Asymptotically efficient estimation of the conditional expected shortfall[☆]

Samantha Leorato[a,*], Franco Peracchi[a,b], Andrei V. Tanase[a]

[a] *University of Rome Tor Vergata, Italy*
[b] *EIEF, Rome, Italy*

Abstract

We propose a procedure for efficient estimation of the trimmed mean of a random variable Y conditional on a set of covariates X. For concreteness, we focus on a financial application where the trimmed mean of interest corresponds to a coherent measure of risk, namely the conditional expected shortfall. Our estimator is based on the representation of the estimand as an integral of the conditional quantile function. We extend the class of estimators originally proposed by Peracchi and Tanase (2008) by introducing a weighting function that gives different weights to different conditional quantiles. Our approach allows for either parametric or nonparametric modeling of the conditional quantiles and the weights, but is essentially nonparametric in spirit. We prove consistency and asymptotic normality of the resulting estimator. Optimizing over the weighting function, we obtain asymptotic efficiency gains with respect to the unweighted estimators. The gains are especially noticeable in the case of fat-tailed distributions.

Keywords: expected shortfall, quantile regression, convex optimization, asymptotic efficiency.

1. Introduction

Quantile regression, introduced by Roger Koenker and Gib Bassett (Koenker and Bassett [15]), has gradually evolved from a robust alternative to least squares to a way of summarizing the conditional distribution of a random variable given a set of covariates. As such, it can be used in a large variety of situations. In this paper we employ quantile regression methods to estimate the trimmed mean of a random variable of interest conditional on a set of covariates. Trimmed means are widely used as alternative location parameters to the ordinary mean because of their robustness and their superior properties under certain types of censoring. They are usually not of direct interest, however, in the sense that, absent other considerations such as robustness or censoring, one would be perfectly happy with the ordinary mean. Here we focus instead on a

[☆] We thank Iván Fernández-Val and Michael Lechner for useful suggestions.

[*] Corresponding author
Email addresses: samantha.leorato@uniroma2.it (Samantha Leorato), franco.peracchi@uniroma2.it (Franco Peracchi), andrei.valentin.tanase@uniroma2.it (Andrei V. Tanase)

financial application where the trimmed mean is of substantive interest in itself as a coherent measure of risk.

Specifically, let the continuous random variable Y_t represent the uncertain return between time t and time $t+1$ on a single asset or a portfolio of assets, and let the set of covariates X_t represent the relevant information available up to time t. This information typically consists of lagged values of other financial or nonfinancial variables, possibly including lagged values of Y_t. The trimmed mean of interest is the mean over the left tail of the conditional distribution of Y_t given $X_t = x$ up to the αth conditional quantile, with $0 < \alpha < 1$, namely

$$\tau(\alpha \,|\, x) = \frac{1}{\alpha} \int_{-\infty}^{Q(\alpha \,|\, x)} y \, \mathrm{d}F(y \,|\, x),$$

where $Q(\alpha \,|\, x)$ is the αth conditional quantile function (CQF) and $F(y \,|\, x)$ the conditional distribution function (CDF) of Y_t given $X_t = x$. In the financial literature, $\tau(\alpha \,|\, x)$ is known as the α-level conditional expected shortfall (CES) of Y_t given $X_t = x$, with α typically set to .05 or .10. The expected shortfall corresponds to the expected loss on holding the asset or the portfolio, given that the loss exceeds the αth quantile of Y_t, a quantity known in the financial literature as the $(1-\alpha)$-percent Value-at-Risk (VaR). The expected shortfall is an increasingly popular measure of risk because it is continuous in α and, unlike the VaR, is coherent, that is, it simultaneously satisfies sub-additivity, monotonicity, positive homogeneity and translation invariance (Artzer et al. [4]). For further references, see Acerbi and Tasche [1], Delbaen [11], and Bertsimas et al. [6], among others. The conditional versions of the expected shortfall and the VaR provide a natural way of incorporating information on economic and market conditions.

A nonparametric estimator of the CES has been proposed by Cai and Wang [7]. Their estimator, called weighted double kernel local linear estimator, combines the attractive features of the double-kernel local linear estimator of Fan and Gijbels [12] and of the weighted Nadaraya-Watson estimator (WNW) of Hall et al. [14], especially the monotonicity and the good boundary behavior of the latter. The main drawback of this estimator is its complexity and the fact that its rate of convergence is slow and decays rapidly with the number of covariates reflecting the curse-of-dimensionality problem.

