Electronic Journal of Statistics
Vol. 13 (2019) 1823–1871
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1560

A joint quantile and expected shortfall regression framework*

Timo Dimitriadis and Sebastian Bayer

University of Konstanz, Department of Economics
78457 Konstanz, Germany
e-mail: timo.dimitriadis@uni.kn; sebastian.bayer@uni.kn

Abstract: We introduce a novel regression framework which simultaneously models the quantile and the Expected Shortfall (ES) of a response variable given a set of covariates. This regression is based on strictly consistent loss functions for the pair consisting of the quantile and the ES, which allow for M- and Z-estimation of the joint regression parameters. We show consistency and asymptotic normality for both estimators under weak regularity conditions. The underlying loss functions depend on two specification functions, whose choices affect the properties of the resulting estimators. We find that the Z-estimator is numerically unstable and thus, we rely on M-estimation of the model parameters. Extensive simulations verify the asymptotic properties and analyze the small sample behavior of the M-estimator for different specification functions. This joint regression framework allows for various applications including estimating, forecasting and backtesting ES, which is particularly relevant in light of the recent introduction of the ES into the Basel Accords. We illustrate this through two exemplary empirical applications in forecasting and forecast combination of the ES.

Keywords and phrases: Expected shortfall, joint elicitability, joint regression, M-estimation, quantile regression.

Received May 2018.

Contents

*We thank the editor and an anonymous referee and associate editor for very helpful comments and suggestions. We further thank Tobias Fissler, Lyudmila Grigoryeva, Roxana Halbleib, Phillip Heiler, Frederic Menninger, Winfried Pohlmeier, Patrick Schmidt, Johanna Ziegel 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 authors acknowledge support by the state of Baden-Württemberg through bwHPC.

1. Introduction

We introduce a novel semiparametric regression framework for the Expected Shortfall (ES) by jointly modeling both, regression equations for the conditional quantile and the conditional ES. The ES at level $\alpha \in (0,1)$ is defined as the expected value of a random variable, given that its realizations exceed the α-quantile of the underlying distribution. We propose both, an M- and a Z-estimator for the joint regression parameters of both models and show that these estimators are consistent and asymptotically normal. Modeling simultaneous regression equations for the quantile and the ES is necessary as M- and Z-estimation of regression parameters of a stand-alone regression framework for the ES is infeasible. The underlying reason is that there does not exist an appropriate loss function that the ES minimizes in expectation and which could be used as the objective function for M-estimation of the regression parameters (Gneiting, 2011). However, Fissler and Ziegel (2016) show that there exists such a loss function if one considers the pair consisting of the quantile and the ES at the same probability level. This result gives rise to the idea of jointly modeling semiparametric models for both, the quantile and the ES and for jointly estimating the regression parameters through M-estimation. The situation for the Z-estimator and the availability of underlying identification functions (moment conditions) is equivalent to the loss functions used for the M-estimator and consequently only allows for joint Z-estimation of both regression equations.

Such a regression framework for the ES is essential for a variety of academic disciplines which consider measuring, forecasting and the evaluation of extreme risks. The most prominent example for this is financial risk management, where the Basel Accords recently proposed to use ES as the standard risk measure (Basel Committee, 2016). The previously used risk measure is the Value-at-Risk (VaR), which is defined as the α-quantile of the return distribution and which has several drawbacks as it is not coherent and fails to capture tail risks beyond itself (Artzner et al., 1999). These deficiencies are overcome by the ES as it 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. Modeling a regression equation for the ES opens up the possibility to extend the existing applications of quantile regression on VaR in the financial literature to ES, such as in Chernozhukov and Umantsev (2001), Engle and Manganelli (2004), Koenker and Xiao (2006), Gaglianone et al. (2011), Halbleib and Pohlmeier (2012), Komunjer (2013), Giacomini and Komunjer (2005), Xiao et al. (2015) and Žikeš and Baruník (2016). Such estimation, forecasting, and backtesting methods for the ES are particularly sought-after in light of the recent shift from VaR to ES in the Basel Accords. Bayer and Dimitriadis (2019) use this regression framework in order to construct an ES version of a Mincer-Zarnowitz backtest, where given ES forecasts are used as covariates and are consequently tested on their correct specification. Taylor (2019), Patton et al. (2019) and Chao et al. (2018) use this regression and estimation method in order to introduce different dynamic models for the ES with autoregressive features.

A further possible field of application for this regression framework arises in microeconometrics where researchers are interested in non-central features of the conditional distribution such as e.g. in income economics and the analysis of social inequalities. In these fields, a traditional method for the comparison of different regions of the conditional distribution is quantile regression (see e.g. Koenker, 2005, Section 1.5.). However, the interpretation of the ES as the mean of the worst α percent is more intuitive as the rather technical interpretation of quantiles, which motivates the use of an ES regression technique in these fields.

M-estimation (Z-estimation) of regression frameworks for different functionals can usually be applied based on different choices of loss (identification) functions. E.g. mean regression parameters can be estimated by employing any loss function from the Bregman class of loss functions and quantile regression parameters can be estimated by employing any member of the class of generalized piecewise linear loss functions (Gneiting, 2011; Gourieroux et al., 1984; Komunjer and Vuong, 2010). Equivalently, the possible loss and identification functions we employ for the M- and Z-estimator in this paper are not unique as they depend on two specification functions which can be chosen freely subject to some conditions.

Even though consistency and asymptotic normality hold for all applicable choices of these specification functions, the choices affect the asymptotic covariance of the estimators, the necessary moment conditions, the numerical stability of the optimization algorithm and the required computation times. We discuss the choice of these functions in a theoretical context with respect to asymptotic efficiency and necessary regularity conditions, and with respect to the numerical properties of the optimization algorithm. We find that so-called positively homogeneous loss functions (see e.g. Nolde and Ziegel, 2017) perform very well. Positive homogeneity of loss functions means that linear rescaling of the input variables does not alter the ranking of the losses, which is a crucial criteria keeping in mind that financial losses can be measured in different currencies which constitutes such a linear transformation.

The estimation of the asymptotic covariance matrix of the regression parameters imposes some difficulties. The first occurs in the estimation of the density quantile function, analogous to quantile regression (cf. Koenker, 2005) and thus, we utilize estimation procedures stemming from this literature. The second issue is the estimation of the variance of the negative quantile residuals conditional on the covariates, a nuisance quantity which is new to the literature. 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. Furthermore, we estimate the covariance matrix using the bootstrap. For ease of application, we provide an R package (Bayer and Dimitriadis, 2019) which contains the implementation of the M-estimator and where the user can choose the specification functions and the estimation method for the covariance matrix of the parameter estimates.

We conduct a Monte-Carlo simulation study where we consider four data generating processes with different properties. We numerically verify consistency and asymptotic normality of the M-estimator for a range of different choices of the specification functions. Furthermore, we find that the Z-estimator is numerically unstable due to the redescending nature of the utilized identification functions and consequently, we rely on M-estimation of the regression parameters. Moreover, we find that the performance of the M-estimator strongly depends on the specification functions, where choices resulting in positively homogeneous loss functions (Nolde and Ziegel, 2017; Efron, 1991) lead to a superior performance in terms of asymptotic efficiency, computation times, and mean squared error of the estimator.

We illustrate the usage of this joint regression framework for the VaR and the ES through two empirical applications. First, we jointly forecast the VaR and the ES based on lagged Realized Volatility (RV) estimates in a linear fashion, where the forecasting weights are estimated by our joint regression framework. For the evaluation of the forecasts, we apply strictly consistent loss functions for the VaR and ES (Fissler and Ziegel, 2016) together with the Model Confidence Set (Hansen et al., 2011) and Murphy Diagrams (Ziegel et al., 2019). We find that the forecasts stemming from our regression method outperform classical forecasting methods such as the Historical Simulation (HS), the GARCH and a HAR model based on RV estimates (Corsi, 2009). Second, we perform a joint forecast combination exercise where we apply our regression technique to estimate the combination weights. For this, we consider the RiskMetrics (RM) and HS forecasting methods and generate combined forecasts of these two models. We find that the combined forecasts outperform the stand-alone methods by means of the same evaluation criteria as in the first application.

Nadarajah et al. (2014) provide an overview of existing estimation methods for the ES. However, the reviewed approaches are only applicable for univariate data and not suitable for estimating the conditional ES through a regression technique. Nevertheless, there are some approaches for the ES which incorporate explanatory variables through indirect estimation procedures. Taylor (2008b) proposes an implicit approach for forecasting ES using exponentially weighted quantile regression and Taylor (2008a) introduces a procedure based on expectile regression and a relationship between the ES and expectiles. Taylor (2019) suggests a joint modeling technique for the quantile and the ES based on maximum likelihood estimation of the asymmetric Laplace distribution. However, asymptotic statistical theory for these estimation approaches of the conditional ES is not available. Barendse (2017) proposes a two-step estimation approach for a regression framework for the interquantile expectation.

In a simultaneous and independent work, Patton et al. (2019) introduce a similar M-estimator for such joint regression models for the quantile and the ES in the autoregressive context and also show its asymptotic behavior. This approach differs from our paper in the following ways. The class of M-estimators we propose is more general as we apply the full class of strictly consistent loss functions for the quantile and the ES, whereas Patton et al. (2019) only consider one special case. We also treat the class of corresponding Z-estimators and show their asymptotic properties. Furthermore, we consider several different estimators for the resulting asymptotic covariance matrix. In contrast, the work of Patton et al. (2019) is more general regarding the underlying assumptions on the data generating process, which allows for autoregressive modeling, which is the main focus of their work. In contrast, our approach focuses on modeling of the conditional quantile and the ES based on exogenous covariates.

The rest of the paper is organized as follows. In Section 2, we introduce the joint regression framework, the underlying regularity conditions together with the asymptotic properties of our estimators and discuss the choice of the specification functions. Section 3 provides details on the numerical implementation of the estimators and on the estimation of the asymptotic covariance matrix. Section 4 presents an extensive simulation study. Section 5 presents two exemplary empirical applications and Section 6 provides concluding remarks. The proofs are deferred to Appendix B.

2. Methodology

2.1. The joint regression framework

Following Lambert et al. (2008), Gneiting (2011) and Fissler and Ziegel (2016), we introduce the concept of (multivariate) p-elicitability. We consider random variables $Z : \Omega \to \mathbb{R}^d$, defined on some complete probability space $(\Omega, \mathcal{F}, P)$, a class of distributions $\mathcal{P}$ on $\mathbb{R}^d$, equipped with the Borel σ-field and a functional $\Gamma : \mathcal{P} \to D$ with its domain of action $D \subseteq \mathbb{R}^p, p \in \mathbb{N}$. We call an integrable loss function $\rho : \mathbb{R}^d \times D \to \mathbb{R}$ *strictly consistent* for the functional Γ relative to the class of distributions $\mathcal{P}$, if Γ is the unique minimizer of $\mathbb{E}\big[\rho(Z, \cdot)\big]$ for all distributions $F \in \mathcal{P}$, where F is the distribution of Z. Furthermore, we call a p-dimensional functional Γ *p-elicitable* relative to the class $\mathcal{P}$, if there exists a loss function ρ which is strictly consistent for Γ relative to $\mathcal{P}$. If the dimension p is clear from the context, we simply call the functional elicitable instead of p-elicitable.

Given the generalized α-quantile $Q_\alpha(Z) = F^{-1}(\alpha) = \inf\big\{z \in \mathbb{R} : F(z) \geq \alpha\big\}$ for some $\alpha \in (0, 1)$, the ES of the random variable Z at level α is defined as

$\mathrm{ES}_\alpha(Z) = \frac{1}{\alpha}\int_0^\alpha Q_u(Z)\,\mathrm{d}u$. If the distribution function of Z is continuous at its α-quantile, this definition can be simplified to the conditional tail expectation $\mathrm{ES}_\alpha(Z) = \mathbb{E}\big[Z \,\big|\, Z \le Q_\alpha(Z)\big]$. Gneiting (2011) shows that the ES is not 1-elicitable with respect to any class $\mathcal{P}$ of probability distributions on intervals $I \subseteq \mathbb{R}$, which contains measures with finite support or finite mixtures of absolutely continuous distributions with compact support (see also Weber, 2006). 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, and more consequential for this work, estimating the parameters of a stand-alone regression model for the functional ES in the sense that $\mathrm{ES}_\alpha(Y|X) = X'\theta_0^e$ by means of M-estimation, i.e. by minimizing some strictly consistent loss function, is infeasible. Even though the ES is not 1-elicitable, Fissler and Ziegel (2016) show that the pair consisting of the ES and the quantile at common probability level α is 2-elicitable relative to the class of distributions with finite first moments and unique α-quantiles and they characterize the full class of strictly consistent loss functions for this pair subject to some regularity conditions. Since the definition of the ES already depends on the respective quantile, the fact that the ES is only elicitable jointly with the quantile is not surprising.

We utilize this joint elicitability result for the introduction of a new joint regression framework for the quantile and the ES where the aforementioned class of strictly consistent loss functions serves as the basis for the M-estimation of the joint regression parameters. For this, let $Y : \Omega \to \mathbb{R}$ and $X : \Omega \to \mathbb{R}^k$ be random variables defined on the some complete probability space $\big(\Omega, \mathcal{F}, P\big)$ as above. Furthermore, we split $X = (X_q, X_e)$, where $X_q \in \mathbb{R}^{k_1}$ and $X_e \in \mathbb{R}^{k_2}$ such that $k_1 + k_2 = k$. This construction allows for using different explanatory variables for the models for the quantile and the ES. Henceforth, the transpose of X will be denoted by X', the cumulative distribution function of Y given X by $F_{Y|X}$ and the conditional density function by $f_{Y|X}$. For an l-times differentiable real-valued function $G : \mathbb{R} \to \mathbb{R}$, we denote the l-th derivative by $G^{(l)}(\cdot)$.

Assumption 2.1 (The joint regression model). The regression framework which jointly models the conditional quantile and ES of Y given the covariates X_q and X_e for some fixed level $\alpha \in (0, 1)$ is given by

$$Y = X_q'\theta_0^q + u^q \qquad \text{and} \qquad Y = X_e'\theta_0^e + u^e, \tag{1}$$

where $Q_\alpha(u^q|X) = 0$ and $\mathrm{ES}_\alpha(u^e|X) = 0$. The model is parametrized by $\theta_0 = (\theta_0^{q\prime}, \theta_0^{e\prime})' \in \Theta \subset \mathbb{R}^k$, where the parameter space Θ is compact, convex and has nonempty interior, $\mathrm{int}(\Theta) \neq \emptyset$.

This assumption implies that the model is correctly specified in the sense that there exists a true parameter vector $\theta_0 \in \Theta$ for which the joint model equals the true quantile and ES of the conditional distribution of Y given X. This model is semiparametric in the sense that we specify parametric models for the conditional quantile and the conditional ES of $F_{Y|X}$ without fully specifying this conditional distribution through parametric restrictions.

The model construction given in (1) allows the quantile and the ES models to depend on different covariate vectors X_q and X_e respectively. However, the conditions $Q_\alpha(u^q|X) = 0$ and $\mathrm{ES}_\alpha(u^e|X) = 0$ in Assumption 2.1 ensure that the joint model is correctly specified in the sense that $Q_\alpha(Y|X) = X_q'\theta_0^q$ and $\mathrm{ES}_\alpha(Y|X) = X_e'\theta_0^e$. Thus, including some covariates into the ES model, but not in the quantile model (or vice versa), is only valid if the true conditional quantile does not depend of the ES covariates. We present one example where the quantile and the ES depend on different covariates and the model is still correctly specified in the simulation study in Section 4. Furthermore, the empirical application of joint forecasting combination in Section 5.2 also depends on different covariates while preserving Assumption 2.1. The simple case where both models depend on the same vector of explanatory variables is naturally contained in this model assumption by the choice $X = (\tilde{X}, \tilde{X})$ for some vector of covariates $\tilde{X}$.

