KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

High dimensional quantile regression: model averaging and composite estimation

Dissertation presented to obtain the degree of Doctor in Business Economics

by

Jing Zhou

Number 695 Year 2020

FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

KU LEUVEN

High dimensional quantile regression: model averaging and composite estimation

Proefschrift Voorgedragen tot
het Behalen van de Graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Jing Zhou

NUMMER 695 2020

Committee

Prof. Dr. Gerda Claeskens (advisor)	*KU Leuven*
Prof. Dr. Ingrid Van Keilegom	*KU Leuven*
Prof. Dr. Katrien Antonio	*KU Leuven*
Prof. Dr. Anneleen Verhasselt	*Hasselt University*
Prof. Dr. Maarten Jansen	*Université libre de Bruxelles*

Acknowledgements

I am glad to present this thesis on high dimensional statistics, model averaging, and quantile regression after near three and a half years of an intensive yet fruitful and inspiring journey.

It is my greatest pleasure to thank my supervisor Prof. Dr. Gerda Claeskens, for her consistent support, encouragement, and guidance. As a fantastic researcher, you have always been available for my questions and premature thoughts. I have been extremely fortunate to be your student.

I am grateful to have the opportunity to collaborate with Prof. Dr. Daumantas Bloznelis, Prof. Dr. Jelena Bradic, Prof. Dr. Claudia Czado, and Marija Tepegjozova, from whom I received valuable comments and feedback for three chapters of this thesis. I would also like to thank all members of the doctoral committee, Prof. Dr. Ingrid Van Keilegom, Prof. Dr. Maarten Jansen, Prof. Dr. Anneleen Verhasselt, and Prof. Dr. Katrien Antonio for the thorough reading of my thesis. In addition, I appreciate that Prof. Dr. Sebastian Scherr offered me a rewarding opportunity of an interdisciplinary collaboration.

I have also been extremely fortunate to share an office with Prof. Dr. Eugen Pircalabelu and later Frits and Álvaro, who created a friendly office atmosphere. And I am lucky to have great colleagues (Prof. Dr. Martina Vandebroek, Andrea, Elif, Negara, Eni, Min, Motahareh, Vivien, Pieter, Ozan, Leonard, Samson, Abolghassem) who created a supportive working environment in the ORStat research group.

My parents have always been my biggest support. It is my greatest honor to thank them for their unconditional love and trust. Furthermore, I also want to thank my friends from the earlier stages of life who supported and helped me generously. This thesis cannot be done without them.

Jing Zhou Leuven, April 2020.

Table of contents

Introduction

Traditional data collection makes decisions on collecting certain predictive variables, which are potentially associated with a variable that is of primary interest, i.e., the response variable. A standard procedure after data collection is to find the association between the predictive variables and the response variables. In the past decades, numerous models have been proposed and intensively studied for different data structures to model the relationship of collected variables. Traditional linear regression is still among the most popular techniques. It assumes linear relationships between predictive variables and the response variable. However, linear combinations of the predictive variables are not expected to perfectly coincide with the observed values; the deviations are represented in the model by the error variables. The above description can be expressed following

$$Y = X\beta + \boldsymbol{\varepsilon}, \tag{0.1}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix consisting of p predictive variables $X_{\cdot j}, j = 1, \ldots, p$ and n samples $X_{i\cdot}, i = 1, \ldots, n$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the error vector, and $Y \in \mathbb{R}^n$ is the response variable. The components of the coefficient vector $\beta \in \mathbb{R}^p$ of the linear combinations of the predictive variables $X_{\cdot j}$'s, which reflect the associations, are parameters to be estimated. Two pertinent questions arise here: "How to estimate those parameters?" and "How reliable are those estimators?". When the errors follow a Gaussian distribution, a common choice for estimation is the method of ordinary least squares (OLS) using a least squares loss function $\rho_{\mathrm{LS}}(x) = x^2$, to estimate

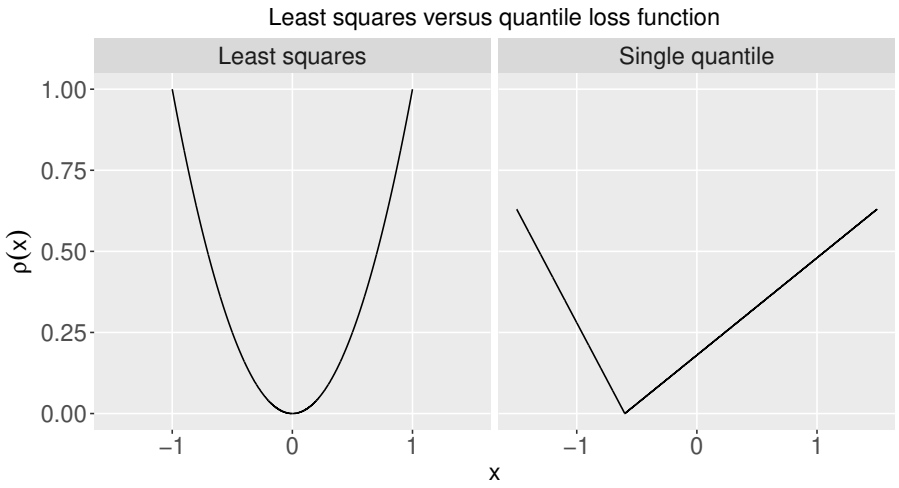


Figure 0.1: *Example figures of the least squares loss and the quantile loss functions at quantile level* $\tau = 0.3$.

the parameter vector β by

$$\widehat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_{i\cdot}\beta)^2 = (X^\top X)^{-1} X^\top Y.$$

However, the least squares loss function is known to be sensitive to non-Gaussian distributed errors. To obtain reliable parameter estimations allowing non-Gaussian distributed errors, quantile regression (Koenker and Bassett, 1978a; Koenker, 2005a) is developed. The quantile regression estimates the parameter vector β at quantile level τ by minimizing the quantile loss function

$$\rho_\tau(x) = (x - u_\tau)(\tau - I\{x \leq u_\tau\}), \tag{0.2}$$

where $u_\tau = F_\varepsilon^{-1}(\tau)$, F_ε is the distribution function of the error ε, $I\{\cdot\}$ is the indicator function. Example figures of the least squares loss and quantile loss functions are in Figure 0.1. Assessing the reliability of the estimators of the parameters is often achieved by hypothesis testing and confidence intervals at a certain significance level.

Nowadays, data collection has become cheaper as technology advances.

Consequently, new data structures, i.e., high dimensional data, appear in various fields, e.g., biology, astronomy, pattern recognition, climate research, etc. Compared to traditional data structures, high dimensional data have significantly more variables, typically of a much larger order than the sample size. Additionally, the number of variables of high dimensional data either is fixed or grows accordingly with the sample size following a fixed ratio, depending on the data collection procedure. Due to a large number of variables in high dimensional data, classical regression theory assuming a fixed number of variables and a sample size that goes to infinity is no longer applicable. Accordingly, statistical theory is developed for high dimensional data with excessive variables. Extensive research has been done on answering the two critical regression questions regarding parameter estimation and the reliability of estimators, in high dimensions, see Candes et al. (2007); Bickel et al. (2009); Rigollet et al. (2011); Vershynin (2018); Bühlmann and van de Geer (2011).

A standard estimation approach is regularization, which estimates regression coefficients by combining a loss function $\rho(x)$ (e.g., least squares loss, Huber's loss, quantile loss, absolute deviation function, etc.) and a regularizer R_λ (e.g., l_1, l_2, SCAD, etc.) with regularization parameter λ. Numerous regularized estimators are derived based on the least squares loss function (Hoerl and Kennard, 1970; Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005), modeling the mean, which is sensitive to non-Gaussian distributed errors. Similar to the classical regression setting, a robust alternative modeling quantiles in high dimensions is the regularized quantile estimator defined as

$$\widehat{\beta}_\tau = \arg\min_{\beta} \Big\{ \sum_{i=1}^{n} \rho_\tau(Y_i - X_{i\cdot}\beta) + R_\lambda(\beta) \Big\},$$

where $\rho_\tau(x)$ is the quantile loss function at quantile level τ and R_λ is a regularizer. Further extensions include two types of weighted quantile regression estimators. In this thesis, I will particularly focus attention to composite and model averaged quantile estimators. By using different quantile

levels, both approaches aim at modeling a conditional mean. Indeed, for a continuous random variable Y, it holds that $E(Y) = \int_0^1 F_Y^{-1}(\tau)d\tau$ with F_Y^{-1} the quantile function. The regularized composite quantile estimator is obtained by minimizing K weighted quantile loss functions

$$\widehat{\beta}_{\mathrm{C}} = \arg\min_{\beta}\{\sum_{k=1}^{K} w_k \sum_{i=1}^{n} \rho_{\tau_k}(Y_i - X_{i\cdot}\beta) + R_\lambda(\beta)\}.$$

And the regularized model averaged quantile estimator is obtained by weighting multiple regularized quantile estimators $\widehat{\beta}_{\tau_k}, k = 1, \ldots, K$, defined as

$$\widehat{\beta}_{\mathrm{MA}} = \sum_{k=1}^{K} w_k \widehat{\beta}_{\tau_k}.$$

A question then arises: how do we choose weights $w = (w_1, \ldots, w_K)^\top$? Weight selection for model averaging, also known as "forecast combination" (Bates and Granger, 1969; Cheng et al., 2015) or "multimodel inference" (Burnham and Anderson, 2002), has been studied substantially in low dimensions. Different weighting schemes have been proposed such as Mallow's C_p (Hansen, 2007), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the focused information criterion (FIC) (Claeskens and Hjort, 2008), adaptive forecast combination via exponential re-weighting (Yang, 2004) combining jackknife or leave-one-out cross-validation (Hansen, 2007; Hansen and Racine, 2012), or sample splitting called "adaptive regression by mixing" (Yang, 2001). For model-averaged and composite quantile estimators, Koenker (2005b) derived weight expressions achieving the lower bound of the corresponding asymptotic variance. However, adapting the above weighting schemes to high dimensions has not been thoroughly studied and requires cautious investigation.

Assessing the reliability of the estimators by statistical inference often relies on asymptotic distributions. However, due to the bias introduced to the estimators by regularization, obtaining asymptotic distributions of the regularized estimators becomes challenging in high dimensions. Sev-

eral popular approaches have been proposed to tackle difficulties arising from the biased estimators in high dimensions. Earlier literature derives the so-called oracle-properties of the non-zero components of the true regression coefficient vector by ignoring the selection uncertainty, meaning the estimated (non-)zero components correspond to the true (non-)zeros of the regression coefficient vector. An alternative approach considering the complete coefficient vector is by debiasing (or desparsifying) (van de Geer et al., 2014; Javanmard and Montanari, 2014b) the regularized estimators. By adding a term compensating the bias caused by regularization, the asymptotic normality of the bias-corrected estimator can be obtained under certain sparsity conditions. Based on the debiased estimators, confidence intervals and hypothesis tests are constructed. Simultaneous hypothesis tests are first constructed following conservative Bonferroni-type adjustments (Javanmard and Montanari, 2014a; van de Geer et al., 2014), and later are improved by bootstrapping the debiased estimators (Zhang and Cheng, 2017; Dezeure et al., 2017).

A major limitation of regularization is strict assumptions on the distributions and the existence of moments of certain variables. In practice, those assumptions may or may not be satisfied, which makes the validity of the estimators doubtful. Without specifying distributions, assuming homoscedastic errors or linearity between response and predictive variables, flexible modeling using copulas has attracted attention. A multivariate joint distribution is linked to a multivariate copula by the probability integral transform, which transforms any univariate variable to a uniformly distributed variable on the interval $[0, 1]$ by applying its marginal distribution functions. For a multivariate joint distribution of the response variable and predictive variables, Sklar's Theorem (Sklar, 1973) in Lemma 0.1 states that there exists an associated copula mapping probability integral transformed variables to the joint distribution.

Lemma 0.1 (Sklar's Theorem)**.** *Let the inverse of the marginal distributions of Y and $X_{\cdot j}$ be F_Y^{-1} and F_j^{-1} respectively. And let the marginal density function of Y and $X_{\cdot j}$ be f_Y and $f_j, j = 1, \ldots, p$. Further, let the*

joint distribution function of Y and $X_{\cdot j}$'s be F with density function f. It holds that

$$C_{V,1,\ldots,p}(v, u_1, \ldots, u_p) = F\Big(F_Y^{-1}(v), F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\Big),$$

and the copula density is determined by

$$c_{V,1,\ldots,p}(v, u_1, \ldots, u_p) = \frac{f\Big(F_Y^{-1}(v), F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)\Big)}{f_Y\big(F_Y^{-1}(v)\big) f_1\big(F_1^{-1}(u_1)\big) \cdots f_p\big(F_p^{-1}(u_p)\big)}$$

Sklar's Theorem bridges multivairate joint distributions with possibly complex forms and copulas with uniformly distributed margins, providing an option of modeling multivariate copulas in hypercubes. However, when the number of variables involved gets large, modeling a multivariate copula is still challenging and does not escape from issues due to high dimensionality. To model multivariate copulas, a pair-copula construction, i.e., vine copulas (Joe, 1996; Bedford and Cooke, 2001, 2002), can be obtained by recursive conditioning and consists of only pair copulas. This pair copula construction avoids modeling multivariate copulas in hypercubes and deals with multiple pair copulas. Since only pair copulas are involved in the modeling process, vine copulas do not suffer from common issues in high dimensions such as "curse of dimensionality". As an extension of vine copulas, flexible quantile modeling as an alternative for regularized quantile estimator has been investigated by Noh et al. (2013, 2015); Chen et al. (2009).

High dimensional statistics has attracted lots of attention due to technical challenges in deriving inferences. Model averaging combines multiple estimators for achieving a more accurate prediction. Quantile regression that models a quantile as opposed to the mean, is known to be less sensitive to non-Gaussian distributed errors. This thesis combines the above three topics discussing the quantile estimator, as well as its model-averaged and composite versions, in high dimensions. Sparse high dimensional quantile regression is the major focus of the first three chapters. We derive oracle-

type asymptotic normality, asymptotic mean squared error (AMSE), as well as the corresponding weight choices for the model-averaged and composite quantile estimators in high dimensions. Confidence intervals and hypothesis testing will be discussed for l_1-regularized single quantile estimators. To relax strict model assumptions, we discuss a flexible quantile modeling using vine copulas in the last chapter.

Chapter 1 assumes a sparse high dimensional linear model allowing for an exponential order of the number of parameters p and the sample size n, such that $\log(p) = O(n^{\delta})$ with $\delta \in (0, 1)$. The model-averaged and composite quantile estimators are robust alternatives to the least squares estimator, while the model-averaged quantile estimator is computationally much faster than the composite estimator. Under the perfect selection assumption assuming only the true non-zero coefficients (the so-called active set of coefficients) are estimated to be non-zeros, we show asymptotic normality of model-averaged quantile predictions with fixed weights. Further, we derive an oracle-type weight expression for the regularized model-averaged quantile estimator achieving the lower bound of the asymptotic variance of the active set of coefficients. We also investigate the effect of equal weights and estimated oracle-type weights on the efficiency of the model-averaged and composite quantile estimators by simulation.

Chapter 2 focuses on the sparse high dimensional setting where the sample size n and the dimension p grow accordingly with the ratio $n/p \to \delta \in (0, 1)$. To relax the perfect selection assumption, we consider the robust approximate message passing algorithm (RAMP) which offers the expression of asymptotic mean squared error (AMSE) for the l_1-regularized M-estimators (e.g., least squares estimators, quantile estimators, etc.) with any convex loss function. The loss functions are not required to be differentiable. We obtain expressions of the AMSE of the l_1-regularized model-averaged and composite M-estimators with fixed weights. An AMSE-type weight choice minimizing the AMSE is derived for the model-averaged and composite M-estimators, and a Stein-type estimator is derived for practically using the expressions of the AMSE and the AMSE-type weight. The

efficiency of the model-averaged and composite quantile estimators using the estimated AMSE-type weights are compared with estimators using estimated oracle-type weights and equal weights.

Chapter 3 is an extension of Chapter 2 focusing on a sequence denoted as $\widetilde{\beta}_{(t)}$ with iteration $t = 0, 1, \ldots$ at convergence in the RAMP iteration. The sequence was shown to be linked to the corresponding debiased estimator (Mousavi et al., 2013; Javanmard and Montanari, 2018) and is asymptotically normally distributed at each iteration t. The asymptotic distribution of $\widetilde{\beta}_{(t)}$ centers at the true regression coefficient and has variance $\bar{\zeta}^2_{\text{emp},(t)}$, which can be obtained directly from the RAMP iteration. We construct componentwise confidence intervals and two-sided individual hypothesis tests by asymptotic normality of $\widetilde{\beta}_{(t)}$ and $\bar{\zeta}^2_{\text{emp},(t)}$ at convergence. The effect of testing multiple hypotheses is handled by a Holm-Bonferroni correction.

Chapter 4 proposes to model a conditional quantile by C-vine copulas, i.e., pair-copula constructions obtained by recursive conditioning, where the bivariate copulas are estimated nonparametrically. The structure of the C-vine is obtained by a new algorithm maximizing truncated conditional log-likelihoods gradually. The proposed method can be adapted to both low and high dimensional data allowing heteroscedastic errors and non-Gaussian copulas. The performance of the proposed estimator is evaluated by out-of-sample prediction, compared with the D-vine based quantile estimator (Kraus and Czado, 2017), using data with different distributions, nonlinear relationships between predictors and response variable, and heteroscedastic errors. Further, we construct a flexible nonparametric mean estimator using the constructed C-vine copulas. The performance of the C-vine based mean estimator is compared with its D-vine based alternative, the ordinary least squares estimator in low dimensions and the Lasso estimator in high dimensions, as well as the nonparametric regression with the least squares cross-validated bandwidths.

The various chapters in this thesis can be found in

(i) Bloznelis D., Claeskens G. and Zhou J. (2019) Composite versus

model-averaged quantile regression. *Journal of Statistical Planning and Inference, 200, 32-46.*

(ii) Zhou, J., Claeskens, G. and Bloznelis, D. (2018). Weight choice for penalized composite quantile regression and for model averaging. *Proceedings of the 33rd International Workshop on Statistical Modelling, University of Bristol, UK, July 16-20, 2018. Pages 219-224.*

(iii) Zhou, J., Claeskens, G. and Bradic, J. (2019). Detangling robustness in high dimensions: composite versus model-averaged estimation. *Submitted to Electronic Journal of Statistics.*

(iv) Zhou, J., Tepegjozova M., Claeskens G. and Czado, C.(2020). Sparse C- and D-vine copula selection in high dimensions. *Technical report.*

(v) Zhou, J. and Claeskens G. (2020). Componentwise confidence intervals and hypothesis testing in high dimensions – a computational approach. *Technical report.*

Chapter 1

Composite versus model-averaged quantile regression

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.

This chapter is based on

Bloznelis D., Claeskens G. and Zhou J. (2019) Composite versus model-averaged quantile regression. *Journal of Statistical Planning and Inference, 200, 32-46.*

Zhou, J., Claeskens, G. and Bloznelis, D. (2018). Weight choice for regularized composite quantile regression and for model averaging. *Proceedings of the 33rd International Workshop on Statistical Modelling, University of Bristol, UK, July 16-20, 2018. Pages 219-224.*

1.1 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 (1978a) 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 the 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 (1978a), while the idea of CQR was proposed by R.V. Hogg in 1979; see (Koenker, 1984, 2005a). 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 1.2 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 1.3 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 1.4 and a data example in Section 1.5 show the estimators' performance in practice.

1.2 Low dimensional linear quantile regression

Consider a linear model $Y = X\beta + \boldsymbol{\varepsilon}$ as in (0.1). The components ε_j's of the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ are independent copies of a random variable denoted by ε. We denote the common cumulative distribution of the ε_i's by F_ε with corresponding density function by f_ε.

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

$$(\hat{u}_\tau, \hat{\beta}_\tau^\top) = \arg\min_{u_\tau, \beta} \sum_{i=1}^{n} \rho_\tau(Y_i - X_{i\cdot}\beta), \tag{1.1}$$

where $\rho_\tau(x) = (x - u_\tau)(\tau - I\{x \leq u_\tau\})$ is the quantile loss function as in (0.2), $\hat{u}_\tau$ is an estimator of the quantile intercept u_τ, and $\hat{\beta}_\tau$ is the estimator of the vector β. The true $100\tau\%$ quantile of Y given $X_{i\cdot}$ is $X_{i\cdot}^\top\beta + u_\tau$, where $u_\tau = F_\varepsilon^{-1}(\tau)$ is the $100\tau\%$ quantile of the distribution of the error ε_i's. Hence, the only regression parameter that depends on the quantile level is the quantile intercept u_τ, while the true β is the same for all quantiles.

1.2.1 Composite and model-averaged estimators

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

$$\begin{aligned}(\hat{u}_{\tau_1,\text{comp}}, \ldots, \hat{u}_{\tau_k,\text{comp}}, \hat{\beta}_{\text{comp}}^\top)(\mathbf{1}_k/k) = \\ \arg\min_{u_{\tau_1},\ldots,u_{\tau_k},\beta} \textstyle\sum_{l=1}^{k} \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - X_{i\cdot}\beta),\end{aligned} \tag{1.2}$$

where the vector $(\hat{u}_{\tau_1,\text{comp}}, \ldots, \hat{u}_{\tau_k,\text{comp}}, \hat{\beta}_{\text{comp}}^\top)$ depends on the weights $\mathbf{1}_k/k$, and $\mathbf{1}_k$ is a vector of length k consisting of ones. In (1.2), the estimating functions from the single quantile regression models as in equation (1.1) are simply summed, or equivalently, all given the same weight. A more general result, see Koenker (1984) and Koenker (2005a, Sec. 5.5), is to allow for different weights $w_\text{C} = (w_{\text{C},1}, \ldots, w_{\text{C},k})^\top$, resulting in the

weighted CQR estimator

$$(\hat{u}_{\tau_1,\text{comp}},\ldots,\hat{u}_{\tau_k,\text{comp}},\hat{\beta}^{\top}_{\text{comp}})(w) = \arg\min_{u_{\tau_1},\ldots,u_{\tau_k},\beta} \sum_{l=1}^{k} w_{\text{C},l} \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - X_{i\cdot}\beta).$$

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 regularized CQR in a high dimensional setting as an alternative to regularized least squares estimation. They develop an oracle estimator that is at worst 30% less efficient than the regularized 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 regularized least squares estimator fails. Bradic et al. (2011) consider a weighted regularized 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 heteroscedastic 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 (1.1) and take their weighted average, to arrive at a weighted MAQR estimator

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

where $w_{\mathrm{MA}} = (w_{\mathrm{MA},1}, \ldots, w_{\mathrm{MA},k})^\top$ is a vector of weights. In model averaging, one usually restricts the weights to sum to one, $\sum_{l=1}^{k} w_{\mathrm{MA},l} = 1$. 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 $\hat{\beta}_{\mathrm{mod.avg}}$, or (2) a single, though more complicated, estimation with a weighted loss function that immediately results in an estimator $\hat{\beta}_{\mathrm{comp}}$?

1.2.2 Asymptotic relative efficiency

For the low dimensional case, part of the answer to this question has been given by Koenker (1984), see also Koenker (2005a, Th. 5.2). Under the assumption that $\sum_{l=1}^{k} w_{\mathrm{C},l} = 1$ 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 $\sqrt{n}(\hat{\beta}_{\mathrm{comp}}(w_{\mathrm{C}}) - \beta)$ with asymptotic covariance matrix

$$Q^{-1} \frac{\sum_{l,l'=1}^{k} w_{\mathrm{C},l} w_{\mathrm{C},l'} \min(\tau_l, \tau_{l'})\{1 - \max(\tau_l, \tau_{l'})\}}{\{\sum_{l,l'=1}^{k} w_{\mathrm{C},l} w_{\mathrm{C},l'} f_\varepsilon(b_{\tau_l}) f_\varepsilon(b_{\tau_{l'}})\}^2} = Q^{-1} w_{\mathrm{C}}^\top A w_{\mathrm{C}} / (w_{\mathrm{C}}^\top \boldsymbol{f}_\varepsilon)^2,$$

where $Q = \lim_{n\to\infty}\{\frac{1}{n} X^\top X\}$, $\boldsymbol{f}_\varepsilon = (f_\varepsilon(b_{\tau_1}), \ldots, f_\varepsilon(b_{\tau_k}))^\top$, 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 (1978a) 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 $a_{ll'}/\{f_\varepsilon(u_{\tau_l}) f_\varepsilon(u_{\tau_{l'}})\}$. This immediately leads to the asymptotic normality of $\hat{\beta}_{\text{mod.avg}}$. It follows that, when $n \to \infty$, with $\tilde{\beta} = (\sum_{l=1}^k w_{\text{MA},l})\beta$,

$$\sqrt{n}(\hat{\beta}_{\text{mod.avg}}(w_{\text{MA}}) - \tilde{\beta}) \xrightarrow{d} N(0, w_{\text{MA}}^\top \Omega w_{\text{MA}} Q^{-1}). \tag{1.3}$$

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

$$\begin{aligned}\text{ARE}\{\hat{\beta}_{\text{mod.avg}}(w_{\text{MA}}), \hat{\beta}_{\text{comp}}(w_{\text{C}})\} &= \frac{\text{asyVar}\, \hat{\beta}_{\text{mod.avg}}(w_{\text{MA}})}{\text{asyVar}\, \hat{\beta}_{\text{comp}}(w_{\text{C}})} \\ &= \sum_{l,l'=1}^k w_{\text{MA},l} w_{\text{MA},l'} \frac{a_{l,l'}}{f_\varepsilon(u_{\tau_l}) f_\varepsilon(u_{\tau_{l'}})} \frac{\{\sum_{j=1}^k w_{\text{C},j} f_\varepsilon(u_{\tau_j})\}^2}{\sum_{j,j'=1}^k w_{\text{C},j} w_{\text{C},j'} a_{j,j'}}.\end{aligned}$$

Using inequalities related to eigenvalues of the matrix A, Koenker (2005a, Th. 5.2, Cor. 5.1) explains that there exists a choice of the weight vectors w_{MA} and w_{C} such that both estimators, $\hat{\beta}_{\text{mod.avg}}(w_{\text{MA}})$ and $\hat{\beta}_{\text{comp}}(w_{\text{C}})$, achieve the same lower bound for the asymptotic variance, asyVar $= (\boldsymbol{f}_\varepsilon^\top A^{-1} \boldsymbol{f}_\varepsilon)^{-1}$. 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 w_{MA} and w_{C} follow expressions

$$w_{\text{MA,opt}} = (\boldsymbol{f}_\varepsilon^\top A^{-1} \boldsymbol{f}_\varepsilon)^{-1} \text{diag}(\boldsymbol{f}_\varepsilon) A^{-1} \boldsymbol{f}_\varepsilon \text{ and } w_{\text{C,opt}} = A^{-1} \boldsymbol{f}_\varepsilon.$$

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_ε.

Note that some components of the optimal weight vectors $w_{\text{MA,opt}}$ and $w_{\text{C,opt}}$ 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.

1.2.3 Weights: theoretically optimal versus estimated optimal versus equal

The optimal weights $w_{\text{MA,opt}}$ and $w_{\text{C,opt}}$ are often not computable due to an incompletely specified density function f_ε, 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 $\hat{w}_{\text{MA}}$ and $\hat{w}_{\text{C}}$ are obviously random. While $w_{\text{MA,opt}}$ and $w_{\text{C,opt}}$ minimize the asymptotic variance of, respectively, $\hat{\beta}_{\text{mod.avg}}(w_{\text{MA}})$ and $\hat{\beta}_{\text{comp}}(w_{\text{C}})$, no such guarantee can be given for their estimated counterparts $\hat{w}_{\text{MA}}$ and $\hat{w}_{\text{C}}$. 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,

$$\begin{aligned} E[\hat{\beta}_{\text{mod.avg}}(\hat{w}_{\text{MA}})] &= \sum_{l=1}^{k} E[\hat{w}_{\text{MA},l}\hat{\beta}_{\tau_l}], \\ \text{Var}[\hat{\beta}_{\text{mod.avg}}(\hat{w}_{\text{MA}})] &= \sum_{l=1}^{k} \text{Var}[\hat{w}_{\text{MA},l}\hat{\beta}_{\tau_l}] \\ &\quad +2\sum_{l=1}^{k}\sum_{l'=1,\ l'<l}^{k} \text{Cov}[\hat{w}_{\text{MA},l}\hat{\beta}_{\tau_l}, \hat{w}_{\text{MA},l'}\hat{\beta}_{\tau_{l'}}]. \end{aligned}$$

Both quantities depend on the joint distribution of the weight vector $\hat{w}_{\text{MA}}$ 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 weights $\hat{w}_{\text{C}}$. Since $w_{\text{MA,opt}}$ and $w_{\text{C,opt}}$ are the theoretical minimizers of the fixed-weight asymptotic variances, using estimated weights $\hat{w}_{\text{MA}}$ and $\hat{w}_{\text{C}}$ 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 1.4.

To avoid the estimation of optimal weights when the error distribution is unknown, we might use the equally-weighted MAQR estimator where $w_{\text{MA},l} = 1/k$ for $l \in \{1, \ldots, k\}$; or the equally-weighted CQR estimator (1.2) of Zou and Yuan (2008). First, we find the choice of the weights w_{MA} of the model-averaged estimator that achieves the same ARE as the equally-weighted composite estimator with $w_{\text{C},l} = \mathbf{1}_k/k$. The equally-weighted composite estimator gains precisely the same asymptotic variance as a weighted model-averaged estimator with weights proportional to the density $f_\varepsilon(u_{\tau_l})$, denoted $w^{[1]}_{\text{MA},l}$, that is, $w^{[1]}_{\text{MA},l} = f_\varepsilon(u_{\tau_l})/\{\sum_{j=1}^{k} f_\varepsilon(u_{\tau_j})\}$. Thus more weight is assigned to the quantile estimator $\hat{u}_{\tau_l}$ for which the

density $f_\varepsilon(u_{\tau_l})$ is larger. Indeed, it can be verified that

$$\text{ARE}\{\hat{\beta}_{\text{mod.avg}}(w_{\text{MA}}^{[1]}), \hat{\beta}_{\text{comp}}(\mathbf{1}_k/k)\} = 1.$$

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

$$\text{ARE}\{\hat{\beta}_{\text{comp}}(w_{\text{C}}^{[1]}), \hat{\beta}_{\text{mod.avg}}(\mathbf{1}_k/k)\} = 1.$$

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 1.1 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 $\text{ARE}(\{\hat{\beta}_{\text{mod.avg}}(\mathbf{1}_k/k), \hat{\beta}_{\text{mod.avg}}(w_{\text{MA,opt}})\}$ and $\text{ARE}(\{\hat{\beta}_{\text{comp}}(\mathbf{1}_k/k), \hat{\beta}_{\text{comp}}(w_{\text{C,opt}})\}$. 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 $k = 15$.

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-weighted composite and model-averaged estimators:

$$\text{ARE}\{\hat{\beta}_{\text{mod.avg}}(\mathbf{1}_k/k), \hat{\beta}_{\text{comp}}(\mathbf{1}_k/k)\} < 1 \Leftrightarrow \left(\frac{\overline{f}_\varepsilon}{f_\varepsilon}\right)^\top \cdot A \cdot \frac{\overline{f}_\varepsilon}{f_\varepsilon} < \mathbf{1}_k^\top \cdot A \cdot \mathbf{1}_k, \quad (1.4)$$

where $\overline{f}_\varepsilon/f_\varepsilon = (\overline{f}_\varepsilon/f_\varepsilon(u_{\tau_1}), \ldots, \overline{f}_\varepsilon/f_\varepsilon(u_{\tau_k}))^\top$ and $\overline{f}_\varepsilon = \sum_{l=1}^k f_\varepsilon(u_{\tau_l})/k$. If this condition holds, the equally-weighted model-averaged estimator is more efficient than its equally-weighted composite counterpart. Condition (1.4) 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.

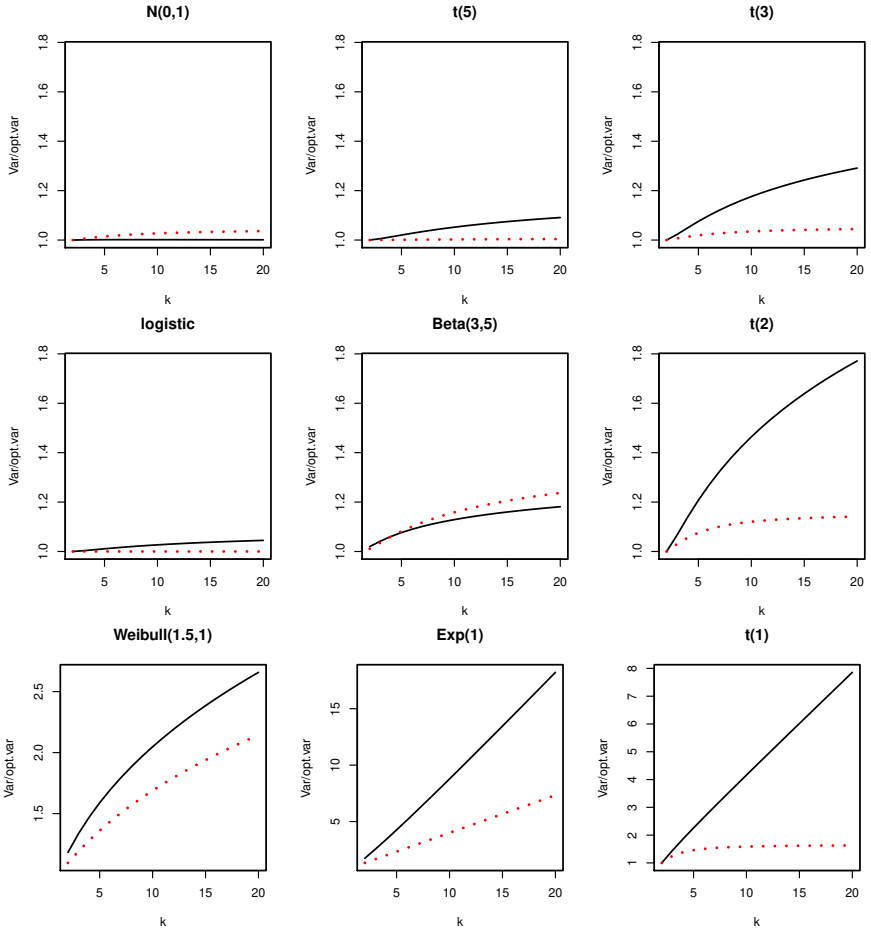


Figure 1.1: *Asymptotic variance of the equally-weighted model-averaged estimator (solid line) and the equally-weighted composite estimator (dotted line) over the optimal variance for various distributions and different numbers k of equally-spaced quantiles.*

1.3 Weighted quantile estimators in high dimensions

Consider now a sparse *high dimensional* linear model as in Bradic et al. (2011) following $Y = X\beta + \boldsymbol{\varepsilon}$ in (0.1). We assume independent and identically distributed mean-zero errors ε_i's; and with the number of parameters p 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, regularized estimation is employed.

1.3.1 Regularized composite quantile estimator

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

$$\begin{aligned}&(\hat{u}_{\tau_1,\text{comp,pen}}, \ldots, \hat{u}_{\tau_k,\text{comp,pen}}, \hat{\beta}^{\top}_{\text{comp,pen}})(\nu) \qquad (1.5)\\ &= \arg \min_{u_{\tau_1},\ldots,u_{\tau_k},\beta} \left\{ \sum_{l=1}^{k} \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - X_{i\cdot}^{\top}\beta) + n \sum_{j=1}^{p} \gamma_{\lambda}(|\beta_j^{(0)}|)|\beta_j| \right\},\end{aligned}$$

where $\rho_{\tau_l}(x) = (x - u_{\tau_l})(\tau_l - I\{x \leq u_{\tau_l}\})$ is the quantile loss function at quantile level τ_l in (0.2). For the regularizer term, $\beta_j^{(0)}$ is an initial slope estimator (e.g., the Lasso, regularized quantile estimator at single quantile levels, etc.) and γ_{λ} is some function, e.g. the derivative of some regularizer function, allowing for (adaptive) lasso (Tibshirani, 1996; Zou, 2006) where $\gamma_{\lambda}(z) = \lambda|z|^{-a}$ for some constant $a \geq 0$ and SCAD (Fan and Li, 2001) where $\gamma_{\lambda}(z) = \lambda[I(z \leq \lambda) + \max(a\lambda - z, 0)I(z > \lambda)/\{(a-1)\lambda\}]$ for $a > 0$.

A related work of Jiang et al. (2012) addresses estimation of a high dimensional nonlinear regression model of the form $Y = h(X;\beta) + \boldsymbol{\varepsilon}$ 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 regularized 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,

$$w_{\text{C}} = (\boldsymbol{f}_{\varepsilon}^{\top} A^{-2} \boldsymbol{f}_{\varepsilon})^{-1/2} A^{-1} \boldsymbol{f}_{\varepsilon}.$$

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

In Bradic et al. (2011), the optimal value of the weights is given by $w_{\text{C}} = A^{-1}\boldsymbol{f}_{\varepsilon}$ to achieve the lower bound for the variance, $(\boldsymbol{f}_{\varepsilon}^{\top} A^{-1} \boldsymbol{f}_{\varepsilon})^{-1}$. While such weights may be negative and thus lead to a nonconvex objective function that is hard to optimize, an alternative weight vector $w_{\text{C},+}$ is

obtained by minimizing $w_{\mathrm{C}}^{\top} A w_{\mathrm{C}}$ subject to having all weights nonnegative and $\boldsymbol{f}_{\varepsilon}^{\top} w_{\mathrm{C}} = 1$. There is no explicit expression for the nonnegative optimal weights $w_{\mathrm{C},+}$. 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.

1.3.2 Model-averaged regularized 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,

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

where

$$(\hat{u}_{\tau_l}, \hat{\beta}_{\tau_l,\text{pen}}^{\top}) = \arg\min_{u_{\tau_l},\beta} \left\{ \sum_{i=1}^{n} \rho_{\tau_l}(Y_i - X_{i\cdot}^{\top}\beta) + n \sum_{j=1}^{p} \gamma_{\lambda_l}(|\beta_j^{(0)}|)|\beta_j| \right\}. \quad (1.6)$$

Note that different regularization constants can be used for the separate quantile estimators, allowing for high flexibility. A major advantage of using the model-averaged regularized 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 regularized 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 $\tilde{x}^\top\beta$ constructed with each of the estimated vectors $\hat{\beta}_{\tau_{l,\text{pen}}}$ for the different values of l, where $\tilde{x}$ is a known vector.

Given the regularity conditions 1 – 4 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 regularity conditions 1 – 5 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 $\tilde{X}_a$ be any design matrix of dimension $r \times s$, containing the information about which $r \geq 1$ predictions we wish to make.

Proposition 1.1. *Under the above-mentioned assumptions and the assumptions of Theorem 2 of Bradic et al. (2011), for the model-averaged regularized quantile predictions it holds that*

$$\sqrt{n}\Big(\frac{1}{n}\tilde{X}_a(X_a^\top X_a)^{-1}\tilde{X}_a^\top\Big)^{-1/2}\{w_{\text{MA},1}\tilde{X}_a(\hat{\beta}_{a,\tau_1,\text{pen}} - \beta_a) + \ldots \\ + w_{\text{MA},k}\tilde{X}_a(\hat{\beta}_{a,\tau_k,\text{pen}} - \beta_a)\} \to_d N_r(0, (w_{\text{MA}}^\top \Omega w_{\text{MA}})I_r)$$

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

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

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

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

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

$$\sum_{l=1}^{k} w_{\text{MA},l}\tilde{X}_a\{\hat{\beta}_{a,\tau_l,\text{pen}} - \beta_a\} = (w_{\text{MA}}^\top \otimes \tilde{X})(\hat{\beta}_{a,\tau,\text{pen}} - \mathbf{1}_k \otimes \beta_a),$$

by the following convergence in distribution

$$\sqrt{n}(\frac{1}{n}\tilde{X}_a(X_a^\top X_a)^{-1}\tilde{X}_a^\top)^{-1/2}\tilde{X}_a(w_{\text{MA}} \otimes I_s)(\hat{\beta}_{a,\tau,\text{pen}} - \mathbf{1}_k \otimes \beta_a) \to_d N_s(0, (w_{\text{MA}}^\top \Omega w_{\text{MA}}) \otimes I_r).$$

Under the assumption that $\alpha_0 \in [0, 2/3)$, the estimation bias asymptotically disappears (Bradic et al., 2011, Th. 2). □

Note that here we make the same assumptions as in Bradic et al. (2011), namely, the effect of the regularizer 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 (1.3), from which we readily arrive at the following expression:

$$w_{\mathrm{MA,opt}} = (\boldsymbol{f}_\varepsilon^\top A^{-1} \boldsymbol{f}_\varepsilon)^{-1} \mathrm{diag}(\boldsymbol{f}_\varepsilon) A^{-1} \boldsymbol{f}_\varepsilon.$$

We might restrict the weights to be nonnegative to mimick 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 $\tilde{X}$ in dimension $r \times p$ and denote by $\tilde{X}_{a_l}$ the restriction of $\tilde{X}$ to the estimated active set for the lth regularized quantile estimator. We end up with the model-averaged prediction $\sum_{l=1}^k w_{\mathrm{MA},l} \tilde{X}_{a_l} \hat{\beta}_{a_l,\tau_l,\mathrm{pen}}$ instead of $\sum_{l=1}^k w_{\mathrm{MA},l} \tilde{X}_a \hat{\beta}_{a,\tau_l,\mathrm{pen}}$.

1.4 Simulation study

1.4.1 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 as in (0.1) with $X \sim \mathcal{N}(0, \Sigma_X)$ and $(\Sigma_X)_{i,j} = (0.5)^{|i-j|}, i, j \in \{1, \ldots, p\}$.

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 es-

timator. 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:

(i) Apply OLS to (0.1) to obtain the residuals. Estimate the optimal weights from the empirical distribution of the residuals.

(ii) Estimate individual quantile regressions for all quantiles.

(iii) 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 1.1 and 1.2. Table 1.1 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.

Table 1.2 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

$f(\varepsilon)$ / 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 1.1: *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.*

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 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 1.2 and the asymptotic analysis are mainly due to the use of estimated weights instead of theoretically optimal weights. The theo-

$f(\varepsilon)$ / 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 1.2: *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.*

retical 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.

1.4.2 Quantile estimators in high dimensions

In the simulation study for the high dimensional case, we consider a linear model given in (0.1), 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 $n_{\text{sim}} = 1000$ times for each error distribution.

We implement regularized composite and regularized 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 regularized composite and model-averaged estimators with nonequal weights follows a 2-step procedure:

(i) Apply regularized estimation, e.g. Lasso or regularized quantile regression, on the model in (0.1) to obtain initial slope estimates $\hat{\beta}^{(0)}$ and residuals; the initial quantile intercepts $\hat{u}_{\tau_l}^{(0)}$ are estimated as the empirical τ_l-level quantile of the residuals. Lasso with 5-fold cross-validation choosing the tuning parameter is considered for the initial estimation here. Obtain the nonnegative optimal weights for the composite estimator as

$$w_{\mathrm{C,opt+}} = \arg \min_{w_{\mathrm{C}}, w_{\mathrm{C}}>0, \mathbf{1}_k^\top w_{\mathrm{C}}=1} \{w_{\mathrm{C}}^\top A w_{\mathrm{C}} / (w_{\mathrm{C}}^\top \boldsymbol{f}_\varepsilon)^2\}$$

and for the model-averaged estimator as

$$w_{\mathrm{MA,opt+}} = \arg \min_{w_{\mathrm{MA}}, w_{\mathrm{MA}}>0, \mathbf{1}_k^\top w_{\mathrm{MA}}=1} \{w_{\mathrm{MA}}^\top \mathrm{diag}(\boldsymbol{f}_\varepsilon)^{-1} A \mathrm{diag}(\boldsymbol{f}_\varepsilon)^{-1} w_{\mathrm{MA}}\}$$

as in Sections 1.3.1 and 1.3.2, respectively.

(ii) Optimize the objectives in (1.5) and (1.6) using a local linear approximation of the SCAD regularizer with the starting values $\beta^{(0)}$ and $u_{\tau_l}^{(0)}$ for the slope and intercept estimators, respectively. Employ 5-fold cross validation to find the optimal tuning parameters.

The relative efficiency of the regularized 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,

$$\begin{aligned} &\mathrm{RE}\{\hat{\beta}_{\mathrm{comp.pen}}(w_{\mathrm{C}}), \hat{\beta}_{\mathrm{mod.avg.pen}}(w_{\mathrm{MA}})\} \\ &= \frac{\sum_{r=1}^{n_{\mathrm{sim}}} \sum_{j=1}^{p} \{\hat{\beta}^r_{\mathrm{comp.pen},j}(w_{\mathrm{C}}) - \beta_j\}^2}{\sum_{r=1}^{n_{\mathrm{sim}}} \sum_{j=1}^{p} \{\hat{\beta}^r_{\mathrm{mod.avg.pen},j}(w_{\mathrm{MA}}) - \beta_j\}^2}. \end{aligned}$$

The superscript r indicates that the estimator is obtained in the rth simulation run. Similarly, the relative efficiency of the equally-weighted estima-

tors compared to nonnegative optimally-weighted estimators is calculated as

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

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 1.3 and 1.4, respectively. Furthermore, the relative efficiency of the equally-weighted estimators compared to the nonnegative optimally-weighted estimators is presented in Tables 1.5 and 1.6. From Tables 1.3 and 1.4, 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 1.3) 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.

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 1.4).

Considering estimated nonnegative optimal weights against equal weights,

$f(\varepsilon)$ / 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 1.3: *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 1.4: *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.*

$f(\varepsilon)$ / 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 1.5: *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.*

$f(\varepsilon)$ / 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 1.6: *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.*

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 1.5, 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 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 1.3 to 1.6 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 1.7 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 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.

$f(\varepsilon)$ / 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 1.7: *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* $k = 30$*.*

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 1.8. 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 $k = 30$.

(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 1.8: *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.*

1.4.3 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 (0.1) 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 $n_{\text{sim}} = 1000$ times for each error distribution, and the results are reported in Table 1.9.

In the high-sparsity case of $s = 3$, the estimated nonnegative optimal weights outperform 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 1.4.2, Table 1.5. 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 per-

	$s = 3$								
$f(\varepsilon)$ / 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
	$s = 100$								
$f(\varepsilon)$ / 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 1.9: *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.*

formance 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.

1.5 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 1.4, except that the Lasso is replaced by regularized 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,

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

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

$$\text{MAPE}_\tau = \text{median}\{|Y_i - \hat{u}_\tau - X_{i\cdot}\hat{\beta}|, i = 1, \ldots, n\},$$

with $\tau \in \{1/(k+1), \ldots, k/(k+1)\}$, and the prediction error from Wang and Wang (2016),

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

where $\hat{\rho}_\tau(x) = (x - \hat{u}_\tau)(x - I\{x \leq \hat{u}_\tau\})$. The latter two measures being perhaps more relevant than MSPE for this dataset.

