Graduate School of Decision Sciences

Universität Konstanz

GSDS
Working Paper
No. 2017-08

A Joint Quantile and Expected Shortfall Regression Framework

Timo Dimitriadis
Sebastian Bayer

April 2017

Graduate School of Decision Sciences

All processes within our society are based on decisions – whether they are individual or collective decisions. Understanding how these decisions are made will provide the tools with which we can address the root causes of social science issues.

The GSDS offers an open and communicative academic environment for doctoral researchers who deal with issues of decision making and their application to important social science problems. It combines the perspectives of the various social science disciplines for a comprehensive understanding of human decision behavior and its economic and political consequences.

The GSDS primarily focuses on economics, political science and psychology, but also encompasses the complementary disciplines computer science, sociology and statistics. The GSDS is structured around four interdisciplinary research areas: (A) Behavioural Decision Making, (B) Intertemporal Choice and Markets, (C) Political Decisions and Institutions and (D) Information Processing and Statistical Analysis.

GSDS – Graduate School of Decision Sciences
University of Konstanz
Box 146
78457 Konstanz

Phone: +49 (0)7531 88 3633
Fax: +49 (0)7531 88 5193
E-mail: gsds.office@uni-konstanz.de

-gsds.uni-konstanz.de

ISSN: 2365-4120
February 2017

© 2017 by the author(s)

A Joint Quantile and Expected Shortfall Regression Framework

Timo Dimitriadis[†] and Sebastian Bayer

University of Konstanz, Department of Economics, 78457 Konstanz, Germany

This Version: 07.04.2017

Abstract

We introduce a novel regression framework which simultaneously models the quantile and the Expected Shortfall of a response variable given a set of covariates. The foundation for this joint regression is a recent result by Fissler and Ziegel (2016), who show that the quantile and the ES are jointly elicitable. This joint elicitability allows for M- and Z-estimation of the joint regression parameters. Such a parameter estimation is not possible for an Expected Shortfall regression alone as Expected Shortfall is not elicitable. We show consistency and asymptotic normality for the M- and Z-estimator under standard regularity conditions. The loss function used for the M-estimation depends on two specification functions, whose choices affect the properties of the resulting estimators. In an extensive simulation study, we verify the asymptotic properties and analyze the small sample behavior of the M-estimator under different choices for the specification functions. This joint regression framework allows for various applications including estimating, forecasting and backtesting Expected Shortfall, which is particularly relevant in light of the upcoming introduction of Expected Shortfall in the Basel Accords.

Keywords: Expected Shortfall, Joint Elicitability, Joint Regression, M-estimation

1. Introduction

In the past few years, Expected Shortfall (ES) has increasingly become the object of interest for practitioners, academics and regulators due to its upcoming introduction into the Basel Accords (Basel Committee, 2016). The ES at some fixed probability level α is defined as the mean of the returns which are smaller than the α-quantile of the return distribution. This risk measure has the desired ability to capture information from the whole left tail of the return distribution, which is particularly important for measuring extreme financial risks. So far, the most commonly used risk measure in the financial literature is the Value-at-Risk (VaR), which is the α-quantile of the return distribution. Its popularity is mainly due to the fact that so far the Basel Accords stipulate its use for the calculation of capital requirements for banks. The main drawback of the VaR is its inability to capture tail risks beyond itself (Artzner et al., 1999; Basel Committee, 2013), a deficiency which is overcome by the ES. However, in contrast to VaR, ES is not elicitable (Gneiting, 2011), which means that there exists no loss function which the ES uniquely minimizes in expectation. Consequently, modeling the ES of a random variable Y given a vector of covariates X through a linear regression, $\mathrm{ES}_\alpha(Y|X) = X'\theta^e$ is infeasible since estimation of the regression parameter vector θ^e through M-estimation requires such a loss function. Most recently, Fissler and Ziegel (2016) show that the quantile (VaR) and ES are jointly elicitable by proposing a joint loss function whose expectation is minimized by these functionals.

[†]Corresponding author
Email addresses: timo.dimitriadis@uni-konstanz.de, sebastian.bayer@uni-konstanz.de

In this paper we introduce a joint regression framework where we simultaneously model a linear quantile regression and a linear ES regression in the sense that $Q_\alpha(Y|X) = X'\theta^q$ and $\mathrm{ES}_\alpha(Y|X) = X'\theta^e$. The existence of the aforementioned joint loss function enables us to jointly estimate the regression parameters through minimization of this loss function. This estimation technique is in general known as M-estimation. We also introduce a Z-estimator for the regression parameters which, instead of minimizing a loss function, finds the multivariate root of a set of estimating equations. To the best of our knowledge such a joint regression and the associated joint parameter estimation are new to the literature. We show consistency and asymptotic normality for both estimators under standard regularity conditions.

The joint loss function, the estimating equations and also the parameter estimates depend on two specification functions, which can be chosen freely from some class of functions. Even though consistency and asymptotic normality hold for all applicable choices of these specification functions, the necessary moment conditions and the resulting asymptotic covariance matrix of the estimators depend on them. We discuss the choices of these functions in the context of asymptotic efficiency, minimal regularity conditions and the numerical performance of the estimators.

The estimation of the asymptotic covariance matrix imposes difficulties, especially for extreme probability levels of the quantile and ES and for small sample sizes. First, estimation of the density quantile function causes problems since estimating the tails of a density function is inaccurate due to the small amount of observations in the tails. This issue is well known from quantile regression and consequently, we rely on estimation methods from this field (Koenker, 2005). Second, we have to estimate the (truncated) variance of the negative quantile residuals conditional on the covariates, a nuisance quantity which is new to the literature. This task especially suffers from limited sample sizes since for extreme quantile levels, we are left with very few negative quantile residuals. We introduce several estimators for this quantity which are able to cope with limited sample sizes and which can model the dependency of the negative quantile residuals on the covariates.

We conduct a Monte-Carlo simulation study where we compare different choices for the specification functions, investigate the small sample behavior of our estimator, verify its asymptotic properties and compare the performance of the different estimators for the asymptotic covariance matrix. Our simulations include one regression design with homoskedastic and two more complicated designs with heteroskedastic error terms, which complicate the estimation of the asymptotic covariance matrix. For all three data generating processes, we can empirically verify consistency and asymptotic normality of the M-estimator depending on several different choices of the specification functions. However, we find that its small sample performance highly depends on these choices, whereas choices resulting in a positively homogeneous loss function (Nolde and Ziegel, 2017; Efron, 1991) result in a superior performance in terms of numerical stability of the estimator, asymptotic efficiency and computation times. Eventually, we also evaluate joint estimation of the quantile and ES in a univariate setting and compare our M-estimation approach to existing approaches in the literature from Chen (2008) and Brazauskas et al. (2008). We find a similar performance of all three estimators in estimating the sample ES, however, our M-estimation technique dominates the other approaches in terms of estimating the asymptotic variance of the ES estimates.

This joint regression technique for the quantile and ES has a wide range of potential applications, as it generalizes quantile regression to the pair consisting of the quantile and the ES. It opens up the possibility to extend all existing applications of quantile regression on VaR in the financial literature to ES such as e.g. in Gaglianone et al. (2011), Koenker and Xiao (2006), Engle and Manganelli (2004) and Halbleib and Pohlmeier (2012). Extensions for estimation, forecasting and backtesting methods for ES are particularly sought-after in light of the upcoming shift from VaR to ES in the recent Basel Accords.

The rest of the paper is organized as follows. In Section 2, we introduce the joint regression framework, the underlying regularity conditions and the asymptotic properties of our estimator. The proofs are deferred to the Appendix. Section 3 provides details on the numerical implementation of the parameter estimation and on the estimation of the asymptotic covariance matrix. Section 4 presents an extensive Monte-Carlo simulation and Section 5 concludes.

2. Methodology

2.1. Elicitability and Identifiability of Expected Shortfall

Regression techniques such as mean or quantile regression and the underlying M-estimation of the parameters are closely related to the theory of elicitable functionals and strictly consistent loss functions. The M-estimation technique is always based on a certain loss function, whose expectation must be minimized by the quantity we intend to estimate. Thus, the infeasibility of estimating regression parameters for the ES stand-alone is due to the unavailability of such a loss function (Gneiting, 2011).

Following Lambert et al. (2008), Gneiting (2011) and Fissler and Ziegel (2016), we introduce the concept of (multivariate) p-elicitability and p-identifiability. We consider a random variable $Z \in \mathcal{Z} \subseteq \mathbb{R}^d$, a class of distributions $\mathcal{P}$ on $\mathcal{Z}$ and a domain of action $D \subseteq \mathbb{R}^p, p \in \mathbb{N}$ for the functional $T : \mathcal{P} \to D$. We call a loss function $L : D \times \mathcal{Z} \to \mathbb{R}$ *strictly consistent* for the functional T relative to the class of distributions $\mathcal{P}$, if $T(F)$ is the unique minimizer of $\mathbb{E}_F\big[L(x, Z)\big]$ for all distributions $F \in \mathcal{P}$. Furthermore, we call a p-dimensional functional *p-elicitable* relative to the class $\mathcal{P}$, if there exists a loss function L which is strictly consistent for T relative to $\mathcal{P}$. If the dimension p is clear from the context, we simply call the functional elicitable instead of p-elicitable.

Closely related to elicitability is the concept of identifiability (see e.g. Steinwart et al., 2014, Fissler and Ziegel, 2016 and Nolde and Ziegel, 2017). A functional T is *identifiable* w.r.t. $\mathcal{P}$, if there exists a function $V : D \times \mathcal{Z} \to \mathbb{R}^p$ such that $\mathbb{E}_F\big[V(t, Z)\big] = 0$ if and only if $t = T(Z)$ for all distributions $F \in \mathcal{P}$. For univariate functionals, the concept of identifiability implies elicitability under some additional assumptions (Steinwart et al., 2014). However, for multivariate functionals, this relationship is still under research.

Given the quantile $Q_\alpha(Z)$ at probability level α, the corresponding ES at level α is defined as

$$\mathrm{ES}_\alpha(Z) = \mathbb{E}\big[Z \,\big|\, Z \le Q_\alpha(Z)\big]. \tag{2.1}$$

Gneiting (2011) shows that the ES is not elicitable with respect to any class $\mathcal{P}$ of probability distributions on the interval $I \subseteq \mathbb{R}$ that contains the measures with finite support, or the finite mixtures of the absolutely continuous distributions with compact support. This result has several consequences for the risk measure ES. First, consistent and meaningful ranking of competing forecasts for the functional ES is infeasible. Second, estimation of ES by means of M-estimation, i.e. by minimizing some associated loss function cannot be carried out. Closely related, the estimation of parameters of an ES regression framework in the sense that $\mathrm{ES}_\alpha(Y|X) = X'\theta^e$ by means of M-estimation is also infeasible.

Fissler and Ziegel (2016) show that the pair consisting of the quantile and the corresponding ES at common probability level α is 2-elicitable relative to the class of all distributions with finite first moments and unique α-quantiles. For a random variable $Z \in \mathbb{R}$ with distribution F with finite first moments and unique α-quantiles, the corresponding class of strictly consistent loss functions is given by

$$\begin{aligned} L(Z, q, e) = {} & \big(\mathbb{1}_{\{Z \le q\}} - \alpha\big)\, G_1(q) - \mathbb{1}_{\{Z \le q\}} G_1(Z) \\ & + G_2(e)\left(e - q + \frac{(q - Z)\mathbb{1}_{\{Z \le q\}}}{\alpha}\right) - \mathcal{G}_2(e) + a(Z), \end{aligned} \tag{2.2}$$

where $\mathcal{G}_2' = G_2$, $\mathcal{G}_2$ is strictly increasing and strictly convex, G_1 is increasing, and a and G_1 are F-integrable. Since the definition of ES already depends on the respective quantile, the fact that ES is only jointly elicitable with the quantile is quite intuitive. Even though the authors present this class of loss functions in the context of forecast ranking (Fissler et al., 2016), we apply this function for M-estimation of the joint regression parameters. We obtain the identification functions for the pair consisting of the quantile and the ES by differentiating (2.2) for all $Z \neq q$ (Nolde and Ziegel, 2017),

$$V(Z, q, e) = \begin{pmatrix} \frac{1}{\alpha}\big(\mathbb{1}_{\{Z \le q\}} - \alpha\big)\big[\alpha G_1'(q) + G_2(e)\big] \\ G_2'(e)\big[e - q + \frac{1}{\alpha}(q - Z)\mathbb{1}_{\{Z \le q\}}\big] \end{pmatrix}, \tag{2.3}$$

where the functions G_1 and G_2 are given as in (2.2) and are furthermore assumed to be continuously differentiable. In our context, (2.3) is used for specifying the estimating equations for the Z-estimator.

2.2. The Joint Regression Framework

In this section, we formally introduce the joint linear quantile and ES regression framework. For that, we assume that there is a real-valued response variable Y and a k-dimensional vector of explanatory variables $X = (X_1, \dots, X_k)' \in \mathcal{X} \subseteq \mathbb{R}^k$. The cumulative distribution function of the conditional distribution of Y given X will henceforth be denoted by $F_{Y|X}$ and the conditional density function by $f_{Y|X}$ (whenever it exists).

The regression framework which jointly models the conditional quantile and the conditional Expected Shortfall of Y given X for some fixed $\alpha \in (0, 1)$ is given by the linear regression equations

$$\begin{aligned} Y &= X'\theta^q + u^q, \quad \text{and} \\ Y &= X'\theta^e + u^e, \end{aligned} \tag{2.4}$$

with parameters $\theta = (\theta^{q\prime}, \theta^{e\prime})' \in \Theta \subseteq \mathbb{R}^{2k}$, where

$$Q_\alpha(u^q|X) = 0, \qquad \text{and} \qquad \mathrm{ES}_\alpha(u^e|X) = 0. \tag{2.5}$$

We propose both, an M-estimation and a Z-estimation procedure for the compound regression parameter vector θ. For the M-estimation procedure, we adapt the strictly consistent joint loss function for the quantile and ES as given in (2.2), and generalize it such that it can be applied to a regression framework,

$$\begin{aligned} \rho(Y, X, \theta) = {} & \big(\mathbb{1}_{\{Y \le X'\theta^q\}} - \alpha\big) G_1(X'\theta^q) - \mathbb{1}_{\{Y \le X'\theta^q\}} G_1(Y) \\ & + G_2(X'\theta^e)\left(X'\theta^e - X'\theta^q + \frac{(X'\theta^q - Y)\mathbb{1}_{\{Y \le X'\theta^q\}}}{\alpha}\right) \\ & - \mathcal{G}_2(X'\theta^e) + a(Y), \end{aligned} \tag{2.6}$$

where the function G_1 is twice continuously differentiable, $\mathcal{G}_2$ is three times continuously differentiable, $\mathcal{G}_2' = G_2$, G_2 and G_2' are strictly positive, G_1 is increasing and a and G_1 are integrable. We discuss the choice of the *specification functions* G_1 and $\mathcal{G}_2$ in a theoretical context in Section 2.4 and by their numerical performance in Section 4.2. The corresponding (ρ-type) M-estimator is defined by a sequence $\hat{\theta}_{\rho,n}$, such that

$$\hat{\theta}_{\rho,n} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta). \tag{2.7}$$

Instead of minimizing some objective function $\rho(Y, X, \theta)$, such as in (2.6) and (2.7), we can also define the corresponding Z-estimator (or ψ-type M-estimator), which sets a vector of *estimating equations*, denoted by $\psi(Y, X, \theta)$, to zero. More general, it suffices that these estimating equations converge to zero almost surely (or even in probability). Formally, the Z-estimator is a sequence $\hat{\theta}_{\psi,n}$, such that

$$\frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, X_i, \hat{\theta}_{\psi,n}) \to 0 \tag{2.8}$$

almost surely (or in probability), where

$$\psi(Y, X, \theta) = \begin{pmatrix} \frac{1}{\alpha}\big(\mathbb{1}_{\{Y \le X'\theta^q\}} - \alpha\big)\big(\alpha X G_1'(X'\theta^q) + X G_2(X'\theta^e)\big) \\ X G_2'(X'\theta^e)\left(X'\theta^e - X'\theta^q + \frac{1}{\alpha}(X'\theta^q - Y)\mathbb{1}_{\{Y \le X'\theta^q\}}\right) \end{pmatrix}, \tag{2.9}$$

and where the functions G_1 and G_2 are given above.

When the loss function $\rho(Y, X, \theta)$ is continuously differentiable in θ, it is obvious that the two estimation approaches given in (2.7) and (2.8) are equivalent. However, in this case the loss function $\rho(Y, X, \theta)$ is not differentiable and $\psi(Y, X, \theta)$ is even discontinuous at the points where $Y = X'\theta^q$. Thus, we treat these two estimation approaches as different estimators and show their asymptotic behavior separately in the following section.

2.3. Asymptotic Properties

In this section, we present the asymptotic properties of both, the Z- and the M-estimator of the regression parameters given in (2.7) and (2.8). We show consistency of the estimators in Theorem 2.2 and Theorem 2.3 and asymptotic normality in Theorem 2.4 and Theorem 2.5, respectively. We impose the following set of regularity conditions for our estimators.

Assumption 2.1 (Regularity Conditions).

($\mathcal{A}$-1) Let (Y_i, X_i) for $i = 1, \dots, n$ be an iid series of random variables, distributed such as (Y, X) where $Y \in \mathbb{R}$ and $X \in \mathcal{X} \subseteq \mathbb{R}^k$. We assume that the conditional distribution $F_{Y|X}$ has finite first moments and is absolutely continuous with probability density function $f_{Y|X}$, which is strictly positive, continuous and bounded in a neighbourhood of $X'\theta_0^q$.

($\mathcal{A}$-2) The parameter space $\Theta \subset \mathbb{R}^{2k}$ is compact with non-empty interior.

($\mathcal{A}$-3) Let the conditional quantile and conditional ES at level α of Y given X be linear functions in X given by the regression equations

$$Y = X'\theta_0^q + u^q, \quad \text{and} \tag{2.10}$$
$$Y = X'\theta_0^e + u^e, \tag{2.11}$$

where

$$Q_\alpha(u^q|X) = 0, \qquad \text{and} \qquad \mathrm{ES}_\alpha(u^e|X) = 0, \tag{2.12}$$

and where the true regression parameters $\theta_0 \in \mathrm{int}(\Theta)$ are in the interior of Θ.

