Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1223–1246
https://doi.org/10.1093/jrsssb/qkad063
Advance access publication 15 June 2023
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Robust estimation and inference for expected shortfall regression with many regressors

Xuming He[1], Kean Ming Tan[2] and Wen-Xin Zhou[3]

[1]Department of Statistics and Data Science, Washington University in St. Louis, St. Louis, MO, USA
[2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA
[3]Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL, USA

Address for correspondence: Wen-Xin Zhou, Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA. Email: wenxinz@uic.edu

Abstract

Expected shortfall (ES), also known as superquantile or conditional value-at-risk, is an important measure in risk analysis and stochastic optimisation and has applications beyond these fields. In finance, it refers to the conditional expected return of an asset given that the return is below some quantile of its distribution. In this paper, we consider a joint regression framework recently proposed to model the quantile and ES of a response variable simultaneously, given a set of covariates. The current state-of-the-art approach to this problem involves minimising a non-differentiable and non-convex joint loss function, which poses numerical challenges and limits its applicability to large-scale data. Motivated by the idea of using Neyman-orthogonal scores to reduce sensitivity to nuisance parameters, we propose a statistically robust and computationally efficient two-step procedure for fitting joint quantile and ES regression models that can handle highly skewed and heavy-tailed data. We establish explicit non-asymptotic bounds on estimation and Gaussian approximation errors that lay the foundation for statistical inference, even with increasing covariate dimensions. Finally, through numerical experiments and two data applications, we demonstrate that our approach well balances robustness, statistical, and numerical efficiencies for expected shortfall regression.

Keywords: expected shortfall, heavy-tailed distribution, Huber loss, Neyman orthogonality, quantile regression

1 Introduction

Expected shortfall (ES), also known as superquantile or conditional value-at-risk (VaR), has been recognised as an important risk measure with versatile applications in finance (Acerbi & Tasche, 2002; Rockafellar & Uryasev, 2002), management science (Ben-Tal & Teboulle, 1986; Du & Escanciano, 2017), operations research (Rockafellar et al., 2014; Rockafellar & Uryasev, 2000), and clinical studies (He et al., 2010). For example, in finance, expected shortfall refers to the expected return of an asset or investment portfolio conditional on the return being below a lower quantile of its distribution, namely its VaR. In their Fundamental Review of the Trading Book (Basel Committee, 2016, 2019), the Basel Committee on Banking Supervision confirmed the replacement of VaR with ES as the standard risk measure in banking and insurance.

Let Y be a real-valued random variable with finite first-order absolute moment, $\mathbb{E}|Y| < \infty$, and let F_Y be its cumulative distribution function (CDF). For any $\alpha \in (0, 1)$, the quantile and ES at level α are defined as

$$Q_\alpha(Y) = F_Y^{-1}(\alpha) = \inf\{y \in \mathbb{R} : F_Y(y) \geq \alpha\} \quad \text{and} \quad \mathrm{ES}_\alpha(Y) = \mathbb{E}\{Y | Y \leq Q_\alpha(Y)\}, \tag{1}$$

respectively. If F_Y is continuous, the α-level ES is equivalently given by (see, e.g. Lemma 2.16 of McNeil et al., 2015)

$$\mathrm{ES}_\alpha(Y) = \frac{1}{\alpha}\int_0^\alpha Q_u(Y)\,\mathrm{d}u. \tag{2}$$

Received: May 15, 2022. **Revised:** May 1, 2023. **Accepted:** May 25, 2023

For instance, in socio-economics applications, Y is the income and $\mathrm{ES}_\alpha(Y)$ can be interpreted as the average income for the sub-population whose income falls below the α-quantile of the entire population. We refer the reader to Chapter 6 of Shapiro et al. (2014) and Rockafellar and Royset (2014) for a thorough discussion of ES and its mathematical properties.

With the increasing focus on ES and its desired properties as a risk measure, it is natural to examine the impact of a p-dimensional explanatory vector X, on the tail behaviour of Y through ES. One motivating example is the Job Training Partnership Act (JTPA), a large publicly funded training programme that provides training for adults with barriers to employment and out-of-school youths. The goal is to examine whether the training programme improves future income for adults with low-income earnings (Bloom et al., 1997), for which quantile regression (QR)-based approaches have been proposed (Abadie et al., 2002; Chernozhukov & Hansen, 2008). For example, the 0.05-quantile of the post-programme income refers to the highest income earning of those who have the 5% lowest income among the entire population, while the 0.05-ES concerns the average income earning within this sub-population and therefore is more scientifically relevant in the JTPA study.

Compared to the substantial body of literature on QR, extant works on ES estimation and inference in the presence of covariates are scarce. We refer the reader to Scaillet (2004), Cai and Wang (2008), Kato (2012), Linton and Xiao (2013), and Martins-Filho et al. (2018) for nonparametric conditional ES estimation, and more recently to Dimitriadis and Bayer (2019), Patton et al. (2019), and Barendse (2020) in the context of (semi-)parametric models. As suggested in Patton et al. (2019), this is partly because regulatory interest in ES as a risk measure is only recent, and also due to the fact that this measure is not *elicitable* (Gneiting, 2011). Let $\mathcal{P}$ be a class of distributions on $\mathbb{R}^d$. We say that a statistical functional $\theta : \mathcal{P} \to \mathcal{D}$ with $D \subseteq \mathbb{R}^p$ ($p \geq 1$) is elicitable relative to the class $\mathcal{P}$ if there exists a loss function $\rho : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ such that $\theta(F) = \mathrm{argmin}_{\theta \in D} \mathbb{E}_{Z\sim F} \rho(Z, \theta)$ for any $F \in \mathcal{P}$. Here, $\mathbb{E}_{Z\sim F}$ means that the expectation is taken with respect to the random variable Z that follows the distribution F. For example, the mean is elicitable using the L_2-loss, and the median is elicitable using the L_1-loss. Although the ES is not elicitable on its own, it is jointly elicitable with the quantile using a class of strictly consistent joint loss functions (Fissler & Ziegel, 2016). Based on this result, Dimitriadis and Bayer (2019) and Patton et al. (2019) proposed a joint regression model for the conditional α-level quantile and ES of Y, given the covariates $X \in \mathbb{R}^p$. In this work, we focus on (conditional) linear joint quantile-ES models:

$$Q_\alpha(Y|X) = X^{\mathrm{T}}\beta^* \quad \text{and} \quad \mathrm{ES}_\alpha(Y|X) = X^{\mathrm{T}}\theta^*. \tag{3}$$

Equivalently, we have $\varepsilon = Y - X^{\mathrm{T}}\beta^*$ and $\xi = Y - X^{\mathrm{T}}\theta^*$, where the conditional α-quantile of ε and the conditional α-level expected shortfall of ε, given $X \in \mathbb{R}^p$, are zero. More generally, one may allow the quantile and the ES models to depend on different covariate vectors X_q and X_e, respectively. In this case, the conditional α-quantile and α-ES of ε and ξ, respectively, given $X = (X_q^{\mathrm{T}}, X_e^{\mathrm{T}})^{\mathrm{T}}$, are assumed to be zero.

To jointly estimate β^* and θ^*, Dimitriadis and Bayer (2019) and Patton et al. (2019) considered an M-estimator, defined as the global minimum of any member of a class of strictly consistent joint loss functions over some compact set (Fissler & Ziegel, 2016). The joint loss function, which will be specified in equation (5), is non-differentiable and non-convex. Dimitriadis and Bayer (2019) employed the derivative-free Nelder–Mead algorithm to minimise the resulting non-convex loss, which is a heuristic search method that may converge to non-stationary points. From a statistical perspective, they further established consistency and asymptotic normality for the global minima. However, from a computational perspective, finding the global minimum of a non-convex function is generally intractable: approximating the global minimum of a k-times continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ to ϵ-accuracy requires at least as many as $(1/\epsilon)^{p/k}$ evaluations (ignoring problem-dependent constants) of the function and its first k-derivatives (Nemirovski & Yudin, 1983). The lack of differentiability makes this problem even more challenging numerically. To mitigate the computational effort, Barendse (2020) proposed a two-step procedure by first estimating the quantile parameters via standard QR, followed by least squares regression with generated response variables. Although computationally efficient, the ensuing

estimator is sensitive to heavy-tailed error distributions due to the use of L_2-loss for fitting possibly highly skewed data in the second step; see Section 3.1 for a rigorous statement.

In this paper, we propose a new two-stage method for robust estimation and inference under a joint quantile and expected shortfall regression model (3), with a particular focus on the latter. Compared to existing approaches, our proposed method is robust against heavy-tailed errors without compromising statistical efficiency under light-tailed distributions. Computationally, our method can be implemented via fast and scalable gradient-based algorithms. The main contributions of this work are summarised as follows:

- Our method builds upon a recent approach to joint quantile and expected shortfall regression via a two-step procedure (Barendse, 2020). However, a general non-asymptotic theory for this approach has yet to be established. To fill this gap, we establish a finite-sample theoretical framework for the two-step ES estimator when the dimension of the model, p, increases with the number of observations, n. Specifically, we provide explicit upper bounds, as a function of (n, p), on the estimation error (under L_2-risk) and (uniform) Gaussian approximation errors; see Online Supplementary Section A. We also construct asymptotically valid (entrywise) confidence intervals for the ES parameters. The main computational effort of this two-step procedure is the QR fit in stage one. Therefore, we recommend using the convolution-smoothed QR method (Fernandes et al., 2021), which can be solved using fast first-order algorithms that are scalable to very large-scale problems (He et al., 2023). Our non-asymptotic theory allows the dimension p to grow with the sample size, which paves the way for analysing series/projection estimators under joint nonparametric quantile-ES models (Belloni et al., 2019) and penalised estimators under high-dimensional sparse models.
- The standard two-step estimator is a least squares estimator (LSE) with generated response variables. As a result, it is sensitive to the tails of the distribution of Y. We propose a robust ES regression method that applies adaptive Huber regression (Zhou et al., 2018) in the second step to address this issue. The resulting estimator achieves sub-Gaussian deviation bounds even when the (conditional) distribution of $Y|X$ only has Pareto-like tails. To achieve a trade-off between bias and robustness, we propose using a diverging robustification parameter $\tau = \tau(n, p) > 0$. In practice, we choose this hyper-parameter using a data-driven mechanism (L. Wang et al., 2021), guided by the non-asymptotic results presented in Section 4 and inspired by the censored equation approach introduced in Hahn et al. (1990). We have also developed efficient algorithms to compute standard and robust two-step ES estimators under additional constraints. These constraints ensure that the fitted ES does not exceed the fitted quantile at each observation. We refer the reader to the Online Supplementary Section D for more details.
- We conduct thorough numerical comparisons between the two-step estimator and the proposed robust variant with the joint M-estimator of Dimitriadis and Bayer (2019) on large synthetic data sets generated from a location-scale model, with both light- and heavy-tailed error distributions. To compute the joint M-estimator, we use the R package `esreg`, which is available at https://cran.r-project.org/package=esreg. To implement the proposed robust two-step procedure, we use a combination of R packages, `quantreg` or `conquer` and `adaHuber`. Our results show that the proposed robust ES regression approach achieves satisfying statistical performance, a higher degree of robustness against heavy-tailed error distributions, and superior computational efficiency and stability. We also demonstrate the effectiveness of our approach through numerical experiments and two real data examples.

In this work, the term 'robustness' specifically refers to the robustness against heavy-tailed distributions, as revealed by non-asymptotic deviation analysis dating back to Catoni (2012). In Catoni (2012)'s study of univariate mean estimation, it was found that while the sample mean has the optimal minimax mean squared error among all mean estimators, its deviation is worse for non-Gaussian samples than for Gaussian ones. Moreover, the worst-case deviation is sub-optimal when the sampling distribution has heavy tails. To be more specific, let $X_1, \ldots, X_n$ be independent copies of X with mean μ and variance $\sigma^2 > 0$. Applying Chebyshev's inequality to the empirical

mean $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ yields a polynomial-type deviation bound: $|\bar{X}_n - \mu| \le \sigma\sqrt{1/(n\delta)}$ holds with probability at least $1-\delta$ for any $\delta \in (0, 1)$. Furthermore, if X is sub-Gaussian, meaning that $\mathbb{E}e^{\lambda(X-\mu)} \le e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$, then $\bar{X}_n$ can be referred to as a sub-Gaussian estimator, as it satisfies with probability $1-\delta$ that $|\bar{X}_n - \mu| \le \sigma\sqrt{2\log(2/\delta)/n}$. In order to obtain sub-Gaussian deviations under a condition of bounded second moments, Fan et al. (2017) considered the Huber mean estimator $\widehat{\mu}_\tau = \operatorname{argmin}_{\theta\in\mathbb{R}} \sum_{i=1}^n \ell_\tau(X_i - \theta)$, which is closely related to the method proposed by Catoni (2012). Here $\ell_\tau(\cdot)$ denotes the Huber loss; see definition (18). Theorem 5 in Fan et al. (2017) establishes that for any $v \ge \sigma$, $\widehat{\mu}_\tau$ with $\tau = v\sqrt{n/\log(1/\delta)}$ and $\delta \in (0, 1)$ satisfies $|\widehat{\mu}_\tau - \mu| \le 4v\sqrt{\log(1/\delta)/n}$ with probability at least $1-2\delta$ as long as $n \ge 8\log(1/\delta)$. While Fan et al. (2017) does not explicitly state this, the divergence of τ in this context is also intended to strike a balance between bias and robustness. In comparison to (univariate) mean estimation, the problem of regression with growing dimensions and generated response variables present new technical challenges and requires more nuanced analysis. Nevertheless, the underlying phenomenon is quite similar.