Table 1.10 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 1.2 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

	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 1.10: *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.*

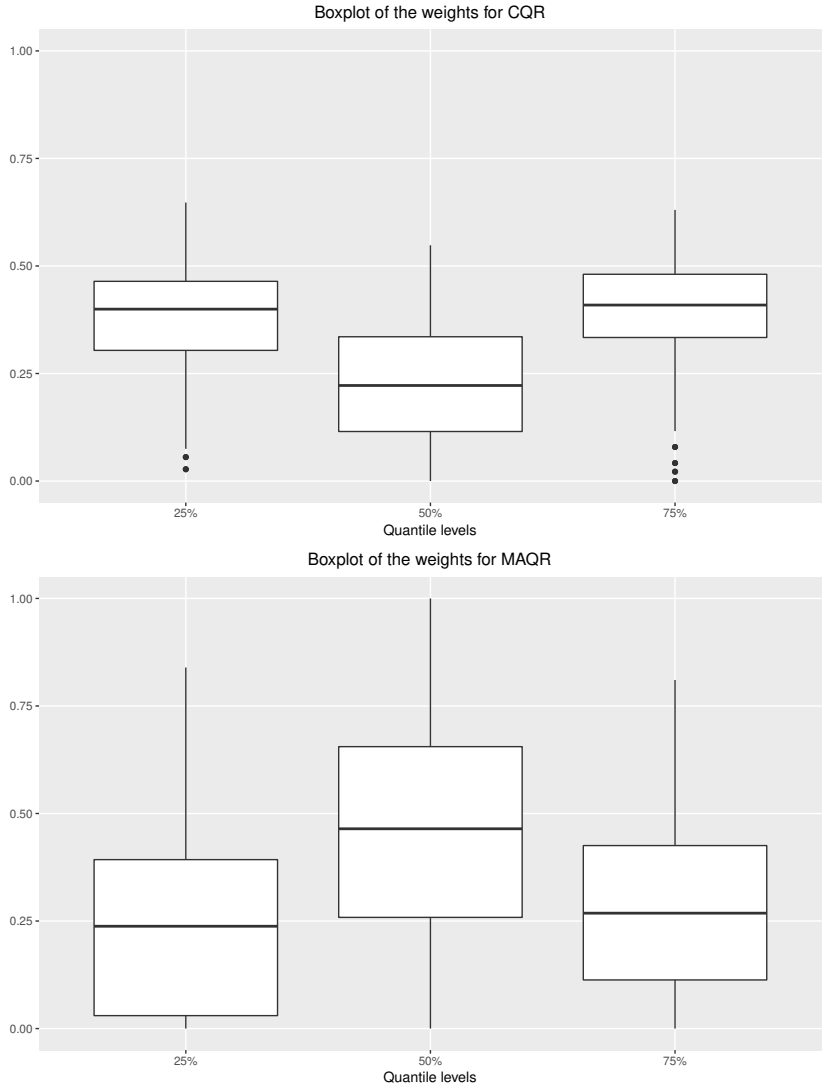


Figure 1.2: *Boxplots of the non-negative optimal weights for CQR (left) and MAQR (right). In each boxplot, weights for* 25%, 50% *and* 75% *quantile levels are placed from right to bottom.*

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.

1.6 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 $Y_i = h(X_{i\cdot}; \beta) + \varepsilon_i$, including the linear model as a special case when $h(X_{i\cdot}; \beta) = X_{i\cdot}^\top \beta$. For nonregularized 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 regularized 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 regularization in order to obtain results for a model-averaged estimator in a high dimensional setting.

Chapter 2 includes investigation of a proper weight choice in high dimensions by taking the shrinkage effects of the regularized 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 regularized 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 (2014a). Then, different sets of weights are expected to be preferable than under the assumption of perfect variable selection.

Chapter 2

Detangling robustness in high dimensions: composite versus model-averaged estimation

Robust methods, though ubiquitous in practice, are yet to be fully understood in the context of regularized estimation and high dimensions. Even simple questions become extremely challenging very quickly. For example, classical statistical theory identifies equivalence between model-averaged and composite quantile estimation. However, little to nothing is known about such equivalence between methods that encourage sparsity. This chapter provides a toolbox to further study robustness in these settings and focuses on prediction without imposing valid model selection. It is well known that prediction is an easier task, and that model selection is often merely a tool for achieving superior prediction.

In particular, we study optimally weighted model-averaged as well as composite l_1-regularized estimation. Optimal weights are determined by minimizing the asymptotic mean squared error. This approach incorporates the effects of regularization without the assumption of perfect selection, as is often used in practice. Such weights are then optimal for prediction quality. Through an extensive simulation study, we show that no single method systematically outperforms others. We find, however, that model-averaged and composite quantile estimators often outperform least squares methods, even in the case of Gaussian model noise. In fact, composite estimators perform better than their model-averaged counterparts. Real data application witnesses the method's practical use through the reconstruction of compressed audio signals.

This chapter is based on

Zhou, J., Claeskens, G. and Bradic, J. (2019). Detangling robustness in high dimensions: composite versus model-averaged estimation. *Submitted to Electronic Journal of Statistics.*

Zhou, J., Claeskens, G. and Bloznelis, D. (2018). Weight choice for penalized composite quantile regression and for model averaging. *Proceedings of the 33rd International Workshop on Statistical Modelling, University of Bristol, UK, July 16-20, 2018. Pages 219-224.*

2.1 Introduction

We investigate the benefits of model-averaged as well as composite estimators in high dimensional problems where the underlying goal is superior prediction quality. Robustness in data analysis with potentially more parameters than samples is a critical practical question. This is of particular

interest in constructing recoveries of compressed images and signals which should have high precision.

Model averaging, as a first tool to improve quality of estimation, forms a weighted average of estimators and is here utilized for regularized sparsity-encouraging estimation in a high dimensional regression setting. Model averaging is also well-known in the Bayesian setting (Hoeting et al., 1999), though we focus on its frequentist version in which a user determines the weights assigned to the separate estimators (e.g., see Claeskens and Hjort, 2008; Hjort and Claeskens, 2003; Yuan and Yang, 2005). Model averaging enjoys a wide application, see, for example, the recent overview chapter for model averaging in ecology by Dormann et al. (2018) and for application to hydrology by Höge et al. (2019). In econometrics, the terminology "forecast combinations" appears (e.g., in Cheng et al., 2015; Bates and Granger, 1969); whereas "multimodel inference" is another commonly used term for this procedure (Burnham and Anderson, 2002).

While the technique is quite thoroughly investigated for low dimensional models, much fewer results have been obtained in high dimensions. Ando and Li (2014) consider high dimensional linear regression. By computing the marginal correlation between each covariate and the response and forming groups according to the obtained values, regularized estimation is avoided, and a fixed number of low-dimensional models are fit by the least squares method and subsequently averaged. Zhao et al. (2016) extend this method to dependent data, while Ando and Li (2017) extend this approach to generalized linear models, again by only fitting low-dimensional models, this time via maximum likelihood estimation. In these papers, the choice of the weights is obtained via cross-validation; see also Hansen (2007) and Hansen and Racine (2012) for similar weight finding approaches in low-dimensional models.

Our setting is different and is theoretically valid (see Theorem 2.1 below). We explicitly work with l_1-regularized estimators that are averaged, and we do not rely on the correct low-dimensional representation of the model. For the choice of the optimal weights, we explicitly take variable

selection effects (of regularization itself) into account. Is the dependence among regularized estimators an impediment or a hidden benefit in obtaining robust predictions, i.e., predictions that do not change much when the data is changed a little?

A second approach to robustness is through composite estimation. While model averaging combines estimators after optimization of their respective loss functions, composite estimation weights the loss functions directly (before optimization). For quantile regression in low dimensions, Koenker (2005b, Theorem 5.2) stated the asymptotic equivalence of model-averaged and composite quantile regression estimators, provided each method uses its own, optimal set of weights that minimize the asymptotic variance. Hence, with optimal weights, in low dimensions, there is no asymptotic preference between the two methods. For high dimensional quantile regression, when one restricts the attention to inference regarding the true nonzero part of the regression coefficient and ignores the variable selection effect, Bloznelis et al. (2019) obtained the same equivalence for high dimensional quantile regression using different types of regularizations (e.g., SCAD, lasso, adaptive lasso, etc).

In practice, however, one works with an estimated coefficient vector for which one is not sure that the regularization has led to the correct selection. Therefore, incorporating imperfections of variable selection is especially important for achieving robustness. This is where our approach differs from Bloznelis et al. (2019) or Bradic et al. (2011), where an irrepresentable condition (needed for consistent model or asymptotically perfect selection) has been used to specify weights and analyze robustness.

The approximate message passing (AMP) algorithm is crucial in our approach to take the variable selection into account when studying the estimators' asymptotic mean squared errors. The use of such algorithms has been investigated by Donoho et al. (2009) and Bayati and Montanari (2011a) for compressed sensing. Donoho and Montanari (2016) explain the use of AMP algorithms for obtaining the variance of high dimensional M-estimators for which $n/p \to \delta \in (1, \infty)$. However, the robustness of

sparsity encouraging AMP estimators is still largely unknown.

In this chapter, we first extend the robust AMP (RAMP) of Bradic (2016) to regularized composite estimation. Second, we construct estimators and develop new theory for the asymptotic mean squared error (AMSE) both for model-averaged and for composite estimators. Note that model-averaged AMSE required an extension of AMP theory for a challenging case of dependent estimates. Besides, we establish new Stein-type risk estimates of the AMSE in both cases.

The new estimates of the AMSE of the model averaged and composite estimators enable a theoretically justified and data-driven, optimal weight choice by minimizing the estimated AMSE (without relying on perfect variable selection). The estimated AMSE gives more information regarding the estimators as compared to merely considering which variables have been selected.

Organization of the chapter. First, in Section 2.2 we detail the model-averaged and composite estimators in a high dimensional setup. Next, we explain the model-averaged robust message passing algorithm in Section 2.3. he limiting behaviour of the estimators in the algorithm is studied by state evolution parameters in Section 2.4. We obtain the estimators' asymptotic mean squared error as well as an estimator of that quantity in Section 2.5. We showcase the procedure for high dimensional regularized quantile regression in Section 2.6 and present numerical results in Section 2.7. Section 2.8 concludes. All proofs, together with the conditions and some technical lemmas are collected in Sections 2.9 and 2.10.

2.2 Model-averaged and composite estimation

We consider a high dimensional linear model $Y = X\beta + \boldsymbol{\varepsilon}$ as in (0.1). We assume the components of $\boldsymbol{\varepsilon}$ to be independent and identically distributed with mean zero, cumulative distribution function F_ε and probability density function f_ε. We allow for a sparse high dimensional setup. Denote by s the l_0 norm of the parameter vector, $s = \|\beta\|_0$, which counts the

number of nonzero components of the vector β. We assume that the ratios $n/p \to \delta \in (0,1)$ and $n/s \to a \in (1,\infty)$ when p, n, s tend to ∞.

We consider two types of weighted estimation methods. First, model-averaged estimation where estimators from different models or estimation methods are weighted and summed to arrive at a final estimator, see (2.2). Second, composite estimation where a weighted average of loss functions is minimized, see (2.3).

For model-averaged estimation of the parameter β, define for $k = 1, \ldots, K$ the regularized estimators

$$\widehat{\beta}_k(\lambda_k) \quad = \quad \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho_k(Y_i - X_{i\cdot}\beta) + \lambda_k \|\beta\|_1 \right\}, \tag{2.1}$$

where $\rho_1, \ldots, \rho_K$ are nonnegative convex loss functions and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^\top$ is a vector of possibly different nonnegative regularization parameters. For a set of weights $w = (w_1, \ldots, w_K)^\top$, the model-averaged estimator is defined as

$$\widehat{\beta}_{\mathrm{MA}}(\boldsymbol{\lambda}) \quad = \quad \sum_{k=1}^{K} w_k \widehat{\beta}_k(\lambda_k). \tag{2.2}$$

Often one assumes that the weights $w_1, \ldots, w_K$ are all nonnegative and sum to 1, although this is not necessary for the computation of the estimator.

For composite estimation we consider again K loss functions, though only with a single nonnegative regularization parameter λ, such that the regularized composite estimator is defined as

$$\widehat{\beta}_{\mathrm{C}}(\lambda) \quad = \quad \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n} w_k \rho_k(Y_i - X_{i\cdot}\beta) + \lambda \|\beta\|_1 \right\}. \tag{2.3}$$

Computationally, composite estimation is harder than model-averaged estimation and requires that all weights are positive to ensure a nonnegative and convex weighted loss function, even when all ρ_k are nonnegative and

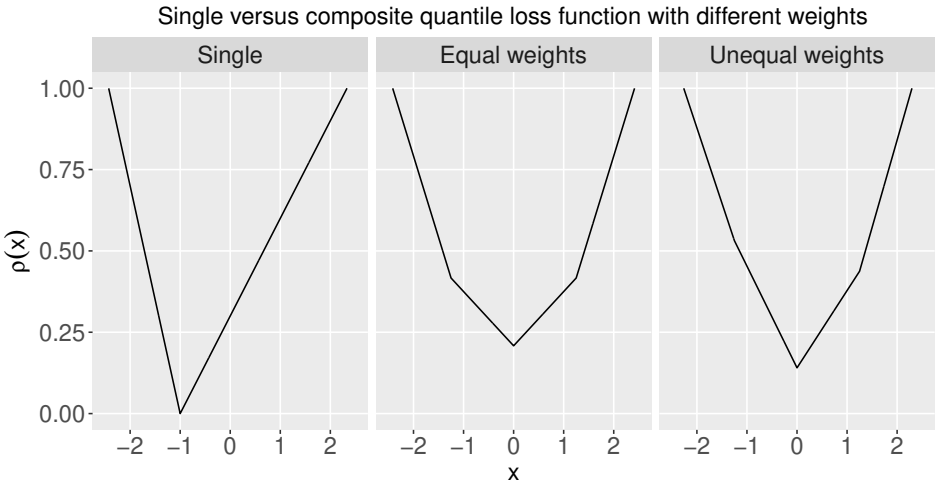


Figure 2.1: *Examples of quantile loss functions. Left:* $\tau = 0.3$ *quantile loss function. Middle: Composite quantile loss function at quantile levels 0.25, 0.5, 0.75 with equal weights* $w = (1/3, 1/3, 1/3)^\top$*. Right: Composite quantile loss function at quantile levels 0.25, 0.5, 0.75 with weights* $w = (0.15, 0.55, 0.3)^\top$.

convex. Hence, for composite estimation it is required that the weight vector $w \in [0,1]^K$ such that $\sum_{k=1}^K w_k = 1$.

As a worked-out scenario throughout the chapter we consider quantile loss functions $\rho_k(\cdot), k = 1, \ldots, K$ that are defined below. For more information about quantile regression with i.i.d. errors, see Koenker (2005b, Sec. 3.2.2). In this chapter we assume that the design matrix X does not contain a column of ones, see Condition (A1) in Section 2.9. With $\tau \in (0,1)$, the τ-quantile of the response Y is obtained as $X\beta + F_\varepsilon^{-1}(\tau) = X\beta + u_\tau$.

Figure 2.1 presents first a single quantile loss function in (0.2) with $\tau = 0.3$. For model averaging we specify K different quantile levels and use K different such quantile loss functions for estimation of β:

$$\rho_k(x) = (x - u_{\tau_k})(\tau_k - I\{x \leq u_{\tau_k}\}), \quad k \in \{1, \ldots, K\}.$$

For composite quantile estimation we assume that the K quantile levels $\tau_1 < \cdots < \tau_K$, then also the quantiles of ε are sorted $u_{\tau_1} < \cdots < u_{\tau_K}$. Define $u_{\tau_0} = -\infty$ and $u_{\tau_{K+1}} = \infty$.

The middle panel of Figure 2.1 depicts such a composite quantile loss function $\rho_C = \sum_{k=1}^{K} w_k \rho_k$ for $K = 3$ quantile levels 0.25, 0.5 and 0.75 with equal weights $w = (1/3, 1/3, 1/3)^\top$. The panel on the right in Figure 2.1 uses the same quantile levels but depicts the quantile loss function ρ_C with weights $w = (0.15, 0.55, 0.3)^\top$.

In general, the composite quantile loss function can be rewritten in the following way,

$$\rho_{\mathrm{C}}(x) = \begin{cases} \displaystyle\sum_{k=1}^{K} w_k(1-\tau_k)(u_{\tau_k} - x), & x < u_{\tau_1} \\ \displaystyle\sum_{k=1}^{K} w_k \tau_k (x - u_{\tau_k}), & x \geq u_{\tau_K} \\ \displaystyle\sum_{k=1}^{\ell} w_k \tau_k |x - u_{\tau_k}| + \sum_{k=\ell+1}^{K} w_k(1-\tau_k)|x - u_{\tau_k}|, & x \in [u_{\tau_\ell}, u_{\tau_{\ell+1}}) \\ & \ell = 1, \ldots, K-1. \end{cases} \tag{2.4}$$

Note that a single quantile loss function can be seen as a special case of a composite loss function with $K = 1$ and the single weight $w_1 = 1$. Theoretical results regarding regularized estimation for a single quantile loss function can be found in Bradic (2016). Henceforth, we concentrate in the example on the composite case.

One aim of this chapter is to investigate the weight choice w by minimizing the asymptotic mean squared error of the estimators $\widehat{\beta}_{\mathrm{MA}}(\boldsymbol{\lambda})$ and $\widehat{\beta}_{\mathrm{C}}(\lambda)$.

2.3 Robust approximate message passing

The idea behind approximate message passing algorithms is to provide an iterative procedure that has as its fixed point the estimator of interest; in

this case the minimizer (2.2) of the regularized loss function in the case of model averaging, and the estimator (2.3) in the case of composite estimation. Due to a mean squared convergence (Serfling, 2009, Sec. 1.2.3, convergence in rth mean) between the solution of the approximate message passing algorithm and the estimator (2.2), respectively (2.3), the asymptotic mean squared error that holds for the solution to the approximate message algorithm, is also the asymptotic MSE of the other estimator. Studying effects of regularization while allowing $n/p \to \delta \in (0, 1)$ is extremely challenging. In these cases, AMP provides theoretical advantages as it enables a complete (not yet easy but tractable) structure for obtaining AMSE. This chapter is the first to not only obtain but also use the asymptotic mean squared error of the regularized estimators to optimize the weight choice of both the model-averaged estimator and the composite estimator. We extend the theory of the RAMP to apply to the model-averaged estimator, see Theorem 2.1. Challenges arise with incorporating dependence into the AMSE expression. This, in turn, leads to a new Stein-type estimator of RAMPs asymptotic MSE, see Theorem 2.2. While we focus on the weight choice, the availability of an estimated AMSE may be used in other contexts too such as for the construction of confidence intervals.

2.3.1 Notation

When the composite loss function $\rho_{\mathrm{C}} = \sum_{k=1}^{K} w_k \rho_k$ is used in the RAMP algorithm with tuning parameter α we denote the estimator at iteration number t by $\widehat{\beta}_{\mathrm{C},(t)}(\alpha)$. When the value of the tuning parameter is clear from the context we also denote the RAMP estimator by $\widehat{\beta}_{\mathrm{C},(t)}$.

For constructing the model averaging estimator we denote the separate estimators from the RAMP algorithm using regularity parameters α_k, $k = 1, \ldots, K$ by $\widehat{\beta}_{k,(t)}(\alpha_k)$ and the model averaged estimator is denoted by $\widehat{\beta}_{\mathrm{MA},(t)}(\boldsymbol{\alpha}) = \sum_{k=1}^{K} w_k \widehat{\beta}_{k,(t)}(\alpha_k)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$. When the value of the tuning parameters is clear from the context we denote the model averaging RAMP estimator by $\widehat{\beta}_{\mathrm{MA},(t)}$.

A generic estimator, without referring to a specific loss function or construction is denoted by $\widehat{\beta}_{(t)}$, using tuning parameter α; the subscript (t) refers to the iteration number.

2.3.2 The robust approximate message passing algorithm

We first revise the (robust) approximate message passing algorithm which consists of three steps which are iterated until convergence. In comparison with the simpler AMP for the case with a differentiable convex loss function (Donoho et al., 2009), this procedure for robust high dimensional parameter estimation (Donoho and Montanari, 2016; Bradic, 2016) adjusts the residuals to incorporate the effective score directly. While more details are given in Algorithm 1, which is applied to the different loss functions $\rho_1, \ldots, \rho_k$ and to their weighted sum $\rho_{\mathrm{C}} = \sum_{k=1}^{K} w_k \rho_K$, we here provide the main outline. The used notation does not explicitly indicate a dependence on the number of coefficients p to not overcomplicate the formulas.

Donoho and Montanari (2016) proposed to use the following proximal mapping operator to adjust the residuals. With $b > 0$,

$$\mathrm{Prox}(z, b) = \arg\min_{x \in \mathbb{R}} \{ b\rho(x) + \frac{1}{2}(x - z)^2 \}$$

which minimizes the square loss regularized by the non-differentiable loss. The parameter b controls how the proximal operator map points to the minimum of the non-differentiable loss, where small values correspond to a small movement towards the minimum of ρ. The fixed point solution of the proximal operator coincides with the minimum of the loss function ρ. For more information, see Parikh and Boyd (2014).

We continue with the worked out example on quantile regression, see (2.4). For $\ell = 0, \ldots, K$, define $h(\ell) = \sum_{k=1}^{\ell} w_k \tau_k - \sum_{k=\ell+1}^{K} w_k (1 - \tau_k)$, where we define a summation sign to be equal to zero in case the upper summation index is smaller than the lower one, that is, $\sum_{i=a}^{b} x_i = 0$ if

$b < a$. The proximal operator for the composite quantile case, see (2.4), is

$$\text{Prox}(z;b) = \begin{cases} z - bh(\ell), & z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_{\ell+1}} + bh(\ell)),\ \ell = 0, \ldots, K \\ u_{\tau_\ell} & z \in [u_{\tau_\ell} + bh(\ell-1), u_{\tau_\ell} + bh(\ell)],\ \ell = 1, \ldots, K. \end{cases} \tag{2.5}$$

See Section 2.10.2 for the derivation of the algorithm.

We now describe the three steps in more detail.

Step 1: Create adjusted residuals.
We use the estimates $\widehat{\beta}_{(t-1)}$ and $\widehat{\beta}_{(t)}$ from iteration steps $t-1$ and t to compute the adjusted residuals

$$z_{(t)} = Y - X\widehat{\beta}_{(t)} + \tag{2.6}$$
$$n^{-1}G(z_{(t-1)}; b_{(t-1)}) \sum_{j=1}^{p} I\left\{\eta\big(\widehat{\beta}_{(t-1),j} + X_{\cdot j}G(z_{(t-1)}; b_{(t-1)}); \theta_{t-1}\big) \neq 0\right\},$$

where the soft-thresholding function $\eta(x;\theta) = \text{sign}(x)\max(|x| - \theta, 0)$. In Algorithm 1, see Section 2.4, we give details on how to set the soft-thresholding parameter θ, which might change in each iteration, and we explain that a proper choice of θ as a function of the regularity constant λ leads to an equivalence of the RAMP estimator and the regularized estimator. This step adds a product related to n, p, s to the ordinary residual $Y - X\widehat{\beta}$. Bradic (2016) recognized that this adjustment is similar to the proximal gradient descent (Beck and Teboulle, 2009) using the step size $\sum_{j=1}^{p} I\left\{\eta\big(\widehat{\beta}_{(t-1),j} + X_{\cdot j}G(z_{(t-1)}; b_{(t-1)}); \theta_{t-1}\big) \neq 0\right\} \cdot p/s$.

The score function G is defined in (2.10). And we explain its construction below, starting from the effective score function $\widetilde{G}$. The effective score function used in Donoho and Montanari (2016) is

$$\widetilde{G}(z;b) = b \cdot \partial\rho(x)|_{x=\text{Prox}(z;b)},\ \text{with } b > 0; \tag{2.7}$$

a subgradient is used in case of nondifferentiability. That is, for a value x

where ρ is non-differentiable

$$\partial\rho(x) = \{y : \rho(u) \geq \rho(x) + y(u - x), \forall u\}.$$

Throughout, we use ∂_1 as the notation for the partial derivative or partial subgradient of a function with respect to its first argument. Functions (e.g. $\widetilde{G}$) are applied componentwise to vectors.

For the example on composite quantile regression the subgradient of ρ_{C} is computed as,

$$\partial\rho_{\mathrm{C}}(x) \begin{cases} = h(\ell), & x \in (u_{\tau_\ell},\ u_{\tau_{\ell+1}}),\ \text{for } \ell = 0, \ldots, K, \\ \in [h(\ell-1), h(\ell)], & x = u_{\tau_\ell},\ \text{for } \ell = 1, \ldots, K. \end{cases} \tag{2.8}$$

The effective score function for composite quantile regression, see Section 2.10.2, is

$$\widetilde{G}(z; b) = \begin{cases} bh(\ell), & z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_{\ell+1}} + bh(\ell)),\ \ell = 0, \ldots, K \\ z - u_{\tau_\ell}, & z \in [u_{\tau_\ell} + bh(\ell-1), u_{\tau_\ell} + bh(\ell)],\ \ell = 1, \ldots, K. \end{cases} \tag{2.9}$$

To incorporate the sparsity, Bradic (2016), see also Bayati and Montanari (2011a), used the rescaled, min regularized effective score function,

$$G(z; b) = \delta\omega^{-1}\widetilde{G}(z; b) \tag{2.10}$$

where ω corresponds to the limit of s/p with $s = \|\beta\|_0$ the true number of nonzero components as $s, p \to \infty$. Condition (A2) in Section 2.9 formalizes the limit of the coefficient vector β by placing the sequence of uniform distributions on the components; the components of the p-vector β converge to a distribution F_{B_0} and can be represented by a random variable B_0. In addition, Condition (A2) provides an alternative definition of the ratio $\omega = P(B_0 \neq 0)$ using the limit random variable B_0.

Step 2: Use the effective score function to set b.
We choose the scalar $b_{(t)}$ such that the empirical average of the effective score function $G(z; b)$ has slope 1, thus $n^{-1}\sum_{i=1}^{n} \partial_1 G(z_{i,(t)}; b_{(t)}) = 1$. In

the case of a non-differentiable loss function, Bradic (2016) proposed to solve $\widehat{\nu}(b_{(t)}) = 1$ with

$$\widehat{\nu}(b_{(t)}) = \frac{b_{(t)}\delta}{\omega}\big(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2}\partial v_j\{z_{(t),i}\} + \sum_{l=1}^{L-1}\gamma_l\{\widehat{f}_P(r_{l+1}) - \widehat{f}_P(r_l)\}\big). \quad (2.11)$$

Condition (A3) in Section 2.9 (See also Condition (R) of Bradic, 2016) writes the subgradient $\partial\rho$ as a sum of three functions: the differentiable functions v_1 and v_2, where v_1 has an absolutely continuous derivative and v_2 is continuous consisting piecewise linear parts, and non-decreasing step functions $\gamma_l\{\widehat{f}_P(r_{l+1}) - \widehat{f}_P(r_l)\}, l = 1, \ldots, L-1$, with step magnitude γ_l, endpoints r_l, and $\widehat{f}_P$ the estimated density of $\text{Prox}(z_{i,(t)}; b_{(t)})$ for $i = 1, \ldots, n$.

The derivation of the estimator $\widehat{\nu}(b_{(t)})$, see also Section 2.10.2, relies on the limiting behaviour of the system, see Section 2.4.

For the composite quantile loss, see (2.4), we clearly see the dependence on the quantiles. The estimator of ν in (2.11) uses $v_1(z) = 0$ and $v_2(z) = z - u_{\tau_\ell}\ z \in [u_{\tau_\ell} + bh(\ell-1), u_{\tau_\ell} + bh(\ell)],\ \ell = 1, \ldots, K$, corresponding to the differentiable pieces in (2.9). The step functions $v_3(z) = bh(\ell)$ when $z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_\ell+1} + bh(\ell)),\ \ell = 0, \ldots, K$. Solving for b in the equation $\widehat{\nu}(b) = 1$ is equivalent to solve for b in the following equation,

$$\begin{aligned}\frac{s}{n} \;=\; & b\big[\sum_{k=0}^{K-1} h(k) f_z\{u_{\tau_{k+1}} + bh(k)\} - \sum_{k=1}^{K} h(k) f_z\{u_{\tau_k} + bh(k)\}\big] \\ & + F_z\{bh(K)\} - F_z\{bh(0)\}, \end{aligned} \quad (2.12)$$

where F_z is the cumulative distribution function and f_z the density function of the adjusted residuals. In practice, a grid search is performed to approximate the solution $\widehat{b}_t$. For each b in the grid, we use the empirical cumulative distribution, that is, $\widehat{F}_z(bh(K)) = n^{-1}\sum_{i=1}^{n} I\{z_{i;(t)} \le bh(K)\}$. A kernel density estimator of f_z with the Gaussian kernel estimates defined as $\widehat{f}_z\{u_{\tau_k} + bh(k)\} = (nh)^{-1}\sum_{i=1}^{n}\phi\{(z_{i;(t)} - u_{\tau_k} - bh(k))/h\}$ with ϕ being the standard normal density function. The solution $\widehat{b}_t$ is taken to be the

average of the smallest b in the grid that makes the righthand side of (2.12) smaller than $\frac{s}{n}$ and the next value in the grid.

Step 3: Update the estimator of β.
Use the estimated $b_{(t)}$ from the previous step to update the estimate of β to

$$\widehat{\beta}_{(t+1)} = \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}), \text{ where } \widetilde{\beta}_{(t)} = \widehat{\beta}_{(t)} + X^\top G(z_{(t)}; b_{(t)}). \tag{2.13}$$

A similar iterative thresholding algorithm is proposed in Daubechies et al. (2004) for parameter estimation with sparsity constraint. The estimator $\widetilde{\beta}_{(t)}$, before applying the soft-thresholding function, is of interest too since it can be interpreted as a debiased estimator (Javanmard and Montanari, 2014a,b; van de Geer et al., 2014; Javanmard and Montanari, 2018). A thorough study, however, is beyond the case of the current work.

2.4 State evolution

Within each iteration step t of the approximate message passing algorithm, state evolution studies the limiting behaviour of the estimators when the sample size goes to infinity. We now define the state evolution parameter $\bar{\zeta}^2_{(t)}$ which is critical for Algorithm 1. We start by defining the empirical version as follows

$$\bar{\zeta}^2_{\text{emp},(t)} = \frac{1}{n}\sum_{i=1}^{n} G(z_{i,(t)}; b_{(t)})^2. \tag{2.14}$$

This quantity is linked to the state evolution recursion which describes the limiting behaviour of large systems, see Theorem 2 in Bayati and Montanari (2011a) and Lemma 1 in Bradic (2016). It holds that, see Section 2.10.2 for details,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} G\big(z_{i,(t)}; b_{(t)}\big)^2 \overset{a.s.}{=} E[G(\varepsilon - \bar{\sigma}_{(t)}Z; b_{(t)})^2] = \bar{\zeta}^2_{(t)}, \tag{2.15}$$

Algorithm 1: RAMP algorithm for a single loss function with tuning parameter α

Function `singleRAMP(`α`)`**:**

Initialization: $\widehat{\beta}_{(0)} \leftarrow 0 \in \mathbb{R}^p$,
iteration index $t \leftarrow 0$, final iteration $t_{\text{final}} \leftarrow 0$,
adjusted residuals $z_{(0)} \leftarrow Y \in \mathbb{R}^n$,
empirical state evolution $\bar{\zeta}^2_{(0)}$ using (2.14),
tuning parameter of the soft-thresholding function $\theta_{(0)} = \alpha\bar{\zeta}_{(0)}$

while *iteration* $t \leq T$ *and tolerance* $tol > \varepsilon_{\text{tol}}$ **do**

(i) *Adjust residuals:* adjust the residuals $z_{(t)} \in \mathbb{R}^n$:

$$z_{(t)} \leftarrow Y - X\widehat{\beta}_{(t)} + \frac{1}{n}G(z_{(t-1)}; b_{(t-1)}) \sum_{j=1}^{p} I\left\{\eta\big(\widehat{\beta}_{j,(t-1)} + X_{\cdot j}^{\top} G(z_{(t-1)}; b_{(t-1)}); \theta_{(t-1)}\big) \neq 0\right\}.$$

(ii) *Effective score:*

(a) choose the scalar $b_{(t)}$ satisfying
if G differentiable **then** $1 = \frac{1}{n}\sum_{i=1}^{n} \partial_1 G(z_{i,(t)}; b_{(t)})$
else $1 = \widehat{\nu}(b_{(t)})$, see (2.11);

(b) update the state evolution parameter $\bar{\zeta}^2_{(t)}$ using (2.14)

(c) update the tuning parameter $\theta_{(t)} \leftarrow \alpha\bar{\zeta}_{(t)}$.

(iii) *Estimation:* Update the coefficient estimation

$$\widetilde{\beta}_{(t)} \leftarrow \widehat{\beta}_{(t)} + X^{\top} G(z_{(t)}; b_{(t)}) \text{ and } \widehat{\beta}_{(t+1)} \leftarrow \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}),$$

(iv) *Adjust iteration index:* $t \leftarrow t + 1$; $t_{\text{final}} \leftarrow t$.

(v) *Calculate tolerance:* $tol = \|\widehat{\beta}_{(t)} - \widehat{\beta}_{(t-1)}\|^2/p$

end

return $\widehat{\beta} \leftarrow \widehat{\beta}_{(t_{\text{final}})}$, $\widetilde{\beta} \leftarrow \widetilde{\beta}_{t_{\text{final}}}$,
the estimated $\text{AMSE}(\widehat{\beta}; \beta)$ *for* $\widehat{\beta}$*, see Theorems 2.1 and 2.2.*

which, together with $\bar{\sigma}^2_{(t)}$ in (2.17), are the state evolution parameters for the large system, Z is a random variable with standard normal distribution independent of everything else.

Due to the symmetry of Z, the state evolution parameter is formally defined as

$$\bar{\zeta}^2_{(t)} = E[G(\varepsilon + \bar{\sigma}_{(t)}Z; b_{(t)})^2]. \tag{2.16}$$

This definition explicitly features the extra Gaussian component $\bar{\sigma}_{(t)}Z$ in the limiting version, with variance

$$\bar{\sigma}^2_{(t)} = \delta^{-1}E[(\eta(B_0 + \bar{\zeta}_{(t-1)}Z; \theta_{(t-1)}) - B_0)^2] \tag{2.17}$$

with B_0 defined in Condition (A2) in Section 2.9. To connect the theoretical expression of $\bar{\sigma}^2_{(t)}$ to Algorithm 1, we apply Bayati and Montanari (2011a, Eq.(3.6)) and Bradic (2016, Eqs.(7.10), (7.19)). This leads to

$$\delta^{-1}\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\{\eta(\widehat{\beta}_{(t),j} + X_{\cdot j}^{\top}G(z_{i,(t)}; b_{(t)}); \theta_{(t)}) - \beta_j\}^2 \overset{a.s.}{=} \bar{\sigma}^2_{(t)}. \tag{2.18}$$

Note that (2.18) features the debiased estimator from (2.13).

We now explain the connection between the estimators that explicitly use an l_1 regularization and the corresponding estimators from the RAMP algorithm.

By applying Theorem 2 of (Bradic, 2016) we get the immediate connection between the regularized estimators $\widehat{\beta}_k(\lambda_k)$ for $k = 1, \ldots, K$ and the corresponding estimators obtained by applying the RAMP algorithm with a suitable choice of its regularity parameter α. We explain this below. Since the regularized estimators $\widehat{\beta}_k(\lambda_k)$ for $k = 1, \ldots, K$ are used for $\widehat{\beta}_{\mathrm{MA}}(\boldsymbol{\lambda})$, (2.2), the connection between the model-averaged estimators from regularization and from application of the RAMP algorithm, follows immediately from the connections between the K separate estimators. The composite estimator $\widehat{\beta}_{\mathrm{C}}(\lambda)$, (2.3), is a special case of a model averaged estimator with $K = 1$, weight equal to one, and loss function $\rho_{\mathrm{C}} = \sum_{k=1}^{K} w_k\rho_k$.

Denote $(\bar{\zeta}^2, b)$ as the fixed point solution when the iteration number $t \to \infty$ of the following equations,

$$\bar{\zeta}^2_{(t)} = E[G(\varepsilon + \bar{\zeta}_{(t)}Z; b_{(t)})^2] = (\delta/\omega)^2 E[\widetilde{G}(\varepsilon + \bar{\zeta}_{(t)}Z; b_{(t)})^2] \quad (2.19)$$

$$1 = E[\partial_1 G(\varepsilon + \bar{\zeta}_{(t)}Z; b_{(t)})] = (\delta/\omega) E[\partial_1 \widetilde{G}(\varepsilon + \bar{\zeta}_{(t)}Z_k; b_{(t)})]. \quad (2.20)$$

Note that (2.19) is the state evolution recursion for the large system while in (2.20) the first equality is the population version of the requirement in step 2 in Algorithm 1 which states that $n^{-1}\sum_{i=1}^{n}\partial_1 G(z_{i,(t)}; b_{(t)}) = 1$. The second equalities of both (2.19) and (2.20) follow by using the definition of G in (2.10), with $\widetilde{G}$ being defined in (2.7).

Then, under conditions (A1)–(A5) (see Section 2.9), for the RAMP algorithm with $\theta = \alpha\bar{\zeta}$, where the tuning parameter $\alpha > 0$ (which motivates the definition of $\theta_{(t)} = \alpha\bar{\zeta}_{(t)}$ in Algorithm 1), and for the l_1-optimization with

$$\lambda = \frac{\alpha\bar{\zeta}}{b\delta} P(|B_0 + \bar{\zeta}Z| \geq \alpha\bar{\zeta}), \quad (2.21)$$

it follows by Theorem 2 of Bradic (2016) that

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\{\widehat{\beta}_{\mathrm{C},j}(\lambda) - \widehat{\beta}_{\mathrm{C},(t),j}(\alpha)\}^2 = 0 \text{ a.s.} \quad (2.22)$$

The convergence in (2.22) explicitly connects the two composite estimators: one estimator uses an explicit l_1-regularization as in (2.3), the other estimator is obtained via the RAMP algorithm.

Similar results can be found in Huang (2020, Theorem 2.2) for a generalized AMP algorithm with non-negative convex loss function, and in Bayati and Montanari (2011b, Theorem 1.8) for the AMP algorithm with least squares loss function.

For the model averaging estimator we use such an equivalence for estimation with each separate loss function ρ_k, $k = 1, \ldots, K$. When using explicit l_1-regularization as in (2.1) with the regularization constants λ_k

matching as in (2.21) the values $\theta_k = \alpha_k \bar{\zeta}$, for $k = 1, \ldots, K$ that are used in the RAMP algorithm, again Theorem 2 of Bradic (2016) applies. It hence follows that

$$\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \{\widehat{\beta}_{\mathrm{MA},j}(\boldsymbol{\lambda}) - \widehat{\beta}_{\mathrm{MA},(t),j}(\boldsymbol{\alpha})\}^2 = 0, \text{a.s.}$$

2.5 Theoretical contributions

2.5.1 Asymptotic mean squared error

We first define the asymptotic mean squared error as

$$\mathrm{AMSE}(\widehat{\beta}_{(t)}, \beta) = \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} (\widehat{\beta}_{(t),j} - \beta_j)^2. \tag{2.23}$$

Combining (2.18) and (2.15), we obtain

$$\begin{aligned} \mathrm{AMSE}(\widehat{\beta}_{(t)}, \beta) &= \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \left(\eta(\widetilde{\beta}_{(t-1),j}; \theta_{(t-1)}) - \beta_j\right)^2 \\ &\overset{a.s.}{=} E[\{\eta(B_0 - \bar{\zeta}_{(t-1)} Z; \theta_{(t-1)}) - B_0\}^2], \end{aligned} \tag{2.24}$$

which corresponds to Bradic (2016, Eq.(3.4)) with $\widetilde{\beta}_{(t),j}$ the debiased estimator in (2.13).

In Section 2.4, we defined the empirical state evolution parameter $\bar{\zeta}^2_{\mathrm{emp,(t)}}$, and we described the connections between the empirical updates in Algorithm 1 and the theoretical state evolution recursion, which connects to the theoretical expression of the AMSE. While Algorithm 1 and the theoretical state evolution recursion involve only a single estimator, the model-averaged estimator, on the other hand, is the weighted sum of K such estimators $\widehat{\beta}_k$, $k = 1, \ldots, K$, each obtained by Algorithm 1. Consequently, the estimators $\widehat{\beta}_k$, $k = 1, \ldots, K$ are correlated.

Lemma 2.1 extends Theorem 2 in Bayati and Montanari (2011a) and (3.16) in Lemma 1(b) in Bayati and Montanari (2011a) to the almost sure

convergence of the product for any two recursions among K paralleled recursions. All proofs are contained in Section 2.10.

Lemma 2.1. *Let the sequences of design matrices $\{X(p)\}$, coefficient vectors $\{\beta(p)\}$, error vectors $\{\varepsilon(p)\}$, initial condition vectors $\{q_0(p)\}$ be the common sequences for K recursions satisfying conditions (A1)–(A4) in Section 2.9. Let $\{\bar{\sigma}^2_{k,(t)}, \bar{\zeta}^2_{k,(t)}\}$ be defined uniquely by the recursions in (2.16) and (2.17). These are the state evolution parameters for the kth estimation with initialization $\bar{\sigma}^2_{k,(0)} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} q^2_{(0),i}/\delta$. Then Lemma 1 in Bayati and Montanari (2011a) holds individually for each of the K recursions; additionally, for all pseudo-Lipschitz functions $\tilde{\psi}_{\mathrm{c}} : \mathbb{R}^{t+2} \to \mathbb{R}$ of order κ_{c} for some $1 \le \kappa_{\mathrm{c}} \le \kappa/2$ with κ as in Condition (A4) and t a natural number larger than or equal to 0,*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \tilde{\psi}_{\mathrm{c}}(h_{k_1,(1),j}, \ldots, h_{k_1,(t+1),j}, \beta_j)\tilde{\psi}_{\mathrm{c}}(h_{k_2,(1),j}, \ldots, h_{k_2,(t+1),j}, \beta_j) \overset{a.s.}{=}$$
$$E[\tilde{\psi}_{\mathrm{c}}(\bar{\zeta}_{k_1,(0)}Z_{k_1,(0)}, \ldots, \bar{\zeta}_{k_1,(t)}Z_{k_1,(t)}, B_0)\tilde{\psi}_{\mathrm{c}}(\bar{\zeta}_{k_2,(0)}Z_{k_2,(0)}, \ldots, \bar{\zeta}_{k_2,(t)}Z_{k_2,(t)}, B_0)]$$

where $(Z_{k,(0)}, \ldots, Z_{k,(t)}) \sim \mathcal{N}(0, I_{t+1})$, $k = k_1, k_2$, is a $(t+1)$-dimensional zero-mean multivariate standard normal vector independent of B_0, ε; at iteration t, $(Z_{k_1,(t)}, Z_{k_2,(t)})$ is a bivariate standard normal vector with covariance not necessarily equal to zero.

Note that Algorithm 1 belongs to the general recursion in Bayati and Montanari (2011a), the initial condition takes $q_{(0)} = -\beta$ and the kth estimator calculated by Algorithm 1 takes $h_{k,(t+1)} = \beta - X^\top G(z_{k,(t)}; b_{k,(t)}) - \beta_{k,(t)}$.

We obtain at iteration t, for $k_1, k_2 \in \{1, \ldots, K\}$,

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p}(\widehat{\beta}_{k_1,(t),j} - \beta_j)(\widehat{\beta}_{k_2,(t),j} - \beta_j)$$
$$\overset{a.s.}{=} E\Big[\prod_{r=1}^{2}\big\{\eta(B_0 + \bar{\zeta}_{k_r,(t-1)}Z_{k_r}; \theta_{k_r,(t-1)}) - B_0\big\}\Big],$$

where Z_{k_1} and Z_{k_2} are possibly dependent standard normal random variables.

Since the estimators $\widehat{\beta}_{k_r}$, $r = 1, 2$ use the same design matrix, a correlation between Z_{k_1} and Z_{k_2} exists (see Corollary 2.2) and contributes to the correlation between $\widehat{\beta}_{k_1}$ and $\widehat{\beta}_{k_2}$. Using Lemma 2.1, we obtain the theoretical AMSE for the regularized model-averaged estimator.

Theorem 2.1. *Assume conditions (A1)–(A5) in Section 2.9 . At Algorithm 1's iteration step t for the estimator $\widehat{\beta}_{k,(t)}$, for each $k = 1, \ldots, K$, and for a weight vector $w = (w_1, \ldots, w_K)^\top$, the model-averaged estimator $\widehat{\beta}_{\mathrm{MA},(t)} = \sum_{k=1}^{K} w_k \widehat{\beta}_{k,(t)}$ has asymptotic mean squared error*

$$\begin{aligned} \mathrm{AMSE}(\widehat{\beta}_{\mathrm{MA},(t)}, \beta) &= \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} (\widehat{\beta}_{\mathrm{MA},(t),j} - \beta_j)^2 \\ &= \lim_{p\to\infty} w^\top \Sigma_{0,(t)}(p) w \overset{a.s.}{=} w^\top \Sigma_{(t)} w \end{aligned} \tag{2.25}$$

where $\Sigma_{0,(t)}(p)$ is a $K \times K$ matrix with (k_1, k_2)th component

$$(\Sigma_{0,(t)})_{(k_1,k_2)}(p) = p^{-1} \sum_{j=1}^{p} (\widehat{\beta}_{k_1,(t),j} - \beta_j)(\widehat{\beta}_{k_2,(t),j} - \beta_j); \tag{2.26}$$

similarly, $\Sigma_{(t)}$ is a $K \times K$ matrix with the (k_1, k_2)th component

$$(\Sigma_{(t)})_{(k_1,k_2)} = E\Big[\prod_{r=1}^{2} \{\eta(B_0 + \bar{\zeta}_{k_r,(t-1)} Z_{k_r}; \theta_{k_r,(t-1)}) - B_0\} \Big].$$

Since the AMSE expression of the regularized model-averaged estimator is a quadratic function of the weight vector w, Corollary 2.1 readily provides the lower bound of the AMSE as well as the weight vector reaching this lower bound. We define the K-vector $\mathbf{1}_K$ to consist of ones only.

Corollary 2.1. *Constraining the weights to sum to one, the lower bound of the AMSE at iteration t for the model-averaged estimator as in (2.25) is equal to $\left(\mathbf{1}_K^\top (\Sigma_{(t)})^{-1} \mathbf{1}_K\right)^{-1}$. This lower bound is attained for the theoretical*

optimal weights $w_{\mathrm{MA}} = \big(\Sigma_{(t)}\big)^{-1}\mathbf{1}_K\big(\mathbf{1}_K^\top(\Sigma_{(t)})^{-1}\mathbf{1}_K\big)^{-1}$.

2.5.2 Estimating optimal weights

The expression of the core matrix $\Sigma_{(t)}$, which is the limit matrix for $n, p \to \infty$, contains the random variable B_0 which satisfies Condition (A2) in Section 2.9. Likewise, $\Sigma_{0,(t)}$ which is the limit matrix for fixed p while $n \to \infty$, contains the true coefficient β (see (2.26)). In practice, neither the true coefficient vector β, nor the random variable B_0 is known. To make practical use of the expressions of the AMSE, we derive an estimator of the matrix $\Sigma_{0,(t)}$ relying only on sequences generated in Algorithm 1.