In this paper we propose an alternative estimator based on the fact that the CES may equivalently be represented as the following integral of the CQF

$$\tau(\alpha \,|\, x) = \frac{1}{\alpha} \int_0^{\alpha} Q(p \,|\, x) \, \mathrm{d}p.$$

Our estimator of the CES generalizes the integrated conditional quantile function (ICQF) estimator proposed by Peracchi and Tanase [20] in two directions. First, we consider a weighted version of

the estimator characterized by higher asymptotic efficiency. Second, we embed the estimators in a broader class that includes semi-parametric estimators. Although we focus on estimating the CES, our method applies with minor changes to more general trimmed means, for example two-sided trimmed means with limits defined by conditional quantiles or other functions of X_t.

More precisely, given a function $W\colon [0,\alpha] \mapsto [0,1]$, replacing the CDF $F(y \mid x)$ in the definition of the CES by its transformed version $W(F(y \mid x))$ gives the weighted conditional expected shortfall (WCES), which may equivalently be represented as an integral of the CQF with nonuniform weights. The idea of introducing a set of weights to increase asymptotic efficiency when estimating a population parameter of interest is widely used in parametric and nonparametric statistics, and is a key feature of generalized method of moments and minimum distance methods. To our knowledge, however, the idea of applying a weighting function to estimate efficiently a functional of the CQF is new.

The use of a weighted version of the CDF in the definition of the WCES is also related to the theory of non-expected utility of Yaari [23] and Prelec [21], where modifying the distribution of the returns accommodates risk aversion of the investor. We do not pursue this subjective interpretation of the WCES and confine ourselves to weighting as a way of improving asymptotic efficiency of estimation. Intuitively, introducing a weighting function enables one to compensate the inefficiency of quantile estimators at very extreme quantiles by giving more weight to the quantiles near α, which are estimated more precisely. Weighting does not affect consistency of estimation because the weighting function is chosen such that the WCES and the CES coincide.

Building on results by Peracchi and Tanase [20], we consider a class of analog estimators based on representing the WCES as an integral of the CQF with nonuniform weights. These estimators are called weighted integrated conditional quantile function (WICQF) estimators. Our approach allows for either parametric or nonparametric modeling of the CQF and the weights, but is essentially nonparametric in spirit.

A drawback of conventional quantile regression estimators is that they do not guarantee monotonicity. In our case, the problem is likely to be mitigated by integration over the conditional quantiles. In the case of linear quantile regression estimators, the linearity assumption is an additional problem. Angrist et al. [3] show that, although biased under misspecification, linear quantile regression estimators remain asymptotically normal. In general, despite its limitations, the linear quantile regression model is widely used because of parsimony and computational convenience. Further, estimates of a linear quantile regression model can be used as preliminary non-monotonic curves to be rearranged according to the method recently proposed by Chernozhukov et al. [9]. For

these reasons, although presenting the asymptotic results for arbitrary estimators of the CQF, we focus on the case when the quantile regression model is linear.

The remainder the paper is organized as follows. Section 2 defines the WCES and suggests a class of analog estimators based on a given weighting function. Section 3 analyzes the asymptotic behavior of the proposed estimators for a fixed grid. Section 4 studies their asymptotic behavior for a data-dependent grid. Section 5 considers choosing the weighting function in order to maximize asymptotic efficiency. Section 6 presents a study of the asymptotic efficiency gains that can be attained for different distributions of the returns. Section 7 present an application to real data to highlight the potentials of our procedure. Finally, Section 8 concludes. All proofs are collected in the appendix.

2. Estimation of the WCES

The weighted conditional expected shortfall (WCES) is the weighted version of the conditional expected shortfall (CES). It is obtained by modifying the conditional distribution of the returns on an asset/portfolio through a suitable transformation or, equivalently, by giving nonuniform weights to the various conditional quantiles.