($\mathcal{A}$-4) Let the functions $\rho(Y, X, \theta)$ be given as in (2.6) and $\psi(Y, X, \theta)$ be given as in (2.9), where the function G_1 is twice continuously differentiable, $\mathcal{G}_2$ is three times continuously differentiable, $\mathcal{G}_2' = G_2$, G_2 and G_2' are strictly positive, G_1 is increasing and a and G_1 are integrable.

($\mathcal{A}$-5) We assume that the matrix $\mathbb{E}\left[XX'\right]$ is positive definite.

($\mathcal{A}$-6) We assume that certain moments of X are finite. For the exact moment conditions, we refer to Appendix A.

Assumption ($\mathcal{A}$-1) is a combination of typical conditions in the context of mean and quantile regression. The condition that the distribution $F_{Y|X}$ is absolutely continuous and has a strictly positive, bounded and continuous density function in a neighborhood of $X'\theta_0^q$ is also imposed for the asymptotic theory of quantile regression. Existence of the conditional moments of Y given X appears in mean regression and is subject to our regularity conditions since ES is a truncated mean.

Compactness of the parameter space Θ generally simplifies the proofs of the asymptotic results. However, in this setup, it is crucial for a flexible choice of the function G_2. Assume we choose a function G_2 such that $\lim_{z\to-\infty} zG_2'(z) = 0$, which holds for many of the possible choices for G_2. Then, we get that the function ψ_2 is redescending to zero for all θ^e such that $X'\theta^e \to -\infty$, which makes it impossible to show consistency of the Z-estimator for non-compact parameter spaces.

The full rank condition ($\mathcal{A}$-5) is typical for any regression design with stochastic regressors in order to exclude perfect multicollinearity of the regressors. Furthermore, the existence of certain moments of the explanatory variables as assumed in conditions ($\mathcal{M}$-1) - ($\mathcal{M}$-4) in Appendix A is also standard in any regression theory relying on stochastic regressors.

Theorem 2.2. Assume the regularity conditions given in Assumption 2.1 and the Moment Conditions ($\mathcal{M}$-1) in Appendix A hold true. Then, for every sequence $\hat{\theta}_{\psi,n} \in \Theta$ satisfying $\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, X_i, \hat{\theta}_{\psi,n}) \to 0$ almost surely, it holds that

$$\hat{\theta}_{\psi,n} \to \theta_0 \tag{2.13}$$

almost surely.

The proof for this theorem is given in Appendix B.1

Theorem 2.3. Assume that the regularity conditions in Assumption 2.1 and the Moment Conditions ($\mathcal{M}$-2) in Appendix A hold true. Then, for every sequence $\hat{\theta}_{\rho,n} \in \Theta$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i, \hat{\theta}_{\rho,n}) \leq \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i, \theta_0) + o_P(1), \tag{2.14}$$

it holds that

$$\hat{\theta}_{\rho,n} \xrightarrow{\mathbb{P}} \theta_0. \tag{2.15}$$

The proof for this theorem is given in Appendix B.2.

Theorem 2.4. Assume the regularity conditions given in Assumptions 2.1 and the Moment Conditions ($\mathcal{M}$-3) in Appendix A hold true. Then, for every sequence $\hat{\theta}_{\psi,n} \in \Theta$ satisfying $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(Y_i, X_i, \hat{\theta}_{\psi,n}) \xrightarrow{\mathbb{P}} 0$ and $\hat{\theta}_{\psi,n} \xrightarrow{\mathbb{P}} \theta_0$, it holds that

$$\sqrt{n}\left(\hat{\theta}_{\psi,n} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Lambda^{-1}C\Lambda^{-1}\right), \tag{2.16}$$

with

$$\Lambda = \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix}, \tag{2.17}$$

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{2.18}$$

where

$$\Lambda_{11} = \mathbb{E}\left[(XX')\left(\frac{\alpha G_1'(X'\theta_0^q) + G_2(X'\theta_0^e)}{\alpha}\right) f_{Y|X}(X'\theta_0^q)\right], \tag{2.19}$$

$$\Lambda_{22} = \mathbb{E}\left[(XX')G_2'(X'\theta_0^e)\right], \tag{2.20}$$

and

$$C_{11} = \frac{1-\alpha}{\alpha}\mathbb{E}\left[(XX')\left(\alpha G_1'(X'\theta_0^q) + G_2(X'\theta_0^e)\right)^2\right], \tag{2.21}$$

$$C_{12} = C_{21} = \frac{1-\alpha}{\alpha}\mathbb{E}\left[(XX')\left(\alpha G_1'(X'\theta_0^q) + G_2(X'\theta_0^e)\right)G_2'(X'\theta_0^e)\left(X'\theta_0^q - X'\theta_0^e\right)\right], \tag{2.22}$$

$$C_{22} = \mathbb{E}\left[(XX')G_2'^2(X'\theta_0^e)\left(\frac{1}{\alpha}\operatorname{Var}\left(Y - X'\theta_0^q \middle| Y \leq X'\theta_0^q, X\right) + \frac{1-\alpha}{\alpha}\left(X'\theta_0^q - X'\theta_0^e\right)^2\right)\right]. \tag{2.23}$$

The proof for this theorem is given in Appendix B.3.

Theorem 2.5. Assume that the regularity conditions in Assumptions 2.1 and the Moment Conditions ($\mathcal{M}$-4) in Appendix A hold true. Then, for every sequence $\hat{\theta}_{\rho,n} \in \Theta$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i, \hat{\theta}_{\rho,n}) \leq \inf_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i, \theta) + o_P(n^{-1}) \tag{2.24}$$

and $\hat{\theta}_{\rho,n} \xrightarrow{\mathbb{P}} \theta_0$, it holds that

$$\sqrt{n}(\hat{\theta}_{\rho,n} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0, \Lambda^{-1}C\Lambda^{-1}\big), \tag{2.25}$$

where the matrices Λ and C are given as in Theorem 2.4.

The proof for this theorem is given in Appendix B.4. The preceding theorems show that even though the estimators $\hat{\theta}_{\rho,n}$ and $\hat{\theta}_{\psi,n}$ are not identical, their asymptotic distribution coincides. For our numerical implementation in the subsequent chapters, we will use the M-estimator $\hat{\theta}_{\rho,n}$ as given in (2.7) since numerically, optimization of an objective function is easier than root-searching as further described in Section 3.1. This holds especially since the empirical estimating equations $\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, X_i, \theta)$ do not necessarily exhibit an exact root (for the quantile parameters θ^q) due to its discontinuity induced by the indicator function.

Remark 2.6 (Quantile Regression). Notice that the asymptotic covariance matrix of the quantile-specific parameter estimates $\hat{\theta}^q$ is given by

$$\alpha(1-\alpha)D_1^{-1}D_0D_1^{-1} \tag{2.26}$$

where

$$D_1 = \mathbb{E}\left[(XX')\big(\alpha G_1'(X'\theta_0^q) + G_2(X'\theta_0^e)\big)f_{Y|X}(X'\theta_0^q)\right] \qquad \text{and} \tag{2.27}$$

$$D_0 = \mathbb{E}\left[(XX')\big(\alpha G_1'(X'\theta_0^q) + G_2(X'\theta_0^e)\big)^2\right]. \tag{2.28}$$

This simplifies to the well-known covariance matrix of quantile regression parameter estimates by setting $G_1(z) = z$ and $G_2(z) = 0$, which means ignoring the ES-specific part of our loss function and estimating equations. This demonstrates that the quantile regression method is nested in our regression procedure, also in terms of its asymptotic distribution.

Remark 2.7 (Estimation of the quantile and ES). We can use this regression framework to jointly estimate the quantile and ES of an iid sample $Y_1, \ldots, Y_n$ by simply choosing $X = \mathbf{1}_n$. Then, the asymptotic covariance matrix given in Theorem 2.4 and Theorem 2.5 simplifies to Σ, where

$$\Sigma_{11} = \frac{\alpha(1-\alpha)}{f_Y^2(\theta_0^q)}, \tag{2.29}$$

$$\Sigma_{12} = (1-\alpha)\frac{\theta_0^q - \theta_0^e}{f_Y(\theta_0^q)}, \tag{2.30}$$

$$\Sigma_{22} = \frac{1}{\alpha}\operatorname{Var}(Y - \theta_0^q | Y \leq \theta_0^q) + \frac{1-\alpha}{\alpha}(\theta_0^q - \theta_0^e)^2. \tag{2.31}$$

Notice that in this simplified case, the asymptotic covariance matrix is independent of the specification functions G_1 and $\mathcal{G}_2$ used in the loss function and in the estimating equations. Furthermore, (2.29) implies that quantile estimates stemming from our joint estimation procedure have the same asymptotic efficiency

as quantile estimates stemming from minimizing the check loss (or any choice from the class of generalized piecewise linear loss functions) and from sample quantiles (cf. Koenker, 2005).

The same holds true for the asymptotic efficiency of ES estimates stemming from our joint optimization procedure and from estimating the sample ES. Brazauskas et al. (2008) estimate the sample ES of an iid sample $Y_1, \dots Y_n$ based on the empirical distribution function $\hat{F}_n$ by

$$\widehat{\mathrm{ES}}_\alpha = \frac{1}{\alpha} \int_0^\alpha \widehat{F}_n^{-1}(u)\,\mathrm{d}u, \tag{2.32}$$

and show asymptotic normality of the estimator with asymptotic variance

$$\mathrm{AVar}\left(\widehat{\mathrm{ES}}_\alpha\right) = \frac{1}{\alpha^2} \int_{-\infty}^{F^{-1}(\alpha)} \int_{-\infty}^{F^{-1}(\alpha)} F(x \wedge y) - F(x)F(y)\,\mathrm{d}x\mathrm{d}y. \tag{2.33}$$

We show in Proposition B.5 in Appendix B.5 that under mild regularity conditions, this expression equals the ES-specific asymptotic variance part of our M-estimator as given in (2.31).

Chen (2008) estimate the sample ES based on the sample quantile estimator, $\hat{Q}_\alpha(Y)$ by

$$\widetilde{\mathrm{ES}}_\alpha = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{\{Y_i \le \hat{Q}_\alpha(Y)\}}}{\sum_{i=1}^n \mathbb{1}_{\{Y_i \le \hat{Q}_\alpha(Y)\}}}. \tag{2.34}$$

The author shows asymptotic normality of the estimator under more general dependence conditions on the data, however, for iid data the asymptotic variance boils down to

$$\mathrm{AVar}\left(\widetilde{\mathrm{ES}}_\alpha\right) = \mathrm{Var}\left(Y - Q_\alpha(Y) \middle| Y \le Q_\alpha(Y)\right). \tag{2.35}$$

We compare the numerical performance of our joint estimator to both estimators for the sample ES and also compare the performance of the associated estimators for the asymptotic variance in in Section 4.4.

2.4. Choice of the Specification Functions

The loss function and the estimating equations, given in (2.6) and (2.9) depend on two specification functions, G_1 and $\mathcal{G}_2$, which have to fulfill some regularity conditions for the asymptotic results, given in Assumption 2.1, condition ($\mathcal{A}$-4). The choice of these functions highly influences the performance of our regression procedure in terms of asymptotic efficiency, the necessary moment conditions of the regressors and the numerical performance of the optimization algorithm.

Efron (1991) and Nolde and Ziegel (2017) argue that for the estimation of regression parameters it is crucial that the associated loss functions are positively homogeneous of order b in the sense that for all $c > 0$,

$$L(cZ, ct) = c^b L(Z, t). \tag{2.36}$$

This is a typical property for loss functions since the ordering of the losses should be independent of the unit of measurement, e.g. the currency we measure the prices and risk forecasts with. Loss functions following this property guarantee that we can change the scaling and still obtain the same optima and consequently the same parameter estimates.

Nolde and Ziegel (2017) show that the only way for obtaining a strictly consistent and positively homogeneous[1] loss function of order b for the pair consisting of the quantile and the ES are given by the following choices:

$$b < 0: \quad G_1(x) = -c_0, \qquad \mathcal{G}_2(z) = c_1(-z)^b + c_0, \tag{2.37}$$

$$b = 0: \quad G_1(x) = d_0 \mathbb{1}_{\{x \le 0\}} + d_0' \mathbb{1}_{\{x > 0\}}, \qquad \mathcal{G}_2(z) = -c_1 \log(-z) + c_0, \tag{2.38}$$

$$b \in (0,1): \quad G_1(x) = \left(d_1 \mathbb{1}_{\{x \le 0\}} + d_1' \mathbb{1}_{\{x > 0\}}\right)|z|^b - c_0, \qquad \mathcal{G}_2(z) = -c_1(-z)^b + c_0, \tag{2.39}$$

[1] For $b = 0$, only the score differences are positively homogeneous. However, the ordering of the losses is still unaffected under this slightly weaker property.

for some constants $c_0, d_0, d_0' \in \mathbb{R}$ with $d_0 \leq d_0'$, $d_1, d_1' \geq 0$ and $c_1 > 0$.

Notice that for the above choices, we have to restrict the domain of $\mathcal{G}_2$ to the negative real line. However, since the (conditional) ES of financial assets for small probability levels is always negative, this is no critical restriction. In the estimation procedure, we have restrict the parameter space Θ such that $X'\theta^e < 0$ for all $\theta \in \Theta$ and for all $X \in \mathcal{X}$. Even though this restriction imposes theoretical problems for distributions of X with unbounded support, this issue can easily be solved for the numerical implementation of our estimator by simply transforming the data Y before the estimation such that $Y < 0$. For details on this, we refer to Section 3.1.

For the numerical implementation of our estimators, we use different choices for the specification functions where we choose the constants such that

$$b = -1: \quad G_1(x) = 0, \quad \mathcal{G}_2(z) = (-z)^{-1}, \tag{2.40}$$
$$b = 0: \quad G_1(x) = 0, \quad \mathcal{G}_2(z) = -\log(-z), \tag{2.41}$$
$$b = 0.5: \quad G_1(x) = 0, \quad \mathcal{G}_2(z) = -\sqrt{-z}. \tag{2.42}$$

Another popular choice for G_1 and $\mathcal{G}_2$ arises from the moment conditions of the regressors as given by the conditions ($\mathcal{M}$-1) - ($\mathcal{M}$-4) in Appendix A. By choosing $G_1(z) = 0$ and by choosing $\mathcal{G}_2$ such that G_2 and its first and second derivatives are bounded, the respective moment conditions simplify to

($\mathcal{M}$-1)' For the proof of Theorem 2.2, we assume that the following moments are finite:

- $\mathbb{E}\left[||X||^3\right]$
- $\mathbb{E}\left[||X||^2\mathbb{E}\left[|Y|\middle|X\right]\right]$

($\mathcal{M}$-2)' For the proof of Theorem 2.3, we assume that the following moments are finite:

- $\mathbb{E}\left[||X||^2\right]$
- $\mathbb{E}\left[|Y|\right]$
- $\mathbb{E}\left[\left|G_1(Y)\right|\right]$
- $\mathbb{E}\left[\left|a(Y)\right|\right]$

($\mathcal{M}$-3)' For the proof of Theorem 2.4, we assume that the following moments are finite:

- $\mathbb{E}\left[||X||^5\right]$
- $\mathbb{E}\left[||X||^4\mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^3\mathbb{E}\left[Y^2\middle|X\right]\right]$

($\mathcal{M}$-4)' For the proof of Theorem 2.5, we assume that the following moments are finite:

- $\mathbb{E}\left[\left|G_1(Y)\right|\right]$
- $\mathbb{E}\left[\left|a(Y)\right|\right]$
- $\mathbb{E}\left[||X||^4\right]$
- $\mathbb{E}\left[||X||^3\mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^2\mathbb{E}[Y^2|X]\right]$

Besides simplifying the notation of the moment conditions, the boundedness assumption also relaxes these moment conditions. Consider e.g. the choice $G_2(z) = \exp(z)$, which is proposed by Fissler et al. (2016). Then, the moment conditions given in Appendix A are of exponential order and thus rule out many fat-failed distributions for X. This motivates the usage of bounded functions for G_2 as e.g. the second proposal of Fissler et al. (2016): the cdf of the logistic distribution $G_2(z) = \exp(z)/\big(1 + \exp(z)\big)$. Further examples for bounded G_2 functions include the cdfs of any absolutely continuous distribution with the whole real line as support. In the simulation study in Section 4.2, we compare five choices for the specification functions in terms of mean square error, asymptotic efficiency of the estimator and computation times.

3. Numerical Estimation of the Model

3.1. Optimization

The parameter vector $\theta = (\theta^{q\prime}, \theta^{e\prime})'$ can either be estimated by M-estimation (by minimizing the ρ-function given in (2.6)), or by Z-estimation (by finding the multivariate root of the ψ-function given in (2.9)). As shown in Section 2.3, both estimators have the same asymptotic efficiency. However, in practice the implementation of the Z-estimator imposes some problems.

The numerical implementation of the Z-estimator relies on root-finding of the ψ function, which is implemented as in Method of Moments estimation by minimizing the function $\sum_i \psi(Y_i, X_i, \theta) \cdot \sum_i \psi(Y_i, X_i, \theta)'$ numerically. However, for many valid choices of the function G_2, it holds that $zG_2'(z) \to 0$ for $z \to -\infty$ and thus $\psi_2(Y, X, \theta) \to 0$ for $X'\theta^e \to -\infty$, which means that ψ_2 is redescending to zero. Furthermore, if we choose $G_1(z) = 0$ and $G_2(z) \to 0$ for $z \to -\infty$, it also holds that $\psi_1(Y, X, \theta) \to 0$ for $X'\theta^e \to -\infty$. Consequently, for estimates $\hat{\theta}^e_{\psi,n}$ such that $X'\hat{\theta}^e_{\psi,n} \to -\infty$, we get the same minimal value of the objective function $\sum_i \psi(Y_i, X_i, \hat{\theta}^e_{\psi,n}) \cdot \sum_i \psi(Y_i, X_i, \hat{\theta}^e_{\psi,n})' \to 0$ as for a sequence converging to the true regression parameters θ_0. Thus, the optimization algorithm for the Z-estimator is numerically very unstable and diverges in many simulation setups.