1.1 Notation

For any two vectors $u = (u_1, \ldots, u_k)^{\mathrm{T}}$ and $v = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$, we define their inner product as $u^{\mathrm{T}}v = \langle u, v\rangle = \sum_{j=1}^k u_j v_j$. We use $\|\cdot\|_p$ $(1 \le p \le \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|u\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ for $p \ge 1$ and $\|u\|_\infty = \max_{1\le i\le k}|u_i|$. Let $\mathbb{S}^{k-1} = \{u \in \mathbb{R}^k : \|u\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^k$ under ℓ_2-norm. Given a positive semi-definite matrix $A \in \mathbb{R}^{k\times k}$ and $u \in \mathbb{R}^k$, let $\|u\|_A := \|A^{1/2}u\|_2$. Given an event/subset $\mathcal{A}$, $\mathbb{1}(\mathcal{A})$ or $\mathbb{1}_{\mathcal{A}}$ denotes the zero-one indicator function for $\mathcal{A}$. For two real numbers a and b, we write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For two sequences $\{a_n\}_{n\ge1}$ and $\{b_n\}_{n\ge1}$ of non-negative numbers, we write $a_n \lesssim b_n$ if $a_n \le Cb_n$ for some constant $C > 0$ independent of n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2 Preliminaries and background

2.1 The joint regression framework

Assume we observe a sequence of data vectors $\{(Y_i, X_i)\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ is the response variable, and $X_i \in \mathbb{R}^p$ is a p-dimensional vector of explanatory variables (covariates). For some fixed probability level $\alpha \in (0, 1)$, denote the conditional α-level quantile and ES of Y_i given the covariates X_i as $Q_\alpha(Y_i|X_i)$ and $\mathrm{ES}_\alpha(Y_i|X_i)$, respectively. For the latter, we adhere to the definition $\mathrm{ES}_\alpha(Y_i|X_i) = \mathbb{E}\{Y_i|Y_i \le Q_\alpha(Y_i|X_i), X_i\}$.

We consider the joint regression framework introduced in Dimitriadis and Bayer (2019) for modelling the conditional quantile and expected shortfall. For some probability level $\alpha \in (0, 1)$, assume that

$$Q_\alpha(Y_i|X_i) = X_i^{\mathrm{T}}\beta^*, \qquad \mathrm{ES}_\alpha(Y_i|X_i) = X_i^{\mathrm{T}}\theta^*, \tag{4}$$

where $\beta^*, \theta^* \in \mathbb{R}^p$ are the unknown true underlying parameters for quantile and ES, respectively. Fissler and Ziegel (2016) explained that quantile and ES are jointly *elicitable* and proposed a class of *strictly consistent* joint loss functions for quantile and ES estimation. Let G_1 be an increasing and integrable function, and let $\mathcal{G}_2$ be a three times continuously differentiable function such that both $\mathcal{G}_2$ and its derivative $G_2 = \mathcal{G}_2'$ are strictly positive. The proposed joint loss function in Fissler and Ziegel (2016) takes the form

$$\begin{aligned} s(\beta, \theta; Y, X) = {} & \{\alpha - \mathbb{1}(Y \le X^{\mathrm{T}}\beta)\}\{G_1(Y) - G_1(X^{\mathrm{T}}\beta)\} \\ & + \frac{G_2(X^{\mathrm{T}}\theta)}{\alpha}\Bigg\{\underbrace{\alpha X^{\mathrm{T}}(\theta-\beta) - (Y - X^{\mathrm{T}}\beta)\mathbb{1}(Y \le X^{\mathrm{T}}\beta)}_{=:\, s_0(\beta,\theta;Y,X)}\Bigg\} - \mathcal{G}_2(X^{\mathrm{T}}\theta). \end{aligned} \tag{5}$$

This general form also includes the joint loss function proposed by Acerbi and Székely (2014) by taking $G_1(x) = -(W/2)x^2$ for some $W \in \mathbb{R}$ and $\mathcal{G}_2(x) = \alpha x^2/2$.

In the regression framework with a fixed number of covariates, Dimitriadis and Bayer (2019) established the consistency and asymptotic normality of the M-estimator $(\widetilde{\beta}^{\mathrm{T}}, \widetilde{\theta}^{\mathrm{T}})^{\mathrm{T}}$, defined as

$$\begin{pmatrix} \widetilde{\beta} \\ \widetilde{\theta} \end{pmatrix} \in \underset{\beta,\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} s(\beta, \theta; Y_i, X_i), \tag{6}$$

where $\Theta \subseteq \mathbb{R}^p$ is the parameter space, assumed to be compact, convex, and has non-empty interior. The main challenge of the aforementioned approach is that the objective function in equation (6) is non-differentiable and non-convex for any feasible choice of the functions G_1 and G_2 (Fissler & Ziegel, 2016). Note from definition (1) that the expected shortfall depends on the quantile, not vice versa. The estimation and inference of θ^* is thus the main challenge. It is, however, infeasible to estimate a single regression model for ES through M-estimation, that is, by minimising some strictly consistent loss function (Dimitriadis & Bayer, 2019).

In the joint regression framework, if the main goal is to estimate and forecast ES, then β^* can be naturally viewed as a nuisance parameter. Motivated by the idea of using Neyman-orthogonal scores to reduce sensitivity with respect to nuisance parameters (Barendse, 2020; Chernozhukov et al., 2018; Neyman, 1979) proposed a two-stage procedure that bypasses non-convex optimisation problems. In the first stage, an estimate $\widehat{\beta}$ of β^* is obtained via standard QR. The second step employs an orthogonal score with fitted thresholding quantiles to estimate θ^*. The key observation is as follows. Define the function

$$\begin{aligned} \psi_0(\beta, \theta; X) &= \mathbb{E}\{s_0(\beta, \theta; Y, X)|X\} \\ &= \alpha X^{\mathrm{T}}\theta - \mathbb{P}(Y \le X^{\mathrm{T}}\beta|X)\mathbb{E}(Y|Y \le X^{\mathrm{T}}\beta, X) + \left\{\mathbb{P}(Y \le X^{\mathrm{T}}\beta|X) - \alpha\right\}X^{\mathrm{T}}\beta, \end{aligned} \tag{7}$$

where s_0 is given in equation (5). Under model (4), we have $\psi_0(\beta^*, \theta^*; X) = 0$ almost surely over X. Let $F_{Y|X}$ be the conditional distribution function of Y given X. Provided that $F_{Y|X}$ is continuously differentiable, taking the gradient with respect to β on both sides of the above equality yields

$$\partial_\beta \psi_0(\beta, \theta; X) = \{F_{Y|X}(X^{\mathrm{T}}\beta) - \alpha\}X, \qquad \text{for any } \beta, \theta \in \mathbb{R}^p.$$

We hence refer to the following property:

$$\partial_\beta \psi_0(\beta, \theta; X)\big|_{\beta=\beta^*} = \{F_{Y|X}(X^{\mathrm{T}}\beta^*) - \alpha\}X = 0 \tag{8}$$

as *Neyman orthogonality*.

2.2 Two-step ES estimation via Neyman-orthogonal score

We start with a detailed overview of the two-step approach proposed by Barendse (2020) using the Neyman-orthogonal score (7) under the joint model (4). In Section 3.1, we will develop a non-asymptotic (finite-sample) theory for the two-step ES estimator, $\widehat{\theta}$, under the regime in which p is allowed to increase with the sample size n. We further develop asymptotic normality results for individual coordinates, or more generally linear projections, of $\widehat{\theta}$, in the increasing-dimension regime '$p^2/n = o(1)$'. Our non-asymptotic results and techniques pave the way for analysing high-dimensional sparse quantile-ES models.

The first step involves computing the standard QR estimator of β^*:

$$\widehat{\beta} \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(Y_i - X_i^{\mathrm{T}}\beta), \tag{9}$$

where $\rho_\alpha(u) = \{\alpha - \mathbb{1}(u < 0)\}u$ is the check function (Koenker & Bassett, 1978). The second step is motivated by the orthogonal score s_0 in equation (5). Specifically, let $\widehat{\mathcal{L}}(\beta, \theta) = (1/n)\sum_{i=1}^{n} s_i^2(\beta, \theta)$

be the joint empirical loss with

$$s_i(\beta, \theta) := s_0(\beta, \theta; Y_i, X_i) = \alpha X_i^{\mathrm{T}}\theta - \mathbb{1}(Y_i \le X_i^{\mathrm{T}}\beta)Y_i + \{\mathbb{1}(Y_i \le X_i^{\mathrm{T}}\beta) - \alpha\}X_i^{\mathrm{T}}\beta. \tag{10}$$

Given $\widehat{\beta}$ obtained from the first step, the ES estimator $\widehat{\theta}$ of θ^* is computed as

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \widehat{\mathcal{L}}(\widehat{\beta}, \theta). \tag{11}$$

For any β fixed, the function $\theta \mapsto \widehat{\mathcal{L}}(\beta, \theta)$ is convex with gradient and Hessian given by

$$\partial_\theta \widehat{\mathcal{L}}(\beta, \theta) = \frac{2\alpha}{n}\sum_{i=1}^n s_i(\beta, \theta)X_i \quad \text{and} \quad \partial_\theta^2 \widehat{\mathcal{L}}(\beta, \theta) = \frac{2\alpha^2}{n}\sum_{i=1}^n X_i X_i^{\mathrm{T}},$$

respectively. By the first-order condition, the ES regression estimator $\widehat{\theta}$ satisfies the moment condition $\partial_\theta \widehat{\mathcal{L}}(\widehat{\beta}, \widehat{\theta}) = 0$, and has a closed-form expression

$$\widehat{\theta} = \widehat{\beta} + \left(\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right)^{-1} \frac{1}{\alpha}\sum_{i=1}^n (Y_i - X_i^{\mathrm{T}}\widehat{\beta})X_i \mathbb{1}(Y_i \le X_i^{\mathrm{T}}\widehat{\beta}), \tag{12}$$

provided that $\mathbb{X} = (X_1, \ldots, X_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ is full-rank.

Remark 1 When p is large, we suggest using the convolution-smoothed QR (conquer) estimator (Fernandes et al., 2021; He et al., 2023) in the first step, which can be computed by fast and scalable gradient-based algorithms. Given a smoothing parameter/bandwidth $h > 0$, the conquer estimator $\widehat{\beta}_h$ minimises the convolution-smoothed loss function $\beta \mapsto \sum_{i=1}^n \rho_{\alpha,h}(Y_i - X_i^{\mathrm{T}}\beta)$ with $\rho_{\alpha,h}(u) = (\rho_\alpha * K_h)(u) = \int_{-\infty}^{\infty} \rho_\alpha(u)K_h(v-u)\mathrm{d}v$, where $K_h(u) := (1/h)K(u/h)$ for some symmetric, non-negative kernel function K, and * is the convolution operator. We refer to Fernandes et al. (2021) and He et al. (2023) for more details, including both asymptotic and finite-sample properties of $\widehat{\beta}_h$ when p is fixed and growing as well as the bandwidth selection.