Let the continous random variable Y_t represent the uncertain return on a given asset or portfolio between time t and time $t+1$, let the k-dimensional random vector X_t represent the relevant information about Y_t available up to time t, let $F(y\,|\,x) = \Pr\{Y_t \le y\,|\,X_t = x\}$ and $Q(\alpha\,|\,x) = \inf\{y\colon F(y\,|\,x) \ge \alpha\}$, $\alpha \in (0,1)$, respectively denote the CDF and the CQF of Y_t given $X_t = x$, and let W be a function with the following properties:

A.1 The function $W\colon [0,\alpha] \mapsto [0,1]$ is continuous, nondecreasing, and satisfies $W(0) = 0$ and $W(\alpha) = 1$.

Definition 1. *For every function W satisfying A.1, the α-level WCES is defined as*

$$\tau_\alpha(x) = \int_0^{Q(\alpha\,|\,x)} y \,\mathrm{d}W(F(y\,|\,x)). \tag{2.1}$$

Any continuous distribution function with support $[0,\alpha]$ can be used to define a WCES. If the function W is concave on $[0,\alpha]$, then the WCES is a coherent risk measure, just like the ordinary CES (see Acerbi 2002).

Now suppose that the function W is regular in the following sense:

A.2 The function W is continuously differentiable on $(0,\alpha)$ with nonnegative derivative $W' = w$.

A function w that satisfies A.2 is a density on $(0, \alpha)$. If A.2 holds then, after a change of variable, we obtain the following equivalent representation of the WCES

$$\tau_\alpha(x) = \int_0^\alpha Q(p \,|\, x) w(p) \, \mathrm{d}p. \tag{2.2}$$

The weighted unconditional expected shortfall and the unweighted CES correspond, respectively, to the case when $X_t = x$ with probability one and the case when $w(p) = \alpha^{-1}$. Notice that the WCES is a function of x for a given α (e.g. $\alpha = .05$ or $.10$). The dependence of the WCES on w will be made explicit only in Section 5, where we discuss how to optimally choose w.

Consistency requires that, in addition to satisfying properties A.1 and A.2, the weight function w must be such that the WCES and the CES coincide. If we impose the condition that WCES = CES, then w must also satisfy the integral equation

$$\int_0^\alpha Q(p \,|\, x) \left(w(p) - \frac{1}{\alpha} \right) \mathrm{d}p = 0. \tag{2.3}$$

The set of weighting functions w satisfying A.2 and (2.3) is

$$\left\{ w(p),\, p \in (0, \alpha] \colon w(p) \geq 0,\, \int_0^\alpha w(p) \, \mathrm{d}p = 1,\, \int_0^\alpha Q(p \,|\, x) \left(w(p) - \frac{1}{\alpha} \right) \mathrm{d}p = 0 \right\}. \tag{2.4}$$

This set may consist of parametric functions, indexed by a finite number of parameters, or of nonparametric functions. Later in Section 5 we permit the weighting function w to take negative values on $(0, \alpha]$, thus replacing the set in (2.4) by the larger set

$$\left\{ w(p),\, p \in (0, \alpha] \colon \int_0^\alpha w(p) \, \mathrm{d}p = 1,\, \int_0^\alpha Q(p \,|\, x) \left(w(p) - \frac{1}{\alpha} \right) \mathrm{d}p = 0 \right\}.$$

Relaxing property A.2 in this way enables us to obtain a closed-form expression for the efficient estimator of the CES (see Theorem 4 below). However, because w is no longer a proper weighting function, the resulting estimator cannot be interpreted as a proper average of the conditional quantiles up to level α.

We now present our class of estimators of the CES. Because they rely on representation (2.2), we call them weighted integrated conditional quantile function (WICQF) estimators. Construction of these estimators is carried out in two steps. In the first step we consider a discrete approximation to $\tau_\alpha(x)$ over a grid of points $0 \leq p_0 < p_1 < \cdots < p_I = \alpha$. A data driven choice of the grid will be discussed in Section 4. Then, in the second step, we estimate the discrete approximation by an analog estimator.

Our discrete approximation to $\tau_\alpha(x)$ is obtained by replacing the integral in (2.2) with the finite sum

$$\tau_\alpha^*(x) = \sum_{i=1}^{I} w_i Q_i(x),$$

where $w_i = W(p_i) - W(p_{i-1})$ and $Q_i(x) = Q(p_i \,|\, x)$. Our estimator of the WCES is then obtained by replacing the unknown conditional quantiles $Q_i(x)$ with consistent estimators $\hat{Q}_i(x) = \hat{Q}(p_i \,|\, x)$. The resulting estimator is

$$\hat{\tau}_\alpha(x) = \sum_{i=1}^{I} w_i \hat{Q}_i(x). \tag{2.5}$$

Different estimators of $Q_i(x)$ may be considered, either parametric or semiparametric.