In contrast, the M-estimator which minimizes the ρ-function does not suffer from this problem since its objective function has a unique global optimum. For this reason, we rely on M-estimation of the regression model. Nevertheless, we face some difficulties in the numerical optimization procedure. Most importantly, the ρ-function is non-convex (Fissler, 2017), which has two main implications: First, it is not possible to solve the minimization problem by applying an efficient optimization technique such as linear or quadratic programming as in e.g. quantile regression (Koenker, 2005). Second, a non-convex objective function can exhibit several local optima so that standard optimization techniques might encounter difficulties in finding the global optimum. Furthermore, the ρ-function is not differentiable everywhere, which implies that we can only apply derivative-free optimization algorithms.

In order to find the global optimum of the ρ function, we estimate θ by repeated optimizations initialized with randomized starting values such as e.g. in Engle and Manganelli (2004). This means we run several optimization procedures using different starting values, compare the losses of the estimates of the different optimizations and select the one with the smallest loss as our final estimate. In practice, the resulting estimates are very close to the ones stemming from global optimization techniques such as e.g. simulated annealing, whereas the major advantage of the repeated optimization technique is its considerably lower computation time.

In detail, for obtaining pre-estimates for the approximate area of our parameters, we run a quantile regression with probability level α for the quantile parameters θ^q. For pre-estimates for the ES-specific parameter vector θ^e, we run a second quantile regression with probability level $\tilde{\alpha}$, where we choose $\tilde{\alpha}$ such that the $\tilde{\alpha}$-quantile and the α-ES coincide under normality, e.g. $\tilde{\alpha} = 1\%$ for $\alpha = 2.5\%$. For the repeated optimizations, we generate $N = 1000$ randomized starting values by adding a $2k$-dimensional normally distributed noise term with a standard deviation of $s = 0.1$ to these pre-estimates. We evaluate the average ρ-function at these randomized values and only keep the $M = 10$ starting values with the smallest average loss. Then, we optimize our objective function using these M different starting values by applying the Nelder-Mead Simplex optimization algorithm. Finally, out of these M estimates, we select the one with the smallest average loss as our final estimate. We provide an R package which implements this optimization procedure (Bayer and Dimitriadis, 2017a) and furthermore allows for flexible choices of the parameters N, M and s.

A further technical problem arises from the choice of the specifications functions which result in a positively homogeneous loss function. As discussed in Section 2.4, these choices require that we restrict θ^e such that $X'\theta^e < 0$. We ensure this condition by estimating the regression model for the transformed dependent variable $Y - \max(Y)$. We undo the data transformation by adding $\max(Y)$ to the estimated intercept parameters.

3.2. Asymptotic Covariance Estimation

While most parts of the asymptotic covariance matrix given in Theorem 2.5 are straightforward to estimate, two nuisance quantities impose some difficulties. The first is the density quantile function,

$$f_{Y|X}(X'\theta_0^q), \tag{3.1}$$

which is also subject to the quantile regression asymptotic covariance matrix and is hence already well investigated in this field (Koenker, 2005).

The second is the variance of the quantile residuals, conditional on the covariates and given that these residuals are negative,

$$\text{Var}\left(Y - X'\theta_0^q \middle| Y \leq X'\theta_0^q, X\right) = \text{Var}\left(u^q \middle| u^q \leq 0, X\right). \tag{3.2}$$

Estimation of this quantity is very demanding for two reasons. First, for very small probability levels such as e.g. $\alpha = 2.5\%$, which is a typical choice in financial risk management, we are left with very little observations (only $\alpha \cdot n$) such that $u^q \leq 0$. Second, modeling this truncated variance conditional on the covariates X is challenging, especially considering the problem of very small sample sizes.

3.2.1. The Density Quantile Function

Estimation of the density quantile function (3.1) is usually based on differentiating both sides of the identify $F(F^{-1}(\alpha)) = \alpha$, which yields

$$F'(F^{-1}(\alpha)) = \left(\frac{d}{d\alpha}F^{-1}(\alpha)\right)^{-1}. \tag{3.3}$$

Thus, it is straightforward to obtain an estimate for the density quantile function by the inverse of the difference quotient of the empirical quantile function,

$$\widehat{f}_{Y|X}(X_i'\theta_0^q) = \left(\frac{\widehat{F}_i^{-1}(\alpha + h_n) - \widehat{F}_i^{-1}(\alpha - h_n)}{2h_n}\right)^{-1}, \tag{3.4}$$

where $\widehat{F}_i^{-1}(\cdot)$ denotes an estimate of the empirical quantile function of the i-th observation. The quantity h_n denotes a bandwidth parameter which we choose according to Hall and Sheather (1988).

There is a vast literature on estimating F_i^{-1} (see Koenker, 2005) and in the following, we present two approaches which perform well in our simulation study. The main difference between the two approaches is that the first assumes that the quantile residuals are independent of the covariates, whereas the second allows for a linear dependence structure.

Under the assumption that the quantile regression residuals u^q are independent of the covariates X, it holds that $\widehat{F}_i^{-1}(\cdot) = \widehat{F}^{-1}(\cdot)$ for all i, i.e. the estimate is the same for all observations. In this case, we linearly interpolate the empirical quantile function of the quantile residuals in order to get an estimate for the function $\widehat{F}^{-1}(\cdot)$ and thus get the values $\widehat{F}^{-1}(\alpha \pm h_n)$ in (3.4) (Koenker, 1994). Henceforth, we refer to this as the *iid* estimator of the density quantile function.

The second approach we consider allows for a linear dependence of u^q on the covariates X, for which we apply the technique of Hendricks and Koenker (1992). They suggest to estimate two additional quantile regressions for the quantile levels $\alpha + h_n$ and $\alpha - h_n$, and we denote these estimated coefficients by $\widehat{\theta}^q_{(\alpha \pm h_n)}$. The estimated empirical quantile function is then simply obtained by $\widehat{F}_i^{-1}(\alpha \pm h_n) = X_i'\widehat{\theta}^q_{(\alpha \pm h_n)}$. As above, we get the estimate for the density quantile function by (3.4). Henceforth, we refer to this as the *nid* estimator of the density quantile function.

3.2.2. The Truncated Conditional Variance

The second quantity which is demanding to estimate is the conditional variance of the quantile residuals $u^q = Y - X'\theta_0^q$ given X and given that $u^q \leq 0$. Depending on the dynamics of the underlying process we present two different estimation approaches.

If the distribution of u^q is independent of the covariates X, we simply estimate (3.2) by the sample variance of the negative quantile residuals and we refer to this estimator as *ind* in the following. However, as only about $\alpha \cdot n$ of the quantile residuals are negative, this estimator suffers from the drawback of very small sample sizes for typical choices of α in the context of financial risk management.

Furthermore, our regression design and the associated asymptotic theory allow for a dependence between u^q and X which the *ind* estimator cannot capture. In order to explicitly model the relationship between the conditional variance of u^q and X, we assume a location-scale process for u^q,

$$u^q = m(X) + \sigma(X) \cdot \varepsilon \tag{3.5}$$

where $m(X) = \mathbb{E}[u^q|X]$ and $\sigma^2(X) = \mathrm{Var}(u^q|X)$ are the conditional mean and variance of u^q given X. However, $\sigma^2(X)$ represents the conditional variance of u^q given X, whereas we are interested in its truncated variant (given $u^q \leq 0$). One possibility is to estimate (3.5) only for those observations where $u^q \leq 0$, but this approach again suffers from very few negative quantile residuals and from additional noise induced by the estimation of the additional regression parameters.

We present a feasible alternative by assuming that ε follows some absolutely continuous distribution $G(0,1)$ with zero mean and unit variance, which implies that $u^q|X \sim G\big(m(X), \sigma(X)\big)$ with mean $m(X)$, variance $\sigma^2(X)$, distribution function F_G and density f_G. The quantities $m(X)$ and $\sigma^2(X)$ can then be estimated using all available observations of u_q and X. Eventually, we obtain the truncated conditional variance by the scaling formula

$$\mathrm{Var}(u^q|u^q \leq 0, X) = \int_{-\infty}^{0} z^2 h(z)\, dz - \left(\int_{-\infty}^{0} z h(z)\, dz\right)^2, \tag{3.6}$$

where $h(z) = f_G(z)/F_G(0)$ is the truncated conditional density function of u^q given X and given that $u^q \leq 0$.

In our implementation we assume, a linear dependence for the conditional mean and standard deviation, i.e. $m(X) = X'\eta$ and $\sigma(X) = X'\gamma$, for some parameter vectors $\eta, \gamma \in \mathbb{R}^k$. For the distribution of ε we assume either the Normal or the more flexible Student-t distribution in order to capture possible overkurtosis of the quantile residuals. We estimate the model parameters, including the degrees of freedom of the t-distribution, by maximum likelihood estimation using the presumed distribution of ε. Positivity of the conditional standard deviation $\sigma(X)$ is ensured by numerically excluding solutions that do not fulfil the condition $X'\gamma \geq 0$. We thus obtain estimates for the functions F_G and f_G and can consequently apply (3.6) using numerical integration techniques and thereby obtain a prediction for the truncated conditional variance. Depending on the distributional assumption on the quantile residuals u^q, we denote these estimators by *scl-N* or *scl-t*.

This approach can easily be extended in two ways. First, one could consider more flexible specifications for $m(X)$ and $\sigma(X)$, e.g. through polynomial regression, regression splines or nonparametric regression techniques (Fan and Yao, 1998). Second, it is straightforward to account for further properties of the distribution of the quantile residuals such as e.g. skewness by employing more flexible distributional choices for ε.

4. Simulation Study

In this section, we investigate the finite sample behavior of the M-estimator for the regression parameters and verify the asymptotic properties derived in Section 2.3 by means of simulations. Furthermore, we

compare the performance of different choices for the specification functions in terms of estimation accuracy, asymptotic efficiency and computation times and evaluate the precision of the different covariance matrix estimators described in Section 3.2.

4.1. Regression Models: Data Generating Processes

In order to assess the numerical properties of our joint regression model, we simulate data from a linear location-scale process,

$$Y|X = X'\gamma + (X'\eta) \cdot \varepsilon, \tag{4.1}$$

where $\varepsilon \sim F(0,1)$ has a zero mean, unit variance distribution, $X = \left(1, X_2, \ldots, X_k\right)'$ and with parameter vectors $\gamma, \eta \in \mathbb{R}^k$. For this process, the true conditional quantile and ES are linear functions in X, given by

$$\begin{aligned} Q_\alpha\left(Y|X\right) &= X'(\gamma + z_\alpha \eta) \quad \text{and} \\ \mathrm{ES}_\alpha\left(Y|X\right) &= X'(\gamma + \xi_\alpha \eta), \end{aligned} \tag{4.2}$$

where z_α and ξ_α are the quantile and ES of the distribution $F(0,1)$, which means that $\theta_0^q = \gamma + z_\alpha \eta$ and $\theta_0^e = \gamma + \xi_\alpha \eta$. Furthermore, the distributions of the quantile- and ES-residuals is given by

$$\begin{aligned} u^q|X &\sim F\left(-z_\alpha(X'\eta),\ (X'\eta)^2\right), \quad \text{and} \\ u^e|X &\sim F\left(-\xi_\alpha(X'\eta),\ (X'\eta)^2\right). \end{aligned} \tag{4.3}$$

This general specification of our data generating process (DGP) allows for a variety of setups with different properties by choosing γ and η accordingly. For the simulation study, we choose γ and η in (4.1) such that we get the following three specifications for the DGPs,

$$\begin{aligned} &\text{DGP-(1)}: \quad \boldsymbol{X} = (1,\ X_2), \qquad X_2 \sim \chi_1^2 \\ &\qquad\qquad Y|\boldsymbol{X} \sim \mathcal{N}\left(-X_2,\ 1\right) \end{aligned} \tag{4.4}$$

$$\begin{aligned} &\text{DGP-(2)}: \quad \boldsymbol{X} = (1,\ X_2), \qquad X_2 \sim \chi_1^2 \\ &\qquad\qquad Y|\boldsymbol{X} \sim \mathcal{N}\left(-X_2,\ (1 + 0.1X_2)^2\right) \end{aligned} \tag{4.5}$$

$$\begin{aligned} &\text{DGP-(3)}: \quad \boldsymbol{X} = (1,\ X_2,\ X_3), \qquad X_2,\ X_3 \sim U[0,1] \quad \text{with} \quad \mathrm{corr}(X_2, X_3) = 0.5 \\ &\qquad\qquad Y|\boldsymbol{X} \sim t_5\left(X_2 - X_3,\ \left(0.5 + 0.1X_2 + 0.1X_3\right)^2\right) \end{aligned} \tag{4.6}$$

The most important difference between the first two univariate processes is the distribution of the model residuals: for DGP-(1), the model residuals are independent of X, whereas under DGP-(2) they depend on the regressors X as $\eta_2 \neq 0$. This distinction heavily influences the performance and properties of the different covariances matrix estimators described in Section 3.2. The third DGP depends on two correlated explanatory variables in order to also assess the performance of a regression framework with multiple covariates. Furthermore, the distribution of Y given X is leptokurtic, i.e. it has fatter tails than the normal distribution.

We simulate all three processes 25,000 times with varying sample sizes of $n = 250$, 500, 1000, 2000 and 5000 observations. For each replication, and for each of the sample sizes, we regress the simulated Y's on the covariates X including an intercept term using our joint regression method for the probability level $\alpha = 2.5\%$.

4.2. Comparing the Specification Functions

We start the discussion of the simulation results by numerically investigating the performance of the M-estimator based on different choices for the specification function $\mathcal{G}_2$ used in the loss function in (2.6). Section 2.4 already discusses this choice in terms of the resulting theoretical properties of our estimator. For the empirical comparison study of the M-estimator based on different loss functions, we use the following five choices for the specification functions:

$$G_1(x) = 0, \qquad \mathcal{G}_2(z) = -\frac{1}{z}, \qquad G_2(z) = \frac{1}{z^2}, \tag{4.7}$$

$$G_1(x) = 0, \qquad \mathcal{G}_2(z) = -\log(-z), \qquad G_2(z) = -\frac{1}{z} \tag{4.8}$$

$$G_1(x) = 0, \qquad \mathcal{G}_2(z) = -\sqrt{-z}, \qquad G_2(z) = -\frac{1}{2\sqrt{-z}}, \tag{4.9}$$

$$G_1(x) = 0, \qquad \mathcal{G}_2(z) = \log\big(1 + \exp(z)\big), \qquad G_2(z) = \frac{\exp(z)}{1 + \exp(z)}, \tag{4.10}$$

$$G_1(x) = 0, \qquad \mathcal{G}_2(z) = \exp(z), \qquad G_2(z) = \exp(z). \tag{4.11}$$

The first three choices (4.7) - (4.9) result in positively homogeneous loss functions, whereas (4.10) and (4.11) are examples for a bounded and an unbounded choice for G_2 respectively. We fix the function G_1 to be constant zero since other choices do not result in a better numerical performance of the estimators and this choice is consistent with the homogeneity result of Nolde and Ziegel (2017).

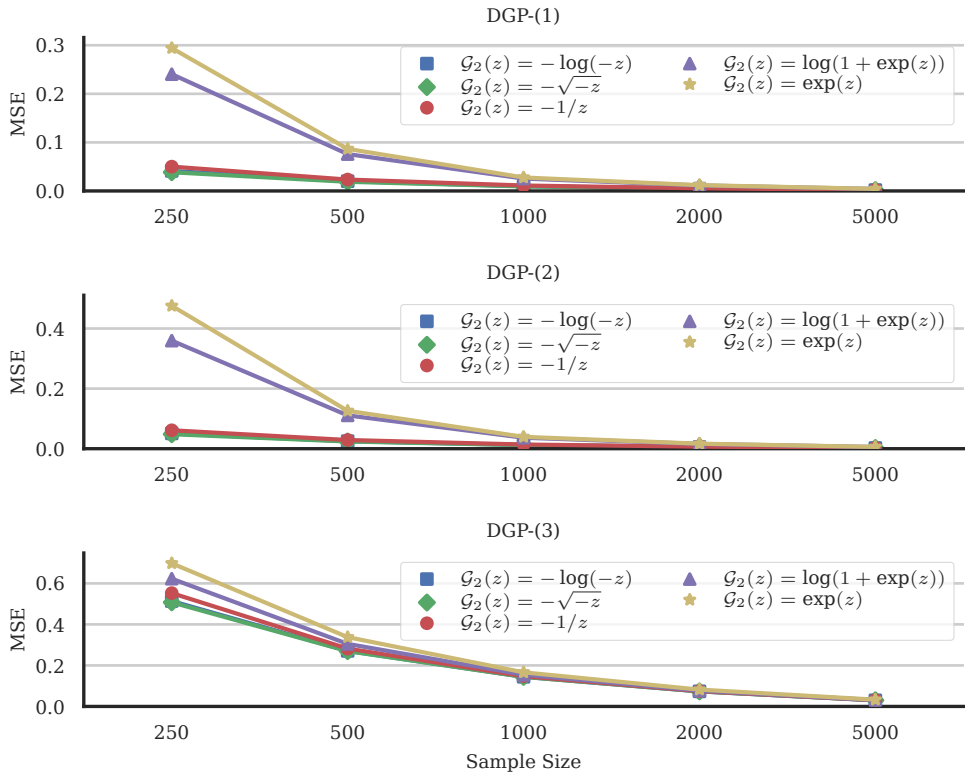

Figure 1: Sum of the mean squared errors of the parameters estimates using different choices for the specification functions in the loss function.

Figure 1 presents the sum (over the different regression parameters) of the mean squared errors (MSE) of the regression parameters for the three GDPs described above, different sample sizes and for the five choices of the specification functions. As implied by the asymptotic theory, we obtain consistent parameter

estimates for all five choices of the specification functions as their MSE converges to zero for all three DGPs. However, they differ substantially with respect to their small sample properties. The three positively homogeneous specifications result in the most accurate estimates, whereas the choices $\mathcal{G}_2(z) = -\sqrt{-z}$ and $\mathcal{G}_2(z) = -\log(-z)$ tend to perform slightly better than the choice $\mathcal{G}_2(z) = -\frac{1}{z}$. Furthermore, the bounded choice $\mathcal{G}_2(z) = \log\big(1 + \exp(z)\big)$ still performs better than the unbounded exponential function. Figure 6 in Appendix D furthermore presents the relative bias of the individual parameter estimates, which are in line with the MSE results.

