally-weighted estimator are up to six or

$f(\varepsilon) \diagdown k:$	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	1.00	1.00	1.01	1.01	1.01	1.02	1.02	1.02	1.02
t_5	1.00	1.01	1.01	1.01	1.00	1.00	1.00	0.99	0.99
t_3	1.00	0.99	0.98	0.96	0.95	0.94	0.93	0.92	0.91
t_2	1.01	0.98	0.94	0.91	0.88	0.85	0.84	0.82	0.81
t_1	1.04	1.12	1.20	1.32	1.48	1.53	1.55	1.52	1.50
Logistic(0,1)	1.00	1.01	1.00	1.00	1.00	1.01	1.01	1.01	1.01
Beta(3,5)	0.98	0.97	0.95	0.94	0.94	0.93	0.92	0.92	0.92
Weibull(1.5,1)	0.85	0.75	0.68	0.64	0.60	0.57	0.55	0.53	0.52
Exp(1)	0.58	0.40	0.31	0.26	0.23	0.20	0.18	0.16	0.15

Table 2: Simulated relative efficiency of the optimally-weighted model-averaged estimator to the equally-weighted model-averaged estimator for different distributions and numbers of quantiles. Ratios less than 1 (colored in gray) indicate that the optimally-weighted estimator has lower simulated MSE than the equally-weighted estimator.

seven times lower than for the estimator with equal weights. The difference in performance becomes larger at larger numbers of quantiles k. On the contrary, the optimally-weighted estimator underperforms for the t_1 distribution with MSEs up to 55% above those of the equally-weighted estimator. For normal, t_5, and logistic distributions, the performance of the two estimators is on par. Results at larger sample sizes ($n = 10^3$ and $n = 10^4$) are similar.

An additional simulation (not shown) using the theoretical optimal weights and large sample sizes reveals a close correspondence to the theoretically expected ratios. Substituting the OLS by a quantile regression at the median in the first stage of the optimally-weighted MAQR estimation causes no important difference. Hence, the differences between the results in Table 2 and the asymptotic analysis are mainly due to the use of estimated weights instead of theoretically optimal weights. The theoretical results do not take the randomness of the weights into account and might therefore be unrealistic for practical use. A deeper study of the finite sample estimation effects is an interesting topic for further research.

4.2 Quantile estimators in high dimensions

In the simulation study for the high-dimensional case, we consider a linear model given in (8), as in Bradic et al. (2011). We set the dimension $p = 150$; the number of observations $n = 100$; and the true coefficient vector $\beta = (3, 1.5, 0, 2, 0, \ldots, 0)^\top$. The error distributions considered are the same as in the low-dimensional case. The design matrix X is generated once by random sampling, then fixed thereafter. The simulation is repeated $n_{\mathrm{sim}} = 1000$ times for each error distribution.

We implement penalized composite and penalized model-averaged estimators using equal weights and nonnegative optimal weights, while the case of unrestricted optimal weights is skipped due to the nonconvexity of the objective function of the composite quantile regression. Equally-spaced quantiles ($\frac{1}{k+1}, \ldots, \frac{k}{k+1}$) for $k = 2, \ldots, 10$ are used for all estimators. Estimating the penalized composite and model-averaged estimators with nonequal weights follows a 2-step

procedure:

1. Apply penalized estimation, e.g. Lasso or penalized quantile regression, on the model in (5) to obtain initial slope estimates $\hat{\beta}^{(0)}$ and residuals; Lasso with 5-fold cross-validation is considered for the initial estimation here. Estimate the quantile intercepts $\hat{b}_{\tau_l}^{(0)}$ as the empirical τ_l-level quantile of the residuals. Obtain the nonnegative optimal weights for the composite estimator as $\nu_{\mathrm{opt}+} = \arg\min_\nu\{\nu^\top A\nu/(\nu^\top \boldsymbol{f})^2\}$ subject to $\nu > 0$ and for the model-averaged estimator as $\omega_{\mathrm{opt}+} = \arg\min_\omega\{\omega^\top \mathrm{diag}(\boldsymbol{f})^{-1} \times A\mathrm{diag}(\boldsymbol{f})^{-1}\omega\}$ subject to $\omega > 0$ as in Sections 3.1 and 3.2, respectively.