We propose both, an M-estimation and a Z-estimation procedure for the compound regression parameter vector θ_0. For the M-estimation, we utilize the class of strictly consistent joint loss functions[1] for the quantile and ES as given in Fissler and Ziegel (2016) such that it can be used in a regression framework,

$$\begin{aligned}\rho(Y, X, \theta) = {} & \big(\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \alpha\big) G_1(X_q'\theta^q) - \mathbb{1}_{\{Y \le X_q'\theta^q\}} G_1(Y) \\ & + G_2(X_e'\theta^e)\left(X_e'\theta^e - X_q'\theta^q + \frac{(X_q'\theta^q - Y)\mathbb{1}_{\{Y \le X_q'\theta^q\}}}{\alpha}\right) \\ & - \mathcal{G}_2(X_e'\theta^e) + a(Y),\end{aligned} \tag{2}$$

where the function G_1 is twice continuously differentiable, $\mathcal{G}_2$ is three times continuously differentiable, $\mathcal{G}_2^{(1)} = G_2$, G_2 and $G_2^{(1)}$ are strictly positive, G_1 is increasing and a and G_1 are integrable functions. Fissler and Ziegel (2016) also show that given some regularity conditions, there exist no strictly consistent loss functions outside the class of functions given in (2) which implies that this is the most general class of objective functions for the M-estimator of this regression framework. We discuss the choice of the *specification functions* G_1 and $\mathcal{G}_2$ in a theoretical context in Section 2.3 and by their numerical performance in Section 4.2. The function a only depends on Y and thus, it does not influence the estimated regression parameters and is usually set to zero. The corresponding (ρ-type) M-estimator is formally defined as a sequence $\hat{\theta}_{\rho,n}$ such that

$$\hat{\theta}_{\rho,n} = \mathrm{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i, \theta). \tag{3}$$

Instead of minimizing some objective function $\rho(Y, X, \theta)$ such as in (2) and (3), we can also define the corresponding Z-estimator (or ψ-type M-estimator), which sets a vector of identification functions (moment conditions) to zero. In the case of our joint quantile and ES regression, these identification functions

[1] One can interpret the structure of this loss function as follows (Fissler et al., 2016): The first summand in (2) is a strictly consistent loss function for the quantile (Gneiting, 2011) and hence only depends on the quantile, whereas the second summand cannot be split into a part depending only on the quantile and one depending only on the ES. This illustrates the fact that the ES itself is not 1-elicitable, but 2-elicitable together with the respective quantile.

are given by

$$\psi(Y,X,\theta) = \begin{pmatrix} X_q\big(G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)\big(\mathbb{1}_{\{Y\le X_q'\theta^q\}} - \alpha\big) \\ X_e G_2^{(1)}(X_e'\theta^e)\left(X_e'\theta^e - X_q'\theta^q + \frac{1}{\alpha}(X_q'\theta^q - Y)\mathbb{1}_{\{Y\le X_q'\theta^q\}}\right) \end{pmatrix}, \tag{4}$$

where the functions G_1 and G_2 are given as above. More generally, it suffices that these identification functions converge to zero almost surely and thus, we formally define the Z-estimator as a sequence $\hat{\theta}_{\psi,n}$, such that

$$\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, X_i, \hat{\theta}_{\psi,n}) \to 0 \quad \text{a.s.} \tag{5}$$

Heuristically, identification functions for a regression framework are usually obtained as the derivatives (with respect to θ) of some corresponding loss function. Furthermore, when the loss function $\rho(Y,X,\theta)$ is continuously differentiable in θ, it is obvious that the M- and Z-estimation approaches are equivalent, also in terms of their asymptotic distribution. In these cases, standard textbook asymptotic theory can be applied. However, for this joint quantile and ES regression, the loss functions $\rho(Y,X,\theta)$ given in (2) are not differentiable for the points where $Y = X_q'\theta^q$, and the identification functions are even discontinuous at these points. As this set of points forms a nullset with respect to the absolutely continuous distribution of Y given X, the identification functions $\psi(Y,X,\theta)$ are still almost surely the derivative of $\rho(Y,X,\theta)$. Consequently, for the proofs of the asymptotic theory presented in the subsequent section, we rely on asymptotic theory of M- and Z-estimation which allows for the case of non-differentiable objective functions as given in Huber (1967).

2.2. Asymptotic properties

In this section, we present the asymptotic properties of the M- and Z-estimator of the regression parameters. Consistency and asymptotic normality hold under the following set of weak regularity conditions, which are natural for this regression framework.

Assumption 2.2 (Regularity Conditions).

($\mathcal{A}$-1) The data (Y_i, X_i) for $i = 1,\dots,n$ is an iid series of random variables, distributed such as (Y,X) given above. Furthermore, the conditional distribution $F_{Y|X}$ has finite second moments and is absolutely continuous with probability density function $f_{Y|X}$, which is strictly positive, continuous and bounded in a neighbourhood of the true conditional quantile, $X_q'\theta_0^q$.

($\mathcal{A}$-2) The matrices $\mathbb{E}\big[X_q X_q'\big]$ and $\mathbb{E}\big[X_e X_e'\big]$ are positive definite.

($\mathcal{A}$-3) The functions $\rho(Y,X,\theta)$ and $\psi(Y,X,\theta)$ are given as in (2) and (4), where the function G_1 is twice continuously differentiable, $\mathcal{G}_2$ is three times continuously differentiable, $\mathcal{G}_2^{(1)} = G_2$, G_2 and $G_2^{(1)}$ are strictly positive, G_1 is increasing and a and G_1 are integrable.

Remark 2.3 (Finite Moment Conditions). We further have to assume that certain moments of X are finite. We specify the Finite Moment Conditions ($\mathcal{M}$-1)–($\mathcal{M}$-4) in Appendix A. Note that these general moment conditions simplify substantially for sensible choices of the specification functions G_1 and $\mathcal{G}_2$ as further outlined in Section 2.3.

Assumption ($\mathcal{A}$-1) is a combination of typical regularity conditions of mean and quantile regression. Absolute continuity of $F_{Y|X}$ with a strictly positive, bounded and continuous density function in a neighborhood of the true conditional quantile is also imposed for the asymptotic theory of quantile regression. Existence of the conditional moments of Y given X is subject to the conditions of mean regression and is included in our regularity conditions since the ES is a truncated mean. The positive definiteness (full rank condition) in ($\mathcal{A}$-2) is common for any regression design with stochastic regressors in order to exclude perfect multicollinearity of the regressors. The conditions for the specification functions G_1 and $\mathcal{G}_2$ in ($\mathcal{A}$-3) mainly originate from the conditions for the joint elicitability of the quantile and ES in Fissler and Ziegel (2016). Differentiability of these functions is required in this setup for obtaining the identification functions and for the differentiations in the computation of the asymptotic covariance in Theorem 2.6 and Theorem 2.7. The existence of certain moments of the explanatory variables as in conditions ($\mathcal{M}$-1)–($\mathcal{M}$-4) in Appendix A is also standard in any regression design relying on stochastic regressors. Even though compactness of the parameter space Θ in Assumption 2.1 generally simplifies the proofs, in this setup it is crucial for consistency of the Z-estimator as the identification functions ψ_2 are redescending to zero for many reasonable choices of the G_2 function such as e.g. the choices resulting in positively homogeneous loss functions. For details on this, we refer to Section 3.1.

The following Theorem shows consistency of the Z-estimator based on the full class of identification functions given in (4).

Theorem 2.4. Assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-1) in Appendix A hold true. Then, for the Z-estimator defined in (5), it holds that $\hat{\theta}_{\psi,n} \xrightarrow{\mathbb{P}} \theta_0$.

The proof of the Theorem is given in Appendix B. Equivalently, the subsequent Theorem shows consistency of the corresponding M-estimators.

Theorem 2.5. Assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-2) in Appendix A hold true. Then, for the M-estimator defined in (3), it holds that $\hat{\theta}_{\rho,n} \xrightarrow{\mathbb{P}} \theta_0$.

The proof of the Theorem is given in Appendix B. For the validity of this Theorem, it remains to assume that $\hat{\theta}_{\rho,n}$ *nearly* minimizes the loss function ρ in the sense that $\frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i, \hat{\theta}_{\rho,n}) \leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i, \theta_0) + o_P(1)$ rather than the strict definition of the M-estimator in (3).

We now turn to asymptotic normality of the Z-estimator as shown in the following Theorem.

Theorem 2.6. Assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-3) in Appendix A hold true. Then, for every sequence $\hat\theta_{\psi,n} \in \Theta$ satisfying $\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(Y_i, X_i, \hat\theta_{\psi,n}) \xrightarrow{\mathbb{P}} 0$, it holds that

$$\sqrt{n}\big(\hat\theta_{\psi,n} - \theta_0\big) \xrightarrow{d} \mathcal{N}\left(0, \Lambda^{-1} C \Lambda^{-1}\right), \tag{6}$$

with

$$\Lambda = \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix} \qquad \text{and} \qquad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{7}$$

where

$$\Lambda_{11} = \frac{1}{\alpha}\mathbb{E}\left[(X_q X_q') f_{Y|X}(X_q'\theta_0^q)\big(\alpha G_1^{(1)}(X_q'\theta_0^q) + G_2(X_e'\theta_0^e)\big)\right], \tag{8}$$

$$\Lambda_{22} = \mathbb{E}\big[(X_e X_e') G_2^{(1)}(X_e'\theta_0^e)\big], \tag{9}$$

$$C_{11} = \frac{1-\alpha}{\alpha}\mathbb{E}\left[(X_q X_q')\big(\alpha G_1^{(1)}(X_q'\theta_0^q) + G_2(X_e'\theta_0^e)\big)^2\right], \tag{10}$$

$$\begin{aligned} C_{12} = C_{21}' = \frac{1-\alpha}{\alpha}\mathbb{E}\,\big[&(X_q X_e')\big(X_q'\theta_0^q - X_e'\theta_0^e\big)\times \\ &\big(\alpha G_1^{(1)}(X_q'\theta_0^q) + G_2(X_e'\theta_0^e)\big) G_2^{(1)}(X_e'\theta_0^e)\big], \end{aligned} \tag{11}$$

$$\begin{aligned} C_{22} = \mathbb{E}\,\Big[&(X_e X_e')\big(G_2^{(1)}(X_e'\theta_0^e)\big)^2 \times \\ &\left(\frac{1}{\alpha}\operatorname{Var}\left(Y - X_q'\theta_0^q \big| Y \le X_q'\theta_0^q, X\right) + \frac{1-\alpha}{\alpha}\big(X_q'\theta_0^q - X_e'\theta_0^e\big)^2\right)\Big]. \end{aligned} \tag{12}$$

The proof of the Theorem is given in Appendix B. For asymptotic normality of the Z-estimator, we have to strengthen condition (5) to $\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(Y_i, X_i, \hat\theta_{\psi,n}) \xrightarrow{\mathbb{P}} 0$. The next Theorem shows asymptotic normality of the M-estimator and we can see that both estimators are subject to the same asymptotic covariance matrix.

Theorem 2.7. Assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-4) in Appendix A hold true. Then, for the M-estimator defined in (3), it holds that

$$\sqrt{n}\big(\hat\theta_{\rho,n} - \theta_0\big) \xrightarrow{d} \mathcal{N}\big(0, \Lambda^{-1} C \Lambda^{-1}\big), \tag{13}$$

where the matrices Λ and C are given as in Theorem 2.6.

The proof of the Theorem is given in Appendix B. Similar to the consistency statement of Theorem 2.5, it is possible to relax the minimization condition (3) to some *near* minimization. However, for this theorem it is required that the sequence $\hat\theta_{\rho,n}$ is such that $\frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \hat\theta_{\rho,n}) \le \inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \theta) + o_P(n^{-1})$, which is a stronger condition as required for Theorem 2.5.

Remark 2.8 (Quantile Regression). Notice that the asymptotic covariance matrix of the quantile-specific parameter estimates $\hat{\theta}^q$ is given by $\alpha(1-\alpha)D_1^{-1}D_0D_1^{-1}$, where

$$D_1 = \mathbb{E}\left[(X_qX_q')f_{Y|X}(X_q'\theta_0^q)\big(\alpha G_1^{(1)}(X_q'\theta_0^q) + G_2(X_e'\theta_0^e)\big)\right] \qquad \text{and} \tag{14}$$

$$D_0 = \mathbb{E}\left[(X_qX_q')\big(\alpha G_1^{(1)}(X_q'\theta_0^q) + G_2(X_e'\theta_0^e)\big)^2\right]. \tag{15}$$

This simplifies to the 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 identification functions. This demonstrates that the quantile regression method is nested in our regression procedure, also in terms of its asymptotic distribution.

Remark 2.9 (Asymptotic Covariance of the ES and the Oracle Estimator). The ES-specific part of the asymptotic covariance is mainly governed by the term C_{22}, which depends on the quantity

$$\begin{aligned} &\frac{1}{\alpha}\operatorname{Var}\big(Y - X_q'\theta_0^q \big| Y \le X_q'\theta_0^q, X\big) + \frac{1-\alpha}{\alpha}\big(X_q'\theta_0^q - X_e'\theta_0^e\big)^2 \\ &= \frac{1}{\alpha^2}\operatorname{Var}\left((Y - X_q'\theta_0^q)\mathbb{1}_{\{Y \le X_q'\theta_0^q\}}\,\middle|\, X\right). \end{aligned} \tag{16}$$

It is reasonable that the asymptotic covariance of ES regression parameters depends on the truncated variance of Y given X as the asymptomatic covariance of mean regression parameters is driven by the conditional (non-truncated) variance of Y given X. The second term $\frac{1-\alpha}{\alpha}\big(X_q'\theta_0^q - X_e'\theta_0^e\big)^2$ in (16) is included since the ES represents a truncated mean where the truncation point itself is a statistical functional (the quantile). In comparison, we consider an oracle M-estimator for the ES-specific regression parameters θ^e, given by the loss function

$$\rho_{\text{Oracle}}(Y, X, \theta^e) = (Y - X_e'\theta^e)^2\mathbb{1}_{\{Y \le X_q'\theta_0^q\}}, \tag{17}$$

where we assume that the true quantile regression parameters θ_0^q are known. The resulting asymptotic covariance is given by

$$\begin{aligned} \operatorname{AVar}\left(\widehat{\theta}^e_{\text{Oracle}}\right) = &\frac{1}{\alpha}\mathbb{E}\big[X_eX_e'\big]^{-1}\mathbb{E}\left[(X_eX_e')\operatorname{Var}\left(Y - X_e'\theta_0^e | \, Y \le X_q'\theta_0^q, X\right)\right] \\ &\times \mathbb{E}\big[X_eX_e'\big]^{-1}, \end{aligned} \tag{18}$$

which shows that the additional term $\big(X_q'\theta_0^q - X_e'\theta_0^e\big)^2$ is not included for this estimator with fixed truncation point $X_q'\theta_0^q$.

Remark 2.10 (Joint Estimation of the Sample Quantile and ES). We can use this regression framework to jointly estimate the quantile and ES of an identically distributed sample $Y_1, \dots, Y_n$ by regressing on a constant only.

The asymptotic covariance matrix given in Theorem 2.6 and Theorem 2.7 then simplifies to Σ with components

$$\Sigma_{11} = \frac{\alpha(1-\alpha)}{f_Y^2(\theta_0^q)}, \tag{19}$$

$$\Sigma_{12} = \Sigma_{21} = (1-\alpha)\frac{\theta_0^q - \theta_0^e}{f_Y(\theta_0^q)}, \tag{20}$$

$$\Sigma_{22} = \frac{1}{\alpha}\operatorname{Var}(Y - \theta_0^q | Y \le \theta_0^q) + \frac{1-\alpha}{\alpha}(\theta_0^q - \theta_0^e)^2, \tag{21}$$

where θ_0^q and θ_0^e are the true quantile and ES of Y. The same result is obtained by Zwingmann and Holzmann (2016), who further allow for a distribution function for Y which is not differentiable at the quantile with strictly positive derivative. Notice that in this simplified case without covariates, the asymptotic covariance matrix is independent of the specification functions G_1 and $\mathcal{G}_2$ used in the loss and identification functions. Furthermore, (19) implies that quantile estimates stemming from our joint estimation procedure have the same asymptotic efficiency as quantile estimates stemming from minimizing the generalized piecewise linear loss (Gneiting, 2011) and as sample quantiles (cf. Koenker, 2005). The same holds true for the efficiency of the sample ES estimators (based on the sample quantile) of Brazauskas et al. (2008) and Chen (2008).

Remark 2.11 (Pseudo-R^2 and the choice of $a(Y)$). By choosing $a(Y) = \alpha G_1(Y) + \mathcal{G}_2(Y)$ in (2), we can guarantee non-negative losses $\rho(Y, X, \theta) \ge 0$. This choice enables us to define a pseudo-R^2 for our joint regression framework in the sense of Koenker and Machado (1999),

$$R^{QE} = 1 - \frac{\rho(Y, X, \hat{\theta})}{\rho(Y, X, \tilde{\theta})}, \tag{22}$$

where $\hat{\theta}$ denotes the parameter estimates of the full regression model and $\tilde{\theta}$ denotes the parameter estimates of a regression model restricted to an intercept term only. However, this choice of $a(Y)$ comes at the cost of more restrictive moment conditions, since we need to impose that $\mathbb{E}\big[G_1(Y) + \mathcal{G}_2(Y)\big] < \infty$.

2.3. Choice of the specification functions

The loss and identification functions given in (2) and (4) depend on two specification functions, G_1 and $\mathcal{G}_2$ (with derivative G_2), which have to fulfill the regularity conditions ($\mathcal{A}$-3) in Assumption 2.2. Fissler et al. (2016) already mention the feasible choices $G_1(z) = 0$, $G_1(z) = z$, $G_2(z) = \exp(z)$ and $G_2(z) = \exp(z)/\big(1 + \exp(z)\big)$ in order to show that this class is non-empty. In contrast to the loss functions of mean, quantile and expectile regression, there is no natural choice for these specification functions for the quantile and ES yet (Nolde and Ziegel, 2017). However, as the choice of these functions strongly influences the performance of our regression procedure in terms of its asymptotic efficiency, the necessary moment conditions of the regressors and the numerical performance of the optimization algorithm, we discuss sensible selection criteria in the following.