We also discuss the asymptotic efficiency of the M-estimator based on the different choices for the specification functions as the asymptotic covariance, given in Theorem 2.4 depends on these choices. As we cannot compute the expectations (with respect to X) in the true asymptotic covariance matrix analytically, we estimate it through Monte-Carlo integration with a sample size of 10^9 using the formulas in Theorem 2.4 and by using the true density and conditional variance function, as both are analytically known for the underlying DGPs. Table 1 reports the diagonal elements and the Frobenius norm of the true asymptotic covariance matrix of the parameter estimates for the different choices for the specification functions for the three DGPs. For comparison, we also report the same quantities for quantile regression parameter estimates. We see that on average, the specification functions $\mathcal{G}_2(z) = -\log(-z)$ and $\mathcal{G}_2(z) = -\sqrt{-z}$ exhibit the smallest entries for the asymptotic variances, closely followed by the third choice for a positively homogeneous loss function, $\mathcal{G}_2(z) = -1/z$. The other non-homogeneous choices lead to considerably larger asymptotic variances for all considered parameters and DGPs. Furthermore, by comparing the asymptotic efficiency of our estimation approach (of the quantile-specific parameters) to quantile regression, we see that we roughly obtain the same asymptotic efficiency.

Table 1: Diagonal entries and Frobenius norms of the true covariance of the stabilizing transformation for all DGPs and all choices of the $\mathcal{G}_2$ function. For comparison, the table also reports the values for quantile regression.

DGP-(1)	θ_0^q	θ_1^q	θ_0^e	θ_1^e	Norm_F
$\mathcal{G}_2(z) = -\log(-z)$	10.92	3.79	16.33	6.28	7.90
$\mathcal{G}_2(z) = -\sqrt{-z}$	10.77	3.62	15.97	5.79	7.67
$\mathcal{G}_2(z) = -1/z$	11.38	4.38	17.13	7.59	8.52
$\mathcal{G}_2(z) = \log(1 + \exp(z))$	15.36	18.33	21.82	25.43	16.58
$\mathcal{G}_2(z) = \exp(z)$	15.51	18.96	22.25	27.19	17.25
Quantile Regression	10.70	3.57	–	–	–
DGP-(2)	θ_0^q	θ_1^q	θ_0^e	θ_1^e	Norm_F
$\mathcal{G}_2(z) = -\log(-z)$	12.36	7.13	17.83	10.38	9.61
$\mathcal{G}_2(z) = -\sqrt{-z}$	12.34	7.11	17.69	10.17	9.53
$\mathcal{G}_2(z) = -1/z$	12.71	7.79	18.47	11.65	10.20
$\mathcal{G}_2(z) = \log(1 + \exp(z))$	17.08	31.13	23.75	39.06	22.95
$\mathcal{G}_2(z) = \exp(z)$	17.24	32.16	24.19	41.59	23.93
Quantile Regression	12.55	7.38	–	–	–
DGP-(3)	θ_0^q	θ_1^q	θ_2^q	θ_0^e	θ_1^e	θ_2^e	Norm_F
$\mathcal{G}_2(z) = -\log(-z)$	24.65	90.92	90.95	83.93	307.70	308.07	115.70
$\mathcal{G}_2(z) = -\sqrt{-z}$	24.66	90.93	90.95	83.99	307.75	308.04	115.70
$\mathcal{G}_2(z) = -1/z$	24.64	90.94	90.99	83.81	307.74	308.26	115.75
$\mathcal{G}_2(z) = \log(1 + \exp(z))$	24.99	98.77	98.85	83.52	320.82	322.29	121.63
$\mathcal{G}_2(z) = \exp(z)$	25.25	102.95	102.96	84.61	344.21	345.57	130.29
Quantile Regression	24.67	90.96	90.96	–	–	–	–

In Figure 2 we provide the average computation time for fitting the regression model depending on the choice of the $\mathcal{G}_2$-function. We can see that the three positively homogeneous loss functions also exhibit the smallest computation times.

Due to their superior performance in terms of asymptotic efficiency, numerical stability and computation times, we suggest using a specification function which results in a positively homogeneous loss function

for the implementation of the M-estimator of the regression parameters. This is also in line with the theoretical findings of Section 2.4.

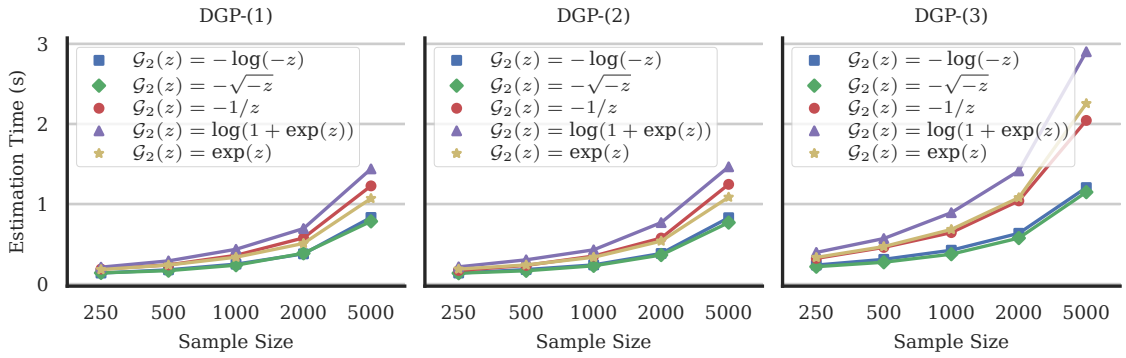

Figure 2: Average estimation time for all five $\mathcal{G}_2$-functions and the three DGPs.

4.3. Comparing the Variance-Covariance Estimators

In this section, we compare the empirical performance of the asymptotic covariance estimators discussed in Section 3.2. For the comparison of their precision, we compute the differences between the average of the estimated asymptotic covariances and the empirical covariance of the estimated parameters over the 25,000 Monte-Carlo repetitions. In order to present the results in a parsimonious way, we report the Frobenius norm of the lower triangular part of this matrix difference (such that the covariance terms enter only once). On the one hand, implied by the asymptotic theory, this quantity should converge to zero for large sample sizes. On the other hand, we also evaluate the performance of the estimated asymptotic covariances in finite samples, which is particularly important in possible applications involving confidence bounds and testing.

The results for the three DGPs and the three homogeneous choices for the loss function are provided in Figure 3. Given the poor performance in the preceding section, the results for the choices $\mathcal{G}_2(z) = \log(1 + \exp(z))$ and $\mathcal{G}_2(z) = \exp(z)$ are not provided here but are available upon request. Each of the plots presents the results for the three covariance estimators (*iid/nid*, *nid/scl-N* and *nid/scl-t*) and for different sample sizes.

For the univariate and homoskedastic DGP-(1) shown in the first row of Figure 3, we find that the norms of all three estimators and all $\mathcal{G}_2$-functions converge to zero, i.e. the average estimated covariances and the empirical covariance coincide for large sample sizes. However, for this DGP the *iid/nid* estimator converges faster than the scaling based estimators. As there is no dependency between u^q and X, the *iid/nid* estimator is able to capture the full dynamics of the underlying data and consequently, the other two scaling based estimators suffer from estimating additional (and redundant) parameters. Consequently, in this simple case the *iid/nid* estimator performs best in finite samples.

The second row of Figure 3 presents the results for DGP-(2), which is of heteroskedastic nature since in contrast to DGP-(1), the distribution of the model residuals u^q depends on X. Consequently, the estimates of the *iid/nid* approach do not converge to the empirical covariance for large samples as this estimator fails to capture the underlying dynamics of the data. In contrast, the two scaling approaches are able to model these dynamics and consequently, the norms of the matrix differences converge to zero for both estimation methods. However, for the smallest considered sample size, the *iid/nid* estimator outperforms these scaling variants. This can be explained by the additional estimation noise stemming from estimation of the additional parameters which is particularly prominent in small samples.

The last row of Figure 3 presents the results for DGP-(3). We find that the *nid/scl-N* estimator does not converge to the empirical covariance at all as the underlying distributional assumption used for the

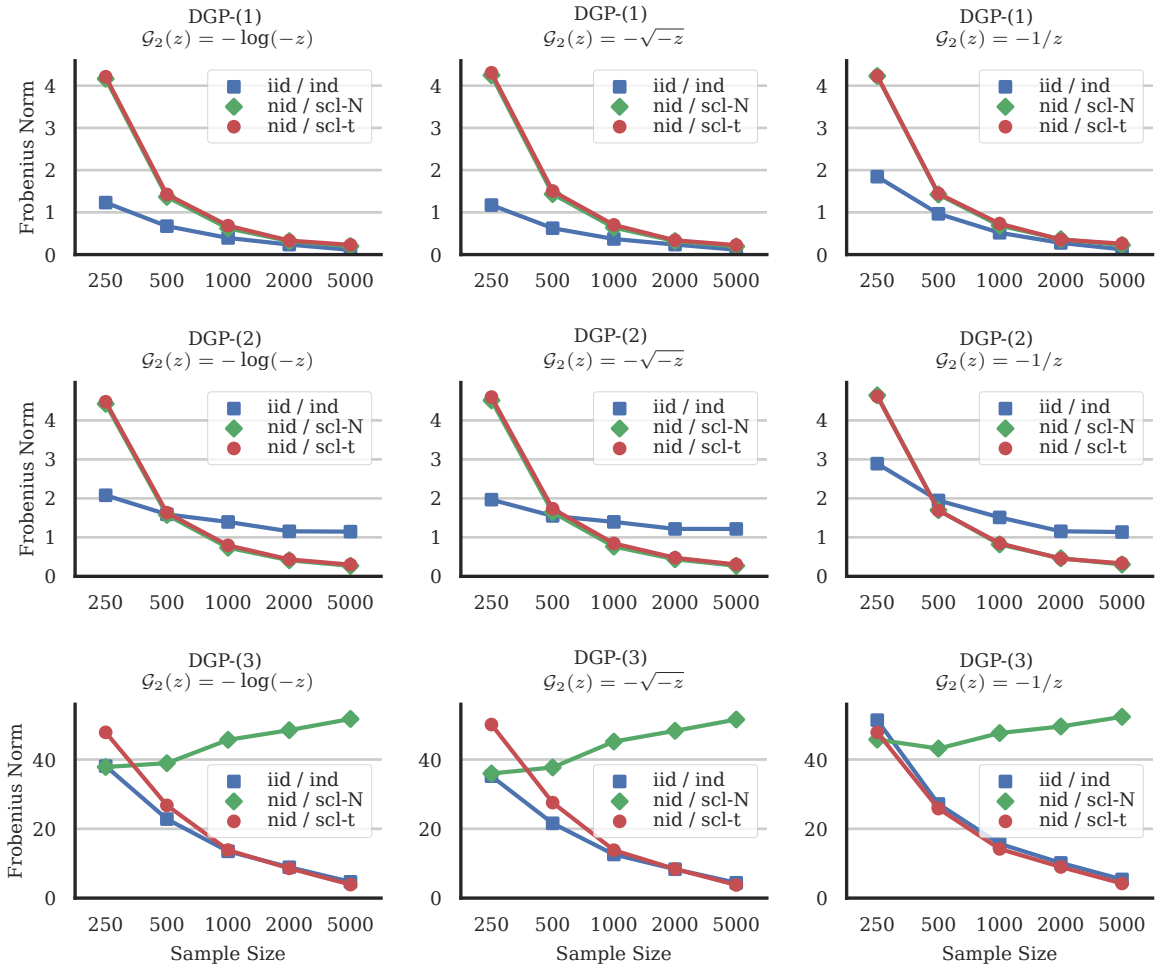

Figure 3: Frobenius Norm of the element-wise differences between the empirical and the estimated covariance matrices for three different covariance estimators.

scaling relation does not capture the fat tails in the data. In contrast, the *nid/scl-t* approach estimates the degrees of freedom for the presumed t-distribution and is consequently able to account for this property of the data. In the scenario, the *iid/nid* estimator performs well, which can be explained by the relatively mild heteroskedastic nature of this DGP. Since the covariates follow a uniform distribution, we do not have to deal with very large conditional variances as compared to DGP-(2), where X follows a fat-tailed χ^2_1-distribution.

In summary, the performance of the *nid/scl-t* estimator is the most stable among all considered choices. However, for data that does not exhibit heteroskedasticity or in the case of very small sample sizes, one can also use the *iid/nid* estimator. These covariance estimators are implemented in the provided R package.

4.4. Estimating the Sample Quantile and Expected Shortfall

Besides modeling conditional regression equations for the pair quantile and ES based on a set of covariates, our regression framework also nests joint estimation of the sample quantile and ES as already discussed in Remark 2.7. Following the results from Section 4.2, we choose the specification functions $G_1(z) = 0$ and $\mathcal{G}_2(z) = -\log(-z)$ for the joint M-estimation.

We compare the accuracy of the ES-part from our joint estimation approach with two alternative estimators for the sample ES from Brazauskas et al. (2008) and Chen (2008), given in (2.32) and (2.34). For that purpose, we simulate data from the standard Normal and from a Student-t distribution with five

degrees of freedom for a variety of sample sizes and estimate the 2.5%-ES and the associated asymptotic variances of the respective estimators.

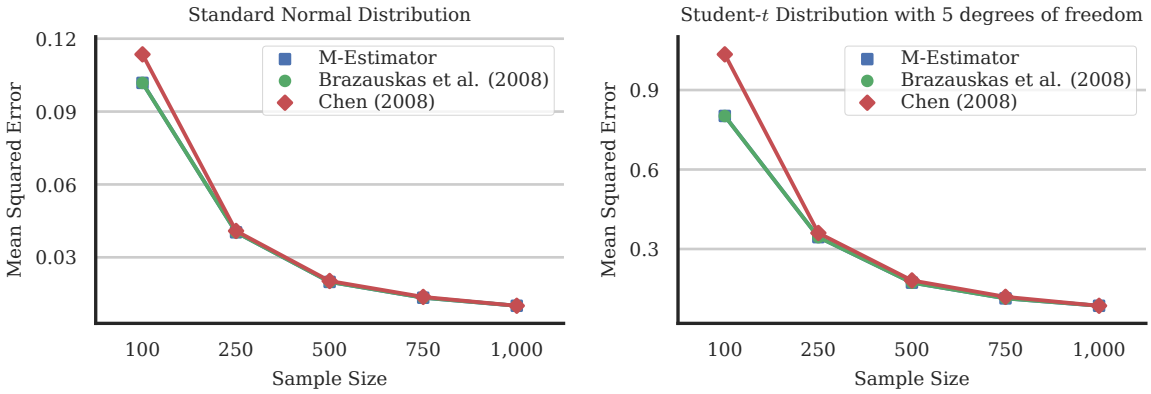

Figure 4: Mean squared errors of the different ES estimation approaches and two different distributions.

Figure 4 presents the MSE for the three ES-estimators for the Normal and the t_5 distribution. As all three estimators are consistent, we observe that for an increasing sample size the MSE converges to zero for all three approaches. Also for small sample sizes, the performance of all three estimators is similar.

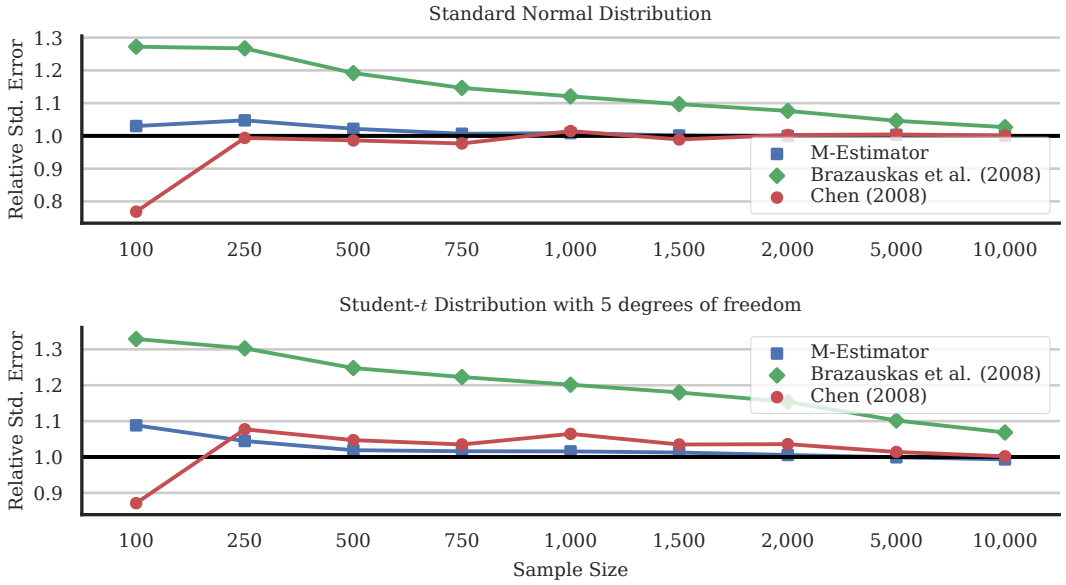

Figure 5: Relative standard errors for the different ES estimation approaches and two different distributions.

We also compare these three different approaches in terms of the accuracy of estimating the asymptotic variance of the ES estimates. For that, we use the estimators for the asymptotic variance proposed by Brazauskas et al. (2008) and Chen (2008) and for our M-estimator, we report the results for the *iid*/*ind* approach.

In Figure 5 we compare the relative standard errors of the ES estimates, which are defined as the average of the estimated asymptotic standard deviations divided by the empirical standard deviation of the estimates. We presents these results for sample sizes up to 10,000 observations due to the very slow convergence of the estimator of Brazauskas et al. (2008). Even though all three estimators converge to a relative standard error of one, the estimator of Brazauskas et al. (2008) substantially overestimates the variance of the ES estimator for realistic sample sizes. In comparison, our approach and the method of Chen (2008) are relatively accurate in small samples.