2. Optimize the objectives in (6) and (7) using a local linear approximation of the SCAD penalty with the starting values $\beta^{(0)}$ and $b_{\tau_l}^{(0)}$ for the slope and intercept estimators, respectively. Employ 5-fold cross validation to find the optimal tuning parameters.

The relative efficiency of the penalized composite estimator compared to the model-averaged estimator is calculated as the ratio of the trace of the empirical MSEs of the two estimators where we consider the full vector, including the components that are estimated as zero,

$$\mathrm{RE}\{\hat{\beta}_{\mathrm{comp.pen}}(\nu), \hat{\beta}_{\mathrm{mod.avg.pen}}(\omega)\} = \frac{\sum_{r=1}^{n_{\mathrm{sim}}} \sum_{j=1}^{p} \{\hat{\beta}_{\mathrm{comp.pen},j}^{r}(\nu) - \beta_j\}^2}{\sum_{r=1}^{n_{\mathrm{sim}}} \sum_{j=1}^{p} \{\hat{\beta}_{\mathrm{mod.avg.pen},j}^{r}(\omega) - \beta_j\}^2}.$$

The superscript r indicates that the estimator is obtained in the rth simulation run. Similarly, the relative efficiency of the equally-weighted estimators compared to nonnegative optimally-weighted estimators is calculated as

$$\mathrm{RE}\{\hat{\beta}_{\mathrm{pen}}(1_k/k), \hat{\beta}_{\mathrm{pen}}(\nu_{\mathrm{opt}+})\} = \frac{\sum_{r=1}^{n_{\mathrm{sim}}} \sum_{j=1}^{p} \{\hat{\beta}_{\mathrm{pen},j}^{r}(1_k/k) - \beta_j\}^2}{\sum_{r=1}^{n_{\mathrm{sim}}} \sum_{j=1}^{p} \{\hat{\beta}_{\mathrm{pen},j}^{r}(\nu_{\mathrm{opt}+}) - \beta_j\}^2},$$

which is applicable to both the MAQR and the CQR estimator (thus the subscripts in the formula above are not specific to either).

The relative efficiency of the composite estimator compared to the model-averaged estimator using equal weights and nonnegative optimal weights is reported in Tables 3 and 4, respectively. Furthermore, the relative efficiency of the equally-weighted estimators compared to the nonnegative optimally-weighted estimators is presented in Tables 5 and 6. From Tables 3 and 4, we observe that the relative efficiency of the model-averaged estimator compared to the composite estimator, using equal or nonnegative optimal weights, is rarely close to 1, but none of the estimators generally dominates the other. Also, for a particular distribution, the relative efficiency is not necessarily a monotone function of the number of quantiles k.

The performance of the equally-weighted MAQR estimator in relation to its CQR counterpart (Table 3) is superior for Beta, Weibull and exponential distributions with up to 1.64-fold gains in efficiency; but inferior for $t(1)$ and $t(2)$ distributions with up to 2.18-fold losses, except for the case of only two quantiles. The cases of normal, logistic, $t(3)$ and $t(5)$ are mixed. There, having a small or a large number of quantiles favours the model-averaged estimator while having a medium number of quantiles favours the composite estimator.

$f(\varepsilon)$ / k :	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	0.92	0.97	1.01	1.06	1.01	0.97	0.85	0.78	0.70
t_5	0.89	0.98	1.08	1.09	1.10	1.02	0.97	0.88	0.78
t_3	0.88	1.07	1.10	1.23	1.22	1.15	1.08	0.99	0.90
t_2	0.92	1.13	1.31	1.32	1.49	1.39	1.36	1.22	1.16
t_1	0.98	1.17	1.72	1.73	1.98	2.07	2.01	2.03	2.18
Logistic(0,1)	0.86	0.99	1.16	1.21	1.25	1.11	1.06	0.97	0.87
Beta(3,5)	0.95	0.90	0.86	0.87	0.84	0.79	0.77	0.70	0.64
Weibull(1.5,1)	0.94	0.92	0.86	0.87	0.82	0.77	0.73	0.68	0.61
Exp(1)	0.89	1.02	0.99	1.01	0.96	0.95	0.89	0.80	0.73

Table 3: Simulated relative efficiency of the equally-weighted model-averaged estimator compared to the equally-weighted composite estimator for different distributions and numbers of quantiles. Ratios less than 1 (colored in gray) indicate that the model-averaged estimator has lower simulated MSE than the composite estimator.