Efron (1991) and Nolde and Ziegel (2017) argue that for M-estimation of regression parameters it is crucial that the utilized loss function is positively homogeneous of some order $b \in \mathbb{R}$ in the sense that

$$\rho(cY, X, c\theta) = c^b \rho(Y, X, \theta) \tag{23}$$

for all $c > 0$. This is an important 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. For the pair consisting of the quantile and the ES, Nolde and Ziegel (2017) characterize the full class of positively homogeneous[2] loss functions of order b for the case where we restrict the domain of $\mathcal{G}_2$, i.e. the conditional ES to the negative real line[3],

$$b < 0: \quad G_1(z) = -c_0 \qquad \mathcal{G}_2(z) = c_1(-z)^b + c_0 \tag{24}$$

$$b = 0: \quad G_1(z) = d_0 \mathbb{1}_{\{z \le 0\}} + d_0' \mathbb{1}_{\{z > 0\}} \qquad \mathcal{G}_2(z) = -c_1 \log(-z) + c_0 \tag{25}$$

$$b \in (0,1): G_1(z) = \big(d_1 \mathbb{1}_{\{z \le 0\}} + d_1' \mathbb{1}_{\{z > 0\}}\big)|z|^b - c_0 \quad \mathcal{G}_2(z) = -c_1(-z)^b + c_0 \tag{26}$$

for some constants $c_0, d_0, d_0' \in \mathbb{R}$ with $d_0 \le d_0'$, $d_1, d_1' \ge 0$ and $c_1 > 0$. There are no positively homogeneous loss functions for the cases $b \ge 1$. This results implies that the choice $G_1(z) = 0$ is a good candidate for all three homogeneous specifications and this is also a common choice in the existing literature (see Fissler et al., 2016; Nolde and Ziegel, 2017; Ziegel et al., 2019). We use this, but also the second prominent choice, $G_1(z) = z$ in the remainder of the paper.

A different natural guiding principle for selecting the specification functions is induced by choosing $\mathcal{G}_2$ (and G_1) such that the moment conditions ($\mathcal{M}$-1)–($\mathcal{M}$-4) in Appendix A are as least restrictive and as parsimonious as possible. For instance, choosing $\mathcal{G}_2$ such that G_2 and its first and second derivatives are bounded functions (and $G_1(z) = 0$) results in the moment condition $\mathbb{E}\big[||X||^5 + ||X||^4 \mathbb{E}\big[|Y|\big|X\big] + ||X||^3 \mathbb{E}\big[Y^2\big|X\big] + |a(Y)|\big] < \infty$. This motivates the usage of bounded functions[4] for G_2 such as e.g. the second example of Fissler et al. (2016), $G_2(z) = \exp(z)/\big(1 + \exp(z)\big)$, which is the distribution function

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

[3] Since the conditional ES of financial assets for small probability levels is always negative, this is no critical restriction. However, for the numerical parameter estimation, we have to restrict the parameter space Θ such that $X_{e,i}'\theta^e < 0$ for all $\theta \in \Theta$ and for all $X_{e,i}$ in the underlying sample. For details on this, we refer to Section 3.1.

[4] Note that the positively homogeneous loss functions exhibit unbounded $\mathcal{G}_2$ functions. However, as the function $\mathcal{G}_2(z)$ does not grow faster than linear as z tends to infinity, the resulting finite moment conditions are not too restrictive.

of the standard logistic distribution. Further examples of bounded G_2 functions include the distribution functions of absolutely continuous distributions on the real line. In the simulation study in Section 4.2, we compare the performance of different specification functions in terms of the mean squared error and the asymptotic efficiency of the estimator.

3. Numerical estimation of the model

In this section, we discuss the difficulties one encounters and the solutions we propose for estimating the joint regression model. Section 3.1 illustrates the numerical optimization procedure we employ for estimating the regression parameters and Section 3.2 discusses different estimation methods for the covariance matrix of the estimator.

3.1. Optimization

Theorem 2.6 and Theorem 2.7 imply that both, M-estimation and Z-estimation of the regression parameters θ have the same asymptotic efficiency and consequently, we discuss these estimation approaches in terms of their numerical performance in the following. The numerical implementation of the Z-estimator relies on root-finding of the identification functions given in (4), which we implement as in GMM-estimation by minimizing the inner product $\sum_i \psi(Y_i, X_i, \theta)' \cdot \sum_i \psi(Y_i, X_i, \theta)$. However, the identification functions are redescending to zero for many attractive choices of $\mathcal{G}_2$ in the sense that $\psi_2(Y, X, \theta) \to 0$ for $X_e'\theta^e \to -\infty$. Consequently, for θ such that $\theta^q = \theta_0^q$ and $X_e'\theta^e \to -\infty$, we get the same minimal value of the Z-estimation objective function $\sum_i \psi(Y_i, X_i, \theta)' \cdot \sum_i \psi(Y_i, X_i, \theta)$ as for the true regression parameters θ_0. Thus, the Z-estimator is numerically unstable and diverges in many setups.

Consequently, we rely on M-estimation of the regression parameters in the following. As the loss functions given in (2) are not differentiable and non-convex for all applicable choices of the specification functions (Fissler, 2017), we apply a derivative-free global optimization technique. More specifically, we use the Iterated Local Search (ILS) meta-heuristic of Lourenço et al. (2003), which successively refines the parameter estimates by repeated optimizations with iteratively perturbed starting values. Our exact implementation consists of the following steps. First, we obtain starting values for θ^q and θ^e from two quantile regressions of Y on X for the probability levels α and $\tilde{\alpha}$, where we choose $\tilde{\alpha}$ such that the $\tilde{\alpha}$-quantile and the α-ES coincide under normality. Second, using these starting values we minimize the loss function with the derivative-free and robust Nelder-Mead Simplex algorithm (Nelder and Mead, 1965). Third, we perturb the resulting parameter estimates by adding normally distributed noise with zero mean and standard deviation equal to the estimated asymptotic standard errors of the initial quantile regression estimates. Fourth, we re-optimize the model with the perturbed parameter estimates as new starting values. If the loss is further decreased by this re-optimization, we update the estimates and otherwise, we retain the previous ones. Fifth, we iterate over the previous two steps until the loss does not decrease in $m = 10$ consecutive iterations. Our numerical experiments indicate that this repeated optimization procedure yields estimates very close to the ones stemming from other global optimization techniques such as e.g. simulated annealing, whereas the major advantage of ILS is the considerably lower computation time.

For the choices of the specification functions which result in positively homogeneous loss functions, we have to restrict the domain of $\mathcal{G}_2$ to the negative real line as already discussed in Section 2.3. Thus, we have to restrict Θ such that $X_{e,i}'\theta^e < 0$ for all $\theta \in \Theta$ and for all $i = 1, \ldots, n$ during the optimization process. Even though in financial risk management the response variable Y is usually given by financial returns where the true (conditional) ES is strictly negative, there might still be some outliers $X_{e,i}$ such that $X_{e,i}'\theta_0^e \geq 0$. In such a case, imposing the restriction $X_{e,i}'\theta^e < 0$ for all $i = 1, \ldots, n$ during the optimization process generates substantially biased estimates for θ^e. In order to avoid this, we estimate the regression model for the transformed dependent variables $Y - \max(Y)$ for the positively homogeneous loss functions and add $\max(Y)$ to the estimated intercept parameters to undo the transformation[5].

We provide an R package for the estimation of the regression parameters (see Bayer and Dimitriadis, 2019). This package contains an implementation of the M-estimator and allows for choosing the specification functions G_1 and $\mathcal{G}_2$. Furthermore, the covariance matrix of the parameter estimates can be estimated either by using the asymptotic theory and the techniques we discuss in the next section, or by using the nonparametric iid bootstrap (Efron, 1979). We recommend applying the M-estimator with the ILS algorithm as this procedure exhibits the best performance in our numerical experiments with respect to accuracy, stability and computation times.

3.2. Asymptotic covariance estimation

While most parts of the asymptotic covariance matrix given in Theorem 2.6 and Theorem 2.7 are straightforward to estimate, two nuisance quantities impose some difficulties. The first is the density quantile function $f_{Y|X}(X_q'\theta_0^q)$, which is already well investigated in the quantile regression literature. In particular, we consider the estimators proposed by Koenker (1994), henceforth denoted by *iid* and by Hendricks and Koenker (1992), henceforth denoted by *nid*. The main difference between these is that the first is based on the assumption that the quantile residuals are independent of the covariates, whereas the second allows

[5]Note that this data transformation changes the average loss function as the applied loss functions are in general not translation invariant. Thus, optimizing the translated loss function can lead to different parameter estimates. However, we do not face the risk of obtaining substantially biased estimates in cases where $X_{e,i}'\theta_0^e \geq 0$ for some $i \in \{1, \ldots n\}$. Our numerical experiments indicate that the difference between estimating the model for Y and for $Y - \max(Y)$ is small when $X_{e,i}'\theta_0^e < 0$ for all $i \in \{1, \ldots n\}$, but can be quite substantial if there is an outlier for $X_{e,i}$ such that $X_{e,i}'\theta_0^e \geq 0$.

for a linear dependence structure. Both approaches depend on a bandwidth parameter which we choose according to Hall and Sheather (1988).

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

$$\operatorname{Var}\big(Y - X_q'\theta_0^q \big| Y \le X_q'\theta_0^q, X\big) = \operatorname{Var}\big(u^q \big| u^q \le 0, X\big). \tag{27}$$

Estimation of this quantity is demanding for two reasons. First, for very small probability levels which are typical in financial risk management such as e.g. $\alpha = 2.5\%$, the truncation $u^q \le 0$ cuts off all but very few (about $\alpha \cdot n$) observations. Second, modeling this truncated variance conditional on the covariates X is challenging, especially considering the very small sample sizes. Under the assumption of homoscedasticity, i.e. that the distribution of u^q is independent of the covariates X, we can simply estimate (27) by the sample variance of the negative quantile residuals and we refer to this estimator as *ind* in the following.

We propose two further estimators which allow for a dependence of the quantile residuals on the covariates. For this purpose, we assume a location-scale process with linear[6] specifications of the conditional mean and standard deviation in order to explicitly model the conditional relationship of u^q on X_q,

$$u^q = X_q'\zeta + X_q'\phi \cdot \varepsilon, \tag{28}$$

for some parameter vectors $\zeta, \phi \in \mathbb{R}^{k_1}$ and where $\varepsilon \sim G(0,1)$ follows a zero mean, unit variance distribution, such that $u^q | X_q \sim G\big(X_q'\zeta, (X_q'\phi)^2\big)$ with distribution function F_G and density f_G. As we need to estimate the truncated variance of u^q given $u^q \le 0$, i.e. a truncated variant of $(X_q'\phi)^2$, one possibility is to estimate (28) only for those observations where $u^q \le 0$. However, this approach particularly suffers from the very few negative quantile residuals as we need to estimate additional parameters compared to the *ind* approach.

We present a feasible alternative by estimating the parameters ζ and ϕ using all available observations of u^q and X_q by quasi generalized pseudo maximum likelihood (Gourieroux and Monfort, 1995, Section 8.4.4) and we obtain the truncated conditional variance by the scaling formula $\operatorname{Var}(u^q | u^q \le 0, X_q) = \int_{-\infty}^0 z^2 h(z)\, dz - (\int_{-\infty}^0 z h(z)\, dz)^2$, where $h(z) = f_G(z)/F_G(0)$ is the truncated conditional density of u^q given X_q and $u^q \le 0$. We propose one parametric estimator, henceforth denoted by *scl-N*, where we assume that the distribution G is the normal distribution and apply a closed-form solution to the scaling formula. We further propose a semiparametric estimator, henceforth denoted by *scl-sp*, where we estimate the distribution G nonparametrically and then apply the scaling formula for this estimated density by numerical integration.

We further propose to use the iid bootstrapping procedure (Efron, 1979). For this, we generate $B \in \mathbb{N}$ bootstrap samples, i.e. for each $b = 1, \ldots, B$, we take

[6]This approach can further be generalized by considering more general specifications for the conditional mean and standard deviation. However, our numerical experiments indicate that the estimation accuracy for the asymptotic covariance matrix does not increase by deviating from these linear specifications.

the original data $(Y_i, X_{q,i}, X_{e,i})$ for all $i = 1, \dots, n$ and resample n such triples with replacement. Taking this resampled data set, we compute the bootstrapped parameter estimate $\hat{\theta}^b_{\rho,n}$ for all $b = 1, \dots, B$. Then, the boostrapped covariance is given by the sample covariance over all bootstrapped parameter estimates $\hat{\theta}^b_{\rho,n}$ for $b = 1, \dots, B$.

4. Simulation study

In this section, we investigate the finite sample behavior of the M-estimator and verify the asymptotic properties derived in Section 2.2 through simulations. Furthermore, we compare the performance of different choices for the specification functions and evaluate the precision of the different covariance matrix estimators described in Section 3.2.

4.1. Data generating process

In order to assess the numerical properties of estimating the joint regression, we simulate data from a linear location-scale data generating process (DGP),

$$Y = Z'\gamma + (Z'\eta) \cdot v, \tag{29}$$

where $v \sim F(0, 1)$ follows some distribution with zero mean and unit variance, $Z = \big(1, Z_2, \dots, Z_s\big)'$ and $\gamma, \eta \in \mathbb{R}^s$ for some $s \in \mathbb{N}$. For this process, the true conditional quantile and ES are linear functions in Z, given by

$$Q_\alpha\left(Y|Z\right) = Z'(\gamma + z_\alpha \eta) \qquad \text{and} \qquad \mathrm{ES}_\alpha\left(Y|Z\right) = Z'(\gamma + \xi_\alpha \eta), \tag{30}$$

where z_α and ξ_α are the α-quantile and α-ES of the distribution $F(0, 1)$, which implies that $\theta^q_0 = \gamma + z_\alpha \eta$ and $\theta^e_0 = \gamma + \xi_\alpha \eta$. Furthermore, the conditional distributions of the quantile- and ES-residuals are given by

$$u^q|Z \sim F\left(-z_\alpha(Z'\eta), \, (Z'\eta)^2\right) \quad \text{and} \quad u^e|Z \sim F\left(-\xi_\alpha(Z'\eta), \, (Z'\eta)^2\right). \tag{31}$$

For the simulation study, we want to assess the performance of our regression procedure in various setups. Thus, we choose four different specifications for Z, γ, η and F in the following such that we get data generating processes (DGP) with different properties,

DGP-(1): $Z = (1, \, Z_2)$, $\quad Z_2 \sim \chi^2_1$ and $Y|Z \sim \mathcal{N}\big(-Z_2, \, 1\big)$

DGP-(2): $Z = (1, \, Z_2)$, $\quad Z_2 \sim \chi^2_1$ and $Y|Z \sim \mathcal{N}\big(-Z_2, \, (1 + 0.5 Z_2)^2\big)$

DGP-(3): $Z = (1, \, Z_2, \, Z_3)$ $\quad Z_2, \, Z_3 \sim U[0, 1]$ with $\mathrm{corr}(Z_2, \, Z_3) = 0.5$ and $Y|Z \sim t_5\Big(Z_2 - Z_3, \, (1 + Z_2 + Z_3)^2\Big)$.

DGP-(4): $Z = (1, \, Z_2, \, Z_3)$ $\quad Z_2, \, Z_3 \sim U[0, 1]$ with $\mathrm{corr}(Z_2, \, Z_3) = 0.5$ and $Y|Z \sim \mathcal{N}\Big(-z_\alpha Z_2 - \xi_\alpha Z_3, \, (1 + Z_2 + Z_3)^2\Big)$.

While DGP-(1) is homoscedastic, i.e. the model residuals are independent of Z, DGP-(2) is heteroscedastic, i.e. the conditional variance of the error terms depends on the covariates. Furthermore, with DGP-(3) we include a regression setup with multiple, correlated regressors and a leptocurtic conditional distribution. For these first three DGPs, we estimate the regression model by choosing the same vector of explanatory variables for both, the quantile and ES equation, $X = \big(X_q, X_e\big) = (Z, Z)$.

The last regression setup, DGP-(4) is specified such that the true conditional quantile and conditional ES depend on different vectors of explanatory variables. For this, we specify the parameters γ and η such that $\theta_0^q = \gamma + z_\alpha \eta = \big(z_\alpha, 0, z_\alpha - \xi_\alpha\big)$ and $\theta_0^e = \gamma + \xi_\alpha \eta = \big(\xi_\alpha, \xi_\alpha - z_\alpha, 0\big)$. We estimate this DGP by setting $X_q = (1, Z_3)$ and $X_e = (1, Z_2)$, as the true regression parameters of the second explanatory variable of the quantile model is zero, and equivalently the third parameter of the ES model.

We simulate all four processes 10,000 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 $X = (X_q, X_e)$ using our joint regression method for the probability level $\alpha = 2.5\%$. We choose this probability level in accordance with the choice of the Basel Committee which stipulate $\alpha = 2.5\%$ for the reported ES forecasts. Choosing more conservative probability levels towards the median of the distribution further stabilizes the simulation results, as by choosing $\alpha = 2.5\%$, we already consider a very extreme level which results in a lot of estimation uncertainty.

4.2. Comparing the specification functions

We start the discussion of the simulation results by investigating the numerical performance of the M-estimator based on different choices of the specification functions G_1 and $\mathcal{G}_2$ used in the loss function in (2). We use two choices for the first specification function,

$$G_1(z) = 0 \qquad \text{and} \qquad G_1(z) = z. \tag{32}$$

The first choice, $G_1(z) = 0$, follows the reasoning of Section 2.3 and Nolde and Ziegel (2017); Ziegel et al. (2019). It is interesting to notice that in classical quantile regression, the choice $G_1(z) = 0$ is not possible, however, as the second line of the loss function (2) also contains a quantile-specific piecewise linear loss function, this choice is actually feasible here. The second choice, $G_1(z) = z$, corresponds to the function used in classical quantile regression.