5. Conclusion

In this paper, we introduce a joint regression technique for the quantile (VaR) and ES. This regression approach relies on the joint loss function introduced by Fissler and Ziegel (2016), which permits the joint elicitation of the quantile and ES. We introduce an M- and Z-estimator for the parameters of the joint regression model. Given a set of standard regularity conditions, we show consistency and asymptotic normality for both estimators. The underlying loss function, estimating equations and the asymptotic covariance matrix of the estimators depend on two specification functions, which we investigate in terms of asymptotic efficiency, numerical performance and computation times. While the asymptotic distribution coincides for both estimators, we find the Z-estimator to be numerically unstable and consequently focus on the M-estimator in the numerical illustration. In an extensive simulation study, we verify the asymptotic distribution and investigate the small sample properties of the M-estimator. We furthermore evaluate the choice of the specification functions numerically and find that choices resulting in positively homogeneous loss functions lead to better estimates in terms of efficiency, numerical accuracy and computation times.

This joint regression technique allows for various financial applications for the risk measures VaR and ES. Bayer and Dimitriadis (2017b) use this regression to develop an ES backtest which is particularly relevant in terms of the recent introduction of ES into the Basel regulatory framework. The proposed regression can further be used to model ES by generalizing existing applications of quantile regression on VaR (e.g. Engle and Manganelli, 2004; Koenker and Xiao, 2006; Halbleib and Pohlmeier, 2012; Komunjer, 2013; Žikeš and Baruník, 2016).

Acknowledgements

We would like to thank Patrick Schmidt, Lyudmila Grigoryeva, Tobias Fissler, Frederic Menninger, Winfried Pohlmeier, Roxana Halbleib, Phillip Heiler and the participants of the Stochastics Colloquium on 11/30/2016 at the University of Konstanz for fruitful discussions and suggestions which inspired some of the results of this paper. Financial support by the Heidelberg Academy of Sciences and Humanities (HAW) within the project "Analyzing, Measuring and Forecasting Financial Risks by means of High-Frequency Data", by the German Research Foundation (DFG) within the research group "Robust Risk Measures in Real Time Settings" and general support by the Graduate School of Decision Sciences (University of Konstanz) is gratefully acknowledged. The computation in this work was performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

Appendix A General Moment Conditions

($\mathcal{M}$-1) For Theorem 2.2, we assume that the following moments are finite for some constant $d_0 > 0$:

- $\mathbb{E}\left[||X||^2 \sup_{U_{d_0}(\theta_0)} \left|G_1'(X'\theta^q)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{U_{d_0}(\theta_0)} \left|G_1''(X'\theta^q)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{U_{d_0}(\theta_0)} \left|G_2(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X||^3 \sup_{U_{d_0}(\theta_0)} \left|G_2'(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X||^3 \sup_{U_{d_0}(\theta_0)} \left|G_2''(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{U_{d_0}(\theta_0)} \left|G_2'(X'\theta^e)\right| \mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^2 \sup_{U_{d_0}(\theta_0)} \left|G_2''(X'\theta^e)\right| \mathbb{E}\left[|Y|\middle|X\right]\right]$

($\mathcal{M}$-2) For Theorem 2.3, we assume that the following moments are finite:

- $\mathbb{E}\left[||X||^2\right]$
- $\mathbb{E}\left[\sup_{\theta\in\Theta}\left|G_1(X'\theta^q)\right|\right]$
- $\mathbb{E}\left[\left|G_1(Y)\right|\right]$
- $\mathbb{E}\left[\left|a(Y)\right|\right]$
- $\mathbb{E}\left[||X|| \sup_{\theta\in\Theta}\left|G_2(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[\sup_{\theta\in\Theta}\left|G_2(X'\theta^e)\right| \mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\mathcal{G}_2(X'\theta^e)\right|\right]$

($\mathcal{M}$-3) For Theorem 2.4, we assume that the following moments are finite for some constant $d_0 > 0$ and for all $\theta \in \bar{U}_{d_0}(\theta_0)$:

- $\mathbb{E}\left[||X||^3 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_1'(X'\tau^q)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_1''(X'\tilde{\tau}^q)\right)\right]$
- $\mathbb{E}\left[||X||^3 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_1'(X'\tau^q)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_1''(X'\tilde{\tau}^q)\right)\right]$
- $\mathbb{E}\left[||X||^3 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_1'(X'\tau^q)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tilde{\tau}^e)\right)\right]$
- $\mathbb{E}\left[||X||^3 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2(X'\tau^e)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_1''(X'\tilde{\tau}^q)\right)\right]$
- $\mathbb{E}\left[||X||^3 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2(X'\tau^e)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tilde{\tau}^e)\right)\right]$
- $\mathbb{E}\left[||X||^3 \sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_1'^2(X'\tau^q)\right]$
- $\mathbb{E}\left[||X||^3 \sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2^2(X'\tau^e)\right]$
- $\mathbb{E}\left[||X||^3 \sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_1'(X'\tau^q)G_2(X'\tau^e)\right]$
- $\mathbb{E}\left[||X||^5 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tau^e)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2''(X'\tilde{\tau}^e)\right)\right]$
- $\mathbb{E}\left[||X||^5 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tau^e)\right)^2\right]$
- $\mathbb{E}\left[||X||^4 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tau^e)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2''(X'\tilde{\tau}^e)\right)\mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^3 G_2'(X'\theta^e)\left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tau^e)\right)\mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^3 G_2'(X'\theta^e)\left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2''(X'\tau^e)\right)\mathbb{E}\left[Y^2\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^3 \left(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2'(X'\tau^e)\right)\left(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2''(X'\tilde{\tau}^e)\right)\mathbb{E}\left[Y^2\middle|X\right]\right]$

($\mathcal{M}$-4) For Theorem 2.5, we assume that the following moments are finite for some constant $d_0 > 0$:

- $\mathbb{E}\left[\left|G_1(Y)\right|\right]$
- $\mathbb{E}\left[\left|a(Y)\right|\right]$
- $\mathbb{E}\left[||X|| \sup_{\theta \in U_{d_0}(\theta_0)} \left|G_1'(X'\theta^q)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{\theta \in \bar{U}_{d_0}(\theta_0)} \left|G_1'^2(X'\theta^q)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{\theta \in \bar{U}_{d_0}(\theta_0)} \left|G_1'(X'\theta^q) G_2(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X|| \sup_{\theta \in U_{d_0}(\theta_0)} \left|G_2(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{\theta \in U_{d_0}(\theta_0)} \left|G_2'(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X||^2 \sup_{\theta \in \bar{U}_{d_0}(\theta_0)} \left|G_2^2(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X||^4 \sup_{\theta \in \bar{U}_{d_0}(\theta_0)} \left|G_2'^2(X'\theta^e)\right|\right]$
- $\mathbb{E}\left[||X|| \sup_{\theta \in U_{d_0}(\theta_0)} \left|G_2'(X'\theta^e)\right| \mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^3 \sup_{\theta \in \bar{U}_{d_0}(\theta_0)} \left|G_2'^2(X'\theta^e)\right| \mathbb{E}\left[|Y|\middle|X\right]\right]$
- $\mathbb{E}\left[||X||^2 \sup_{\theta \in \bar{U}_{d_0}(\theta_0)} \left|G_2'^2(X'\theta^e)\right| \mathbb{E}[Y^2|X]\right]$

Appendix B Proofs

For the subsequent proofs we use the following notation. Let $Y \in \mathbb{R}$ and $X \in \mathcal{X} \subseteq \mathbb{R}^k$ be random variables on some probability space equipped with the Borel σ-field. Let $\theta = (\theta^{q\prime}, \theta^{e\prime})' \in \Theta \subseteq \mathbb{R}^{2k}$ be a joint parameter vector, whereas $\theta_0 = (\theta_0^{q\prime}, \theta_0^{e\prime})' \in \text{int}(\Theta)$ denote the true parameter vector of our joint regression. Henceforth, for a vector $v \in \mathbb{R}^k$, $||v||$ denotes the maximum norm $||v||_{\max} = \max_j |v_j|$ and for a matrix A, $||A||$ denotes the row-sum matrix norm which is induced by the maximum norm for vectors. For convenience of the supremum notation, for all $\theta \in \text{int}(\Theta)$ and for some $d > 0$, we define the open neighborhood $U_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| < d\}$ and its closure $\bar{U}_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| \leq d\}$.

B.1 Proof of Theorem 2.2

Lemma B.1. Let

$$u(Y, X, \theta, d) = \sup_{\tau \in \bar{U}_d(\theta)} \left|\left|\psi(Y, X, \tau) - \psi(Y, X, \theta)\right|\right| \tag{B.1}$$

and assume that the regularity conditions in Assumption 2.1 and the Moment Conditions ($\mathcal{M}$-1) in Appendix A hold. Then, there are strictly positive real numbers b and d_0, such that

$$\mathbb{E}\big[u(Y, X, \theta, d)\big] \leq b \cdot d \quad \text{for} \quad ||\theta - \theta_0|| + d \leq d_0, \quad \text{and} \tag{B.2}$$

for all $d \geq 0$.

Proof. Let in the following $d > 0$ and $\theta \in \Theta$ such that $||\theta - \theta_0|| + d \leq d_0$. We first notice that for some fixed $X \in \mathcal{X}$ and for all $\tau \in \bar{U}_d(\theta)$, it holds that

$$\left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \leq \mathbb{1}_{\{X'\theta_-^q \leq Y \leq X'\theta_+^q\}} \tag{B.3}$$

for all $Y \in \mathbb{R}$ and for some $\theta_-^q, \theta_+^q \in \bar{U}_d(\theta)$. Since $\bar{U}_d(\theta)$ is compact, we get that

$$\sup_{\tau \in \bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \leq \mathbb{1}_{\{X'\theta_-^q \leq Y \leq X'\theta_+^q\}} \tag{B.4}$$

for all $Y \in \mathbb{R}$ and for some values $\theta_-^q, \theta_+^q \in \bar{U}_d(\theta)$. Note that the values θ_-^q and θ_+^q depend on X and θ, however they are independent of Y. Consequently, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \middle| X\right] \leq \mathbb{E}\left[\mathbb{1}_{\{X'\theta_-^q \leq Y \leq X'\theta_+^q\}} \middle| X\right] \\
&= F_{Y|X}\big(X'\theta_+^q\big) - F_{Y|X}\big(X'\theta_-^q\big) = f_{Y|X}(X'\tilde{\theta}^q)\big(X'\theta_+^q - X'\theta_-^q\big) \\
&\leq 2||X|| \cdot \sup_{\tau \in \bar{U}_d(\theta)} f_{Y|X}(X'\tau^q) \cdot d,
\end{aligned} \tag{B.5}$$

where we apply the mean value theorem for some $\tilde{\theta}^q$ on the line between θ^q_- and θ^q_+, i.e. $\tilde{\theta}^q \in \bar{U}_d(\theta)$.

For the first component of ψ, we get that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|\psi_1(Y,X,\theta)-\psi_1(Y,X,\tau)\right\|\right] \\
&\leq \mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(G_1'(X'\theta^q)-G_1'(X'\tau^q)+\frac{G_2(X'\theta^e)-G_2(X'\tau^e)}{\alpha}\right)\right\|\right] \\
&\quad+\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(G_1'(X'\tau^q)+\frac{G_2(X'\tau^e)}{\alpha}\right)\right\|\cdot\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|\mathbb{1}_{\{Y\leq X'\theta^q\}}-\mathbb{1}_{\{Y\leq X'\tau^q\}}\right|\middle|X\right]\right].
\end{aligned} \tag{B.6}$$

The first term in (B.6) is $\mathcal{O}(d)$ since $G_1'(X'\theta^q)$ and $G_2(X'\theta^e)$ are continuously differentiable functions w.r.t θ and thus, by the mean value theorem we get that

$$\begin{aligned}
\sup_{\tau\in\bar{U}_d(\theta)}\left|G_1'(X'\theta^q)-G_1'(X'\tau^q)\right| &\leq \sup_{\tilde{\tau}\in\bar{U}_d(\theta)}\left\|XG_1''(X'\tilde{\tau}^q)\right\|\cdot\sup_{\tau\in\bar{U}_d(\theta)}\left\|\theta^q-\tau^q\right\| \\
&\leq \sup_{\tilde{\tau}\in\bar{U}_d(\theta)}\left\|XG_1''(X'\tilde{\tau}^q)\right\|\cdot d,
\end{aligned} \tag{B.7}$$

and the respective moments are finite by assumption. The same arguments hold for the function G_2. For the second term in (B.6), we apply (B.5) and thus get that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(G_1'(X'\tau^q)+\frac{G_2(X'\tau^e)}{\alpha}\right)\right\|\cdot\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|\mathbb{1}_{\{Y\leq X'\theta^q\}}-\mathbb{1}_{\{Y\leq X'\tau^q\}}\right|\middle|X\right]\right] \\
&\leq \mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(G_1'(X'\tau^q)+\frac{G_2(X'\tau^e)}{\alpha}\right)\right\|\,||X||\cdot\sup_{\tau\in\bar{U}_d(\theta)}f_{Y|X}(X'\tau^q)\right]\cdot d.
\end{aligned} \tag{B.8}$$

Since the density $f_{Y|X}(X'\theta)$ is bounded in a neighborhood of $X'\theta_0$ and the respective moments are finite by assumption, we get that this term is also $\mathcal{O}(d)$.

For the second component of ψ, we get that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|\psi_2(Y,X,\theta)-\psi_2(Y,X,\tau)\right\|\right] \\
&\leq \mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X(X'\theta^e-X'\theta^q)G_2'(X'\theta^e)-X(X'\tau^e-X'\tau^q)G_2'(X'\tau^e)\right\|\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)X'\theta^q}{\alpha}\right\|\cdot\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|\left(\mathbb{1}_{\{Y\leq X'\theta^q\}}-\mathbb{1}_{\{Y\leq X'\tau^q\}}\right)\right|\middle|X\right]\right] \\
&\quad+\mathbb{E}\left[\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|\mathbb{1}_{\{Y\leq X'\tau^q\}}\left(\frac{XG_2'(X'\theta^e)X'\theta^q}{\alpha}-\frac{XG_2'(X'\tau^e)X'\tau^q}{\alpha}\right)\right\|\middle|X\right]\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|\cdot\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|Y\left(\mathbb{1}_{\{Y\leq X'\theta^q\}}-\mathbb{1}_{\{Y\leq X'\tau^q\}}\right)\right|\middle|X\right]\right] \\
&\quad+\mathbb{E}\left[\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|\frac{Y\mathbb{1}_{\{Y\leq X'\tau^q\}}}{\alpha}\left(XG_2'(X'\theta^e)-XG_2'(X'\tau^e)\right)\right\|\middle|X\right]\right] \\
&= \text{(i)}+\text{(ii)}+\text{(iii)}+\text{(iv)}+\text{(v)}.
\end{aligned}$$

The first, third and fifth term are linearly bounded by (B.7) since the functions $(X'\theta^e-X'\theta^q)G_2'(X'\theta^e)$ and $(X'\theta^q)G_2'(X'\theta^e)$ and $G_2'(X'\theta^e)$ are continuously differentiable. For the second term, we use the

arguments from (B.5). For the fourth term, we use similar arguments as in (B.5), and get that there exist some $\theta^q_-, \theta^q_+ \in \bar{U}_d(\theta)$ and a value $\tilde{\theta}^q$ on the line between θ^q_- and θ^q_+, such that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|Y\left(\mathbb{1}_{\{Y\le X'\theta^q\}}-\mathbb{1}_{\{Y\le X'\tau^q\}}\right)\right|\middle| X\right]\right]\\
&\le \mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|\mathbb{E}\left[|Y|\,\mathbb{1}_{\{X'\theta^q_-\le Y\le X'\theta^q_+\}}\middle| X\right]\right]\\
&= \mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|\int_{X'\theta^q_-}^{X'\theta^q_+}|y|f_{Y|X}(y)\mathrm{d}y\right]\\
&\le \mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\||X'\tilde{\theta}^q|f_{Y|X}(X'\tilde{\theta}^q)\left(X'\theta^q_+-X'\theta^q_-\right)\right]\\
&\le \frac{2}{\alpha}\mathbb{E}\left[G_2'(X'\theta^e)\|X\|^2\sup_{\tau\in\bar{U}_d(\theta)}|X'\tau^q|f_{Y|X}(X'\tau^q)\right]\cdot d = \mathcal{O}(d)
\end{aligned} \tag{B.9}$$

since $f_{Y|X}$ is bounded in a neighborhood of $X'\theta_0$ and the respective moments exist by assumption. This concludes the proof of the lemma. □

Lemma B.2. Let the random variable $X \in \mathcal{X}$ with distribution $\mathbb{P}$ be such that its second moments exist and the matrix $\mathbb{E}[XX']$ is positive definite. Furthermore, let $\tilde{\Theta} \subset \mathbb{R}^k$ be a compact subspace with nonempty interior and let $g : \mathcal{X} \times \tilde{\Theta} \to \mathbb{R}$ be a strictly positive function. Then, the matrix

$$\mathbb{E}\big[(XX')g(X,\theta)\big] \tag{B.10}$$

is also positive definite.

Proof. Since $\mathbb{E}[XX']$ is positive definite, we know that for all $z \in \mathbb{R}^k$ with $z \neq 0$, it holds that $0 < z'\mathbb{E}[XX']z = \mathbb{E}[z'(XX')z] = \mathbb{E}[(X'z)^2]$ and consequently $\mathbb{P}\big(X'z \neq 0\big) > 0$. Since $\sqrt{g(X,\theta)}$ is a strictly positive scalar for all $\theta \in \tilde{\Theta}$, it also holds that $\mathbb{P}\big((X'z)\sqrt{g(X,\theta)} \neq 0\big) > 0$ and thus, for all $z \neq 0$,

$$z'\mathbb{E}\big[(XX')g(X,\theta)\big]z = \mathbb{E}\left[\left(X'z\sqrt{g(X,\theta)}\right)^2\right] > 0. \tag{B.11}$$

This positivity statement holds since $\left(X'z\sqrt{g(X,\theta)}\right)^2$ is a non-negative random variable and $\mathbb{P}\big((X'z)\sqrt{g(X,\theta)} \neq 0\big) > 0$.