Model-averaged estimator

Before deriving the estimator of the AMSE for the model-averaged estimator, we first define $\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t)}$ which is an estimator of the parameter $\bar{\zeta}_{(k_1,k_2),(t)}$, a quantity similar to the state evolution parameter $\bar{\zeta}^2_{k,(t)}$, which records the covariance between the unbiased sequences $\widetilde{\beta}_{k_1,(t)}$ and $\widetilde{\beta}_{k_2,(t)}$ generated in (2.13) in Algorithm 1 when $p \to \infty$. Since model-averaged estimators combine estimators constructed from the same data into one weighted average, the correlation between $\widehat{\beta}_{k_1}$ and $\widehat{\beta}_{k_2}$ is needed to understand the AMSE of the model-averaged estimator.

Notice that the unbiasedness of the sequence $\widetilde{\beta}_{k,(t)}$ follows from the argument that $\widetilde{\beta}_{k,j,(t)}$ converges weakly to $B_0 + \bar{\zeta}_{k,(t)}Z_k$ when $p \to \infty$, while assigning $1/p$ point mass to each entry of the vector. Then, $\widetilde{\beta}_{k,j,(t)}|(B_0 = \beta_j) \sim N(\beta_j, \bar{\zeta}^2_{k,(t)})$ for large p indicates first that $\widetilde{\beta}_{k,j,(t)}$ centers at β_j ensuring the unbiasedness. Also the vector $\widetilde{\beta}_{k,(t)}$ is Gaussian distributed. By applying the soft-thresholding function η on $\widetilde{\beta}_{k,j,(t)}$ in Lemma 2.4, we avoid the true coefficient vector β in $\Sigma_{0,(t)}$ resulting in a Stein-type risk estimator requiring only arguments from Algorithm 1. A Gaussianity argument has also been used in Bayati and Montanari (2011b); Bayati et al. (2013); Mousavi et al. (2013, 2018) to derive a similar Stein-type risk estimator for the Lasso. Details can be found in Section 2.10.2. The bias of the esti-

mator $\widehat{\beta}_{k,(t)}$ is introduced in Algorithm 1 by applying the soft-thresholding function componentwise to the unbiased sequence $\widetilde{\beta}_{k,(t)}$.

Corollary 2.2. *Assume conditions (A1)–(A5) in Section 2.9. For any $k_1, k_2 = 1, \ldots, K$, at iteration t,*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p}(\widetilde{\beta}_{k_1,(t),j} - \beta)(\widetilde{\beta}_{k_2,(t),j} - \beta) \overset{a.s.}{=} \bar{\zeta}_{k_1,(t)}\bar{\zeta}_{k_2,(t)} Cov(Z_{k_1}, Z_{k_2}),$$

where $\bar{\zeta}_{k,(t)}, k = k_1, k_2$ is the state evolution parameter corresponding to $\widehat{\beta}_k$.

Corollary 2.2 indicates both the existence and a feasible estimation of the covariance between Z_{k_1} and Z_{k_2}. As an estimator for

$$\bar{\zeta}_{(k_1,k_2),(t)} = \bar{\zeta}_{k_1,(t)}\bar{\zeta}_{k_2,(t)}\mathrm{Cov}(Z_{k_1}, Z_{k_2})$$

we define

$$\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t)} = \frac{1}{p-1}\sum_{j=1}^{p}\big(\widetilde{\beta}_{k_1,(t),j} - \frac{1}{p}\sum_{j=1}^{p}\widetilde{\beta}_{k_1,(t),j}\big)\big(\widetilde{\beta}_{k_2,(t),j} - \frac{1}{p}\sum_{j=1}^{p}\widetilde{\beta}_{k_2,(t),j}\big). \tag{2.27}$$

We now state an unbiased estimator for the matrix $\Sigma_{0,(t)}$, and a consistent estimator for the matrix $\Sigma_{(t)}$ upon convergence of Algorithm 1.

Theorem 2.2. *Assume conditions (A1)–(A5) in Section 2.9, and that the state evolution parameter in (2.14) satisfies $\bar{\zeta}^2_{\mathrm{emp},(t)} - \bar{\zeta}^2_{\mathrm{emp},(t-1)} = o(1)$. For any $k_1, k_2 = 1, \ldots, K$, define*

$$\begin{aligned}(\widehat{\Sigma}_0)_{(k_1,k_2),(t)} = &-\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t-1)} \\ &+ \frac{1}{p}\sum_{j=1}^{p}\prod_{r=1}^{2}\{\eta(\widetilde{\beta}_{k_r,(t-1),j}; \theta_{k_r,(t-1)}) - \widetilde{\beta}_{k_r,(t-1),j}\} \\ &+ \bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t-1)} \cdot \frac{1}{p}\sum_{j=1}^{p}\sum_{r=1}^{2} I\{|\widetilde{\beta}_{k_r,(t-1),j}| \geq \theta_{k_r,(t-1)}\},\end{aligned}$$

with $\widetilde{\beta}_{k_1,(t-1)}$, $\widetilde{\beta}_{k_2,(t-1)}$ *in (2.13) Then,* $(\widehat{\Sigma}_0)_{(k_1,k_2),(t)}$ *is an unbiased estimator of component* (k_1, k_2) *of the matrix* $\Sigma_{0,(t)}$ *at iteration* t*. Further,* $(\widehat{\Sigma}_0)_{(k_1,k_2),(t)}$ *is a consistent estimator of the matrix* $\Sigma_{(t)}$ *in Theorem 2.1.*

This new estimator can be compared to the estimator used in Bayati et al. (2013, Def. 2) and Mousavi et al. (2018, Eq.(9)) for the case of a single estimator ($K = 1$). The proof of Theorem 2.2, see Section 2.10.2 uses Stein's lemma (see Lemma 2.4) to estimate the covariances that appear in the matrix $\Sigma_{0,(t)}$. The soft-thresholding function $\eta(\cdot;\theta)$ that appears in the estimator $\widehat{\Sigma}_{0,(t)}$ links the estimator $\widehat{\beta}_k$ to the estimator $\widetilde{\beta}_k$. The proof also uses the joint asymptotic normality of the jth components of the vectors $\widetilde{\beta}_{k_1}$ and $\widetilde{\beta}_{k_2}$. The obtained estimator for $\Sigma_{0,(t)}$ in the case $K > 1$ is nontrivial and new to the literature.

Estimated AMSE-type optimal weights for the model-averaged estimator are obtained by using the estimator $\widehat{\Sigma}_{0,(t)}$ at the final iteration in Theorem 2.2. In combination with the sum-to-one constrained weights this gives the estimated weights that minimize the estimated AMSE for the model averaged estimator

$$\widehat{w}_{\mathrm{MA}} = \big(\widehat{\Sigma}_{(t)}\big)^{-1}\mathbf{1}_K\big(\mathbf{1}_K^\top(\widehat{\Sigma}_{(t)})^{-1}\mathbf{1}_K\big)^{-1}.$$

When additional constraints such as positivity are needed, the optimal weights no longer have an explicit formula, but they are straightforward to compute, see (2.30).

Composite estimator

The AMSE of a composite estimator can be obtained from Theorem 2.1 as a special case, treating the composite loss function as a single loss function with weight one, thus $\rho_{\mathrm{C}} = \sum_{k=1}^K w_k\rho_k$ as in (2.3). At iteration t,

$$\Sigma_{(t)} = E[\{\eta(B_0+\bar{\zeta}_{(t-1)}Z;\theta_{(t-1)})-B_0\}]^2, \text{ and } \Sigma_{0,(t)} = p^{-1}\sum_{j=1}^p(\widehat{\beta}_{(t),j}-\beta_j)^2.$$

The matrices $\Sigma_{(t)}, \Sigma_{0,(t)}$ are now real numbers and coincide with the AMSE of the estimator in (2.24). We obtain the corresponding estimator for the AMSE

$$\widehat{\Sigma}_{\mathrm{C},0} = \widehat{\mathrm{AMSE}}_{\mathrm{C}}(w) \tag{2.28}$$
$$= -\bar{\zeta}^2_{\mathrm{emp}}(w) + \frac{1}{p}\sum_{j=1}^{p}\Big[\big\{\eta(\widetilde{\beta}_j(w);\theta) - \widetilde{\beta}_j(w)\big\}^2 + 2\bar{\zeta}^2_{\mathrm{emp}}(w) I\{|\widetilde{\beta}_j(w)| \geq \theta\}\Big].$$

For the single loss function ρ_{C}, the estimator of AMSE in (2.28) can be compared to the Stein-type estimator that has been obtained in Definition 2 in Bayati et al. (2013) for the AMP algorithm using the least squares loss, which is a special case of Algorithm 1.

Finding optimal weights is more complicated for the composite estimator. Indeed, while the model-averaged estimator has an AMSE which is a quadratic function in the weights, see (2.25), the composite estimator and its AMSE depend on the weights in a highly nonlinear fashion; e.g., observe that the soft-thresholding function in (2.28) depends on w.

Therefore, optimization of the estimated AMSE with respect to the weights proceeds numerically;

$$w_{\mathrm{C},1} = \arg\min_{w} \widehat{\mathrm{AMSE}}_{\mathrm{C}}(w).$$

See Section 2.6.2 for more details.

2.5.3 The case of dense (non-sparse) linear models with $n/p \to \delta \geq 1$: asymptotic variance optimality

Donoho and Montanari (2016) and El Karoui et al. (2013) showed that the asymptotic variance of the M-estimators in the case where $p, n \to \infty$ and $n/p \to \delta \in [1, \infty)$ contains an extra Gaussian component. Recently, Lei et al. (2018) obtained the coordinate-wise asymptotic normality of regression M-estimators in the moderate p/n regime for a fixed design matrix. In the sparse high dimensional linear model setting where $\delta \in (0, 1)$, it was

shown that the sequence $\widetilde{\beta}_{(t)}$ in (2.13) follows for the Lasso estimator (Bayati et al., 2013) a similar normal distribution with the variance containing an extra Gaussian component. The above-mentioned literature focuses on the asymptotics for a single M-estimator, we extend the asymptotic result to the model-averaged estimator. In this section, we only characterize the asymptotic variance of the model-averaged estimator for dense linear models with $n/p \to \delta \geq 1$, following Donoho and Montanari (2016).

Under the dense linear model with $n \geq p$, the soft-thresholding function $\eta(\cdot;\theta)$ is replaced by the identity function and the ratio $\omega = E[\|B_0\|_0] = 1$. Consequently, Algorithm 1 is adjusted to estimate

$$\widehat{\beta}_k = \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \sum_{i=1}^{n} \rho_k(Y_i - X_i\beta) \Big\},$$

where β is dense. It is trivial to show that Algorithm 1 still belongs to the general recursion in Bayati and Montanari (2011a). For a single estimator at iteration t denoted as $\widehat{\beta}_{k,(t)}$, the two state evolution parameters $\bar{\zeta}^2_{k,(t)}$ and $\bar{\sigma}^2_{k,(t)}$ coincide and Theorem 4.1 in Donoho and Montanari (2016) holds.

Theorem 2.3. *Assume conditions (A1)–(A5) in Section 2.9 . Let $n/p \to \delta \geq 1$ when $n, p \to \infty$. For the asymptotic variance of the model-averaged estimator $\widehat{\beta}_{\mathrm{MA}}$ holds that*

$$\lim_{n,p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \mathrm{Var}(\widehat{\beta}_{\mathrm{MA},j}) \overset{a.s}{=} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \mathrm{Cov}(Z_{k_1}, Z_{k_2}) \prod_{r=1}^{2} \{w_{k_r} V^{1/2}(\widetilde{G}_{k_r}; \widetilde{F}_{k_r})\} \tag{2.29}$$

for differentiable $\widetilde{G}$, where $V(\widetilde{G}_k; F_k) = (\int \widetilde{G}_k^2 dF_k)/(\int \partial_1 \widetilde{G}_k dF_k)^2$ denotes the Huber asymptotic variance formula for M-estimators.
For non-differentiable $\widetilde{G}$, we replace V in (2.29) *by the consistent estimator $\widehat{V}(\widetilde{G}_k; F_k) = (\int \widetilde{G}_k^2 dF_k)/\widehat{\nu}(b_k)^2$. The extra Gaussian component is identified in the convolution of the regression noise distribution and a Gaussian distribution: $\widetilde{F}_k = F_\varepsilon \star N(0, \bar{\zeta}_k^2)$.*

Recall that the componentwise empirical distribution of $\widehat{\beta}_k(p)$, when $p \to \infty$, converges weakly to $B_0 + \bar{\zeta}_k Z_k$ following Bayati and Monta-

nari (2011a) and Donoho and Montanari (2016). Then for large p, while the iteration $t \to \infty$, $\widehat{\beta}_k(p) \sim N(\beta, \zeta_k^2 I_p)$ (Donoho and Montanari, 2016; Mousavi et al., 2013) with I_p the $p \times p$ identity matrix. The (k_1, k_2)th component of the empirical variance matrix is denoted by $\big(\Sigma_{\text{emp}}(p)\big)_{(k_1,k_2)} = p^{-1}\sum_{j=1}^{p}(\widehat{\beta}_{k_1,j} - \beta_j)(\widehat{\beta}_{k_2,j} - \beta_j)$, which is unbiasedly estimated by

$$\big(\widehat{\Sigma}_{\text{emp}}(p)\big)_{(k_1,k_2)} = \sum_{j=1}^{p}(\widehat{\beta}_{k_1,j} - \frac{1}{p}\sum_{j=1}^{p}\widehat{\beta}_{k_1,j})(\widehat{\beta}_{k_2,j} - \frac{1}{p}\sum_{j=1}^{p}\widehat{\beta}_{k_2,j})/(p-1).$$

Note that this estimator coincides with (2.27) for the special case that $n \geq p$ and the soft-thresholding function is replaced by the identity function.

2.6 Computational details

2.6.1 Regularized model-averaged quantile estimation

The estimation of the quantile $u_{\tau_k} = F_\varepsilon^{-1}(\tau_k)$ follows a two-step procedure.

(i) Obtain an initial slope estimate $\widehat{\beta}_{\text{init}}$ and calculate the residuals. Example initial slope estimates are the Lasso or regularized quantile estimation with a single quantile level.

(ii) For $k = 1, \ldots, K$, estimate the quantile intercepts $\widehat{u}_{\tau_k}$ by taking the corresponding $\tau_k \times 100\%$ quantile of the residuals from the previous step.

The regularized model-averaged estimator is obtained by averaging over K paralleled estimators. See Algorithm 2 for the pseudo code, of which the core is Algorithm 1 in which the effective score function G is that for a single quantile loss function using $K = 1$, see also Example 2 in Bradic (2016). In our numerical work the upper bound for the number of iteration steps T is set to be 50 in both the simulation and the data analysis sections. With $K = 1$, this algorithm applies to the regularized composite estimator too.

Algorithm 2: RAMP algorithm for K paralleled estimations with tuned α's

Function `KparallelRAMP(`K`)`**:**
  for k *in* $\{1, \ldots, K\}$ **do**
    Initialization: $\widehat{\beta}(\alpha_{k,\text{opt}}) \leftarrow 0 \in \mathbb{R}^p$, $\widetilde{\beta}_k(\alpha_{k,\text{opt}}) \leftarrow 0 \in \mathbb{R}^p$ and $\text{AMSE}(\widehat{\beta}_k(\alpha_{k,\text{opt}}); \beta) \leftarrow 0$
    for α in candidate set $\mathcal{A}$ **do**
      `singleRAMP`(α) in Algorithm 1
      if $\text{AMSE}(\widehat{\beta}_k(\alpha); \beta) \leq \text{AMSE}(\widehat{\beta}_k(\alpha_{k,\text{opt}}); \beta)$ **then**
        $\widehat{\beta}_k(\alpha_{k,\text{opt}}) \leftarrow \widehat{\beta}_k(\alpha)$, $\widetilde{\beta}_k(\alpha_{k,\text{opt}}) \leftarrow \widetilde{\beta}_k(\alpha)$, $\text{AMSE}(\widehat{\beta}_k(\alpha_{k,\text{opt}}); \beta) \leftarrow \text{AMSE}(\widehat{\beta}_k(\alpha); \beta)$
      end
    end
  end
  return $(\widehat{\beta}_1(\alpha_{1,\text{opt}}), \ldots, \widehat{\beta}_K(\alpha_{K,\text{opt}}))$, $(\widetilde{\beta}_1(\alpha_{1,\text{opt}}), \ldots, \widetilde{\beta}_K(\alpha_{K,\text{opt}}))$, *and* $(\text{AMSE}(\widehat{\beta}_1(\alpha_{1,\text{opt}}); \beta), \ldots, \text{AMSE}(\widehat{\beta}_K(\alpha_{K,\text{opt}}); \beta))$

AMSE refers to the estimated version. The $\widetilde{\beta}_k$s are recorded for calculating the weights in Corollary 2.1.

The tuning parameter α of Algorithm 2 controls the sparsity of the estimators and requires a tuning procedure to choose it in practice. In Section 2.7, we consider the one dimensional Golden-section search algorithm (Kiefer, 1953) for tuning the value α in the range $[\alpha_{\min}, \alpha_{\max}]$ that minimize the estimated MSE of $\widehat{\beta}$ using the estimator derived in Section 2.5.2. The upper bound $\alpha_{\max}$ is chosen to be 2.3 for the simulations and data analysis. The lower bound $\alpha_{\min}$ in the data analysis follows the lower bound in Proposition 9.2 in Eldar and Kutyniok (2012) and is chosen to be the unique non-negative solution to the equation $(1+\alpha^2)\Phi(-\alpha)-\alpha\phi(\alpha) = \delta/2$, where $\phi(x)$ and $\Phi(x)$ denote the p.d.f and c.d.f of the standard normal distribution respectively. In the simulation study, the lower bound $\alpha_{\min}$ is chosen to be 1.3 for computational efficiency purpose, since the optimal tuning parameter for those settings was rarely less than 1.3.

2.6.2 Optimization of the weights

To obtain the regularized model-averaged quantile estimations with the AMSE-type weight derived in Corollary 2.1, we follow the following procedure:

(i) Obtain optimally tuned paralleled regularized quantile estimates, see (2.13), $(\widehat{\beta}_{\tau_1}(\alpha_{1,\text{opt}}), \ldots, \widehat{\beta}_{\tau_K}(\alpha_{K,\text{opt}}))$, and the additional K estimates $(\widetilde{\beta}_{\tau_1}(\alpha_{1,\text{opt}}), \ldots, \widetilde{\beta}_{\tau_K}(\alpha_{K,\text{opt}}))$ from the converged iterations using Algorithm 2.

(ii) Estimate the AMSE-type optimal weight $\widehat{w}_{\text{MA},1}$ with constraints by

$$\widehat{w}_{\text{MA},1} = \arg \min_{w \geq 0, \mathbf{1}_K^\top w = 1} w^\top \widehat{\Sigma}_0 w \tag{2.30}$$

where the $K \times K$ matrix $\widehat{\Sigma}_0$ is the consistent estimator of Theorem 2.2.

(iii) Obtain the regularized model-averaged estimate (2.2) with the estimated AMSE-type optimal weight.

It is worth mentioning that $\widehat{w}_{\mathrm{MA},1}$ is a constrained version of w_{MA} attaining the lower bound of the AMSE in Corollary 2.1. $\widehat{w}_{\mathrm{MA},1}$ focuses on approximating the lower bound of the AMSE of the sparse coefficient vector β without assuming that the nonzero entries are selected perfectly; whereas another type of weight choice derived in Bradic et al. (2011) and Chapter 1 aims at the lower bound of the variance of the nonzero part of β by imposing the perfect selection assumption. A numerical comparison of these two types of weight choices is presented in Section 2.7.

To equip the regularized composite quantile estimator with the weight minimizing the estimated AMSE we cannot make use of an analytical solution to the weight minimization problem. Instead, a numerical search for a better weight choice in the neighbourhood of an initial weight proposal is employed. The basic idea is that the estimator $\widehat{\beta}_{\mathrm{C}}(w_{\mathrm{C}})$ is treated as a function of the weights. We propose a collection of candidate weight vectors in the neighbourhood of the weight chosen in the previous step. The weight for $\widehat{\beta}_{\mathrm{C}}(w_{\mathrm{C}})$ is updated in each step by the one having the lowest estimated AMSE, i.e.,

$$w_{\mathrm{C},1} = \arg\min_{w_{\mathrm{cand}}} \widehat{\mathrm{AMSE}}(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}}); \beta).$$

A more detailed search procedure is as follows.

(i) Propose a reasonable initial weight vector $w_{\mathrm{C,init}}$, e.g. the vector of equal weights; estimate $\widehat{\beta}_{\mathrm{C}}$ at the initial weight $w_{\mathrm{C,init}}$ and obtain the estimate of $\mathrm{AMSE}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{C,init}}); \beta\big)$.

(ii) Initiate the searching step calculator $s_{\mathcal{D}} = 0$, the candidate optimal weight $w_{\mathrm{C},1} = w_{\mathrm{C,init}}$, and the corresponding candidate minimum MSE

$$\mathrm{AMSE}(w_{\mathrm{C},1}) = \mathrm{AMSE}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{C,init}}); \beta\big)$$

estimated by the AMSE estimator in Theorem 2.2 for $K = 1$, the collection of the used weight vectors $\mathcal{V}_w = \{w_{\mathrm{C},1}\}$.

(iii) Propose a set of candidate weight vectors $\mathcal{V}_{w_{\mathrm{cand}}}$, which excludes

those recorded in the collection of the used weight vectors $\mathcal{V}_w$, in the neighbourhood of the current optimal weight $w_{\mathrm{C},1}$. Rules of proposing candidate weight vectors are user-decided; here, we consider a $(K-1)$-dimensional grid search centering at $w_{\mathrm{C},1}$.

(iv) Obtain the regularized composite quantile estimates at all candidate weight vectors in $\mathcal{V}_{w_{\mathrm{cand}}}$ by Algorithm 2. Update the used weight vector collection $\mathcal{V}_w$, searching step calculator $s_{\mathcal{V}} = s_{\mathcal{V}} + 1$, the candidate optimal weight $w_{\mathrm{C},1}$ by the weight with the lowest estimated AMSE in $\mathcal{V}_w = \{w_{\mathrm{C},1}\}$, and the candidate minimum AMSE value $\mathrm{AMSE}(w_{\mathrm{C},1})$.

(v) Stop the iteration if the searching step calculator $s_{\mathcal{V}} > S_{\mathcal{V}}$ or the candidate weight vector collection $\mathcal{V}_{w_{\mathrm{cand}}} = \emptyset$; otherwise repeat steps 3 and 4.

The pseudocode of the search procedure is stated in Algorithm 3.

2.7 Numerical results

2.7.1 Simulation study

In this section, we consider the following setup under the high dimensional linear model setting.

(i) Fix the dimension $p = 500$, the sample size $n = 250$, the ratio $\delta = 0.5$. The number of non-zero components s is taken to be 5 for the high-sparsity setting and 50 for the medium-sparsity setting; the non-zero part is generated from the Dirac distribution with a point mass equally distributed on -1 and 1, or a standard normal distribution.

(ii) In each repetition, we generate a new dataset by randomly generating a sensing matrix X, a coefficient vector β, and an error vector $\boldsymbol{\varepsilon}$. The components of the sensing matrix X are independent and generated from $N(0, 1/250)$.

Algorithm 3: Weight search for regularized composite estimator

Function `Weight Search`**:**

Initialization: Better weight recorder $w_{\mathrm{C},1} \leftarrow w_{\mathrm{C,init}}$, step calculator $s_{\mathcal{V}} \leftarrow 0$, MSE recorder $\mathrm{AMSE}(w_{\mathrm{C},1}) \leftarrow \widehat{\mathrm{AMSE}}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{C,init}}); \beta\big)$, and the collection of the used weight vectors $\mathcal{V}_w = \{w_{\mathrm{C},1}\}$.

while *searching step* $s_{\mathcal{V}} \leq S_{\mathcal{V}}$ *or candidate weight collection* $\mathcal{V}_{w_{\mathrm{cand}}} = \emptyset$ **do**

(i) Propose a new $\mathcal{V}_{w_{\mathrm{cand}}}$ in the neighbourhood of $w_{\mathrm{C},1}$. Rules of proposing candidate weight vectors are user-decided; here, we consider a $(K-1)$-dimensional grid search centering at $w_{\mathrm{C},1}$.

(ii) **for** w_{cand} *in* $\mathcal{V}_{w_{\mathrm{cand}}} \cap \mathcal{V}_w^{\complement}$ **do**

Estimate $\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}})$ and $\widehat{\mathrm{AMSE}}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}}); \beta\big)$

if $\widehat{\mathrm{AMSE}}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}}); \beta\big) < \mathrm{AMSE}(w_{\mathrm{C},1})$ **then**

$w_{\mathrm{C},1} \leftarrow w_{\mathrm{cand}}$,

$\mathrm{AMSE}(w_{\mathrm{C},1}) \leftarrow \widehat{\mathrm{AMSE}}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}}); \beta\big)$

end

end

(iii) Update $s_{\mathcal{V}} = s_{\mathcal{V}} + 1$.

end

return $w_{C,1}$, $\widehat{\beta}(\alpha_{\mathrm{opt}}, w_{\mathrm{C},1})$, *and* $\widehat{\mathrm{AMSE}}(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}}); \beta)$

A possible initial weight vector $w_{\mathrm{C,init}}$ is the vector of equal weights or the weight proposed in Bradic et al. (2011); $\widehat{\beta}_C$ is estimated by Algorithm 2, and $\mathrm{AMSE}(\widehat{\beta}_{\mathrm{C}}; \beta)$ is estimated by (2.28).

(iii) As error distributions, we take the standard normal $N(0,1)$, student-t with degrees of freedom 3, and the mixture of normal distributions $0.5N(0,1) + 0.5N(5,9)$; errors generated in Step 2 are centered and rescaled to have standard deviation 0.2.

The objective is to compare the performance of the regularized model-averaged estimator and the composite estimator with different weights, with emphasis on the weights where the selection uncertainty is taken into account. The simulation is repeated 500 times for each setup. For both the regularized model-averaged and composite quantile estimator, the weights considered are

(1) The estimated AMSE-type weights (i.e. $w_{\mathrm{MA},1}$ for the model-averaged quantile estimator and $w_{\mathrm{C},1}$ for the composite quantile estimator).

(2) The estimated weights based on minimizing the asymptotic variance of the estimators of only the active set of coefficients, denoted by $w_{\mathrm{MA},2}$ (Bloznelis et al., 2019) and $w_{\mathrm{C},2}$ (Bradic et al., 2011) where, with the (k_1,k_2)th component of A being $A_{k_1,k_2} = \min(\tau_{k_1},\tau_{k_2})\{1 - \max(\tau_{k_1},\tau_{k_2})\}$, $A_\varepsilon = \mathrm{diag}(f_\varepsilon(u_{\tau_1}),\ldots,f_\varepsilon(u_{\tau_K}))$, and $a_\varepsilon = (f_\varepsilon(u_{\tau_1}),\ldots,f_\varepsilon(u_{\tau_K}))^\top$

$$w_{\mathrm{MA},2} = \arg \min_{w, \mathbf{1}_K^\top w=1, w_k \geq 0} \left\{ w^\top A_\varepsilon^{-1} A A_\varepsilon^{-1} w \right\} \tag{2.31}$$

and

$$w_{\mathrm{C},2} = \arg \min_{w, a_\varepsilon^\top w=1, w_k \geq 0} \left[w^\top A w \right].$$

Only considering the variance has been the standard practice so far.

(3) Equal weights $1/K$ for each component.

The number of quantiles K for both estimators is taken to be 3, with quantile levels $25\%, 50\%, 75\%$.

We present the empirical MSEs of the abovementioned estimators for estimation of three vectors of coefficients. First, we consider the estimator

of the subvector of the full coefficient vector that consists of only the non-zero true coefficients, we refer to this as the "non-zero part". Second, we consider the estimator of the subvector of the coefficients that are truly zero. This is referred to as the "zero part". Third, we consider the full vector of estimated coefficients. Note that some truly zero coefficients might have a non-zero estimate, while some truly non-zero coefficients might be estimated as zero. For each of these three vectors, "parts", we compare the estimated values with the true values to get

$$\mathrm{MSE}(\widehat{\beta}_{\mathrm{part}}) = \sum_{j_{\mathrm{part}}=1}^{p_{\mathrm{part}}} (\widehat{\beta}_{j_{\mathrm{part}}} - \beta_{j_{\mathrm{part}}})^2 / p_{\mathrm{part}}$$

for the appropriate part of the full vectors. Results for the regularized model-averaged quantile estimator with different weights are presented in Table 2.1. We observe that the model-averaged quantile estimator using the weight in (2.30) has lower MSEs for estimating the non-zero part of β and for the full vector β, and this for t_3 and the mixture of normally distributed errors in the high-sparse case where the number of non-zero components $s = 5$. Using equal weights leads to a fair performance of the model-averaged quantile estimator especially for estimating the all-zero part of β. The Lasso estimator is considered as the baseline comparison which from Table 2.1 seems to have a competitive performance, especially in the medium sparsity settings. However, the Lasso mostly gives over-sparse estimations, which can be observed in the top half of Table 2.2 summarizing the averaged true positive (TP) and true negative (TN) recovery rates which are defined as

$$\mathrm{TP\ (TN)} = \frac{\text{number of correctly identified as non-zeros (zeros)}}{\text{number of true non-zeros (zeros)}}$$

The Lasso has consistently the highest TN rate and mostly the lowest TP rate. Further, while increasing the standard deviation of the errors, the overly-sparse estimation of the Lasso becomes clearer, i.e., Lasso gives sparser estimations and becomes all-zeros eventually. The regularized

model-averaged estimator with equal weights mostly has the highest TP rate, except for the medium sparsity settings where the non-zero part of the true regression coefficient is sampled from a Dirac distribution at -1 and 1, and the errors are sampled from $N(0,1)$ or $0.5N(0,1)+0.5N(5,9)$. The model-averaged estimator with the weight in (2.30) has consistently the second highest TN rate.

Since there is no analytical expression for the selection incorporated weight of the regularized composite quantile estimator $w_{\mathrm{C},1}$, the choice of weights can only be determined numerically by an exhaustive search. To reduce the searching time of the composite quantile estimator, we set the stopping criterion $S_{\mathcal{V}}$ to be 5 and only randomly select 4 points in the neighbourhood $\mathcal{V}_{w_{\mathrm{cand}}}$; the tuning parameter α of the soft-thresholding function is tuned once for the regularized composite quantile estimator with the weight $w_{\mathrm{C},2}$, then fixed thereafter.

Table 2.3 summarizes the empirical MSEs of the regularized composite quantile estimator with different weights. Since the tuning parameter α is selected for $w_{\mathrm{C},2}$ and a fixed tuning parameter is used for obtaining the regularized composite quantile estimates with other weights, it is not surprising that using $w_{\mathrm{C},2}$ leads to lower MSEs in most cases. However, it is worth noticing that using equal weights, while α is not optimally tuned, leads to fair performances of the regularized composite quantile estimator. The Lasso estimator consistently has the lowest empirical MSEs recovering the all-zero parts, through the largest empirical MSEs recovering the non-zero parts. This is caused by overly sparse estimations of the Lasso, which is indicated in the bottom half of Table 2.2. The regularized composite estimator with locally optimized $w_{\mathrm{C},1}$ consistently has the highest TP rate, and second highest TN rate among all competitors, except the TN rate for t_3 distributed errors and TP rate for $0.5N(0,1)+0.5N(5,9)$ distributed errors, while the non-zero parts of β are generated from Dirac distribution at -1 and 1.

Tables 2.2 and 2.3 illustrates that the regularized composite quantile estimator mostly improves the performance of regularized single quantile

f_ε	part	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$)	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{MA},2}}$)	MSE($\widehat{\beta}_{\widehat{w}^{\mathrm{MA}}_{\mathrm{eq}}}$)	MSE($\widehat{\beta}_{\mathrm{Lasso}}$)
Non-zero part of β: **Dirac distribution at -1 and 1** ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)					
$\mathbf{s=5}$	Non-zero	0.312	0.299	0.306	0.480
$N(0,1)$	Zero (‡)	6.812	6.436	5.276	0.526
	Full vec (†)	3.790	3.630	3.585	4.854
	Non-zero	0.167	0.168	0.182	0.681
t_3	Zero (‡)	4.051	3.579	3.041	0.106
	Full vec (†)	2.078	2.039	2.121	6.816
	Non-zero	0.247	0.355	0.314	0.412
$0.5N(0,1)+$	Zero (‡)	4.593	7.516	5.418	0.791
$0.5N(5,9)$	Full vec (†)	2.920	4.294	3.680	4.207
$\mathbf{s=50}$	Non-zero	0.487	0.502	0.526	0.376
$N(0,1)$	Zero (†)	5.498	4.438	3.675	5.710
	Full vec ($*$)	5.364	5.419	5.590	4.275
	Non-zero	0.399	0.427	0.452	0.384
t_3	Zero (†)	4.976	3.945	3.412	5.317
	Full vec ($*$)	4.436	4.630	4.832	4.318
	Non-zero	0.504	0.517	0.540	0.371
$0.5N(0,1)+$	Zero (†)	5.303	4.386	3.635	5.913
$0.5N(5,9)$	Full vec ($*$)	5.514	5.566	5.724	4.241
Non-zero part of β: $\mathbf{N(0,1)}$ ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)					
$\mathbf{s=5}$	Non-zero	0.206	0.197	0.203	0.378
$N(0,1)$	Zero (‡)	5.624	5.683	4.439	0.158
	Full vec (†)	2.613	2.537	2.465	3.800
	Non-zero	0.123	0.126	0.132	0.540
t_3	Zero (‡)	3.727	3.153	2.752	0.017
	Full vec (†)	1.601	1.574	1.590	5.403
	Non-zero	0.159	0.230	0.204	0.313
$0.5N(0,1)+$	Zero (‡)	3.788	6.723	4.720	0.348
$0.5N(5,9)$	Full vec (†)	1.969	2.970	2.511	3.162
$\mathbf{s=50}$	Non-zero	0.257	0.256	0.265	0.216
$N(0,1)$	Zero (†)	3.377	2.835	2.401	2.445
	Full vec ($*$)	2.870	2.819	2.870	2.376
	Non-zero	0.201	0.207	0.216	0.244
t_3	Zero (†)	2.831	2.275	2.009	1.859
	Full vec ($*$)	2.264	2.278	2.336	2.611
	Non-zero	0.275	0.278	0.285	0.220
$0.5N(0,1)+$	Zero (†)	3.571	2.945	2.536	2.530
$0.5N(5,9)$	Full vec ($*$)	3.076	3.049	3.077	2.423

Table 2.1: *The mean, over 500 simulation repetitions, of the empirical MSE of the regularized* model-averaged *quantile estimator with $K = 3$ for three error distributions. Empirical MSEs are calculated for the non-zero parts, all-zero parts, and the full vector of the true coefficient β. The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (top half), or standard normal distribution (bottom half). Smaller values of MSE among competitors indicate more accurate estimations.*

Non-zero part of β:			Dirac distribution at -1 and 1						N(0, 1)					
f_ε		rate	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},2}}$	$\widehat{\beta}^{\mathrm{MA}}_{\widehat{w}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},2}}$	$\widehat{\beta}^{\mathrm{MA}}_{\widehat{w}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$
$\mathbf{s=5}$ $N(0,1)$		TP	0.992	0.991	0.993	0.906	0.982	0.677	0.683	0.688	0.419	0.660
		TN	0.904	0.903	0.896	0.995	0.940	0.916	0.912	0.907	0.998	0.945
	t_3	TP	0.999	0.999	0.999	0.663	1.000	0.754	0.762	0.765	0.294	0.739
		TN	0.910	0.903	0.896	0.999	0.941	0.913	0.905	0.899	1.000	0.943
$0.5N(0,1)+$ $0.5N(5,9)$		TP	0.992	0.984	0.992	0.942	0.820	0.719	0.711	0.724	0.486	0.482
		TN	0.922	0.912	0.906	0.992	0.942	0.927	0.916	0.911	0.997	0.946
$\mathbf{s=50}$ $N(0,1)$		TP	0.836	0.847	0.854	0.889	0.548	0.647	0.658	0.666	0.619	0.600
		TN	0.843	0.830	0.823	0.868	0.606	0.843	0.832	0.807	0.883	0.894
	t_3	TP	0.892	0.899	0.904	0.882	0.453	0.696	0.706	0.715	0.590	0.622
		TN	0.833	0.816	0.807	0.873	0.707	0.839	0.822	0.811	0.931	0.842
$0.5N(0,1)+$ $0.5N(5,9)$		TP	0.822	0.837	0.843	0.892	0.531	0.633	0.643	0.650	0.621	0.539
		TN	0.845	0.833	0.826	0.864	0.601	0.846	0.834	0.827	0.911	0.834
Non-zero part of β:			$\widehat{\beta}_{\widehat{w}_{\mathrm{C},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},2}}$	$\widehat{\beta}^{\mathrm{C}}_{\widehat{w}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},2}}$	$\widehat{\beta}^{\mathrm{C}}_{\widehat{w}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$
f_ε		rate										
$\mathbf{s=5}$ $\mathrm{N}(0,1)$		TP	0.993	0.991	0.991	0.911	0.982	0.664	0.656	0.657	0.444	0.660
		TN	0.946	0.946	0.946	0.994	0.940	0.963	0.963	0.963	0.998	0.945
	t_3	TP	1.000	1.000	1.000	0.675	1.000	0.740	0.729	0.736	0.303	0.739
		TN	0.945	0.944	0.944	0.998	0.941	0.957	0.957	0.956	1.000	0.943
$0.5N(0,1)+$ $0.5N(5,9)$		TP	0.990	0.982	0.968	0.939	0.820	0.699	0.650	0.626	0.485	0.482
		TN	0.955	0.953	0.951	0.991	0.942	0.965	0.966	0.967	0.993	0.946
$\mathbf{s=50}$ $N(0,1)$		TP	0.903	0.895	0.891	0.883	0.548	0.669	0.668	0.665	0.613	0.600
		TN	0.824	0.823	0.824	0.872	0.606	0.848	0.846	0.846	0.916	0.894
	t_3	TP	0.939	0.933	0.934	0.871	0.453	0.724	0.722	0.720	0.590	0.622
		TN	0.819	0.817	0.820	0.879	0.707	0.835	0.834	0.835	0.929	0.842
$0.5N(0,1)+$ $0.5N(5,9)$		TP	0.886	0.881	0.875	0.891	0.531	0.651	0.648	0.642	0.615	0.539
		TN	0.827	0.824	0.825	0.867	0.601	0.855	0.852	0.852	0.914	0.834

Table 2.2: *The mean, over 500 simulation repetitions, of the true positive (TP) and true negative (TN) rate of the regularized* model-averaged *(top half) and* composite *(bottom half) quantile estimator with $K = 3$ for three error distributions. The TP and TN rates of the regularized single quantile estimator at quantile level 0.5 are presented in the 7th and 12th columns.* The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (left), or standard normal distribution (right). Larger values of TP and TN indicate a better identification power; the largest values among competitors are highlighted in green, whereas the second largest values are highlighted in yellow.

f_ε	part	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{C},1}}$)	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{C},2}}$)	MSE($\widehat{\beta}_{\widehat{w}^{\mathrm{C}}_{\mathrm{eq}}}$)	MSE($\widehat{\beta}_{\mathrm{Lasso}}$)	MSE($\widehat{\beta}_{0.5}$)
Non-zero part of β: **Dirac distribution at -1 and 1** ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)						
$\mathbf{s = 5}$	Non-zero	0.226	0.246	0.249	0.479	0.272
$N(0,1)$	Zero (‡)	6.566	7.119	7.199	0.571	11.641
	Full vec (†)	2.906	3.163	3.202	4.847	3.752
	Non-zero	0.122	0.135	0.133	0.674	0.142
t_3	Zero (‡)	3.782	4.148	4.109	0.112	5.650
	Full vec (†)	1.593	1.756	1.740	6.747	1.911
	Non-zero	0.184	0.246	0.311	0.420	0.461
$0.5N(0,1)+$	Zero (‡)	4.635	6.165	7.353	1.015	20.016
$0.5N(5,9)$	Full vec (†)	2.301	3.068	3.839	4.303	6.011
$\mathbf{s = 50}$	Non-zero	0.342	0.359	0.367	0.384	0.310
$N(0,1)$	Zero (†)	8.722	9.273	9.536	5.572	4.368
	Full vec ($*$)	4.203	4.423	4.524	4.339	4.308
	Non-zero	0.280	0.294	0.298	0.398	0.598
t_3	Zero (†)	7.026	7.511	7.618	5.159	19.415
	Full vec ($*$)	3.429	3.617	3.663	4.443	5.709
	Non-zero	0.358	0.376	0.385	0.375	0.318
$0.5N(0,1)+$	Zero (†)	9.099	9.644	10.007	5.750	4.301
$0.5N(5,9)$	Full vec ($*$)	4.403	4.631	4.751	4.268	4.360
Non-zero part of β: $\mathbf{N(0,1)}$ ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)						
$\mathbf{s = 5}$	Non-zero	0.157	0.173	0.175	0.363	0.177
$N(0,1)$	Zero (‡)	3.782	4.311	4.327	0.220	9.528
	Full vec (†)	1.946	2.153	2.178	3.655	2.555
	Non-zero	0.099	0.110	0.108	0.527	0.107
t_3	Zero (‡)	2.575	2.861	2.826	0.043	5.169
	Full vec (†)	1.245	1.382	1.360	5.273	1.492
	Non-zero	0.134	0.176	0.212	0.317	0.406
$0.5N(0,1)+$	Zero (‡)	2.787	3.772	4.550	0.476	23.379
$0.5N(5,9)$	Full vec (†)	1.615	2.135	2.568	3.213	4.469
$\mathbf{s = 50}$	Non-zero	0.174	0.182	0.186	0.216	0.230
$N(0,1)$	Zero (†)	4.925	5.220	5.369	2.299	3.970
	Full vec($*$)	2.181	2.289	2.342	2.371	2.571
	Non-zero	0.130	0.138	0.139	0.236	0.168
t_3	Zero (†)	3.849	4.077	4.152	1.888	2.887
	Full vec ($*$)	1.645	1.744	1.765	2.531	1.925
	Non-zero	0.189	0.198	0.205	0.214	0.236
$0.5N(0,1)+$	Zero (†)	5.149	5.507	5.738	2.386	3.984
$0.5N(5,9)$	Full vec ($*$)	2.354	2.476	2.562	2.353	2.776

Table 2.3: *The mean, over 500 simulation repetitions, of the empirical MSE of the regularized* composite *quantile estimator with $K = 3$ and the regularized single quantile estimator at quantile level 0.5 for three error distributions. Empirical MSEs are calculated for the non-zero parts, all-zero parts, and the full vector of the true coefficient β. The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (top half), or standard normal distribution (bottom half). Smaller values of MSE among competitors indicate more accurate estimations.*

estimator. For the same simulations settings, we compare the averaged empirical MSEs, true positive and true negative rates of the regularized composite quantile estimator, see Tables 2.2, column 7 and 12, and 2.3, column 7, with the single regularized quantile estimator at the median $\tau = 0.5$. For settings where $s = 5$, the composite quantile estimator clearly dominates the single quantile estimator for all three error distributions. For settings where $s = 50$, the composite estimator still mostly outperforms the single quantile estimator, except for the following cases : (1) the MSE for the non-zero and zero estimated subvector of β in settings where errors are generated from $N(0, 1)$ and $0.5N(0, 1) + 0.5N(5, 9)$ distribution and the true non-zero subvector of β is generated from a Dirac distribution; (2) TN rates in settings where errors are generated from $N(0, 1)$ and t_3 distribution and the true non-zero subvector of β is generated from $N(0, 1)$.

Convergence rates for the model-averaged and composite estimator, while setting the tolerance to be 10^{-6} for different error distributions are included in Table 2.4.

(%)	$s = 10$		$s = 50$	
f_ε	MAQR	CQR	MAQR	CQR
$N(0, 1)$	76.39	89.84	69.30	86.42
t_3	78.05	77.81	71.43	81.65
$0.5N(0, 1) + 0.5N(5, 9)$	77.00	86.38	71.43	85.08

Table 2.4: *Convergence rates of both regularized model-averaged and composite quantile estimator while the convergence tolerance is set to be* 10^{-6}*. The convergence rate of the regularized model-averaged estimator is calculated by including only those of which all single quantile components converge before 50 iterations.*

2.7.2 Data analysis

We consider the example audio wave file of a waveshape from Octave in the R package `signal`. The dataset is a list of 3 elements; the audio wave

sample is a vector of 17380 entries stored in the element "sound", the sample rate is 22050 Hz stored in the element "rate", and the resolution of the wave file is 16 bits recorded in the element "bits". To alleviate the computational burden of the signal compression and reconstruction, we only consider the signal from the 6145th entry to the 8192th entry of the original sound wave signal.

The preprocessing – discrete wavelet transform

Originated from the compressed sensing problem, the sparse linear model $Y = X\beta + \boldsymbol{\varepsilon}$ describes the image or signal compression. The s-sparse p-dimensional input signal β is first compressed by a known sensing matrix $X \in \mathbb{R}^{n\times p}$ with $n < p$; the compressed signal vector $X\beta \in \mathbb{R}^n$ can be corrupted by the noise ε with ε_i's i.i.d. via transmission. Notice that the p-dimensional input signal vector β is assumed to be s-sparse which is usually unsatisfied by signals expressed in the standard basis. To obtain the sparse representation of β in practice, an intermediate stage of expressing the natural non-sparse vector β^* in a proper orthonormal basis $\Psi^* = (\psi_1^*, \ldots, \psi_p^*)$ is required. Examples of such orthonormal basis include the orthonormal wavelet basis, the Fourier basis, etc. To perform the discrete wavelet transform, we use the R package `wavethresh`. The collection of the coefficients at all resolution levels is used for further compression.

The artificially corrupted compression

To imitate the compressed sensing process, we process the audio wave signal vector as follows:

(i) Perform the Daubechies' least asymmetric wavelet transform with 8 vanishing moments using the `wd` function in the R package `wavethresh` on the original signal $\beta^* \in \mathbb{R}^{2048}$ and obtain the corresponding wavelet coefficient vector $\beta \in \mathbb{R}^{2047}$ with $p = 2047$.

(ii) Randomly generate the sensing matrix X with i.i.d components $X_{ij} \sim$

$N(0, 1/n)$, where $n = \lfloor \delta' p \rfloor$ and δ' is the undersampling ratio chosen to be 0.5 here; compress the corresponding wavelet coefficients β by computing $X\beta$.

(iii) Corrupt the compressed wavelet coefficients by error vector ε with i.i.d. components ε_i following p.d.f f_ε; obtain the artificial observed signal vector $Y = X\beta + \boldsymbol{\varepsilon}$. Additionally, the standard normal $N(0, 1)$, student-t with 3 degrees of freedom, and the bimodal mixed normal $0.5N(0, 1) + 0.5N(5, 9)$ are used as the corruption error distributions; the errors are sampled according to the distributions first, then centered and rescaled to have standard deviation 0.03.