For the second specification function $\mathcal{G}_2$, we use three natural examples resulting in positively homogeneous loss functions of order $b = -1$, $b = 0$ and $b = 0.5$ respectively[7], a bounded G_2 function and the (unbounded) exponential

[7]Our numerical simulations show that the numerical results are unaffected by different choices of the associated constants in (24)–(26).

function:

$$\mathcal{G}_2(z) = -1/z, \qquad \mathcal{G}_2(z) = -\log(-z), \qquad \mathcal{G}_2(z) = -\sqrt{-z},$$
$$\mathcal{G}_2(z) = \log\big(1 + \exp(z)\big), \qquad \text{and} \qquad \mathcal{G}_2(z) = \exp(z). \tag{33}$$

Figure 1 presents the sum (over the k regression parameters) of the mean squared errors (MSE) of the regression parameters for the four DGPs described above, different sample sizes and for the five choices of the specification functions given in (33). As implied by the asymptotic theory, we obtain consistent parameter estimates for all five choices of the specification functions as the MSEs converge to zero for all four 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 $\mathcal{G}_2(z) = -1/z$. Furthermore, the bounded choice $\mathcal{G}_2(z) = \log\big(1+\exp(z)\big)$ still performs better than the unbounded exponential function. The results of the average estimation time in Figure 3 also favor the homogeneous loss functions.

These results can mainly be explained by the different true asymptotic covariances. Table 1 reports the Frobenius norms of the lower triangular parts of the true asymptotic covariance matrices and of the respective (lower triangular) quantile-specific and the ES-specific sub-matrices for the four DGPs and for the five choices of the specification functions given in (33). For comparison, we also report the Frobenius norm of the lower triangular part of the asymptotic covariance of the quantile regression estimator. We approximate the true asymptotic covariance matrix through Monte-Carlo integration with a sample size of 10^8 using the formulas in Theorem 2.6 and by using the true density and conditional truncated variance. On average, the specification functions $\mathcal{G}_2(z) = -\log(-z)$ and $\mathcal{G}_2(z) = -\sqrt{-z}$ exhibit the smallest asymptotic covariances, closely followed by the third choice for a positively homogeneous loss function, $\mathcal{G}_2(z) = -1/z$. The non-homogeneous choices lead to considerably larger asymptotic variances for all considered DGPs and sub-matrices, which confirms the results of Figure 1. Furthermore, by comparing the quantile-specific parameters of the joint estimation approach (from the positively homogeneous loss functions) to quantile regression estimates, we roughly obtain the same asymptotic efficiency.

In order to analyze the effect the individual parameters have on the average MSE, Figure 2 presents the MSEs of the individual regression parameters for different choices of the specification functions, for DGP-(3) and a fixed sample size of $n = 2000$. We present the results in a stacked bar plot, where the quantile parameters are shown in blueish colors whereas the ES parameters have reddish colors. This plot reveals several insights. First, the ES-specific parameters exhibit a larger MSE compared to their quantile counterparts. This observation is again driven by the larger true asymptotic variances of the ES-specific parameters as outlined in Table 1. This fact is explained as by definition, for any $\alpha < 0.5$ the ES is considerably further in the tail of the return distribution and thus, its estimation is subject to more noise. Second, for the first three choices

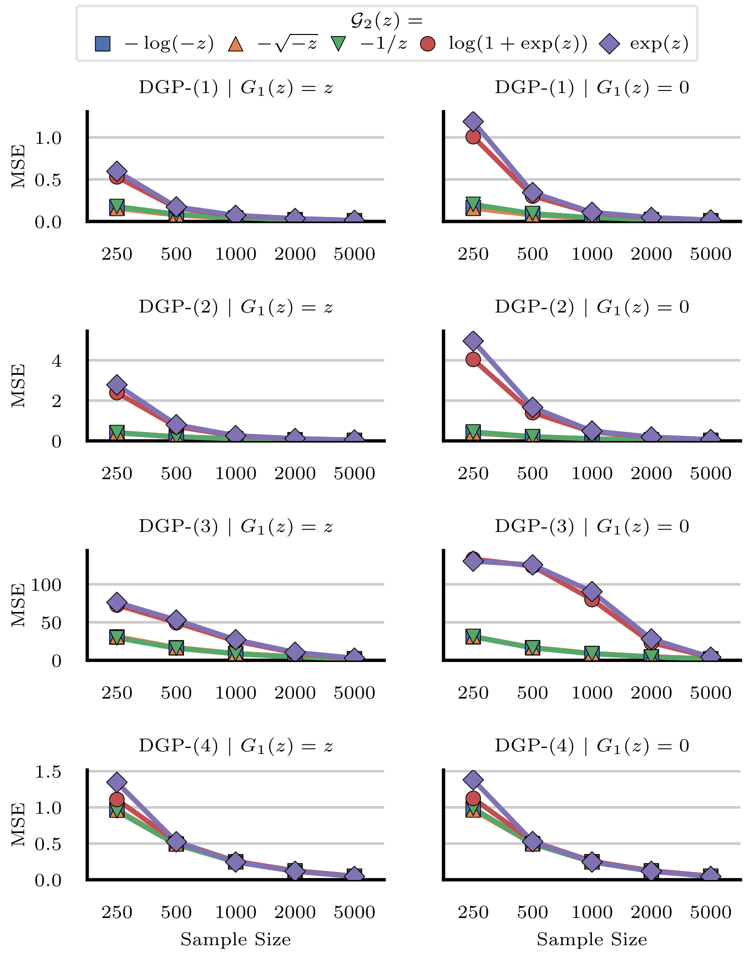

Fig 1. *This figure shows the average mean squared errors of the parameter estimates for all four DGPs. The results are shown for the five choices of the specification functions given in (33) and a range of sample sizes.*

of $\mathcal{G}_2$, resulting in homogeneous loss functions, the results are largely unaffected by the two choices of G_1. However, for the last two choices of $\mathcal{G}_2$, $G_1(z) = z$ results in a considerably smaller MSE. This fact can again be explained by the form of the loss function given in (2). For the choice $G_1(z) = 0$, the quantile parameters are only identified by the *hidden* piecewise linear loss function in the second line of (2), which is scaled by the function $G_2(X_e'\theta^e)$. This works fine for the benign homogeneous choices of $\mathcal{G}_2$, but not for the last two choices. In contrast, the first line of (2) does not vanish for the choice $G_1(z) = z$ and thus, it is able to identify the quantile parameters properly, resulting in a drastically smaller MSE for the quantile parameters.

TABLE 1

This table reports the Frobenius norms of the lower triangular parts of the true asymptotic covariance matrices and the respective quantile-specific and the ES-specific sub-matrices for the four DGPs and for the two times five choices of the specification functions given in (33). For comparison, we report the same quantity for the asymptotic covariance of the quantile regression estimator.

	DGP-(1)			DGP-(2)		
$G_1(z) = z$	Q	ES	Full	Q	ES	Full
$\mathcal{G}_2(z) = -\log(-z)$	12.0	18.4	24.8	34.3	59.4	77.0
$\mathcal{G}_2(z) = -\sqrt{-z}$	11.9	17.9	24.2	35.0	66.6	83.8
$\mathcal{G}_2(z) = -1/z$	11.9	19.9	25.8	36.7	50.8	71.1
$\mathcal{G}_2(z) = \log(1+\exp(z))$	13.2	37.3	42.4	32.6	138.8	145.8
$\mathcal{G}_2(z) = \exp(z)$	13.4	39.2	44.2	32.7	146.4	153.3
Quantile Regression	11.8	–	–	37.1	–	–
$G_1(z) = 0$	Q	ES	Full	Q	ES	Full
$\mathcal{G}_2(z) = -\log(-z)$	12.1	18.4	24.9	32.8	59.4	75.6
$\mathcal{G}_2(z) = -\sqrt{-z}$	11.9	17.9	24.2	34.5	66.6	83.4
$\mathcal{G}_2(z) = -1/z$	12.9	19.9	26.7	31.0	50.8	66.6
$\mathcal{G}_2(z) = \log(1+\exp(z))$	26.6	37.3	52.4	125.4	138.8	212.1
$\mathcal{G}_2(z) = \exp(z)$	27.3	39.2	54.5	129.3	146.4	221.5
Quantile Regression	11.8	–	–	37.1	–	–

	DGP-(3)			DGP-(4)		
$G_1(z) = z$	Q	ES	Full	Q	ES	Full
$\mathcal{G}_2(z) = -\log(-z)$	1469.8	4769.6	5329.0	43.7	180.0	193.7
$\mathcal{G}_2(z) = -\sqrt{-z}$	1469.5	4783.7	5341.8	43.8	180.8	194.5
$\mathcal{G}_2(z) = -1/z$	1471.0	4742.0	5304.1	43.8	178.3	192.2
$\mathcal{G}_2(z) = \log(1+\exp(z))$	1430.0	5861.9	6300.4	41.9	181.0	193.6
$\mathcal{G}_2(z) = \exp(z)$	1431.3	5978.0	6411.6	40.4	162.0	174.6
Quantile Regression	1471.0	–	–	43.9	–	–
$G_1(z) = 0$	Q	ES	Full	Q	ES	Full
$\mathcal{G}_2(z) = -\log(-z)$	1466.4	4769.6	5326.6	43.6	180.0	193.6
$\mathcal{G}_2(z) = -\sqrt{-z}$	1468.7	4783.7	5341.3	43.7	180.8	194.5
$\mathcal{G}_2(z) = -1/z$	1462.0	4742.0	5298.0	43.3	178.3	191.9
$\mathcal{G}_2(z) = \log(1+\exp(z))$	2419.8	5861.9	6857.5	41.8	181.0	193.5
$\mathcal{G}_2(z) = \exp(z)$	2454.1	5978.0	6987.1	40.3	162.0	174.5
Quantile Regression	1471.0	–	–	43.9	–	–

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, Figure 4 reports the average of the Frobenius norm of the lower triangular parts of the differences between the estimated covariances and the empirical covariance of the estimated parameters.[8] We report results for the three considered homogeneous loss functions and the four DGPs, where each of the plots presents

[8]This Frobenius norm corresponds to the MSEs averaged over the lower-triangular matrix entries, which makes this norm appropriate for evaluating covariance matrices.

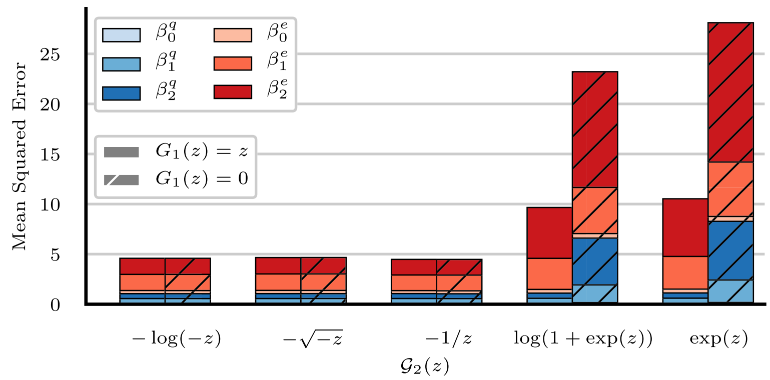

Fig 2. *This figure shows the mean squared errors of the individual parameter estimates for DGP-(3) and for the sample size $n = 2000$. The quantile parameters are shown in blueish colors whereas the ES parameters have reddish colors. The results are shown for the two times five choices of the specification functions given in (33).*

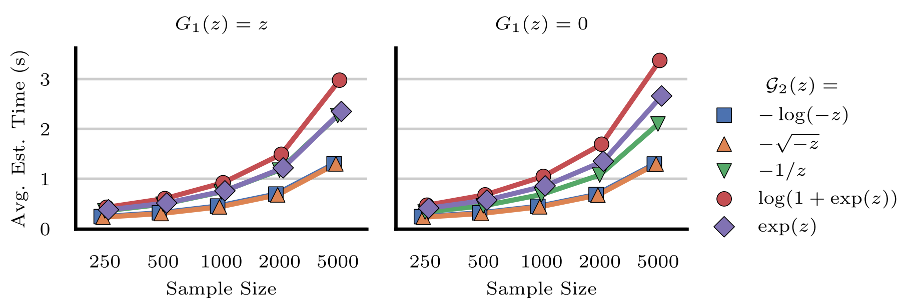

Fig 3. *This figure shows the estimation times averaged over all MC replications and over the four DGPs depending on the two times five choices of specification functions and for different sample sizes.*

the average norm differences for the four covariance estimators (*iid*/*nid*, *nid*/*scl-N*, *nid*/*scl-sp* and the iid bootstrap) depending on the sample size. We also report relative standard errors of the individual parameter estimates for all four DGPs, the four covariance estimators and for a fixed sample size of $n = 2000$ in Figure 5. This plots allows to consider the estimation accuracy of the individual (diagonal) entries of the covariance matrices individually, whereas a value of one indicates perfect estimation accuracy.

Combining the results from Figure 4 and Figure 5, we find that the *iid*/*nid* estimator performs well for the first, homoscedastic DGP whereas for the other two DGPs, it fails to capture the underlying more complicated dynamics of the data. The *nid*/*scl-N* estimator outperforms the other estimation approaches in the first, second and last DGPs, where the underlying conditional distribution indeed follows a normal distribution whereas its performance drops for the third DGP, which follows a Student-t distribution. The performance of the

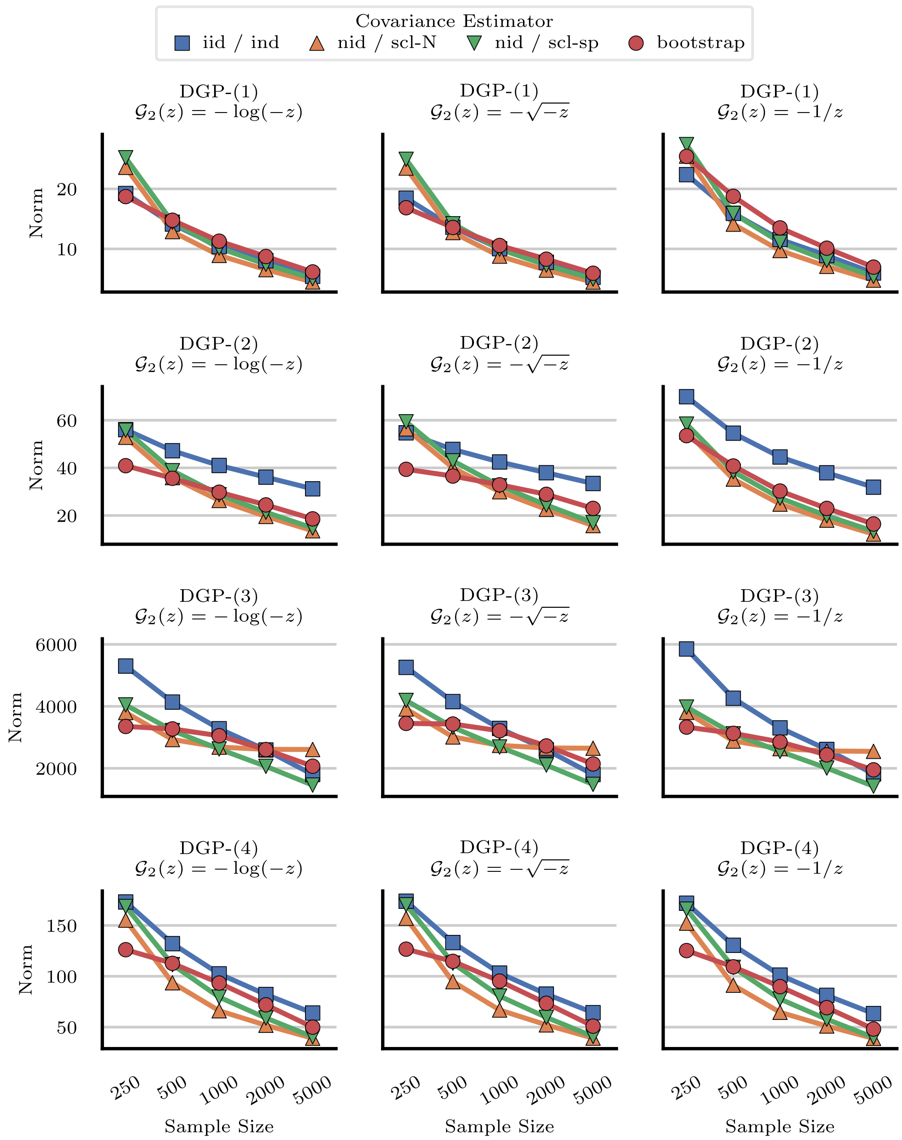

Fig 4. *This figure compares the four covariance estimation approaches described in Section 3.2 for the four data generating processes, a range of sample sizes and the three positively homogeneous choices of the loss function. We report the average of the Frobenius norm of the lower triangular part of the differences between the estimated asymptotic covariances and the empirical covariance of the M-estimator.*

flexible *nid*/*scl-sp* estimator is the most stable throughout all four DGPs. Eventually, the bootstrap estimator accurately estimates the covariance for all four DGPs and in comparison to the other estimators, it performs well even in small

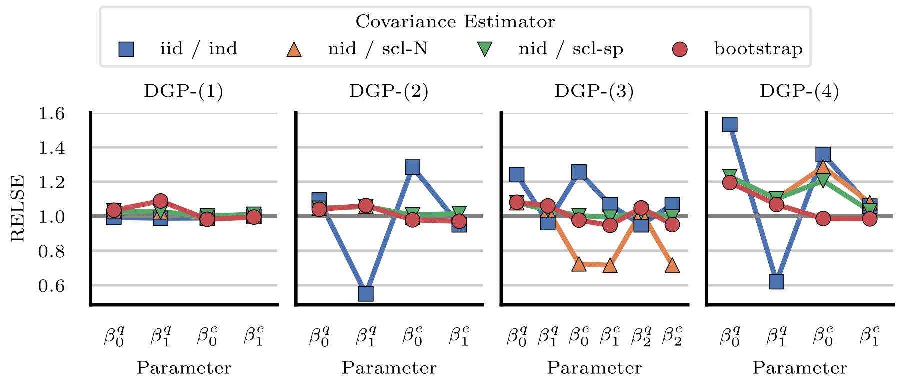

FIG 5. *This figure shows the relative standard errors of the individual model parameters for the four different DGPs and the four different covariance estimation methods. We fix the sample size* $n = 2000$ *and the specification functions* $G_1(z) = 0$ *and* $\mathcal{G}_2(z) = -\log(-z)$*.*

samples. Figure 5 further allows to disentangle the results for the different regression parameters where we cannot find that any of the parameters influence the covariance estimation predominantly throughout the DPGs and covariance estimation methods. Especially for the well-performing *nid/scl-sp* and bootstrap estimators, the relative standard errors are close to unity for all model parameters. The provided R package contains all four covariance estimators discussed in this section.