$f(\varepsilon)$ / k :	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	0.99	0.99	1.01	0.93	0.97	0.99	0.93	0.86	0.89
t_5	0.95	0.92	0.88	0.92	0.79	0.77	0.71	0.74	0.77
t_3	0.98	0.89	0.78	0.77	0.73	0.66	0.66	0.64	0.65
t_2	1.09	0.83	0.72	0.61	0.65	0.58	0.58	0.53	0.53
t_1	1.06	0.60	0.59	0.51	0.48	0.43	0.52	0.40	0.42
Logistic(0,1)	0.91	0.87	0.83	0.79	0.73	0.73	0.72	0.73	0.67
Beta(3,5)	1.12	1.16	1.21	1.23	1.29	1.29	1.29	1.30	1.30
Weibull(1.5,1)	1.03	1.11	1.18	1.20	1.25	1.25	1.30	1.29	1.32
Exp(1)	0.99	1.09	1.27	1.29	1.48	1.48	1.61	1.63	1.76

Table 4: Simulated relative efficiency of the model-averaged estimator compared to the composite estimator, both with nonnegative optimal weights, for different distributions and numbers of quantiles. Ratios less than 1 (colored in gray) indicate that the model-averaged estimator has lower simulated MSE than the composite estimator.

The MAQR estimator with estimated nonnegative optimal weights is favoured over the corresponding CQR estimator by logistic and all t distributions (with the exception of $k = 2$) with gains in efficiency of up to 2.38 times; the converse is true for Beta, Weibull and exponential distributions with up to 1.76-fold losses in efficiency (Table 4).

Considering estimated nonnegative optimal weights against equal weights, we confirm that using equal weights can lead to higher estimation efficiency for both MAQR and CQR, just as the forecast combination puzzle suggests. The MAQR estimator with estimated weights is compared to its equally-weighted counterpart in Table 5, where we see a rather mixed picture. Estimated nonnegative optimal weights are clearly superior in the case of Weibull and exponential distributions; the converse is true for normal, $t(5)$, logistic, and Beta distributions; while $t(1)$, $t(2)$, and $t(3)$ favour either of the estimators depending on the number of quantiles considered.

The CQR estimator with estimated nonnegative optimal weights outperforms its equally-weighted counterpart for Beta, Weibull and exponential distributions, and the gains in efficiency

$f(\varepsilon) \diagdown k:$	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	1.02	1.14	1.26	1.23	1.27	1.29	1.32	1.25	1.32
t_5	1.08	1.23	1.28	1.36	1.22	1.22	1.18	1.18	1.21
t_3	0.97	1.11	1.10	1.10	1.10	1.05	1.04	0.98	0.95
t_2	1.19	1.17	1.05	1.00	0.97	0.97	0.93	0.87	0.90
t_1	1.22	0.95	0.95	0.87	0.78	0.84	0.82	0.81	0.82
Logistic(0,1)	1.08	1.19	1.24	1.25	1.23	1.30	1.24	1.22	1.25
Beta(3,5)	1.13	1.15	1.25	1.26	1.29	1.29	1.30	1.31	1.29
Weibull(1.5,1)	0.88	0.85	0.85	0.83	0.84	0.83	0.82	0.82	0.81
Exp(1)	0.60	0.54	0.51	0.49	0.47	0.45	0.44	0.43	0.42

Table 5: Simulated relative efficiency of the nonnegative optimally-weighted model-averaged estimator to the equally-weighted model-averaged estimator for different distributions and numbers of quantiles. Ratios less than 1 (colored in gray) indicate that the estimator with nonnegative optimal weights has lower simulated MSE than the equally-weighted estimator.