In practice, the artificial vector Y and the sensing matrix X are observed. The accurate recovery of the original wavelet coefficient vector β is of practical interest. To obtain an impression on the performance of the AMSE-type optimal weight, we generate the sensing matrix X under a fixed seed number which is set to be 1 in our case; then generate the error vector $\boldsymbol{\varepsilon}$ under various seed numbers. However, we only present the reconstructions under one seed for each setting in Section 2.7.2.

Signal recovery

To reconstruct the signal vector β expressed in the wavelet basis from the sensing matrix X and the observed compressed signal vector Y corrupted by potentially non-Gaussian distributed error $\boldsymbol{\varepsilon}$, we consider the regularized model-averaged and the composite quantile estimator weighting over three equally-spaced quantiles $(25\%, 50\%, 75\%)$ using equal weights, the oracle-type weights and the AMSE-type weights. The Lasso estimator is considered as the baseline comparison. Notice that the regularized estimates $\widehat{\beta}_{\mathrm{MA}}$ and $\widehat{\beta}_{\mathrm{C}}$ after reconstruction are the representations in the wavelet domain. To compare the accuracy of the reconstruction, we perform a back-transform on the estimates and obtain the corresponding signal vectors $\widehat{\beta}^*_{\mathrm{MA}}$ and $\widehat{\beta}^*_{\mathrm{C}}$ with representations in the natural basis.

Example reconstructions of the audio signal for $K = 3$ using the reg-

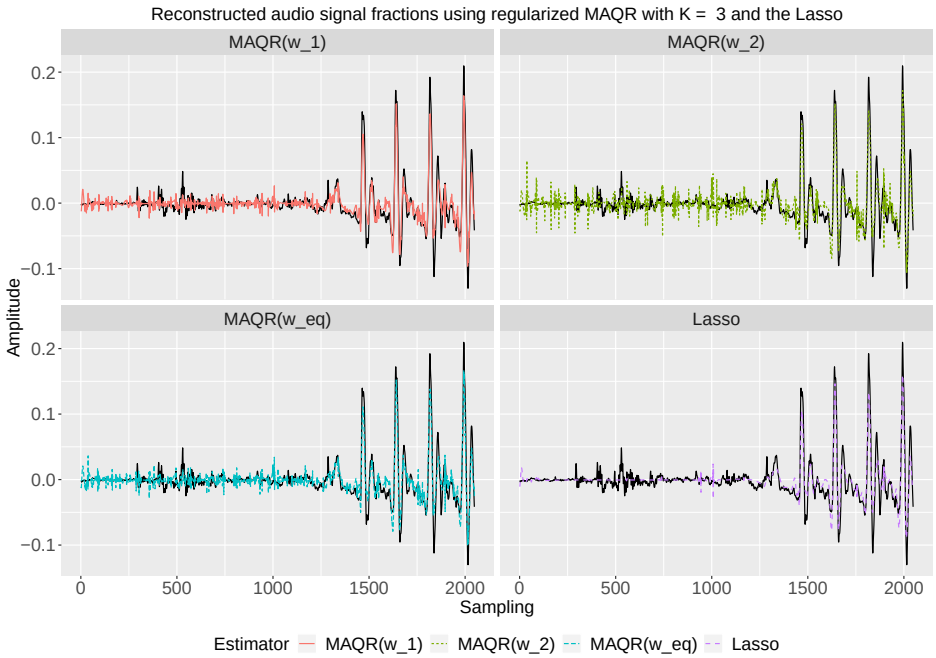


Figure 2.2: *Reconstructed audio signal from using the regularized* model-averaged *estimator with the estimated AMSE-type weights in (2.30), oracle-type optimal weights in (2.31), and equal weights. The original audio curve is depicted in black. The Lasso reconstruction is presented at bottom-right. The error used for corruption follows the mixture of normals distribution* $0.5N(0,1) + 0.5N(5,9)$.

ularized model-averaged estimator equipped with different weights, with the baseline recovery from the Lasso represented in the natural basis are presented in Figure 2.2 for the mixture of normals distributed error, and in Figure 2.3 for the t_3 distributed error. We observe that the strong signals corresponding to large values located at the end of the sound signal are well captured by the model-averaged quantile estimator using different weights for both error distributions. For the weak signals clustering at the front of the signal, the model-averaged estimators using $\widehat{w}_{\mathrm{MA},1}$ and equal weights outperform the counterpart with $\widehat{w}_{\mathrm{MA},2}$ for $0.5N(0,1)+0.5N(5,9)$ distributed errors; recovery differences for the weak signals of the model-

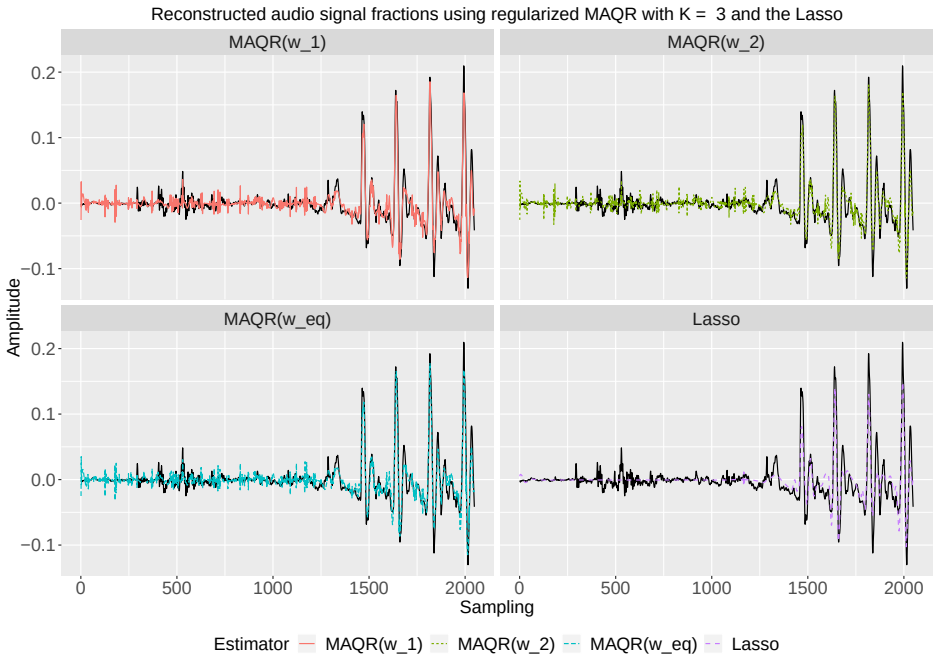


Figure 2.3: *Reconstructed audio signal using the regularized* model-averaged *estimator with the estimated AMSE-type weights in (2.30), oracle-type optimal weights in (2.31), and equal weights. The original audio curve is depicted in black. The Lasso reconstruction is presented at bottom-right. The error used for corruption is* t_3 *distributed.*

averaged estimator using different weights are hardly observable for the t_3 distributed errors. Recovery using the Lasso is competitive to the model-averaged estimator using $w_{\mathrm{MA},1}$ for strong signals. However, the Lasso estimates the signals in an over-sparse way with too many zeros entries; one can observe the almost flat recovery for the weak signals for both error distributions.

Bates and Granger (1969) provide an alternative weight choice for the model-averaged estimator obtained by considering only the variances of $\widehat{\beta}_k$'s and ignoring the covariances. This leads to

$$\widehat{w}_{\mathrm{MA},3} = \arg \min_{w \geq 0, \mathbf{1}_K^\top w = 1} w^\top \mathrm{diag}(\widehat{\Sigma}_{0,(t)}) w, \tag{2.32}$$

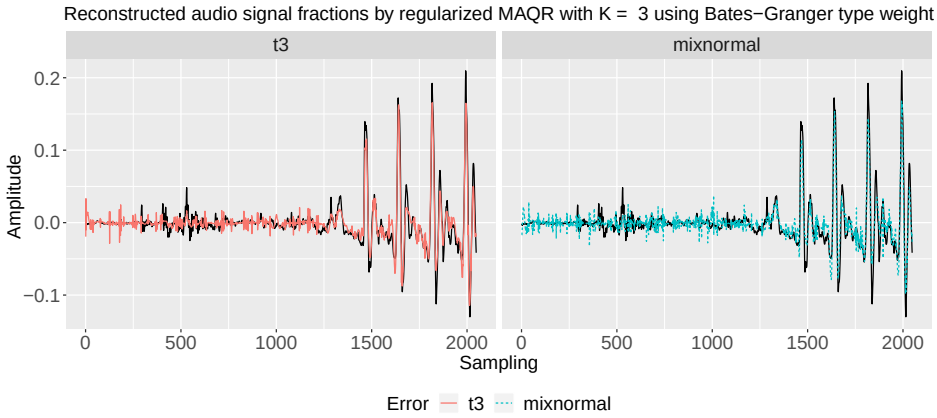


Figure 2.4: *Reconstructed audio signal using the regularized* model-averaged *estimator with Bates-Granger type weight in* (2.32)*. The original audio curve is depicted in black. The left figure uses* t_3 *distributed corruption error, while the figure on the right used* $0.5N(0,1)+0.5N(5,9)$ *distributed corruption error.*

where $\mathrm{diag}(\widehat{\Sigma}_{0,(t)})$ denotes the diagonal matrix obtained from $\widehat{\Sigma}_{0,(t)}$ which keeps the diagonal and has zeros in all off-diagonal entries. Figure 2.4 contains the recovery of the audio signal using the model-averaged estimator using this weight.

For the composite quantile estimator $\widehat{\beta}_{\mathrm{C}}$, we performed the same weight searching method as for the simulation study. This is, $S_{\mathcal{V}} = 5$ and randomly select 4 candidate weights in the neighbourhood of the previous value. We select the tuning parameter α once for the starting weight $\widehat{w}_{\mathrm{C},2}$, it remains unchanged thereafter. The recovered signals by the composite estimator with different weights are very similar in all cases.

To compare the recovery of the regularized model-averaged and composite estimator combined with different weights, as well as the Lasso estimator, we present the mean absolute percentage error (MAPE) in Table 2.5 where the MAPE is defined as

$$\mathrm{MAPE}(\widehat{\beta},\beta) = \frac{1}{p}\sum_{j=1}^{p}\left|\left(\widehat{\beta}_j-\beta_j\right)\Big/\beta_j\right| \tag{2.33}$$

Notice that the MAPE in this Chapter has a different definition from the MAPE_τ, i.e., median absolute prediction error, in Chapter 1. Table 2.6 reports the MSE.

f_ε	t_3				0.5N(0,1) + 0.5 N(5, 9)			
est: MA / C	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$
MAQR	3.177	3.339	3.341	3.005	2.934	5.746	3.722	4.152
CQR	2.798	2.730	2.671	-	6.100	6.090	6.462	-
Lasso	1.346				1.536			

Table 2.5: *The MAPE defined in (2.33) of the audio signal recovered by the regularized model-averaged and composite estimators with different weights, and the Lasso estimator. The seed number used to generated the errors for corrupting the compressed signal vector is 37 for both t_3 and mixed normal distributed errors.*

f_ε	t_3				0.5N(0,1) + 0.5 N(5, 9)			
est: MA / C ($\times 10^{-4}$)	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$
MAQR	1.286	1.288	1.274	1.304	2.044	2.417	2.051	2.006
CQR	1.279	1.273	1.271	-	2.363	2.068	2.120	-
Lasso	2.566				2.003			

Table 2.6: *The MSE of the audio signal recovered by the regularized model-averaged and composite estimators with different weights, and the Lasso estimator. The seed number used to generated the errors for corrupting the compressed signal vector is 37 for both t_3 and mixed normal distributed errors.*

We see that the Lasso has the lowest MAPE for both t_3 and mixed normal distributed errors; at the same time, it estimates the weak signals in an over-sparse way and is not capable of capturing the weak signals. Comparing the effect of different weight choices on the regularized model-averaged quantile estimator with its composite quantile counterpart, we see that the MAPEs of the composite quantile estimators are fairly stable using different weights. The model-averaged estimator with the AMSE-type weight $\widehat{w}_{\text{MA},1}$ has a fair performance compared to the composite estimator, espe-

cially for the mixed normal distributed error. The Bates-Granger weighting provides good results regarding MAPE for the t_3 error case, but not for the mixed normal. Regaring MSE it performs well for the mixed normal case, but is worst for the t_3 errors, where in this example the equal weights perform best, although all results are close. Searching for the selection incorporated weight $\widehat{w}_{\mathrm{C},1}$ for the regularized composite quantile estimator is computationally infeasible for large p (2047 in our case). Estimating the regularized model-averaged quantile estimator averaging 3 quantiles here takes approximately 4 – 5 hours whereas estimating the regularized composite quantile estimator takes more than 16 hours with only 5 steps in a nearby search with 4 surrounding candidate weights, and the tuning parameter α tuned only once for the starting weight.

Additionally, we present the estimated weights for both regularized model-averaged and composite estimators in Table 2.7. An interesting observation is made by comparing the estimated weights $\widehat{w}_{\mathrm{MA},1}$ and $\widehat{w}_{\mathrm{MA},2}$ for the mixed normal distributed error. The weight $\widehat{w}_{\mathrm{MA},1}$ presented here is quite representative; it assigns weight 0 to the quantile estimate at 50% quantile level suggesting the final model-averaged estimate is obtained by averaging estimates at 25% and 75% quantile levels. On the contrary, $\widehat{w}_{\mathrm{MA},2}$ assigns the largest weight to the estimate at 50% quantile level indicating the largest contribution to the final model-averaged estimate.

f_ε	est: MA / C	MAQR	CQR
t_3	$w_{\mathrm{est},1}$	(0.156, 0.725, 0.119)	(0.089, 0.492, 0.419)
	$w_{\mathrm{est},2}$	(0.077, 0.650, 0.273)	(0.314, 0.267, 0.467)
$0.5N(0,1)+$	$w_{\mathrm{est},1}$	(0.548, 0, 0.452)	(0.469, 0.495, 0.036)
$0.5N(5,9)$	$w_{\mathrm{est},2}$	(0.147, 0.843, 0.010)	(0.369, 0.345, 0.286)

Table 2.7: *The estimated weights $\widehat{w}_{\mathrm{MA},1}$ and $\widehat{w}_{\mathrm{MA},2}$ for the model-averaged estimator, and $\widehat{w}_{\mathrm{C},1}$ and $\widehat{w}_{\mathrm{C},2}$ for the composite estimator. The seed number used to generated the errors for corrupting the compressed signal vector is 37 for both t_3 and mixed normal distributed errors.*

2.8 Discussion

This chapter is the first one to take the selection uncertainty due to regularization into account when computing the weights used in model-averaged and composite estimation. While we have studied both composite estimation and model-averaged estimation, the flexibility of allowing for a parallel computation and a component-specific choice of regularization, combined with an explicit expression of the optimal weights for model averaging, places this method in a preferred position from a computational point of view.

It would be interesting to investigate whether AMSE expressions for other types of regularization may be obtained in a similar fashion. Going yet one step further would be incorporating the effect of data-driven values of the regularization parameters λ (for composite estimation) and $\lambda_1, \ldots, \lambda_K$ (for model-averaged estimation) on the choice of the weights. To further study the weight selection and the effect of using data driven weights, one should study the joint distribution of the estimated weights and the estimators of interest. To simplify such matters, sample splitting could be used such that the weights are computed on a hold-out sample and the estimation using those weights proceeds on the rest of the sample. In this chapter we used the same dataset for estimating both β and w.

To avoid overly complicated mathematical expressions, we followed earlier literature in the use of a design matrix where $X_{ij} \sim N(0, 1/n)$. Other applications might require studying, for example, fixed designs, which is beyond the scope of the current chapter.

2.9 Conditions

(A1) Design: The elements of the design matrix X, that is X_{ij} for $i = 1, \ldots, p$ and $j = 1, \ldots, n$, are independent and identically distributed according to a $N(0, 1/n)$ which is also called a standard Gaussian design.

(A2) Coefficients: The p-vector β is such that the sequence of uniform distributions that is placed on its components converges, for p tending to infinity, to a distribution with a bounded $(2k-2)$th moment for $k \geq 2$. Denote by B_0 a random variable with this limiting distribution function F_{B_0}.

(A3) Loss function: (i) The subgradient $\partial\rho(u) = \sum_{j=1}^{3} v_j(u)$ where v_1 has an absolutely continuous derivative, v_2 is continuous and consists of piecewise linear parts and is constant outside a bounded interval, and v_3 is a non-decreasing step function. Denote $v_2'(u) = \alpha_l$ and $v_3(u) = \gamma_l$ when $u \in (r_l, r_{l+1}]$ where $\alpha_0 = \alpha_L = 0$, $-\infty = r_0 < r_1 < \ldots < r_L < r_{L+1} = \infty$ and $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_L < \gamma_{L+1} = \infty$. (ii) The subgradient's absolute value $|\partial\rho(u)|$ is bounded for all $u \in \mathbb{R}$. (iii) $h(t) = \int \rho(z-t)dF_\varepsilon(z)$ has a unique minimum at $t = 0$. (iv) There exists a $\delta > 0$ and $\eta > 1$ such that $E[\{\sup_{|u|\leq\delta} |v_1''(z+u)|\}^\eta]$ is finite.

(A4) We assume that for some $\kappa > 1$,

(a) $\lim_{p\to\infty} E_{\widehat{f}_\beta}(B_0^{2\kappa-2}) = E_{f_{B_0}}(B_0^{2\kappa-2}) < \infty$

(b) $\lim_{p\to\infty} E_{\widehat{f}_\varepsilon}(\varepsilon^{2\kappa-2}) = E_{f_\varepsilon}(\varepsilon^{2\kappa-2}) < \infty$

(c) $\lim_{p\to\infty} E_{\widehat{f}_{q_0}}(B_0^{2\kappa-2}) < \infty$.

(A5) The regression errors $\varepsilon_1, \ldots, \varepsilon_n$ and ε are i.i.d. random variables with mean zero and finite 2nd moment. Assume ε has cumulative distribution function F_ε and probability density function f_ε. Let F_ε have bounded derivatives f_ε and ∂f_ε; further, let $f_\varepsilon > 0$ in the neighbourhood of $r_1, \ldots, r_L$ in (A3).

Condition (A1) has been used by Bayati and Montanari (2011a); Donoho and Montanari (2016); Bradic (2016), Condition (A2) has been used by Bayati and Montanari (2011a); Bradic (2016); while conditions (A3) and (A5) correspond to conditions (R) and (D) of Bradic (2016). Condition (A4) is used in Lemma 2.1, in addition to the moment condition stated in (A2) and (A5). We take $\kappa = 2$ for Algorithm 1.

2.10 Lemmas and Proofs

2.10.1 Auxiliary definitions and lemmas

Definition 1. (Pseudo Lipschitz function) A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order $\kappa \geq 1$, if there exists a constant $L > 0$, such that $\forall x, y \in \mathbb{R}^m$

$$|\phi(x) - \phi(y)| \leq L(1 + \|x\|^{\kappa-1} + \|y\|^{\kappa-1})\|x - y\|.$$

It follows that if ϕ is a pseudo-Lipschitz function of order κ, then there exists a constant L' such that $\forall x \in \mathbb{R}^m : |\phi(x)| \leq L'(1 + \|x\|^{\kappa})$.

Lemma 2.2 (Theorem 1 in Jameson (2014))**.** *If $x_i \geq 0$ where $i = 1, \ldots, n$ and $p \geq 1$. Then*

$$\sum_{i=1}^{n} x_i^p \leq (\sum_{i=1}^{n} x_i)^p \leq n^{p-1} \sum_{i=1}^{n} x_i^p.$$

The reversed inequality holds for $p \in (0, 1)$

Lemma 2.3 (Extrema of quadratic forms in Rao (1973))**.** *Let A be a $m \times m$ matrix , B be a $m \times k$ matrix, and U be a k-vector. Denote by S^- any generalized inverse of $B^\top A^{-1} B$. Then*

$$\inf_{B^\top X = U} X^\top A X = U^\top S^- U$$

where X is a column vector and the infimum is attained at $A^{-1}BS^-U$.

Lemma 2.4 (Stein's lemma in Stein (1981))**.** *Let X_1, X_2 jointly Gaussian distributed. Let $g : \mathbb{R} \to \mathbb{R}$ be absolutely continuous with derivative ∂g and $E|\partial g(X_1)| < \infty$. Then*

$$\mathrm{Cov}(g(X_1), X_2) = \mathrm{Cov}(X_1, X_2) E[\partial g(X_1)].$$

Lemma 2.5 (Lemma 4 in Bayati and Montanari (2011a))**.** *Let $\kappa \geq 2$ and a sequence of vectors $\{\beta(p)\}_{p \geq 0}$ whose empirical distribution converges*

weakly to probability measure f_{B_0} *on* $\mathbb{R}$ *with bounded* κ*th moment; additionally, assume that* $\lim_{p\to\infty} E_{\widehat{f}_\beta}(B_0^\kappa) = E_{f_{B_0}}(B_0^\kappa)$*. Then for any pseudo-Lipschitz function* $\psi : \mathbb{R} \to \mathbb{R}$ *of order* κ*:*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\beta_j) \overset{\text{a.s.}}{=} E[\psi(B_0)].$$

2.10.2 Proofs

Proof of (2.5)

Proof. By definition, the proximal mapping operator is the minimizer of the function $b\rho_{\mathrm{C}}(x) + 0.5(x-z)^2$ which is non-differentiable but subdifferentiable, with subgradient $b \cdot \partial\rho_{\mathrm{C}}(x) + x - z$. $\mathrm{Prox}(z; b)$ is the minimizer if and only if $0 \in \{b \cdot \partial\rho_{\mathrm{C}}(x)|_{x=\mathrm{Prox(z;b)}} + \mathrm{Prox}(z; b) - z\}$. We distinguish between intervals where ρ_{C} is differentiable and non-differentiable points. For $x \in (u_{\tau_\ell}, u_{\tau_{\ell+1}}), \ell = 0, \ldots, K$ the function ρ_C is differentiable. Using the expression of the subgradient in (2.8), we obtain $0 = bh(\ell) + x - z$, which is solved for x to get that $\mathrm{Prox}(z; b) = z - bh(\ell)$. From $\mathrm{Prox}(z; b) \in (u_{\tau_\ell}, u_{\tau_{\ell+1}})$ it follows that $z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_{\ell+1}} + bh(\ell))$. For the non-differentiable points, that is $x = u_{\tau_\ell}, \ell = 1, \ldots, K$, having $0 \in \{b[h(\ell-1), h(\ell)] + x - z\}$ leads to $u_{\tau_\ell} = \mathrm{Prox}(z; b) \in [z - bh(\ell), z - bh(\ell-1)]$. This implies that $z \in [u_{\tau_\ell} + bh(\ell-1), b_{\tau_\ell} + bh(\ell)]$. □

Proof of (2.9)

Proof. By definition, $\widetilde{G}(z; b) = b \cdot \partial\rho(x)|_{x=\mathrm{Prox}(z;b)}$, and, see the Proof of (2.5) hereabove, $0 \in \{b \cdot \partial\rho(x)|_{x=\mathrm{Prox}(z;b)} + \mathrm{Prox}(z, b) - z$. Without loss of generality, we show the calculation for the cases where $z < u_{\tau_1} + bh(0)$ and where $z \in [u_{\tau_1} + bh(0), u_{\tau_1} + bh(1)]$.

For $z < u_{\tau_1} + bh(0)$ it holds that $\mathrm{Prox}(z; b) = z - bh(0) < u_{\tau_1}$, which leads to $\partial\rho(x)|_{x=\mathrm{Prox(z;b)}} = h(0)$. Hence, $\widetilde{G}(z; b) = b \cdot \partial\rho(x)|_{x=\mathrm{Prox(z;b)}} = bh(0)$.

Having $z \in [u_{\tau_1} + bh(0), u_{\tau_1} + bh(1)]$ corresponds to taking the non-differentiable point $u_{\tau_1} = \mathrm{Prox}(z; b)$, see the proof of (2.5). We have

$\partial\rho(x)|_{x=\text{Prox}(z;b)} \in [h(0), h(1)]$. The subgradient $\partial\rho_{\text{C}}$ is non-decreasing (Condition (A3)) and linear. From (2.5) the proximal operator is also a linear function. An intuitive choice for $\widetilde{G}(z;b) = b \cdot \partial\rho(x)|_{x=Prox(z;b)}$ with $z \in [u_{\tau_1} + bh(0), u_{\tau_1} + bh(1)]$ is $\widetilde{G}(z;b) = z - b_{\tau_1}$ which keeps the linearity of the composition of the two functions $\partial\rho$ and $\text{Prox}(\cdot; b)$. □

Proof of (2.15)

Proof. Bayati and Montanari (2011a, Theorem 2, Eq.(3.7)) states in our notation that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(\varepsilon_i - z_{(t),i}, \varepsilon_i) \overset{a.s.}{=} E[\psi(\bar{\sigma}_{(t)} Z, \varepsilon)], \tag{2.34}$$

where $\psi : \mathbb{R}^2 \to \mathbb{R}$ is any pseudo-Lipschitz function, $Z \sim N(0,1)$, $\bar{\sigma}_{(t)}$ from (2.18), and ε as in (A5). Motivated by Eqs.(7.16) and (7.18) in Bradic (2016) we take $\psi(d, \varepsilon) = \{G(\varepsilon - d; b_{(t)})\}^2$, with $b_{(t)}$ as in Algorithm 1, step 2. Applying (2.34) we obtain that as $n \to \infty$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} G(\varepsilon_i - (\varepsilon_i - z_{i,(t)}); b_{(t)})^2 &= \frac{1}{n} \sum_{i=1}^{n} G(z_i; b_{(t)})^2 \\ &\overset{a.s.}{\to} E[G(\varepsilon - \bar{\sigma}_{(t)} Z; b_{(t)})^2]. \end{aligned}$$

□

Estimation of $\nu(b)$

The effective score step in Section 2.3.2, in cases where $G(\cdot; b_{(t)})$ is non-differentiable, requires a solution $b_{(t)}$ to the equation $1 = \widehat{\nu}(b_{(t)})$ where $\widehat{\nu}(b_{(t)})$ is a consistent estimator of a population parameter $\nu(b_{(t)})$ defined as

$$\nu(b_{(t)}) = E[\partial_1 \widetilde{G}(C_{(t)}; b_{(t)})] = b_{(t)}(\delta/\omega) E(\partial[\partial\rho\{\text{Prox}(C_{(t)}; b_{(t)})\}])$$

with $C_{(t)} = \varepsilon - \bar{\sigma}_{(t)}Z$ the random variable characterizing the limit distribution of the adjusted residuals $z_{(t)}$ when $p \to \infty$.
Using Condition (A3) and Lemma 3 of Bradic (2016), $\partial\rho$ can be written as a sum of three functions of which v_1 and v_2 are differentiable. For the step function v_3, we use Condition (A3) on ρ, where γ_l is the step height on the interval $(r_l, r_{l+1}]$. Let $f_{C_{(t)}-\widetilde{G}(C_{(t)};b)}$ denote the density of the variable $C_{(t)} - \widetilde{G}(C_{(t)}; b)$ which is equivalent to $\text{Prox}(C_{(t)}; b)$. The equivalence is obtained by setting the derivative of the $b\rho(x) + \frac{1}{2}(x - C_{(t)})^2$ w.r.t. x to zero and evaluate at $\text{Prox}(C_{(t)}; b)$, due to the fact that the proximal operator is the minimizer of the function $b\rho(x) + \frac{1}{2}(x - C_{(t)})^2$. Then we arrive at

$$\begin{aligned}\frac{\omega\nu(b_{(t)})}{\delta b_{(t)}} &= \sum_{j=1}^{2} E[\partial v_j(C_{(t)})] \\ &\quad + \sum_{l=1}^{L-1} \gamma_l \{f_{C_{(t)}-\widetilde{G}(C_{(t)};b_{(t)})}(r_{l+1}) - f_{C_{(t)}-\widetilde{G}(C_{(t)};b)}(r_l)\}.\end{aligned}$$

The consistent estimator in (2.11) is obtained by replacing the expectation above with the empirical mean and replacing the density of the proximal operator $\text{Prox}(C_{(t)}; b)$ with its kernel density estimator.

Proof of Lemma 2.1

Since this proof is based on the general recursion and Lemma 1 in Bayati and Montanari (2011a), we first restate the general recursion to which Algorithm 1 belongs with slight changes in the notations. Given the noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ and the coefficient vector $\beta \in \mathbb{R}^p$, the general recursion defined is

$$h_{(t+1)} = X^\top m_{(t)} - \xi_{1,(t)} q_{(t)}, \quad m_{(t)} = g_{1,t}(d_{(t)}, \varepsilon)$$

$$d_{(t)} = X q_{(t)} - \xi_{2,(t)} m_{(t-1)}, \quad q_{(t)} = g_{2,(t)}(h_{(t)}, \beta)$$

where

$$\xi_{1,(t)} = n^{-1} \sum_{i=1}^{n} \partial_1 g_{1,(t)}(d_{(t),i}, \varepsilon_i),$$
$$\xi_{2,(t)} = (\delta p)^{-1} \sum_{j=1}^{p} \partial_1 g_{2,(t)}(h_{(t),j}, \beta_j).$$

Further, to connect the general recursion to Algorithm 1, we also state the exact form of $h_{(t+1)}, m_{(t)}, d_{(t)}, q_{(t)}$ taken in Algorithm 1. Lemma 1 in Bradic (2016) states that Algorithm 1 takes $h_{(t+1)} = \beta - X^\top G(z_{(t)}; b_{(t)}) - \beta_{(t)}$, $q_{(t)} = \beta_{(t)} - \beta$, from (2.6) $z_{(t)} = \varepsilon - d_{(t)}$, which defines $d_{(t)}$, $m_{(t)} = -G(z_{(t)}; b_{(t)})$ with the functions $g_{1,(t)}(x_1, x_2) = -G(x_2 - x_1; b_{(t)})$, and $g_{2,(t)}(x_1) = \eta(\beta - x_1; \theta) - \beta$. To proceed with the proof of Lemma 2.1, we first recall the technique used for proving Lemma 1 in Bayati and Montanari (2011a), which uses induction on the iteration t. To not fully repeat the long proof and all notations we only give details about where our proof differs from theirs.

(i) $\mathcal{B}_{(0)}$: show properties (3.15), (3.17), (3.19), (3.21), (3.23) and (3.23) of Bayati and Montanari (2011a) which are related to the vectors $b_{(0)}$ and $m_{(0)}$, by conditioning on the σ-algebra $\mathcal{D}_{(0),(0)}$ generated by $\{\beta, \varepsilon, q_{(0)}\}$; obtain the σ-algebra $\mathcal{D}_{(1),(0)}$ by adding $b_{(0)}$ and $m_{(0)}$ to the set $S_{(0),(0)} = \{\beta, \varepsilon, q_{(0)}\}$.

(ii) $\mathcal{H}_1$: show that the properties (3.14), (3.16), (3.18), (3.20), (3.22), (3.24) and (3.25), which are related to the vectors $h_{(1)}$ and $q_{(1)}$, hold by conditioning on the σ-algebra $\mathcal{D}_{(1),(0)}$; obtain the σ-algebra $\mathcal{D}_{(1),(1)}$ by adding $h_{(1)}$ and $m_{(1)}$ to the set $S_{(1),(0)} = \{\beta, \varepsilon, q_{(0)}, d_{(0)}, m_{(0)}\}$.

(iii) $\mathcal{B}_{(t)}$: Similar to $\mathcal{B}_{(0)}$; the proof is conditioning on the σ-algebra $\mathcal{D}_{(t),(t)}$ for the set containing $\beta, \varepsilon, q_{(0)}$ and all previous obtained vectors; obtain the new σ-algebra $\mathcal{D}_{(t+1),(t)}$ by adding $b_{(t+1)}$ and $m_{(t+1)}$ to the set.

(iv) $\mathcal{H}_{(t+1)}$: Similar to $\mathcal{H}_{(1)}$; conditioning on the σ-algebra $\mathcal{D}_{(t+1),(t)}$ for

the set containing $\beta, \varepsilon, q_{(0)}$ and all previous obtained vectors.

Assuming Lemma 1 in Bayati and Montanari (2011a) holds for all K estimators $\widehat{\beta}_k, k = 1, \ldots, K$ in (2.2), we add an additional step considering the correlations between the estimators. The main technique is conditioning on the σ-algebra generated by $\cup_{k=1}^{K} \mathcal{S}_{k,(1),(0)}$ and $\cup_{k=1}^{K} \mathcal{S}_{k,(t+1),(t)}$, where $\mathcal{S}_{k,(1),(0)}$ and $\mathcal{S}_{k,(t+1),(t)}$ are the sets described in step 2 and 4 above for the kth estimator. The proof is similar to that of (3.16) in Lemma 1(b) of Bayati and Montanari (2011a), with different mathematical techniques in order to adjust the original proof from a single sequence of iterations to K paralleled sequences of iterations.

Proof. Idea of the construction: The construction of $\mathcal{B}_{(0)}$, $\mathcal{H}_{(0)}$, $\mathcal{B}_{(t+1)}$ and $\mathcal{H}_{(t+1)}$ depends on the space $\mathcal{D}_{(t+1),(t)}$ which is the space generated by the true coefficient β, the noise ε, the initial condition $q_{(0)}$, and the subsequent terms generated from Algorithm 1. The proof by induction is similar to the proof of Lemma 1(b) in Bayati and Montanari (2011a). We prove that $\mathcal{H}_{(1)}$ holds and if $\mathcal{B}_{(r)}, \mathcal{H}_{(s)}$ holds for all $r \leq t$ and $s \leq t$, then $\mathcal{H}_{(t+1)}$ holds. Let $o_{k,(t)}(1)$ denote a vector in $\mathbb{R}^t$ for the kth estimator such that all of its entries converge to 0 almost surely for $p \to \infty$.

Step 2 from Bayati and Montanari (2011a): $\mathcal{H}_{(1)}$: We know from Bayati and Montanari (2011a, Eq.(3.35)) that for each k and a Gaussian matrix $\tilde{X}_k$ with the same distribution as the design matrix X, see also Bayati and Montanari (2011a) Lemma 2 (1),

$$h_{k,(1)}|_{\mathcal{D}_{k,(1),(0)}} \stackrel{d}{=} (\tilde{X}_k)^\top m_{k,(0)} + o_{k,(1)}(1) q_{k,(0)}.$$

Let $a_{k,j} = ([(\tilde{X}_k)^\top m_{k,(0)}]_j + o_{1,k}(1) q_{k,(0),j}, \beta_j)$ and $c_{k,j} = ([(\tilde{X}_k)^\top m_{k,(0)}]_j, \beta_j)$ where $k = k_1, k_2$. We first show that for any two k_1, $k_2 \in \{1, \ldots, K\}$.

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \left[\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) \tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j}) \tilde{\psi}_{\mathrm{c}}(c_{k_2,j}) \right] = 0. \tag{2.35}$$

Since $\tilde{\psi}_{\mathrm{c}}$ is κ_{c} order pseudo-Lipschitz, hence we have

$$|\tilde{\psi}_{\mathrm{c}}(a_{k,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k,j})| \leq L\{1 + \max(\|a_{k,j}\|^{\kappa_{\mathrm{c}}-1}, \|c_{k,j}\|^{\kappa_{\mathrm{c}}-1})\}|q^0_{k,j}|o_{1,k}(1);$$
$$|\tilde{\psi}_{\mathrm{c}}(a_{k,j})| \leq L'(1 + \|a_{k,j}\|^{\kappa_{\mathrm{c}}}), \quad |\tilde{\psi}_{\mathrm{c}}(c_{k,j})| \leq L''(1 + \|c_{k,j}\|^{\kappa_{\mathrm{c}}});$$

meanwhile, from the proof in $\mathcal{H}_0$ in Lemma 1 in Bayati and Montanari (2011a), we have for an arbitrary κ_{c} order pseudo-Lipschitz function $\tilde{\psi}_{\mathrm{c}}$

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k,j})| = 0. \tag{2.36}$$

Notice that

$$\begin{aligned}
&|\tilde{\psi}_{\mathrm{c}}(a_{k_1,j})\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})\tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&= |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j})\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(a_{k_2,j})\tilde{\psi}_{\mathrm{c}}(c_{k_1,j}) \\
&\quad + \tilde{\psi}_{\mathrm{c}}(a_{k_2,j})\tilde{\psi}_{\mathrm{c}}(c_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})\tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\leq |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| + |\tilde{\psi}_{\mathrm{c}}(c_{k_1,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})|.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j})\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})\tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\leq \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| + |\tilde{\psi}_{\mathrm{c}}(c_{k_1,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\leq \max_j |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j})| \cdot \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| \\
&\quad + \max_j |\tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| \cdot \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\leq L'_2\{1 + \max_j(\|a_{k_2,j}\|^{\kappa_{\mathrm{c}}})\}\frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| \\
&\quad + L''_1\{1 + \max_j(\|c_{k_1,j}\|^{\kappa_{\mathrm{c}}})\}\frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})|.
\end{aligned} \tag{2.37}$$

By (2.36), for $k = k_1, k_2$, $p^{-1}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k,j})|$ tends to 0 as $p \to +\infty$. The remaining two factors are finite almost surely: $[(\tilde{X}^k)^\top m_{k,(0)}]_j$ is a Gaussian random variable which is finite almost surely; $\beta_{0,j}$ is finite almost surely since its limiting distribution has bounded moments up to $(2\kappa - 2)$ by condition (A2). Hence, for any pairs $k_1, k_2 \in \{1, \ldots, K\}$ (2.35) holds.

From here, we consider $\tilde{h}_{k,(1)}|_{\mathcal{D}_{k,(1),(0)}} \stackrel{d}{=} (\tilde{X}_k)^\top m_{k,(0)}$ of which the components have the same distribution as $\|m_{k,(0)}\| Z_k/\sqrt{n}$ for $Z_k \sim N(0,1)$. Conditioning on $\mathcal{D}_{k_1,(1),(0)}$ and $\mathcal{D}_{k_2,(1),(0)}$, we use the strong law of large numbers for triangular arrays in Theorem 3 of Bayati and Montanari (2011a) to obtain that

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \Big\{ \prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j}, \beta_j) - E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})}[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j}, \beta_j)] \Big\} \stackrel{\text{a.s.}}{=} 0. \tag{2.38}$$

We first prove (2.38). For $k_1 \neq k_2$, we show that the condition in Theorem 3 of Bayati and Montanari (2011a) holds. To simplify the notation, we denote the independent copies of the matrices $\tilde{X}_{k_1}, \tilde{X}_{k_2}$ to be X_{k_1}, X_{k_2}. We take the random variables in the triangular array to be

$$\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j}, \beta_j)\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j}, \beta_j) - E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})}[\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j}, \beta_j)\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j}, \beta_j)] \tag{2.39}$$

and let $0 < \rho < 1$ then

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p} E\Big| \prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j}, \beta_j) - E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})}[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j}, \beta_j)] \Big|^{2+\rho} \\
&= \frac{1}{p}\sum_{j=1}^{p} E_{(X_{k_1}, X_{k_2}, \tilde{X}_{k_1}, \tilde{X}_{k_2})} \Big[\Big| \tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j, \beta_j)\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j, \beta_j) \\
&\quad - \tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j, \beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j, \beta_j) \Big|^{2+\rho} \Big] \\
&= \frac{1}{p}\sum_{j=1}^{p} E\Big[\Big| \tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j, \beta_j)\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j, \beta_j)
\end{aligned}$$

$$
\begin{aligned}
&\quad -\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\\
&\quad +\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\\
&\quad -\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
\leq\ & \frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\\
&\quad -\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\\
&+\frac{1}{p}\sum_{j=1}^{p} E\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\\
&\quad -\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
\leq\ & \frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\Big|^{2+\rho}\\
&\quad \times\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
&+\frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\\
&\quad \times\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
\leq\ & \frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|L'\Big(1+|[X_{k_1}^\top m_{k_1,(0)}]_j|^{\kappa_c}+|\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho}\\
&\quad \times\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
&+\frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|L'\Big(1+|[\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j|^{\kappa_c}+|\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho}\\
&\quad \times\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
\leq\ & \max_{j=1,\ldots,p} E\Big[\Big|L'\Big(1+|[X_{k_1}^\top m_{k_1,(0)}]_j|^{\kappa_c}+|\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho}\Big]\\
&\quad \times\frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([(\tilde{X}_{k_2})^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]
\end{aligned}
$$

$$+ \max_{j=1,\ldots,p} E\Big[\Big|L'\Big(1+|[\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j|^{\kappa_c}+|\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho}\Big]$$
$$\times\frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big].$$

For the first term in the last inequality above, we see that the expectation $E[|L'(1+|[\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j|^{\kappa_c}+|\beta_j|^{\kappa_c})|^{2+\rho}]$ is bounded by some constant, since the expectation is with respect to the matrices $X_{k_1}, X_{k_2}, \tilde{X}_{k_1}, \tilde{X}_{k_2}$ of which the components are Gaussian distributed with mean 0 and variance $1/n$; the rest terms are bounded by a constant; the moments of Gaussian distributed r.v. are all finite. Let us denote the upper bound of this expectation by L'', then the first term of the inequality above is bounded by

$$L''\frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big],$$

which can be shown to be bounded by $cp^{\rho/2}$ following a similar argument as in Lemma 1(b) in Bayati and Montanari (2011a). The second term similarly can be shown to be bounded by $c'p^{\rho/2}$. Hence the variable defined in (2.39) satisfies the condition in Theorem 3 in Bayati and Montanari (2011a); thus the a.s. convergence holds.

In the special case where $k_1 = k_2$, we show that the square of ψ_c is still pseudo-Lipschitz of order $2\kappa_c \leq \kappa$, then the almost sure convergence hold by directly applying the result in Lemma 1 in Bayati and Montanari (2011a).

To simplify the notation, we use ψ to denote any pseudo-Lipschitz function here. For any pairs $x, y \in \mathbb{R}^m$, we have

$$\begin{aligned}
&|\psi^2(x)-\psi^2(y)| \\
&\leq |\psi(x)+\psi(y)||\psi(x)-\psi(y)| \leq (|\psi(x)|+|\psi(y)|)|\psi(x)-\psi(y)| \\
&\leq L'(1+\|x\|^\kappa+1+\|y\|^\kappa)\cdot L(1+\|x\|^{\kappa-1}+\|y\|^{\kappa-1})\|x-y\| \\
&\leq LL''(1+\|x\|^\kappa+\|y\|^\kappa)(1+\|x\|^{\kappa-1}+\|y\|^{\kappa-1})\|x-y\|
\end{aligned}$$

$$\begin{aligned} &\leq LL^{''}(1+\|x\|+\|y\|)^{2\kappa-1}\|x-y\| \\ &\leq LL^{''}3^{\kappa-1}(1+\|x\|^{2\kappa-1}+\|y\|^{\kappa-1})\|x-y\|. \end{aligned}$$

Since $\kappa \geq 1$, $\|x\|, \|y\| \geq 0$, the last two inequalities are obtained by applying the first and second inequality in Lemma 2, respectively. Hence, the square of any arbitrary pseudo-Lipschitz function of order κ is still pseudo-Lipschitz with order 2κ. This proves (2.38).