5. Empirical applications

In this section, we illustrate the use of this regression model through two practical applications in jointly forecasting VaR and ES. Section 5.1 considers forecasting VaR and ES based on Realized Volatility estimates, where the forecasting weights are estimated through our regression procedure. Section 5.2 introduces joint forecast combination for the VaR and ES, where we use this regression in order to estimate the combination weights. The latter presents an application relying on different explanatory variables for the quantile and the ES regression equations.

5.1. Forecasting VaR and ES by means of Realized Volatility

In this empirical application, we use our joint regression framework for forecasting VaR and ES of log returns of the S&P 500 stock market index. For this purpose, we adopt the VaR forecasting framework of Žikeš and Baruník (2016) and jointly forecast the VaR and ES of daily financial returns r_t by

$$Q_\alpha(r_t|\mathrm{RV}_{t-1}) = \theta_1^q + \theta_2^q \mathrm{RV}_{t-1} \quad \text{and} \quad \mathrm{ES}_\alpha(r_t|\mathrm{RV}_{t-1}) = \theta_1^e + \theta_2^e \mathrm{RV}_{t-1}, \quad (34)$$

where $\text{RV}_t = (\sum_i r_{t,i}^2)^{1/2}$ denotes the Realized Volatility (RV) estimator (Andersen and Bollerslev, 1998) for day t, where $r_{t,i}$ denotes the i-th high-frequency return of day t.

We extract close-to-close log returns and RV estimates based on 5 min returns of the S&P 500 index from the Realized Library (Heber et al., 2009) from January 4, 2000 until January 10, 2019 with a total of 4774 days. We estimate the model parameters using a rolling window of 1000 days and evaluate the forecasts on the remaining 3774 days.

We compare the predictive power of this model against three standard models from the literature. The first is the Historical Simulation (HS) approach, which forecasts the VaR and ES for day t as the sample quantile and ES of the daily returns of the past 250 trading days. The second is an AR(1)-GARCH(1,1)-t model (Bollerslev, 1986), and the third is the Heterogeneous Auto-Regressive (HAR) model of Corsi (2009), based on the RV estimates given above. Forecasts of the VaR and ES for the HAR model are obtained from the volatility forecasts and by assuming a standard normal distribution. While the first two of these approaches rely on daily data only, the third one incorporates the same high frequency information as our approach.

Table 2

This table shows the joint losses for the VaR and ES forecasts stemming from the GARCH, HS, and RV-HAR model. We display the losses for the five choices of the specification function G_2 given in (33).

$\mathcal{G}_2(z) =$		$-\log(-z)$	$-\sqrt{-z}$	$-1/z$	$\log(1+\exp(z))$	$\exp(z)$
$G_1(z) = z$	ESR	0.991**	1.685**	−0.355**	−0.025**	−0.032**
	HS	1.197	1.857	−0.260	0.021	0.020
	GARCH	1.022*	1.703*	−0.324*	−0.016*	−0.020*
	HAR	1.172	1.777	−0.178	0.028	0.039
$G_1(z) = 0$	ESR	0.922**	1.616**	−0.424**	−0.094**	−0.101**
	HS	1.114	1.774	−0.343	−0.061	−0.063
	GARCH	0.952*	1.633*	−0.394*	−0.086*	−0.090*
	HAR	1.099	1.704	−0.251	−0.045	−0.034

We evaluate the forecasting power of the VaR and ES of these models by using strictly consistent loss functions for this pair. These loss functions are given in (2), where the respective linear quantile and ES models are replaced by the issued VaR and ES forecasts (Fissler and Ziegel, 2016; Fissler et al., 2016). Table 2 shows the average losses for the respective models and for the ten choices of specification functions already used in the MC study. Two asterisks denote the model with the smallest average loss, one asterisk denotes models contained in the 10% Model Confidence Set (MCS) (Hansen et al., 2011). We find that our regression based forecasting model outperforms all three competitors for all considered strictly consistent loss functions as our regression model exhibits the smallest average losses. We outperform the HS and HAR models significantly as they are not contained in any MCS for all ten loss functions.

As the comparison results are always influenced by the more or less arbitrary choice of utlizied (strictly consistent) loss function, Ziegel et al. (2019) introduce the concept of *forecast domincance* for the VaR and ES, which tests whether one forecast dominates another one for an entire class of loss functions (also see Ehm et al. (2016)). They introduce *Murphy diagrams* as a graphical tool in order to test for forecast dominance by plotting *elementary score differences* together with their pointwise confidence bands. In fact, one forecasting model significantly dominates another one with respect to the considered class of loss functions if and only if the elementary score differences plotted in the Murphy diagrams are significantly negative (positive).

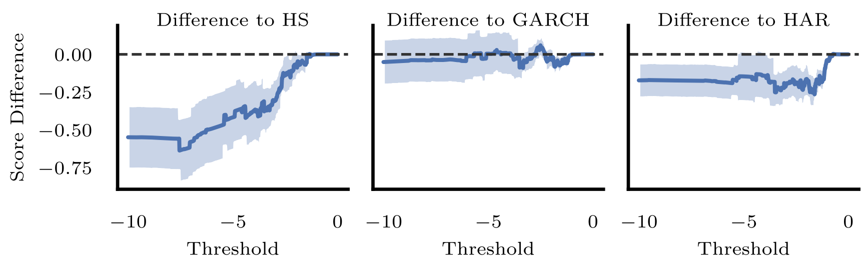

Fig 6. *This figure shows elementary score differences (Murphy diagrams, Ziegel et al., 2019) of the VaR/ES Regression and the respective comparison models. The shaded area depicts the pointwise* 90% *confidence band.*

Figure 6 displays the Murphy diagrams for the VaR and ES losses generated by our ESR forecasting model against the three alternative models together with the respective 90% pointwise confidence bands. This analysis strengthens the results from Table 2 as we can see that the ESR model dominates the HS and the HAR forecasting approaches for almost all threshold values, i.e. for almost all loss functions. Even though we also observe mostly negative elementary score differences in the comparison to the GARCH model, these differences are not significant uniformly over all threshold values and consequently, we cannot dominate this model. It is particularly remarkable that we can significantly dominate the HAR model as both, this model and our ESR model rely on the same underlying data, namely the prior RV estimates. This illustrates that in terms of the VaR and ES, estimating the forecasting weights through this joint regression framework significantly improves the forecasting performance in comparison to applying an indirect approach through a location-scale model as in the HAR model.

5.2. Joint VaR and ES forecast combination

In this section, we consider combination of VaR and ES forecasts by employing our joint VaR and ES regression method. Combining forecasts stemming from different models, estimation approaches, data or information sets has several advantages over standalone forecasts, where e.g. Timmermann (2006) provides three arguments in favor of forecast combination: first, there are diversification gains stemming from the combination of forecasts computed from different assumptions, specifications or information sets. Second, combined forecasts tend to be robust against structural breaks. Third, the influence of potential misspecification of the individual models is reduced due to averaging over a set of forecasts stemming from several models.

Giacomini and Komunjer (2005) further argue that forecast combination is particularly advantageous for risk measures (quantiles) with small probability levels, as it is customary for the VaR and the ES. These extreme risk measure are very sensitive to the few observations in the tails of the empirical distribution of the sample, and thus, forecast combinations based on different information sets can be seen as a way to make the forecast performance more robust to the effects of sample-specific factors. Halbleib and Pohlmeier (2012) and Bayer (2018) provide further empirical evidence in favor of forecast combinations, especially for financial risk measures and in turbulent financial times.

In the following, we consider two series of given forecasts $\big(\hat{v}_{t,1}, \hat{e}_{t,1}\big)$ and $\big(\hat{v}_{t,2}, \hat{e}_{t,2}\big)$ for all $t = 1, \dots, T$ which are $\mathcal{F}_{t-1}$-measureable, i.e. they are issued with the available information in $\mathcal{F}_{t-1}$. We are generally agnostic about the underlying methods used to generate these forecasts and which can be parametric, semiparametric or nonparametric. We consider joint forecast combination of the VaR and ES forecasts, given by specifying the regression with $Y_t = r_t$, $X_{q,t} = \big(1, \hat{v}_{t,1}, \hat{v}_{t,2}\big)'$ and $X_{e,t} = \big(1, \hat{e}_{t,1}, \hat{e}_{t,2}\big)'$ for all $t = 1, \dots, T$. The combination weights are then given by the regression parameters $\theta^q = \big(\theta^q_1, \theta^q_2, \theta^q_3\big)'$ and $\theta^e = \big(\theta^e_1, \theta^e_2, \theta^e_3\big)'$.

This forecast combination constitutes a sensible real-world application of a regression setup relying on different covariate vectors, as the covariates $X_{q,t}$, and $X_{e,t}$ are different in this setup. Under the assumption that at least one linear forecast combination of the given VaR and ES forecasts is correctly specified, it holds that $Q_\alpha(u^q_t|X_t) = 0$ and $\mathrm{ES}_\alpha(u^e_t|X_t) = 0$ almost surely. This implies that the model specification of Assumption 2.1 is reasonable in this setting.

We use daily close-to-close log returns from the S&P500, i.e. the same data as in Section 5.1. We generate one day ahead forecasts stemming from the RiskMetrics (RM) and the HS model. The combination weights $\big(\theta^{q\prime}, \theta^{e\prime}\big)'$ are estimated using the daily returns and the issued forecasts of the first 1000 trading days, resulting in the estimates $\hat{\theta}^q = (-0.23,\, 0.63,\, 0.25)'$ and $\hat{\theta}^e = (0.08,\, 0.62,\, 0.42)'$. The combined forecasts are generated using these estimated parameters for the remaining 3774 trading days. We evaluate the standalone and the combined forecasts using strictly consistent loss functions as in Section 5.1, together with the MCS (Hansen et al., 2011) and Murphy diagrams (Ziegel et al., 2019).

Table 3 depicts the average loss for the ten different combinations of specification functions where two asterisks mark the smallest loss and one asterisk denotes additional models contained in the MCS. We find that in nine of the ten different choices of strictly consistent loss functions, the combined forecasts exhibit the smallest loss compared to the stand-alone forecasts, implying a better average forecasting performance. Furthermore, the stand-alone forecasts are not contained in the MCS for roughly half of the cases, which implies that the results are also significant for these instances.

Table 3
*This table shows the average losses for the stand-alone HS and RM and the combined forecasts for the ten choices of specification functions given in (33). ** denotes the model with the smallest loss and * denotes models contained in the 90% MCS.*

$\mathcal{G}_2(z) =$		$-\log(-z)$	$-\sqrt{-z}$	$-1/z$	$\log(1+\exp(z))$	$\exp(z)$
$G_1(z) = z$	RM	1.153*	1.779**	−0.214	0.021*	0.028
	HS	1.246	1.897	−0.237*	0.033	0.033
	Comb	1.126**	1.779*	−0.268**	0.011**	0.011**
$G_1(z) = 0$	RM	1.079*	1.705*	−0.288	−0.053*	−0.047
	HS	1.161	1.811	−0.322*	−0.052*	−0.053*
	Comb	1.051**	1.704**	−0.344**	−0.064**	−0.064**

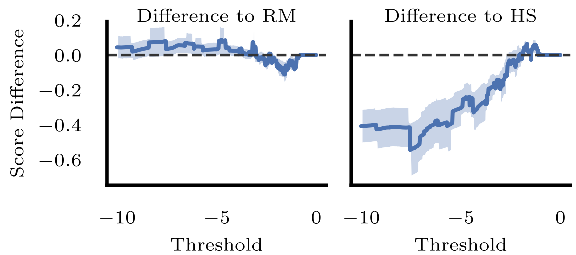

Fig 7. *This figure shows elementary score differences (Murphy diagrams, Ziegel et al., 2019) which compare the baseline forecast combination model with the two stand-alone RM and HS forecasting models. The shaded area depicts the pointwise* 90% *confidence band.*

We also present Murphy diagrams in Figure 7 in order to show the losses for the full class of strictly consistent loss functions. We find that the combined forecasts dominate the HS approach. However, even though the RM approach can be outperformed significantly for some of the loss functions, we cannot dominate this forecast in the strong sense of Ziegel et al. (2019) through Murphy diagrams.

6. Conclusion

In this paper, we introduce a joint regression technique for the quantile (the VaR) and the ES. This regression approach relies on the class of strictly consistent joint loss functions introduced by Fissler and Ziegel (2016), which permits the joint elicitation of the quantile and the ES. We introduce an M- and a 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, which we also verify numerically through extensive simulations. The underlying loss and identification functions and the asymptotic covariance matrices of the estimators depend on the choice of two specification functions, which we investigate in terms of the resulting moment conditions, asymptotic efficiency, numerical performance and computation times. In our numerical simulations, we find that choices resulting in positively homogeneous loss functions dominate other choices with respect to the aforementioned criteria. Furthermore, we propose several estimation methods for the asymptotic covariance matrix, which are able to cope with different properties of the underlying data. We provide an R package (Bayer and Dimitriadis, 2019) which provides the M-estimation procedure where one can choose the underlying specification functions and the estimation method for the asymptotic covariance matrix.

Our new joint regression technique allows for a wide range of applications for the risk measures VaR and ES. We illustrate two empirical applications of this regression method by setting up a joint forecasting model for the VaR and the ES based on past RV estimates and by considering joint forecast combination. In both applications, our regression-based methods can outperform their competitors. The new regression technique is already used by Taylor (2019), Patton et al. (2019) and Chao et al. (2018) for the estimation of autoregressive models for the ES (jointly with the VaR). Bayer and Dimitriadis (2019) use this regression to develop an ES backtest which is particularly relevant in light of the recent introduction of ES into the Basel regulatory framework and the present lack of accurate backtesting methods for the ES. Furthermore, this regression approach can be used to model the ES (jointly with the VaR) by generalizing existing applications of quantile regression on VaR, such as e.g. in Koenker and Xiao (2006), Engle and Manganelli (2004), Chernozhukov and Umantsev (2001), Žikeš and Baruník (2016), Halbleib and Pohlmeier (2012), Komunjer (2013), Giacomini and Komunjer (2005) and Xiao et al. (2015).

Appendix A: Finite moment conditions

For convenience of the supremum notation, for all $\theta \in \text{int}(\Theta)$ and for $d > 0$, we define the open neighborhood $U_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| < d\}$ and its closure $\bar{U}_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| \leq d\}$. In the following, if not stated otherwise, the sup notation is understood as the supremum in the neighborhood $\sup_\theta = \sup_{\theta \in \bar{U}_{d_0}(\theta_0)}$ (or equivalently $\sup_\tau = \sup_{\tau \in \bar{U}_{d_0}(\theta_0)}$).