$f(\varepsilon) \diagdown k:$	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	0.94	1.12	1.26	1.41	1.33	1.26	1.21	1.13	1.04
t_5	1.00	1.31	1.57	1.62	1.69	1.62	1.60	1.40	1.22
t_3	0.87	1.34	1.54	1.74	1.84	1.84	1.70	1.51	1.32
t_2	1.00	1.60	1.91	2.17	2.24	2.30	2.18	2.02	1.95
t_1	1.13	1.83	2.75	2.97	3.20	4.03	3.19	4.09	4.29
Logistic(0,1)	1.02	1.35	1.75	1.91	2.09	1.98	1.84	1.61	1.61
Beta(3,5)	0.97	0.89	0.89	0.89	0.84	0.79	0.77	0.71	0.64
Weibull(1.5,1)	0.80	0.70	0.62	0.60	0.54	0.51	0.46	0.43	0.37
Exp(1)	0.53	0.50	0.39	0.39	0.31	0.29	0.24	0.21	0.17

Table 6: Simulated relative efficiency of the nonnegative optimally-weighted composite estimator to the equally-weighted composite estimator for different distributions and numbers of quantiles. Ratios less than 1 (colored in gray) indicate that the estimator with nonnegative optimal weights has lower simulated MSE than the equally-weighted estimator.

increase with the number of quantiles. It underperforms under normal, logistic and t distributions, except for the case of only two quantiles. The losses in efficiency are the largest for $t(1)$, $t(2)$, and logistic distributions. The effect of changing the number of quantiles is nonmonotonic.

Tables 3 to 6 show the relative performance of pairs of estimators (model-averaged against composite, and optimally-weighted against equally-weighted). But which estimator is the overall best for a given distribution and a given number of quantiles? Table 7 provides a summary. On the whole, composite estimation (denoted by hollow symbols) tends to dominate model-averaged estimation (denoted by filled-in symbols). Next, equal weights (denoted by circles) tend to dominate estimated nonnegative optimal weights (denoted by triangles). However, each estimator gets to be the best at least in a few cases, thus there is room for all of them in practice. The underperformance of the model-averaged estimator should be weighted against its computational efficiency, such that in order to save time less accurate estimation could sometimes be acceptable. Also, over a fixed time interval the model-averaged estimator with

$f(\varepsilon)\ \big/\ k:$	2	3	4	5	6	7	8	9	10	30
$\mathcal{N}(0,1)$	●	●	○	○	○	●	●	●	●	●
t_5	●	●	○	○	○	○	●	●	●	●
t_3	▲	○	○	○	○	○	○	▲	▲	▲
t_2	●	○	○	○	○	○	○	○	○	▲
t_1	●	○	○	○	○	○	○	○	○	▲
Logistic(0,1)	●	●	○	○	○	○	○	●	●	●
Beta(3,5)	●	△	●	●	●	●	●	●	△	●
Weibull(1.5,1)	△	△	△	△	△	△	△	△	△	▲
Exp(1)	▲	△	△	△	△	△	△	△	△	▲

Table 7: Estimator with the lowest simulated MSE among the four estimators for different distributions and numbers of quantiles. ○– composite estimator with equal weights; △– composite estimator with estimated nonnegative optimal weights; ●– model-averaged estimator with equal weights; ▲– model-averaged estimator with estimated nonnegative optimal weights. The last column only considers the model-averaged estimators with equal and estimated nonnegative optimal weights because the composite estimators are too expensive to compute for $k = 30$.

many quantiles would be competing against a composite estimator with few quantiles. Therefore, for some distributions a relatively higher efficiency of the model-averaged estimator against the composite one could be expected as compared to the values indicated in the tables.

We have repeatedly noted the computational advantages of the MAQR estimator over the CQR estimator. A quantitative assessment of the high-dimensional case is given in Table 8. While the computational time does not differ much for $k = 2$ quantiles, the MAQR estimator is between 2 and 9 times faster for $k = 10$, depending on the distribution. For MAQR the computation time increases linearly with the number of quantiles, while the increase in execution time may be more rapid for CQR (the precise speed depends on the algorithms used for optimizing the CQR objective). In our example with the Weibull distribution, CQR estimation takes 6.4 seconds at $k = 2$ and 514.7 seconds at $k = 30$. On the other hand, MAQR takes 7.7 seconds for $k = 2$ and only 63.9 seconds for $k = 30$.