Using Lemma 2.5 for $v = \beta$ and

$$\psi(\beta_j) = E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})} \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j}, \beta_j) \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j}, \beta_j),$$

the following convergence holds

$$\begin{aligned} &\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})} \Big[\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j}, \beta_j) \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j}, \beta_j) \Big] \\ &\overset{\text{a.s.}}{=} E_{B_0} \Bigg[E_{(Z_{k_1,(0)}, Z_{k_2,(0)})} \Big[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\| \frac{m_{k_r,(0)}}{\sqrt{n}} \| Z_{k_r,(0)}, B_0) \Big] \Bigg] \\ &\overset{\text{a.s.}}{=} E \Big[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\bar{\zeta}_{k_r,(0)} Z_{k_r,(0)}, B_0) \Big]. \end{aligned}$$

Step 4 from Bayati and Montanari (2011a): $\mathcal{H}_{t+1}$: Following the first expression in the proof of Lemma 1(b) in step 4 in Bayati and Montanari (2011a), for any index $k = 1, \ldots, K$

$$\begin{aligned} &\tilde{\psi}_{\mathrm{c}}(h_{k,(1),j}, \ldots, h_{k,(t+1),j}, \beta_j)|_{\mathcal{D}_{k,(t+1),(t)}} \overset{d}{=} \\ &\quad \tilde{\psi}_{\mathrm{c}}\Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \\ &\qquad \Big[\sum_{r=0}^{t-1} \alpha_r h_{k,(r+1)} + (\tilde{X}_k)^{\top} m_{k,(t)} + \tilde{Q}_{k,(t+1)} o_{k,(t+1)}(1) \Big]_j, \beta_j \Big). \end{aligned}$$

The columns of $\tilde{Q}_{k,(t+1)}$ form an orthogonal basis for the column space of $Q_{k,(t+1)} = [q_{k,(0)} \ldots q_{k,(t)}]$. Define the matrix $M_{k,(t)} = [m_{k,(0)} \ldots m_{k,(t-1)}]$, the vector $(m_{k,(t)})_{\|} = \sum_{r=0}^{t-1} \delta_r m_{k,(r)}$ as the projection of $m_{k,(t)}$ on the

column space of $M_{k,(t)}$ and the vector $(m_{k,(t)})_\perp = m_{k,(t)} - (m_{k,(t)})_\parallel$. Similar to the proof in $\mathcal{H}_1$, we first show that the error term $\tilde{Q}_{k,(t+1)} o_{k,(t+1)}(1)$ can be dropped. Let

$$a_{k,j} = \Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \\ \Big[\sum_{r=0}^{t-1} \delta_r h_{k,(r+1)} + (\tilde{X}_k)^\top (m_{k,(t)})_\perp + \tilde{Q}_{k',(t+1)} o_{k,(t+1)}(1)\Big]_j, \beta_j\Big),$$

and $c_{k,j} = \Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \Big[\sum_{r=0}^{t-1} \delta_r h_{k,(r+1)} + (\tilde{X}_k)^\top (m_{k,(t)})_\perp\Big]_j, \beta_j\Big)$. To show that the left hand-side of (2.37) is finite for the new $a_{k,j}$ and $c_{k,j}$, it suffices to show that both $\max_j(\|a_{k_2,j}\|^{\kappa_c})$ and $\max_j(\|c_{k_1,j}\|^{\kappa_c})$ are finite almost surely. By Lemma 2.2, we obtain the following inequality

$$\begin{aligned} \max_j(\|a_{k_2,j}\|^{\kappa_c}) &= \max_j \Big(C(\sum_{r=0}^{t} |h_{k_2,(r+1),j}|^{\kappa_c} + |\beta_j|^{\kappa_c})\Big) \\ &\leq C\Big(\sum_{r=0}^{t} \max_j |h_{k_2,(r+1),j}|^{\kappa_c} + \max_j |\beta_j|^{\kappa_c}\Big) \end{aligned}$$

for some constant C. The finiteness of $\max_j |\beta_j|^{\kappa_c}$ has been discussed in $\mathcal{H}_1$; $\max_j |h_{k_2,(r+1),j}|$ is finite almost surely since Lemma 1 in Bayati and Montanari (2011a) states that for a higher order $l = k-1$, $\lim \sum_{p\to\infty} p^{-1} \sum_{j=1}^{p} (h_{k_2,(t+1),j})^{2l} < \infty$.
The almost-sure finiteness of $\max_j |h_{k_2,(r+1),j}|$ follows by a simple contradiction: assume $P(\max_j |h_{k_2,(r+1),j}| = \infty) = P(|h_{k_2,(r+1),j_{\max}}| = \infty) > 0$, then

$$P\Big(\sup_{p'\geq p} \frac{1}{p'} \sum_{j=1}^{p'} (h_{k_2,(t+1),j})^{2l} < \infty\Big)$$
$$= P\Big(\sup_{p'\geq p} \tfrac{p'-1}{p'} (\tfrac{1}{p'-1} \textstyle\sum_{j\neq j_{\max}} (h_{k_2,(t+1),j})^{2l}) + \tfrac{1}{p'} (h_{k_2,(t+1),j_{\max}})^{2l} < \infty\Big) < 1.$$

The above equation contradicts the result in Lemma 1(e) in Bayati and Montanari (2011a). Follow similar arguments, we have $\max_j(\|c_{k_1,j}\|^{\kappa_c})$

finite almost surely. Now we consider the random variable

$$\tilde{A}_{k,j} = \tilde{\psi}_{\mathrm{c}}\Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \\ \Big[\sum_{r=0}^{t-1} \delta_r h_{k,(r+1)} + (\tilde{X}_k)^\top (m_{k,(t)})_\perp + \tilde{Q}_{k,(t+1)} o_{k,(t+1)}(1)\Big]_j, \beta_j\Big).$$

Following arguments as in $\mathcal{H}_1$, it is easy to show that

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \Big[\tilde{A}_{k_1,j}\tilde{A}_{k_2,j} - E_{(\tilde{X}_{k_1},\tilde{X}_{k_2})}\tilde{A}_{k_1,j}\tilde{A}_{k_2,j}\Big] \overset{\text{a.s.}}{=} 0. \tag{2.40}$$

By Lemma 2.5 and arguments as in the proof of Lemma 1 (b) in Bayati and Montanari (2011a),

$$\begin{aligned}
&\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \tilde{\psi}_{\mathrm{c}}\Big(h_{k_1,(1),j}, \ldots, h_{k_1,(t),j}, \big[\sum_{r=0}^{t-1} \delta_{k_1,(r)} h_{k_1,(r+1)} + (\tilde{X}_{k_1})^\top (m_{k_1,(t)})_\perp\big]_j, \beta_j\Big) \\
&\qquad \times \tilde{\psi}_{\mathrm{c}}\Big(h_{k_2,(1),j}, \ldots, h_{k_2,(t),j}, \big[\sum_{r=0}^{t-1} \delta_{k_2,(r)} h_{k_2,(r+1)} + (\tilde{X}_{k_2})^\top (m_{k_2,(t)})_\perp\big]_j, \beta_j\Big) \\
&\overset{\text{a.s.}}{=} E_{B_0} E_{(Z_{k_1,(0)},\ldots,Z_{k_1,(t)},Z_{k_2,(0)},\ldots,Z_{k_2,(t)})} \\
&\qquad \Big[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}\Big(\bar{\zeta}_{k_r,(0)} Z_{k_r,(0)}, \ldots, \bar{\zeta}_{k_r,(t)} Z_{k_r,(t)}, B_0\Big)\Big] \\
&= E\Big[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}\Big(\bar{\zeta}_{k_r,(0)} Z_{k_r,(0)}, \ldots, \bar{\zeta}_{k_r,(t)} Z_{k_r,(t)}, B_0\Big)\Big].
\end{aligned}$$

□

Proof of Corollary 2.2

Proof. The almost sure convergence holds by choosing $\tilde{\psi}_{\mathrm{c}}(y_{(0)}, \ldots, y_{(t)}, \beta_j) = \psi_{\mathrm{c}}(y_{(t)}, \beta_j) = \Big(\beta_j - y_{(t)}\Big) - \beta_j$ in Lemma 2.1. □

Proof of Theorem 2.1

Proof. By Lemma 2.1 and choosing $\tilde{\psi}_{\mathrm{c}}(y_{(0)}, \ldots, y_{(t)}, \beta_j) = \psi_{\mathrm{c}}(y_{(t)}, \beta_j) = \eta(\beta_j - y_{(t)}; \theta_{(t)}) - \beta_j$ which is a pseudo-Lipschitz function of order $\kappa_c = 1$ the convergence in (2.25) is obtained. □

Proof of Theorem 2.2

Proof. Theorem 2 in Bayati and Montanari (2011a) showed that when assigning $1/p$ point mass to each entry of the vector, $\widetilde{\beta}_{k,j,(t-1)}(p)$ converges weakly to $B_0 + \bar{\zeta}_{k,(t-1)} Z_k$ for $p \to \infty$ where $Z_k \sim N(0,1)$ and B_0 has p.d.f. f_{B_0}. When p is large, $\widetilde{\beta}_{k,(t-1)} \mid (B_0 = \beta) \approx N(\beta, \bar{\zeta}^2_{k,(t-1)} I_p)$; the normality comes from $Z_k \sim N(0,1)$. Similar results for the Lasso estimator can be found in Bayati et al. (2013) and Donoho and Montanari (2016). The normality of $\widetilde{\beta}_{k,(t-1)}$ ensures that the Stein's unbiased risk estimate is applicable for constructing the AMSE estimator. We choose μ, x, $\widehat{\mu}(x)$ and $g(x)$ in Lemma 2.4 to be β, $\widetilde{\beta}_{k,(t-1)}$, $\eta(\widetilde{\beta}_{k,(t-1)}; \theta_{k,(t-1)})$ and $(\eta(\widetilde{\beta}_{k,(t-1)}; \theta_{k,(t-1)}) - \widetilde{\beta}_{k,(t-1)})$, respectively. Recall that $\eta(\widetilde{\beta}_{k,(t-1)}; \theta_{k,(t-1)})$ refers to applying the soft-thresholding function with parameter θ_t to each entry of the vector $\widetilde{\beta}_{(t-1)}$. Then the function $\eta(\cdot; \theta_{k,(t-1)})$ is weakly differentiable with the derivative defined almost everywhere on $\mathbb{R}^p$ except at $-\theta_{k,(t-1)}$ and $\theta_{k,(t-1)}$ in each coordinate.

Next, consider any pair (k_1, k_2) with $k_1, k_2 \in \{1, \ldots, K\}$. The conditional normality holds for $\widetilde{\beta}_{k_r,(t-1)}$ $(r = 1, 2)$. Each component of the sequence $\widetilde{\beta}_{k_r,(t-1)}$ is independent of the remaining entries. Hence, the dependence between $\widetilde{\beta}_{k_1,(t-1)}$ and $\widetilde{\beta}_{k_2,(t-1)}$ comes from the entry-wise dependence of the two variables. In other words, there is only dependence between $\widetilde{\beta}_{k_1,(t-1),j_1}$ and $\widetilde{\beta}_{k_2,(t-1),j_2}$ when $j_1 = j_2$. The covariance between the two sequences is

$$\bar{\zeta}_{(k_1,k_2),(t-1)} = \mathrm{Cov}(\widetilde{\beta}_{k_1,(t-1)}, \widetilde{\beta}_{k_2,(t-1)}).$$

For $\widetilde{\beta}_{k_r,(t-1),-j}$ the vector obtained by excluding the jth entry of $\widetilde{\beta}_{k_r,(t-1)}$ we can easily check that the function $f(x, \widetilde{\beta}_{k_r,(t-1),-j}) : x \to \big(\eta(x; \theta) - x\big)$ is univariate and satisfies the condition in Lemma 2.4 (Stein, 1981).

Meanwhile, since $\bar{\zeta}^2_{\text{emp},(t)} = \bar{\zeta}^2_{\text{emp},(t-1)} + o(1)$ by assumption, $\theta_{k_r,(t)} = \alpha\bar{\zeta}_{k_r,(t)}$ where α is fixed for the different iterations, we obtain

$$\begin{aligned}
&\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \mid \widehat{\beta}_{k_r,(t),j} - \widehat{\beta}_{k_r,(t-1),j} \mid \\
&\overset{a.s}{=} E \mid \eta(B_0 + \bar{\zeta}_{k_r,(t)} Z_{k_r,(t),j}; \theta_{k_r,(t)}) - \eta(B_0 + \bar{\zeta}_{k_r,(t-1)} Z_{k_r,(t-1),j}; \theta_{k_r,(t-1)}) \mid \\
&= E \mid \eta(B_0 + \bar{\zeta}_{k_r,(t)} Z_{k_r,(t),j}; \theta_{k_r,(t)}) \\
&\qquad -\eta(B_0 + \bar{\zeta}_{k_r,(t)} Z_{k_r,(t-1),j}; \theta_{k_r,(t)}) + o(1) \mid \\
&= 0.
\end{aligned}$$

The almost sure convergence holds by Lemma 1(b) (Bayati and Montanari, 2011a). The next equality holds by $\bar{\zeta}^2_{(t)} = \bar{\zeta}^2_{(t-1)} + o(1)$ and the definition of $\theta_{k_r,(t)}$. The last equality holds because both $Z_{k_r,(t-1),j}$ and $Z_{k_r,(t),j}$ are standard Gaussian distributed. Thus, $\widehat{\beta}_{k_r,(t),j} - \widehat{\beta}_{k_r,(t-1),j} | (B_0 = \beta_j)$ converges to 0 almost surely. Further, by (2.13), $\widehat{\beta}_{k_r,(t-1),j} - \widetilde{\beta}_{k_r,(t-1),j} \overset{d}{=} \bar{\zeta}_{k_r,(t-1)} Z_{k,j}$ where $Z_{k_r,j} \sim N(0,1)$. Then,

$$\begin{aligned}
&\widehat{\beta}_{k_r,(t),j} - \widetilde{\beta}_{k_r,(t-1),j} \\
&\quad = (\widehat{\beta}_{k_r,(t),j} - \widehat{\beta}_{k_r,(t-1),j}) + (\widehat{\beta}_{k_r,(t-1),j} - \widetilde{\beta}_{k_r,(t-1),j}) \overset{d}{=} \bar{\zeta}_{k_r,(t-1)} Z_{k,j},
\end{aligned}$$

where $\widehat{\beta}_{k_r,(t),j} = \eta(\widetilde{\beta}_{k_r,(t-1),j}; \theta_{k_r,(t-1)})$, by Slutsky's theorem. Next, Stein's lemma is applied. We denote by A_j, conditioning on $(B_0 = \beta_j)$ and $\widetilde{\beta}_{k_r,(t-1),-j}, r = 1, 2$. It holds that

$$\begin{aligned}
&E\Big[\{\eta(\widetilde{\beta}_{k_1,(t-1),j}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1),j}\}(\widetilde{\beta}_{k_2,(t-1),j} - \beta_j) | A_j\Big] \\
&\quad = \text{Cov}\Big(\eta(\widetilde{\beta}_{k_1,(t-1),j}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1),j}, \widetilde{\beta}_{k_2,(t-1),j} | A_j\Big) \\
&\quad = \text{Cov}(\widetilde{\beta}_{k_1,(t-1),j}, \widetilde{\beta}_{k_2,(t-1),j} | A_j) E\Big[\partial_1 \eta(\widetilde{\beta}_{k_1,(t-1),j}; \theta_{k_1,(t-1)}) - 1 | A_j\Big].
\end{aligned}$$

Below we condition everywhere on B which denotes the event that $B_{0,j} = \beta_j$ for $j = 1, \ldots, p$ where $B_{0,j}$ are independent copies of B_0. Taking expec-

tation w.r.t. $\widetilde{\beta}_{k_r,(t-1),-j}$, we obtain for the whole vector,

$$\begin{aligned}
&E\Big[\{\eta(\widetilde{\beta}_{k_1,(t-1)};\theta_{k_1,(t-1)})-\widetilde{\beta}_{k_1,(t-1)}\}(\widetilde{\beta}_{k_2,(t-1)}-\beta)|B\Big]\\
&\quad= \bar{\zeta}_{(k_1,k_2),(t-1)}E\Big[\partial_1\eta(\widetilde{\beta}_{k_1,(t-1)};\theta_{k_1,(t-1)})-\mathbf{1}_p|B\Big]\\
&E\Big[\{\eta(\widetilde{\beta}_{k_2,(t-1)};\theta_{k_2,(t-1)})-\widetilde{\beta}_{k_2,(t-1)}\}(\widetilde{\beta}_{k_1,(t-1)}-\beta)|B\Big]\\
&\quad= \bar{\zeta}_{(k_1,k_2),(t-1)}E\Big[\partial_1\eta(\widetilde{\beta}_{k_2,(t-1)};\theta_{k_2,(t-1)})-\mathbf{1}_p|B\Big].
\end{aligned}$$

Next, we show the construction of the estimator for $(\Sigma_0)_{(k_1,k_2),(t)}$ at iteration t. The product-sign notation $\prod_{r=1}^2 v_r = v_1^\top v_2$.

$$\begin{aligned}
&E\Big[(\widehat{\beta}_{k_1,(t)}-\beta)^\top(\widehat{\beta}_{k_2,(t)}-\beta)|B\Big] = E\Big[\prod_{r=1}^2\{\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})-\beta\}|B\Big]\\
&= E\Big[\prod_{r=1}^2\{\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})-\widetilde{\beta}_{k_r,(t-1)}\}|B\Big]+E\Big[\prod_{r=1}^2(\widetilde{\beta}_{k_r,(t-1)}-\beta)|B\Big]\\
&\quad+E\Big[\{\eta(\widetilde{\beta}_{k_1,(t-1)};\theta_{k_1,(t-1)})-\widetilde{\beta}_{k_1,(t-1)}\}^\top(\widetilde{\beta}_{k_2,(t-1)}-\beta)|B\Big]\\
&\quad+E\Big[(\widetilde{\beta}_{k_1,(t-1)}-\beta)^\top\{\eta(\widetilde{\beta}_{k_2,(t-1)};\theta_{k_2,(t-1)})-\widetilde{\beta}_{k_2,(t-1)}\}|B\Big]\\
&= E\Big[\prod_{r=1}^2\{\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})-\widetilde{\beta}_{k_r,(t-1)}\}|B\Big]+\bar{\zeta}_{(k_1,k_2),(t-1)}\\
&\quad+\bar{\zeta}_{(k_1,k_2),(t-1)}\sum_{r=1}^2 E\Big[\partial_1\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})-\mathbf{1}_p|B\Big]\\
&= -\bar{\zeta}_{(k_1,k_2),(t-1)}+E\Big[\prod_{r=1}^2\{\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})-\widetilde{\beta}_{k_r,(t-1)}\}|B\Big]\\
&\quad+\bar{\zeta}_{(k_1,k_2),(t-1)}\sum_{r=1}^2 E\Big[\partial_1\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})|B\Big].
\end{aligned}$$

Replacing the expectations and the covariance $\bar{\zeta}_{(k_1,k_2),(t-1)}$ with their corresponding empirical versions leads to the unbiased estimator of $\big(\Sigma_0\big)_{(k_1,k_2),(t)}$,

$$\big(\widehat{\Sigma}_{0,(t)}(p)\big)_{(k_1,k_2)} = -\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t-1)}$$

$$+\frac{1}{p}\sum_{j=1}^{p}\prod_{r=1}^{2}\Big\{\eta(\widetilde{\beta}_{k_r,(t-1),j};\theta_{k_r,(t-1)})-\widetilde{\beta}_{k_r,(t-1),j}\Big\}$$
$$+\frac{\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t-1)}}{p}\sum_{j=1}^{p}\sum_{r=1}^{2}I\Big\{|\widetilde{\beta}_{k_r,(t-1),j}|\geq\theta_{k_r,(t-1)})\Big\}.$$

The consistency of the estimator $(\widehat{\Sigma}_{0,(t)}(p))_{(k_1,k_2)}$ follows since

$$\lim_{p\to\infty}\big(\widehat{\Sigma}_{0,(t)}(p)\big)_{(k_1,k_2)}=\lim_{p\to\infty}\big(\Sigma_{0,(t)}(p)\big)_{(k_1,k_2)}=\big(\Sigma_{(t)}\big)_{(k_1,k_2)}$$

holds with probability one for for all $k_1,k_2=1,\ldots,K$. The first equality follows by the unbiasedness of $\big(\widehat{\Sigma}_{0,(t)}(p)\big)_{(k_1,k_2)}$ for $(\Sigma_{0,(t)}(p))_{(k_1,k_2)}$, and the second equality holds by Lemma 2.1. The proof is completed by realizing the above equality shows almost sure convergence which indicates convergence in probability. □

Proof of Theorem 2.3

Proof. Under the assumption that $n>p$, the model-averaged estimator is unbiased. Hence

$$\mathrm{AMSE}(\widehat{\beta}_{\mathrm{MA}},\beta)=\lim_{n,p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\mathrm{Var}(\widehat{\beta}_{\mathrm{MA},j})\overset{a.s.}{=}w^{\top}\Sigma_{(\infty)}w,\tag{2.41}$$

where $\Sigma_{(\infty)}$ is a $K\times K$ matrix with (k_1,k_2)th component

$$\begin{aligned}\big(\Sigma\big)_{(k_1,k_2)}&=E\Big[\{I(B_0+\bar{\zeta}_{k_1}Z_{k_1})-B_0\}\{I(B_0+\bar{\zeta}_{k_2}Z_{k_2})-B_0\}\Big]\\&=\ \mathrm{Cov}\big(\bar{\zeta}_{k_1}Z_{k_1},\bar{\zeta}_{k_2}Z_{k_2}\big)=\mathrm{Cov}(Z_{k_1},Z_{k_2})\bar{\zeta}_{k_1}\bar{\zeta}_{k_2}.\end{aligned}\tag{2.42}$$

Combining (2.19) and (2.20), we obtain that

$$\begin{aligned}\bar{\zeta}_k&=\delta\Big\{E\big[\widetilde{G}(\varepsilon+\bar{\zeta}_kZ_k;b_k)^2\big]\Big\}^{1/2}\\&=\Big\{E\big[\widetilde{G}(\varepsilon+\bar{\zeta}_kZ_k;b_k)^2\big]\Big\}^{1/2}\Big\{E\big[\partial_1\widetilde{G}(\varepsilon+\bar{\zeta}_kZ_k;b_k)\big]\Big\}^{-1}.\end{aligned}\tag{2.43}$$

The expressions of the asymptotic variance of the model-averaged estimator in Theorem 2.3 hold by combining (2.41), (2.42), and (2.43).

□

Chapter 3

Componentwise confidence intervals and hypothesis testing in high dimensions – a computational approach

Recent literature on debiasing (desparsifying) the regularized estimators focuses on the debiased Lasso estimator and uses it to construct confidence intervals and hypothesis tests. In this chapter, we deeper investigate the estimator $\widetilde{\beta}$ from the robust approximate message passing algorithm, see Chapter 2. Based on the asymptotic normality of the estimator $\widetilde{\beta}$, we propose a computational approach constructing componentwise confidence intervals and hypothesis tests for the coefficients of sparse high dimensional models with growing dimensions. Numerical results show that the constructed componentwise confidence intervals and hypothesis tests have reasonable accuracy, especially in high-sparsity settings. In addition, a residual

bootstrap procedure is proposed for improving the accuracy of the confidence intervals and hypothesis tests in medium sparsity settings with small sample sizes.

This chapter is based on

Zhou, J. and Claeskens G. (2020). Componentwise confidence intervals and hypothesis testing in high dimensions – a computational approach. *Technical report.*

3.1 Introduction

Due to the simultaneous selection and estimation aspects of regularized procedures it is challenging to study the asymptotic properties of the resulting estimators. Indeed, at first the literature focused only on properties of the non-null part of the estimators, resulting in so-called oracle properties where one assumed that the selection has been done perfectly, e.g. Zou (2006). That is, one assumes that all true zero coefficients have been discarded by applying the regularization and that all true non-zero coefficients have been estimated by a non-zero value. Later, studying compressed sensing, Donoho et al. (2009); Bayati and Montanari (2011a); Donoho and Montanari (2016) found that the mean squared error of regularized estimators (the full vector, not only the non-zero part) could be obtained by using another estimation scheme, namely the approximate message passing algorithm. While such algorithm produces another estimator, that estimator converges to the regularized estimator in mean-square (provided some choices regarding tuning parameters hold), see discussion after (2.21) in Section 2.4 in Chapter 2. In Chapter 2 we followed Bradic (2016) in the construction of a robust approximate message passing algorithm in order to investigate the mean squared error of estimators resulting from minimizing an objective function formed by using a convex non-differentiable loss function in combination with an l_1-regularization. See Chapter 2 for a study of the mean squared error of regularized composite estimators and of model-averaged regularized estimators for high dimensional data. We

explicitly showcased the method for high dimensional quantile regression.

As a by-product of constructing the estimator via the (robust) approximate message passing algorithm, there is another sequence of estimators (one estimator for each iteration of the algorithm), which possesses interesting properties regarding bias and which can be compared to the debiased estimators that have been constructed earlier. Javanmard and Montanari (2014a); van de Geer et al. (2014) worked in high dimensional linear models with homoscedastic Gaussian errors and studied a debiasing (or desparsification) of the l_1-regularized least squares estimator, i.e. the Lasso estimator. The desparsified estimator in van de Geer et al. (2014) is obtained by inverting the Karush–Kuhn–Tucker characterization of the Lasso estimator. The debiased estimator in Javanmard and Montanari (2014a) is obtained by adding a term proportional to the subgradient of the l_1-regularizer of the Lasso method. Both approaches showed the asymptotic normality of the debiased (desparsified) Lasso assuming the sparsity s, the number of true non-zero coefficients, is of order $o(\sqrt{n}/\log p)$, and the sample size requirement for the asymptotic properties to hold is $s^2 \log p/n \to 0$ for constructing confidence regions. A joint coverage probability for multiple confidence intervals is taking care of by adjusting the nominal levels via a conservative Bonferroni correction method. Relaxing the sparsity condition $o(\sqrt{n}/\log p)$ in the above-mentioned work, Javanmard and Montanari (2018) showed that the debiased Lasso estimator is asymptotically Gaussian under $s = o(n/(\log p)^2)$. Li (2017) proposed a bootstrap double debiased Lasso estimator to further improve the sample size condition to $n \gg \max\{s \log p, (\tilde{s} \log p)^2\}$, where $\tilde{s}$ is the number of coefficients with magnitudes less than a certain threshold proportional to $\sqrt{\log p/n}$. The desparsification of the l_1-regularized estimators, on the one hand, reduces the bias caused by the shrinkage effect of the regularization and on the other hand, obtains an estimator with a limiting normal distribution, not being hindered by the presence of zeros for the components of the parameter vector that did not survive the selection.

While a study of the asymptotic mean squared error of the l_1-regularized

estimator is taken care of by an application of the RAMP algorithm, see Chapter 2, in this chapter we focus on a study of a debiased, or bias-corrected estimator obtained from the RAMP algorithm. We use this estimator to construct confidence intervals for components of the parameter, and for the construction of hypothesis tests. The quantile loss function is considered again as a working example. We evaluate the componentwise confidence intervals and hypothesis tests by both simulated data and audio signals. The method works well in the sense that the constructed confidence intervals and hypothesis tests are observed to have simulated coverage probabilities, respectively type I errors, close to the nominal values already when the sample size is small in settings with high sparsity.

The literature is not only limited to debiasing the Lasso estimators obtained from least squares loss functions. A popular robust alternative is to use quantile loss functions (Koenker and Bassett, 1978b; Koenker, 2005b). Debiasing the regularized quantile estimator was investigated in Zhao et al. (2014, 2019); Bradic and Kolar (2017). Zhao et al. (2014) considered a very similar debiasing procedure as in Javanmard and Montanari (2014a) for l_1-regularized composite quantile estimator by using a consistent approximation of the inverse of the covariance matrix of the predictive variables. The debiased estimator was used to construction confidence intervals. Simultaneous confidence intervals are constructed similar to Zhang and Cheng (2017) by using a Gaussian multiplier bootstrap. Alternatively, the debiasing term in Bradic and Kolar (2017) consists of the quantile loss function, an inverse of the covariance matrix of the predictive variables, and an estimator of the sparsity function (Tukey, 1965), which is shown to be equivalent to the derivative of the quantile function in Koenker (2005b). To estimate the sparsity function, the high dimensional rank score was developed. Further, uniform confidence bands were constructed based on a Bahadur representation of the debiased estimator.

Recent developments on simultaneous hypothesis testing also consider bootstrap procedures to accompany the debiasing approach. Belloni et al. (2018) construct simultaneous confidence bands for parameters of interest

in moment condition models with functional response data by bootstrapping only the linearized part of the asymptotic linear expansion of the bias. The moment condition models assume that the observations are generated by a known function with some parameters, and the expectation of the data generating function is equal to zero. A special case of the moment condition model is the l_1-regularized M-estimator. Similarly, Zhang and Cheng (2017) focuses on the debiased Lasso estimator which is asymptotically normal centering at the true coefficient vector. They bootstrap only the linearized part of the remainder term of the difference of debiased estimator and the true coefficients. They proposed for sparse high dimensional linear models with non-Gaussian distributed errors, a simultaneous test statistic performing multiple two-sided hypothesis tests. The test statistic is shown to be asymptotically nonconservative as compared to the Bonferroni adjustment in van de Geer et al. (2014); Javanmard and Montanari (2014a). Dezeure et al. (2017) proposed to construct the test statistic for the same multiple two-sided hypothesis tests in Zhang and Cheng (2017) using residual bootstrap, the non-Gaussian multiplier wild bootstrap, or the xyz-paired bootstrap for both homoscedastic and heteroscedastic residuals. Different from Zhang and Cheng (2017); Belloni et al. (2018), the method by Dezeure et al. (2017) considered bootstrapping the entire de-sparsified estimator; and the sparsity assumption on the design matrix was shown to be weaker than in Zhang and Cheng (2017).

In this chapter we propose to use a residual bootstrap procedure to improve the accuracy of the confidences intervals and hypothesis tests in medium sparsity settings.

Different from the previous literature, our approach relies heavily on the RAMP algorithm. Bradic (2016) incorporated arbitrary convex non-differentiable loss function to the general recursion in Bayati and Montanari (2011a), and the resulting algorithm approximates the l_1-regularized M-estimators. Chapter 2 (see also Zhou et al., 2019) further discussed the model-averaged version of Bradic (2016). However, the above two papers mainly focus on the asymptotic mean squared error (AMSE) of the regu-

larized estimators. A sequence at iteration t denoted as $\widetilde{\beta}_{(t)}$ in Chapter 2, Section 2.3.2, (2.13) was shown to be asymptotically normal centering at the true regression coefficient vector, see Section 2.5.2 before Corollary 2.2. Similar asymptotic normality of the sequence $\widetilde{\beta}_{(t)}$ with iteration index t in the AMP algorithm approximating the Lasso estimator, which is a special case of the l_1-regularized M-estimator, was argued in Mousavi et al. (2013, 2018); Javanmard and Montanari (2018). However, the above two works only used the asymptotic normality of $\widetilde{\beta}_{(t)}$ for deriving a Stein-type estimator for the AMSE.

In this chapter, we further investigate the sequence $\widetilde{\beta}_{(t)}$, $t = 1, 2, \ldots$ at convergence. Based on the asymptotic normality of the converged estimator $\widetilde{\beta}$, the confidence intervals and individual hypothesis tests are constructed componentwise for the true regression coefficient vector β. Multiple tests testing a group of individual hypothesis tests, are adjusted using a Holm-Bonferroni correction. Additionally, we consider the bootstrap in small sample cases.

3.2 Setup

We consider the same sparse high dimensional linear model $Y = X\beta + \boldsymbol{\varepsilon}$ as in Chapter 2, Section 2.2 with $Y, \boldsymbol{\varepsilon} \in \mathbb{R}^n$, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$ satisfying assumptions Condition (A1) - Condition (A5) in Chapter 2, Section 2.9. The components of $\boldsymbol{\varepsilon}$ are independent and identically distributed with cumulative distribution function F_ε and density function f_ε. The random variable B_0 has distribution F_{B_0} to which $\beta_j, j = 1, \ldots, p$ converges by assigning $1/p$ point mass to each component of β. We denote the number of non-zero components of β by its l_0 norm $s = \|\beta\|_0$; and we assume that the ratios $n/p \to \delta \in (0, 1)$, $n/s \to a \in (1, \infty)$, $s/p \to \omega = P(B_0 \neq 0)$ when $n, p, s \to \infty$.

We define l_1-regularized M-estimator $\widehat{\beta}$ as the solution to the following

minimization problem, that is,

$$\widehat{\beta}(\lambda) \quad = \quad \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho(Y_i - X_{i\cdot}^{\top}\beta) + \lambda\|\beta\|_1 \right\}, \tag{3.1}$$

where the nonnegative convex possibly non-differentiable loss function ρ satisfies Condition (A3). As in Chapter 2, we consider the (robust) approximate message passing algorithm (Bradic, 2016; Zhou et al., 2019; Bayati and Montanari, 2011a; Donoho and Montanari, 2016; Donoho et al., 2009) to obtain a sequence of estimators $\widehat{\beta}_{(t)}$, with iteration number $t = 1, 2, \ldots$, with the estimator at convergence denoted by $\widehat{\beta}$.

For completeness, we here briefly revise the main ingredients of this algorithm. To incorporate non-differentiable loss functions, the proximal mapping operator (Donoho and Montanari, 2016) with parameter $b > 0$ is used to adjust the residuals in the algorithm,

$$\mathrm{Prox}(z, b) = \arg\min_{x\in\mathbb{R}}\{b\rho(x) + \frac{1}{2}(x - z)^2\},\ b > 0;$$

and the effective score function (Donoho and Montanari, 2016) $\widetilde{G}(z; b) = b \cdot \partial\rho(x)|_{x=\mathrm{Prox}(z;b)}$, where $\partial\rho(x) = \{y : \rho(u) \geq \rho(x) + y(u - x), \forall u\}$ is the subgradient at non-differentiable points x and the gradient at differentiable points. To incorporate the sparsity s, the rescaled effective score function is defined as

$$G(z; b) = \delta\omega^{-1}\widetilde{G}(z; b).$$

The RAMP algorithm, with a fixed tuning parameter α and iterations indexed by t, starts from $\widehat{\beta}_{(0)} = 0$ and updates iteratively using the following three steps:

Step 1 **Adjusted residuals:** adjust the empirical residuals by

$$z_{(t)} = Y - X\widehat{\beta}_{(t)} + \\ n^{-1}G(z_{(t-1)}; b_{(t-1)}) \sum_{j=1}^{p} I\left\{\eta\big(\widehat{\beta}_{(t-1),j} + X_j^{\top}G(z_{(t-1)}; b_{(t-1)}); \theta_{t-1}\big) \neq 0\right\};$$

with $X_{\cdot j}$ being the jth column of X, I denotes the indicator function, and the soft-thresholding function $\eta(x;\theta) = \text{sign}(x) \cdot \max(|x| - \theta, 0)$.

Step 2 **Effective score:** choose the scalar $b_{(t)}$ such that the empirical average of the rescaled effective score function $G(z;b)$ has slope 1; update the tuning parameter $\theta_{(t)} = \alpha\bar{\zeta}_{\text{emp},(t)}$ where

$$\bar{\zeta}^2_{\text{emp},(t)} = \frac{1}{n}\sum_{i=1}^{n} G(z_{i,(t)}; b_{(t)})^2, \tag{3.2}$$

with a limit version, when $n, p \to \infty$, denoted by

$$\bar{\zeta}^2_{(t)} = E\Big[G(\varepsilon + \bar{\sigma}_{(t)}Z; b_{(t)})\Big], \tag{3.3}$$

with

$$\bar{\sigma}^2_{(t)} = \delta^{-1}E[(\eta(B_0 + \bar{\zeta}_{(t-1)}Z; \theta_{(t-1)}) - B_0)^2],$$

where $Z \sim N(0,1)$, B_0 is defined in Condition (A2), and ε is defined in Condition (A5).

Step 3 **Estimation:** update the estimator of β

$$\widehat{\beta}_{(t+1)} = \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}), \text{ where } \widetilde{\beta}_{(t)} = \widehat{\beta}_{(t)} + X^\top G(z_{(t)}; b_{(t)}). \tag{3.4}$$

Since a bias is introduced by applying the soft-thresholding function η in Step 3, the estimator $\widetilde{\beta}_{(t)}$ which is obtained before applying the thresholding can be interpreted as a debiased estimator. This estimator is of main interest in this chapter.

3.3 Componentwise confidence intervals

Instead of the sequence $\widehat{\beta}_{(t)}$ indexed by iteration t, the main focus of this chapter is the other sequence $\widetilde{\beta}_{(t)}$ in (3.4) in the three-steps iteration of the RAMP algorithm, which was argued in Chapter 2, Section 2.5.2, before Corollary 2.2 to converge weakly to $B_0 + \bar{\zeta}_{(t)}Z_{(t)}$, $n, p \to \infty$, where

B_0 is defined in Condition (A2), $\bar{\zeta}_{(t)}$ is the square-root of state evolution parameter in (3.3), and Z is an independent standard normally distributed variable. Thus, the conditional distribution $\widetilde{\beta}_{(t),j}|(B_0 = \beta_j)$ approximately follows a $N(\beta_j, \bar{\zeta}^2_{(t)})$ distribution. With the center being the true regression coefficient vector β, $\widetilde{\beta}_{(t)}$ is an asymptotically unbiased estimator for β with common variance $\bar{\zeta}_{(t)}$ for each component with $j = 1, \ldots, p$. Equivalently, let

$$\widetilde{B}_j = \frac{\widetilde{\beta}_j - \beta_j}{\bar{\zeta}}\Big|(B_0 = \beta_j),$$

we obtain the following asymptotic normality, when $p \to \infty$

$$\widetilde{B}_j \overset{d}{\approx} N(0,1),\ j = 1, \ldots, p. \tag{3.5}$$

Since the asymptotic normality in (3.5) is obtained by rewriting the conditional distribution $\widetilde{\beta}_{(t)}|(B_0 = \beta_j)$, our expression appears to be different from the common asymptotic normality expressions containing $\sqrt{n}$. In fact, the scaling term $\sqrt{n}$ is in the state evolution parameter $\bar{\zeta}_{(t)}$, see Bayati and Montanari (2011a, Proof of Lemma 1 (b), e.g. the last almost sure convergence on p775). The variance $\bar{\zeta}^2_{(t)}$ is the limit version of the estimator $\bar{\zeta}^2_{\text{emp},(t)}$ in (3.2), which is directly obtainable from the RAMP algorithm.

We denote the sequences $\widetilde{\beta}_{(t)}$ in (3.4) and $\bar{\zeta}^2_{\text{emp},(t)}$ in (3.2), with iteration index t, from a RAMP iteration at convergence by $\widetilde{\beta}$, $\bar{\zeta}^2_{\text{emp}}$. Then, componentwise confidence intervals can be constructed by plugging $\widetilde{\beta}_j$'s and $\bar{\zeta}^2_{\text{emp}}$ in the expression of $\widetilde{B}_j$'s (3.5). A confidence interval for β_j with asymptotic confidence level $1 - \alpha \in (0, 1)$ can be constructed as

$$\widehat{\text{CI}}_j(1-\alpha) = \big[\widetilde{\beta}_j - \Phi^{-1}(1-\alpha/2)\bar{\zeta}_{\text{emp}}, \widetilde{\beta}_j + \Phi^{-1}(1-\alpha/2)\bar{\zeta}_{\text{emp}}\big], \tag{3.6}$$

where Φ is the cumulative distribution function of the standard normal distribution.

3.4 Hypothesis testing

Sparse high dimensional models assume that there is a large proportion of the components of the true regression coefficient vector that is equal to zero and that there are only a few non-zero components. Consequently, in practice, we are often interested in testing if one (or a few) components are equal to zero. The classical approach is a two-sided hypothesis test for testing whether β_j is equal to 0. A test statistic can be constructed based on the asymptotic distribution of $\widetilde{\beta}_j$. Rejecting or not rejecting the null hypothesis at a given significance level can be decided upon using the p-value. We describe in this section first a two-sided individual testing problem where the null hypothesis states that β_j is equal to a given value $\beta_{0,j}$. Next, we construct a multiple test testing a family of hypothesis for $\{\beta_j, j \in M\}$, where $M \subseteq \{1, \ldots, p\}$.

The individual hypotheses are stated as follows. For each component $\beta_j, j = 1, \ldots, p$, we are interested in testing the null hypothesis

$$H_{0,j} : \beta_j = \beta_{0,j} \text{ versus } H_{a,j} : \beta_j \neq \beta_{0,j}.$$

The p-value of the test statistic

$$T_j = \frac{\widetilde{\beta}_j - \beta_{0,j}}{\bar{\zeta}_{\text{emp}}} \tag{3.7}$$

for $H_{0,j}$ can be computed using the conditional asymptotic normality in (3.5) in the following way by inserting the values $\widetilde{\beta}_j$ and $\bar{\zeta}^2_{\text{emp}}$ from the RAMP algorithm at convergence,

$$P_j = 2\Big(1 - \Phi\Big(\frac{|\widetilde{\beta}_j - \beta_{0,j}|}{\bar{\zeta}_{\text{emp}}}\Big)\Big). \tag{3.8}$$

As usual, for a given significance level α, the null hypothesis $H_{0,j}$ is rejected if $P_j \leq \alpha$.

In practice, we are often interested in testing multiple hypotheses $H_{0,j}$. To control the error rate of a multiple test, a procedure adjusting the

p-values (e.g., Bonferroni, Holm-Bonferroni (Holm, 1979), Šidák (Šidák, 1967), the Hochberg (Hochberg, 1988) procedure, etc.) should be considered in order to control the Type I error rate of the combined tests. Here, we follow the choice of van de Geer et al. (2014) and adjust the p-values by the Holm-Bonferroni procedure. For a family of hypotheses $\{H_{0,j}, j \in M\}$, $M \subseteq \{1, \ldots, p\}$ with a nominal probability of a type I error α, the adjustment works as follows:

(i) obtain the p-values P_j for testing the individual hypothesis $H_{0,j}$ by (3.8);

(ii) sort the p-values in ascending order and denote the sorted p-values by $P_{[j]}$;

(iii) the significance levels for the individual tests are adjusted to $\frac{\alpha}{m-[j]+1}$;

(iv) reject the null hypothesis $H_{0,[j]}$ if $P_{[j]} \leq \frac{\alpha}{m-[j]+1}$.

3.5 Simulation study

In this section, we investigate the finite sample performance of the componentwise confidence intervals in (3.6) in Section 3.3 and the hypothesis tests constructed in Section 3.4. These matters have not been investigated before by using the estimator $\widetilde{\beta}$ resulting from the RAMP algorithm.

We first describe the simulation procedure.

(i) Fix the dimension p, sample size n, compression ratio δ, and the number of non-zero components s. We randomly generate a sensing matrix X and a coefficient vector β, which are used in all replications for the same simulation setting. The components of the sensing matrix X are independent and generated from $N(0, 1/n)$. The subvector of β consisting of non-zero components is generated from a Dirac distribution with point mass equally distributed on -1 and 1, or from a $N(0, 1)$. We consider a high-sparsity setting with $s = 5$

and a medium sparsity setting with $s = 50$. We choose $p = 500$, $n = 250$ or 100, and accordingly $\delta = 0.5$ or 0.2.

(ii) In each simulation replication r for one setting, we generate an error vector $\boldsymbol{\varepsilon}_b$. The considered error distributions are the standard normal $N(0, 1)$, student-t with 3 degrees of freedom, and the mixture of normal distributions $0.5N(0, 1)+0.5N(5, 9)$. The errors are centered and rescaled to have standard deviation 0.2 after sampling.

(iii) Construct the response vector $Y_r = X\beta + \boldsymbol{\varepsilon}_r$.

(iv) Obtain the converged $\widetilde{\beta}_r$ by the RAMP algorithm; compute the $(1-\alpha)\%$ confidence interval for each component β_j by plugging $\widetilde{\beta}_{r,j}$ and $\bar{\zeta}^2_{\text{emp}}$ in (3.6).

(v) Repeat R' times Step 2 - 4; leave R converged records, and report averaged coverage probability for the subvectors of β consisting of non-zero components, of only the components that are zero and of the full vector β.

This procedure is tested for l_1-regularized quantile estimators at quantile level 0.5. We consider $R = 500$ replications for each setting, and choose the confidence level $1-\alpha$ to be 0.95 and 0.99. Example 95% componentwise confidence intervals of the subvector consisting of non-zero components of β in different settings are included in Figure 3.1. This example is chosen by taking the replication of which the coverage probability is the closest to 95% nominal coverage probability from $R = 500$ replications. The non-zero components of β for Figure 3.1 are randomly generated from $N(0, 1)$ using the random seed number 5; the same data will be reused for Figure 3.2. We observe that the true values of the non-zero components of the vector β, depicted using the red dots, are all located in the 95% confidence intervals in the high-sparsity settings with $s = 5$ for $N(0, 1)$ and t_3 distributed errors, and 2 out of 5 are located outside of the 95% confidence intervals for $0.5N(0, 1) + 0.5N(5, 9)$ distributed errors. In the medium sparsity settings with $s = 50$, there are 3 out of 50 non-zero components of

β which are not located in the 95% confidence intervals for $N(0,1)$ and t_3 distributed errors, and only 1 out of 50 non-zero component of β that are located outside of the 95% confidence intervals for $0.5N(0,1)+0.5N(5,9)$ distributed errors.

Since componentwise confidence intervals are based on the asymptotic normality of the estimator $\widetilde{B}_j, j=1,\ldots,p$ in (3.5), we also present example QQ-plots of $\widetilde{B}_j$'s, i.e., pooling $\widetilde{\beta}_j$'s corresponding to the components of β, and plot empirical quantiles of $\widetilde{B}_j$'s against theoretical standard normal quantiles. We use the same data, i.e., same β, $\widetilde{\beta}$, $\bar{\zeta}_{\text{emp}}$, for Figure 3.2. The non-zero components of β are generated from $N(0,1)$ under the seed number 5. Example QQ-plots for different settings are included in Figure 3.2, and all six plots, as well as p-values of the Shapiro-Wilk tests presented at the end of the titles of each plot in Figure 3.2, suggest normality of $\widetilde{B}_j$'s.

We also calculate the average coverage probabilities and average lengths of the confidence intervals. The average coverage probabilities $\widehat{\text{CP}}_{\text{vec}}(1-\alpha)$ for subvectors with length p_{vec} of the full vector β are calculated as follows

$$\widehat{\text{CP}}_{\text{vec}}(1-\alpha)=\sum_{j=1}^{p_{\text{vec}}}\widehat{\text{CP}}_j(1-\alpha)/p_{\text{vec}}, \tag{3.9}$$

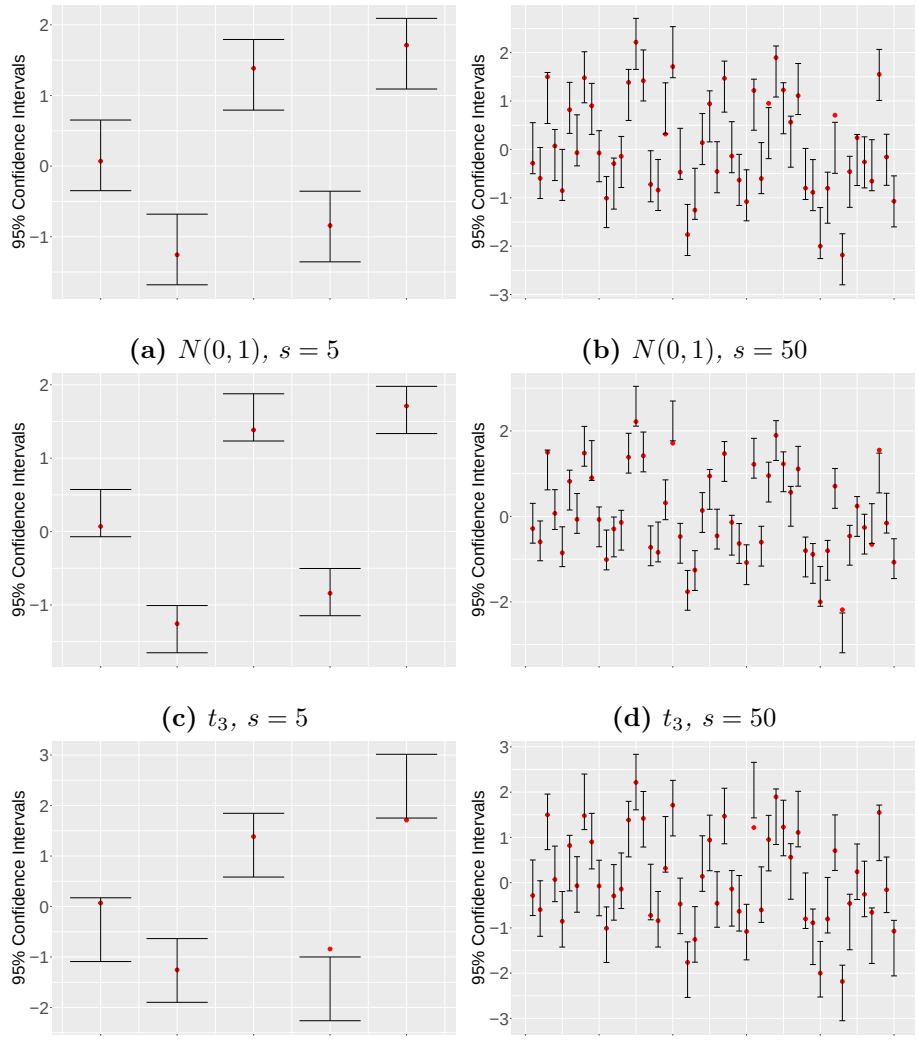
where

$$\widehat{\text{CP}}_j(1-\alpha)=\sum_{r=1}^{R} I\{\beta_j\in\widehat{\text{CI}}_{r,j}(1-\alpha)\}/R, \tag{3.10}$$

and the confidence interval of the jth component in the rth simulation replication $\widehat{\text{CI}}_{r,j}(1-\alpha)$ is obtained using (3.6). Since the confidence intervals for each component of β_j are constructed using a common variance $\bar{\zeta}^2_{\text{emp}}$ using (3.6), we obtain only one averaged length of the confidence intervals

$$\widehat{\mathcal{L}}(1-\alpha)=\frac{2\Phi(1-\alpha/2)}{R}\sum_{r=1}^{R}\bar{\zeta}_{\text{emp},r}.$$

The average coverage probabilities and averaged lengths of the confidence intervals, averaging over the components of β are presented in Table 3.1.