Define $p \times p$ matrices $\Sigma = \mathbb{E}(XX^{\mathrm{T}})$ and $\Omega = \mathbb{E}(\omega^2 XX^{\mathrm{T}})$ with $\omega := (Y - X^{\mathrm{T}}\beta^*)\mathbb{1}(Y \le X^{\mathrm{T}}\beta^*) + \alpha X^{\mathrm{T}}(\beta^* - \theta^*)$ satisfying $\mathbb{E}(\omega|X) = 0$ under model (4). Provided that $p = p_n$ satisfies $p^2/n \to 0$, we will show in the Online Supplementary Theorem A.3 that $\widehat{\theta}_j$ is asymptotically normal:

$$\frac{\alpha\sqrt{n}(\widehat{\theta}_j - \theta_j^*)}{\sqrt{(\Sigma^{-1}\Omega\Sigma^{-1})_{jj}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1) \quad \text{as } n, p \to \infty.$$

As a direct implication, an asymptotically valid entrywise confidence interval for θ^* can be constructed as follows. Recall that $(\widehat{\beta}, \widehat{\theta})$ is the joint quantile-ES regression estimators given in equations (9) and (11), respectively. Define the estimated 'residuals' as

$$\widehat{\varepsilon}_i = Y_i - X_i^{\mathrm{T}}\widehat{\beta} \quad \text{and} \quad \widehat{\omega}_i = \widehat{\varepsilon}_i \wedge 0 + \alpha X_i^{\mathrm{T}}(\widehat{\beta} - \widehat{\theta}). \tag{13}$$

We then use the sample analogue of Σ and a plug-in estimator of Ω, namely, $\widehat{\Sigma} = \sum_{i=1}^n X_i X_i^{\mathrm{T}}/n$ and $\widehat{\Omega} = \sum_{i=1}^n \widehat{\omega}_i^2 X_i X_i^{\mathrm{T}}/n$. Consequently, we construct (approximate) 95% confidence interval for each

coefficient as

$$\left[\widehat{\theta}_j - \frac{1.96}{\alpha\sqrt{n}}(\widehat{\Sigma}^{-1}\widehat{\Omega}\widehat{\Sigma}^{-1})_{jj}^{1/2}, \widehat{\theta}_j + \frac{1.96}{\alpha\sqrt{n}}(\widehat{\Sigma}^{-1}\widehat{\Omega}\widehat{\Sigma}^{-1})_{jj}^{1/2}\right], \quad j = 1, \ldots, p. \tag{14}$$

3 Robust expected shortfall regression

3.1 Motivation

The two-step estimator $\widehat{\theta}$ given in equation (12) is essentially an LSE with generated response variables. While the two-step procedure is computationally efficient and enjoys nice asymptotic properties, due to the use of the least squares type loss, it is sensitive to outliers or heavy-tailed data that is ubiquitous in various areas such as climate, insurance claims, and genomics data. In particular, heavy-tailedness has become a well-known stylised fact of financial returns and stock-level predictor variables (Cont, 2001). Since the expected shortfall is a quantity that describes the tail behaviour of a distribution, it is important to construct an estimator that is robust to the power-law or Pareto-like tails.

To motivate the need for a robust ES estimator, we start with the non-regression setting in which $X_i \equiv 1$. The two-step ES estimator (12) can then be simplified as

$$\widehat{\mathrm{ES}}_\alpha = \frac{1}{\alpha n}\sum_{i=1}^n Y_i \mathbb{1}\{Y_i \le \widehat{Q}_\alpha\} + \widehat{Q}_\alpha\{1 - \widehat{F}(\widehat{Q}_\alpha)/\alpha\}, \tag{15}$$

where $\widehat{F}$ is the empirical CDF of Y and $\widehat{Q}_\alpha = \widehat{F}^{-1}(\alpha)$ is the sample quantile. The estimator $\widehat{\mathrm{ES}}_\alpha$ (15) coincides with the ES estimate (4) in Bassett et al. (2004), although the latter is motivated differently by the following property:

$$\mathrm{ES}_\alpha(Y) = \mathbb{E}(Y) - \frac{1}{\alpha}\min_{\beta\in\mathbb{R}} \mathbb{E}\rho_\alpha(Y - \beta).$$

Since $|\widehat{F}(\widehat{Q}_\alpha) - \alpha| \le 1/n$, up to higher order terms, $\widehat{\mathrm{ES}}_\alpha$ equals $(\alpha n)^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{Y_i \le \widehat{Q}_\alpha\}$ which, by the consistency of sample quantiles, is first-order equivalent to the 'oracle' ES estimator $\widehat{\mathrm{ES}}_\alpha^{\mathrm{ora}} := (\alpha n)^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{Y_i \le Q_\alpha(Y)\}$.

Since the truncated variable $Y_i\mathbb{1}\{Y_i \le Q_\alpha(Y)\}$ can be highly left-skewed with heavy tails, the corresponding empirical mean is sensitive to the (left) tails of the distribution of Y, and hence lacks robustness against heavy-tailed data. Specifically, let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance $\sigma^2 > 0$. When X_i is sub-Gaussian (i.e. $\mathbb{E}(e^{\lambda X_i}) \le e^{\lambda^2\sigma^2/2}$ for any $\lambda \in \mathbb{R}$), it follows from the Chernoff bound (Chernoff, 1952) that

$$\mathbb{P}\left\{|\bar{X}_n - \mu| \ge \sigma\sqrt{2\log(2/\delta)/n}\right\} \le \delta, \quad \text{valid for any } \delta \in (0, 1). \tag{16}$$

In other words, the sample mean $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ satisfies the sub-Gaussian deviation bound. On the other hand, the following proposition provides a lower bound for the deviations of the empirical mean $(1/n)\sum_{i=1}^n Y_i\mathbb{1}\{Y_i \le Q_\alpha(Y)\}$ when the distribution of Y is the least favourable among all heavy-tailed distributions with mean zero and variance σ^2.

Proposition 1 For any value of the standard deviation $\sigma > 0$ and any probability level $\delta \in (0, e^{-1}]$, there exists some distribution with mean zero and variance σ^2 such that for any $\alpha \in (0, 1)$, the i.i.d. sample $\{Y_i\}_{i=1}^n$ of size n drawn from it satisfies

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n Y_i\mathbb{1}(Y_i \le Q_\alpha) - \mathbb{E}\{Y\mathbb{1}(Y \le Q_\alpha)\} \le -\sigma\sqrt{\frac{1}{\delta n}\cdot\frac{1 - e\delta}{\sqrt{2e}}}\right] \ge \delta, \tag{17}$$

as long as $n \geq e\delta/\alpha$, where $Q_\alpha = Q_\alpha(Y)$ is the αth quantile of Y.

Together, the upper and lower bounds (16) and (17) show that the worst case deviations of the empirical mean are sub-optimal when the underlying distribution is heavy-tailed (as opposed to having Gaussian-like thin tails). If Y follows a heavy-tailed distribution, such as the t- or Pareto distributions, then the left-truncated variables $Z_i := Y_i \mathbb{1}\{Y_i \leq Q_\alpha(Y)\}$ have not only heavy but also asymmetric tails. In this case, the empirical mean $(\alpha n)^{-1} \sum_{i=1}^n Z_i$ can be a sub-optimal estimator of $\mathrm{ES}_\alpha(Y)$.

3.2 Robust estimation and inference via adaptive Huber regression

To robustify the ES regression estimator (12) in the presence of skewed heavy-tailed observations, we utilise the idea of adaptive Huber regression in Zhou et al. (2018). For some $\tau > 0$, the Huber loss (Huber, 1973) takes the form

$$\ell_\tau(u) = \begin{cases} u^2/2 & \text{if } |u| \leq \tau, \\ \tau|u| - \tau^2/2 & \text{if } |u| > \tau. \end{cases} \tag{18}$$

We propose a robust/Huberised ES regression estimator defined as

$$\widehat{\theta}_\tau \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell_\tau(s_i(\widehat{\beta}, \theta)), \tag{19}$$

where $s_i(\widehat{\beta}, \theta)$ is as defined in equation (10), and $\tau > 0$ is a robustification parameter that should be calibrated adaptively from data.

To see this, we consider the oracle Huber ES estimator defined as

$$\widehat{\theta}_\tau^{\mathrm{ora}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell_\tau(s_i(\beta^*, \theta)) = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell_\tau(Z_i - \alpha X_i^{\mathrm{T}} \theta), \tag{20}$$

where $Z_i = (Y_i - X_i^{\mathrm{T}}\beta^*)\mathbb{1}(Y_i \leq X_i^{\mathrm{T}}\beta^*) + \alpha X_i^{\mathrm{T}}\beta^*$. For any $\tau > 0$, $\widehat{\theta}_\tau^{\mathrm{ora}}$ is an M-estimator of its population counterpart

$$\theta_\tau^* = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathbb{E}\{\ell_\tau(Z_i - \alpha X_i^{\mathrm{T}} \theta)\}.$$

Let $\psi_\tau(t) = \ell_\tau'(t) = \mathrm{sign}(t) \min(|t|, \tau)$ be the derivative of the Huber loss. By the convexity of the Huber loss, θ_τ^* must satisfy the first-order condition $\mathbb{E}\{\psi_\tau(Z_i - \alpha X_i^{\mathrm{T}}\theta_\tau^*)X_i\} = 0$. On the other hand, define the ES deviations $\omega_i = Z_i - \alpha X_i^{\mathrm{T}}\theta^*$, satisfying $\mathbb{E}(\omega_i | X_i) = 0$ and $\mathbb{E}(\omega_i) = 0$. Since the conditional distribution of ω_i given X_i is asymmetric, in general we have $\mathbb{E}\{\psi_\tau(Z_i - \alpha X_i^{\mathrm{T}}\theta^*)X_i\} = \mathbb{E}\{\psi_\tau(\omega_i)X_i\} \neq 0$, which in turn implies that $\theta_\tau^* \neq \theta^*$. We thus refer to their difference under the ℓ_2-norm, $\|\theta_\tau^* - \theta^*\|_2$, as the robustification bias. Proposition 2 provides an upper bound for the robustification bias, which depends on τ and some moment parameters. In particular, τ needs to diverge for the robustification bias to diminish.

Proposition 2 Assume that $\varepsilon := Y - X^{\mathrm{T}}\beta^*$ satisfies $\mathrm{var}_X(\varepsilon \wedge 0) \leq \overline{\sigma}^2$ almost surely for some constant $\overline{\sigma} > 0$, and that $\kappa_4 = \sup_{u \in \mathbb{S}^{p-1}} \mathbb{E}\langle u, \Sigma^{-1/2}X\rangle^4 < \infty$, where $\Sigma = \mathbb{E}(XX^{\mathrm{T}})$ is positive definite. Then, for any $\tau \geq 2\kappa_4^{1/4}\overline{\sigma}$, we have $\|\theta_\tau^* - \theta^*\|_\Sigma \leq 2\overline{\sigma}^2/(\alpha\tau)$.

In Section 4, we investigate the finite-sample properties of the robust ES estimator $\widehat{\theta}_\tau$ obtained via equations (9) and (19): our results include a deviation inequality for $\|\widehat{\theta}_\tau - \theta^*\|_\Sigma$ (Theorem 1),

the Bahadur representation (Theorem 2), and a Berry–Esseen bound for linear projections of $\widehat{\theta}_\tau$ and $\widehat{\theta}_\tau^{\mathrm{ora}}$ (Theorem 3). With a properly chosen τ that is of order $\tau \asymp \overline{\sigma}\sqrt{n/p}$, we will show that $\alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \lesssim \overline{\sigma}\sqrt{p/n}$ with high probability. Moreover, for any deterministic vector $a \in \mathbb{R}^p$, the standardised statistic $\alpha\sqrt{n}\langle a, \widehat{\theta}_\tau - \theta^*\rangle/\varrho_a$ converges in distribution to $\mathcal{N}(0, 1)$, where $\varrho_a^2 = a^{\mathrm{T}}\Sigma^{-1}\Omega\Sigma^{-1}a$ and $\omega = (Y - X^{\mathrm{T}}\beta^*)\mathbb{1}(Y \le X^{\mathrm{T}}\beta^*) + \alpha X^{\mathrm{T}}(\beta^* - \theta^*)$. Our theoretical analysis reveals two attractive properties of the adaptive Huberised ES estimator $\widehat{\theta}_\tau$: (i) the non-asymptotic deviation upper bounds for $\widehat{\theta}_\tau$ are much smaller in order than those for $\widehat{\theta}$ at any given confidence level and (ii) the asymptotic relative efficiency of $\widehat{\theta}_\tau$ to $\widehat{\theta}$ is one. Moreover, Theorem 3 shows that the two-step robust estimator (with estimated conditional quantiles) is asymptotically equivalent to the oracle Huberised estimator (20) (assuming β^* were known). This further justifies the usefulness of the Neyman-orthogonal score, which makes the QR estimation error first-order negligible.

Consistent estimators of Σ and $\Omega = \mathbb{E}(\omega^2 XX^{\mathrm{T}})$ are useful for statistical inference. Given the pair of quantile-ES regression estimators $(\widehat{\beta}, \widehat{\theta}_\tau)$, with a slight abuse of notation we use $\widehat{\varepsilon}_i$ and $\widehat{\omega}_i$ to denote the fitted QR and ES residuals as in equation (13) except with $\widehat{\theta}$ replaced by $\widehat{\theta}_\tau$. As discussed in Section 2.2, a naive estimate of Ω is $\widehat{\Omega} = (1/n)\sum_{i=1}^n \widehat{\omega}_i^2 X_i X_i^{\mathrm{T}}$. In the presence of heavy-tailed errors ε_i, even the 'oracle' estimate $\widetilde{\Omega} = (1/n)\sum_{i=1}^n \omega_i^2 X_i X_i^{\mathrm{T}}$ performs poorly and tends to overestimate. Motivated by Huber regression, we further propose a simple truncated estimator of Ω given by

$$\widehat{\Omega}_\gamma = \frac{1}{n}\sum_{i=1}^n \psi_\gamma^2(\widehat{\omega}_i) X_i X_i^{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^n \min\{|\widehat{\omega}_i|, \gamma\}^2 X_i X_i^{\mathrm{T}}, \tag{21}$$

where $\gamma = \gamma(n, p) > 0$ is a second robustification parameter. Consequently, we construct approximate 95% robust confidence intervals for θ_j^*'s as

$$\left[\widehat{\theta}_{\tau,j} - \frac{1.96}{\alpha\sqrt{n}}(\widehat{\Sigma}^{-1}\widehat{\Omega}_\gamma\widehat{\Sigma}^{-1})_{jj}^{1/2}, \widehat{\theta}_{\tau,j} + \frac{1.96}{\alpha\sqrt{n}}(\widehat{\Sigma}^{-1}\widehat{\Omega}_\gamma\widehat{\Sigma}^{-1})_{jj}^{1/2}\right], \quad j = 1, \ldots, p. \tag{22}$$

The convergence rate of $\widehat{\Omega}_\gamma$ with a suitably chosen γ will be discussed in Section 4.