This shows that the matrix $\mathbb{E}\big[(XX')g(X,\theta)\big]$ is positive definite. □

Proof of Theorem 2.2. We use Theorem 2 from Huber (1967) and show that the function $\psi(Y,X,\theta)$ as given in (2.9) satisfies its assumptions. We do not need to show Lemma 2 from Huber (1967) since the parameter space Θ is assumed to be compact and consequently every sequence $\hat{\theta}_{\psi,n}$ satisfying (2.8) ultimately stays in the compact set Θ. As the product of continuous functions and the indicator function $\mathbb{1}_{\{Y\le X'\theta^q\}}$, the ψ-function is $\mathcal{F}$-measurable, where $\mathcal{F}$ is the Borel σ-field on the space $\mathbb{R}\times\mathcal{X}$. Separability can be concluded with the help of Corollary C.3 since the process ψ is almost surely continuous in θ.

For the proof that ψ has a unique root at θ_0, let us first define the sets

$$U = \left\{\omega \in \Omega \middle| X(\omega)'\theta^q \neq X(\omega)\theta_0^q\right\}, \quad \text{and} \tag{B.12}$$

$$W = \left\{\omega \in \Omega \middle| X(\omega)'\theta^q = X(\omega)\theta_0^q\right\}, \tag{B.13}$$

such that $\Omega = W \cup U$ and $W \cap U = \emptyset$. We first show that $\mathbb{P}(U) > 0$. In order to see this, we assume the converse, i.e. let us assume that $\mathbb{P}(W) = \mathbb{P}\big(X'\theta^q = X'\theta_0^q\big) = 1$, which implies that

$$(\theta^q - \theta_0^q)'\,\mathbb{E}[XX']\,(\theta^q - \theta_0^q) = \mathbb{E}\big[\big(X'\theta^q - X'\theta_0^q\big)^2\big] = 0. \tag{B.14}$$

However, since $\theta^q \neq \theta_0^q$, this contradicts the assumption that the matrix $\mathbb{E}[XX']$ is positive definite and we can conclude that $\mathbb{P}(U) > 0$.

The quantity

$$\lambda_1(\theta) = \mathbb{E}\left[X\big(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e)\big)\frac{F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)}{\alpha}\right] \tag{B.15}$$

exists under the moment conditions ($\mathcal{M}$-1) in Appendix A and if $\theta^q = \theta_0^q$, it obviously holds that $\lambda_1(\theta) = 0$. Now, let us assume that $\theta \in \Theta$ such that $\theta^q \neq \theta_0^q$. By splitting the expectation, we get that

$$\begin{aligned}\lambda_1(\theta)'(\theta^q - \theta_0^q) = \mathbb{E}&\left[\big(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e)\big)\big(X'\theta^q - X'\theta_0^q\big)\frac{F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)}{\alpha}\mathbb{1}_{\{\omega\in W\}}\right]\\ &+ \mathbb{E}\left[\big(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e)\big)\big(X'\theta^q - X'\theta_0^q\big)\frac{F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)}{\alpha}\mathbb{1}_{\{\omega\in U\}}\right].\end{aligned}$$

The first summand is obviously zero since for all $\omega \in W$, $F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q) = 0$. Since the distribution of Y given X has strictly positive density in a neighbourhood of $X'\theta_0^q$, we get that $F_{Y|X}$ is strictly increasing in a neighbourhood of $X'\theta_0^q$ and thus

$$\big(X'\theta^q - X'\theta_0^q\big)\frac{F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)}{\alpha} > 0 \tag{B.16}$$

for all $\omega \in U$. Since furthermore $\alpha G_1'(X'\theta^q) + G_2(X'\theta^e) > 0$ for all $\theta \in \Theta$ and $\mathbb{P}(U) > 0$, we get that

$$\lambda_1(\theta)'(\theta^q - \theta_0^q) = \mathbb{E}\left[\big(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e)\big)\big(X'\theta^q - X'\theta_0^q\big)\frac{F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)}{\alpha}\mathbb{1}_{\{\omega\in U\}}\right] > 0, \tag{B.17}$$

and consequently $\lambda_1(\theta) \neq 0$. This implies that $\lambda_1(\theta) = 0$ if and only if $\theta^q = \theta_0^q$.

Furthermore,

$$\lambda_2(\theta) = \mathbb{E}\left[XG_2'(X'\theta^e)\left(X'\theta^q\frac{F_{Y|X}(X'\theta^q) - \alpha}{\alpha} + X'\theta^e - \frac{1}{\alpha}\mathbb{E}\big[Y\mathbb{1}_{\{Y\leq X'\theta^q\}}\big|X\big]\right)\right]. \tag{B.18}$$

Assuming that $\theta^q = \theta_0^q$ (from $\lambda_1(\theta) = 0$), we get that $F_{Y|X}(X'\theta^q) = F_{Y|X}(X'\theta_0^q) = \alpha$ and $\frac{1}{\alpha}\mathbb{E}\big[Y\mathbb{1}_{\{Y\leq X'\theta_0^q\}}\big|X\big] = X'\theta_0^e$. Thus, (B.18) simplifies to

$$\mathbb{E}\big[(XX')G_2'(X'\theta^e)\big]\big(\theta^e - \theta_0^e\big), \tag{B.19}$$

and by applying Lemma B.2, we get that the matrix $\mathbb{E}\big[(XX')G_2'(X'\theta^e)\big]$ is positive definite for all $\theta \in \Theta$. Consequently, $\lambda_2(\theta) = 0$ if and only if $\theta^e = \theta_0^e$ and together with the arguments for λ_1, we get that $\lambda(\theta) = 0$ if and only if $\theta = \theta_0$.

Eventually, assumption (B-2)' from Theorem 2 of Huber (1967) follows directly from Lemma B.1, which concludes this proof. □

B.2 Proof of Theorem 2.3

Proof of Theorem 2.3. For this proof, we apply Theorem 5.7 from van der Vaart (1998).

We start by showing uniform convergence in probability of the empirical mean of the objective function, i.e. $\sup_{\theta\in\Theta}\big|\frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \theta) - \mathbb{E}\big[\rho(Y, X, \theta)\big]\big| \xrightarrow{\mathbb{P}} 0$, by the help of Lemma 2.4 of Newey and

McFadden (1994). Since we have iid data, a compact parameter space Θ and $\rho(Y, X, \theta)$ is continuous for all $\theta \in \Theta$, it remains to show that there exists a dominating function $d(Y, X) \geq |\rho(Y, X, \theta)|$ for all $\theta \in \Theta$ with $\mathbb{E}\big[d(Y, X)\big] < \infty$.

For this, we split the ρ-function by defining

$$\rho(Y, X, \theta) = \rho_1(Y, X, \theta) + \rho_2(Y, X, \theta), \tag{B.20}$$

where

$$\rho_1(Y, X, \theta) = \mathbb{1}_{\{Y \leq X'\theta^q\}} \left(G_1(X'\theta^q) - G_1(Y) + \frac{1}{\alpha} G_2(X'\theta^e)(X'\theta^q - Y) \right), \tag{B.21}$$

$$\rho_2(Y, X, \theta) = G_2(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) - \mathcal{G}_2(X'\theta^e) - \alpha G_1(X'\theta^q) + a(Y). \tag{B.22}$$

We now get that

$$\begin{aligned} \big|\rho_1(Y, X, \theta)\big| &\leq \sup_{\theta \in \Theta} \left| G_1(X'\theta^q) + \frac{1}{\alpha} G_2(X'\theta^e)(X'\theta^q - Y) \right| + \big|G_1(Y)\big| && \text{(B.23)} \\ &=: d_1(Y, X), && \text{(B.24)} \end{aligned}$$

and equivalently,

$$\begin{aligned} \rho_2(Y, X, \theta) &\leq \sup_{\theta \in \Theta} \big|G_2(X'\theta^e)\big(X'\theta^e - X'\theta^q\big)\big| + \sup_{\theta \in \Theta} \big|\mathcal{G}_2(X'\theta^e)\big| + \big|\alpha G_1(Y) + a(Y)\big| && \text{(B.25)} \\ &=: d_2(Y, X). && \text{(B.26)} \end{aligned}$$

The functions $d_1(Y, X)$ and $d_2(Y, X)$ have finite expectation by the moment conditions ($\mathcal{M}$-2) in Appendix A, which eventually proofs the requirements for Lemma 2.4 of Newey and McFadden (1994) and we can conclude that

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta) - \mathbb{E}\big[\rho(Y, X, \theta)\big] \right| \xrightarrow{\mathbb{P}} 0. \tag{B.27}$$

It remains to show that $\mathbb{E}\big[\rho(Y, X, \theta)\big]$ has a unique and global minimum at $\theta = \theta_0$. For this, we assume that $\theta \in \Theta$ such that $\theta \neq \theta_0$ and we define the sets

$$U = \big\{\omega \in \Omega \big| X(\omega)'\theta^q \neq X(\omega)'\theta_0^q \quad \text{or} \quad X(\omega)'\theta^e \neq X(\omega)'\theta_0^e\big\} \quad \text{and} \tag{B.28}$$

$$W = \big\{\omega \in \Omega \big| X(\omega)'\theta^q = X(\omega)'\theta_0^q \quad \text{and} \quad X(\omega)'\theta^e = X(\omega)'\theta_0^e\big\}, \tag{B.29}$$

such that $\Omega = U \cup W$ and $U \cap W = \emptyset$.

We first show that $\mathbb{P}(U) > 0$. In order to see this, we assume the converse, i.e. we assume that $\mathbb{P}(W) = 1$, which implies that

$$(\theta^q - \theta_0^q)' \, \mathbb{E}[XX'] \, (\theta^q - \theta_0^q) = \mathbb{E}\left[\big(X'\theta^q - X'\theta_0^q\big)^2 \right] = 0, \tag{B.30}$$

since $\mathbb{P}\big(X'\theta^q = X\theta_0^q\big) = 1$, and equivalently

$$(\theta^e - \theta_0^e)'\mathbb{E}[XX'](\theta^e - \theta_0^e) = 0. \tag{B.31}$$

However, since $\theta \neq \theta_0$ and consequently either $\theta^q \neq \theta_0^q$ or $\theta^e \neq \theta_0^e$, this contradicts the assumption that the matrix $\mathbb{E}[XX']$ is positive definite and we can conclude that $\mathbb{P}(U) > 0$.

From the joint elicitability property of VaR and ES of Fissler and Ziegel (2016), Corollary 5.5 we get that for all $x \in \mathcal{X}$ such that $x'\theta \neq x'\theta_0$, it holds that

$$\mathbb{E}\big[\rho(Y, X, \theta_0)\big|X = x\big] < \mathbb{E}\big[\rho(Y, X, \theta)\big|X = x\big], \tag{B.32}$$

and thus for all $\omega \in U$,

$$\mathbb{E}\big[\rho(Y, X, \theta_0)\big|X\big](\omega) < \mathbb{E}\big[\rho(Y, X, \theta)\big|X\big](\omega), \tag{B.33}$$

since the distribution of Y given X has finite first moment and a unique α-quantile. We define the random variable

$$h(X, \theta, \theta_0) = \mathbb{E}\big[\rho(Y, X, \theta_0) - \rho(Y, X, \theta)\big|X\big], \tag{B.34}$$

and (B.33) implies that $h\big(X(\omega), \theta, \theta_0\big) < 0$ for all $\omega \in U$. Since $\mathbb{P}(U) > 0$, this implies that

$$\mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in U\}}\right] < 0. \tag{B.35}$$

Furthermore, for all $\omega \in W$, it obviously holds that $h(X(\omega), \theta, \theta_0) = 0$ and consequently,

$$\mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in W\}}\right] = 0. \tag{B.36}$$

Thus, we get that

$$\mathbb{E}\big[h(X, \theta, \theta_0)\big] = \mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in U\}}\right] + \mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in W\}}\right] < 0 \tag{B.37}$$

for all $\theta \in \Theta$ such that $\theta \neq \theta_0$, which shows that $\mathbb{E}\big[\rho(Y, X, \theta)\big]$ has a unique minimum at $\theta = \theta_0$. □

B.3 Proof of Theorem 2.4

For the proof of this theorem, we need the following two lemmas.

Lemma B.3. Assume that the regularity conditions in Assumption 2.1 and Moment Conditions ($\mathcal{M}$-3) in Appendix A hold. Then, for

$$\lambda(\theta) = \mathbb{E}\big[\psi(Y, X, \theta)\big], \tag{B.38}$$

there are strictly positive numbers a, d_0, such that

$$||\lambda(\theta)|| \geq a \cdot ||\theta - \theta_0|| \qquad \text{for} \quad ||\theta - \theta_0|| \leq d_0. \tag{B.39}$$

Proof. Let $d_0 > 0$ and let $||\theta - \theta_0|| \leq d_0$. Then, applying the mean value theorem, we get that

$$\lambda_1(\theta) = \frac{1}{\alpha}\mathbb{E}\left[(XX')\big(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e)\big) f_{Y|X}(X'\tilde{\theta}^q)\right](\theta^q - \theta_0^q) \tag{B.40}$$

for some $\tilde{\theta}^q$ on the line between θ^q and θ_0^q. Similarly, for the second component we get that

$$\begin{aligned}\lambda_2(\theta) = \mathbb{E}\left[X\frac{G_2'(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha}\big[X'(\theta^q - \theta_0^q)\big]\big[X'(\tilde{\theta}^q - \theta^q)\big]\right] \\ + \mathbb{E}\big[(XX')G_2'(X'\theta^e)\big](\theta^e - \theta_0^e),\end{aligned} \tag{B.41}$$

where $\tilde{\theta}^q$ lies on the line between θ^q and θ_0^q.

We first assume that $||\theta - \theta_0|| = ||\theta^q - \theta_0^q||$, i.e. $||\theta^q - \theta_0^q|| \geq ||\theta^e - \theta_0^e||$. Since the matrix

$$A(\theta) := \mathbb{E}\left[(XX')\frac{\big(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e)\big)}{\alpha} f_{Y|X}(X'\tilde{\theta}^q)\right] \tag{B.42}$$

exists and has full rank for all $\theta \in \Theta$ by Lemma B.2 and is obviously symmetric, A has strictly positive real Eigenvalues $\gamma_1(\theta), \dots, \gamma_k(\theta)$ with minimum $\gamma_{(1)}(\theta)$ and we thus get that[2]

$$||\lambda(\theta)|| \geq ||\lambda_1(\theta)|| = ||A(\theta)(\theta^q - \theta_0^q)|| \geq \gamma_{(1)}(\theta) \cdot ||\theta^q - \theta_0^q|| \tag{B.43}$$

$$\geq \left(\inf_{||\theta-\theta_0|| \leq d_0} \gamma_{(1)}(\theta) \right) \cdot ||\theta^q - \theta_0^q|| = c_1 \, ||\theta - \theta_0||. \tag{B.44}$$

Since $||\theta - \theta_0|| \leq d_0$ is a compact set and the function $\theta \mapsto \inf_{||\theta-\theta_0|| \leq d_0} \gamma_{(1)}(\theta)$, where $\gamma_{(1)}(\theta)$ is the smallest Eigenvalue of the matrix $A(\theta)$, is continuous[3], we get that the infimum coincides with the minimum and thus, the constant $c_1 := \inf_{||\theta-\theta_0|| \leq d_0} \gamma_{(1)}(\theta)$ is strictly positive and does not depend on θ.

Now, we assume that $||\theta - \theta_0|| = ||\theta^e - \theta_0^e|| \leq d_0$, i.e. $||\theta^e - \theta_0^e|| \geq ||\theta^q - \theta_0^q||$. For the first term of $\lambda_2(\theta)$, given in (B.41), we define the vector

$$b(\theta) := \mathbb{E}\left[X \frac{G_2'(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha} \left[X'(\theta^q - \theta_0^q)\right] \left[X'\tilde{\theta}^q - X'\theta^q)\right] \right], \tag{B.45}$$

and for its l-th component, we get that

$$\begin{aligned} |b_l(\theta)| &= \left| \sum_{i,j} (\theta_i^q - \theta_{0i}^q)(\tilde{\theta}_j^q - \theta_j^q) \mathbb{E}\left[X_i X_j X_l \frac{G_2'(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha} \right] \right| \\ &\leq \sum_{i,j} \mathbb{E}\left[\left| X_i X_j X_l \frac{G_2'(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha} \right| \right] \cdot |\theta_i^q - \theta_{0i}^q| \cdot |\tilde{\theta}_j^q - \theta_j^q| \\ &\leq c_2 \sum_{i,j} |\theta_i^q - \theta_{0i}^q| \cdot |\tilde{\theta}_j^q - \theta_j^q| \\ &\leq c_2 k^2 ||\theta - \theta_0||^2, \end{aligned} \tag{B.46}$$

for all $l = 1, \dots, k$, which implies that

$$||b(\theta)|| \leq c_3 ||\theta - \theta_0||^2, \tag{B.47}$$

for some $c_3 > 0$. For $D(\theta) := \mathbb{E}\left[(XX')G_2'(X'\theta^e)\right]$, it holds that $||D(\theta)(\theta^e - \theta_0^e)|| \geq c_4 ||\theta^e - \theta_0^e|| = c_4 ||\theta - \theta_0||$ for $c_4 > 0$ by the same arguments as in (B.43). From (B.46), we can choose d_0 small enough such that

$$2||b(\theta)|| \leq 2c_3 ||\theta - \theta_0||^2 \leq c_4 ||\theta - \theta_0|| \leq ||D(\theta)(\theta^e - \theta_0^e)||. \tag{B.48}$$

Furthermore, by the submultiplicativity of the matrix norm, we also get that $||D(\theta)(\theta^e - \theta_0^e)|| \leq ||D(\theta)|| \cdot ||\theta^e - \theta_0^e|| = c_5 ||\theta^e - \theta_0^e||$ and by the inverse triangle inequality, we get that

$$||\lambda(\theta)|| \geq ||\lambda_2(\theta)|| = \left|\left| D(\theta)(\theta^e - \theta_0^e) + b(\theta) \right|\right| \geq \left| ||D(\theta)(\theta^e - \theta_0^e)|| - ||b(\theta)|| \right|. \tag{B.49}$$

From (B.48), we can choose d_0 small enough such that $||D(\theta^e - \theta_0^e)|| > 2||b||$ and thus

$$\left| ||D(\theta^e - \theta_0^e)|| - ||b|| \right| = ||D(\theta^e - \theta_0^e)|| - ||b|| \geq \frac{1}{2} ||D(\theta^e - \theta_0^e)|| \tag{B.50}$$

$$\geq \frac{c_4}{2} ||\theta^e - \theta_0^e|| = \frac{c_4}{2} ||\theta - \theta_0||. \tag{B.51}$$

□

[2] For a symmetric matrix A with full rank, we can find an orthogonal basis of Eigenvectors $\{v_1, \dots, v_k\}$ with corresponding nonzero Eigenvalues $\{\gamma_1(\theta), \dots, \gamma_k(\theta)\}$ such that $x = \sum b_j v_j$ with $b_j \in \mathbb{R}$. Then, $||Ax|| = ||A \sum b_j v_j|| = ||\sum b_j A v_j|| = ||\sum b_j \gamma_j v_j|| \geq \min |\gamma_j| \cdot ||\sum b_j v_j|| = \min |\gamma_j| \cdot ||x||$.