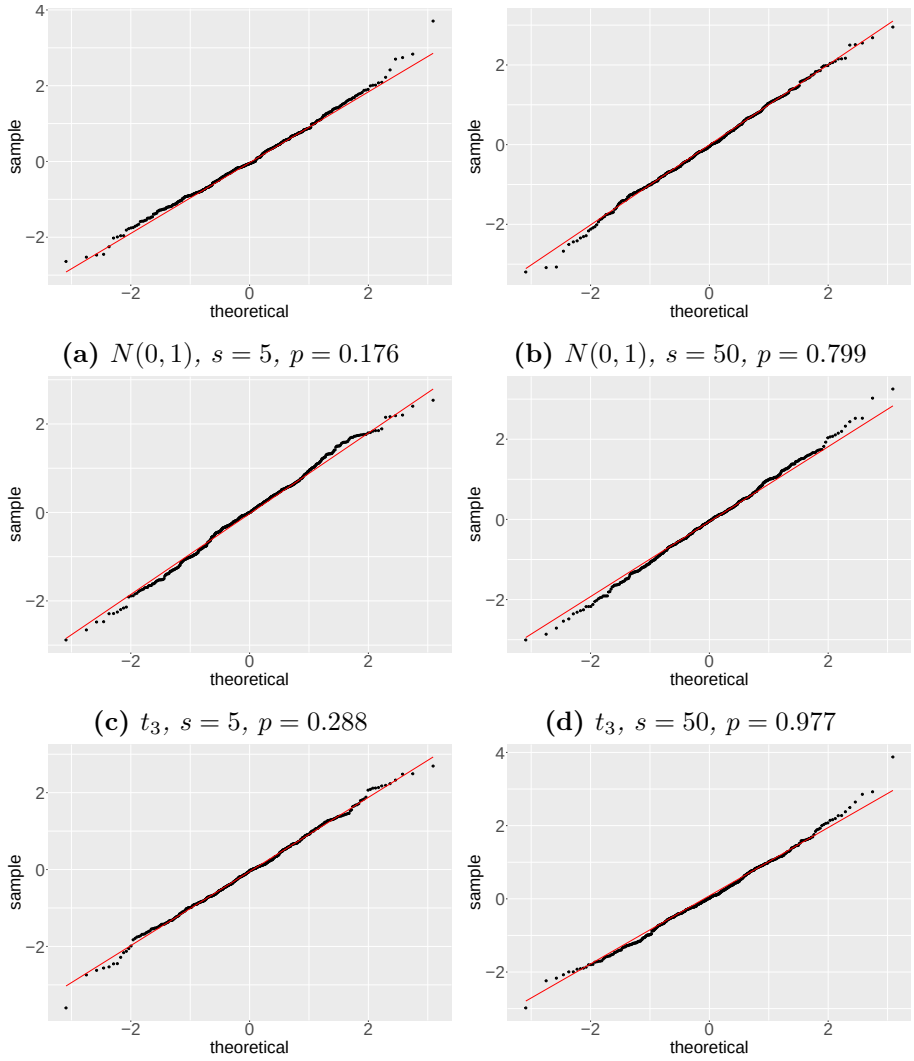


(a) $N(0,1)$, $s = 5$ **(b)** $N(0,1)$, $s = 50$

(c) t_3, $s = 5$ **(d)** t_3, $s = 50$

(e) $0.5N(0,1) + 0.5N(5,9)$, $s = 5$ **(f)** $0.5N(0,1) + 0.5N(5,9)$, $s = 50$

Figure 3.1: *Example 95% confidence intervals plots of non-zero components of β; the non-zero values are randomly generated from $N(0,1)$ using the seed number 5. The example is chosen by taking the replication of which the coverage probability is the closest to 95% nominal coverage probability from $R = 500$ replications. Plots for $s = 5$ are in the left column and for $s = 50$ are in the right column. Each row corresponds to plots for one error distribution, i.e., $N(0,1)$ – (a), (b), t_3 – (c), (d), $0.5N(0,1) + 0.5N(5,9)$ – (e), (f).*



(a) $N(0,1)$, $s = 5$, $p = 0.176$ **(b)** $N(0,1)$, $s = 50$, $p = 0.799$

(c) t_3, $s = 5$, $p = 0.288$ **(d)** t_3, $s = 50$, $p = 0.977$

(e) *mixed normal*, $s = 5$, $p = 0.654$ **(f)** *mixed normal*, $s = 50$, $p = 0.227$

Figure 3.2: *Example QQ-plots of* $\widetilde{B}_j$ *in* (3.5)*; non-zero components of* β *are randomly generated from* $N(0,1)$ *using the seed number 5. The example is chosen by taking the first replication from* $R = 500$ *replications. Plots for* $s = 5$ *are on the left column and for* $s = 50$ *are on the right column. Each row corresponds to plots for one error distribution, i.e.,* $N(0,1)$ *– (a), (b),* t_3 *– (c), (d),* $0.5N(0,1) + 0.5N(5,9)$ *– (e), (f). The p-values of the Shapiro-Wilks test for each setting are presented in the end of the titles of each plot.*

From Table 3.1, we observe that: (1) the zero subvector of β and the full vector β mostly have average coverage probabilities $\widehat{\mathrm{CP}}_{\mathrm{vec}}(1-\alpha)$ close to the nominal coverage probability $1-\alpha$ both for 0.95 and 0.99; (2) $\widehat{\mathrm{CP}}_{\mathrm{vec}}(1-\alpha)$ for the subvector consisting of the non-zero part of β is closer to the nominal coverage probability in the high-sparsity setting where $s = 5$, as compared to the medium sparsity setting where $s = 50$; (3) t_3 distributed errors have slightly shorter averaged lengths of the confidence intervals $\widehat{\mathcal{L}}(1-\alpha)$ and lower average coverage probabilities $\widehat{\mathrm{CP}}_{\mathrm{vec}}(1-\alpha)$ for both high and medium sparsity settings; (4) average coverage probabilities $\widehat{\mathrm{CP}}_{\mathrm{vec}}(1-\alpha)$ are closer to the nominal coverage probabilities and the lengths of confidence intervals $\widehat{\mathcal{L}}(1-\alpha)$ gets shorter, when increasing the sample size n from 100 to 250, especially in the medium sparsity settings where $s = 50$.

We also report the observed extrema and median calculated for the simulation study $\max_{j\in\{1,\ldots,p\}} \widehat{\mathrm{CP}}_j(1-\alpha)$, $\min_{j\in\{1,\ldots,p\}} \widehat{\mathrm{CP}}_j(1-\alpha)$, and $\mathrm{median}_{j\in\{1,\ldots,p\}}\widehat{\mathrm{CP}}_j(1-\alpha)$. Extrema and medians of the coverage probabilities of different settings are reported in Table 3.2 for $n = 100$ and $\delta = 0.2$, and in Table 3.3 for $n = 250$ and $\delta = 0.5$. In each simulation setting, the extrema suggest the worst empirical coverage probabilities of β_j's. The range of the maximum and minimum, as well as the ranges of the extrema and the nominal coverage probabilities, quantifies the worst deviations of the empirical coverage probabilities from the nominal coverage probabilities. A larger range suggests a more severe deviation from the nominal probability. We see a much wider range, i.e., minima and maxima deviate much farther from the medians, of $\widehat{\mathrm{CP}}_j(1-\alpha)$ in the median sparsity settings where $s = 50$. Differences due to different error distributions are not considerable when the sample size is small $n = 100$. However, t_3 distributed errors mostly lead to the largest range among three error distributions, when the sample size is increased to $n = 250$. We observe that large sample sizes lead to a smaller range of values of $\widehat{\mathrm{CP}}_j(1-\alpha)$ which are close to the corresponding nominal coverage probabilities in the medium sparsity settings, whereas the values $\widehat{\mathrm{CP}}_j(1-\alpha)$ in the high-sparsity set-

$n = 100$	$\delta = 0.2$	95% nominal				99% nominal			
		CP_{vec}			$\mathcal{L}$	CP_{vec}			$\mathcal{L}$
f_ε	s	non-zero	zero	full vector		non-zero	zero	full vector	
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**									
$N(0,1)$	5	0.96	0.95	0.95	0.91	0.99	0.99	0.99	1.20
	50	0.88	0.94	0.93	2.20	0.97	0.98	0.98	2.90
t_3	5	0.94	0.95	0.95	0.86	0.98	0.99	0.99	1.13
	50	0.88	0.93	0.93	2.19	0.97	0.98	0.98	2.88
$0.5N(0,1)$	5	0.96	0.95	0.95	0.91	0.99	0.99	0.99	1.20
$+0.5N(5,9)$	50	0.89	0.95	0.94	2.22	0.97	0.99	0.98	2.92
Subvector of β of nonzeros: $\mathbf{N(0,1)}$									
$N(0,1)$	5	0.92	0.95	0.95	0.88	0.98	0.99	0.99	1.16
	50	0.92	0.94	0.93	2.27	0.98	0.98	0.98	2.99
t_3	5	0.94	0.95	0.95	0.83	0.98	0.99	0.99	1.16
	50	0.92	0.94	0.93	2.23	0.98	0.98	0.98	2.94
$0.5N(0,1)$	5	0.92	0.95	0.95	0.88	0.98	0.99	0.99	1.16
$+0.5N(5,9)$	50	0.92	0.94	0.93	2.24	0.98	0.98	0.98	2.95
$n = 250$	$\delta = 0.5$	95% nominal				99% nominal			
		CP_{vec}			$\mathcal{L}$	CP_{vec}			$\mathcal{L}$
f_ε	s	non-zero	zero	full vector		non-zero	zero	full vector	
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**									
$N(0,1)$	5	0.93	0.95	0.95	0.90	0.98	0.99	0.99	1.18
	50	0.91	0.95	0.95	1.45	0.98	0.99	0.99	1.91
t_3	5	0.92	0.94	0.94	0.64	0.97	0.98	0.98	0.84
	50	0.91	0.95	0.94	1.36	0.98	0.99	0.99	1.79
$0.5N(0,1)$	5	0.94	0.95	0.95	1.16	0.99	0.99	0.99	1.52
$+0.5N(5,9)$	50	0.92	0.95	0.95	1.49	0.98	0.99	0.99	1.96
Subvector of β of nonzeros: $\mathbf{N(0,1)}$									
$N(0,1)$	5	0.93	0.95	0.95	0.89	0.99	0.99	0.99	1.17
	50	0.95	0.95	0.95	1.23	0.99	0.99	0.99	1.62
t_3	5	0.93	0.94	0.94	0.63	0.98	0.98	0.98	0.83
	50	0.95	0.95	0.95	1.04	0.99	0.99	0.99	1.36
$0.5N(0,1)$	5	0.94	0.95	0.95	1.16	0.99	0.99	0.99	1.53
$+0.5N(5,9)$	50	0.95	0.95	0.95	1.28	0.99	0.99	0.99	1.69

Table 3.1: *The average coverage probabilities* $CP_{\text{vec},j}(1-\alpha), j = 1,\ldots,p_{\text{vec}}$ *and average length* $\mathcal{L}(1-\alpha)$ *of confidence intervals of subvectors of* β *for* $n = 100$ $(\delta = 0.2)$ *and* $n = 250$ $(\delta = 0.5)$ *for* $1-\alpha = 0.95$ *and 0.99.*

tings are already close to the nominal confidence level when sample size is small.

Next we consider hypothesis testing for a sparse high dimensional model where we test whether $\beta_{0,j} = 0$. We denote the set of indices of non-zero and zero components of β by S with size s and $S^c = \{1, \ldots, p\} \backslash S$ with size $p - s$, respectively. To evaluate the performance of the individual hypothesis tests at significance level α, we calculate averaged false positive (FP) and true positive (TP) rates defined as

$$\mathrm{FP}(\alpha) = \frac{\sum_{j \in S^c} \sum_{r=1}^{R} I\{P_{r,j} \leq \alpha\}/R}{p - s},$$

and

$$\mathrm{TP}(\alpha) = \frac{\sum_{j \in S} \sum_{r=1}^{R} I\{P_{r,j} \leq \alpha\}/R}{s}.$$

Averaged FP and TP rates for different settings are presented in Table 3.4. We see that for high-sparsity settings where $s = 5$, the FP rates are already close to the nominal significance levels in cases where the sample size is small, i.e., $n = 100$. Also, the TP rates remain stable and have no significant improvement when the sample size n is increased to 250. However, in the medium sparsity settings where $s = 50$, the FP and TP rates are largely improved by increasing the sample size from 100 to 250. Another observation is that the TP rates in settings with t_3 distributed errors are mostly the highest among all three errors for the same sparsity s and the same distribution of non-zero components of β.

We consider a multiple test $\{H_{0,j}, j \in \{1, \ldots, p\}\}$ including all individual hypotheses. The multiple test is evaluated by the empirical version of the familywise error rate (FWER) defined as

$$\mathrm{FWER}(\alpha) = \frac{1}{R} \sum_{r=1}^{R} I\{\exists H_{0,j}^{(r)} \text{ is rejected at adjusted } \alpha, j \in S^c\},$$

and the rejection percentage (RP) observed in the simulation study is

95% nominal				
f_ε	s	non-zero	zero	full vector
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**				
$N(0,1)$	5	(0.95, 0.96, 0.96)	(0.61, 0.96, 1.00)	(0.61, 0.96, 1.00))
	50	(0.08, 0.97, 1.00)	(0.04, 1.00, 1.00)	(0.04, 1.00, 1.00)
t_3	5	(0.90, 0.93, 0.96)	(0.67, 0.96, 0.99)	(0.67, 0.96, 0.99)
	50	(0.09, 0.98, 1.00)	(0.03, 0.99, 1.00)	(0.03, 0.99, 1.00)
$0.5N(0,1)$	5	(0.93, 0.96, 0.98)	(0.62, 0.96, 1.00)	(0.62, 0.96, 1.00)
$+0.5N(5,9)$	50	(0.07, 0.99, 1.00)	(0.03, 1.00, 1.00)	(0.03, 1.00, 1.00)
Subvector of β of nonzeros: $\mathbf{N(0,1)}$				
$N(0,1)$	5	(0.88, 0.91, 0.96)	(0.77, 0.96, 0.99)	(0.77, 0.96, 0.99)
	50	(0.26, 1.00, 1.00)	(0.00, 1.00, 1.00)	(0.00, 1.00, 1.00)
t_3	5	(0.89, 0.95, 0.96)	(0.80, 0.96, 1.00)	(0.80, 0.96, 1.00)
	50	(0.26, 1.00, 1.00)	(0.00, 1.00, 1.00)	(0.00, 1.00, 1.00)
$0.5N(0,1)$	5	(0.89, 0.92, 0.95)	(0.81, 0.96, 0.99)	(0.81, 0.96, 0.99)
$+0.5N(5,9)$	50	(0.31, 1.00, 1.00)	(0.00, 1.00, 1.00)	(0.00, 1.00 1.00)
99% nominal				
f_ε	s	non-zero	zero	full vector
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**				
$N(0,1)$	5	(0.99, 0.99, 1.00)	(0.84, 0.99, 1.00)	(0.84, 0.99, 1.00)
	50	(0.44, 1.00, 1.00)	(0.21, 1.00, 1.00)	(0.21, 1.00, 1.00)
t_3	5	(0.97, 0.98, 0.99)	(0.87, 0.99, 1.00)	(0.87, 0.99, 1.00)
	50	(0.44, 1.00, 1.00)	(0.20, 1.00, 1.00)	(0.20, 1.00, 1.00)
$0.5N(0,1)$	5	(0.97, 0.99, 0.99)	(0.83, 0.99, 1.00)	(0.83, 0.99, 1.00)
$+0.5N(5,9)$	50	(0.46, 1.00, 1.00)	(0.20, 1.00, 1.00)	(0.20, 1.00, 1.00)
Subvector of β of nonzeros: $\mathbf{N(0,1)}$				
$N(0,1)$	5	(0.97, 0.98, 0.99)	(0.93, 0.99, 1.00)	(0.93, 0.99, 1.00)
	50	(0.65, 1.00, 1.00)	(0.13, 1.00, 1.00)	(0.13, 1.00, 1.00)
t_3	5	(0.97, 0.98, 0.99)	(0.93, 0.99, 1.00)	(0.93, 0.99, 1.00)
	50	(0.67, 1.00, 1.00)	(0.09, 1.00, 1.00)	(0.09, 1.00, 1.00)
$0.5N(0,1)$	5	(0.97, 0.98, 0.99)	(0.94, 0.99, 1.00)	(0.94, 0.99, 1.00)
$+0.5N(5,9)$	50	(0.71, 1.00, 1.00)	(0.08, 1.00, 1.00)	(0.08, 1.00, 1.00)

Table 3.2: *Extrema and medians of* $CP_{\text{vec},j}(1-\alpha), j = 1, \ldots, p_{\text{vec}}$. *Nominal coverage probabilities considered are 0.95 and 0.99. Values in the parentheses follow (minimum, median, maximum). For this table, we consider sample size* $\underline{n = 100}$ *with* $\underline{(\delta = 0.2)}$.

95% nominal				
f_ε	s	non-zero	zero	full vector
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**				
$N(0,1)$	5	(0.92, 0.93, 0.93)	(0.88, 0.95,0.99)	(0.88, 0.95, 0.99))
	50	(0.34, 0.97, 1.00)	(0.06, 0.98, 1.00)	(0.06, 0.98, 1.00)
t_3	5	(0.88, 0.93, 0.94)	(0.85, 0.94, 0.97)	(0.85, 0.94, 0.97)
	50	(0.30, 0.97, 1.00)	(0.05, 0.99, 1.00)	(0.05, 0.98, 1.00)
$0.5N(0,1)$	5	(0.93, 0.94, 0.96)	(0.88, 0.95, 0.98)	(0.88, 0.95, 0.98)
$+0.5N(5,9)$	50	(0.37, 0.97, 1.00)	(0.09, 0.98, 1.00)	(0.09, 0.98, 1.00)
Subvector of β of nonzeros: $\mathbf{N(0,1)}$				
$N(0,1)$	5	(0.91, 0.93, 0.97)	(0.86, 0.95, 0.98)	(0.86, 0.95, 0.98)
	50	(0.82, 0.96, 0.99)	(0.41, 0.97, 1.00)	(0.41, 0.97, 1.00)
t_3	5	(0.91, 0.93, 0.94)	(0.84, 0.94, 0.97)	(0.84, 0.94, 0.97)
	50	(0.85, 0.96, 0.99)	(0.39, 0.98, 1.00)	(0.39, 0.98, 1.00)
$0.5N(0,1)$	5	(0.92, 0.94, 0.96)	(0.88, 0.95, 0.98)	(0.88, 0.95, 0.98)
$+0.5N(5,9)$	50	(0.82, 0.96, 1.00)	(0.42, 0.97, 1.00)	(0.42, 0.97, 1.00)
99% nominal				
f_ε	s	non-zero	zero	full vector
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**				
$N(0,1)$	5	(0.98, 0.98, 1.00)	(0.96, 0.99, 1.00)	(0.96, 0.99, 1.00)
	50	(0.71, 0.99, 1.00)	(0.22, 1.00, 1.00)	(0.22, 1.00, 1.00)
t_3	5	(0.96, 0.97, 0.99)	(0.95, 0.98, 1.00)	(0.95, 0.98, 1.00)
	50	(0.67, 1.00, 1.00)	(0.20, 1.00, 1.00)	(0.20, 1.00, 1.00)
$0.5N(0,1)$	5	(0.98, 0.99, 1.00)	(0.96, 0.99, 1.00)	(0.96, 0.99, 1.00)
$+0.5N(5,9)$	50	(0.72, 1.00, 1.00)	(0.31, 1.00, 1.00)	(0.31, 1.00, 1.00)
Subvector of β of nonzeros: $\mathbf{N(0,1)}$				
$N(0,1)$	5	(0.98, 0.99, 1.00)	(0.96, 0.99, 1.00)	(0.96, 0.99, 1.00)
	50	(0.96, 0.99, 1.00)	(0.71, 0.99, 1.00)	(0.71, 0.99, 1.00)
t_3	5	(0.97, 0.97, 0.99)	(0.94, 0.98, 0.99)	(0.94, 0.98, 0.99)
	50	(0.96, 0.96, 1.00)	(0.71, 1.00, 1.00)	(0.71, 1.00, 1.00)
$0.5N(0,1)$	5	(0.98, 0.99, 1.00)	(0.96, 0.99, 1.00)	(0.96, 0.99, 1.00)
$+0.5N(5,9)$	50	(0.94, 0.94, 1.00)	(0.74, 1.00, 1.00)	(0.74, 1.00, 1.00)

Table 3.3: *Extrema and medians of $CP_{\text{vec},j}(1-\alpha), j = 1, \ldots, p_{\text{vec}}$. Nominal coverage probabilities considered are 0.95 and 0.99. Values in the parentheses follow (minimum, median, maximum). For this table, we consider sample size $n = 250$ with $(\delta = 0.5)$.*

defined as

$$\mathrm{RP}(\alpha) = \frac{1}{s} \sum_{j \in S} \Big\{ \sum_{r=1}^{r} I\{H_{0,j}^{(r)} \text{ is rejected at adjusted } \alpha\}/R \Big\}.$$

Values of FWER and RP for different settings are included in Table 3.4. We observe similar patterns as those shown for the individual hypothesis testing. In the high-sparsity settings where $s = 5$, values of FWER and RP are already stable when the sample size is small ($n = 100$), and are not significantly improved when increasing the sample size to $n = 250$. However, the FWERs are larger than the nominal significance level. In the medium sparsity settings where $s = 50$, values of FWER and RP can be largely improved by increasing the sample size.

3.6 Bootstrap confidence intervals for small samples

In Section 3.5, we have seen that the coverage probabilities of componentwise confidence intervals, the true positive and true negative rates of individual hypothesis tests and the familywise error rate and simulated rejection percentages for multiple tests can be improved for the medium sparsity setting with $s = 50$ when the sample size is small $n = 100$. The accuracy can be improved by increasing the sample size to $n = 250$. Here, we describe a residual bootstrap procedure, instead of using the confidence interval based on the asymptotic normal distribution of the estimator as in (3.6). For each replication r in the simulation procedure in Section 3.3, we state the following bootstrap procedure replacing Step 4. Bootstrap residuals can be constructed by a suitable estimator of β. Similar to Dezeure et al. (2017) who construct bootstrap residuals using the Lasso estimator, we use the converged vectors $\widehat{\beta}_r$ of the rth replication from the RAMP algorithm; the components of the $\widehat{\beta}_r$ are denoted by $\widehat{\beta}_{r,j}$. Another possible choice of an estimator of β for constructing bootstrap residuals is the value of $\widetilde{\beta}_r$ at convergence.

$n = 100$	$\delta = 0.2$	Significance $\alpha = 0.05$				Significance $\alpha = 0.01$			
f_ε	s	FP	TP	FWER	RP	FP	TP	FWER	RP
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**									
$N(0,1)$	5	0.05	0.99	0.10	0.63	0.01	0.96	0.05	0.49
	50	0.06	0.28	0.23	0.04	0.02	0.17	0.17	0.03
t_3	5	0.05	0.99	0.10	0.70	0.01	0.95	0.03	0.57
	50	0.07	0.29	0.26	0.04	0.02	0.18	0.19	0.03
$0.5N(0,1)$	5	0.05	0.99	0.10	0.63	0.01	0.97	0.04	0.47
$+0.5N(5,9)$	50	0.05	0.27	0.17	0.03	0.02	0.16	0.10	0.02
Subvector of β of nonzeros: $\mathbf{N(0,1)}$									
$N(0,1)$	5	0.05	0.80	0.10	0.67	0.01	0.76	0.03	0.64
	50	0.06	0.28	0.21	0.05	0.02	0.16	0.07	0.03
t_3	5	0.05	0.81	0.11	0.69	0.01	0.78	0.02	0.66
	50	0.07	0.27	0.22	0.05	0.02	0.16	0.08	0.04
$0.5N(0,1)$	5	0.05	0.80	0.11	0.67	0.01	0.76	0.04	0.64
$+0.5N(5,9)$	50	0.07	0.28	0.22	0.05	0.02	0.16	0.08	0.04
$n = 250$	$\delta = 0.5$	Significance $\alpha = 0.05$				Significance $\alpha = 0.01$			
f_ε	s	FP	TP	FWER	RP	FP	TP	FWER	RP
Subvector of β of nonzeros: **Dirac distribution at -1 and 1**									
$N(0,1)$	5	0.05	0.99	0.09	0.66	0.01	0.95	0.02	0.53
	50	0.05	0.67	0.20	0.13	0.01	0.47	0.07	0.07
t_3	5	0.06	1.00	0.20	0.96	0.02	1.00	0.10	0.93
	50	0.05	0.73	0.22	0.18	0.01	0.54	0.08	0.12
$0.5N(0,1)$	5	0.05	0.91	0.11	0.33	0.01	0.77	0.04	0.25
$+0.5N(5,9)$	50	0.05	0.65	0.16	0.12	0.01	0.46	0.06	0.07
Subvector of β of nonzeros: $\mathbf{N(0,1)}$									
$N(0,1)$	5	0.05	0.81	0.10	0.70	0.01	0.79	0.04	0.66
	50	0.05	0.56	0.06	0.27	0.01	0.46	0.02	0.22
t_3	5	0.07	0.82	0.21	0.79	0.02	0.81	0.12	0.77
	50	0.05	0.68	0.07	0.39	0.01	0.58	0.03	0.34
$0.5N(0,1)$	5	0.05	0.79	0.07	0.52	0.01	0.74	0.02	0.44
$+0.5N(5,9)$	50	0.05	0.60	0.10	0.28	0.01	0.49	0.03	0.23

Table 3.4: *Average FP and TP rates for individual hypothesis testing; as well as FWER and RP for multiple testing. Rates are calculated for* $n = 100$ $(\delta = 0.2)$ *and* $n = 250$ $(\delta = 0.5)$.

However, if $\widetilde{\beta}_r$, i.e., the vector which is linked to the debiased estimator of the regularized estimator of the rth simulation replications, is used to construct bootstrap residuals for the subsequent bootstrap replications, the RAMP algorithm estimating $\widehat{\beta}^*$ of the residual bootstrap replications has very low convergence rates (around 20%) in all settings. If $\widehat{\beta}_r$, i.e., the l_1-regularized estimator of the rth simulation replication, is used to construct the bootstrap residuals for the following bootstrap replication, convergence of the RAMP algorithm of the bootstrap replications is not an issue. Thus, we choose to use converged $\widehat{\beta}_r$'s here.

We consider B bootstrap replications for each simulation replication $r, r = 1, \ldots, R$, and choose $B = 500$ here. The procedure of simulation replications is described in Section 4.5.

(i) For the rth simulation replications with samples $(Y_{r,i}, X_{r,i\cdot})$, $i = 1, \ldots, n$, first obtain converged vector $\widehat{\beta}_r$; next, construct residuals by $\widehat{\varepsilon}_{r,i} = Y_{r,i} - X_{r,i\cdot}\widehat{\beta}_r$.

(ii) For each bootstrap replication b, sample $\widehat{\varepsilon}^*_{r,b,i}, i = 1, \ldots, n$ with replacement from $\widehat{\varepsilon}_{r,i}$'s; construct bootstrap pairs $(Y^*_{r,b,i}, X^*_{r,b,i\cdot}), i = 1, \ldots, n$ by $Y^*_{r,b,i} = X^*_{r,b,i\cdot}\widehat{\beta}_r + \varepsilon^*_{r,b,i}$. Then, obtain $\widetilde{\beta}^*_{r,b}$ at convergence;

(iii) Repeat steps 2 and 3 $B' > B$ times; keep $B = 500$ converged bootstrap records (convergence rates are in Table 2.4 in Chapter 2) to construct componentwise confidence intervals using the bootstrap quantiles in the following way: For the jth component of β, obtain the empircal quantile estimators at quantile levels $\alpha/2$ and $(1-\alpha/2)$ using the estimators $\widetilde{\beta}^*_{r,b,j}$'s corresponding to B times bootstrap replication. Denote the empirical $\alpha/2 \times 100\%$th and $(1-\alpha/2) \times 100\%$th quantile estimators of $\widetilde{\beta}^*_{r,b,j}$'s by $\widetilde{\beta}^*_{r,(\alpha/2),j}$ and $\widetilde{\beta}^*_{r,(1-\alpha/2),j}$, respectively. Then, the bootstrap componentwise confidence intervals for the jth component of β can be constructed as $[\widetilde{\beta}^*_{r,(\alpha/2),j}, \widetilde{\beta}^*_{r,(1-\alpha/2),j}]$, where $j = 1, \ldots, p$.

Tables 3.1 and 3.2 indicate that t_3 distributed errors in the medium sparsity setting $s = 50$, with a small sample size $n = 100$, mostly have

estimated coverage probabilities that are least close to the nominal coverage probabilities. We consider the above stated bootstrap procedure for the setting with t_3 distributed errors and the non-zero components of the true regression coefficient vector are generated from $N(0,1)$. To reduce the computational burden, we consider $R = 100$ simulation replications where a new dataset is generated in each replication. Further, we consider $B = 500$ bootstrap replication for each dataset. The coverage probabilities $\widehat{\beta}_{\text{vec}}$ and averaged length of the confidence intervals $\widehat{\mathcal{L}}(1-\alpha)$ are presented in Table 3.5 for $\alpha = 0.05, \text{or } 0.01$. Compared to Table 3.1 (2.23 for $\alpha = 0.05$ and 2.94 for $\alpha = 0.01$), the average lengths of the confidence intervals become shorter (1.28 for $\alpha = 0.05$ and 1.66 for $\alpha = 0.01$), which suggests a smaller standard deviation. The average coverage probability of the subvectors consisting of non-zero components decreases significantly from 0.92 (0.98) to 0.74 (0.84) for the nominal level $\alpha = 0.05$ (0.01). The averaged coverage probability of the zero subvectors of β increases from 0.94 (0.98) to 0.99 (1.00) for the nominal level $\alpha = 0.05(0.01)$. The averaged coverage probability of the full vector of β increases from 0.93 (0.98) to 0.97 (0.98) for $\alpha = 0.05$ (0.01). Notice that we only considered $R = 50$ times simulation replications combining $B = 500$ times bootstrap replications for each simulation replication r, whereas Table 3.1 is reported using $R = 500$ times simulation replications. In principle, results in Table 3.5 and 3.1 are not comparable. However, we do observe acceptable averaged coverage probabilities with narrower averaged confidence intervals by bootstrapping, which is worth further investigation.

Nominal $1-\alpha$	$\text{CP}_{\text{vec}}(\alpha)$			$\mathcal{L}(\alpha)$
	non-zero	zero	full vector	
95%	0.74	0.99	0.97	1.28
99%	0.84	1.00	0.98	1.66

Table 3.5: *The average coverage probabilities and averaged length of bootstrap confidence intervals of subvectors of β for t_3 distributed errors where $n = 100$ ($\delta = 0.2$).*

3.7 Sparse signal recovery

We consider again the audio wave signals example used in Chapter 2, Section 2.7.2. The artificial compressed sensing process involves a wavelet transform of the original audio signal for obtaining a "sparse" representation of $\beta \in \mathbb{R}^{2047}$. The artificial compressed sensing process is as follows:

(i) The sparse signal β is compressed by a randomly generated compression matrix $X \in \mathbb{R}^{1024\times 2047}$; components of the matrix X_{ij} are i.i.d. with a $N(0, 1/1024)$ distribution.

(ii) The compressed signal $X\beta$ is then sent to a receiver; the received signal Y from transmission is corrupted by error $\boldsymbol{\varepsilon}$. Components of the error vector $\boldsymbol{\varepsilon}$ are randomly generated from either t_3 or mixed normal distribution $0.5N(0, 1) + 0.5N(5, 9)$, and are rescaled to have standard deviation 0.03.the transformed audio signal.

In practice, we are interested in recovering the sparse signal β from the compression matrix X and the received signal Y. Here, the wavelet audio signal β is known, and is artificially compressed and corrupted.

We first construct componentwise confidence intervals. For a clearer presentation, we only plot confidence intervals of the last 20 entries of β, see Figure 3.3.

Notice that the componentwise confidence intervals are constructed based on the asymptotic normality $\widetilde{B}_j \sim N(0, 1), j = 1, \ldots, p$ in (3.5), where $\widetilde{B}_j$'s are constructed using the converged estimators $\widetilde{\beta}_j$'s of β_j's, i.e., the wavelet coefficients of the audio signal. We first construct $\widetilde{B}_j$'s using the components of the converged $\widetilde{\beta}$, which follow a standard normal distribution according to theory. To check the normality of $\widetilde{B}_j$'s, we plot the constructed $\widetilde{B}_j$'s against the standard normal Figure 3.4. From the figures, we see that the test statistic is roughly normal distributed for both t_3 and $0.5N(0, 1) + 0.5N(5, 9)$ distributed corruption errors. The Shapiro-Wilk test for normality gives p-values 0.460 for t_3 distributed errors and 0.615 for $0.5N(0, 1) + 0.5N(5, 9)$ distributed errors suggesting normality of

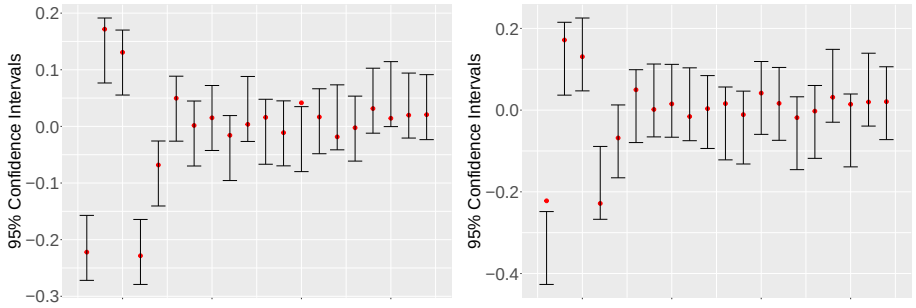


Figure 3.3: *The 95% confidence intervals of the last 20 entries of β. True values are depicted by red round dots. The plot on the left corresponds to t_3 distributed errors, and the plot on the right corresponds to $0.5N(0,1) + 0.5N(5,9)$ distributed errors.*

$\widetilde{B}_j$'s.

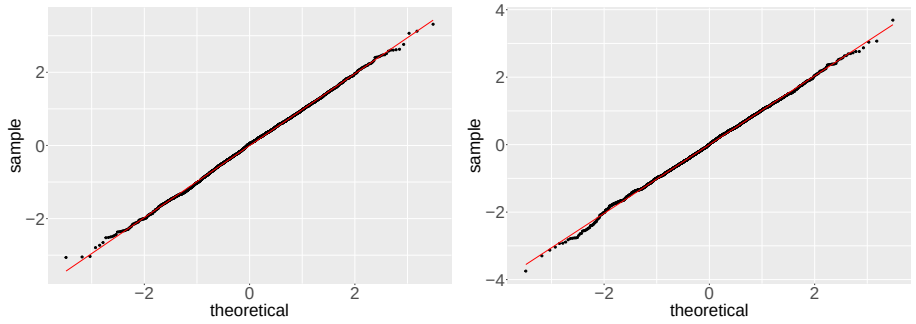


Figure 3.4: *QQ plots of the test statistic calculated by* (3.7)*. The plot on the left corresponds to t_3 distributed errors, and the plot on the right corresponds to $0.5N(0,1) + 0.5N(5,9)$ distributed errors.*

Next, we consider a multiple testing scenario including individual null hypotheses $H_{0,j} : \beta_j = 0$ versus two-sided alternative hypotheses $H_{a,j} : \beta_j \neq 0$ for all components of the wavelet coefficient β of the audio signal fraction. Since most β_j's are close to zero with countable non-negligible entries, we set cut-off values by taking the $(\tau/2)$th and $(1-\tau/2)$th empirical quantile of β_j's. Our goal is to identify β_j's with magnitude exceeding the two cut-off values by the multiple hypothesis test. We consider the

cut-off level $\tau = 0.01, 0.05$. The nominal significance levels of the test are $\alpha = 0.01$ or 0.05. We replicate the artificial compressed sensing process $R = 200$ times, and evaluate the performance of the multiple hypothesis test by the familywise error rate (FWER) and rejection percentage (RP). Table 3.6 reports the FWER and RP of the multiple test. We observe that when setting the level of the cut-off value τ to be 0.05 (i.e., select β_j's whose magnitude greater than 97.5% or less than 2.5% of β_j's magnitude), the FWER and RP of the test are both low for the significance level $\alpha = 0.01$ and 0.05. On the contrary, the FWER and RP of the test are both high when the cut-off level $\tau = 0.01$ (i.e., select β_j's whose magnitude exceed the range of 99% of β_j's magnitude). This observation is not surprising: The cut-off level τ decides if β_j's are counted as non-negligible entries; higher cut-off level τ results in less non-negligible entries. When an individual null hypothesis $H_{0,j}$ is rejected, it can be counted as a Type I error when $\tau = 0.01$, resulting in high FWER, but counted as a correct rejection when $\tau = 0.05$. This situation happens for β_j's whose magnitudes are not high enough to be counted as non-negligible entries when $\tau = 0.01$, but are counted as non-negligible entries when $\tau = 0.05$. Similarly, when magnitudes of β_j's are low enough for rejecting $H_{0,j}$ but high enough to exceed cut-off level 0.05, the Type II error increases resulting in low RP.

	Cut-off level τ	0.05		0.01	
	Significance α	0.05	0.01	0.05	0.01
t_3	FWER	0.11	0.11	0.93	0.81
	RP	0.22	0.20	0.88	0.85
$0.5N(0,1)$	FWER	0.06	0.00	0.38	0.14
$+0.5N(5,9)$	RP	0.13	0.11	0.58	0.51

Table 3.6: *FWER and RP of multiple hypothesis test finding variables with magnitude exceeding certain cut-off values. The cut-off values are determined by $(\tau/2)$th and $(1-\tau/2)$th quantile of β_j's. Significance levels α considered are 0.05 and 0.01.*

3.8 Discussion

This chapter focuses on the estimator $\widetilde{\beta}$ from the RAMP algorithm at convergence in Chapter 2, and is the first one proposing a computational approach constructing componentwise confidence intervals and conducting hypothesis tests for l_1-regularized estimators. The proposed confidence intervals and hypothesis tests are shown to have fair accuracy, especially in high-sparsity settings.

Since Chapter 2 discusses the model-averaged and composite versions of the l_1-regularized estimators, a natural extension would be componentwise confidence intervals and hypothesis testing for the model-averaged and composite estimators. Further, we have seen in the simulation section that the accuracy of the confidence intervals and hypothesis tests depends on the sample size and the sparsity. It would be interesting to derive theory on the optimal sample size given a certain sparsity.

For multiple testing, we considered a conservative Bonferroni adjustment following van de Geer et al. (2014); Javanmard and Montanari (2014a). It is of interest to develop and study other non-conservative adjustments.

Also, the numerical results in the residual bootstrap section do not show a clear evidence that bootstrapping improves accuracy (e.g., averaged coverage probabilities) of the componentwise confidence intervals when the sample size is small. A theoretical study on the consistency of the bootstrap confidence intervals is worth investigating.

Chapter 4

C-vine copula based quantile regression

Classical approaches on quantile regression make strict distributional and moment assumptions on the linear model, which are usually violated in practice. We propose here a flexible nonparametric quantile regression approach based on C-vine copulas without imposing additional assumptions and allowing for conditional heteroscedasticity. As a subclass of regular vine copulas, C-vine copulas formulate multivariate copulas by using only bivariate copulas, which is referred to as the pair-copula construction. The proposed algorithm incorporates a new variable selection approach, namely a "one-step-ahead" selection, by maximizing the conditional log-likelihood of one level of the C-vine tree sequence taking also the next level of trees into account. The performance of the proposed method is evaluated in both low and high dimensional settings using simulated and real data.

This chapter is based on:

Zhou, J., Tepegjozova M., Claeskens G. and Czado, C.(2020). Sparse C- and D-vine copula selection in high dimensions. *Technical report.*

4.1 Introduction

As a robust alternative for the ordinary least squares regression which estimates the conditional mean of a response variable given the predictive variables, quantile regression (Koenker and Bassett, 1978b) focuses on the conditional quantiles. This method has been studied extensively in statistics, economics and finance. Under strict distributional assumptions on the variables, properties such as asymptotic normality and consistency of the estimator are obtained (Koenker, 2005b). Several extensions of quantile regression exist, such as adapting quantile regression in the Bayesian framework (Yu and Moyeed, 2001), for longitudinal data (Koenker, 2004), time-series models (Xiao and Koenker, 2009), high dimensional models with l_1-regularizer (Belloni and Chernozhukov, 2011), to nonparametric estimation by kernel weighted local linear fitting (Yu and Jones, 1998) and by additive models (Koenker, 2011; Fenske et al., 2011), etc. Most of these approaches are derived under strict model assumptions which are often violated in practice.

In addition to the above-mentioned issue, Bernard and Czado (2015) addressed other potential concerns such as quantile crossings and model-misspecification when the dependence structure of the response variables and the predictive variables does not follow Gaussian copulas for Gaussian distributed margins. Hence, flexible models without assuming specific distributions for the data, homoscedasticity, nor a linear relationship between response and predictive variables, are of interest and motivate research on modeling conditional quantiles using copulas, see also Noh et al. (2013,

2015); Chen et al. (2009). In cases where the number of predictive variables is large, a pair-copula construction, namely a vine copula, of a multivariate copula is developed in Joe (1996); Bedford and Cooke (2001, 2002) through recursive conditioning. By this pair-copula construction, a multivariate density function can be constructed using only bivariate copula densities and marginal densities. Different pair-copula constructions can be obtained by changing the order of conditioning. A crucial assumption, a.k.a, the simplifying assumption, is often made for reducing computational burden and increasing modelling flexibility. This assumption states that the conditional copula does not depend on the value of the conditioning variable. Haff et al. (2010); Stoeber et al. (2013) further investigated the validity of this assumption and found that it is not a severe restriction in practice. By assuming that the conditional copula depends on the conditioning variables only through the conditional densities, we obtain simplified pair-copula constructions, i.e., simplified vine copulas. A graphical representation was developed in Bedford and Cooke (2001, 2002) for representing different pair-copula constructions based on graph theory. A general construction using such pair copulas can be represented by a regular (R-) vine tree sequence using nodes and edges; edges in each level of trees correspond to bivariate copulas in the pair-copula construction, and become nodes in the next tree. Two subclasses of R-vines are canonical (C-) vines and drawable (D-) vines, i.e., in each level of trees, C-vine copulas follow a star structure with a root node connected to all other nodes, and D-vine copulas follows a sequential structure. For recent literature about predicting conditional quantiles using simplified vine copulas, see Kraus and Czado (2017) using D-vines and Chang and Joe (2019) using R-vines. The proposed method in this chapter uses C-vines and incorporates a search algorithm maximizing the conditional log-likelihood step by step in the sequence of trees of which the C-vines are composed.

Similar to the method of Kraus and Czado (2017) which is based on D-vine copulas, we propose in this chapter a flexible nonparametric approach estimating conditional quantiles using C-vine copulas without strict model

assumptions. Since this approach does not set strict restriction to the underlying copulas between the predictive variables and the response variable (Bernard and Czado, 2015, also see discussion above), issues such as quantile crossings are avoided. The main contribution which is novel compared to other papers is a "one-step-ahead" approach maximizing the conditional log-likelihood, of which the next level of trees are taken into account in each level of root node selection. Additionally, all marginal densities and copulas are estimated nonparametrically, thus this allows more flexibility as opposed to parametric specifications and the construction permits a large variety of dependence structures, resulting in a well-performing conditional quantile estimator.

This chapter is organized as follows. Section 4.2 introduces the general setup and concept of C-vine copulas. Sections 4.4.2 describes in detail the construction of the conditional quantile estimator using C-vine copulas. Since all densities involved in the model construction are estimated nonparametrically, we investigate in Theorem 4.1 the consistency of the conditional quantile esimator for given variables orders. The new "one-step-ahead" approach maximizing the conditional log-likelihood gradually is described in detail in Section 4.4. Performance of the constructed C-vine based conditional quantile estimator is evaluated in Section 4.5 by several quantile related measurements in various simulation settings, and compared with the D-vine based quantile estimator in Kraus and Czado (2017). We also evaluate the constructed C-vine based quantile estimator using low and high dimensional datasets in Section 4.6.

4.2 Setup

In a regression framework some predictive variables in the p-vector $X = (X_{\cdot 1}, \ldots, X_{\cdot p})$ have predictive ability for the response $Y \in \mathbb{R}$. Denote the cumulative distribution function of Y as F_Y, the marginal cumulative distribution function of $X_{\cdot j}$ as F_j for $j \in \{1, \ldots, p\}$, and the joint distribution function of (Y, X) as F. Further we assume that F_Y, F_j, and F are con-

tinuous, then there exists a copula C associated with the joint distribution of (Y, X) such that

$$F(y, x_1, \ldots, x_p) = C(F_Y(y), F_1(x_1), \ldots, F_p(x_p))$$

with the joint density function

$$f(y, x_1, \ldots, x_p) = c(F_Y(y), F_1(x_1), \ldots, F_p(x_p)) \cdot f_Y(y) \cdot f_1(x_1) \cdot \ldots \cdot f_p(x_p).$$

To inspect the dependence structure of (Y, X), we transform all variables marginally to a uniform distribution on the unit interval $[0, 1]$ (the so-called u-scale) by the probability integral transform (PIT). The corresponding transformed variables are defined as $V = F_Y(Y)$ and $U_j = F_j(X_{\cdot j}), j = 1, \ldots, p$ which are uniformly distributed on $[0, 1]$. The joint distribution of $(V, U) = (V, U_1, \ldots, U_p)$ is a copula denoted by $C_{V,1,\ldots,p}$ with density $c_{V,1,\ldots,p}$.

Additionally, we introduce some notations following Kraus and Czado (2017); Czado (2019). Let a set $D \subset \{1, \ldots, p\}$ be the index set of a sub-vector of the p-vector X, such that X_D is a sub-vector of X consisting of the variables indexed by elements in D. For example, the sub-vector is denoted as $X_{\{1,2,3,4\}}$ with the corresponding transformed sub-vector denoted as $U_{\{1,2,3,4\}}$, when $D = \{1, 2, 3, 4\}$. Let $j, j' \in \{1, \ldots, p\} \backslash D$ be two indices not belonging to D. Then we have the following definitions.

(i) $C_{j,j';D}$ (respectively $C_{V,j;D}$) denotes the copula associated with the conditional distribution of $(X_{\cdot j}, X_{\cdot j'})$ (respectively $(Y, X_{\cdot j})$) conditioning on X_D with the corresponding copula density $c_{j,j';D}$ ($c_{V,j;D}$).

(ii) $F_{j|D}$ (respectively $F_{Y|D}$) denotes the conditional distribution of $X_{\cdot j}$ (respectively Y) conditional on X_D.

(iii) $C_{j|D}$ (respectively $C_{V|D}$) denotes the conditional distribution of the PIT transformed variable U_j (respectively V) conditional on U_D with corresponding density function $c_{j|D}$ (respectively $c_{V|D}$).