Things are straightforward when the $Q_i(x)$ are specified as linear in parameters, that is, $Q_i(x) = \beta_i^\top x$. In this case, the WICQF estimator takes the particularly simple form

$$\hat{\tau}_\alpha(x) = \bar{\beta}^\top x, \tag{2.6}$$

where $\bar{\beta} = \sum_{i=1}^{I} w_i \hat{\beta}_i$ and the vectors $\hat{\beta}_i$ are obtained by minimizing the asymmetric absolute loss function

$$\sum_{t=1}^{T} \ell_{p_i}(Y_t - \beta^\top X_t) = \sum_{t=1}^{T} (Y_t - \beta^\top X_t)(p_i - \mathbb{1}\{Y_t - \beta^\top X_t < 0\}), \qquad i = 1, \ldots, I$$

(see Koenker [16]). A drawback of this estimator is that the estimated conditional quantiles need not be monotonic, that is, there may exist $p_j < p_i$ such that $\hat{Q}_j(x) > \hat{Q}_i(x)$ for some x. When estimated conditional quantiles cross each other, monotonicity of $\hat{\tau}_\alpha(x)$ is lost. However, this effect is likely to be mitigated by the fact that $\hat{\tau}_\alpha(x)$ is an average of the I estimated conditional quantiles $\hat{Q}_1(x), \ldots, \hat{Q}_I(x)$. To avoid the drawbacks of linear quantile regression estimators, semiparametric estimators of the $Q_i(x)$ may also be considered.

The choice of weights in the discrete approximation $\tau_\alpha^*(x)$ to $\tau_\alpha(x)$ is simple and intuitive. The asymptotic results in the next sections hold, more generally, for all approximations where the weights sum to one. These include, for example, generalizations of the approximation $I^{-1} \sum_{i=1}^{I} \bar{Q}_i(x)$, where $\bar{Q}_i(x) = [Q_i(x) + Q_{i-1}(x)]/2$, such as $\sum_{i=1}^{I} w_i \bar{Q}_i(x) = \sum_{i=0}^{I} \bar{w}_i Q_i(x)$, where $\bar{w}_0 = [W(p_1) - W(p_0)]/2$, $\bar{w}_I = [W(p_I) - W(p_{I-1})]/2$, and $\bar{w}_i = [W(p_{i+1}) - W(p_{i-1})]/2$ for $i = 1, \ldots, I-1$. This choice of weights is more convenient when the approximation error is of concern. Thus it will be used in Section 4, where we let the size of the grid increase with the dimension of the sample, to get a better rate in terms of approximation bias and to obtain less restrictive conditions on the rate of convergence of the sequence of grid-points.

3. Asymptotic properties

Because a WICQF estimator has essentially the structure of an L-statistic, consistency and asymptotic normality may be derived following with minor modifications the method adopted by Csörgö et al. [10] and Mason and Shorack [17].

In this section we treat W and w as given and consider the asymptotic behavior of the WICQF estimator when the grid of points $p_0, \ldots, p_I$ is fixed and does not depend on the available data. In this case, we are allowed to assume $p_0 = 0$. Section 4 deals with the case when the dimension I of the grid increases with the sample size, whereas Section 5 considers the problem of choosing w optimally.

Throughout this section we weaken the requirements on the function W to the following condition:

A.3 The function $W\colon [0,\alpha] \mapsto [0,1]$ satisfies $W(0) = 0$ and $W(\alpha) = 1$, and is continuously differentiable on $(0,\alpha)$ with derivative $w = W'$.

Condition A.3 does not require the weighting function w to be nonnegative. This is useful because it enables us to obtain in Section 5 a closed-form expression for the asymptotically efficient estimator.