[3] This follows since the entries of the matrix $A(\theta)$ are continuous in θ as the expectation of a continuous function which is dominated by an integrable function is again a continuous by the dominated convergence theorem. Furthermore, the Eigenvalues of a matrix are the solution of the characteristic polynomial, which has continuous coefficients since our matrix entries are continuous in θ. Eventually, since the roots of any polynomial with continuous coefficients are again continuous, we can conclude that the Eigenvalues of $A(\theta)$ are continuous in θ.

Lemma B.4. Let

$$u(Y, X, \theta, d) = \sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi(Y, X, \tau) - \psi(Y, X, \theta) \right\|. \tag{B.52}$$

and assume that the regularity conditions in Assumption 2.1 and Moment Conditions ($\mathcal{M}$-3) in Appendix A hold. Then, there are strictly positive numbers c and d_0, such that

$$\mathbb{E}\left[u(Y, X, \theta, d)^2 \right] \leq c \cdot d \quad \text{for} \quad ||\theta - \theta_0|| + d \leq d_0, \tag{B.53}$$

for all $d \geq 0$.

Proof. Let in the following $d > 0$ and $\theta \in \Theta$ such that $||\theta - \theta_0|| + d \leq d_0$. It holds that

$$\left(\sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi(Y, X, \tau) - \psi(Y, X, \theta) \right\| \right)^2 = \sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi(Y, X, \tau) - \psi(Y, X, \theta) \right\|^2 \tag{B.54}$$

and we consequently show that

$$\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi_j(Y, X, \tau) - \psi_j(Y, X, \theta) \right\|^2 \right] = \mathcal{O}(d) \tag{B.55}$$

for both components $j = 1, 2$ and for some $d > 0$ small enough.

For the first squared component, we get that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi_1(Y, X, \tau) - \psi_1(Y, X, \theta) \right\|^2 \right] \\
\leq\ & \max\left(\left| \frac{1-\alpha}{\alpha} \right|^2, 1 \right) \cdot \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| X\left(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e) - \alpha G_1'(X'\tau^q) - G_2(X'\tau^e) \right) \right\|^2 \right] \\
&+ \frac{2}{\alpha^2} \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| X\left(\alpha G_1'(X'\tau^q) + G_2(X'\tau^e) \right) \right\|^2 ||X|| \sup_{\tau \in \bar{U}_d(\theta)} f_{Y|X}(X'\tau^q) \right] \cdot d \\
&+ \frac{2}{\alpha^2} \max\left(1-\alpha, \alpha\right) \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| X\left(\alpha G_1'(X'\theta^q) + G_2(X'\theta^e) - \alpha G_1'(X'\tau^q) - G_2(X'\tau^e) \right) \right\| \right. \\
&\qquad \left. \cdot \left\| X\left(\alpha G_1'(X'\tau^q) + G_2(X'\tau^e) \right) \right\| \right],
\end{aligned}$$

where we apply (B.5) for the second summand. The remaining two summands can be bounded linearly by the arguments given in (B.7) since G_1' and G_2 are continuously differentiable functions and the respective moments are finite.

For the second component of ψ, we get that

$$\begin{aligned}
&\left\| \psi_2(Y, X, \tau) - \psi_2(Y, X, \theta) \right\| \\
&\quad \leq \left\| X(X'\theta^e - X'\theta^q) G_2'(X'\theta^e) - X(X'\tau^e - X'\tau^q) G_2'(X'\tau^e) \right\| \\
&\qquad + \left\| \frac{X G_2'(X'\theta^e) X'\theta^q}{\alpha} \left(\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}} \right) \right\| \\
&\qquad + \left\| \mathbb{1}_{\{Y \leq X'\tau^q\}} \left(\frac{X G_2'(X'\theta^e) X'\theta^q}{\alpha} - \frac{X G_2'(X'\tau^e) X'\tau^q}{\alpha} \right) \right\| \\
&\qquad + \left\| \frac{X G_2'(X'\theta^e)}{\alpha} Y \left(\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}} \right) \right\| \\
&\qquad + \left\| \frac{Y \mathbb{1}_{\{Y \leq X'\tau^q\}}}{\alpha} \left(X G_2'(X'\theta^e) - X G_2'(X'\tau^e) \right) \right\| \\
&\quad = \text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)} + \text{(v)}.
\end{aligned} \tag{B.56}$$

Thus, in order to evaluate $\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|\psi_2(Y,X,\tau)-\psi_2(Y,X,\theta)\right\|^2\right]$, we have to consider all the cross products out of the five summands in (B.56). Since the techniques applied are very similar, we only show details for two of the cross products.

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\text{(ii)}\cdot\text{(v)}\right]\\
&=\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|\frac{XG_2'(X'\theta^e)X'\theta^q}{\alpha}\left(\mathbb{1}_{\{Y\le X'\theta^q\}}-\mathbb{1}_{\{Y\le X'\tau^q\}}\right)\right\|\right.\\
&\qquad\left.\cdot\left\|\frac{Y\mathbb{1}_{\{Y\le X'\tau^q\}}}{\alpha}\left(XG_2'(X'\theta^e)-XG_2'(X'\tau^e)\right)\right\|\right]\\
&\le\frac{1}{\alpha^2}\mathbb{E}\left[\left\|XG_2'(X'\theta^e)X'\theta^q\right\|\cdot\mathbb{E}\left[|Y|\middle\|X\right]\cdot\|X\|\cdot\sup_{\tau\in\bar{U}_d(\theta)}\left\|G_2'(X'\theta^e)-G_2'(X'\tau^e\right\|\right]\\
&\le\frac{1}{\alpha^2}\mathbb{E}\left[\left\|XG_2'(X'\theta^e)X'\theta^q\right\|\cdot\mathbb{E}\left[|Y|\middle\|X\right]\cdot\|X\|\cdot\sup_{\tau\in\bar{U}_d(\theta)}\left\|XG_2''(X'\tau^e)\right\|\right]\cdot d\\
&=\mathcal{O}(d),
\end{aligned}
$$

by (B.7) since G_2' is continuously differentiable.

The following crossproducts can be bounded analogously by bounding the indicator functions and by applying the mean value theorem as in (B.7): (i)^2, (iii)^2, (v)^2, (i)·(iii), (i)·(iv), (i)·(v), (ii)·(iv), (ii)·(v), (iii)·(iv), (iii)·(v) and (iv)·(v).

A second type of technique, similar to the arguments in (B.9) arises in the cases (ii)^2, (iv)^2 and (ii)·(iv). We get that there exists $\theta_-^q,\theta_+^q\in\bar{U}_d(\theta)$ and a value $\tilde{\theta}^q$ on the line between θ_-^q and θ_+^q, such that

$$
\begin{aligned}
\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\text{(iv)}^2\right]&\le\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|^2\mathbb{E}_{Y|X}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|Y\left(\mathbb{1}_{\{Y\le X'\theta^q\}}-\mathbb{1}_{\{Y\le X'\tau^q\}}\right)\right|^2\right]\right]\\
&\le\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|^2\mathbb{E}_{Y|X}\left[Y^2\,\mathbb{1}_{\{X'\theta_-^q\le Y\le X'\theta_+^q\}}\right]\right]\\
&=\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|^2\int_{X'\theta_-^q}^{X'\theta_+^q}y^2f_{Y|X}(y)\mathrm{d}y\right]\\
&\le\mathbb{E}\left[\left\|\frac{XG_2'(X'\theta^e)}{\alpha}\right\|^2(X'\tilde{\theta}^q)^2f_{Y|X}(X'\tilde{\theta}^q)\left(X'\theta_+^q-X'\theta_-^q\right)\right]\\
&\le\frac{2}{\alpha}\mathbb{E}\left[\|X\|^3G_2'^2(X'\theta^e)\cdot\sup_{\tau\in\bar{U}_d(\theta)}(X'\tau^q)^2f_{Y|X}(X'\tau^q)\right]\cdot d\\
&=\mathcal{O}(d),
\end{aligned}
$$

where we apply a multivariate version of the mean value theorem and notice that $f_{Y|X}$ is bounded. □

Proof of Theorem 2.4. We apply Theorem 3 of Huber (1967) for the ψ-function as given in (2.9) and show the respective assumptions of this theorem.

For the measureability and separability of the ψ function, we refer to the proof of Theorem 2.2. It is already shown in the proof of Theorem 2.2 that there exists a $\theta_0\in\Theta$ such that $\lambda(\theta_0)=0$. For the technical conditions (N-3), we apply Lemma B.3, Lemma B.1 and Lemma B.4. It remains to show that $\mathbb{E}\left[||\psi(Y,X,\theta_0)||^2\right]<\infty$, which follows from the subsequent computation of C and the Moment Conditions ($\mathcal{M}$-3) in Appendix A.

The asymptotic covariance matrix is given by $\Lambda^{-1}C\Lambda^{-1}$, where

$$C = \mathbb{E}\big[\psi(Y,X,\theta_0)\,\psi(Y,X,\theta_0)'\big] \tag{B.57}$$

and

$$\Lambda = \frac{\partial\lambda(\theta)}{\partial\theta}\bigg|_{\theta=\theta_0} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix} = \begin{pmatrix}\left[\frac{\partial\lambda_1(\theta)}{\partial\theta^q}\right]_{\theta_0} & \left[\frac{\partial\lambda_1(\theta)}{\partial\theta^e}\right]_{\theta_0}\\ \left[\frac{\partial\lambda_2(\theta)}{\partial\theta^q}\right]_{\theta_0} & \left[\frac{\partial\lambda_2(\theta)}{\partial\theta^e}\right]_{\theta_0}\end{pmatrix}. \tag{B.58}$$

Straightforward calculations yield the matrix C as given in (2.21) - (2.23). For the computation of Λ, we first notice that the function

$$\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big] = \begin{pmatrix}\frac{1}{\alpha}\big(F_{Y|X}(X'\theta^q)-\alpha\big)\big(\alpha XG_1'(X'\theta^q)+XG_2(X'\theta^e)\big)\\ XG_2'(X'\theta^e)\Big(X'\theta^e - X'\theta^q + \frac{1}{\alpha}\mathbb{E}\big[(X'\theta^q - Y)\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big]\Big)\end{pmatrix} \tag{B.59}$$

is continuously differentiable for all θ in some neighborhood $U_d(\theta_0)$ around θ_0. The computation of $\frac{\partial}{\partial\theta}\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]$ is straight-forward apart from the following term. Let us choose a value $\tilde\theta$ in some small neighborhood of θ such that the distribution of Y given X has a density for all $\tau\in U_d(\theta)$ and such that $X'\tilde\theta \le X'\theta$. Then,

$$\begin{aligned}
&\frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big]\\
&= \frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X'\tilde\theta^q\}}\big|X\big] + \frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{X'\tilde\theta^q<Y\le X'\theta^q\}}\big|X\big]\\
&= \frac{\partial}{\partial\theta^q}\int_{X'\tilde\theta^q}^{X'\theta^q} y f_{Y|X}(y)\mathrm{d}y\\
&= X(X'\theta^q)f_{Y|X}(X'\theta^q).
\end{aligned} \tag{B.60}$$

We consequently get that for all $\theta\in U_d(\theta_0)$,

$$\frac{\partial}{\partial\theta^q}\mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] = (XX')\left(\frac{\alpha G_1'(X'\theta^q)+G_2(X'\theta^e)}{\alpha}\right)f_{Y|X}(X'\theta^q) \tag{B.61}$$

$$+ (XX')G_1''(X'\theta^q)\frac{F_{Y|X}(X'\theta^q)-\alpha}{\alpha}, \tag{B.62}$$

$$\frac{\partial}{\partial\theta^e}\mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] = \frac{\partial}{\partial\theta^q}\mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big] = (XX')G_2'(X'\theta^e)\frac{F_{Y|X}(X'\theta^q)-\alpha}{\alpha}, \tag{B.63}$$

$$\frac{\partial}{\partial\theta^e}\mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big] = (XX')G_2''(X'\theta^e)\left(X'\theta^q\frac{F_{Y|X}(X'\theta^q)-\alpha}{\alpha} + X'\theta^e - \frac{1}{\alpha}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big]\right) \tag{B.64}$$

$$+ (XX')G_2'(X'\theta^e). \tag{B.65}$$

In order to conclude that

$$\frac{\partial}{\partial\theta}\mathbb{E}\big[\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\big] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\right], \tag{B.66}$$

we apply a measure-theoretical version of the Leibniz integration rule, which mainly requires that the derivative of the integrand exists and is absolutely bounded by some integrable function $d(Y,X)$, independent of θ. For the first term, this can easily be obtained by defining

$$d(Y,X) = \sup_{U_d(\theta_0)}\left\|(XX')\left(\frac{\alpha G_1'(X'\theta^q)+G_2(X'\theta^e)}{\alpha}\right)f_{Y|X}(X'\theta^q)\right. \tag{B.67}$$

$$\left. + (XX')G_1''(X'\theta^q)\frac{F_{Y|X}(X'\theta^q)-\alpha}{\alpha}\right\|, \tag{B.68}$$

which has finite expectation by assumption. The other two terms follow the same reasoning. Inserting $\theta = \theta_0$ eventually shows (2.19) and (2.20). □

B.4 Proof of Theorem 2.5

Proof of Theorem 2.5. For this proof, we apply Theorem 5.23 from van der Vaart (1998). For that, the map $(Y, X) \mapsto \rho(Y, X, \theta)$ is obviously measurable as the sum of measurable functions. Furthermore, the map $\theta \mapsto \rho(Y, X, \theta)$ is almost surely differentiable since the only point of non-differentiability occurs where $Y = X'\theta^q$ which is a nullset with respect to the distribution of $\{Y, X\}$ and for all $\theta \in \Theta$ such that $Y \neq X'\theta^q$, its derivative is given by $\psi(Y, X, \theta)$.

For the local Lipschitz continuity of $\rho(Y, X, \theta)$ (with measurable and square-integrable Lipschitz-constant), we split the ρ-function such as in (B.20), i.e.

$$\rho(Y, X, \theta) = \rho_1(Y, X, \theta) + \rho_2(Y, X, \theta), \tag{B.69}$$

where

$$\rho_1(Y, X, \theta) = \mathbb{1}_{\{Y \leq X'\theta^q\}} \left(G_1(X'\theta^q) - G_1(Y) + \frac{1}{\alpha} G_2(X'\theta^e)(X'\theta^q - Y) \right), \tag{B.70}$$

$$\rho_2(Y, X, \theta) = G_2(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) - \mathcal{G}_2(X'\theta^e) - \alpha G_1(X'\theta^q) + a(Y). \tag{B.71}$$

Local Lipschitz continuity of ρ_2 follows since it is a continuously differentiable function and thus locally Lipschitz. We consequently get that for some $d > 0$ and for all $\theta_1, \theta_2 \in U_d(\theta_0)$, it holds that

$$\big|\rho_2(Y, X, \theta_1) - \rho_2(Y, X, \theta_2)\big| \leq \big\|\theta_1 - \theta_2\big\| \cdot \sup_{\theta \in U_d(\theta_0)} \left\| \begin{pmatrix} -XG_2(X'\theta^e) - \alpha X G_1'(X'\theta^q) \\ XG_2'(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) \end{pmatrix} \right\|, \tag{B.72}$$

with square-integrable Lipschitz-constant

$$K(Y, X) = \sup_{\theta \in U_d(\theta_0)} \left\| \begin{pmatrix} -XG_2(X'\theta^e) - \alpha X G_1'(X'\theta^q) \\ XG_2'(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) \end{pmatrix} \right\|. \tag{B.73}$$

For the function ρ_1, we consider three cases. First, let $\theta_1, \theta_2 \in \Theta$ such that $X'\theta_1^q \leq X'\theta_2^q < Y$. Then it holds that,

$$\rho_1(Y, X, \theta_1) = \rho_1(Y, X, \theta_2) = 0, \tag{B.74}$$

since $\mathbb{1}_{\{Y \leq X'\theta_1^q\}} = \mathbb{1}_{\{Y \leq X'\theta_2^q\}} = 0$, which is obviously a Lipschitz continuous function.

Second, let $\theta_1, \theta_2 \in \Theta$ such that $Y \leq X'\theta_1^q \leq X'\theta_2^q$. Then, for $\theta = \theta_1, \theta_2$,

$$\rho_1(Y, X, \theta) = G_1(X'\theta^q) - G_1(Y) + \frac{1}{\alpha} G_2(X'\theta^e)(X'\theta^q - Y), \tag{B.75}$$

which is a continuously differentiable function and thus

$$\big|\rho_1(Y, X, \theta_1) - \rho_1(Y, X, \theta_2)\big| \leq \big\|\theta_1 - \theta_2\big\| \cdot \sup_{\theta \in U_d(\theta_0)} \left\| \begin{pmatrix} XG_1'(X'\theta^q) + \frac{1}{\alpha} XG_2(X'\theta^e) \\ \frac{1}{\alpha} XG_2'(X'\theta^e)(X'\theta^q - Y) \end{pmatrix} \right\|. \tag{B.76}$$

Finally, let $\theta_1, \theta_2 \in \Theta$ such that $X'\theta_1^q < Y \leq X'\theta_2^q$. Then, since G_1 is increasing, we get that

$$\begin{aligned} \big|\rho_1(Y, X, \theta_1) - \rho_1(Y, X, \theta_2)\big| &= \left| G_1(X'\theta_2^q) - G_1(Y) + \frac{1}{\alpha} G_2(X'\theta_2^e)(X'\theta_2^q - Y) \right| \\ &\leq \big|G_1(X'\theta_2^q) - G_1(X'\theta_1^q)\big| + \left| \frac{1}{\alpha} G_2(X'\theta_2^e)(X'\theta_2^q - X'\theta_1^q) \right| \\ &\leq \big\|\theta_1^q - \theta_2^q\big\| \cdot \sup_{\theta \in U_d(\theta_0)} \left(\big\|XG_1'(X'\theta^q)\big\| + \frac{1}{\alpha} \big\|XG_2(X'\theta^e)\big\| \right). \end{aligned}$$

Thus, the function $\rho(Y, X, \theta)$ is locally Lipschitz continuous in θ with square-integrable Lipschitz constants, $\mathbb{E}\big[K(Y,X)^2\big] < \infty$ by the Moment Conditions ($\mathcal{M}$-4) in Appendix A.