As previously discussed, the robustification parameter τ plays a crucial role in achieving a balance between bias and robustness against heavy-tailed error distributions. This balance is necessary because of the asymmetric nature of the ES residual $\omega = \varepsilon \wedge 0 + \alpha X^{\mathrm{T}}(\beta^* - \theta^*)$ with $\varepsilon = Y - X^{\mathrm{T}}\beta^*$. Assuming that the (conditional) variance of $\varepsilon_- = \varepsilon \wedge 0$ is bounded, i.e. $\mathrm{var}_X(\varepsilon_-) \le \overline{\sigma}^2$ (almost surely) for some $\overline{\sigma} > 0$, Theorem 1 suggests that to achieve a tight deviation bound at the $1 - \delta$ confidence level for any given $\delta \in (0, 1)$, the robustification parameter $\tau = \tau(n, p)$ should be of order $\overline{\sigma}\sqrt{n/(p + \log\delta^{-1})}$. In practice, the scale of $\overline{\sigma}$ is typically unknown. A useful heuristic is to substitute it with the sample standard deviation of the negative QR residuals $\{\widehat{\varepsilon}_{i,-} = \min(Y_i - X_i^{\mathrm{T}}\widehat{\beta}, 0)\}_{i=1}^n$, which we denote by $\widehat{\sigma}$. Here, $\widehat{\beta}$ refers to the first-stage QR estimator. Using $\widehat{\tau} = \widehat{\sigma}\sqrt{n/(p + \log\delta^{-1})}$ as a data-driven proxy for τ, the resulting estimator is also location and scale equivariant.

In the following, we present a refined data-driven approach for selecting τ that consistently outperforms the previously mentioned rule of thumb in the numerical experiments conducted in Section 6. This approach is adapted from the method proposed in L. Wang et al. (2021) and draws inspiration from the censored equation approach originally introduced by Hahn et al. (1990) as a proof technique for deriving robust weak convergence theory for self-normalised sums. Note that for each $\tau > 0$, the Huber ES estimator $\widehat{\theta}_\tau$ can be defined equivalently as the solution to the estimating equation $\sum_{i=1}^n \psi_\tau(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta)X_i = 0, \theta \in \mathbb{R}^d$, where $\widehat{Z}_i = \widehat{\varepsilon}_{i,-} + \alpha X_i^{\mathrm{T}}\widehat{\beta}$ are the generated response variables, and $\widehat{\beta}$ denotes the initial QR estimator. Since the optimal choice of τ is proportional to the noise scale, we propose to estimate θ^* and the unknown noise scale simultaneously by solving

the following system of equations for $(\theta, s) \in \mathbb{R}^p \times (0, \infty)$:

$$g_1(\theta, s) := \frac{1}{n}\sum_{i=1}^{n} \psi_k\left(\frac{\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta}{s}\right) X_i = 0,$$
$$g_2(\theta, s) := \frac{1}{n}\sum_{i=1}^{n} \psi_k^2\left(\frac{\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta}{s}\right) - 1 = 0,$$

where $k = k(n, p) = \sqrt{n/(p + \log n)}$. Since the Huber loss is convex, the first (vector) equation in θ with s fixed can be solved using either the iteratively reweighted least squares algorithm or the Broyden–Fletcher–Goldfarb–Shanno algorithm. For the second equation with θ fixed, it can be shown that the function $s \mapsto g_2(\theta, s)$ is non-increasing, as demonstrated by its derivative

$$\frac{\partial}{\partial s} g_2(\theta, s) = -\frac{2}{ns^3}\sum_{i=1}^{n} (\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta)^2 \mathbb{1}\{|\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta| \le sk\} \le 0.$$

Proposition 3 in L. Wang et al. (2021) further guarantees that the equation $g_2(\theta, s) = 0$ with θ fixed has a unique solution, provided that $\sum_{i=1}^{n} \mathbb{1}(|\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta| > 0) > n/k^2 = p + \log n$. Based on these observations, we propose the following alternating algorithm, which begins at iteration 0 with an initial estimate $\theta^0 = \widehat{\theta}$, the two-step LSE given in equation (11), or equivalently equation (12). At each iteration $t = 1, 2, \ldots$, the procedure involves two steps:

(i) Compute the ES 'residuals' $\omega_i^t = \widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta^{t-1}$ using the previous estimate θ^{t-1}. Let s^t be the solution to the equation $(1/n)\sum_{i=1}^{n} (|\omega_i^t/s| \wedge k)^2 = 1$, $s > 0$.
(ii) Compute the updated estimate $\theta^t \in \operatorname{argmin}_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell_{\tau^t}(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta)$, where $\tau^t = s^t k$.

Given a prespecified tolerance $\epsilon > 0$ (e.g. $\epsilon = 10^{-5}$), the algorithm will terminate at the tth iteration if $\max\{\|g_1(\theta^t, s^t)\|_2, |g_2(\theta^t, s^t)|\} \le \epsilon$, or if the maximum number of iterations is reached. Our numerical experiments in Section 6 show that this algorithm generally achieves convergence after only a small number of iterations. Intuitively, we attribute the algorithm's fast convergence to the observation that $\widehat{\theta}_\tau$ changes gradually as τ varies. This graduate change cause the residuals to behave similarly over a range of τ values. We refer the reader to the Online Supplementary Section B for a detailed elaboration on the motivations behind our proposed data-driven method.

4 Statistical theory

This section presents non-asymptotic high probability bounds for the error $\|\widehat{\theta}_\tau - \theta^*\|_2$ of the Huberised two-step ES estimator $\widehat{\theta}_\tau$, as defined in equation (19). Additionally, we establish a non-asymptotic Bahadur representation for $\widehat{\theta}_\tau$, which is a crucial step towards obtaining a Berry–Esseen-type bound for Gaussian approximation. Throughout this section, we write $X = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ with $x_1 \equiv 1$. Without loss of generality, we assume that the random predictors $x_2, \ldots, x_p$ have zero means, that is, $\mu_j = \mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$. This makes the later *sub-Gaussian* assumption more reasonable; see Condition 2 below. Otherwise, we set $Z = (1, z_2, \ldots, z_p)^{\mathrm{T}} = (1, x_2 - \mu_2, \ldots, x_p - \mu_p)^{\mathrm{T}}$. With this notation, the joint model (4) becomes $Q_\alpha(Y|Z) = \beta_0^\dagger + \sum_{j=2}^{p} z_j \beta_j^*$ and $\mathrm{ES}_\alpha(Y|Z) = \theta_0^\dagger + \sum_{j=2}^{p} z_j \theta_j^*$, where $\beta_1^\dagger = \mu^{\mathrm{T}}\beta^*$ and $\theta_1^\dagger = \mu^{\mathrm{T}}\theta^*$ with $\mu = (1, \mu_2, \ldots, \mu_p)^{\mathrm{T}}$. The sub-Gaussian assumption can then be imposed on Z, and our analysis naturally applies to $\{(Y_i, Z_i)\}_{i=1}^n$.

In the context of a joint (linear) quantile and ES regression model, we initiate by establishing a high probability bound, explicitly dependent on n and p, for the QR estimator $\widehat{\beta}$ (9). To this end, we impose the following conditions on the covariates and the conditional distribution of Y given X.

Condition 1 The conditional density function of $\varepsilon := Y - X^{\mathrm{T}}\beta^*$ given X, denoted by $f_{\varepsilon|X}$, exists and is continuous on its support. Moreover, there exist constants $\underline{f}, l_0 > 0$ such that $f_{\varepsilon|X}(0) \geq \underline{f}$ and $|f_{\varepsilon|X}(t) - f_{\varepsilon|X}(0)| \leq l_0|t|$ for all $t \in \mathbb{R}$ almost surely (over X).

Condition 2 The random covariate vector $X \in \mathbb{R}^p$ is sub-Gaussian, that is, there exists some (dimension-free) constant $\upsilon_1 \geq 1$ such that $\mathbb{P}(|u^{\mathrm{T}}W| \geq \upsilon_1 t) \leq 2e^{-t^2/2}$ for all $t \geq 0$ and $u \in \mathbb{S}^{p-1}$, where $W = \Sigma^{-1/2}X$ and $\Sigma = \mathbb{E}(XX^{\mathrm{T}})$ is positive definite. Let $\kappa_l = \sup_{u\in\mathbb{S}^{p-1}} \mathbb{E}|u^{\mathrm{T}}W|^l$ for $l \geq 1$.

Condition 1 imposes regularity conditions on the random error distributions, accommodating heteroskedastic error distributions and not requiring the existence of any moment. Condition 2 is used to guarantee that population and empirical quantities (e.g. the objective or gradient function or the gradient function) are uniformly close to each other in a compact region. It can be replaced by a boundedness assumption, which will lead to similar results. For example, $X = (x_1, \ldots, x_p)^{\mathrm{T}}$ is compactly supported with either $\|X\|_\infty \leq C_X$ or $\|\Sigma^{-1/2}X\|_2 \leq B_X$, where C_X is an absolute constant and B_X is usually proportional to $\sqrt{p}$.

Proposition 3 Under Conditions 1 and 2, the QR estimator $\widehat{\beta}$ given in equation (9) satisfies, for any $t \geq 0$, that $\|\widehat{\beta} - \beta^*\|_\Sigma \leq C_1\underline{f}^{-1}\sqrt{(p+t)/n}$ holds with probability at least $1 - e^{-t}$ as long as $n \geq C_2 l_0^2 \underline{f}^{-4}(p+t)$, where $C_1, C_2 > 0$ are constants depending only on υ_1.

While QR has been extensively studied since the seminal work of Koenker and Bassett (1978), there remains a paucity of literature that addresses its finite-sample properties, particularly in terms of high probability bounds. Proposition 3 revisits Theorem 2.1 originally presented in Pan and Zhou (2021). For the sake of completeness, we provide a self-contained and simplified proof in the Online Supplementary Section G.9. Shifting our focus to ES regression, which involves conditional expectations, we additionally impose the following moment condition on the random error ε.

Condition 3 The conditional CDF $F_{\varepsilon|X}$ of ε given X is continuously differentiable and satisfies $|F_{\varepsilon|X}(t) - F_{\varepsilon|X}(0)| \leq \overline{f}|t|$ for all $t \in \mathbb{R}$. Moreover, the negative part of ε, denoted by $\varepsilon_- = \varepsilon \wedge 0$, satisfies $\mathrm{var}_X(\varepsilon_-) \leq \overline{\sigma}^2$ almost surely (over X), where var_X denotes the conditional variance given X.

Condition 3 asserts that the conditional variance of the negative part of the QR residual $\varepsilon = Y - X^{\mathrm{T}}\beta^*$ is bounded. In our theoretical analysis, we assume $\overline{\sigma}$ to be a constant for convenience. More generally, one can assume a form of $\varepsilon = \sigma(X)\eta$, where $\sigma : \mathbb{R}^p \to (0, \infty)$ is a positive function on $\mathbb{R}^p$ (not necessarily bounded), and η is independent of X satisfying $\mathrm{var}(\eta\mathbb{1}(\eta \leq 0)) \leq \overline{\sigma}^2$. In this case, an additional moment assumption on $\sigma(X)$, such as boundedness $\mathbb{E}\{\sigma(X)^4\}$, would suffice.

Our next result establishes high probability bounds for the estimation error of ES regression, conditioning on the event that $\widehat{\beta}$ falls within a local neighbourhood of β^*.

Theorem 1 Assume Conditions 2 and 3 hold. For any $t > 0$, let $r_0 > 0$ be such that $r_0 \lesssim \overline{\sigma}$ and $\overline{f}r_0^2 \lesssim \overline{\sigma}\sqrt{(p+t)/n}$. Then, the two-step robust α-ES ($0 < \alpha \leq 1/2$) estimator $\widehat{\theta}_\tau$ with $\tau = c_0\overline{\sigma}\sqrt{n/(p+t)}$ (for any $c_0 \geq 1$) satisfies that, with probability at least $1 - 3e^{-t}$ conditioned on the event $\{\|\widehat{\beta} - \beta^*\|_\Sigma \leq r_0\}$,

$$\alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \leq C_1\overline{\sigma}\sqrt{\frac{p+t}{n}} + C_2\left(\sqrt{\frac{p+t}{n}}r_0 + \overline{f}r_0^2\right) \tag{23}$$

provided that the sample size obeys $n \geq C_3(p+t)$, where $C_1 > 0$ is a constant depending on (υ_1, c_0) and $C_2, C_3 > 0$ depend only on υ_1.