Given a function $W\colon [0,\alpha] \mapsto [0,1]$, we define the vector of weights $\mathbf{w} = (w_1, \ldots w_I)^\top$, where $w_i = W(p_i) - W(p_{i-1})$ with $0 = p_0 < p_1 < \cdots < p_I \leq \alpha$. Let $\tau^*_\alpha(x) = \sum_{i=1}^I w_i Q_i(x)$ denote the approximation to the WCES by a finite sum. The next theorem gives sufficient conditions for $\hat{\tau}_\alpha(x)$ to be a $\sqrt{T}$-consistent and asymptotically normal estimator of the discrete approximation $\tau^*_\alpha(x)$ when a $\sqrt{T}$-consistent and asymptotically normal estimator of the CQF is used in (2.5). The conditions of the theorem are very general. In particular, they cover the case when the data $\{(X_t, Y_t), t = 1, \ldots, T\}$ are independently and identically distributed (iid). They also cover the case when the data consist of T identically distributed and weakly dependent observations, which is a more relevant case for the financial applications in Section 7, where it is natural to expect some form of time dependence. The proof of the theorem is straightforward and is therefore omitted.

Theorem 1. *Let $\hat{Q}(p \,|\, x)$ be an estimator of $Q(p \,|\, x)$ that is $\sqrt{T}$-consistent for all $p \in (0,\alpha]$, and assume that for every I-tuple $(p_1, \ldots, p_I)$ the random vector $\{\sqrt{T}[\hat{Q}_i(x) - Q_i(x)], i = 1, \ldots, I\}$, converges in distribution to a multivariate Gaussian vector with mean zero and covariance matrix $\boldsymbol{V}$. Then*

$$\sqrt{T}[\hat{\tau}_\alpha(x) - \tau^*_\alpha(x)] \xrightarrow{d} \mathcal{N}(0, \mathbf{w}^\top \boldsymbol{V} \mathbf{w}),$$

The next result considers the case when $\hat{\tau}_\alpha(x)$ is a weighted sum of linear quantile regression estimators. Although our presentation assumes for simplicity that the (X_t, Y_t) are iid, all the results can be shown to hold also for a weakly dependent sequence of random vectors. For example,

Theorem 2.2 of Fitzenberger [13] establishes asymptotic normality of linear quantile regression estimators under a strongly mixing data generating process. Let $\bar{\tau}_\alpha(x) = \sum_{i=1}^I w_i \beta_i^\top x$, where

$$\beta_i = \arg\min_\beta \mathbb{E}\, \ell_{p_i}(Y - \beta^\top X), \qquad i = 1, \ldots, I. \tag{3.1}$$

Clearly $\bar{\tau}_\alpha(x) \neq \tau_\alpha^*(x)$ unless the linear specification of conditional quantiles is correct. In this case, although $\hat{\tau}_\alpha(x)$ estimates the "wrong" estimand, its asymptotic normality is unaffected because linear quantile regression estimators remain asymptotically normal (Angrist et al. [3]).

Theorem 2. *Suppose that:*

(i) $F(\cdot \,|\, x)$ *is absolutely continuous, with continuous density* $f(\cdot \,|\, x)$ *uniformly bounded away from zero at all points* $y \in [\varepsilon, 1-\varepsilon]$ *for every* $\varepsilon > 0$.

(ii) $\lim_{T\to\infty} T^{-1} \sum_t X_t X_t^\top = \mathbf{D}$, *where* $\mathbf{D}$ *is a finite positive definite matrix.*

(iii) For any $0 < p_i < 1$, $\lim_{T\to\infty} T^{-1} \sum_t f(\beta_i^\top X_t \,|\, X_t) X_t X_t^\top = \mathbf{J}_i$, *where* $\mathbf{J}_i$ *is a finite positive definite matrix.*

(iv) $\max_{t=1,\ldots,T} \|X_t\|/\sqrt{T} \to 0$, *where* $\|\cdot\|$ *is the Euclidean norm of a vector in* $\mathbb{R}^k$.

Then Theorem 1 holds for $\sqrt{T}[\hat{\tau}_\alpha(x) - \bar{\tau}_\alpha(x)]$ *with* $\boldsymbol{V} = (\mathcal{I}_I \otimes x^\top)\, \boldsymbol{\Omega}\, (\mathcal{I}_I \otimes x)$, *where* $\mathcal{I}_I$ *is the* I*-dimensional identity matrix,* $\boldsymbol{\Omega}$ *is an* $Ik \times Ik$ *matrix consisting* $k \times k$ *submatrices of the form*

$$\Omega_{ij} = \mathbf{J}_i^{-1}\, \Sigma_{ij}\, \mathbf{J}_j^{-1}, \tag{3.2}$$

and

$$\Sigma_{ij} = \mathbb{E}[(p_i - \mathbf{1}\{Y < \beta_i^\top X_t\})(p_j - \mathbf{1}\{Y_t < \beta_j^\top X_t\}) X_t X_t^\top] \tag{3.3}$$

is a positive definite $k \times k$ *matrix for all* i *and* j. *If the linear quantile regression model is correctly specified, that is* $Q_i(x) = \beta_i^\top x$, *then* $\Sigma_{ij} = [\min(p_i, p_j) - p_i p_j](\mathbb{E}\, X_t X_t^\top)$.