4.3 Model-averaged estimator in yet higher dimensions

The computational simplicity of the MAQR estimator allows examining its performance in even higher dimensions. In this section, we analyze the same linear model (8) as before, but increase the column dimension of the design matrix X to $p = 500$ and the number of observations to $n = 200$. We compare three levels of sparsity defined by the number of nonzero elements in the coefficient vector β: a high-sparsity case of $s = 3$, a medium-sparsity case of $s = 100$ and a low-sparsity case of $s = 200$. The error distributions and the numbers of quantiles considered are the same as before. The nonzero components of the true coefficient vector are generated once by randomly sampling s values from 500 independent realizations of a standard normal random variable (the seed number for the random sampling is set to 1 in R software). The simulation is repeated $n_{\mathrm{sim}} = 1000$ times for each error distribution, and the results are reported in Table 9.

In the high-sparsity case of $s = 3$, the estimated nonnegative optimal weights outperform

(Unit: s)	$k = 2$								
	$\mathcal{N}(0,1)$	t_5	t_3	t_2	t_1	Logistic(0,1)	Beta(3,5)	Weibull(1.5,1)	Exp(1)
CQR	6.8	11.5	9.4	20.2	17.5	6.7	10.0	6.4	6.7
MAQR	7.1	8.5	9.6	11.4	13.5	7.9	9.2	7.7	8.8
	$k = 10$								
	$\mathcal{N}(0,1)$	t_5	t_3	t_2	t_1	Logistic(0,1)	Beta(3,5)	Weibull(1.5,1)	Exp(1)
CQR	119.6	193.0	201.9	339.7	283.6	153.3	80.0	103.2	69.3
MAQR	28.2	38.2	42.0	50.1	54.7	31.0	36.0	32.4	35.1
	$k = 30$								
	$\mathcal{N}(0,1)$	t_5	t_3	t_2	t_1	Logistic(0,1)	Beta(3,5)	Weibull(1.5,1)	Exp(1)
CQR	784.1	304.2	643.6	816.7	950.0	765.6	346.1	541.7	337.2
MAQR	65.8	72.5	84.3	99.3	94.6	59.2	71.8	63.9	68.3

Table 8: Execution time (in seconds) for the nonnegative optimally-weighted composite and model-averaged estimators for different error distributions using 2, 10, and 30 quantiles (average over three runs). We use an Intel i7-6700 (Quad-core 3.40GHz) processor to carry out the experiment.

	$s = 3$								
$f(\varepsilon)\,/\quad k:$	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	1.02	1.05	1.05	1.06	1.07	1.06	1.05	1.04	1.05
t_5	1.04	1.06	1.07	1.06	1.05	1.05	1.06	1.05	1.05
t_3	1.03	1.05	1.03	1.03	1.02	1.01	1.00	0.99	0.98
t_2	1.02	0.99	0.94	0.92	0.90	0.89	0.88	0.87	0.85
t_1	1.12	0.80	0.75	0.70	0.68	0.65	0.65	0.64	0.63
Logistic(0,1)	1.02	1.07	1.06	1.06	1.04	1.04	1.02	1.02	1.02
Beta(3,5)	1.09	1.13	1.16	1.20	1.20	1.22	1.26	1.27	1.26
Weibull(1.5,1)	0.87	0.82	0.78	0.77	0.79	0.78	0.78	0.77	0.80
Exp(1)	0.73	0.65	0.64	0.64	0.64	0.63	0.63	0.64	0.63
	$s = 100$								
$f(\varepsilon)\,/\quad k:$	2	3	4	5	6	7	8	9	10
$\mathcal{N}(0,1)$	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
t_5	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
t_3	1.00	1.01	1.00	1.00	1.01	1.01	1.01	1.01	1.02
t_2	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.02	1.02
t_1	1.02	1.00	0.99	0.99	0.99	0.98	0.98	0.98	0.98
Logistic(0,1)	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.02	1.02
Beta(3,5)	1.00	1.03	1.03	1.02	1.02	1.02	1.02	1.02	1.03
Weibull(1.5,1)	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
Exp(1)	1.00	1.01	1.01	1.00	1.01	1.01	1.01	1.01	1.01