($\mathcal{M}$-1) For Theorem 2.4, we assume that the following moments are finite for some $d_0 > 0$:

- $\mathbb{E}[||X||^2 \sup_\theta |G_1^{(1)}(X_q'\theta^q)|]$
- $\mathbb{E}[||X||^2 \sup_\theta |G_1^{(2)}(X_q'\theta^q)|]$
- $\mathbb{E}[||X||^2 \sup_\theta |G_2(X_e'\theta^e)|]$
- $\mathbb{E}[||X||^3 \sup_\theta |G_2^{(1)}(X_e'\theta^e)|]$
- $\mathbb{E}[||X||^3 \sup_\theta |G_2^{(2)}(X_e'\theta^e)|]$
- $\mathbb{E}[||X||^2 \sup_\theta |G_2^{(1)}(X_e'\theta^e)| \, \mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^2 \sup_\theta |G_2^{(2)}(X_e'\theta^e)| \, \mathbb{E}[|Y||X]]$

($\mathcal{M}$-2) For Theorem 2.5, we assume that the following moments are finite:

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

($\mathcal{M}$-3) For Theorem 2.6, we assume that the following moments are finite for some constant $d_0 > 0$ and for all $\theta \in \bar{U}_{d_0}(\theta_0)$:

- $\mathbb{E}[||X||^3(\sup_\tau G_1^{(1)}(X_q'\tau^q))(\sup_{\tilde{\tau}} G_1^{(2)}(X_q'\tilde{\tau}^q))]$
- $\mathbb{E}[||X||^3(\sup_\tau G_1^{(1)}(X_q'\tau^q))(\sup_{\tilde{\tau}} G_2^{(1)}(X_e'\tilde{\tau}^e))]$
- $\mathbb{E}[||X||^3(\sup_\tau G_2(X_e'\tau^e))(\sup_{\tilde{\tau}} G_1^{(2)}(X_q'\tilde{\tau}^q))]$
- $\mathbb{E}[||X||^3(\sup_\tau G_2(X_e'\tau^e))(\sup_{\tilde{\tau}} G_2^{(1)}(X_e'\tilde{\tau}^e))]$
- $\mathbb{E}[||X||^3 \sup_\tau(G_1^{(1)}(X_q'\tau^q))^2]$
- $\mathbb{E}[||X||^3 \sup_\tau(G_2(X_e'\tau^e))^2]$
- $\mathbb{E}[||X||^3 \sup_\tau G_1^{(1)}(X_q'\tau^q)G_2(X_e'\tau^e)]$
- $\mathbb{E}[||X||^5(\sup_\tau G_2^{(1)}(X_e'\tau^e))(\sup_{\tilde{\tau}} G_2^{(2)}(X_e'\tilde{\tau}^e))]$
- $\mathbb{E}[||X||^5(\sup_\tau G_2^{(1)}(X_e'\tau^e))^2]$
- $\mathbb{E}[||X||^4(\sup_\tau G_2^{(1)}(X_e'\tau^e))(\sup_{\tilde{\tau}} G_2^{(2)}(X_e'\tilde{\tau}^e))\mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^3 G_2^{(1)}(X_e'\theta^e)(\sup_\tau G_2^{(1)}(X_e'\tau^e))\mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^3 G_2^{(1)}(X_e'\theta^e)(\sup_\tau G_2^{(2)}(X_e'\tau^e))\mathbb{E}[Y^2|X]]$
- $\mathbb{E}[||X||^3(\sup_\tau G_2^{(1)}(X_e'\tau^e))(\sup_{\tilde{\tau}} G_2^{(2)}(X_e'\tilde{\tau}^e))\mathbb{E}[Y^2|X]]$

($\mathcal{M}$-4) For Theorem 2.7, we assume that the following moments are finite for some constant $d_0 > 0$:

- $\mathbb{E}[|G_1(Y)|]$
- $\mathbb{E}[|a(Y)|]$
- $\mathbb{E}[||X|| \sup_\theta |G_1^{(1)}(X_q'\theta^q)|]$
- $\mathbb{E}[||X||^2 \sup_\theta (G_1^{(1)}(X_q'\theta^q))^2]$
- $\mathbb{E}[||X||^2 \sup_\theta |G_1^{(1)}(X_q'\theta^q)G_2(X_e'\theta^e)|]$
- $\mathbb{E}[||X|| \sup_\theta |G_2(X_e'\theta^e)|]$
- $\mathbb{E}[||X||^2 \sup_\theta |G_2^{(1)}(X_e'\theta^e)|]$
- $\mathbb{E}[||X||^2 \sup_\theta (G_2(X_e'\theta^e))^2]$
- $\mathbb{E}[||X||^4 \sup_\theta (G_2^{(1)}(X_e'\theta^e))^2]$
- $\mathbb{E}[||X|| \sup_\theta |G_2^{(1)}(X_e'\theta^e)|\ \mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^3 \sup_\theta (G_2^{(1)}(X_e'\theta^e))^2\ \mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^2 \sup_\theta (G_2^{(1)}(X_e'\theta^e))^2\ \mathbb{E}[Y^2|X]]$

Appendix B: Proofs

Henceforth, $||v||$ denotes the maximum norm for a vector $v \in \mathbb{R}^k$ 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 $\bar{U}_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| \leq d\}$.

Proof of Theorem 2.4. We apply Theorem 2 from Huber (1967) and show that the function $\psi(Y, X, \theta)$ as given in (4) satisfies the respective assumptions of this theorem. Note that the parameter space Θ is assumed to be compact and thus, we do not have to show condition (B-4) in the notation of Huber (1967). As the product of continuous functions and the indicator function $\mathbb{1}_{\{Y \le X_q'\theta^q\}}$, the function ψ is measurable and regarded as a stochastic process in θ, ψ is separable in the sense of Doob as it is almost surely continuous in θ (Gikhman and Skorokhod, 2004, p.164). This condition assures measurability of the suprema[9] given below and in Lemma B.1.

In oder to show that ψ has a unique root at θ_0, let us first define the sets

$$\begin{aligned} U_\theta &= \big\{\omega \in \Omega \big| X_q(\omega)'\theta^q \neq X_q(\omega)'\theta_0^q\big\}, \qquad \text{and} \\ W_\theta &= \big\{\omega \in \Omega \big| X_q(\omega)'\theta^q = X_q(\omega)'\theta_0^q\big\}, \end{aligned} \tag{35}$$

for all $\theta \in \Theta$ such that $\Omega = W_\theta \cup U_\theta$ and $W_\theta \cap U_\theta = \emptyset$. We first show that $\mathbb{P}(U_\theta) > 0$ for all $\theta \neq \theta_0$. In order to see this, we assume the converse, i.e. let us assume that for a fixed $\theta \neq \theta_0$, it holds that $\mathbb{P}(W_\theta) = \mathbb{P}\big(X_q'\theta^q = X_q'\theta_0^q\big) = 1$, which implies that

$$(\theta^q - \theta_0^q)' \, \mathbb{E}[X_q X_q'] \, (\theta^q - \theta_0^q) = \mathbb{E}\big[\big(X_q'\theta^q - X_q'\theta_0^q\big)^2\big] = 0. \tag{36}$$

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

The quantity

$$\begin{aligned} \lambda_1(\theta) &= \mathbb{E}\big[\psi_1(Y, X, \theta)\big] \\ &= \mathbb{E}\Big[X_q\big(\alpha G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)\big(F_{Y|X}(X_q'\theta^q) - F_{Y|X}(X_q'\theta_0^q)\big)\Big] \end{aligned}$$

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

$$\begin{aligned} \lambda_1(\theta)'(\theta^q - \theta_0^q) = \mathbb{E}\Big[&\big(G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)\big(X_q'\theta^q - X_q'\theta_0^q\big)\big(F_{Y|X}(X_q'\theta^q) \\ &- F_{Y|X}(X_q'\theta_0^q)\big)\mathbb{1}_{\{\omega \in W_\theta\}}\Big] \\ + \mathbb{E}\Big[&\big(G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)\big(X_q'\theta^q - X_q'\theta_0^q\big)\big(F_{Y|X}(X_q'\theta^q) \\ &- F_{Y|X}(X_q'\theta_0^q)\big)\mathbb{1}_{\{\omega \in U_\theta\}}\Big]. \end{aligned}$$

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

[9] Many other authors such as e.g. Newey and McFadden (1994); Andrews (1994); van der Vaart (1998) rely on outer probability in order to avoid these measurability issues.

neighborhood of $X_q'\theta_0^q$ and thus

$$\big(X_q'\theta^q - X_q'\theta_0^q\big)\big(F_{Y|X}(X_q'\theta^q) - F_{Y|X}(X_q'\theta_0^q)\big) > 0 \tag{37}$$

for all $\omega \in U_\theta$. Furthermore, since $\alpha G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e) > 0$ for all $\theta \in \Theta$ and $\mathbb{P}(U_\theta) > 0$, we get that

$$\begin{aligned}\lambda_1(\theta)'(\theta^q - \theta_0^q) = \mathbb{E}\Big[&\big(G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)\big(X_q'\theta^q - X_q'\theta_0^q\big)\big(F_{Y|X}(X_q'\theta^q) \\ &- F_{Y|X}(X_q'\theta_0^q)\big)\mathbb{1}_{\{\omega \in U_\theta\}}\Big]\end{aligned}$$

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

$$\begin{aligned}\lambda_2(\theta) = \mathbb{E}\Big[X_e G_2^{(1)}(X_e'\theta^e)\Big(&X_q'\theta^q\,\big(F_{Y|X}(X_q'\theta^q) - \alpha\big)/\alpha + X_e'\theta^e \\ &- 1/\alpha\,\mathbb{E}\big[Y\mathbb{1}_{\{Y \le X_q'\theta^q\}}\big|X\big]\Big)\Big].\end{aligned} \tag{38}$$

Assuming that $\theta^q = \theta_0^q$, which results from $\lambda_1(\theta) = 0$, we get that $F_{Y|X}(X_q'\theta^q) = F_{Y|X}(X_q'\theta_0^q) = \alpha$ and $1/\alpha\,\mathbb{E}\big[Y\mathbb{1}_{\{Y \le X_q'\theta_0^q\}}\big|X\big] = X_e'\theta_0^e$. Consequently, (38) simplifies to $\mathbb{E}\big[(X_e X_e')G_2^{(1)}(X_e'\theta^e)\big]\big(\theta^e - \theta_0^e\big)$ and as $\mathbb{E}\big[(X_e X_e')\big]$ has full rank by assumption, applying Lemma B.2 yields that the matrix $\mathbb{E}\big[(X_e X_e')G_2^{(1)}(X_e'\theta^e)\big]$ 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. □

Proof of Theorem 2.5. For this proof, we apply Theorem 5.7 from van der Vaart (1998) and show that the respective assumptions of this theorem hold. As in the proof of Theorem 2.6, we can conclude measurability of the suprema since the process ρ is continuous and consequently separable in the sense of Doob. Thus, we do not have to rely on outer probability measures such as in van der Vaart (1998). We start by showing uniform convergence in probability of the empirical mean of the objective function 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 $d(Y, X) \ge |\rho(Y, X, \theta)|$ for all $\theta \in \Theta$ with $\mathbb{E}\big[d(Y, X)\big] < \infty$. We define

$$\begin{aligned}d(Y, X) = &\sup_{\theta\in\Theta}\big|G_1(X_q'\theta^q) + 1/\alpha\,G_2(X_e'\theta^e)(X_q'\theta^q - Y)\big| + \big|G_1(Y)\big| \\ &+ \sup_{\theta\in\Theta}\big|G_2(X_e'\theta^e)\big(X_e'\theta^e - X_q'\theta^q\big)\big| + \sup_{\theta\in\Theta}\big|\mathcal{G}_2(X_e'\theta^e)\big| + \big|\alpha G_1(Y) + a(Y)\big|\end{aligned} \tag{39}$$

and it holds that $d(Y, X) \ge \big|\rho(Y, X, \theta)\big|$ for all $\theta \in \Theta$ and consequently, we can conclude uniform convergence in probability.

We now 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

$$U_\theta = \big\{\omega \in \Omega \big| X_q(\omega)'\theta^q \neq X_q(\omega)'\theta_0^q \quad \text{or} \quad X_e(\omega)'\theta^e \neq X_e(\omega)'\theta_0^e\big\} \quad \text{and} \quad (40)$$

$$W_\theta = \big\{\omega \in \Omega \big| X_q(\omega)'\theta^q = X_q(\omega)'\theta_0^q \quad \text{and} \quad X_e(\omega)'\theta^e = X_e(\omega)'\theta_0^e\big\}, \qquad (41)$$

such that $\Omega = U_\theta \cup W_\theta$ and $U_\theta \cap W_\theta = \emptyset$. We first show that $\mathbb{P}(U_\theta) > 0$ for all $\theta \neq \theta_0$. In order to see this, we assume the converse, i.e. we assume that $\mathbb{P}(W_\theta) = 1$, which implies that $(\theta^q - \theta_0^q)'\,\mathbb{E}[X_q X_q']\,(\theta^q - \theta_0^q) = \mathbb{E}\big[\big(X_q'\theta^q - X_q'\theta_0^q\big)^2\big] = 0$, since $\mathbb{P}\big(X_q'\theta^q = X_q\theta_0^q\big) = 1$ and equivalently $(\theta^e - \theta_0^e)'\mathbb{E}[X_e X_e'](\theta^e - \theta_0^e) = 0$. 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 matrices $\mathbb{E}[X_q X_q']$ and $\mathbb{E}[X_e X_e']$ are positive definite and it follows that $\mathbb{P}(U_\theta) > 0$.

From the joint elicitability property of the quantile and ES of Fissler and Ziegel (2016), Corollary 5.5 we get that for all $x \in \mathbb{R}^k$, $x = (x_q, x_e)$ such that $x_q'\theta^q \neq x_q'\theta_0^q$ or $x_e'\theta^e \neq x_e'\theta_0^e$, it holds that

$$\mathbb{E}\big[\rho(Y,X,\theta_0)\big|X = x\big] < \mathbb{E}\big[\rho(Y,X,\theta)\big|X = x\big], \qquad (42)$$

since the distribution of Y given X has a finite first moment and a unique α-quantile. Thus, for all $\omega \in U_\theta$,

$$\mathbb{E}\big[\rho(Y,X,\theta_0)\big|X\big](\omega) < \mathbb{E}\big[\rho(Y,X,\theta)\big|X\big](\omega). \qquad (43)$$

We now define the random variable

$$h(X,\theta,\theta_0)(\omega) = \mathbb{E}\big[\rho(Y,X,\theta_0)\big|X\big](\omega) - \mathbb{E}\big[\rho(Y,X,\theta)\big|X\big](\omega), \qquad (44)$$

and (43) implies that $h\big(X,\theta,\theta_0\big)(\omega) < 0$ for all $\omega \in U_\theta$. Since $\mathbb{P}(U_\theta) > 0$, this implies that $\mathbb{E}\left[h(X,\theta,\theta_0)\mathbb{1}_{\{\omega \in U_\theta\}}\right] < 0$. Furthermore, for all $\omega \in W_\theta$, it obviously holds that $h(X,\theta,\theta_0)(\omega) = 0$ and consequently $\mathbb{E}\left[h(X,\theta,\theta_0)\mathbb{1}_{\{\omega \in W_\theta\}}\right] = 0$. Thus, we get that

$$\mathbb{E}\big[h(X,\theta,\theta_0)\big] = \mathbb{E}\left[h(X,\theta,\theta_0)\mathbb{1}_{\{\omega \in U_\theta\}}\right] + \mathbb{E}\left[h(X,\theta,\theta_0)\mathbb{1}_{\{\omega \in W_\theta\}}\right] < 0 \qquad (45)$$

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 $\theta = \theta_0$. As we define $\hat{\theta}_{\rho,n} = \operatorname{argmin}_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \theta)$ in (3), it obviously holds that $\frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \hat{\theta}_{\rho,n}) \leq \frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \theta_0) + o_P(1)$ which concludes this proof. □

Proof of Theorem 2.6. We apply Theorem 3 of Huber (1967) for the ψ-function as given in (4) and show the respective assumptions of this theorem. Consistency of the Z-estimator is shown in Theorem 2.4. For the measureability and separability of the ψ function, we refer to the proof of Theorem 2.4. It is already shown in the proof of Theorem 2.4 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 $\mathbb{E}\big[||\psi(Y,X,\theta_0)||^2\big] < \infty$, 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 $C = \mathbb{E}\big[\psi(Y,X,\theta_0)\,\psi(Y,X,\theta_0)'\big]$ and

$$\Lambda = \frac{\partial\lambda(\theta)}{\partial\theta'}\bigg|_{\theta=\theta_0} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix} = \begin{pmatrix}\frac{\partial\lambda_1(\theta)}{\partial\theta^{q\prime}}\Big|_{\theta_0} & \frac{\partial\lambda_1(\theta)}{\partial\theta^{e\prime}}\Big|_{\theta_0}\\ \frac{\partial\lambda_2(\theta)}{\partial\theta^{q\prime}}\Big|_{\theta_0} & \frac{\partial\lambda_2(\theta)}{\partial\theta^{e\prime}}\Big|_{\theta_0}\end{pmatrix}. \tag{46}$$