Except for (i), that requires the conditional distribution of Y_t given $X_t = x$ not to depend on t, the other conditions in Theorem 2 are the same as those in Theorem 4.1 of Koenker [16]. The next result gives a consistent estimator of the asymptotic variance $\mathrm{AV}(\hat{\tau}_\alpha(x)) = \mathbf{w}^\top \boldsymbol{V} \mathbf{w}$, which is needed for inference.

Corollary 1. *If the model is correctly specified, a consistent estimator of* $\mathrm{AV}(\hat{\tau}_\alpha(x))$ *is*

$$\widehat{\mathrm{AV}}(\hat{\tau}_\alpha(x)) = x^\top \left\{ \sum_{i=1}^I \sum_{j=1}^I w_i w_j \left[\min(p_i, p_j) - p_i p_j\right] \hat{\mathbf{J}}_i^{-1}\, \hat{\mathbf{D}}\, \hat{\mathbf{J}}_j^{-1} \right\} x, \tag{3.4}$$

with $\hat{\mathbf{D}} = T^{-1}\sum_{t=1}^{T} X_t X_t^\top$ *and* $\hat{\mathbf{J}}_i = T^{-1}\sum_{t=1}^{T} \hat{f}(\hat{\beta}_i^\top X_t \,|\, X_t)\, X_t X_t^\top$*, where* $\hat{f}$ *is a consistent estimator of the conditional density* f*.*

The proof of the corollary follows immediately from consistency of $\hat{\mathbf{D}}$ and $\hat{\mathbf{J}}_i$ for, respectively, $\mathbf{D}$ and $\mathbf{J}(p_i)$.

4. Data-dependent choice of grid

We now consider the case where the grid of points used to construct the WICQF estimator is allowed to depend on the data and to increase with the sample size T. Our aim is to study the asymptotic behavior of $\hat{\tau}_\alpha(x)$ under this assumption. Conditions on the rate at which the number of grid-points should grow to infinity as $T \to \infty$ provide a rough criterion to find the appropriate dimension of the grid for a given sample size.

Given estimates $\hat{Q}_0(x), \ldots, \hat{Q}_I(x)$ of the value of the CQF $Q(\cdot \,|\, x)$ at the grid-points $p_0, \ldots, p_I$, with $p_0 = \alpha_0$ and $p_I = \alpha$, we construct an estimate $\hat{Q}(\cdot \,|\, x)$ of the CQF over the interval $(\alpha_0, \alpha]$, by setting

$$\hat{Q}(p \,|\, x) = \frac{\hat{Q}_i(x) + \hat{Q}_{i-1}(x)}{2}, \qquad p \in (p_{i-1}, p_i], \qquad i = 1, \ldots, I. \tag{4.1}$$

With this definition, the estimated CQF is a step function on $(\alpha_0, \alpha]$ with jumps at $p_0, p_1, \ldots, p_I$. If $w(p) = W'(p)$, then

$$\int_{\alpha_0}^{\alpha} \hat{Q}(p \,|\, x) w(p) \mathrm{d}p = \sum_{i=0}^{I} \bar{w}_i \hat{Q}_i(x),$$

where $\bar{w}_0 = [W(p_1) - W(p_0)]/2$, $\bar{w}_I = [W(p_I) - W(p_{I-1})]/2$, and $\bar{w}_i = [W(p_{i+1}) - W(p_{i-1})]/2$ for $i = 1, \ldots, I-1$. As already pointed out, this choice of weights enables us to weaken the restrictions on the rate of convergence of I to infinity with respect to the simpler alternative of setting $\hat{Q}(p \,|\, x) = \hat{Q}_i(x)$, $p \in (p_{i-1}, p_i]$, $i = 1, \ldots, I$.