(iv) The h-functions $h_{U_j|U_{j'}}(u_j|u_{j'}) = \frac{\partial}{\partial u_{j'}} C_{j,j'}(u_j, u_{j'})$ and $h_{V|U_j}(v|u_j) = \frac{\partial}{\partial u_j} C_{V,j}(v, u_j)$. Similarly, for the conditioned case,

$$h_{U_j|U_{j'};U_D}(u_j|u_{j'}; u_D) = \frac{\partial}{\partial u_{j'}} C_{j,j';D}(u_j, u_{j'}; u_D),$$

and

$$h_{V|U_j;U_D}(v|u_j; u_D) = \frac{\partial}{\partial u_j} C_{V,j;D}(v, u_j; u_D).$$

Further, for unconditional copulas, $h_{U_j|U_{j'}}(u_j|u_{j'}) = C_{j|j'}(u_j|u_{j'})$ and $h_{V|U_j}(v|u_j) = C_{V|j}(v|u_j)$. When the simplifying assumption holds, that is, the copula function does not depend on the specific conditioning value, then $h_{U_j|U_k;U_l}(u_j|u_k; u_l) = \frac{\partial}{\partial u_k} C_{jk;l}(u_j, u_k)$.

In this chapter we focus on C-vines (Bedford and Cooke, 2001, 2002) which have the advantage of expressing a joint density function by a product of bivariate functions with a special structure, which are easy to handle. The continuous joint density function of $(Y, X_{\cdot 1}, \ldots, X_{\cdot p})$ for the ordered sequence $X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot p}, Y$, by Czado (2019, Thm 4.8), is written as a product of conditional bivariate copula densities and marginal densities in the following way

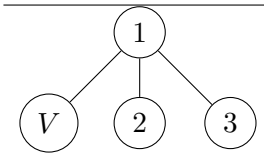
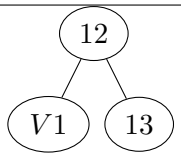
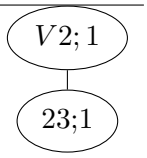
$$f(y, x_1, \ldots, x_p) = \Big[\prod_{j=1}^{p} f_j(x_j) \prod_{j=1}^{p-1} \prod_{j'=1}^{p-j} c_{j,j+j';1,\ldots,j-1} \Big] \Big[f_Y(y) \prod_{j=1}^{p} c_{j,V;1,\ldots,j-1} \Big]. \tag{4.1}$$

This pair-copula construction (Bedford and Cooke, 2001, 2002), is called a C-vine density. When all marginal densities in (4.1) are uniform, the density is a C-vine copula density. An example of a graphical representation of a C-vine is presented in Table 4.1, where each edge corresponds to a pair-copula, and all edges (nodes) are covered by the pair-copulas in the decomposition.

To simplify the notation in the graphical representation, we use the subscripts in (4.1) to denote the nodes (edges) in the graph. For example, "12" denotes the bivariate copula of U_1 and U_2, whereas "23;1" denotes

the bivariate copula of U_2 and U_3 conditional on U_1, etc.

The main focus of this chapter is quantile regression based on C-vine

Tree 1	Tree 2	Tree 3
		

$$c_{V|123}(v|u_1,u_2,u_3) = c_{V1}(v,u_1)\cdot c_{V2;1}(C_{V|1}(v|u_1),C_{2|1}(u_2|u_1))$$
$$\cdot c_{V3;12}(C_{V|1,2}(v|u_1,u_2),C_{3|12}(u_3|u_1,u_2))$$

Table 4.1: *Conditional density of V given U_1, U_2, U_3. Notations in the graphs should be interpreted as follows: j in Tree 1 represents U_j, V stands for the variable V; "12" ("$V1$") in Tree 2 represents the bivariate copula of U_1 and U_2 (resp. V and U_1); "$V2;1$" in Tree 3 represents the bivariate copula of V and U_2 given U_1.*

copulas, discussed in detail in Section 4.3.1, which relies on the conditional distribution $F_{Y|1,\dots,p}$. Let the index $j'' \in D$ and $D_{-j''} = D\backslash\{j''\}$. The pair-copula construction specified for C-vines provides a strategy to estimate the conditional distribution $F_{Y|1,\dots,p}$ by using only the pair-copulas in (4.1), in concert with the following recursion (Joe, 1997)

$$F_{Y|D} = h_{V|j'';D_{-j''}}\Big(F_{Y|D_{-j''}}(y|x_{D_{-j''}})\Big|F_{X_{j''}|D_{-j''}}(x_{j''}|x_{D_{-j''}})\Big).$$

Example 4.3.1 shows that the conditional distribution $F_{Y|1,2,3}$ can be expressed iteratively by a sequence of h-functions $h_{V|U_3;U_1,U_2}$, $h_{V|U_2;U_1}$, $h_{U_3|U_2;U_1}$, $h_{V|U_1}$, $h_{U_2|U_1}$, $h_{U_3|U_1}$, $h_{U_2|U_1}$.

The main contribution of this chapter is a new variable ordering algorithm searching for a root node order of the C-vines described in detail in Section 4.4. Compared to Kraus and Czado (2017); Chang and Joe (2019), the new algorithm maximizes the truncated conditional log-likelihoods of each level of trees taking also the edges of the next level of trees into account.

4.3 C-vine copula based quantile regression model

4.3.1 Conditional quantile function

The main interest in using vine copula-based quantile regression is to predict the $\tau \in (0,1)$ quantile of the response variable Y given X, which can be achieved by a joint modeling of (Y, X) and by using the conditional quantile function $q_\tau(x_{\{1,\dots,p\}}) = F^{-1}_{Y|1,\dots,p}(\tau|x_{\{1,\dots,p\}})$. To motivate the construction of a C-vine copula, we first give the conditional distribution of Y given X on the u-scale and its corresponding inverse, the quantile function, as derived in Kraus and Czado (2017),

$$F_{Y|1,\dots,p}(y|x_{\{1,\dots,p\}}) = C_{V|1,\dots,p}(v|u_{\{1,\dots,p\}}),$$

$$F^{-1}_{Y|1,\dots,p}(\tau|x_{\{1,\dots,p\}}) = F^{-1}_Y\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big). \tag{4.2}$$

The above equations indicate that the conditional quantile of Y given X can be expressed as a composition of the inverse marginal distribution function F^{-1}_Y with the conditional quantile function $C^{-1}_{V|1,\dots,p}$. Estimating the conditional quantile of Y can be achieved by replacing in (4.2) the unknown functions by estimates, that is, we use estimated inverses of the marginal distributions, denoted by $\hat{F}^{-1}_Y$, $\hat{F}^{-1}_j, j = 1,\dots,p$, and an estimated conditional quantile function $\hat{C}^{-1}_{V|1,\dots,p}$.

4.3.2 Variable ordering in C-vines

When C-vines are used to estimate the conditional distribution $C_{V|1,\dots,p}$ and its inverse function $C^{-1}_{V|1,\dots,p}$, an ordering of the variables U_j's is required. This ordering determines the structure of the C-vine. Details on the ordering procedure based on maximizing the conditional log-likelihood of V, are covered in section 4.4.

We demonstrate the connection between the variable ordering and the root node order using the graphical representation of a 4-dimensional C-vine in Table 4.2 which shows the C-vine structure corresponding to a

specific variable order.

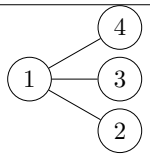
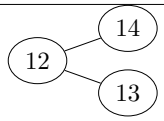
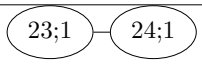
Tree	Root node	Number of choices for the root node
 Tree 1	"1"	4
 Tree 2	"12"	3
 Tree 3	"23;1" or "24;1"	$4 \cdot 3 = 12$

Table 4.2: *Illustration of a variable order in C-vines ($d = 4$). To simplify the notation in the graphical representation, variable U_j is denoted by its subscript j in the graph; e.g., "1" in Tree 1 represents variable U_1, "12" in Tree 2 represents the bivariate copula of U_1 and U_2, "23;1" in Tree 3 represents the bivariate copula of U_2 and U_3 given U_1.*

In the regression framework we require V, the response on the u-scale, to be one of the leaves, not the central node. The specific variable ordering that is used in the C-vine determines how the conditional density and conditional quantile of V given U are estimated, see (4.2). Both functions, $C_{V|1,\dots,p}$ and its inverse $C^{-1}_{V|1,\dots,p}$, can be expressed by means of a product of bivariate copulas.

First we explain the used notation. Denote by $C_{V|1,\dots,p-1}$ the conditional cumulative distribution function of V given $U_1 = u_1, \dots, U_{p-1} = u_{p-1}$, and by $C_{p|1,\dots,p-1}$ the conditional cumulative distribution function of U_p given $U_1 = u_1, \dots, U_{p-1} = u_{p-1}$. Further, denote by $c_{V,p|1,\dots,p-1}$ the conditional density of V and U_p given $U_1, \dots, U_{p-1}$, and by $c_{V,p;1,\dots,p-1}$ the copula density of $c_{V,p|1,\dots,p-1}$.

To obtain the C-vine representation of the conditional density of V given $U_1 = u_1, \dots, U_p = u_p$, we start from

$$c_{V|1,\dots,p}(v|u_{\{1,\dots,p\}})$$

$$= c_{V,p;1,\ldots,p-1}\Big(C_{V|1,\ldots,p-1}(v|u_{\{1,\ldots,p-1\}}), C_{p|1,\ldots,p-1}(u_p|u_{\{1,\ldots,p\}})\Big) \times c_{V|1,\ldots,p-1}(v|u_{\{1,\ldots,p-1\}}). \tag{4.3}$$

By recursively rewriting the conditional density $c_{V|1,\ldots,p-1}(v|u_{\{1,\ldots,p-1\}})$ in (4.3) in a similar way, we obtain the representation of the C-vine in (4.4) consisting of only bivariate copulas,

$$c_{V|1,\ldots,p}(v|u_{\{1,\ldots,p\}}) = c_{V,1}(v,u_1) \times \prod_{p'=2}^{p} c_{V,p';1,\ldots,p'-1}\big(C_{V|1,\ldots,p'-1}(v|u_{\{1,\ldots,p'-1\}}), C_{p'|1,\ldots,p'-1}(u_p|u_{\{1,\ldots,p'-1\}})\big). \tag{4.4}$$

The order of the indices in (4.4) corresponds to a specific variable order from U_1 to U_p; reshuffling the predictive variables will lead to a different copula density representation of V given $U_1 = u_1, \ldots, U_p = u_p$. Obviously, the C-vine representation is not unique.

An example of the conditional density of V given U corresponding to a 4-dimensional C-vine with a predetermined order of variables, is presented in Table 4.1. The conditional quantile function $C^{-1}_{V|1,\ldots,p}$ is rewritten recursively using the inverse of the h-functions (Czado, 2019; Joe, 1997). Example 4.3.1 illustrates a decomposition under the simplifying assumption for the 4-dimensional C-vine of Table 4.1.

Example 4.3.1. *Consider the C-vine in Table 4.1, the conditional distribution of V given $(U_1, U_2, U_3)^\top$ can be expressed using h-functions as*

$$\begin{aligned} C_{V|1,2,3}(v|u_1,u_2,u_3) &= h_{V|U_3;U_1,U_2}\Big(C_{V|1,2}(v|u_1,u_2)|C_{3|1,2}(u_3|u_1,u_2)\Big) \\ &= h_{V|U_3;U_1,U_2}\Big(h_{V|U_2;U_1}\big(C_{V|1}(v|u_1)\big|C_{2|1}(u_2|u_1)\big)\Big|h_{U_3|U_2;U_1}\big(C_{3|1}(u_3|u_1)\big|C_{2|1}(u_2|u_1)\big)\Big) \\ &= h_{V|U_3;U_1,U_2}\Big(h_{V|U_2;U_1}\big(h_{V|U_1}(v|u_1)\big|h_{U_2|U_1}(u_2|u_1)\big)\Big|h_{U_3|U_2;U_1}\big(h_{U_3|U_1}(u_3|u_1)\big|h_{U_2|U_1}(u_2|u_1)\big)\Big). \end{aligned}$$

The conditional quantile function of V given U_1, U_2, U_3 at quantile level τ is

$$\begin{aligned} &C^{-1}_{V|1,2,3}(\tau|u_1,u_2,u_3) \\ &= h^{-1}_{V|U_1}\left(h^{-1}_{V|U_2;U_1}\left(h^{-1}_{V|U_3;U_1,U_2}\big(\tau\big|h_{U_3|U_2;U_1}\big(h_{U_3|U_1}(u_3|u_1)|h_{U_2|U_1}(u_2|u_1)\big)\big)\Big|h_{U_2|U_1}(u_2|u_1)\right)\Big|u_1\right). \end{aligned} \tag{4.5}$$

4.3.3 Nonparametric estimators of the copula densities and h-functions

Here, we explain how the h-functions are estimated in a nonparametric way. By the simplifying assumption, $h_{V|U_j;U_D}(v|u_j;u_D) = \frac{\partial}{\partial u_j} C_{V,j;D}(v,u_j)$ can be estimated by the pair (V,U_j) conditioned on U_D. Recall that U_D is the subvector of U of which the components' indices belong to the set $D \subset \{1,\dots,p\}$. Thus, it is sufficient to show the estimation procedure for the h-function $h_{V|U}$. Consider a pair (U,V) of which the h-function $h_{V|U} = C_{V|U}$ can be estimated by the following rescaled estimator

$$\hat{C}_{V|U}(v|u) = \int_0^v \hat{c}_{V,U}(\tilde{v},u)\mathrm{d}\tilde{v} \Big/ \int_0^1 \hat{c}_{V,U}(\tilde{v},u)\mathrm{d}\tilde{v}, \tag{4.6}$$

where $\hat{c}_{V,U}$ is the nonparametric estimator of the bivariate copula density of (U,V).

Example estimators of the copula density $\hat{c}_{V,U}$ are the transformation estimator (Charpentier et al., 2007), the transformation local likelihood estimator (Geenens et al., 2017), the tapered transformation estimator (Wen and Wu, 2015), the beta kernel estimator (Charpentier et al., 2007), and the mirror-reflection estimator (Gijbels and Mielniczuk, 1990). Among the above-mentioned kernel estimators, the transformation local likelihood estimator (Geenens et al., 2017) considered in this chapter, was found by Nagler et al. (2017) to have an overall best performance. The estimator is implemented in the R packages `kdecopula` and `rvinecopulib` using Gaussian kernels.

For completeness we revise the construction of the transformation local likelihood estimator below. Let the $n \times 2$ transformed sample matrix be

$$D = (S,T) \tag{4.7}$$

where the transformed samples $D_i = \big(S_i = \Phi^{-1}(U_i), T_i = \Phi^{-1}(V_i)\big), i = 1,\dots,n$, and Φ denotes the Gaussian cumulative distribution function. The logarithm of the density $f_{S,T}$ of the transformed samples $(S_i,T_i), i =$

$1, \ldots, n$ is approximated locally by a polynomial expansion $P_{\boldsymbol{a}_m}$ of order m with intercept $\tilde{a}_{m,0}$ such that the approximation denoted by

$$\tilde{f}_{S,T}(\Phi^{-1}(u), \Phi^{-1}(v)) = \exp\left\{\tilde{a}_{m,0}(\Phi^{-1}(u), \Phi^{-1}(v))\right\}.$$

The transformation local likelihood estimator is then defined as

$$\tilde{c}(u, v) = \frac{\tilde{f}_{S,T}(\Phi^{-1}(u), \Phi^{-1}(v))}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))}. \tag{4.8}$$

To get the local polynomial approximation, we need a kernel function $\boldsymbol{K}$ with 2×2 bandwidth matrix $\boldsymbol{B}_n$. For some pair $(\check{s}, \check{t})$ close to (s, t), $\log f_{ST}(\check{s}, \check{t})$ is assumed to be well approximated, locally, by, for instance, a polynomial with $m = 1$ (log-linear)

$$P_{\boldsymbol{a}_1}(\check{s} - s, \check{t} - t) = a_{1,0}(s, t) + a_{1,1}(s, t)(\check{s} - s) + a_{1,2}(s, t)(\check{t} - t),$$

or $m = 2$ (log-quadratic)

$$\begin{aligned} P_{\boldsymbol{a}_2}(\check{s} - s, \check{t} - t) &= a_{2,0}(s, t) + a_{2,1}(s, t)(\check{s} - s) + a_{2,2}(s, t)(\check{t} - t) \\ &+ a_{2,3}(s, t)(\check{s} - s)^2 + a_{2,4}(s, t)(\check{t} - t)^2 + a_{2,5}(s, t)(\check{s} - s)(\check{t} - t). \end{aligned}$$

The estimated coefficient vector

$$\begin{aligned} \tilde{\boldsymbol{a}}_m(s, t) = \arg\max_{a_m} \Bigg\{ &\sum_{i=1}^{n} \boldsymbol{K}\left(\boldsymbol{B}_n^{-1/2} \begin{pmatrix} s - S_i \\ t - T_i \end{pmatrix}\right) P_{\boldsymbol{a}_m}(S_i - s, T_i - t) \\ &- n \int\int_{\mathbb{R}^2} \boldsymbol{K}\left(\boldsymbol{B}_n^{-1/2} \begin{pmatrix} s - \check{s} \\ t - \check{t} \end{pmatrix}\right) \exp\left(P_{\boldsymbol{a}_m}(\check{s} - s, t - t)\right) \mathrm{d}\check{s}\mathrm{d}\check{t} \Bigg\} \end{aligned} \tag{4.9}$$

While it is well-known that kernel estimators suffer from the curse of dimensionality, in the C-vine construction only two-dimensional functions need to be estimated, this thus avoids problems with high-dimensionality. We next explain as in Geenens et al. (2017) how a bandwidth selection is obtained. Consider the principal component decomposition for the $n \times 2$

sample matrix $D = (S, T)$ in (4.7), such that

$$(Q, R)^\top = WD^\top \tag{4.10}$$

where each row of W is an eigenvector of $D^\top D$. We obtain an estimator of f_{ST} through the density estimator of f_{QR} which can be estimated based on a diagonal bandwidth matrix $\text{diag}(h_Q^2, h_R^2)$. Selecting the bandwidths h_Q uses samples $Q_i, i = 1, \ldots, n$ as

$$h_Q = \arg\min_{h>0} \left\{ \int_{-\infty}^{\infty} \left\{ \tilde{f}_Q^{(p)} \right\}^2 dq - \frac{2}{n} \sum_{i=1}^{n} \tilde{f}_{Q(-i)}^{(p)}(\hat{Q}_i) \right\} \tag{4.11}$$

where $\tilde{f}_Q^{(p)} (p = 1, 2)$ are the local polynomial estimators for f_Q, and $\tilde{f}_{Q(-i)}^{(p)}$ is the "leave-one-out" version of $\tilde{f}_Q^{(p)}$ computed by leaving out Q_i. The procedure of selecting h_R is similar. The bandwidth matrix for the bivariate copula density is then given by $\boldsymbol{B}_n = K_n^{(p)} W^{-1} \text{diag}(h_Q^2, h_R^2) W^{-1}$ where $K_n^{(p)}$ takes $n^{1/45}$ for the local log-quadratic case ($p = 2$). Selection for the k-nearest-neighbour type bandwidth is similar. The k-nearest-neighbour bandwidths denoted by h'_Q and h'_R are obtained by restricting the minimization in (4.11) in the interval $(0, 1)$, i.e.,

$$h'_Q = \arg\min_{h'_Q \in (0,1)} \left\{ \int_{-\infty}^{\infty} \left\{ \tilde{f}_Q^{(p)} \right\}^2 dq - \frac{2}{n} \sum_{i=1}^{n} \tilde{f}_{Q(-i)}^{(p)}(\hat{Q}_i) \right\}.$$

Estimating f_{QR} at any (q, r) is obtained by using its $k = K_n^{(p)} \cdot h'_Q \cdot n$ nearest neighbours where $K_n^{(p)}$ takes $n^{-4/45}$ for $p = 2$.

4.3.4 Consistency of the conditional quantile estimator

Example 4.3.1 gives an intuition on expressing the conditional quantile function $C^{-1}_{V|1,\ldots,p}$ recursively by the inverse of h-functions for fixed variable orders. The conditional quantile function on the original scale in (4.2) requires the inverse of the marginal distribution function of Y. Following Kraus and Czado (2017); Noh et al. (2013), the marginal cumulative

distribution functions F_Y and the F_j's are estimated nonparametrically to reduce the bias caused by model misspecification. When also all inverses of the h-functions are estimated nonparametrically, we establish the consistency of the conditional quantile estimator $\hat{F}^{-1}_{Y|1,\dots,p}$ in Proposition 4.1 for fixed variable orders.

Proposition 4.1. *Let the inverse of the marginal distribution functions F_Y and F_j, $j = 1, \dots, p$ be uniformly continuous and estimated nonparametrically, and let the inverse of the h-functions expressing the conditional quantile estimator $C^{-1}_{V|1,\dots,p}$ be uniformly continuous and estimated nonparametrically in the interior of the support of bivariate copulas, i.e., $[\delta, 1-\delta]^2, \delta \to 0_+$.*

1. *If estimators of the inverse of marginal functions $\hat{F}^{-1}_Y$, $\hat{F}^{-1}_j$, $j = 1, \dots, p$, are uniformly strong consistent on the support $[\delta, 1-\delta], \delta \to 0_+$, and the estimators of the inverse of h-functions composing the conditional quantile estimator $C^{-1}_{V|1,\dots,p}$ are uniformly strong consistent, then the estimator $\hat{F}^{-1}_{Y|1,\dots,p}(\tau|x_{\{1,\dots,p\}})$ is also uniformly strong consistent.*

2. *If estimators of the inverse of marginal functions $\hat{F}^{-1}_Y$, $\hat{F}^{-1}_j$, $j = 1, \dots, p$, are at least weak consistent, and the estimators of the inverse of h-functions are also at least weak consistent, the the estimator $\hat{F}^{-1}_{Y|1,\dots,p}(\tau|x_{\{1,\dots,p\}})$ is weak consistent.*

Remark: The uniform strong consistency result in 1 requires additionally that all estimators of the inverse of marginal distributions involved, i.e., $\hat{F}^{-1}_Y$, $\hat{F}^{-1}_j$'s, are uniformly strong consistent on a truncated compact interval $[\delta, 1-\delta], \delta \to 0_+$. Although not directly used in the proof of Proposition 4.1 in Section 4.8, the truncation is an essential condition for guaranteeing the uniform strong consistency of all estimators of the inverse of the marginal distributions (i.e., estimators of quantile functions), see Cheng (1995); Van Keilegom and Veraverbeke (1998); Cheng (1984).

The inverse of the nonparametric estimators above inherit the consistency of the original nonparametric estimators F_Y, F_j's, and the h-

functions. Example of nonparametric estimators estimating the marginal distributions F_Y and F_j's are the continuous kernel smoothing estimator (Parzen, 1962) and the transformed local likelihood estimator in the univariate case (Geenens, 2014). When choosing a Gaussian kernel function for the nonparametric estimators in Parzen (1962); Geenens (2014), estimators of the marginal distributions F_Y, F_j's are uniformly strong consistent.

The inverse of the h-functions can be obtained through a nested sequence of bivariate copula densities (see section 4.4.2 for details). Here, we consider the transformed local likelihood estimator in Geenens et al. (2017), which inherits uniformly (weak or strong) consistency from the estimator $\tilde{f}_{S,T}$ in (4.8). Using a product of univariate Gaussian kernels for the kernel function $\boldsymbol{K}$ in (4.9) for estimating $\tilde{f}_{S,T}$ is discussed in Geenens et al. (2017) and implemented in the R packages `kdecop` and `rvinecopulib`.

Proposition 4.1 shows the uniform consistency and gives an indication on the performance of the conditional quantile estimator $\hat{F}^{-1}_{Y|1,\dots,p}$ for fixed variable orders, while combining the consistent estimators of F_Y, F_j's and the bivariate copula densities. Extensive studies on numerical performance of $\hat{F}^{-1}_{Y|1,\dots,p}$ will be presented in Section 4.5.

4.3.5 Implementation

Implementation of the quantile prediction is based on the R package `vinereg` (Nagler and Kraus, 2018). An outline of the main estimation procedure for predicting the conditional quantile τ of the response Y given predictive variables X is described in the following five steps. Our main contribution in this implementation compared to Kraus and Czado (2017); Chang and Joe (2019) is the variable ordering algorithm in Step 2 as described in Section 4.4, taking trees of the next level into consideration while maximizing the conditional log-likelihoods for selecting a root node for each level of trees. An additional contribution is the pair-copulas construction complying with the C-vine copulas structure.

Step 1 Obtain nonparametric estimators of marginal distributions of predictive variables $X_{\cdot j}$'s and the response variable Y, denote the nonparametric estimators by $\hat{F}_j, j = 1, \ldots, p$, and $\hat{F}_Y$. Then, transform all variables marginally by $U_j = \hat{F}_j(X_{\cdot j}), j = 1, \ldots, p$, and $V = \hat{F}_Y(Y)$.

Step 2 Select a variable order by maximizing the conditional log-likelihood of V based on the selection procedure in Section 4.4.

Step 3 Estimate the C-vine based on the selected variable order in Step 2.

Step 4 Reuse the bivariate copulas estimated in Step 3 to evaluate the inverse h-functions in (4.5) at new samples.

Step 5 Reuse the estimated marginal distribution of Y to transform the predicted conditional quantile from the u-scale to its original scale, i.e., from $C^{-1}_{V|1,\ldots,p}$ to $F^{-1}_{Y|1,\ldots,p}$ by (4.2).

4.4 Procedure of variable ordering

The main idea of ordering variables follows Czado et al. (2012) maximizing the conditional log-likelihood gradually; though we propose in this chapter a "one-step-ahead" maximization approach by taking trees of the next level into account when selecting the root node for each level. Denote by k_1^* the index of that variable among the predictive variables having the highest dependency with all other variables including the response variable V; then by conditioning on $U_{k_1^*}$, the variable indexed by k_2^* among the remaining variables has the highest dependency with other variables. The goal is to find an order of variables in a C-vine while taking into account the dependency with V. In other words, we find an order for the predictive variables $U_1, \ldots, U_p$ according to the maximum of the values of the conditional log-likelihood of V given the ordered U_j's.

Meanwhile, (4.4) suggests that the order of U_j's that V conditions on, decides the values of the conditional density $c_{V|1,\ldots,p}(v|u_{\{1,\ldots,p\}})$ and con-

structs different C-vines. To find the order of the variables, we maximize the conditional log-likelihood of V sequentially. The first variable is selected by maximizing $\sum_{i=1}^{n} \log c_{V|1,\ldots,p}(v_i|u_{i,k_1}, u_{i,j})$ over $k_1 \in \{1,\ldots,p\}$, where $j \in \{1,\ldots,p\}\backslash\{k_1\}$; by this method, we select the root node with the highest dependence between the response variable and one of the remaining variables. It is worth mentioning that, instead of summing up all j ranging in $\{1,\ldots,p\}\backslash\{k_1\}$ while calculating the conditional log-likelihood for each k_1, we calculate the conditional log-likelihood for a single j at a time here; in total $p \cdot (p-1)$ log-likelihoods are calculated in this step, since we have p choices for k_1. Denote the selected first variable as $U_{k_1^*}$, the second variable is selected by maximizing $\sum_{i=1}^{n} \log c(v_i|u_{i,k_1^*}, u_{i,k_2}, u_{i,j})$ where $j \in \{1,\ldots,p\}\backslash\{k_1^*, k_2\}, k_2 \in \{1,\ldots,p\}\backslash\{k_1^*\}$. Under this regime, we order the variables such that their conditional dependence with the response variable is taken into account.

4.4.1 Pre-selection based on partial correlation

Constructing an appropriate C-vine and finding the root node order relies on the pairwise dependence structure. We allow both $p \leq n$ and $p > n$. However, we need to calculate $p \cdot (p-1) + (p-1) \cdot (p-2) + \ldots + 2 \cdot 1 = (p-1) \cdot p \cdot (p+1)/3$ times a log-likelihood before completing the ordering of the variables; and this becomes computationally demanding in cases where $p > n$. To reduce the computational complexity, we perform a pre-selection based on the partial correlation measures, motivated by the fact that the correlation measures such as Kendall's τ and Spearman's ρ can be expressed in terms of pair-copulas (Haff et al., 2010). In the latter context, we consider Spearman's rho denoted as $\rho_{k_t,V;k_1^*,\ldots,k_{t-1}^*}$ where $k_t \in \{1,\ldots,p\}\backslash\{k_1^*,\ldots,k_{t-1}^*\}$ and the k_t^*'s are the selected variables in the previous steps. An exception is the pre-selection for the first variable where the partial correlation measures are replaced by unconditional correlation measures.

The partial Spearman's ρ is calculated using Theorem 2.14 in Czado

(2019) by

$$\rho_{j_1,j_2;\{1,\ldots,p\}\setminus\{j_1,j_2\}} = \frac{\rho_{j_1,j_2;\{1,\ldots,p-1\}\setminus\{j_1,j_2\}} - \rho_{j_1,p;\{1,\ldots,p-1\}\setminus\{j_1\}}\rho_{j_2,p;\{1,\ldots,p-1\}\setminus\{j_2\}}}{\sqrt{1-\rho^2_{j_1,p;\{1,\ldots,p-1\}\setminus\{j_1\}}}\sqrt{1-\rho^2_{j_2,p;\{1,\ldots,p-1\}\setminus\{j_2\}}}}, \tag{4.12}$$

where $\rho_{j_1,j_2;\{1,\ldots,p'\}\setminus\{j_1,j_2\}}$ denotes the correlation of U_{j_1} and U_{j_2} given $U_j, j \in \{1,\ldots,p'\}\setminus\{j_1,j_2\}$, $\rho_{j',p';\{1,\ldots,p'-1\}\setminus\{j_1\}}$ denotes the correlation of $U_{j'}$ and $U_{p'}$ given $U_j \in \{1,\ldots,p'-1\}\setminus\{j_1\}$, $j' = j_1, j_2$, $p' = p-1, p$. Estimation is facilitated with estimated quantities based on the PIT transformed data V and U_j's. We show an example of partial correlation calculation for an additional variable reduction in higher dimensions in the construction of the first C-vine tree in Section 4.4.2.

4.4.2 Construction in more detail

First C-vine tree

We first select K candidate variables with indices $k_{1,1},\ldots,k_{1,K}$ as the candidate of the first root variable. The selection is based on the unconditional correlation of the variables U_j and the response V. The conditional log-likelihood allowing for two truncated C-vine likelihoods with root nodes $U_{k_{1,r}}$ and $U_{j|k_{1,r}}$ based on data $\{(V_i, U_{i,k_{1,r}}, U_{i,j}), i = 1,\ldots,n\}$ is given as

$$\begin{aligned}\sum_{i=1}^{n} l_{i,k_{1,r},j}(v_i, u_{i,k_{1,r}}, u_{i,j}) = \sum_{i=1}^{n} \Big\{ &\log c_{V,k_{1,r}}(v_i, u_{i,k_{1,r}}) \\ &+ \log c_{V,j;k_{1,r}}(C_{V|k_{1,r}}(v_i|u_{i,k_{1,r}}), C_{j|k_{1,r}}(u_{i,j}|u_{i,k_{1,r}}))\Big\},\end{aligned} \tag{4.13}$$

for $j \in \{1,\ldots,p\}\setminus\{k_{1,r}\}, r = 1,\ldots,K$.

The truncated C-vine likelihood in (4.13) takes one step ahead as compared to Kraus and Czado (2017) by taking an additional variable U_j into

consideration. The first selected variable corresponds to

$$\arg \max_{\substack{k_{1,r} \\ r\in\{1,\ldots,K\}}} \max_{\substack{j \\ j\in\{1,\ldots,p\}\setminus\{k_{1,r}\}}} \sum_{i=1}^{n} l_{i,k_{1,r},j}(v_i, u_{i,k_{1,r}}, u_{i,j}).$$

To evaluate the log-likelihood here, the main components are the copula density $c_{V,k_{1,r}}$, the conditional c.d.f's, i.e., $C_{V|k_{1,r}}$ and $C_{j|k_{1,r}}$, and the conditional copula density $c_{V,j;k_{1,r}}$. The copula density $c_{V,k_{1,r}}$ can be estimated nonparametrically by kernel estimators of the copula density, and the conditional copula density $c_{V,k_{1,r}}$ can be estimated similarly once we obtain the pseudo-samples

$$u_{i,V|k_{1,r}} = C_{V|k_{1,r}}(v_i|u_{i,k_{1,r}})$$

and

$$u_{i,j|k_{1,r}} = C_{j|k_{1,r}}(u_{i,j}|u_{i,k_{1,r}})$$

based on evaluating the conditional c.d.f.'s $C_{V|k_{1,r}}$ and $C_{j|k_{1,r}}$ at the ith observations of V and the U_j's.

The variable among the K candidates which has the largest estimated conditional log-likelihood defined in (4.13) is chosen, and the corresponding index of the chosen variable is denoted as k_1^*. Meanwhile, we record the log-likelihood values for $\sum_{i=1}^{n} l_{i,k_{1,r},j}(v_i, u_{i,k_{1,r}}, u_{i,j})$, as well as the pseudo-samples

$$u_{i,V|k_1^*} = C_{V|k_1^*}(v_i|u_{i,k_1^*})$$

and

$$u_{i,j|k_1^*} = C_{j|k_1^*}(u_{i,j}|u_{i,k_1^*})$$

for $i = 1, \ldots, n$. The pseudo samples will be reused when calculating the conditional log-likelihood for the following trees.

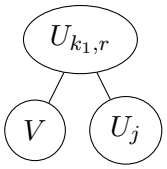


Figure 4.1: *Finding a root node $U_{k_1^*}$ for the first C-vine tree by calculating $\sum_{i=1}^{n} l_{i,k_{1,r},j}(v_i, u_{i,k_{1,r}}, u_{i,j})$ based on $c(v_i|u_{i,k_{1,r}}, u_{i,j})$, where $r = 1, \ldots, K$, $j \in \{1, \ldots, p\}\backslash\{k_{1,r}\}$, and $i = 1, \ldots, n$. The node $U_{k_1,r}$ is a candidate root node, whereas nodes corresponding to the response V and an additional variable U_j are leafs.*

Additional variable reduction in higher dimensions

The search procedure described above requires, for K variables as candidates entering the root node of the first C-vine tree, calculating $j \in \{1, \ldots, p\}\backslash\{k_{1,r}\}$, leading to a total calculation of $(p-1)\times K$ log-likelihoods. In cases where p is large, this search procedure would cause a heavy computational burden. Hence, we suggest two additional variable reduction methods based on partial correlation $\rho_{Vj;k_{1,r}}$, for all $j \in \{1, \ldots, p\}\backslash\{k_{1,r}\}$. The basic idea is similar to the pre-selection in Section 4.4.1. The calculation of the partial correlation $\rho_{Vj;k_{1,r}}$ is based on (4.12) and follows

- Calculate the unconditional correlation ρ_{Vj} using data V and U_j.
- Calculate unconditional correlations $\rho_{V,k_{1,r}}$ using data V and $U_{k_{1,r}}$, and $\rho_{j,k_{1,r}}$ using U_j and $U_{k_{1,r}}$.
- Calculate the partial correlation using (4.12) by

$$\rho_{Vj;k_{1,r}} = \frac{\rho_{Vj} - \rho_{V,k_{1,r}} \cdot \rho_{j,k_{1,r}}}{\sqrt{1-\rho^2_{V,k_{1,r}}} \cdot \sqrt{1-\rho^2_{j,k_{1,r}}}}.$$

For $U_{k_{1,r}}$, we order $U_j, j \in \{1, \ldots, p\}\backslash\{k_{1,r}\}$ according to the partial Spearman correlation. Reducing the number of choices for the variable U_j can be based on either only partial correlation, or a combination of partial correlation and random selection. The selection method and the size of

reduction can be decided by the users.

Second C-vine tree

Similar to the procedure of constructing the first C-vine tree, we first select K candidate variables denoted as $k_{2,1}, \ldots, k_{2,K}$ with the largest absolute value of the partial correlations $\rho_{j,v;k_1^*}$, $j \in \{1, \ldots, p\} \backslash \{k_1^*\}$; the partial correlation is described in Section 4.4.1. For a fixed variable $k_{2,r}$, we calculate the conditional log-likelihood

$$\begin{aligned}\sum_{i=1}^{n} l_{i,k_1^*,k_{2,r},j}(v_i, u_{i,k_1^*}, u_{i,k_{2,r}}, u_{i,j}) = \sum_{i=1}^{n} \Big\{ & \log c_{V,k_1^*}(v_i, u_{i,k_1^*}) \\ & + \log c_{V,j;k_1^*}\big(C_{V|k_1^*}(v_i|u_{i,k_1^*}), C_{j|k_1^*}(u_{i,j}|u_{i,k_1^*})\big)\Big\} \\ + \sum_{i=1}^{n} \Big\{ & \log c_{V,j;k_1^*,k_{2,r}}\big(C_{V|k_1^*,k_{2,r}}(v_i|u_{k_1^*}, u_{i,k_{2,r}}), C_{j|k_1^*,k_{2,r}}(u_{i,j}|u_{i,k_1^*}, u_{i,k_{2,r}})\big)\Big\}\end{aligned}$$

for $j \in \{1, \ldots, p\} \backslash \{k_1^*, k_{2,r}\}, r = 1, \ldots, K$. The first two terms on the right-hand side of the expression of $\sum_{i=1}^{n} l_{i,k_1^*,k_{2,r},j}(v_i, u_{i,k_1^*}, u_{i,k_{2,r}}, u_{i,j})$ above reuse the recorded values calculated while constructing the first tree using (4.13).

Notice that the conditional c.d.f.

$$C_{V|k_1^*,k_{2,r}}(v_i|u_{k_1^*}, u_{i,k_{2,r}}) = \frac{\partial}{\partial w} C_{V,u_{k_{2,r}};u_{k_1^*}}(u_{i,V|k_1^*}, w)|_{w=u_{i,k_{2,r}|k_1^*}}$$

where the pseudo samples $u_{i,V|k_1^*}$ and $u_{i,k_{2,r}|k_1^*}$ were recorded while calculating the first C-vine tree. Estimating the conditional distribution $C_{V|k_1^*,k_{2,r}}$ and evaluating at the ith observation can be done by combining the estimators defined in (4.8) and (4.6).

The second selected variable corresponds to

$$\arg \max_{r \in \{1,\ldots,K\}} \max_{k_{2,r}} \max_{j \in \{1,\ldots,p\} \backslash \{k_1^*, k_{2,r}\}} \sum_{i=1}^{n} l_{i,k_1^*,k_{2,r},j}(v_i, u_{i,k_1^*}, u_{i,k_{2,r}}, u_{i,j})$$

The index of the selected variable is denoted as k_2^*.

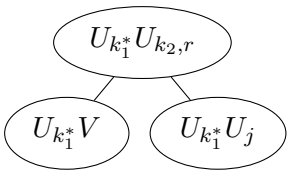


Figure 4.2: *Finding a root node for the second C-vine tree by calculating* $\sum_{i=1}^{n} l_{i,k_1^*,k_{2,r},j}(v_i, u_{i,k_1^*}, u_{i,k_{2,r}}, u_{i,j})$ *based on* $c(v_i|u_{i,k_1^*}, u_{i,k_{2,r}}, u_{i,j})$ *where* $r = 1, \ldots, K$, $j \in \{1, \ldots, p\}\backslash\{k_1^*, k_{2,r}\}$, *and* $i = 1, \ldots, n$. *The node* $U_{k_1*}U_{k_2,r}$ *is a candidate root node where* $U_{k_2,r}$ *is the variable to be selected. The node* $U_{k_1^*}V$ *with the response* V *is a leaf.*

Higher order C-vine tree

The selection of the next ordered variable in the higher order trees is similar to that in Section 4.4.2. The expression of the log-likelihood is calculated gradually based on the conditional copula density in (4.4). More specifically, the first selected variable is based on the conditional copula density $c(v|u_{k_1}, u_j)$, the second selected variable is based on the conditional copula density $c(v|u_{k_1^*}, u_{k_2}, u_j)$, the tth variable selected is based on $c(v|u_{k_1^*}, \ldots, u_{k_{t-1}^*}, u_{k_t}, u_j)$; in each selection step, the selected variable maximizes the corresponding conditional log-likelihood derived from the conditional copula density. Computationally, the additional term of the log-likelihood selecting the tth variable, compared to that selecting the $(t-1)$th variable, is calculated from the conditional c.d.f (the h-function) from the previous selection step.

In the tth step, the candidate tree is as follows

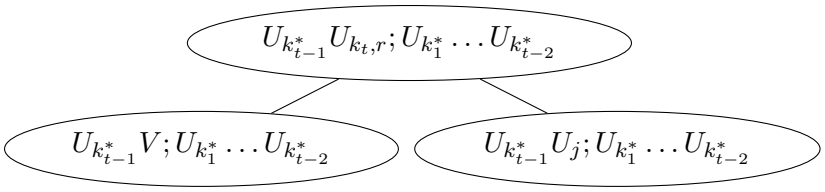


Figure 4.3: *Finding a root node for the tth C-vine tree by calculating* $\sum_{i=1}^{n} l_{i,k_1^*,\dots,k_{t-1}^*,k_{t,r},j}(v_i, u_{i,k_1^*},,\dots,u_{i,k_{t-1}^*},u_{i,k_{t,r}},u_{i,j})$ *based on* $c(v_i|u_{i,k_1^*},\dots,u_{i,k_{t-1}^*},u_{i,k_{t,r}},u_{i,j})$ *where* $r = 1,\dots,K$, $j \in \{1,\dots,p\}\backslash\{k_1^*,\dots,k_{t-1}^*,k_{t,r}\}$, *and* $i = 1,\dots,n$. *The node* $U_{k^*_{t-1}}U_{k_t,r};U_{k^*_1}\dots U_{k^*_{t-2}}$ *is a candidate root node where* $U_{k_t,r}$ *is the variable to be selected. The node with the response* V *is always a leaf.*

4.5 Simulation

We set up the following simulation settings. Each setting is replicated for $R = 100$ times. In each simulation replication, we randomly generate n_{train} samples used for estimation described in Steps 1 – 4 in Section 4.3.5; additionally, another $n_{\text{eval}} = \frac{1}{2}n_{\text{train}}$ samples for setting (a) – (f) and $n_{\text{eval}} = n_{\text{train}}$ for setting (g), (h) are generated for predicting the conditional quantile as described in Step 5. Settings (a) – (f) are designed to test the quantile prediction accuracy of nonparametric C-vine quantile regression in cases where $p \leq n$; hence, we set $n_{\text{train}} = 1000$ or 300. Settings (g) and (h) test the quantile prediction accuracy in cases where $p > n$; hence, we set $n_{\text{train}} = 100$. For settings with homoscedastic noises, i.e., $Y_i = g(X_{i\cdot}) + \varepsilon_i, i = 1,\dots,n$ with $g(\cdot)$ being some mapping and the standard deviation of ε_i's denoted by σ (see Setting (a), (g), (h)), we also report signal-to-noise ratios (SNR) defined as

$$\text{SNR} = \frac{\sum_{i=1}^{n} g(X_{i\cdot})/n}{\sigma}.$$

(a) Simulation setting M5 from Kraus and Czado (2017) where $Y = \sqrt{|2X_{\cdot 1} - X_{\cdot 2} + 0.5|} + (-0.5X_{\cdot 3}+1)(0.1X_{\cdot 4}^3) + \sigma\varepsilon$ with $\varepsilon \sim N(0,1), \sigma \in \{0.1, 1\}$, $(X_{\cdot 1}, X_{\cdot 2}, X_{\cdot 3}, X_{\cdot 4}) \sim N_4(0, \Sigma)$, and the (i,j)th component of the covariance matrix $(\Sigma)_{i,j} = 0.5^{|i-j|}$. The SNR is 10.33 (1.03)

when $\sigma = 0.1(1)$.

(b) $(Y, X_{\cdot 1}, \ldots, X_{\cdot 5})$ follows a mixture of two 6-dimensional t-copulas with degrees of freedom equal to 3 with mixture probabilities 0.3 and 0.7; association matrices R_1, R_2 as in Table 4.3 and marginal distributions as given in Table 4.4.

$$R_1 = \begin{pmatrix} 1 & 0.6 & 0.5 & 0.6 & 0.7 & 0.1 \\ 0.6 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.6 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.7 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.1 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1 & -0.3 & -0.5 & -0.4 & -0.5 & -0.1 \\ -0.3 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ -0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ -0.4 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ -0.5 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ -0.1 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

Table 4.3: *Association matrices for simulation setting (b)*

Y	$X_{\cdot 1}$	$X_{\cdot 2}$	$X_{\cdot 3}$	$X_{\cdot 4}$	$X_{\cdot 5}$
$N(0,1)$	t_4	$N(1,4)$	t_4	$N(1,4)$	t_4

Table 4.4: *Marginal distributions for simulation setting (b)*

(c) Linear and heteroscedastic (Chang and Joe, 2019): $Y = 5(X_{\cdot 1} + X_{\cdot 2} + X_{\cdot 3} + X_{\cdot 4}) + 10(U_1 + U_2 + U_3 + U_4)\varepsilon$ where $(X_{\cdot 1}, X_{\cdot 2}, X_{\cdot 3}, X_{\cdot 4}) \sim N(0, \Sigma)$, $\Sigma_{i,j} = 0.5^{I\{i \neq j\}}$, $\varepsilon \sim N_4(0, 0.5)$, and $U_j, j = 1, \ldots, 4$ are obtained from $X_{\cdot j}$, $j = 1, \ldots, 4$ by the probability integral transform.

(d) Nonlinear and heteroscedastic (Chang and Joe, 2019):

$$Y = U_1 U_2 e^{1.8 U_3 U_4} + 0.5(U_1 + U_2 + U_3 + U_4)\varepsilon$$

where $U_j, j = 1, \ldots, 4$ are probability integral transformed from $N_4(0, \Sigma)$, $\Sigma_{i,j} = 0.5^{I\{i \neq j\}}$, and $\varepsilon \sim N(0, 0.5)$.

(e) Sampling from an R-vine copula (Czado, 2019): $(Y, X_{\cdot 1}, \ldots, X_{\cdot 4})$ follows a R-vine copula structure with R-vine matrix

$$\begin{pmatrix} 1 & 1 & 1 & 4 & 4 \\ & 5 & 5 & 1 & 1 \\ & & 4 & 5 & 3 \\ & & & 3 & 5 \\ & & & & 2 \end{pmatrix}.$$

Details on R-vine matrix representation are in Section 4.10. The graphical representation of this R-vine is in Figure 4.5 in Section 4.12. The corresponding parameter matrix of pair-copulas of the R-vine is introduced in Section 4.11. The pair-copula families correspond to

$$\begin{pmatrix} 0 & 3(4.8) & 1(0.5) & 1(0.9) & 4(3.9) \\ 0 & 0 & 4(1.9) & 24(2.6) & 24(6.5) \\ 0 & 0 & 0 & 23(5.1) & 3(0.9) \\ 0 & 0 & 0 & 0 & 1(0.2) \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where 1 corresponds to a Gaussian copula, 3 a Clayton, 4 a Gumbel, 23 a rotated 90 degree Clayton, and 24 a rotated 90 degree Gumbel; numbers in the parentheses represent the corresponding parameters of the copulas. For example, the $(1, 2)$th element 3(4.8) represents that the unconditional pair copula $C_{1,5}$ is a Gaussian copula with parameter 4.8.

(f) Similar to (e), while $(Y, X_{\cdot 1}, \ldots, X_{\cdot 4})$ follows a D-vine copula struc-

ture with R-vine matrix

$$\begin{pmatrix} 5 & 5 & 4 & 3 & 2 \\ & 4 & 5 & 4 & 3 \\ & & 3 & 5 & 4 \\ & & & 2 & 5 \\ & & & & 1 \end{pmatrix}.$$

The graphical representation of this D-vine is in Figure 4.4 in Section 4.12. The pair-copula families are the same as (e).