Table 9: Simulated relative efficiency of the nonnegative optimally-weighted model-averaged estimator to its equally-weighted counterpart for different distributions and numbers of quantiles, under high and medium sparsity ($s = 3$ and $s = 100$) of the true slope coefficient vector. Ratios less than 1 (colored in gray) indicate that the estimator with nonnegative optimal weights has lower simulated MSE than the equally-weighted estimator.

the equal weights for $t(1)$ and $t(2)$ distributions for all numbers of quantiles but $k = 2$, also for $t(3)$ with $k = 9$ and 10, and for Weibull and logistic distributions. Equal weights are superior elsewhere. The results are similar to these in Section 4.2, Table 5. This is not surprising as the sparsity level considered there is also relatively high with 3 out of 150 true slope coefficients being nonzero.

Meanwhile, in the medium-sparsity case of $s = 100$, the dominance of the estimated nonnegative optimal weights disappears for $t(2)$, $t(3)$, Weibull and exponential distributions and shrinks for $t(1)$ distribution. On the other hand, equal weights are just barely superior in all other cases, with asymptotic relative efficiency never increasing above 1.03.

In the low-sparsity case of $s = 200$, the relative efficiency varies in a narrow band between 1.00 and 1.03 (not shown), indicating that equal weights are the better choice, though only marginally so. A comparison across the different levels of sparsity suggests that the difference in performance between MAQR estimators with estimated nonnegative optimal weights and equal weights grows with the level of sparsity but is negligible when most of the true slope coefficients in the model are nonzero.

5 Quantile estimation for Riboflavin data

In this section, we consider the Riboflavin dataset from Bühlmann and van de Geer (2011) available in R package `hdi` (Dezeure et al., 2015). We aim to predict the logarithm of riboflavin production rate in Bacillus subtilis using a linear model where the regressors are log-transformed expression levels of 4088 genes. CQR and MAQR with different weight choices are employed for estimating the linear model, and their performance is compared via prediction errors.

In the first step, we follow the pre-selection procedure in Bühlmann and van de Geer (2011) for reducing the computational burden; we select the top 150 genes with the highest variance. This results in a sub-dataset of $p = 150$ genes; the number of observations is $n = 71$. In the second step, we randomly split the dataset into a training subsample with 61 observations and a test subsample with the remaining 10 observations. This is done 200 times to assess the variability of the results arising from the random splitting. The training set is used to estimate the slopes and intercepts of the CQR and MAQR with equal weights and nonnegative optimal weights. The estimation follows the two-step procedure described in Section 4, except that the Lasso is replaced by penalized quantile regression at the median for obtaining the initial slope estimator $\beta^{(0)}$. We employ three equally spaced quantiles ($k = 3$).

The 10 observations in the test set are used for calculating prediction errors, and on their basis, three measures of accuracy. In addition to the mean squared prediction error,

$$\text{MSPE}_\tau = \sum_{i=1}^{n}(Y_i - \hat{b}_\tau - X_i\hat{\beta})^2/n, \ \tau \in \{1/(k+1), \ldots, k/(k+1)\},$$

we consider the median absolute prediction error from Xu et al. (2014),

$$\text{MAPE}_\tau = \text{median}\{|Y_i - \hat{b}_\tau - X_i\hat{\beta}|, i = 1, \ldots, n\}, \ \tau \in \{1/(k+1), \ldots, k/(k+1)\},$$

and the prediction error from Wang and Wang (2016),

$$\text{PE}_\tau = \sum_{i=1}^{n} \rho_\tau(Y_i - \hat{b}_\tau - X_i\hat{\beta}), \ \tau \in \{1/(k+1), \ldots, k/(k+1)\},$$

the latter two measures being perhaps more relevant than MSPE for this dataset.