For any $\delta \in (0, 1)$, the robust estimator $\widehat{\theta}_\tau$ with $\tau \asymp \overline{\sigma}\sqrt{n/(p + \log(1/\delta))}$ satisfies with probability at least $1 - \delta$ conditioned on $\{\|\widehat{\beta} - \beta^*\|_\Sigma \le r_0\}$ that

$$\alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \lesssim \overline{\sigma}\sqrt{\frac{p + \log(1/\delta)}{n}} + \sqrt{\frac{p + \log(1/\delta)}{n}} r_0 + \overline{f} r_0^2.$$

The above bound is proportional to $\log(1/\delta)$, in contrast to the bound for the two-step LSE, which is proportional to $1/\delta$, as demonstrated in the Online Supplementary Theorem A.1 in Section A. This observation suggests that the Huberised estimator is much more robust to heavy tails from a non-asymptotic perspective, compared to the two-step LSE. Specifically, in cases where the error variables only have finite variance, the worst-case deviations of $\widehat{\theta}$ are considerably larger than those of $\widehat{\theta}_\tau$.

Remark 2 (Bias-robustness trade-off). The choice of τ stated in Theorem 1 is a reflection of the bias-robustness trade-off. As discussed in Section 3.2, the robust estimator $\widehat{\theta}_\tau$ can be viewed as an M-estimator of $\theta^*_\tau = \operatorname{argmin}_\theta \mathbb{E}\{\ell_\tau(Z_i - \alpha X_i^{\mathrm{T}}\theta)\}$, which differs from the true ES regression coefficient θ^* due to the asymmetry of ES 'residuals' $\omega_i = Z_i - \alpha X_i^{\mathrm{T}}\theta^*$. Consider the decomposition

$$\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \le \underbrace{\|\widehat{\theta}_\tau - \theta^*_\tau\|_\Sigma}_{\text{robustification bias}} + \underbrace{\|\theta^*_\tau - \theta^*\|_\Sigma}_{\text{robust estimation error}}.$$

As long as $\tau \gtrsim \overline{\sigma}$ under Condition 3, Proposition 2 ensures that $\alpha\|\widehat{\theta}_\tau - \theta^*_\tau\|_\Sigma \le 2\overline{\sigma}^2/\tau$. Examining the proof of Theorem 1, we see that

$$\alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \lesssim \overline{\sigma}\sqrt{\frac{p + t}{n}} + \tau\frac{p + t}{n} + \frac{\overline{\sigma}^2}{\tau} + r_0\left(\sqrt{\frac{p + t}{n}} + \frac{\overline{\sigma}}{\tau}\right) + r_0^2$$

with high probability conditioned on the event $\{\|\widehat{\beta} - \beta^*\|_\Sigma \le r_0\}$. We therefore select $\tau \asymp \overline{\sigma}\sqrt{n/(p + t)}$ in order to minimise the upper bound as a function of τ.

Remark 3 (A uniform bound over τ). Recall from Proposition 3 that with probability at least $1 - n^{-1}$, $\|\widehat{\beta} - \beta^*\|_\Sigma \lesssim \underline{f}^{-1}\sqrt{(p + \log n)/n}$ as long as $n \gtrsim p + \log n$. Complementing the proof of Theorem 1 with a discretisation argument, we can obtain a more general result that holds for a range of τ values. Specifically, let $\overline{\tau} \ge \underline{\tau} > 0$ be such that $\overline{\sigma} \lesssim \underline{\tau} \le \overline{\tau} \lesssim \overline{\sigma}\sqrt{n/(p + \log n)}$. Then, with probability at least $1 - Cn^{-1}$ for some absolute constant $C \ge 1$,

$$\sup_{\tau \in [\underline{\tau}, \overline{\tau}]} \alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \lesssim \overline{\sigma}\sqrt{\frac{p + \log n}{n}} + \frac{\overline{\sigma}^2}{\underline{\tau}} + \max\{\overline{f}, 1/\underline{\tau}\}\frac{p + \log n}{\underline{f}^2 n}, \tag{24}$$

as long as $n \gtrsim p + \log n$. The proof of the uniform upper bound in equation (24) is provided in the Online Supplementary Section G.9. As ensured by this uniform bound, a data-driven choice of τ within the aforementioned range can be used.

If, in addition to Condition 3, some higher order moment of ε_- is bounded, namely, $\mathbb{E}_X\{|\varepsilon_- - \mathbb{E}_X(\varepsilon_-)|^k\} \le \alpha_k$ almost surely (over X) for some $k > 2$, the second term on the right-hand side of equation (24) will become $\alpha_k \underline{\tau}^{1-k}$. In order to attain tight (finite-sample) concentration bounds, the robustification parameter $\tau = \tau(n, p)$ should not exceed $\sqrt{n/(p + \log n)}$ in magnitude. Conversely, τ should demonstrate sufficiently rapid growth in order for the

bias term, controlled by $\overline{\sigma}^2\underline{\tau}^{-1}$ or $\alpha_k\underline{\tau}^{1-k}$ (in case higher order moments of ε_- are bounded), to decay at a comparable rate to the stochastic error.

Unlike the two-step LSE $\widehat{\theta}$, the robust counterpart $\widehat{\theta}_\tau$ does not possess a closed-form expression. As a pivotal step in deriving Gaussian approximation results, the following theorem furnishes a non-asymptotic Bahadur representation for $\widehat{\theta}_\tau$, complete with explicit error bounds depending on (n, p) and the first-stage QR estimation error.

Theorem 2 Assume the same conditions as in Theorem 1. For any $t > 0$, the α-ES estimator $\widehat{\theta}_\tau$ with $\tau \asymp \overline{\sigma}\sqrt{n/(p+t)}$ satisfies that, with probability at least $1 - 6e^{-t}$ conditioned on $\{\|\widehat{\beta} - \beta^*\|_\Sigma \le r_0\}$,

$$\left\| \alpha\Sigma^{1/2}(\widehat{\theta}_\tau - \theta^*) - \frac{1}{n}\sum_{i=1}^n \psi_\tau(\omega_i)\Sigma^{-1/2}X_i \right\|_2 \lesssim \overline{\sigma}\frac{p+t}{n} + \overline{f}r_0^2 + r_0\sqrt{\frac{p\log n + t}{n}} \quad (25)$$

as long as $n \gtrsim p + t$, where $\omega_i = \varepsilon_i \wedge 0 + \alpha X_i^{\mathrm{T}}(\beta^* - \theta^*)$.

Lastly, we present the following Gaussian approximation result that bounds the Kolmogorov distance between the distribution of the standardised statistic $\alpha\sqrt{n}a^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*)/\varrho_a$ and the standard normal distribution, uniformly over all deterministic vectors $a \in \mathbb{R}^p$, where $\varrho_a^2 = a^{\mathrm{T}}\Sigma^{-1}\Omega\Sigma^{-1}a$. A similar conclusion applies to the oracle robust estimate $\widehat{\theta}_\tau^{\mathrm{ora}}$ (20). The following theorem shows that the two-step robust estimator obtained via equations (9) and (19) is asymptotically equivalent to the oracle Huberised estimator (20), assuming β^* is known.

Theorem 3 In addition to Conditions 1–3, assume that there exist constants $\overline{\sigma}$, $\alpha_3 > 0$ such that

$$\mathrm{var}_X(\varepsilon_-) \ge \underline{\sigma}^2 \quad \text{and} \quad \mathbb{E}_X\{|\varepsilon_- - \mathbb{E}_X(\varepsilon_-)|^3\} \le \alpha_3 \text{ almost surely over } X. \quad (26)$$

Then, the robust α-level ($\alpha \in (0, 1/2]$) ES estimator $\widehat{\theta}_\tau$ with $\tau \asymp \overline{\sigma}\sqrt{n/(p+\log n)}$ satisfies

$$\begin{aligned} &\sup_{a\in\mathbb{R}^p, t\in\mathbb{R}} \left| \mathbb{P}\left(\alpha\sqrt{n}\langle a, \widehat{\theta}_\tau - \theta^*\rangle/\varrho_a \le t\right) - \Phi(t) \right| \\ &\lesssim \frac{\alpha_3}{\underline{\sigma}^3}\sqrt{\frac{p+\log n}{n}} + (\overline{f}/\underline{f}^2 \vee \alpha_3^{1/3})\frac{p\sqrt{\log n} + \sqrt{p}\log n}{\underline{\sigma}\sqrt{n}}. \end{aligned} \quad (27)$$

Moreover, the oracle Huberised ES estimator $\widehat{\theta}_\tau^{\mathrm{ora}}$ (20) with the same τ satisfies

$$\sup_{a\in\mathbb{R}^p, t\in\mathbb{R}} \left| \mathbb{P}\left(\alpha\sqrt{n}\langle a, \widehat{\theta}_\tau^{\mathrm{ora}} - \theta^*\rangle/\varrho_a \le t\right) - \Phi(t) \right| \lesssim \frac{\alpha_3}{\underline{\sigma}^3}\sqrt{\frac{p+\log n}{n}}. \quad (28)$$

The above Gaussian approximation result lays the theoretical foundation for the statistical inference problems of testing the linear hypothesis $H_0 : a^{\mathrm{T}}\theta^* = c_0$ vs. $H_1 : a^{\mathrm{T}}\theta^* \ne c_0$ and constructing confidence intervals for $a^{\mathrm{T}}\theta^*$, where $a \in \mathbb{R}^p$ and $c_0 \in \mathbb{R}$ are predetermined. Given the joint quantile and ES regression estimates $(\widehat{\beta}, \widehat{\theta}_\tau)$, let $\widehat{\Omega}_\gamma$ be the truncated estimator of $\Omega = \mathbb{E}(\omega^2 XX^{\mathrm{T}})$ defined in equation (21) with $\gamma = \gamma(n, p) > 0$ denoting a second robustification parameter. Then, we consider the robust test statistic $T_a = \alpha\sqrt{n}(a^{\mathrm{T}}\widehat{\theta}_\tau - c_0)/\widehat{\varrho}_{a,\gamma}$ for testing $H_0 : a^{\mathrm{T}}\theta^* = c_0$, and the (approximate) $100(1-c)\%$ confidence interval $a^{\mathrm{T}}\widehat{\theta}_\tau \pm z_{c/2}\widehat{\varrho}_{a,\gamma}/(\alpha\sqrt{n})$ for $a^{\mathrm{T}}\theta^*$, where $\widehat{\varrho}_{a,\gamma}^2 := a^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\Omega}_\gamma\widehat{\Sigma}^{-1}a$ is a robust variance estimator and $z_{c/2}$ is the upper $(c/2)$-percentile of $\mathcal{N}(0, 1)$. The consistency of $\widehat{\varrho}_{a,\gamma}^2$ with a properly chosen γ is investigated in the Online Supplementary Section C.

5 Non-parametric expected shortfall regression

In this section, we consider non-parametric models for joint quantile and expected shortfall regression. For a predetermined quantile level $\alpha \in (0, 1)$, the goal is to estimate the unknown (conditional) quantile and expected shortfall functions $f_q^*(x) = Q_\alpha(Y|X = x)$ and $f_e^*(x) = \mathrm{ES}_\alpha(Y|X = x)$, with an emphasis on the latter. By equation (1), f_q^* and f_e^* can be identified as

$$f_q^* = \underset{f_q}{\operatorname{argmin}}\, \mathbb{E}\rho_\alpha(Y - f_q(X)) \quad \text{and} \quad f_e^* = \underset{f_e}{\operatorname{argmin}}\, \mathbb{E}\{Y - f_e(X)\}^2 \mathbb{1}_{\{Y \le f_q^*(X)\}}.$$

Motivated by the two-step procedure developed under joint linear models, in the following we propose a non-parametric ES estimator using the series regression method (Andrews, 1991; Eubank & Spiegelman, 1990; Newey, 1997). Such a non-parametric estimate is carried out by regressing the dependent variable on an asymptotically growing number of approximating functions of the covariates, and therefore is closely related to the estimator define in equation (11) under the so-called many regressors model (Belloni et al., 2019), that is, the dimension $p = p_n$ is allowed to grow with n. The idea of series estimation is to first approximate f_q^* and f_e^* by their 'projections' on the linear spans of m_1 and m_2 series/basis functions, respectively, and then fit the coefficients using the observed data. Specifically, we approximate functions f_q^* and f_e^* by linear forms $U(x)^{\mathrm{T}}\beta$ and $V(x)^{\mathrm{T}}\theta$, where

$$U(x) = (u_1(x), \ldots, u_{m_1}(x))^{\mathrm{T}} \quad \text{and} \quad V(x) = (v_1(x), \ldots, v_{m_2}(x))^{\mathrm{T}}$$

are two vectors of series approximating functions of dimensions m_1 and m_2. Here both m_1 and m_2 may increase with n. We thus define the vectors of quantile and ES series approximation coefficients as

$$\beta^* \in \underset{\beta \in \mathbb{R}^{m_1}}{\operatorname{argmin}}\, \mathbb{E}\rho_\alpha(Y - U(X)^{\mathrm{T}}\beta) \quad \text{and} \quad \theta^* \in \underset{\theta \in \mathbb{R}^{m_2}}{\operatorname{argmin}}\, \mathbb{E}\{Y - V(X)^{\mathrm{T}}\theta\}^2 \mathbb{1}_{\{Y \le f_q^*(X)\}}. \tag{29}$$