(g) Similar to setting (a).

$$\begin{aligned} Y \quad &= \sqrt{|2X_{\cdot 1} - X_{\cdot 2} + 0.5|} + (-0.5X_{\cdot 3} + 1)(0.1X_{\cdot 4}^3) \\ &\quad + (X_{\cdot 5}, \ldots, X_{\cdot 110})(0, \ldots, 0)^\top + \sigma\varepsilon, \end{aligned}$$

where $(X_{\cdot 1}, \ldots, X_{\cdot 110}) \sim N_{110}(0, \Sigma)$ with the (i, j)th component of the covariance matrix $(\Sigma)_{i,j} = 0.5^{|i-j|}$, $\varepsilon \sim N(0, 1)$, and $\sigma \in \{0.1, 1\}$. Same as setting (a), the SNR is 10.33 (1.03) when $\sigma = 0.1(1)$.

(h) Similar to (g); the true model is $Y = (|X_{\cdot 1}|^{1/2}, \ldots, |X_{\cdot 110}|^{1/2})\beta + \varepsilon$ where the first 5 entries of β are equal to 1, the remaining 105 entries of β are equal to 0, and $\varepsilon \sim N(0, 0.1)$. The SNR is 41.08 (4.11) when $\sigma_\cdot = 0.1(1)$.

Since the true regression quantiles are difficult to obtain in most settings, we consider the averaged check loss (Kraus and Czado, 2017; Komunjer, 2013) and the interval score (Chang and Joe, 2019; Gneiting and Raftery, 2007), instead of the out-of-sample mean averaged square error in Kraus and Czado (2017), to evaluate the performance of the estimation methods. The averaged check loss is defined as

$$\widehat{\mathrm{CL}}_\tau = \frac{1}{R} \sum_{r=1}^{R} \left\{ \frac{1}{n_{\mathrm{eval}}} \sum_{i=1}^{n_{\mathrm{eval}}} \left\{ \rho_\tau\big(Y_{r,i}^{\mathrm{eval}} - \hat{q}_\tau(X_{\cdot r,i}^{\mathrm{eval}})\big) \right\} \right\}. \tag{4.14}$$

A smaller value of the averaged check loss suggests that the corresponding

estimator used for prediction gives a lower value of the loss, this usually suggests that estimator is closer to the global minimizer of the loss function. The interval score for the $(1-\tau)\times 100\%$ prediction interval is defined as

$$\begin{aligned}\widehat{\mathrm{IS}}_\tau = \frac{1}{R}\sum_{r=1}^{R}\Bigg\{\frac{1}{n_{\text{eval}}}\sum_{i=1}^{n_{\text{eval}}}\Big\{&\big(\hat{q}_{\tau/2}(X_{r,i\cdot}^{\text{eval}}) - \hat{q}_{1-\tau/2}(X_{r,i\cdot}^{\text{eval}})\big)\\ &+\frac{2}{\tau}\big(\hat{q}_{1-\tau/2}(X_{r,i\cdot}^{\text{eval}}) - Y_{r,i}^{\text{eval}}\big)I\{Y_{r,i}^{\text{eval}} \le \hat{q}_{1-\tau/2}(X_{r,i\cdot}^{\text{eval}})\}\\ &+\frac{2}{\tau}\big(Y_{r,i\cdot}^{\text{eval}} - \hat{q}_{\tau/2}(X_{r,i\cdot}^{\text{eval}})\big)I\{Y_{r,i}^{\text{eval}} > \hat{q}_{\tau/2}(X_{r,i\cdot}^{\text{eval}})\}\Big\}\Bigg\},\end{aligned} \tag{4.15}$$

and smaller interval scores are better. This expression of interval scores consists of two elements; i.e., the length of the interval corresponding to the first term in (4.15), and a score of if the observations miss the interval corresponding to the last two terms in (4.15). A combination of narrower intervals and lower score on the observations missing intervals leads to a smaller interval score suggesting a more accurate prediction.

For settings (a) – (f), we do not perform pre-selection as described in Section 4.4.1, i.e., we calculate the conditional log-likelihoods of all possible combinations in each searching step corresponding to each level of C-vine trees in Step 2 in Section 4.3.5. On the contrary, due to the computational burden in the search steps for settings (g) and (h), we set the number of candidate variables indexed by k to be $K = 5$, and reduce the choices of the variable indexed by j to 20% of all possible choices where 10% are variables having the highest partial correlation with the response variable and the remaining 10% are chosen randomly from the rest of the variables.

The performance of C-vine quantile regression is compared with D-vine quantile regression as in Kraus and Czado (2017). The performance of the two methods, evaluated by the averaged check loss at the 5%, 50%, 95% quantile levels and the interval score for the 95% prediction interval, are recorded in Table 4.5. All densities are estimated nonparametrically which should give an honest comparison.

Table 4.5 shows that the C-vine copulas based quantile prediction com-

bining the variable ordering in Step 2 in Section 4.3.5 outperforms the D-vine quantile prediction from Kraus and Czado (2017). One exception is setting (b) where the data are generated from a mixture of two 6-dimensional t_3 copulas. Noteworthy, in setting (f) where the true data are generated from a D-vine, our proposed C-vine quantile regression still outperforms its D-vine quantile counterpart on the averaged check loss at quantile levels 5% and 50%.

In high dimensional settings, the D-vine copulas based quantile regressions only outperforms our proposed method in setting (h) on the averaged check loss at quantile level 5%. Another general observation is that our proposed C-vine quantile regression with the "one-step-ahead" variable ordering technique could largely improve the performance of the D-vine quantile regression, especially in cases where the standard deviation of the regression errors is $\sigma = 0.1$. Intuitively, the observations are less disturbed by errors; hence, better performance indicates a more accurate capture of the correlation structure of the predictive variables and the response variable.

4.6 Real data examples

In this section, we test the proposed method on two real datasets, i.e., the abalone dataset (Nash et al., 1994) from Dua and Graff (2019) corresponding to $n \geq p$, and the riboflavin dataset from Bühlmann and van de Geer (2011) corresponding to $n < p$. For both datasets, performances of the proposed C-vine based quantile regression will be evaluated by the averaged check loss defined in (4.14) at 5%, 50% and 95% quantile levels, as well as the 95% prediction interval score defined in (4.15), by randomly splitting the dataset into training and evaluation sets 100 times. Performance of the proposed method will be compared with the D-vine based quantile regression in Kraus and Czado (2017).

		$\widehat{\text{IS}}_{0.05}$		$\widehat{\text{CL}}_{0.05}$		$\widehat{\text{CL}}_{0.5}$		$\widehat{\text{CL}}_{0.95}$	
Setting	n_{train}	C-vine	D-vine	C-vine	D-vine	C-vine	D-vine	C-vine	D-vine
(a)	300	39.034	52.171	0.416	0.605	0.107	0.162	0.384	0.500
$\sigma = 0.1$	1000	39.662	54.822	0.419	0.649	0.093	0.157	0.365	0.495
(a)	300	146.350	150.409	1.553	1.605	0.447	0.457	1.549	1.582
$\sigma = 1$	1000	154.538	159.004	1.617	1.673	0.421	0.435	1.602	1.640
(b)	300	119.063	115.588	1.297	1.259	0.415	0.415	1.300	1.274
	1000	123.665	123.257	1.348	1.345	0.397	0.399	1.346	1.342
(c)	300	1429.667	1449.607	14.838	15.182	4.417	4.517	15.058	15.185
	1000	1520.353	1529.319	15.479	15.625	4.195	4.236	15.752	15.814
(d)	300	85.467	87.901	0.866	0.878	0.247	0.252	0.955	0.996
	1000	90.078	91.578	0.936	0.943	0.226	0.231	0.958	0.981
(e)	300	9.208	9.564	0.086	0.102	0.023	0.028	0.105	0.097
	1000	9.410	9.382	0.086	0.098	0.022	0.024	0.107	0.094
(f)	300	4.952	4.919	0.048	0.049	0.012	0.013	0.053	0.053
	1000	4.679	4.706	0.044	0.044	0.011	0.011	0.050	0.051
(g) $\sigma = 0.1$	100	19.631	36.568	0.238	0.520	0.245	0.457	0.253	0.394
$\sigma = 1$	100	53.638	57.115	0.689	0.772	0.670	0.714	0.652	0.656
(h) $\sigma = 0.1$	100	39.162	55.183	0.474	0.252	0.490	0.690	0.505	1.127
$\sigma = 1$	100	61.104	68.819	0.752	0.525	0.764	0.860	0.776	1.196

Table 4.5: *Out-of-sample prediction $\widehat{IS}_{0.05}$, $\widehat{CL}_{0.05}$, $\widehat{CL}_{0.5}$, $\widehat{CL}_{0.95}$ by C- and D-vine quantile regression in setting (a) – (h). Lower values, indicating better performance among the two, are highlighted in yellow.*

4.6.1 Abalone dataset

The abalone (a type of shellfish) dataset from UCI machine learning repository (Dua and Graff, 2019) was used in Chang and Joe (2019) as an example covering both regression and classification. The dataset has in total 4177 samples including female, male, and infant abalone. Our objective is to obtain quantile predictions of the age of 1528 male abalone, determined by the number of rings counted, using the remaining 7 continuous physical measurement variables. We randomly split the dataset into a training set with 1228 samples and an evaluation set with 300 samples; the random splitting is repeated 100 times. Performance of the proposed C-vine based quantile regression, compared with D-vine based quantile regression (Kraus and Czado, 2017), is evaluated by several performance measurements and is reported in Table 4.6. We see that the proposed C-vine quantile regression and the D-vine quantile regression have a close performance. However, the proposed C-vine quantile regression slightly outperforms the D-vine quantile regression on the averaged check loss at the 5% and 50% quantile levels.

Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
C-vine	308.016	2.560	0.780	3.674
D-vine	304.677	2.637	0.813	3.571

Table 4.6: *Table recording performance of C- and D-vine quantile regression for the Abalone dataset. Out-of-sample prediction* $\widehat{IS}_{0.05}$, $\widehat{CL}_{0.05}$, $\widehat{CL}_{0.5}$, $\widehat{CL}_{0.95}$ *calculated based on* $R = 100$ *times randomly split the dataset into training set with 1228 samples and evaluation set with 300 samples. Lower values indicating better performance, among the two are highlighted in yellow.*

4.6.2 Riboflavin dataset

The riboflavin dataset, available in the R package `hdi`, aims at quantile predictions of log-transformed production rate of Bacillus subtilis using

log-transformed expression levels of 4088 genes. To reduce the computational burden, we perform a pre-selection of the top 100 genes with the highest variance (Bühlmann and van de Geer, 2011), resulting in a subset with $p = 100$ log-transformed gene expressions and $n = 71$ samples. A random splitting of the subset into a training set with 61 samples and an evaluation set with 10 samples, is repeated 100 times. Additionally, to further reduce the computational burden but to keep the accuracy of the fitted C-vine copula model, we perform the pre-selection described in Section 4.4.1 by setting the number of candidate variables indexed by k to be $K = 10$. Further, we reduce the choice of the variables indexed by j to 25% of all possible choices, where 15% of them are the variables that have the highest partial correlations with the log-transformed Bacillus subtilis production rate. The other 10% of the variables were randomly chosen from the remaining variables. The performance of the C- and D-vine based quantile regressions are reported in Table 4.7, where we see that the proposed C-vine copula based quantile regression consistently outperforms the D-vine based quantile regression in Kraus and Czado (2017) to a large extent, i.e., scores of D-vine are approximately 1.5 times of those of C-vines for all four scores.

Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
C-vine	29.849	0.390	0.373	0.356
D-vine	43.948	0.507	0.549	0.592

Table 4.7: *Table recording performance of C- and D-vine quantile regression for the Riboflavin dataset. Out-of-sample prediction* $\widehat{IS}_{0.05}$, $\widehat{CL}_{0.05}$, $\widehat{CL}_{0.5}$, $\widehat{CL}_{0.95}$ *calculated based on* $R = 100$ *times randomly split the dataset into training set with 61 samples and evaluation set with 10 samples. Lower values indicating better performance, among the two are highlighted in yellow.*

4.7 Flexible conditional mean estimator

In the previous sections we derived a conditional quantile estimator of $F^{-1}_{Y|1,\dots,p}(\tau|x_1,\dots,x_p)$ based on C-vine copula modelling. There is an immediate connection between a conditional quantile and a conditional mean. Indeed, by taking $\tau = F_{Y|1,\dots,p}(y|x_1,\dots,x_p)$ and by the change of variable theorem, it holds that

$$\int_0^1 F^{-1}_{Y|1,\dots,p}(\tau|x_1,\dots,x_p)d\tau = \int_{-\infty}^{\infty} y dF_{Y|1,\dots,p}(y|x_{\{1,\dots,p\}}). \tag{4.16}$$

The right-hand side of (4.16) corresponds to the conditional mean of the response variable Y, and a discrete approximation of the left-hand side offers a flexible estimator of the conditional mean of Y in the following way,

$$\hat{E}_{Y|1,\dots,p}\Big[Y|x_1,\dots,x_p\Big] = \sum_{l=1}^{L} \frac{1}{L} \hat{F}^{-1}_{Y|1,\dots,p}(\alpha_l|x_1,\dots,x_p)$$

After the conditional quantile estimators for the different levels $\alpha_1,\dots,\alpha_L$ are constructed, the corresponding conditional mean estimator follows directly. Here, we evaluate the performance of such a conditional mean estimator by an out-of-sample mean prediction via the mean averaged squared error

$$\widehat{\text{MASE}}_{\text{mean}} = \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{n_{\text{eval}}}\sum_{i=1}^{n_{\text{eval}}}\left\{\sum_{l=1}^{L}\frac{1}{L}\hat{q}^{(r)}_{l/11}(X^{\text{eval}}_{r,i\cdot}) - Y^{\text{eval}}_{r,i}\right\}^2\right\}, \tag{4.17}$$

where the conditional mean $\hat{E}_{Y|1,\dots,p}$ is approximated by averaging conditional quantile estimators at $L = 10$ equally spaced quantile levels, i.e., $1/11,\dots,10/11$. This estimator can be seen as a model averaging estimator using equal weights $1/10$.

Table 4.8 reports the comparison of the performance of the conditional mean estimator based on C-vine copulas with the (1) corresponding estimator based on D-vine copulas; (2) the ordinary least squares (OLS) estimator in settings (a) – (f), or the Lasso estimator in settings (g), (h).

Tuning parameters of the Lasso estimators are selected by 10-fold cross-validation, i.e., the one minimizes mean cross-validation errors on a grid; (3) the nonparametric regression implemented in R package `np` (Hayfield and Racine, 2008) with the least squares cross-validated bandwidths. Table 4.8 shows that modelling a conditional mean using the proposed C-vine copula method has overall the best performance when there exists a strong non-linearity between the predictive variables and the response variable. The proposed C-vine copulas outperform the D-vine copula method in all settings. However, using the OLS or the Lasso to modelling the conditional mean in Setting (b), (c), and (g) with $\sigma = 1$, lead to the lowest prediction error, especially in Setting (c) which is a linear model with heteroscedastic errors for which the least squares estimators should perform best.

4.8 Discussion

We proposed a C-vine copula based nonparametric quantile regression model which uses a novel root node order selection algorithm, namely a "one-step-ahead" selection. This "one-step-ahead" selection approach maximizes the conditional log-likelihood of one level of C-vine tree, taking the subsequent level of C-vine tree into account. The pair-copula densities of the C-vine copulas are estimated nonparametrically, which maximizes the flexibility of the model. The proposed model does not impose additional model assumptions; hence it is especially suitable for dealing with an unknown correlation structure of the response variable and the predictive variables or with heteroscedastic errors. The simulation and real data results for both low- and high dimensional settings show a superior finite sample performance of the proposed C-vine copulas models, compared to the D-vine copulas models proposed in Kraus and Czado (2017).

There are still questions to be answered. To ensure the uniqueness of the copulas, we assume in this chapter that all marginal distributions are continuous. However, Chang and Joe (2019) proposed to an R-vine based conditional distribution modeling allowing discrete marginal distributions.

Setting	n_{train}	C-vine	D-vine	OLS	NP
(a)	300	0.161	0.259	1.444	0.887
$\sigma = 0.1$	1000	0.160	0.259	1.425	1.854
(a)	300	1.288	1.368	2.393	1.103
$\sigma = 1$	1000	1.142	1.233	2.421	1.823
(b)	300	1.180	1.208	0.986	1.236
	1000	1.069	1.095	0.975	1.126
(c)	300	142.222	155.671	123.430	589.393
	1000	129.919	133.794	119.422	581.218
(d)	300	0.508	0.536	1.099	2.084
	1000	0.429	0.450	1.102	2.065
(e)	300	0.004	0.010	0.013	0.178
	1000	0.004	0.006	0.013	0.176
(f)	300	0.001	0.003	0.004	0.159
	1000	0.001	0.001	0.004	0.160
Setting	n_{train}	C-vine	D-vine	Lasso	NP
(g) $\sigma = 0.1$	100	0.474	1.532	1.505	0.962
$\sigma = 1$	100	2.904	3.392	2.752	2.901
(h) $\sigma = 0.1$	100	1.520	2.787	17.656	1.631
$\sigma = 1$	100	3.677	4.627	18.662	3.578

Table 4.8: *Prediction* $\widehat{MASE}_{\text{mean}}$ *by C- and D-vine copulas, OLS in settings (a) – (f) and the Lasso in settings (g), (h), and nonparametric regression. Lowest values, indicating stronger out-of-sample prediction ability, are highlighted in yellow.*

Schallhorn et al. (2017); Nagler and Kraus (2018) further incorporated discrete marginal distributions in the D-vine based quantile regression. Incorporating a mix of discrete and continuous distributions in the proposed C-vine based quantile regression would lead to a much more powerful algorithm.

Another question is related to the pre-selection for reducing the computational complexity in high dimensions, i.e., to reduce the number of pair copulas to be calculated in each searching step. This pre-selection is based on the partial Spearman's ρ of the predictive variables with the

response variable. Although Haff et al. (2010) showed the link between partial correlation with log-likelihood, there is no guarantee suggesting that the selected candidate predictive variables are the maximizers of the conditional log-likelihoods in each selection step. Other pre-selection approaches that are more directly relevant to the conditional log-likelihoods should be investigated further.

Also, the current "one-step-ahead" algorithm follows a sequential maximization procedure, which maximizes a truncated conditional log-likelihood in each step. This procedure does not ensure that the resulting root node order is the maximizer of the complete conditional log-likelihood. A backward selection procedure could be incorporated. Another issue regarding root node order selection is that we do not incorporate stopping criteria at the moment, i.e., we include all candidate predictive variables in the final model, which could be time-consuming in ultra-high dimensions.

Further extensions of the algorithm could include more than one step ahead. The cost of an increased computational complexity is to be compared to a potential better selection. Theoretical results regarding the construction of the C-vine with one or more steps ahead in the order selection procedure could shed light to this issue. Such theory is beyond the scope of this work.

Appendix

4.9 Proof of Proposition 4.1

Proof. We first show 1 in Proposition 4.1 on the uniform strong consistency of the inverse of conditional quantile estimator.

By (4.2), the estimator

$$\hat{F}^{-1}_{Y|1,\dots,p}(\tau|x_{\{1,\dots,p\}}) = \hat{F}^{-1}_{Y}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big),$$

where $\hat{u}_j = \hat{F}_j(x_j), j = 1,\dots,p$ denote variables on the u-scale. To avoid

heavy notation, n referring to the sample size will be omitted here. Following Wied and Weißbach (2012); Silverman (1978), to show the uniformly strong consistency of $\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big)$, we show

$$\sup_\tau \left|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \to 0 \ a.s.$$

For all $\epsilon \geq 0$,

$$\begin{aligned}
1 &\geq P\bigg(\sup_\tau \left|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \epsilon\bigg) \\
&= P\bigg(\sup_\tau \left|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - \hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right. \\
&\quad \left. + \hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \epsilon\bigg) \\
&\geq P\bigg(\sup_\tau \Big\{\Big|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - \hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\Big| \\
&\quad + \Big|\hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\Big|\Big\} \leq \epsilon\bigg) \\
&= P\bigg(\sup_\tau \left|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - \hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \\
&\quad + \sup_\tau \left|\hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \epsilon\bigg) \\
&\geq P\bigg(\Big(\sup_\tau \left|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - \hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \tfrac{3}{4}\epsilon\Big) \\
&\quad \bigcap \Big(\sup_\tau \left|\hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \tfrac{1}{4}\epsilon\Big)\bigg) \\
&= P\bigg(\sup_\tau \left|\hat{F}_Y^{-1}\Big(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\Big) - \hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \tfrac{3}{4}\epsilon \ \Big| \\
&\quad \sup_\tau \left|\hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \tfrac{1}{4}\epsilon\bigg) \\
&\cdot P\bigg(\sup_\tau \left|\hat{F}_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big) - F_Y^{-1}\Big(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\Big)\right| \leq \tfrac{1}{4}\epsilon\bigg)
\end{aligned} \tag{4.18}$$

Denote the event

$$A = \sup_{\tau} \left| \hat{F}_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right) - F_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right| \leq \frac{1}{4}\epsilon,$$

then $P(A) = 1$ holds by the uniform strong consistency of the estimator of F_Y^{-1}. We now show that the conditional probability in (4.18) is equal to 1. We start by rewriting the conditional probability.

$$\begin{aligned}
&P\Bigg(\sup_{\tau} \left| \hat{F}_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right) - \hat{F}_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right| \leq \frac{3}{4}\epsilon \Bigg| A \Bigg) \\
&\quad = P\Bigg(\sup_{\tau} \left| \hat{F}_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right) - F_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right)\right. \\
&\qquad \left. + F_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right) - \hat{F}_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right. \\
&\qquad \left. + F_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right) - F_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right| \leq \tfrac{3}{4}\epsilon \Bigg| A \Bigg) \\
&\quad \geq P\Bigg(\sup_{\tau} \left| \hat{F}_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right) - F_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right)\right| \\
&\qquad + \sup_{\tau} \left| F_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right) - \hat{F}_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right| \\
&\quad + \sup_{\tau} \left| F_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right) - F_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right| \leq \tfrac{3}{4}\epsilon \Bigg| A \Bigg)
\end{aligned}$$

This conditional probability is equal to 1, since the first and second supremum are less than or equal to $\frac{1}{4}\epsilon$ by conditioning on A and due to the uniform consistency of $\hat{F}_Y^{-1}$. The last supremum is less than or equal to $\frac{1}{4}\epsilon$ by Bartle and Joichi (1961, Thm.2) on almost uniform convergence, applied to the continuous inverse distribution function F_Y^{-1}, and taking the measurable space to be the probability space. First,

$$P\left(\sup_{\tau} \left| \left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right) - \left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right| \leq \frac{1}{4}\epsilon \right) = 1,$$

which can be obtained similarly to (4.18) using the uniform consistency

and continuity of the inverse of the h-functions. Next, (4.18) states

$$P\left(\sup_{\tau}\left|\hat{F}_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right)-\hat{F}_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right|\leq\epsilon\right)=1.$$

We conclude that $\hat{F}_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right)$ is uniformly strong consistent. To prove the weak consistency in 2, by Wied and Weißbach (2012); Silverman (1978), we only need to show

$$P\left(\left|\hat{F}_Y^{-1}\left(\hat{C}^{-1}_{V|1,\dots,p}(\tau|\hat{u}_{\{1,\dots,p\}})\right)-F_Y^{-1}\left(C^{-1}_{V|1,\dots,p}(\tau|u_{\{1,\dots,p\}})\right)\right|\leq\epsilon\right)\to 1$$

Using the same technique in (4.18) and a similar argument for proving 2 of Proposition 4.1 with Theorem 2 on convergence in measure in Bartle and Joichi (1961), the weak consistency can be obtained. □

4.10 Vine matrices representation

As examples shown in Table 4.2 and 4.1, vines can be illustrated by graphical representations using nodes and edges. The graphical representation of vines is comprehensible, but is inconvenient especially when the number of variables involved gets large. An alternative choice is a matrix representation using a single upper triangular matrix to represent a vine structure (Czado, 2019; Kurowicka, 2009; Dißmann, 2010).

An R-vine with p variables $U_j, j = 1,\dots,p$ can be represented by an upper diagonal matrix $\mathbf{M} = (m_{l',l}) \in \mathbb{R}^{p\times p}$. The matrix can be interpreted as follows: variables U_j's are elements of the matrix $\mathbf{M}$. The element $m_{l',l}$ in the matrix, i.e., the element on the l'th row and lth column of the matrix, is connected to the element $m_{l',l'}$, given the variables $m_{l'',l}, l'' = 1,\dots,l'-1$. For the l'th tree, edges are $[m_{l',l}, m_{l,l}|m_{l'',l}, l''=1,\dots,l'-1]$, for $l'+1\leq l\leq p$, e.g., there is an edge between variables on the $m_{l',l}$th and $m_{l,l}$th elements of $\mathbf{M}$, given variables on $m_{l'',l}, l''=1,\dots,l'-1$.

For example, the C-vine with graphical representation in Table 4.9,

can be represented by $\begin{pmatrix} 1 & 1 & 1 & 1 \\ & 2 & 2 & 2 \\ & & 3 & 3 \\ & & & 4 \end{pmatrix}$, where the numbers corresponds to the indices of variables, e.g., 1 represents U_1. Then, the edges in the first C-vine tree takes $l' = 1$, the edges are $[m_{1,l}, m_{l,l}], l = 2, \ldots, 4$, i.e., $[1,2], [1,3], [1,4]$. The edges in the second C-vine tree takes $l' = 2$, the edges are $[m_{2,l}, m_{l,l} \mid m_{l'',l}, l'' = 1], l = 3, 4$, i.e., $[2,3|1], [2,4|1]$. The edges in the third C-vine tree takes $l'' = 3$, the edge is $[m_{3,l}, m_{l,l} \mid m_{l'',l}, l'' = 1, 2], l = 4$, i.e., $[3,4|1,2]$.

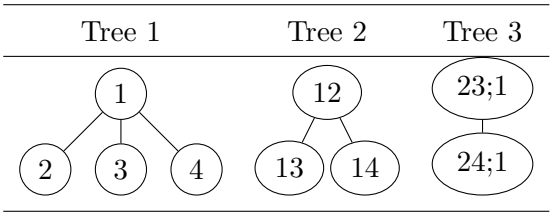


Table 4.9: *The graphical representation of a C-vine ($d = 4$). To simplify the notation in the graphical representation, variable U_j is denoted by its subscript j in the graph; e.g., "1" in Tree 1 represents variable U_1, "12" in Tree 2 represents the bivariate copula of U_1 and U_2, "23;1" in Tree 3 represents the bivariate copula of U_2 and U_3 given U_1.*

4.11 Parameter matrices corresponding to vine matrices

A parameter matrix $\Theta = (\theta_{l',l})$ corresponding to the parameters of pair copulas in each level of trees in an R-vine, is usually given for an R-vine matrix for simulating data. A procedure of simulating vine copulas from an R-vine matrix, see Section 4.10, is described in detail in Czado (2019, Section 6.5). A parameter matrix $\Theta = (\theta_{l',l})$ can be interpreted as follows: we first reorder the variables, such that the diagonal of the R-vine matrix $\mathbf{M}$, corresponding to the indices of the variables, follows

an ascending order, i.e., the index of the m_{ll}th element is l. Then, the (l', l)th element $\theta_{l',l}$ of Θ corresponds to the parameter of the pair copula $C_{m_{l',l},m_{l,l};m_{1,l},\ldots,m_{l',l}}, l' < l \leq p$.

4.12 Visualization of the R- and D-vine in Section 4.5

See Figure 4.5, 4.4 for visualizing the R- and D-vine matrix in setting (e), (f), respectively.

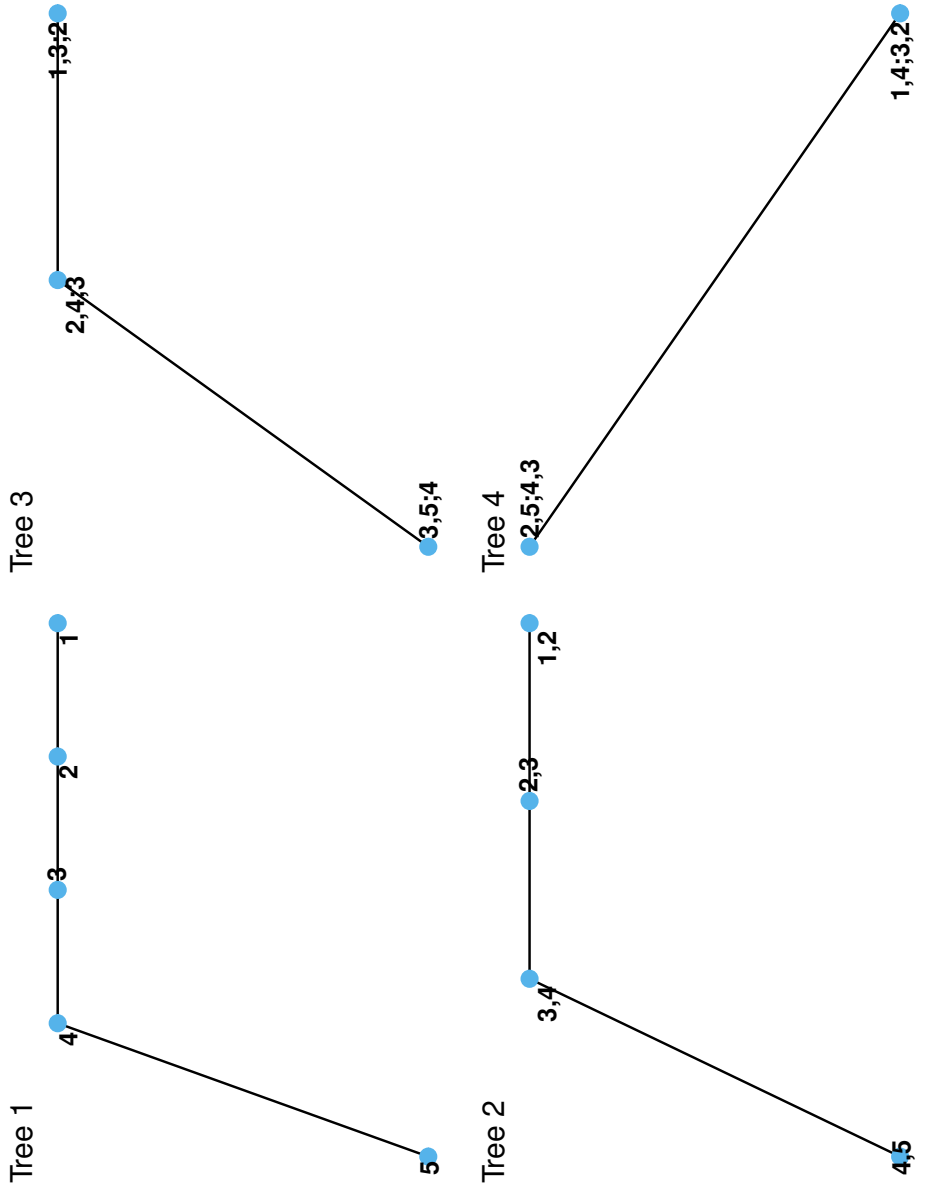


Figure 4.4: *Visualization of the D-vine matrix in setting (e) in Section 4.5. This D-vine has 4 levels of trees. The numbers correspond to indices of variables; the edges correspond to pair copulas and are conditional on the common variable (number).*

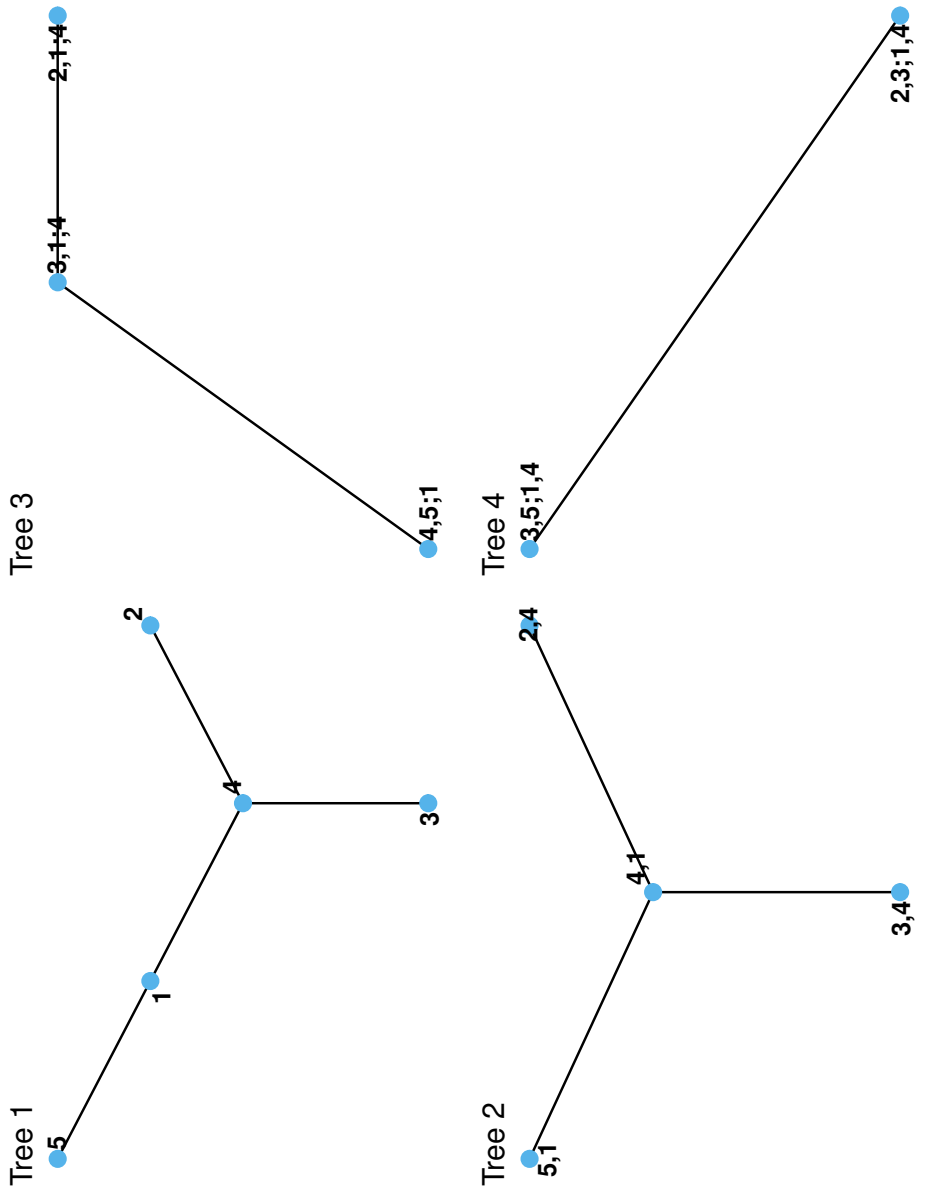


Figure 4.5: *Visualization of the R-vine matrix in setting (f) in Section 4.5. This R-vine has 4 levels of trees. The numbers correspond to indices of variables; the edges correspond to pair copulas and are conditional on the common variable (number). For example, the edge involving "5, 1" and "4, 1" refers to a pair copula* $C_{45;1}$ *between*

Outlook

This thesis contributes to quantile regression, as well as its model-averaged and composite versions, in high dimensions.

The main contribution of Chapter 1 is the asymptotic normality of the model-averaged quantile prediction with fixed weights for high dimensional linear models with fixed dimensions, under the perfect selection assumption assuming variables with true nonzero components (a.k.a., the active set) are selected. An oracle-type weight expression is derived to achieve the lower bound of the asymptotic variance of the active set. However, a major concern is that the perfect selection assumption ignores selection uncertainty. To relax this assumption, Chapter 2 considered the robust approximate message passing (RAMP) algorithm, which provides the expression of the asymptotic mean squared error (AMSE) of the l_1-regularized estimators with an arbitrary convex non-differentiable loss function. The main contribution of Chapter 2 is an AMSE-type weight choice achieving the lower bound of the AMSE of the regularized model-averaged and composite estimator. However, theories of the above two Chapters are derived based on assuming fixed weights, ignoring the correlation of the weights and estimators. A possible extension would be a data-driven weight considering the joint distribution of the weight and estimator. Additionally, both chapters do not consider the selection uncertainty of the tuning parameter of regularization, which is worth further investigation. Another limitation is that we only consider a linear model with homoscedastic errors in the first two chapters. An adaptation of the proposed methodologies to models

with relaxed assumptions (e.g., heteroscedastic or auto-correlated errors) or with a more general form (e.g., generalized linear models, moment condition models in Belloni et al. (2018) for Chapter 1) would be relevant and interesting too.

Chapter 3 further makes use of the asymptotic normality of a sequence $\widetilde{\beta}_{(t)}$ with iteration indexed by t from the RAMP algorithm in Chapter 2. We propose a computational approach constructing componentwise confidence intervals, two-sided individual, and multiple hypothesis tests using $\widetilde{\beta}_{(t)}$ at convergence. Since Chapter 3 only discussed a single l_1-regularized estimator, the confidence intervals and hypothesis testing for regularized model-averaged and composite estimators is a logical extension. A controversial condition of the RAMP algorithm assumes that the components of the design matrix are independent and identically Gaussian distributed. Relaxing this condition by allowing more general design matrices is technically challenging but is of research interest. Also, multiple testing is constructed by adjusting p-values using a conservative Bonferroni-type correction, which can be improved further by other corrections or constructing other test statistics. Another critical development in classical regression is the information criteria, which are underdeveloped in high dimensions due to issues such as a debate about how to properly define the complexity of the models, including a correct specification of the degrees of freedom. It would be interesting to derive different information criteria using the debiased estimators.

The above three chapters assume a sparse high dimensional linear model under strong distributional and moment assumptions, which are often violated in practice. A flexible modeling approach for quantile prediction using nonparametric C-vine couplas (i.e., a bivariate copulas construction obtained by recursive conditioning) is introduced in Chapter 4. The proposed model can be adapted to both low and high dimensional data. However, the major concern is that the proposed algorithm is computational and can only be evaluated by finite sample performance. One possible development would be a study of the asymptotic theory of the proposed

method. Additionally, the proposed model does not perform variable selection, i.e., the final C-vine copulas include all candidate variables, which is computationally inefficient when the number of variables is large. Developing stopping criteria is an interesting research topic. Another limitation is that we perform a pre-selection based on partial correlations of the candidate variables with the response variable for reducing computational complexity. However, this pre-selection approach is purely computational and excludes most variables in each step. Due to this pre-selection and stepwise optimization, a large proportion of C-vine copulas are overlooked. Thus, the global optimality of the final C-vine copulas is not guaranteed. A possible extension is to consider other pre-selection methods or to incorporate backward selection for including overlooked C-vine copulas.

We have seen in the first three chapters that regression-based approaches dealing with high dimensional data has a steady theoretical guarantee, whereas the vine copulas based modeling is more appealing when the objective is prediction. Another popular approach modeling dependence structures is graphical models with the possibility of regularization selecting edges, which are parameters of graphical models. A regularized composite graphical model can be proposed to obtain a robust estimator of the parameters with sparse representation. In addition, a similar debiasing approach, as used in the regression models, was proposed in Janková and van de Geer (2018) for graphical models. Developing information criteria based on the debiased graphical model is another interesting aspect that worth investigating.

List of figures

List of tables

Bibliography

Ando, T. and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265.

Ando, T. and Li, K.-C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, 45(6):2654–2679.

Bartle, R. G. and Joichi, J. T. (1961). The preservation of convergence of measurable functions under composition. *Proceedings of the American Mathematical Society*, 12(1):122–126.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20:451–468.

Bayati, M., Erdogdu, M. A., and Montanari, A. (2013). Estimating Lasso risk and noise level. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, pages 944–952.

Bayati, M. and Montanari, A. (2011a). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

Bayati, M. and Montanari, A. (2011b). The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding

algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Bedford, T. and Cooke, R. (2001). Monte Carlo simulation of vine dependent random variables for applications in uncertainty analysis. *ESREL 2003*.

Bedford, T. and Cooke, R. M. (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068.

Belloni, A. and Chernozhukov, V. (2011). ℓ1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *The Annals of statistics*, 46(6B):3643.

Bernard, C. and Czado, C. (2015). Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, 138:104–126.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bloznelis, D., Claeskens, G., and Zhou, J. (2019). Composite versus model-averaged quantile regression. *Journal of Statistical Planning and Inference*, 200:32 – 46.

Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electronic Journal of Statistics*, 10(2):3894–3944.

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.

Bradic, J. and Kolar, M. (2017). Uniform inference for high-dimensional quantile regression: linear functionals and regression rank scores. *arXiv preprint arXiv:1702.06209*.

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

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Chang, B. and Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics & Data Analysis*, 139:45–63.

Charpentier, A., Fermanian, J.-D., and Scaillet, O. (2007). The estimation of copulas: Theory and practice. In Rank, J., editor, *Copulas: from theory to application in finance*, pages 35–64. London : Risk Books.

Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *The Econometrics Journal*, 12:S50–S67.

Cheng, C. (1995). Uniform consistency of generalized kernel estimators of quantile density. *The Annals of Statistics*, 23(6):2285–2291.

Cheng, G., Wang, S., and Yang, Y. (2015). Forecast combination under heavy-tailed errors. *Econometrics*, 3(4):797–824.

Cheng, K.-F. (1984). On almost sure representation for quantiles of the product limit estimator with applications. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 426–443.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

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.

Czado, C. (2019). *Analyzing Dependent Data with Vine Copulas: A Practical Guide With R*. Lecture Notes in Statistics. Springer International Publishing.

Czado, C., Schepsmeier, U., and Min, A. (2012). Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, 12(3):229–255.

Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457.

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.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719.

Dißmann, J. F. (2010). *Statistical inference for regular vines and application*. Diploma thesis, Technische Universität München.

Donoho, D., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Wood, S. N., Wüest, R. O., and Hartig, F. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88(4):485–504.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge University Press.

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.

Fenske, N., Kneib, T., and Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, 106(494):494–510.

Geenens, G. (2014). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, 109(505):346–358.

Geenens, G., Charpentier, A., and Paindaveine, D. (2017). Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873.

Gijbels, I. and Mielniczuk, J. (1990). Estimating the density of a copula function. *Communications in Statistics-Theory and Methods*, 19(2):445–464.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Haff, I. H., Aas, K., and Frigessi, A. (2010). On the simplified pair-copula construction—simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5):1296–1310.

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

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of statistical software*, 27(5):1–32.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899. With discussion and a rejoinder by the authors.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–417. With discussion and a rejoinder by the authors.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian journal of statistics*, pages 65–70.

Huang, H. (2020). Asymptotic risk and phase transition of l_1-penalized robust estimator. *The Annals of Statistics, to appear.*

Höge, M., Guthke, A., and Nowak, W. (2019). The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, 572:96 – 107.

Jameson, G. (2014). Some inequalities for (a+ b) p and (a+ b) p+(a- b) p. *The Mathematical Gazette*, 98(541):96–103.

Janková, J. and van de Geer, S. (2018). Inference in high-dimensional graphical models. *Handbook of Graphical Models*, pages 325–348.

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

Javanmard, A. and Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

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.

Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, pages 120–141.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506.

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

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

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

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

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262.

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

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

Koenker, R. and Bassett, G. (1978b). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Komunjer, I. (2013). Quantile prediction, Chapter 17 in Handbook of Financial Econometrics, edited by Y. Ait-Sahalia and L.P. Hansen.

Kraus, D. and Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis*, 110:1–18.

Kurowicka, D. (2009). Some results for different strategies to choose optimal vine truncation based on wind spread data. In *3rd Vine Copula Workshop, Oslo*.

Lei, L., Bickel, P. J., and El Karoui, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Probability Theory and Related Fields*, 172(3-4):983–1079.

Li, S. (2017). Debiasing the debiased lasso with bootstrap. *arXiv preprint arXiv:1711.03613*.

Mousavi, A., Maleki, A., and Baraniuk, R. G. (2013). Parameterless optimal approximate message passing. *arXiv preprint arXiv:1311.0035*.

Mousavi, A., Maleki, A., and Baraniuk, R. G. (2018). Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148.

Nagler, T. and Kraus, D. (2018). vinereg: D-vine quantile regression. *R package version 0.5. 0*.

Nagler, T., Schellhase, C., and Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. *Dependence Modeling*, 5(1):99–120.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. (1994). The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. *Sea Fisheries Division, Technical Report*, 48:411.

Noh, H., Ghouch, A. E., and Bouezmarni, T. (2013). Copula-based regression estimation and inference. *Journal of the American Statistical Association*, 108(502):676–688.

Noh, H., Ghouch, A. E., and Van Keilegom, I. (2015). Semiparametric conditional quantile estimation through copula-based multivariate models. *Journal of Business & Economic Statistics*, 33(2):167–178.

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.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Rao, R. C. (1973). *Linear statistical inference and its applications*, volume 2. Wiley New York.

Rigollet, P., Tsybakov, A., et al. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771.

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

Schallhorn, N., Kraus, D., Nagler, T., and Czado, C. (2017). D-vine quantile regression with discrete variables. *arXiv preprint arXiv:1705.08310*.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449–460.

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.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stoeber, J., Joe, H., and Czado, C. (2013). Simplified pair copula constructions—limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118.

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

Tukey, J. W. (1965). Which part of the sample contains the information? *Proceedings of the National Academy of Sciences of the United States of America*, 53(1):127.

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.

Van Keilegom, I. and Veraverbeke, N. (1998). Bootstrapping quantiles in a fixed design regression model with censored data. *Journal of Statistical Planning and Inference*, 69(1):115–131.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

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

Wen, K. and Wu, X. (2015). An improved transformation-based kernel estimator of densities on the unit interval. *Journal of the American Statistical Association*, 110(510):773–783.

Wied, D. and Weißbach, R. (2012). Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53(1):1–21.

Xiao, Z. and Koenker, R. (2009). Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. *Journal of the American Statistical Association*, 104(488):1696–1712.

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.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588.

Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(1):176–222.

Yu, K. and Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.

Zhao, S., Zhou, J., and Li, H. (2016). Model averaging with high-dimensional dependent data. *Economics Letters*, 148:68 – 71.

Zhao, T., Kolar, M., and Liu, H. (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv preprint arXiv:1412.8724*.

Zhao, W., Zhang, F., and Lian, H. (2019). Debiasing and distributed estimation for high-dimensional quantile regression. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Zhou, J., Claeskens, G., and Bradic, J. (2019). Detangling robustness in high-dimensions: composite versus model-averaged estimation. *Technical report*.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

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

Doctoral dissertations of the Faculty of Economics and Business

A list of doctoral dissertations from the Faculty of Economics and Business and be found at the following website:
`http://www.kuleuven.be/doctoraatsverdediging/archief.htm`

KU LEUVEN

FACULTY OF ECONOMICS
AND BUSINESS

AFDELING
Adres
3000 LEUVEN, België
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
@kuleuven.be
www.kuleuven.be