	MSPE				MAPE				PE			
	Equal weights		Optimal weights		Equal weights		Optimal weights		Equal weights		Optimal weights	
τ	CQR	MAQR	CQR	MAQR	CQR	MAQR	CQR	MAQR	CQR	MAQR	CQR	MAQR
0.25	1.103	1.200	1.099	1.196	0.750	0.754	0.763	0.772	3.415	2.998	3.383	3.162
	(0.724)	(0.606)	(0.723)	(0.605)	(0.281)	(0.241)	(0.278)	(0.263)	(1.309)	(0.941)	(1.328)	(1.084)
0.50	1.005	0.905	1.001	0.900	0.639	0.565	0.652	0.612	3.759	3.407	3.798	3.619
	(0.754)	(0.592)	(0.752)	(0.590)	(0.270)	(0.205)	(0.277)	(0.237)	(1.241)	(0.939)	(1.275)	(1.100)
0.75	1.186	1.200	1.182	1.196	0.670	0.623	0.670	0.654	3.135	2.606	3.086	2.820
	(0.715)	(0.606)	(0.715)	(0.605)	(0.247)	(0.187)	(0.249)	(0.236)	(1.319)	(0.779)	(1.310)	(0.947)

Table 10: Accuracy measures MSPE, MAPE and PE at different quantile levels for CQR and MAQR averaged over 200 different splits of the dataset. Values in parentheses are standard deviations.

Table 10 shows the accuracy measures of CQR and MAQR with estimated nonnegative optimal and equal weights. We first compare the performance of composite and model-averaged estimators under the same type of weights. The accuracy measures of MAQR averaged over 200 sample splits are slightly better than those of CQR in 6 out of 9 cases, i.e. with MAPE at the median and the third quartile, PE at all three quantiles, and MSPE at the median. The standard errors of the accuracy measures of MAQR are consistently smaller than those of CQR, suggesting that the performance of MAQR is more stable.

Comparing the prediction accuracy due to estimated nonnegative optimal versus equal weights, we observe that the latter lead to lower MAPE values for all three quantiles and for both estimators. Equal weights also lead to lower values of PE in four cases out of six. Regarding the MSPE, both weight choices produce almost equivalent results, with a slight advantage of estimated nonnegative optimal weights.

It is also interesting to look at the estimated weights themselves since they are not covered in our theoretical analysis, unlike the case of theoretically optimal weights. Figure 2 shows boxplots of the estimated weights for CQR and MAQR. We observe that CQR assigns larger weights to the 25% and 75% quantiles, while MAQR focuses more on the median. The interquartile ranges of the weights for CQR are smaller than those for MAQR; hence, the weights vary less across subsamples for CQR than for MAQR. Curiously, the high variation in estimated weights of MAQR is in contrast to the method's low standard errors of accuracy measures.

In summary, the empirical example illustrates that MAQR is competitive with CQR and that equal weights are competitive with estimated nonnegative optimal weights. In light of the more relevant accuracy measures for this dataset, MAPE and PE, MAQR and equal weights slightly outperform CQR and estimated weights, respectively.
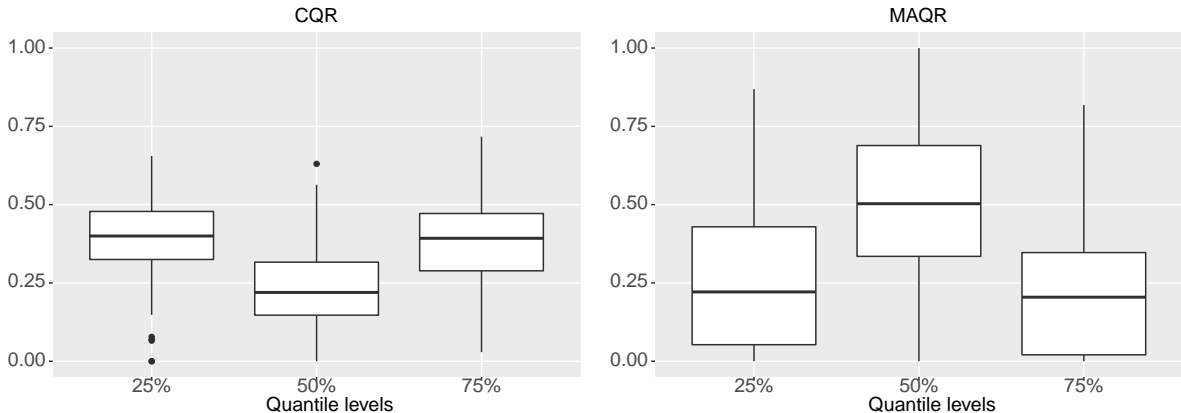
Figure 2: Boxplots of the non-negative optimal weights for CQR (left) and MAQR (right). In each boxplot, weights for 25%, 50% and 75% quantile levels are placed from left to right.