We have already seen in the proof of Theorem 2.3 that the function $\mathbb{E}\big[\rho(Y, X, \theta)\big]$ is uniquely minimized at the point θ_0 and is twice continuously differentiable and consequently admits a second-order Taylor expansion at θ_0. Thus, we have shown the necessary assumptions of Theorem 5.23 from van der Vaart (1998). For the computation of the covariance matrix, we notice that the distribution of Y given X has a density $f_{Y|X}$ in a neighborhood of $X'\theta_0$ and thus, by the same arguments as in (B.60), we get that

$$\frac{\partial}{\partial\theta^q}\mathbb{E}\big[G_1(Y)\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big] = XG_1(X'\theta^q)f_{Y|X}(X'\theta^q). \tag{B.77}$$

Straight-forward calculations yields that for all $\theta \in U_d(\theta_0)$,

$$\begin{aligned}\frac{\partial}{\partial\theta^q}\mathbb{E}\big[\rho(Y,X,\theta)\big|X\big] &= \big(\alpha XG_1'(X'\theta^q) + XG_2(X'\theta^e)\big)\frac{F_{Y|X}(X'\theta^q)-\alpha}{\alpha} &\text{(B.78)}\\ &= \mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] &\text{(B.79)}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta^e}\mathbb{E}\big[\rho(Y,X,\theta)\big|X\big] &= XG_2'(X'\theta^e)\left(X'\theta^e - X'\theta^q + \frac{1}{\alpha}\mathbb{E}\big[(X'\theta^q - Y)\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big]\right) &\text{(B.80)}\\ &= \mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big]. &\text{(B.81)}\end{aligned}$$

By applying the Leibniz integration rule such as in (B.66), we finally get that

$$\frac{\partial}{\partial\theta}\mathbb{E}\big[\rho(Y,X,\theta)\big] = \mathbb{E}\big[\psi(Y,X,\theta)\big], \tag{B.82}$$

and thus, the asymptotic covariance matrix equals the one given in Theorem 2.4. □

B.5 Proof of Proposition B.5

Proposition B.5. Let $Y \in \mathcal{Y}$ be an absolutely continuous random variable with distribution function F and strictly positive density f on the whole support of Y. Then,

$$\frac{1}{\alpha^2}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y) - F(x)F(y)\mathrm{d}x\mathrm{d}y = \frac{1}{\alpha}\operatorname{Var}(Y|Y\le q_\alpha) + \frac{1-\alpha}{\alpha}\big(q_\alpha - es_\alpha\big)^2, \tag{B.83}$$

where $q_\alpha = F^{-1}(\alpha)$ denotes the α-quantile of Y and $es_\alpha = \mathbb{E}\big[Y\big|Y\le q_\alpha\big]$ denotes the ES of Y.

Proof. We proof this proposition by showing some equalities first which will be used later in the proof. Using integration by parts, we get that

$$\begin{aligned}\mathbb{E}\big[Y\big|Y\le q_\alpha\big] &= \int_{-\infty}^{q_\alpha} x\,\frac{f(x)}{F(q_\alpha)}\mathrm{d}x = \frac{1}{\alpha}\int_{-\infty}^{q_\alpha} x\,f(x)\mathrm{d}x\\ &= -\frac{1}{\alpha}\int_{-\infty}^{q_\alpha} F(x)\mathrm{d}x + \alpha q_\alpha,\end{aligned} \tag{B.84}$$

since for any distribution with finite first moments, it holds that $\lim_{x\to-\infty} xF(x) = 0$. Using the same line of argumentation, we get that

$$\mathbb{E}\big[Y^2\big|Y\le q_\alpha\big] = \frac{1}{\alpha}\int_{-\infty}^{q_\alpha} x^2\,f(x)\mathrm{d}x = -\frac{2}{\alpha}\int_{-\infty}^{q_\alpha} xF(x)\mathrm{d}x + \alpha q_\alpha^2, \tag{B.85}$$

since for a distribution having finite second moments, it holds that $\lim_{x\to-\infty} x^2 F(x) = 0$. By applying (B.84), we get that

$$\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x)F(y)\mathrm{d}x\mathrm{d}y = \left(\int_{-\infty}^{q_\alpha} F(x)\mathrm{d}x\right)^2 = \left(\alpha q_\alpha - \alpha\mathbb{E}\left[Y\middle|Y\le q_\alpha\right]\right)^2 = \alpha^2\left(q_\alpha - es_\alpha\right)^2. \tag{B.86}$$

Furthermore, notice that

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y)\mathrm{d}x\mathrm{d}y &= \int_{-\infty}^{q_\alpha}\left(\int_{-\infty}^{y} F(x)\mathrm{d}x + \int_{y}^{q_\alpha} F(y)\mathrm{d}x\right)\mathrm{d}y \\ &= \int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{q_\alpha}\int_{y}^{q_\alpha} F(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{q_\alpha} F(y)(q_\alpha - y)\mathrm{d}y.\end{aligned} \tag{B.87}$$

By rearranging the order of integration, for the first term in (B.87), we get that

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\,\mathrm{d}x\mathrm{d}y &= \iint_{\{(x,y):\, y\le q_\alpha,\, x\le y\}} F(x)\,\mathrm{d}x\mathrm{d}y = \iint_{\{(x,y):\, x\le q_\alpha,\, y\ge x\}} F(x)\,\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{q_\alpha}\int_{x}^{q_\alpha} F(x)\,\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{q_\alpha} F(x)(q_\alpha - x)\,\mathrm{d}y.\end{aligned} \tag{B.88}$$

Thus, by first using (B.87) and (B.88) and by plugging in (B.84) and (B.86), we obtain

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y)\mathrm{d}x\mathrm{d}y &= -2\int_{-\infty}^{q_\alpha} F(y)(q_\alpha - y)\,\mathrm{d}y \\ &= -2q_\alpha\int_{-\infty}^{q_\alpha} F(y)\,\mathrm{d}y + 2\int_{-\infty}^{q_\alpha} yF(y)\,\mathrm{d}y \\ &= 2q_\alpha\left(\alpha q_\alpha - \alpha es_\alpha\right) + \alpha\mathbb{E}\left[Y^2\middle|Y\le q_\alpha\right] - \alpha q_\alpha \\ &= \alpha\mathbb{E}\left[Y^2\middle|Y\le q_\alpha\right] + \alpha q_\alpha^2 - 2\alpha q_\alpha es_\alpha.\end{aligned} \tag{B.89}$$

Eventually, using (B.86) and (B.89), we get that

$$\frac{1}{\alpha^2}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y) - F(x)F(y)\mathrm{d}x\mathrm{d}y \tag{B.90}$$

$$= \frac{1}{\alpha^2}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y)\mathrm{d}x\mathrm{d}y - \frac{1}{\alpha^2}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x)F(y)\mathrm{d}x\mathrm{d}y \tag{B.91}$$

$$= \frac{1}{\alpha}\left(\mathbb{E}\left[Y^2\middle|Y\le q_\alpha\right] + q_\alpha^2 - 2q_\alpha es_\alpha\right) - \left(q_\alpha - es_\alpha\right)^2 \tag{B.92}$$

$$= \frac{1}{\alpha}\operatorname{Var}(Y|Y\le q_\alpha) + \frac{1-\alpha}{\alpha}\left(q_\alpha - es_\alpha\right)^2, \tag{B.93}$$

which concludes the proof. □

Appendix C Separability of almost surely continuous functions

Definition C.1 (Separability of a Stochastic Process). A stochastic process $\psi(x,\theta) : \Omega\times\Theta\to\mathcal{Y}$ is called separable in the sense of Doob, if there exists in Ω an everywhere dense countable set I, and in Ω a nullset N such that for any arbitrary open set $G\subset\Theta$ and every closed set $F\subset\mathcal{Y}$, the two sets

$$\{x|\psi(x,\theta)\in F,\ \forall\theta\in G\} \qquad \text{and} \tag{C.1}$$

$$\{x|\psi(x,\theta)\in F,\ \forall\theta\in G\cap I\} \tag{C.2}$$

differ from each other at most by a subset of N.

Proposition C.2 (Gikhman and Skorokhod (2004)). Let Θ and $\mathcal{Y}$ be metric spaces, Θ be a separable space. The sets (C.1) and (C.2) coincide for all $x \in \Omega$ for which the stochastic process $\psi(x, \theta)$ is continuous in θ.

Proof. It is clear that $\{x|\psi(x,\theta) \in F,\ \forall\theta \in G\} \subseteq \{x|\psi(x,\theta) \in F,\ \forall\theta \in G \cap I\}$. We thus only show the reverse.

Let $G \subset \Theta$ be an arbitrary open set and $F \subset \mathcal{Y}$ an arbitrary closed set. Let furthermore $x \in \Omega$ such that $\psi(x, \theta) \in F$ for all $\theta \in G \cap I$. We have to show that $\psi(x, \tilde{\theta}) \in F$ for all $\tilde{\theta} \in G$ but $\tilde{\theta} \notin I$.

Thus, let $\tilde{\theta} \in G \setminus I$. Since I is a dense set in Θ, there exists a sequence $(\theta_n)_{n\in\mathbb{N}} \in \Theta \cap I$, such that $\theta_n \to \tilde{\theta}$ and since G is an open set in Θ and $\tilde{\theta} \in G$, we can conclude that for $m \in \mathbb{N}$ large enough, $\theta_n \in G$ for all $n \geq m$. Furthermore, by continuity at θ, it holds that $\psi(x, \theta_n) \to \psi(x, \tilde{\theta})$ and since $\theta_n \in G \cap I$ for all n large enough, $\psi(x, \theta_n) \in F$ by assumption. Eventually, since F is a closed set, $\psi(x, \tilde{\theta}) \in F$ which proves the proposition. □

Corollary C.3 (Separability of continuous functions). Let Θ and $\mathcal{Y}$ be metric spaces, Θ be a separable space, and let the stochastic process $\psi(x, \theta)$ be almost surely continuous. Then, ψ is separable.

Proof. Since $\psi(x, \theta)$ is continuous for all $x \in \Omega \setminus N$ for some $N \subset \Omega$ with $\mathbb{P}(N) = 0$. We get from Proposition C.2 that the sets (C.1) and (C.2) coincide for all $x \in \Omega \setminus N$, i.e. they differ only by a subset of N. □

Appendix D Additional Monte-Carlo Simulation Results

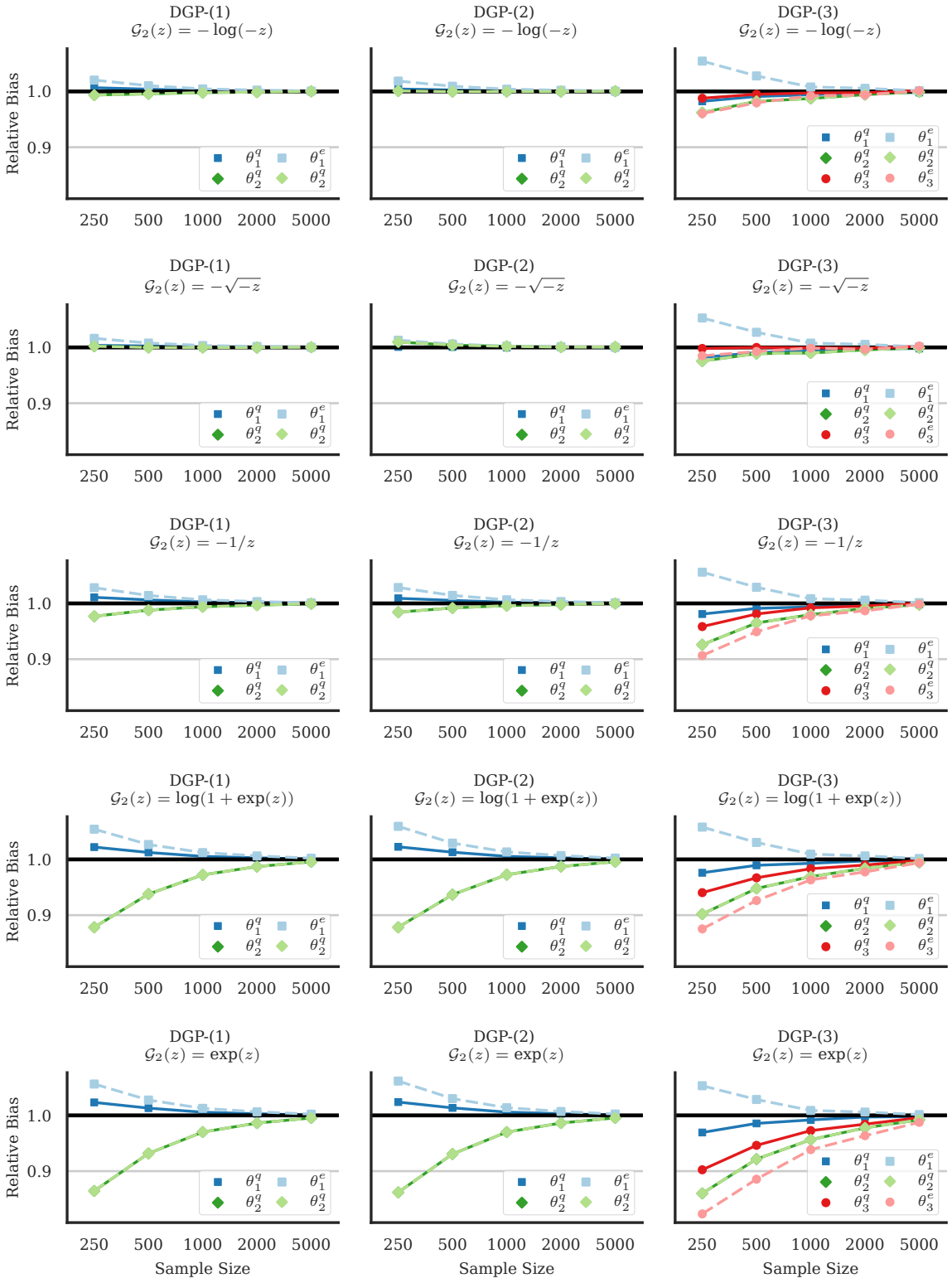

Figure 6: Relative bias (average estimated parameter divided by the true parameter) for all $\mathcal{G}_2$-functions, data generating processes and a range of sample sizes.

References

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228.

Basel Committee (2013). Fundamental review of the trading book: A revised market risk framework. Technical report, Basel Committee on Banking Supervision. Available at http://www.bis.org/publ/bcbs265.pdf.

Basel Committee (2016). Minimum capital requirements for Market Risk. Technical report, Basel Committee on Banking Supervision. Available at http://www.bis.org/bcbs/publ/d352.pdf.

Bayer, S. and Dimitriadis, T. (2017a). *esreg: Joint (VaR, ES) Regression*. R package version 0.1.2, available at https://github.com/BayerSe/esreg.

Bayer, S. and Dimitriadis, T. (2017b). Regression-based Expected Shortfall Backtesting. Working Paper.

Brazauskas, V., Jones, B. L., Puri, M. L., and Zitikis, R. (2008). Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, 138(11):3590–3604.

Chen, S. X. (2008). Nonparametric Estimation of Expected Shortfall. *Journal of Financial Econometrics*, 6(1):87–107.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1:93–125.

Engle, R. and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business and Economic Statistics*, 22(4):367–381.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660.

Fissler, T. (2017). *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space*. PhD thesis, Universität Bern.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4):1680–1707.

Fissler, T., Ziegel, J. F., and Gneiting, T. (2016). Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting. *Risk Magazine*, Janaury 2016.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011). Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business & Economic Statistics*, 29(1):150–160.

Gikhman, I. and Skorokhod, A. (2004). *The Theory of Stochastic Processes I*, volume 210 of *Classics in Mathematics*. Springer Berlin Heidelberg.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Halbleib, R. and Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, 36(8):1212–1228.

Hall, P. and Sheather, S. J. (1988). On the Distribution of a Studentized Quantile. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):381–391.

Hendricks, W. and Koenker, R. (1992). Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity. *Journal of the American Statistical Association*, 87(417):58–68.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. Berkeley: University of California Press.

Koenker, R. (1994). Confidence Intervals for Regression Quantiles. In Mandl, P. and Hušková, M., editors, *Asymptotic Statistics: Proceedings of the Fifth Prague Symposium, held from September 4–9, 1993*, pages 349–359. Physica-Verlag Heidelberg.

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

Koenker, R. and Xiao, Z. (2006). Quantile Autoregression. *Journal of the American Statistical Association*, 101(475):980–990.

Komunjer, I. (2013). Quantile Prediction. In *Handbook of Economic Forecasting*, volume 2, chapter 17, pages 961–994. Elsevier.

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting Properties of Probability Distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138. ACM.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. arXiv:1608.05498 [q-fin.RM].

Steinwart, I., Pasin, C., Williamson, R. C., and Zhang, S. (2014). Elicitation and Identification of Properties. In Balcan, M. F., Feldman, V., and Szepesvari, C., editors, *JMLR Workshop and Conference Proceedings*, volume 35, pages 1–45.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Žikeš, F. and Baruník, J. (2016). Semi-parametric Conditional Quantile Models for Financial Returns and Realized Volatility. *Journal of Financial Econometrics*, 14(1):185–226.