Given independent observations (Y_i, X_i), $1 \le i \le n$ from $(Y, X) \in \mathbb{R} \times \mathcal{X}$ with $\mathcal{X}$ denoting a compact subset of $\mathbb{R}^p$, we write $U_i = U(X_i) \in \mathbb{R}^{m_1}$ and $V_i = V_i(X_i) \in \mathbb{R}^{m_2}$. Extending the two-step approach described in Section 2.2, we first define the (conditional) quantile series estimator of $f_q^*(x) = Q_\alpha(Y|X = x)$ (Belloni et al., 2019):

$$\widehat{f}_q(x) = U(x)^{\mathrm{T}}\widehat{\beta}, \quad x \in \mathcal{X}, \quad \text{where } \widehat{\beta} = \widehat{\beta}_{m_1} \in \underset{\beta \in \mathbb{R}^{m_1}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(Y_i - U_i^{\mathrm{T}}\beta). \tag{30}$$

With generated response variables $\widehat{Z}_i = \alpha\widehat{f}_q(X_i) + \{Y_i - \widehat{f}_q(X_i)\}\mathbb{1}_{\{Y_i \le \widehat{f}_q(X_i)\}}$, the second-stage ES series estimator is given by

$$\widehat{f}_e(x) = V(x)^{\mathrm{T}}\widehat{\theta}, \quad x \in \mathcal{X}, \quad \text{where } \widehat{\theta} = \widehat{\theta}_{m_2} \in \underset{\theta \in \mathbb{R}^{m_2}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} (\widehat{Z}_i - \alpha V_i^{\mathrm{T}}\theta)^2. \tag{31}$$

Commonly used series functions with good approximation properties include B-splines, polynomials, Fourier series and compactly supported wavelets. We refer to Newey (1997) and Chen (2007) for a detailed description of these series functions. In the context of QR, Chen (2007) established the consistency and rate of convergence at a single quantile index. More recently, Belloni et al. (2019) developed a large sample theory for the quantile series coefficient process, including convergence rate and uniform strong approximations. The choice of the parameter m_1, also known as the order of the series estimator, is crucial for establishing the balance between bias and variance.

Note that the quantile series estimator $\widehat{f}_q$ in equation (30) has been well studied by Belloni et al. (2019). Because the number of regressors increases with the sample size, conventional central limit theorems are no longer applicable to capture the joint asymptotic normality of the regression coefficients. The growing dimensionality is the primary source of technical complications. Our theoretical analysis under the joint linear model (4), which leads to novel non-asymptotic high probability bounds, can be used as a starting point for studying the two-step non-parametric ES series estimator $\widehat{f}_e$ defined in equation (31). Of particular interest is to develop a uniform inference procedure for the conditional ES function f_e^* and its theory. That is, at a given confidence level $1-\gamma$, we aim to construct a pair of functional estimates $[\widehat{f}_e^L, \widehat{f}_e^U]$ from $\{(Y_i, X_i)\}_{i=1}^n$ such that

$$\mathbb{P}\left\{\widehat{f}_e^L(x) \le f_e^*(x) \le \widehat{f}_e^U(x) \ \text{ for all } x \in \mathcal{X}\right\} \to 1-\gamma, \quad \text{as } n \to \infty.$$

Since a significant amount of additional work is still needed, including explicit characterisations of the ES series approximation error and the impact of first-stage non-parametric QR estimation error, we leave a rigorous theoretical investigation of $\widehat{f}_e$ to future work. Although we have only focussed on series methods, there are other non-parametric techniques that offer superior empirical and theoretical performance. Among those, deep neural networks have stood out as a promising tool for non-parametric estimation, from least squares, logistic to QR (Farrell et al., 2021; Schmidt-Hieber, 2020; Shen et al., 2021). It is practically useful to construct deep learning implementations of two-step estimators and statistically important to deliver valid inferences on finite-dimensional parameters following first-step estimation (of both quantile and ES functions) using deep learning. A detailed investigation of these problems is beyond the present scope but of future interest.

6 Numerical studies and real data examples

6.1 Monte Carlo experiments

In this section, we assess the numerical performance of the proposed method for fitting expected shortfall regression. For its `R` implementation, we first obtain a QR estimate via the `quantreg` library, and in step two use the `adaHuber` library to solve (19) with the robustification parameter selected adaptively as described in Section 3.2.

We compare the proposed two-step adaptive Huber ES estimator (`2S-AH`) to several competitors: (i) the joint regression estimate (`joint`) via FZ loss minimisation, implemented via the `R` library `esreg` with the default option; (ii) the two-step LSE (12) (`2S-LS`); and (iii) the oracle two-step 'estimator' (`2S-oracle`). Recall that the two-step procedure first obtains a QR estimator $\widehat{\beta}$ via either standard (Koenker & Bassett, 1978) or smoothed QR regression (He et al., 2023), and subsequently computes the ES estimator based on fitted quantile thresholds $\{X_i^{\mathrm{T}}\widehat{\beta}\}_{i=1}^n$. The oracle method refers to the two-step ES estimate based on the true quantile thresholds $\{X_i^{\mathrm{T}}\beta^*\}_{i=1}^n$.

In our simulation studies, we first generate $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^{\mathrm{T}}$ and $\eta^* = (\eta_1^*, \ldots, \eta_p^*)^{\mathrm{T}}$ independently, where γ_j^*s are independent Rademacher random variables and $\eta_j^* \sim_{\text{i.i.d.}} 0.5 \cdot \text{Bernoulli}(1/2)$. Data are then generated from the heteroscedastic model

$$Y_i = X_i^{\mathrm{T}}\gamma^* + X_i^{\mathrm{T}}\eta^* \cdot \varepsilon_i, \tag{32}$$

where $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ with $X_{ij} \sim_{\text{i.i.d.}} \text{Unif}(0, 1.5)$, and the random noise ε_i follows one of the following two distributions: (i) standard normal distribution and (ii) t-distribution with $\nu > 2$ degrees of freedom (t_ν). Given γ^* and η^*, the true quantile and expected shortfall regression coefficients are $\beta^* = \gamma^* + Q_\alpha(\varepsilon) \cdot \eta^*$ and $\theta^* = \gamma^* + \text{ES}_\alpha(\varepsilon) \cdot \eta^*$, where $Q_\alpha(\varepsilon)$ and $\text{ES}_\alpha(\varepsilon)$ are the α-level quantile and expected shortfall of ε, respectively.

We first set the dimension $p = 20$ and sample size $n = \lceil 50p/\alpha \rceil$, where the quantile level α takes values in $\{0.05, 0.1, 0.2\}$. Simulation results on the relative ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2 / \|\theta^*\|_2$, averaged over 200 replications, are reported in Tables 1 and 2 under the $\mathcal{N}(0, 1)$ and $t_{2.5}$ noise model, respectively. All four methods have very similar performance across different quantile levels in the

Table 1. Mean relative ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2/\|\theta^*\|_2$ (and standard error), averaged over 200 replications, when $\varepsilon_i \sim \mathcal{N}(0, 1)$, $p = 20$, $n = \lceil 50p/\alpha \rceil$, and $\alpha = \{0.05, 0.1, 0.2\}$

	$\mathcal{N}(0, 1)$ noise		
Method	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
`2S-AH`	0.130 (0.003)	0.150 (0.003)	0.171 (0.004)
`2S-LS`	0.130 (0.003)	0.150 (0.003)	0.171 (0.004)
`joint`	0.130 (0.003)	0.151 (0.003)	0.177 (0.004)
`2S-oracle`	0.129 (0.003)	0.149 (0.003)	0.171 (0.004)

normal model, while in the presence of heavy-tailed errors, the proposed robust method achieves consistently more favourable performance. This demonstrates that the use of adaptive Huber regression (in stage two) gains robustness against heavy-tailed errors without compromising statistical efficiency when the error distribution is light-tailed.

In a more extreme setting where $\alpha = 0.01$, Figure 1 shows the boxplots of squared ℓ_2-errors for three ES estimates (`2S-LS`, `2S-AH`, and `joint`) under the normal and t_3 models. Although the `2S-LS` estimator is easy-to-compute, it is more sensitive to heavy-tailedness than the joint estimator obtained via FZ loss minimisation. We further compare the proposed method with the joint regression approach in terms of computational efficiency. The computational time in seconds averaged over 50 independent replications, for the two methods with growing (n, p) subject to $n = \lceil 50p/\alpha \rceil$ ($\alpha \in \{0.05, 0.1, 0.2\}$) are reported in Figure 2. These numerical results show evidence that our `R` implementation of the robust two-step method can be faster than the `esreg` library for the joint regression approach by several orders of magnitude.

To shed some light on the drastic difference in numerical efficiency between the two methods, note that the joint regression approach (Dimitriadis & Bayer, 2019) relies on the Nelder–Mead simplex method, which is sensitive to the starting values and not guaranteed to converge to a local minimum. The convergence of the Nelder–Mead method is already very slow for large-scale problems because it is a direct search method based on function comparison. And due to its sensitivity to starting values, Dimitriadis and Bayer (2019) proposed to re-optimise the model (several times) with the perturbed parameter estimates as new starting values. This explains, to some extent, the fast increase in the runtime of `esreg` as both n and p grow. The function in `quantreg` that fits linear QR is coded in `fortran`, and thus is very fast in larger problems. The computation of adaptive Huber regression is based on the Barzilai-Borwein gradient descent method (Barzilai & Borwein, 1988), implemented via RcppArmadillo in `adaHuber`.

Next, we construct entrywise (approximate) 95% confidence intervals (CIs) for the expected shortfall regression parameter θ^*. The CI for the two-step estimator is based on equation (14) (non-robust) and equation (22) (robust), and we use the default option in the `esreg` package to implement Dimitriadis and Bayer (2019)'s method. To evaluate the accuracy and reliability of the CIs, we compute the empirical coverage probability and interval width based on 500 independent replications, then averaged over the p slope coefficients. Results for $p = 20$ and $n = \lceil 50p/\alpha \rceil$ ($\alpha \in \{0.05, 0.1, 0.2\}$) are reported in Tables 3 and 4.

Once again, all three methods perform similarly under normal errors, while the robust approach gives the narrowest CIs while maintaining the desired coverage level under $t_{2.5}$ errors. Together, the results in Tables 2 and 4 demonstrate the robustness of the proposed method, as indicated by the theoretical investigations in Section 4.

6.2 Data application I: health disparity

Iron deficiency is one of the most common nutritional deficiency worldwide and is one of the leading cause of anaemia (Camaschella, 2015). Being able to detect iron deficiency is essential in medical care for patients with inflammation, infection, or chronic disease. It is also important in preventive care since iron deficiency tends to present signs of a more serious illness such as gastrointestinal malignancy (Rockey & Cello, 1993). One measure of iron deficiency that has proven to

Table 2. Mean relative ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2/\|\theta^*\|_2$ (and standard error), averaged over 200 replications, when $\varepsilon_i \sim t_{2.5}$, $p = 20$, $n = \lceil 50p/\alpha \rceil$ and $\alpha = \{0.05, 0.1, 0.2\}$

	$t_{2.5}$ noise		
Method	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
`2S-AH`	0.484 (0.008)	0.470 (0.009)	0.429 (0.008)
`2S-LS`	0.612 (0.013)	0.606 (0.016)	0.532 (0.013)
`joint`	0.581 (0.012)	0.567 (0.014)	0.511 (0.013)
`2S-oracle`	0.612 (0.013)	0.607 (0.016)	0.532 (0.013)

be useful is the soluble transferrin receptor (sTRP), a carrier protein for transferrin (Mast et al., 1998). A high value of sTRP indicates iron deficiency.

The scientific goal here is to assess whether there is any disparity in sTRP levels among four different ethnic groups: Asian, Black, Mexican American, and White. To this end, we analyse a data set obtained from the National Health and Nutrition Examination Survey from 2017 to 2020 (pre-covid). In this data set, the response variable sTRP was measured for female participants who range in age from 20 to 49 years. The covariates of interest are three dummy variables that correspond to Asian, Mexican American, and Black, using White as the baseline. We adjust for demographic variables such as age, education level, and healthy diet throughout our analysis. For simplicity, we remove all participants with missing values on the covariates and the final data set consists of $n = 1{,}689$ observations and $p = 7$ covariates.