6   Discussion and possible extensions

We have compared model-averaged and composite quantile estimators in linear quantile regression models. Such a comparison is possible also for nonlinear models of the form $Y_i = h(X_i; \beta) + \varepsilon_i$, including the linear model as a special case when $h(X_i; \beta) = X_i^t \beta$. For non-penalized estimators, Oberhofer and Haupt (2016) phrased quite weak assumptions regarding dependence of the errors and heterogeneity and obtained consistency and asymptotic normality of a single nonlinear quantile regression estimator. This extends their earlier work (Oberhofer and Haupt, 2005) where under stronger assumptions also the asymptotic distribution of a vector of such quantile estimators has been derived. This may serve as a starting point for obtaining the asymptotic distribution of a model-averaged nonlinear quantile estimator under weak assumptions on the errors. Also of interest would be deriving results of the composite quantile estimator under such error assumptions to further compare both types of estimators.

In the case of penalized estimators, Jiang et al. (2012) studied the composite nonlinear quantile regression estimator. An essential difference between the linear and nonlinear estimator is in the expression of the asymptotic variance. Jiang et al. (2012) could be the starting point for studying the joint distribution of a vector of nonlinear quantile estimators under penalization in order to obtain results for a model-averaged estimator in a high-dimensional setting.

Our current research includes investigation of a proper weight choice in high dimensions by taking the shrinkage effects of the penalized estimators into account. This relates to the topic of post-selection inference where setting a parameter to zero is considered an act of selection. The currently available asymptotic properties of penalized quantile estimators only provide asymptotic normality for the parameter estimators corresponding to the true active set. To take the selection uncertainty into account, one possibility is to use debiasing results as in van de Geer et al. (2014) or Javanmard and Montanari (2014). Then, different sets of weights are expected to be preferable than under the assumption of perfect variable selection.

Acknowledgements

References

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349.

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

Claeskens, G., Magnus, J., Vasnev, A., and Wang, W. (2016). "The forecast combination puzzle: A simple theoretical explanation". *International Journal of Forecasting*, 32:754 – 762.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30(4):533–558.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Jiang, J., Jiang, X., and Song, X. (2014). Weighted composite quantile regression estimation of DTARCH models. *The Econometrics Journal*, 17(1):1–23.

Jiang, X., Jiang, J., and Song, X. (2012). Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statistica Sinica*, 22(4):1479–1506.

Koenker, R. (1984). A note on L-estimates for linear models. *Statistics and Probability Letters*, 2:323–325.

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

Koenker, R. (2017). *quantreg: Quantile Regression.* R package version 5.33.

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

Oberhofer, W. and Haupt, H. (2005). The asymptotic distribution of the unconditional quantile estimator under dependence. *Statistics and Probability Letters*, 73(3):243–250.

Oberhofer, W. and Haupt, H. (2016). Asymptotic theory for nonlinear quantile regression under weak dependence. *Econometric Theory*, 32(3):686–713.

Ruppert, D. and Carroll, R. (1980). Trimmed least-squares estimation in the linear-model. *Journal of the American Statistical Association*, 75(372):828–838.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wang, K. and Wang, H. J. (2016). Optimally combined estimation for tail quantile regression. *Statistica Sinica*, 26:295–311.

Xu, G., Wang, S., and Huang, J. Z. (2014). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics*, 41(2):365–381.

Zhao, Z. and Xiao, Z. (2014). Efficient regressions via optimally combining quantile information. *Econometric theory*, 30(6):1272–1314.

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

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, 36(3):1108–1126.