Straightforward calculations yield the matrix C as given in (10)–(12). For the computation of Λ, we first notice that the function

$$\begin{aligned}&\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\\ &= \begin{pmatrix} X_q\big(G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)\big(F_{Y|X}(X_q'\theta^q)-\alpha\big)\\ X_e G_2^{(1)}(X_e'\theta^e)\Big(X_e'\theta^e - X_q'\theta^q + \frac{1}{\alpha}\mathbb{E}\big[(X_q'\theta^q - Y)\mathbb{1}_{\{Y\le X_q'\theta^q\}}\big|X\big]\Big)\end{pmatrix}\end{aligned} \tag{47}$$

is continuously differentiable for all θ in some neighborhood $U_d(\theta_0)$ around θ_0, since the distribution $F_{Y|X}$ has a density which is strictly positive, continuous and bounded in this area. Let us choose a value $\tilde\theta \in U_d(\theta_0)$ such that $X_q'\tilde\theta^q \le X_q'\theta^q$. Then,

$$\begin{aligned}\frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X_q'\theta^q\}}\big|X\big] &= \frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X_q'\tilde\theta^q\}}\big|X\big] + \frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{X_q'\tilde\theta^q< Y\le X_q'\theta^q\}}\big|X\big]\\ &= \frac{\partial}{\partial\theta^q}\int_{X_q'\tilde\theta^q}^{X_q'\theta^q} y f_{Y|X}(y)\mathrm{d}y = X_q(X_q'\theta^q)f_{Y|X}(X_q'\theta^q).\end{aligned} \tag{48}$$

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

$$\begin{aligned}\frac{\partial}{\partial\theta^{q\prime}}\mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] &= (X_qX_q')\Big[\big(G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)/\alpha\big)f_{Y|X}(X_q'\theta^q)\\ &\qquad + G_1^{(2)}(X_q'\theta^q)\big(F_{Y|X}(X_q'\theta^q)-\alpha\big)\Big],\\ \frac{\partial}{\partial\theta^{e\prime}}\mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] &= \frac{\partial}{\partial\theta^q}\mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big]'\\ &= (X_qX_e')G_2^{(1)}(X_e'\theta^e)\frac{F_{Y|X}(X_q'\theta^q)-\alpha}{\alpha},\\ \frac{\partial}{\partial\theta^e}\mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big] &= (X_eX_e')G_2^{(1)}(X_e'\theta^e)\\ &\quad + 1/\alpha\,(X_eX_e')G_2^{(2)}(X_e'\theta^e)\Big[X_q'\theta^q\big(F_{Y|X}(X_q'\theta^q)-\alpha\big)\\ &\qquad + \alpha(X_e'\theta^e) - \mathbb{E}\big[Y\mathbb{1}_{\{Y\le X_q'\theta^q\}}\big|X\big]\Big]\end{aligned}$$

In order to conclude that $\frac{\partial}{\partial\theta}\mathbb{E}\big[\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\big] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\right]$, we apply a measure-theoretical version of the Leibniz integration rule, which 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

$$d(Y,X) = \sup_{\theta \in U_d(\theta_0)} \Big|\Big| 1/\alpha \left(X_q X_q'\right) \Big[\big(\alpha G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)\big) f_{Y|X}(X_q'\theta^q) + \; G_1^{(2)}(X_q'\theta^q)\big(F_{Y|X}(X_q'\theta^q) - \alpha\big)\Big]\Big|\Big|,$$

which has finite expectation by the Moment Conditions ($\mathcal{M}$-3). The other two terms follow the same reasoning. Inserting $\theta = \theta_0$ eventually shows (8) and (9). $\square$

Proof of Theorem 2.7. For this proof, we apply Theorem 5.23 from van der Vaart (1998) and show that the respective assumptions of this theorem hold. Theorem 2.5 shows consistency of the M-estimator. 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 $Y = X_q'\theta^q$, which is a nullset with respect to the joint distribution of Y and X and for all $\theta \in \Theta$ such that $Y \neq X_q'\theta^q$, its derivative is given by $\psi(Y,X,\theta)$. Local Lipschitz continuity with square-integrable Lipschitz-constant follows from Lemma B.5. We have already seen in the proof of Theorem 2.5 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. The condition $\frac{1}{n}\sum_{i=1}^n \rho(Y_i,X_i,\hat{\theta}_{\rho,n}) \leq \inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \rho(Y_i,X_i,\theta) + o_P(n^{-1})$ is obviously fulfilled as the definition of the M-estimator in (3) implies that $\frac{1}{n}\sum_{i=1}^n \rho(Y_i,X_i,\hat{\theta}_{\rho,n}) = \inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i,\theta)$ as Θ is compact. 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 $X_q'\theta_0^q$, which is strictly positive, continuous and bounded. Therefore, by the same arguments as in (48), we get that $\frac{\partial}{\partial\theta^q}\mathbb{E}\big[G_1(Y)\mathbb{1}_{\{Y\leq X_q'\theta^q\}}\big|X\big] = X_q G_1(X_q'\theta^q) f_{Y|X}(X_q'\theta^q)$. Thus, straight-forward calculations yield that for all $\theta \in U_d(\theta_0)$, it holds that $\frac{\partial}{\partial\theta}\mathbb{E}\big[\rho(Y,X,\theta)\big|X\big] = \mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]$ and by applying the Leibniz integration rule such as in the proof of Theorem 2.6, we finally get that

$$\frac{\partial}{\partial\theta}\mathbb{E}\big[\rho(Y,X,\theta)\big] = \mathbb{E}\big[\psi(Y,X,\theta)\big]. \tag{49}$$

Consequently, the asymptotic covariance matrix equals the one given in Theorem 2.6. $\square$

Lemma B.1. Let

$$u(Y,X,\theta,d) = \sup_{\tau\in\bar{U}_d(\theta)} \big|\big|\psi(Y,X,\tau) - \psi(Y,X,\theta)\big|\big| \tag{50}$$

and assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-1) in Appendix A hold. Then, there are strictly positive real numbers b and d_0, such that

$$\mathbb{E}\big[u(Y,X,\theta,d)\big] \le b\cdot d \quad \text{for} \quad ||\theta-\theta_0|| + d \le d_0, \tag{51}$$

and for all $d \ge 0$.

Proof of Lemma B.1. For measurability of the suprema, we refer to the proof of Theorem 2.4. Let in the following $d > 0$ and $\theta \in \Theta$ such that $||\theta-\theta_0|| + d \le d_0$. We first notice that for some fixed $X_q \in \mathbb{R}^k$ and for all $\tau \in \bar{U}_d(\theta)$, it holds that

$$\left|\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}}\right| \le \mathbb{1}_{\{X_q'\theta_-^q \le Y \le X_q'\theta_+^q\}} \tag{52}$$

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

$$\sup_{\tau\in\bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}}\right| \le \mathbb{1}_{\{X_q'\theta_-^q \le Y \le X_q'\theta_+^q\}} \tag{53}$$

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_q and θ, however they are independent of Y. Consequently, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}}\right| \,\middle|\, X\right] \le \mathbb{E}\left[\mathbb{1}_{\{X_q'\theta_-^q \le Y \le X_q'\theta_+^q\}} \,\middle|\, X\right] \\
&= F_{Y|X}\big(X_q'\theta_+^q\big) - F_{Y|X}\big(X_q'\theta_-^q\big) = f_{Y|X}(X_q'\tilde{\theta}^q)\big(X_q'\theta_+^q - X_q'\theta_-^q\big) \\
&\le 2||X|| \cdot \sup_{\tau\in\bar{U}_d(\theta)} f_{Y|X}(X_q'\tau^q)\cdot d,
\end{aligned} \tag{54}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)} \big|\big|\psi_1(Y,X,\theta) - \psi_1(Y,X,\tau)\big|\big|\right] \\
&\le \mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)} \left|\left|X_q\left(G_1^{(1)}(X_q'\theta^q) - G_1^{(1)}(X_q'\tau^q) + \frac{G_2(X_e'\theta^e) - G_2(X_e'\tau^e)}{\alpha}\right)\right|\right|\right] \\
&\qquad + \mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)} \left|\left|X_q\left(G_1^{(1)}(X_q'\tau^q) + \frac{G_2(X_e'\tau^e)}{\alpha}\right)\right|\right| \right.\\
&\qquad \left.\cdot\,\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}}\right| \,\middle|\, X\right]\right].
\end{aligned} \tag{55}$$

The first term in (55) is $\mathcal{O}(d)$ since $G_1^{(1)}(X_q'\theta^q)$ and $G_2(X_e'\theta^e)$ are continuously differentiable functions w.r.t θ and thus, by the mean value theorem we get that

$$\sup_{\tau\in\bar{U}_d(\theta)}\left|G_1^{(1)}(X_q'\theta^q)-G_1^{(1)}(X_q'\tau^q)\right| \leq \sup_{\tilde{\tau}\in\bar{U}_d(\theta)}\left|\left|X_q G_1^{(2)}(X_q'\tilde{\tau}^q)\right|\right|\cdot \sup_{\tau\in\bar{U}_d(\theta)}\left|\left|\theta^q-\tau^q\right|\right| \\ \leq \sup_{\tilde{\tau}\in\bar{U}_d(\theta)}\left|\left|X_q G_1^{(2)}(X_q'\tilde{\tau}^q)\right|\right|\cdot d, \tag{56}$$

and the respective moments are finite by assumption. The same arguments hold for the function G_2. For the second term in (55), we apply (54) and thus get that

$$\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|\left|X_q\left(G_1^{(1)}(X_q'\tau^q)+\frac{G_2(X_e'\tau^e)}{\alpha}\right)\right|\right| \right. \\ \left. \times\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|\mathbb{1}_{\{Y\leq X_q'\theta^q\}}-\mathbb{1}_{\{Y\leq X_q'\tau^q\}}\right|\,\middle|\, X\right]\right] \\ \leq \mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left|\left|X_q\left(G_1^{(1)}(X_q'\tau^q)+\frac{G_2(X_e'\tau^e)}{\alpha}\right)\right|\right| \left|\left|X_q\right|\right|\cdot\sup_{\tau\in\bar{U}_d(\theta)} f_{Y|X}(X_q'\tau^q)\right]\cdot d. \tag{57}$$

Since the density $f_{Y|X}$ is bounded in a neighborhood of $X_q'\theta_0^q$ 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

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

The first, third and fifth term are linearly bounded by (56) since the functions $(X_e'\theta^e-X_q'\theta^q)G_2^{(1)}(X_e'\theta^e)$ and $(X_q'\theta^q)G_2^{(1)}(X_e'\theta^e)$ and $G_2^{(1)}(X_e'\theta^e)$ are continuously differentiable. For the second term, we use the arguments from (54). For the fourth term, we use similar arguments as in (54), 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

$$\begin{aligned}
&\mathbb{E}\left[\left|\left|\frac{X_e G_2^{(1)}(X_e'\theta^e)}{\alpha}\right|\right| \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|Y\left(\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}}\right)\right| \Big| X\right]\right] \\
\le\ &\mathbb{E}\left[\left|\left|\frac{X_e G_2^{(1)}(X_e'\theta^e)}{\alpha}\right|\right| \mathbb{E}\left[|Y|\ \mathbb{1}_{\{X_q'\theta^q_- \le Y \le X_q'\theta^q_+\}} \Big| X\right]\right] \\
=\ &\mathbb{E}\left[\left|\left|\frac{X_e G_2^{(1)}(X_e'\theta^e)}{\alpha}\right|\right| \int_{X_q'\theta^q_-}^{X_q'\theta^q_+} |y| f_{Y|X}(y)\mathrm{d}y\right] \\
\le\ &\mathbb{E}\left[\left|\left|\frac{X_e G_2^{(1)}(X_e'\theta^e)}{\alpha}\right|\right| |X_q'\tilde{\theta}^q| f_{Y|X}(X_q'\tilde{\theta}^q)\big(X_q'\theta^q_+ - X_q'\theta^q_-\big)\right] \\
\le\ &\frac{2}{\alpha}\mathbb{E}\left[G_2^{(1)}(X_e'\theta^e)\big|\big|X\big|\big|^2 \sup_{\tau \in \bar{U}_d(\theta)} |X_q'\tau^q| f_{Y|X}(X_q'\tau^q)\right] \cdot d = \mathcal{O}(d),
\end{aligned} \tag{58}$$

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

Lemma B.2. Let the random variable $X \in \mathbb{R}^k$ 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 $g : \mathbb{R}^k \times \tilde{\Theta} \to \mathbb{R}$ be a strictly positive function. Then, the matrix

$$\mathbb{E}\big[(XX')g(X,\theta)\big] \tag{59}$$

is also positive definite.

Proof of Lemma B.2. 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 $z \neq 0$,

$$z'\mathbb{E}\big[(XX')g(X,\theta)\big]z = \mathbb{E}\left[\left(X'z\sqrt{g(X,\theta)}\right)^2\right] > 0. \tag{60}$$

This positivity statement holds since $\big(X'z\sqrt{g(X,\theta)}\big)^2$ 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. □

Lemma B.3. Assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-3) in Appendix A hold. Then, for

$$\lambda(\theta) = \mathbb{E}\big[\psi(Y,X,\theta)\big], \tag{61}$$

there are strictly positive numbers a, d_0, such that

$$||\lambda(\theta)|| \ge a \cdot ||\theta - \theta_0|| \qquad \text{for} \quad ||\theta - \theta_0|| \le d_0. \tag{62}$$

Proof of Lemma B.3. Let $d_0 > 0$ and let $||\theta - \theta_0|| \leq d_0$. Then, applying the mean value theorem, we get that

$$\lambda_1(\theta) = \frac{1}{\alpha}\mathbb{E}\left[(X_q X_q')\big(\alpha G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)\big) f_{Y|X}(X_q'\tilde{\theta}^q)\right](\theta^q - \theta_0^q) \quad (63)$$

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

$$\begin{aligned}\lambda_2(\theta) = \mathbb{E}&\left[X_e \frac{G_2^{(1)}(X_e'\theta^e) f_{Y|X}(X_q'\tilde{\theta}^q)}{\alpha}\big[X_q'(\theta^q - \theta_0^q)\big]\big[X_q'(\tilde{\theta^q} - \theta^q)\big]\right] \\ &+ \mathbb{E}\big[(X_e X_e')G_2^{(1)}(X_e'\theta^e)\big](\theta^e - \theta_0^e),\end{aligned} \quad (64)$$

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

$$A(\theta) := \mathbb{E}\left[(X_q X_q')\frac{\big(\alpha G_1^{(1)}(X_q'\theta^q) + G_2(X_e'\theta^e)\big)}{\alpha} f_{Y|X}(X_q'\tilde{\theta}^q)\right] \quad (65)$$

exists and has full rank for all $\theta \in \Theta$ by Lemma B.2 and is obviously symmetric, A has strictly positive real Eigenvalues $\gamma_1(\theta), \ldots, \gamma_k(\theta)$ with minimum $\gamma_{(1)}(\theta)$ and we thus get that[10]

$$\begin{aligned}||\lambda(\theta)|| \geq ||\lambda_1(\theta)|| = ||A(\theta)(\theta^q - \theta_0^q)|| &\geq \gamma_{(1)}(\theta) \cdot ||\theta^q - \theta_0^q|| \quad (66)\\ &\geq \left(\inf_{||\theta - \theta_0|| \leq d_0} \gamma_{(1)}(\theta)\right) \cdot ||\theta^q - \theta_0^q|| = c_1\, ||\theta - \theta_0||. \quad (67)\end{aligned}$$

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[11], 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 (64), we define the vector

$$b(\theta) := \mathbb{E}\left[X_e \frac{G_2^{(1)}(X_e'\theta^e) f_{Y|X}(X_q'\tilde{\theta}^q)}{\alpha}\big[X_q'(\theta^q - \theta_0^q)\big]\big[X_q'\tilde{\theta}^q - X'\theta^q\big]\right], \quad (68)$$

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

[11]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 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.

and for its l-th component, we get that

$$\begin{aligned}
|b_l(\theta)| &= \left|\sum_{i,j}(\theta_i^q - \theta_{0i}^q)(\tilde{\theta}_j^q - \theta_j^q)\mathbb{E}\left[X_{q,i}X_{q,j}X_{e,l}\frac{G_2^{(1)}(X_e'\theta^e)f_{Y|X}(X_q'\tilde{\theta}^q)}{\alpha}\right]\right| \\
&\le \sum_{i,j}\mathbb{E}\left[\left|X_{q,i}X_{q,j}X_{e,l}\frac{G_2^{(1)}(X_e'\theta^e)f_{Y|X}(X_q'\tilde{\theta}^q)}{\alpha}\right|\right]\cdot|\theta_i^q - \theta_{0i}^q|\cdot|\tilde{\theta}_j^q - \theta_j^q| \\
&\le c_2\sum_{i,j}|\theta_i^q - \theta_{0i}^q|\cdot|\tilde{\theta}_j^q - \theta_j^q| \\
&\le c_2 k^2||\theta - \theta_0||^2,
\end{aligned} \tag{69}$$

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

$$||b(\theta)|| \le c_3||\theta - \theta_0||^2, \tag{70}$$

for some $c_3 > 0$. For $D(\theta) := \mathbb{E}\big[(X_e X_e')G_2^{(1)}(X_e'\theta^e)\big]$, it holds that $||D(\theta)(\theta^e - \theta_0^e)|| \ge c_4||\theta^e - \theta_0^e|| = c_4||\theta - \theta_0||$ for $c_4 > 0$ by the same arguments as in (66). From (69), we can choose d_0 small enough such that

$$2||b(\theta)|| \le 2c_3||\theta - \theta_0||^2 \le c_4||\theta - \theta_0|| \le ||D(\theta)(\theta^e - \theta_0^e)||. \tag{71}$$

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

$$||\lambda(\theta)|| \ge ||\lambda_2(\theta)|| = \big|\big|D(\theta)(\theta^e - \theta_0^e) + b(\theta)\big|\big| \ge \Big|||D(\theta)(\theta^e - \theta_0^e)|| - ||b(\theta)||\Big|. \tag{72}$$

From (71), we can choose d_0 small enough such that $||D(\theta^e - \theta_0^e)|| > 2||b||$ and thus

$$\Big|||D(\theta^e - \theta_0^e)|| - ||b||\Big| = ||D(\theta^e - \theta_0^e)|| - ||b|| \ge \frac{1}{2}||D(\theta^e - \theta_0^e)|| \tag{73}$$

$$\ge \frac{c_4}{2}\,||\theta^e - \theta_0^e|| = \frac{c_4}{2}\,||\theta - \theta_0||. \tag{74}$$

□

Lemma B.4. Let

$$u(Y, X, \theta, d) = \sup_{\tau\in\bar{U}_d(\theta)}\big|\big|\psi(Y, X, \tau) - \psi(Y, X, \theta)\big|\big|. \tag{75}$$

and assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-3) in Appendix A hold. Then, there are strictly positive numbers c and d_0, such that

$$\mathbb{E}\big[u(Y, X, \theta, d)^2\big] \le c\cdot d \quad \text{for} \quad ||\theta - \theta_0|| + d \le d_0, \tag{76}$$

and for all $d \ge 0$.