As an exploratory analysis, in Figure 3 we plot the quantile curves of sTRP measurements at levels from 50% to 99% for each of the four different ethnic groups. In this data set, the sTRP values range from 1.24 to 35.1 mg/L. We note that the normal range for females is between 1.9 and 4.4 mg/L (Kratovil et al., 2007), and values that are much higher than 4.4 mg/L indicate severe iron deficiency. We see from Figure 3 that the majority of the population have sTRP levels within the normal range. However, there are large disparities between Black and the other three ethnic groups, reflected in higher quantiles of the marginal distributions of sTRP.

To quantify the statistical significance of the aforementioned disparity, we fit robust expected shortfall regression at $\alpha = 0.75$ (upper tail), with the robustification parameter tuned by the procedure described in Section 3.2. This is equivalent to fitting the proposed `2S-AH` method at level $1 - \alpha$ (see Section 3) after flipping the signs of both the response and the covariates. We also implement the standard QR at level α.

Table 5 reports the estimated coefficients and the associated 95% confidence intervals for the three indicator covariates on the ethnic groups Asian, Mexican American, and Black, using White as a baseline. We see that both the quantile and robust expected shortfall regression methods are able to detect a health disparity between Black and White. Specifically, the estimated robust ES regression coefficient and 95% CI (in the parenthesis) is 3.03 (1.88, 4.19) vs. its QR counterparts' 0.86 (0.37, 1.35). With the use of QR (at level 0.75), we do not observe a statistically significant health disparity between Asian and White. In contrast, `2S-AH` detects health disparity between Asian and White with an estimated coefficient 2.34 (0.59, 4.09). We also see that the QR detects health disparity between Mexican American and White, but the effect size is close to zero. In summary, ES regression complements QR, and can be more effective, as a tool to detect health disparity especially when it only occurs in the tail of the conditional distribution.

6.3 Data application II: JTPA

We consider the JTPA study, a publicly funded training programme that provides training for adults with the goal of improving their earnings. Specifically, we focus on the Title II subprogramme of the JPTA study that is mainly offered to adults with barriers to employment and out-of-school youths. This data set was previously analysed in Bloom et al. (1997). It consists of 30 months of accumulated earnings for 6,102 females and 5,102 males, with 16 covariates

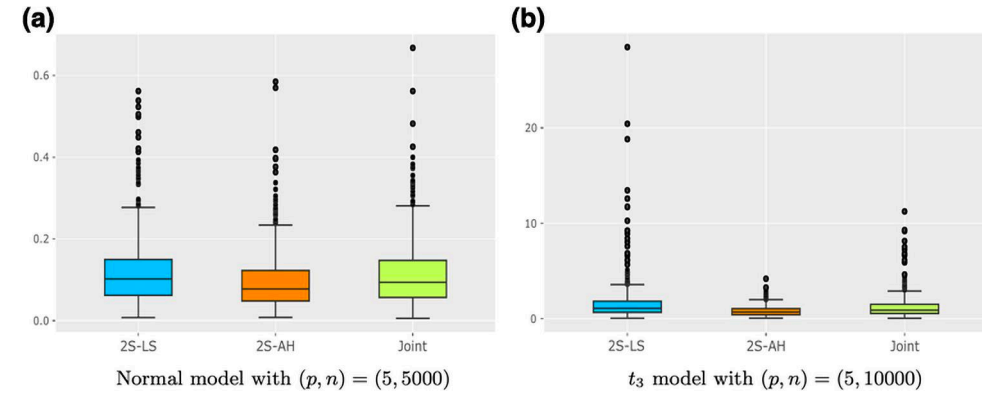

Figure 1. Boxplots of squared total ℓ_2-errors (including the intercept when its true value is 2), based on 500 replications, for three ES regression estimators (`2S-LS`, `2S-AH`, and `joint`) at quantile level $\alpha = 0.01$. The mean squared errors of these three estimators are 0.1219, 0.0983, and 0.1119 in the normal model, and 1.7401, 0.8017, and 1.2542 in the t_3 model.

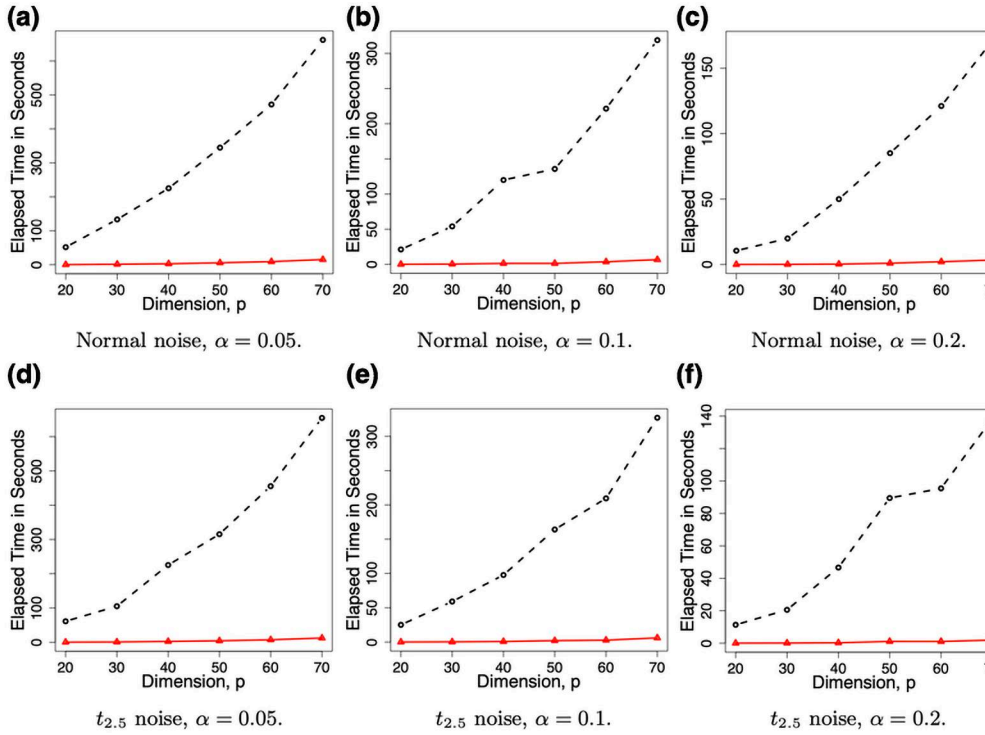

Figure 2. Average elapsed time (in seconds) over 50 replications for the proposed method implemented by a combination of `quantreg` and `adaHuber` and the joint regression approach implemented by `esreg` under $\mathcal{N}(0, 1)$ and $t_{2.5}$ error models when $\alpha \in \{0.05, 0.1, 0.2\}$. The sample size is set to be $n = \lceil 50p/\alpha \rceil$. The solid and dashed lines correspond to the proposed method and the joint regression approach, respectively.

that are related to the demographics of the individuals such as age, race, and the indicator variable that indicates whether the individual received JPTA training. After removing individuals with zero income, there are 4,576 males and 5,296 females. Our goal is to assess the effect of JPTA training on participants' earnings with an emphasis on the low-income population that is employed, for both male and female sub-groups.

Table 3. Empirical coverage probability and mean width (based on 500 replications) of 95% confidence intervals averaged over $p = 20$ variables when $n = \lceil 50p/\alpha \rceil$, $\alpha = \{0.05, 0.1, 0.2\}$, and $\varepsilon_i \sim \mathcal{N}(0, 1)$

$\mathcal{N}(0, 1)$	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.2$	
Method	Coverage	Width	Coverage	Width	Coverage	Width
2S-AH	0.950	0.595	0.949	0.660	0.948	0.744
joint	0.946	0.584	0.944	0.651	0.942	0.740
2S-LS	0.950	0.595	0.949	0.661	0.948	0.745

Table 4. Empirical coverage probability and mean width (based on 500 replications) of 95% confidence intervals averaged over $p = 20$ variables when $n = \lceil 50p/\alpha \rceil$, $\alpha = \{0.05, 0.1, 0.2\}$, and $\varepsilon_i \sim t_{2.5}$

$t_{2.5}$	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.2$	
Method	Coverage	Width	Coverage	Width	Coverage	Width
2S-AH	0.947	3.633	0.946	2.790	0.948	2.243
joint	0.959	5.771	0.959	3.571	0.954	2.872
2S-LS	0.952	4.521	0.950	3.397	0.953	2.687

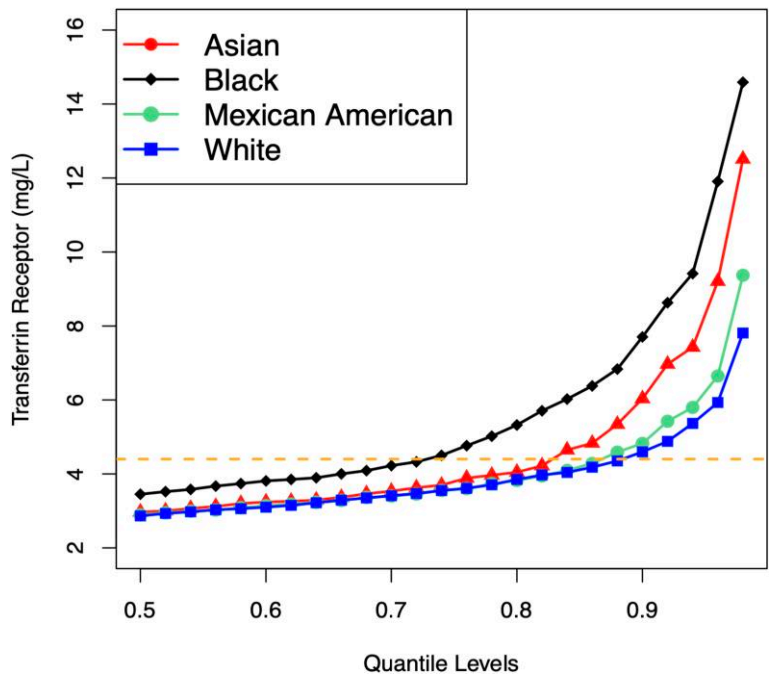

Figure 3. The soluble transferrin receptor levels (mg/L) vs. quantile levels (ranging from 0.5 to 0.99) for the female population in four different ethnic groups: Asian, Black, Mexican American, and White. The orange horizontal dashed line indicates the upper bound of the normal range (1.9–4.4 mg/L) for transferrin receptors among females.

To this end, we fit an expected shortfall regression model using the proposed robust method with $\alpha = \{0.05, 0.1, 0.2\}$. The robustification parameter τ is selected automatically via the procedure described in Section 3.2. Specifically, we regress the 30-month accumulated earnings on the

Table 5. The estimated regression coefficients (and 95% confidence intervals) for three dummy variables: Asian, Black, and Mexican American, using White as a baseline

	Asian	Black	Mexican American
QR	0.31 (−0.02, 0.64)	0.86 (0.37, 1.35)	−0.22 (−0.42, -0.01)
ES regression (`2S-AH`)	2.34 (0.59, 4.09)	3.03 (1.88, 4.19)	0.13 (−0.76, 1.03)

Note. Results of the upper-tail robust ES regression method `2S-AH` and standard QR at quantile level $\alpha = 0.75$ are reported.

JPTA training to assess the effect of JPTA training on low-income individuals, adjusting for whether individuals completed high school, race, Hispanic/non-Hispanic, marital status, working less than 13 weeks in the past year, and age. We report the estimated regression coefficient for the binary variable JPTA training and its associated 95% confidence intervals. The results are summarised in Table 6.

From Table 6, we see that 95% confidence intervals for the robust method do not contain zero for all $\alpha \in \{0.05, 0.1, 0.2\}$. This indicates that the JPTA training is statistically effective to improve earnings for the low-income population. Specifically, for the male sub-population, the estimated ES-effects of JPTA training are 283, 552, and 1,093 dollars at levels 0.05, 0.1, and 0.2, respectively. To further assess whether the estimated effects are scientifically meaningful, we compute the average 30-month accumulated earnings below the quantile levels 0.05, 0.1, and 0.2 for the male sub-group, which are 214, 566, and 1,496, respectively. We find that the JPTA training doubles the average income for individuals with income below the quantile levels 0.05 and 0.1, and becomes less effective for individuals with higher income. Similar findings are also observed for the female sub-group.

7 Conclusion and discussions

This paper considers expected shortfall regression under a joint quantile and ES model recently proposed in Dimitriadis and Bayer (2019) and Patton et al. (2019). The existing approach is based on a joint *M*-estimator, defined as the global minimum of any member of a class of strictly consistent joint loss functions (Fissler & Ziegel, 2016) over some compact set. Since the loss function is non-differentiable and non-convex, the computation of such a joint *M*-estimator is intrinsically difficult especially when the dimensionality is large. To circumvent the aforementioned challenge, Barendse (2020) proposed a two-step procedure for estimating the joint quantile and ES model based on Neyman orthogonalisation: the first step involves fitting the QR, and the second step employs the Neyman-orthogonal scores to estimate the ES parameters. Due to the use of L_2-loss in the second step, the resulting estimator is sensitive to heavy-tailed error distributions.

To address the robustness and computation concerns simultaneously, we propose a robust two-step method that applies adaptive Huber regression (Zhou et al., 2018) in the second step. The key is the use of a diverging robustification parameter for bias-robustness trade-off, tuned by a convenient data-driven mechanism. The proposed method can be efficiently implemented by a combination of `R` packages `quantreg`/`conquer` and `adaHuber`. The Python code that implements both our proposed methods and the existing non-convex optimisation-based methods (Dimitriadis & Bayer, 2019; Peng & Wang, 2022) is now publicly available at https://github.com/WenxinZhou/conquer. We establish a finite-sample theoretical framework for this two-step method, including deviation bound, Bahadur representation and (uniform) Gaussian approximations, in which the dimension of the model, p, may depend on and increase with the sample size, n. Robust confidence intervals/sets are also constructed. Numerical experiments further demonstrate that the proposed robust ES regression approach achieves satisfying statistical performance, high degree of robustness (against heavy-tailed data) and superior computational efficiency and stability. Through two data applications on health disparity and the JTPA study, we illustrate that ES regression complements QR as a useful tool to explore heterogeneous covariate effects on the average tail behaviour of the outcome.

Table 6. The estimated regression coefficient of the binary predictor JPTA training (and its 95% confidence interval) for the proposed robust method and the standard QR at quantile level $\alpha \in \{0.05, 0.1, 0.2\}$

Male sub-group	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
QR	465 (255, 675)	882 (603, 1,161)	2031 (1,431, 2,603)
ES regression (`2S-AH`)	283 (149, 418)	552 (333, 771)	1,093 (641, 1,546)
Female sub-group	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
QR	202 (76, 328)	480 (307, 653)	1,086 (719, 1,452)
ES regression (`2S-AH`)	123 (41, 205)	300 (146, 453)	672 (385, 958)

Note. Results are rounded to the closest integer.

Although we restrict attention to (joint) linear models in this work, our non-asymptotic theory and the underpinning techniques pave the way for analysing (i) series/projection estimators under joint non-parametric quantile-ES models and (ii) penalised estimators under high-dimensional sparse quantile-ES models. We leave these extensions in future research. One limitation in our data analysis for the JTPA study is that we do not account for potential selection bias. Specifically, as pointed out by Abadie et al. (2002), out of all subjects that were assigned to participate in the training programme, only approximately 60% of them (compliers) actually committed to the training programme. These individuals may simply have higher motivation in improving their earnings, and thus, the training status is likely positively correlated with potential income earnings. Generalising the proposed method to estimate the complier expected shortfall treatment effect, using an instrumental variable approach previously considered in Abadie et al. (2002), is another direction for future research.

The ES regression methods considered in this paper are suited for a fixed quantile level $\alpha \in (0, 1)$, independent of the sample size. For extreme quantiles satisfying $\alpha = \alpha_n \to 0$ or 1 as $n \to \infty$, both the FZ loss minimisation method (see equations (5) and (6)) and two-step procedures perform poorly because observations become scarce at that level, i.e. αn is not large enough. In fact, if dimension p is fixed, Online Supplementary Theorem A.1 and Theorem 1 imply that the two-step ES regression estimates, robust, and non-robust, are consistent if $\alpha_n^2 n \to \infty$ as $n \to \infty$. In the case where $\alpha_n^2 n = O(1)$, these methods are no longer useful and one may need to resort to extreme value theory (de Haan & Ferreira, 2006; H. J. Wang et al., 2012), which provides the statistical tools for a feasible extrapolation into the tail of the variable of interest. A more detailed discussion on modelling the extremes is deferred to the Online Supplementary Section E.

Acknowledgments

We sincerely thank the Joint Editor, Qiwei Yao, as well as the Associate Editor and the three reviewers for their numerous constructive comments and valuable suggestions, which have significantly helped us improve the quality of this work.

Conflict of interest: None declared.

Funding

X.H. is supported by NSF Grants DMS-1914496 and DMS-1951980. K.M.T. is supported by NSF DMS-2113346 and NSF CAREER DMS-2238428. W.-X.Z. acknowledges the support from NSF DMS-2113409.

Data availability

The two data sets used in Sections 6.2 and 6.3 are publicly available at https://www.cdc.gov/nchs/nhanes/index.htm and https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive, respectively.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Abadie A., Angrist J., & Imbens G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, *70*(1), 91–117. https://doi.org/10.1111/1468-0262.00270

Acerbi C., & Székely B. (2014). Back-testing expected shortfall. *Risk*, *27*(11), 76–81.

Acerbi C., & Tasche D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, *26*(7), 1487–1503. https://doi.org/10.1016/S0378-4266(02)00283-2

Andrews D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, *59*(2), 307–345. https://doi.org/10.2307/2938259

Barendse S. (2020). Efficiently weighted estimation of tail and interquantile expectations. *Preprint*. https://doi.org/10.2139/ssrn.2937665

Barzilai J., & Borwein J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, *8*(1), 141–148. https://doi.org/10.1093/imanum/8.1.141

Basel Committee. (2016). *Minimum capital requirements for market risk* (Technical Report). Bank for International Settlements. https://www.bis.org/bcbs/publ/d352.pdf.

Basel Committee. (2019). *Minimum capital requirements for market risk* (Technical Report). Bank for International Settlements. https://www.bis.org/bcbs/publ/d457.pdf.

Bassett G., Koenker R., & Kordas G. (2004). Pessimistic portfolio allocation and Choquet expected utility. *Journal of Financial Econometrics*, *2*(4), 477–492. https://doi.org/10.1093/jjfinec/nbh023

Belloni A., Chernozhukov V., Chetverikov D., & Fernández-Val I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, *213*(1), 4–29. https://doi.org/10.1016/j.jeconom.2019.04.003

Ben-Tal A., & Teboulle M. (1986). Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, *32*(11), 1445–1466. https://doi.org/10.1287/mnsc.32.11.1445

Bloom H., Orr L., Bell S., Cave G., Doolittle F., Lin W., & Bos J. (1997). The benefits and costs of JTPA Title II-A programs: Key findings from the national job training partnership act study. *The Journal of Human Resources*, *32*(3), 549–576. https://doi.org/10.2307/146183

Cai Z., & Wang X. (2008). Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics*, *147*(1), 120–130. https://doi.org/10.1016/j.jeconom.2008.09.005

Camaschella C. (2015). Iron-deficiency anemia. *New England Journal of Medicine*, *372*(19), 1832–1843. https://doi.org/10.1056/NEJMra1401038

Catoni O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, *48*(4), 1148–1185. https://doi.org/10.1214/11-AIHP454

Chen X. (2007). Chapter 76 large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, *6*, 5549-5632. https://doi.org/10.1016/S1573-4412(07)06076-X

Chernoff H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507. https://doi.org/10.1214/aoms/1177729330

Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W., & Robins J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov V., & Hansen C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, *142*(1), 379–398. https://doi.org/10.1016/j.jeconom.2007.06.005

Cont R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*(2), 223–236. https://doi.org/10.1080/713665670

de Haan L., & Ferreira A. (2006). *Extreme value theory: An introduction*. Springer-Verlag.

Dimitriadis T., & Bayer S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, *13*(1), 1823–1871. https://doi.org/10.1214/19-EJS1560

Du Z., & Escanciano J. C. (2017). Backtesting expected shortfall: Accounting for tail risk. *Management Science*, *63*(4), 940–958. https://doi.org/10.1287/mnsc.2015.2342

Eubank R. L., & Spiegelman C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association*, *85*(410), 387–392. https://doi.org/10.1080/01621459.1990.10476211

Fan J., Li Q., & Wang Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(1), 247–265. https://doi.org/10.1111/rssb.12166

Farrell M. H., Liang T., & Misra S. (2021). Deep neural networks for estimation and inference. *Econometrica*, *89*(1), 181–213. https://doi.org/10.3982/ECTA16901

Fernandes M., Guerre E., & Horta E. (2021). Smoothing quantile regressions. *Journal of Business & Economic Statistics*, *39*(1), 338–357. https://doi.org/10.1080/07350015.2019.1660177
Fissler T., & Ziegel J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, *44*(4), 1680–1707. https://doi.org/10.1214/16-AOS1439
Gneiting T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762. https://doi.org/10.1198/jasa.2011.r10138
Hahn M. G., Kuelbs J., & Weiner D. C. (1990). The asymptotic joint distribution of self-normalized censored sums and sums of squares. *The Annals of Probability*, *18*(3), 1284–1341. https://doi.org/10.1214/aop/1176990747
He X., Hsu Y.-H., & Hu M. (2010). Detection of treatment effects by covariate-adjusted expected shortfall. *The Annals of Applied Statistics*, *4*(4), 2114–2125. https://doi.org/10.1214/10-AOAS347
He X., Pan X., Tan K. M., & Zhou W.-X. (2023). Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, *232*(2), 367–388. https://doi.org/10.1016/j.jeconom.2021.07.010
Huber P. J. (1973). Robust estimation: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, *1*(5), 799–821. https://doi.org/10.1214/aos/1176342503
Kato K. (2012). Weighted Nadaraya–Watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, *10*(2), 265–291. https://doi.org/10.1093/jjfinec/nbs002
Koenker R., & Bassett G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50. https://doi.org/10.2307/1913643
Kratovil T., DeBerardinis J., Gallagher N., Luban N., Soldin S., & Wong E. (2007). Age specific reference intervals for soluble transferrin receptor (sTfR). *Clinica Chimica Acta*, *380*(1-2), 222–224. https://doi.org/10.1016/j.cca.2007.02.012
Linton O., & Xiao Z. (2013). Estimation and inference about the expected shortfall for time series with infinite variance. *Econometric Theory*, *29*(4), 771–807. https://doi.org/10.1017/S0266466612000692
Martins-Filho C., Yao F., & Torero M. (2018). Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. *Econometric Theory*, *34*(1), 23–67. https://doi.org/10.1017/S0266466616000517
Mast A., Blinder M., Gronowski A., Chumley C., & Scott M. (1998). Clinical utility of the soluble transferrin receptor and comparison with serum ferritin in several populations. *Clinical Chemistry*, *44*(1), 45–51. https://doi.org/10.1093/clinchem/44.1.45
McNeil A. J., Frey R., & Embrechts P. (2015). *Quantitative risk management: Concepts, techniques and tools* (2nd ed.). Princeton University Press.
Nemirovski A., & Yudin D. (1983). *Problem complexity and method efficiency in optimization*. Wiley.
Newey W. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, *79*(1), 147–168. https://doi.org/10.1016/S0304-4076(97)00011-0
Neyman J. (1979). $C(\alpha)$ tests and their use. *Sankhya*, *41*(1/2), 1–21. http://www.jstor.org/stable/25050174
Pan X., & Zhou W.-X. (2021). Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design. *Information and Inference: A Journal of the IMA*, *10*(3), 813–861. https://doi.org/10.1093/imaiai/iaaa006
Patton A. J., Ziegel J. F., & Chen R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, *211*(2), 388–413. https://doi.org/10.1016/j.jeconom.2018.10.008
Peng X., & Wang H. J. (2022). 'Inference for joint quantile and expected shortfall regression', arXiv, arXiv:2208.10586, preprint: not peer reviewed.
Rockafellar R. T., & Royset J. O. (2014). Superquantiles and their applications to risk, random variables, and regression. *INFORMS TutORials in Operations Research, null*(null), 151–167. https://doi.org/10.1287/educ.2013.0111
Rockafellar R. T., Royset J. O., & Miranda S. I. (2014). Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, *234*(1), 140–154. https://doi.org/10.1016/j.ejor.2013.10.046
Rockafellar R. T., & Uryasev S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, *2*(3), 21–41. http://doi.org/10.21314/JOR.2000.038
Rockafellar R. T., & Uryasev S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, *26*(7), 1443–1471. https://doi.org/10.1016/S0378-4266(02)00271-6
Rockey D., & Cello J. (1993). Evaluation of the gastrointestinal tract in patients with iron-deficiency anemia. *New England Journal of Medicine*, *329*(23), 1691–1695. https://doi.org/10.1056/NEJM199312023292303
Scaillet O. (2004). Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance*, *14*(1), 115–129. https://doi.org/10.1111/j.0960-1627.2004.00184.x
Schmidt-Hieber J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, *48*(4), 1875–1897. https://doi.org/10.1214/19-AOS1875
Shapiro A., Dentcheva D., & Ruszczynski A. (2014). *Lectures on stochastic programming: Modeling and theory* (2nd ed.). SIAM.

Shen G., Jiao Y., Lin Y., Horowitz J. L., & Huang J. (2021). 'Deep quantile regression: Mitigating the curse of dimensionality through composition', arXiv, arXiv:2107.04907, preprint: not peer reviewed.

Wang H. J., Li D., & He X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, *107*(500), 1453–1464. https://doi.org/10.1080/01621459.2012.716382

Wang L., Zheng C., Zhou W., & Zhou W.-X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica*, *31*(4), 2153–2177. https://doi.org/10.5705/ss.202019.0045

Zhou W.-X., Bose K., Fan J., & Liu H. (2018). A new perspective on robust *M*-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics*, *46*(5), 1904–1931. https://doi.org/10.1214/17-AOS1606