Proof of Lemma B.4. Let in the following $d > 0$ and $\theta \in \Theta$ such that $||\theta - \theta_0|| + d \le d_0$. It holds that

$$\left(\sup_{\tau \in \bar{U}_d(\theta)} \big|\big|\psi(Y,X,\tau) - \psi(Y,X,\theta)\big|\big| \right)^2 = \sup_{\tau \in \bar{U}_d(\theta)} \big|\big|\psi(Y,X,\tau) - \psi(Y,X,\theta)\big|\big|^2 \tag{77}$$

and consequently, we show that

$$\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \big|\big|\psi_j(Y,X,\tau) - \psi_j(Y,X,\theta)\big|\big|^2 \right] = \mathcal{O}(d) \tag{78}$$

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

For the first squared component, we get that

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

where we apply (54) for the second summand. The remaining two summands can be bounded linearly by the arguments given in (56) since $G_1^{(1)}$ and G_2 are continuously differentiable functions and the respective moments are finite.

For the second component of ψ, we get that

$$\begin{aligned}
&\big|\big|\psi_2(Y,X,\tau) - \psi_2(Y,X,\theta)\big|\big| \\
&\quad \le \big|\big| X_e(X_e'\theta^e - X_q'\theta^q) G_2^{(1)}(X_e'\theta^e) - X_e(X_e'\tau^e - X_q'\tau^q) G_2^{(1)}(X_e'\tau^e) \big|\big| \\
&\qquad + \left|\left| \frac{X_e G_2^{(1)}(X_e'\theta^e) X_q'\theta^q}{\alpha} \left(\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}} \right) \right|\right| \\
&\qquad + \left|\left| \mathbb{1}_{\{Y \le X_q'\tau^q\}} \left(\frac{X_e G_2^{(1)}(X_e'\theta^e) X_q'\theta^q}{\alpha} - \frac{X_e G_2^{(1)}(X_e'\tau^e) X_q'\tau^q}{\alpha} \right) \right|\right|
\end{aligned} \tag{79}$$

$$
\begin{aligned}
&+ \left|\left| \frac{X_e G_2^{(1)}(X_e'\theta^e)}{\alpha} Y \left(\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}} \right) \right|\right| \\
&+ \left|\left| \frac{Y \mathbb{1}_{\{Y \le X_q'\tau^q\}}}{\alpha} \left(X_e G_2^{(1)}(X_e'\theta^e) - X_e G_2^{(1)}(X_e'\tau^e) \right) \right|\right| \\
&= \text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)} + \text{(v)}.
\end{aligned}
$$

Thus, in order to evaluate $\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\left|\psi_2(Y,X,\tau) - \psi_2(Y,X,\theta)\right|\right|^2\right]$, we have to consider all the cross products out of the five summands in (79). Since the techniques applied are very similar, we only show details for two of the cross products.

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

by (56) since $G_2^{(1)}$ is continuously differentiable.

The following crossproducts can be bounded analogously by bounding the indicator functions and by applying the mean value theorem as in (56): (i)^2, (iii)^2, (v)^2, (i) $\cdot$ (iii), (i) $\cdot$ (iv), (i) $\cdot$ (v), (ii) $\cdot$ (iv), (ii) $\cdot$ (v), (iii) $\cdot$ (iv), (iii) $\cdot$ (v) and (iv) $\cdot$ (v).

A second type of technique, similar to the arguments in (58) arises in the cases (ii)^2, (iv)^2 and (ii) $\cdot$ (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

$$
\begin{aligned}
\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \text{(iv)}^2 \right] \le \mathbb{E}\Bigg[&\left|\left| \frac{X_e G_2^{(1)}(X_e'\theta^e)}{\alpha} \right|\right|^2 \\
&\times \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left| Y \left(\mathbb{1}_{\{Y \le X_q'\theta^q\}} - \mathbb{1}_{\{Y \le X_q'\tau^q\}} \right) \right|^2 \middle| X \right] \Bigg]
\end{aligned}
$$

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

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

Lemma B.5. Assume that Assumption 2.1, Assumption 2.2 and the Moment Conditions ($\mathcal{M}$-4) in Appendix A hold. Then, the function $\rho(Y, X, \theta)$, given in (2) is locally Lipschitz continuous in θ in the sense that for all $\theta_1, \theta_2 \in U_d(\theta_0)$ in some neighborhood of θ_0, it holds that

$$
\big|\rho(Y, X, \theta_1) - \rho(Y, X, \theta_2)\big| \leq K(Y, X) \cdot \big|\big|\theta_1 - \theta_2\big|\big|, \tag{80}
$$

where $\mathbb{E}\big[K(Y, X)^2\big] < \infty$.

Proof. We start the proof by splitting the ρ function into two parts,

$$
\rho(Y, X, \theta) = \rho_1(Y, X, \theta) + \rho_2(Y, X, \theta), \tag{81}
$$

where

$$
\rho_1(Y, X, \theta) = \mathbb{1}_{\{Y \leq X_q'\theta^q\}}\Big(G_1(X_q'\theta^q) - G_1(Y) + \frac{1}{\alpha}G_2(X_e'\theta^e)(X_q'\theta^q - Y)\Big), \tag{82}
$$

$$
\rho_2(Y, X, \theta) = G_2(X_e'\theta^e)\big(X_e'\theta^e - X_q'\theta^q\big) - \mathcal{G}_2(X_e'\theta^e) - \alpha G_1(X_q'\theta^q) + a(Y). \tag{83}
$$

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

$$
\begin{aligned}
&\big|\rho_2(Y, X, \theta_1) - \rho_2(Y, X, \theta_2)\big| \\
&\leq \big|\big|\theta_1 - \theta_2\big|\big| \cdot \sup_{\theta \in U_d(\theta_0)} \left|\left|\begin{pmatrix} -X_q G_2(X_e'\theta^e) - \alpha X_q G_1^{(1)}(X_q'\theta^q) \\ X_e G_2^{(1)}(X_e'\theta^e)\big(X_e'\theta^e - X_q'\theta^q\big)\end{pmatrix}\right|\right|,
\end{aligned} \tag{84}
$$

with Lipschitz-constant

$$
K(Y, X) = \sup_{\theta \in U_d(\theta_0)} \left|\left|\begin{pmatrix} -X_q G_2(X_e'\theta^e) - \alpha X_q G_1^{(1)}(X_q'\theta^q) \\ X_e G_2^{(1)}(X_e'\theta^e)\big(X_e'\theta^e - X_q'\theta^q\big)\end{pmatrix}\right|\right|, \tag{85}
$$

which is square-integrable by the moment conditions ($\mathcal{M}$-4).

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

$$\rho_1(Y, X, \theta_1) = \rho_1(Y, X, \theta_2) = 0, \tag{86}$$

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

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

$$\rho_1(Y, X, \theta) = G_1(X_q'\theta^q) - G_1(Y) + \frac{1}{\alpha} G_2(X_e'\theta^e)(X_q'\theta^q - Y), \tag{87}$$

which is a continuously differentiable function and thus

$$\begin{aligned} &\big|\rho_1(Y, X, \theta_1) - \rho_1(Y, X, \theta_2)\big| \\ &\leq \big|\big|\theta_1 - \theta_2\big|\big| \cdot \sup_{\theta \in U_d(\theta_0)} \left|\left| \begin{pmatrix} X_q G_1^{(1)}(X_q'\theta^q) + \frac{1}{\alpha} X_q G_2(X_e'\theta^e) \\ \frac{1}{\alpha} X_e G_2^{(1)}(X_e'\theta^e)(X_q'\theta^q - Y) \end{pmatrix} \right|\right|. \end{aligned} \tag{88}$$

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

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

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. ☐

Proposition B.6. Let Y be a real-valued random variable with distribution function F, finite first and second moments and a unique α-quantile $q_\alpha = F^{-1}(\alpha)$. Then,

$$\frac{1}{\alpha^2} \int_{-\infty}^{q_\alpha} \int_{-\infty}^{q_\alpha} F(x \wedge y) - F(x)F(y) \mathrm{d}x\mathrm{d}y = \frac{1}{\alpha} \operatorname{Var}(Y|Y \leq q_\alpha) + \frac{1-\alpha}{\alpha}\big(q_\alpha - \xi_\alpha\big)^2, \tag{89}$$

where $\xi_\alpha = \mathbb{E}\big[Y\big|Y \leq q_\alpha\big]$ denotes the α-ES of Y.

Proof. We first notice that for a distribution F with finite second moment und unique α-quantile, it holds that

$$\mathbb{E}\big[Y\big|Y \leq q_\alpha\big] = -\frac{1}{\alpha} \int_{-\infty}^{q_\alpha} F(x)\mathrm{d}x + q_\alpha \qquad \text{and} \tag{90}$$

$$\mathbb{E}\big[Y^2\big|Y \le q_\alpha\big] = -\frac{2}{\alpha}\int_{-\infty}^{q_\alpha} xF(x)\mathrm{d}x + q_\alpha^2, \tag{91}$$

which can be obtained by using the identity

$$Y\mathbb{1}_{\{Y\le q_\alpha\}} = \mathbb{1}_{\{Y\le q_\alpha\}}\left(\int_0^\infty \mathbb{1}_{\{Y>t\}}\,\mathrm{d}t - \int_{-\infty}^0 \mathbb{1}_{\{Y\le t\}}\,\mathrm{d}t\right) \tag{92}$$

and by taking expectations on both sides. By applying (90), we get that

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x)F(y)\mathrm{d}x\mathrm{d}y &= \left(\int_{-\infty}^{q_\alpha} F(x)\mathrm{d}x\right)^2 \\ &= \big(\alpha q_\alpha - \alpha\mathbb{E}\big[Y\big|Y\le q_\alpha\big]\big)^2 = \alpha^2\big(q_\alpha - \xi_\alpha\big)^2.\end{aligned} \tag{93}$$

Furthermore, notice that

$$\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{q_\alpha}\int_{y}^{q_\alpha} F(y)\mathrm{d}x\mathrm{d}y, \tag{94}$$

and by rearranging the order of integration for the first term in (94), we get that

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\,\mathrm{d}x\mathrm{d}y &= \iint\limits_{\{(x,y):\, y\le q_\alpha,\, x\le y\}} F(x)\,\mathrm{d}x\mathrm{d}y = \iint\limits_{\{(x,y):\, x\le q_\alpha,\, y\ge x\}} F(x)\,\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{q_\alpha}\int_{x}^{q_\alpha} F(x)\,\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{q_\alpha} F(x)(q_\alpha - x)\,\mathrm{d}x.\end{aligned} \tag{95}$$

Thus, by first using (94) and (95) and by plugging in (90) and (93), we obtain

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y)\mathrm{d}x\mathrm{d}y &= 2\int_{-\infty}^{q_\alpha}\int_{y}^{q_\alpha} F(y)\,\mathrm{d}x\mathrm{d}y \\ &= 2\int_{-\infty}^{q_\alpha} F(y)(q_\alpha - y)\,\mathrm{d}y \\ &= 2q_\alpha\int_{-\infty}^{q_\alpha} F(y)\,\mathrm{d}y - 2\int_{-\infty}^{q_\alpha} yF(y)\,\mathrm{d}y \\ &= 2q_\alpha\big(\alpha q_\alpha - \alpha\xi_\alpha\big) + \alpha\mathbb{E}\big[Y^2\big|Y\le q_\alpha\big] - \alpha q_\alpha^2 \\ &= \alpha\mathbb{E}\big[Y^2\big|Y\le q_\alpha\big] + \alpha q_\alpha^2 - 2\alpha q_\alpha\xi_\alpha.\end{aligned} \tag{96}$$

Eventually, using (93) and (96), straight-forward calculations yield that

$$\frac{1}{\alpha^2}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x\wedge y) - F(x)F(y)\mathrm{d}x\mathrm{d}y = \frac{1}{\alpha}\operatorname{Var}(Y|Y\le q_\alpha) + \frac{1-\alpha}{\alpha}\big(q_\alpha - \xi_\alpha\big)^2, \tag{97}$$

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

$$\{x|\psi(x,\theta) \in F,\ \forall \theta \in G\} \qquad \text{and} \tag{98}$$

$$\{x|\psi(x,\theta) \in F,\ \forall \theta \in G \cap I\} \tag{99}$$

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 (98) and (99) 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 (98) and (99) coincide for all $x \in \Omega \setminus N$, i.e. they differ only by a subset of N. □

References

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–905.

Andrews, D. (1994). Empirical process methods in econometrics. In Engle, R. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 37, pages 2247–2294. Elsevier. MR1315970

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228. MR1850791

Barendse, S. (2017). Interquantile expectation regression. Available at https://ssrn.com/abstract=2937665.

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. (2018). Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and Statistics*, 8:56–77. MR3860596

Bayer, S. and Dimitriadis, T. (2019). Regression based expected shortfall backtesting. arXiv:1801.04112 [q-fin.RM].

Bayer, S. and Dimitriadis, T. (2019). *esreg: Joint Quantile and Expected Shortfall Regression*. R package version 0.4.0, available at https://cran.r-project.org/package=esreg.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327. MR0853051

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. MR2450099

Chao, W., Gerlach, R., and Chen, Q. (2018). A semi-parametric realized joint value-at-risk and expected shortfall regression framework. arXiv:1807.02422 [q-fin.RM].

Chen, S. X. (2008). Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, 6(1):87–107.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, 26(1):271–292.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26. MR0515681

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1:93–125. MR1101317

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B*, 78(3):505–562. MR3506792

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. MR2091566

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. MR3519937

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 and Economic Statistics*, 29(1):150–160. MR2789438

Giacomini, R. and Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, 23:416–431. MR2206011

Gikhman, I. and Skorokhod, A. (2004). *The Theory of Stochastic Processes I*, volume 210 of *Classics in Mathematics*. Springer Berlin Heidelberg. MR2058259

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762. MR2847988

Gourieroux, C. and Monfort, A. (1995). *Statistics and Econometric Models: Volume 1, General Concepts, Estimation, Prediction and Algorithms*. Cambridge University Press.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681–700. MR0740305

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. MR2930045

Hall, P. and Sheather, S. J. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society: Series B*, 50(3):381–391. MR0970974

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497. MR2809377

Heber, G., Lunde, A., Shephard, N., and Sheppard, K. (2009). *Oxford-Man Institute's realized library, version 0.3*. Oxford-Man Institute, University of Oxford.

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. MR0216620

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. MR1311953

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. MR2268657

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310. MR1731491

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, 101(475):980–990. MR2324109

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

Komunjer, I. and Vuong, Q. (2010). Semiparametric efficiency bound in time-series models for conditional quantiles. *Econometric Theory*, 26(02):383–405. MR2600568

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.

Lourenço, H. R., Martin, O. C., and Stützle, T. (2003). Iterated local search. In Glover, F. and Kochenberger, G. A., editors, *Handbook of Metaheuristics*, pages 320–353. Springer US, Boston, MA. MR1975905

Nadarajah, S., Zhang, B., and Chan, S. (2014). Estimation methods for expected shortfall. *Quantitative Finance*, 14(2):271–291. MR3175984

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313. MR3363409

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. MR1315971

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4):1833–1874. MR3743276

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Forthcoming in Journal of Econometrics*.

Taylor, J. W. (2008a). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2):231–252.

Taylor, J. W. (2008b). Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of Financial Econometrics*, 6(3):382–406.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business and Economic Statistics*, 37:1, 121–133. MR3910230

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C. W., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 4, pages 135–196. Elsevier.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. MR1652247

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.

Weber, S. (2006). Distribution invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16(2):419–441. MR2212272

Xiao, Z., Guo, H., and Lam, M. S. (2015). Quantile regression and value at risk. In Lee, C.-F. and Lee, J. C., editors, *Handbook of Financial Econometrics and Statistics*, pages 1143–1167. Springer.

Ziegel, J. F., Krüger, F., Jordan, A., and Fasciati, F. (2019). Robust Forecast Evaluation of Expected Shortfall. *Journal of Financial Econometrics*.

Zwingmann, T. and Holzmann, H. (2016). Asymptotics for the expected shortfall. arXiv:1611.07222 [math.ST].