Throughout this section we assume that, for every fixed x, $\hat{Q}(p \,|\, x)$ is uniformly consistent in p for $Q(p \,|\, x)$ and $\sqrt{T}[\hat{Q}(p \,|\, x) - Q(p \,|\, x)]$ converges weakly in $\mathcal{C}_{(0,\alpha]}$ (the class of continuous function on $(0, \alpha]$) to a Gaussian process with covariance function $V(r, s)$, $r, s \in (0, \alpha]$. This is a stronger assumption than that of Theorem 1, as it is related to stochastic equicontinuity of the functional quantile estimator.

To avoid complications related to the asymptotic behavior of extremal conditional quantile estimators, we limit ourselves to the case when the grid is bounded below by a fixed positive number α_0. The case when $\alpha_0 \to 0$ is not dealt with in this paper. However, when $\alpha_0 T \to \infty$, namely in the case of intermediate order quantiles, the linear quantile regression model $\hat{Q}(\alpha_0 \,|\, x)$

is still asymptotically normal under appropriate conditions on the behavior of the density quantile function $q(\alpha_0 \,|\, x) = 1/f(Q(\alpha_0 \,|\, x))$ (see Chernozukov [8]).

As for the sequence of grid-points, we assume that $0 < \alpha_0 = p_0 < p_1 < \cdots < p_I = \alpha$ and

$$\frac{c_1}{I} = \liminf_I (p_i - p_{i-1}) \leq \limsup_I (p_i - p_{i-1}) = \frac{c_2}{I},$$

where $c_1 < c_2$. This condition is satisfied for the equally-spaced grid, with $p_i - p_{i-1} = (\alpha - \alpha_0)/I$ for all $i = 1, \ldots, I$.

Theorem 3. *Assume that:*

(i) For every fixed x, $Q(p \,|\, x)$ is differentiable in p on $(0, \alpha]$ and its derivative $q(p \,|\, x) = \partial Q(p \,|\, x)/\partial p$ is nondecreasing on $(0, \alpha]$ and bounded on $(\alpha_0, \alpha]$.

(ii) The function W satisfies conditions A.1 and A.2 or A.3 and $|w|$ is bounded above in $(0, \alpha]$.

(iii) The sequence of grid-points $0 < \alpha_0 = p_0 < p_1 < \cdots < p_I = \alpha$ is such that $\lim_{T\to\infty} \sqrt{T}/I^2 = 0$.

Then $\sqrt{T}[\hat{\tau}_\alpha(x) - \tau_\alpha(x)] \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \int_{\alpha_0}^{\alpha} \int_{\alpha_0}^{\alpha} w(p)w(s)V(p,s)\,\mathrm{d}p\,\mathrm{d}s$.

If $\hat{Q}(p \,|\, x)$ is a linear quantile regression estimator and $\{(X_t, Y_t)\}$ are iid, then assumptions (i)–(iv) of Theorem 2 guarantee weak convergence of the quantile regression process, with

$$V(r,s) = x^\top \mathbf{J}^{-1}(r)\, \mathbb{E}[(r - \mathbb{1}\{Y_t < \beta(r)^\top X_t\})(s - \mathbb{1}\{Y_t < \beta(s)^\top X_t\}) X_t X_t^\top] \mathbf{J}^{-1}(s) x$$

(see Angrist et al. [3]). If the CQF is itself linear, then condition (i) in Theorem 3 is equivalent to the assumption that the population quantile regression coefficient $\beta(p)$ is differentiable and concave on $(0, \alpha]$.

5. Asymptotic efficient WICQF estimators

This section presents our proposal for efficient estimation of the CES via minimization of the asymptotic variance of the WICQF estimator with respect to the choice of weights. Throughout this section the grid-points $p_0, p_1, \ldots, p_I$ are taken as fixed.

We begin by discretizing the problem. Given a function W satisfying either A.1-A.2 or A.3 and an integer I, we consider the vector of weights $\mathbf{w} = (w_1, \ldots, w_I)^\top$, where $w_i = W(p_i) - W(p_{i-1})$. The vector of uniform weights corresponds to the choice $W(p) = p/\alpha$ and $p_i = \alpha i/I$, which gives $w_i = 1/I$. To stress the dependence of the estimand and the estimator on these weights, we omit

the explicit reference to the level α and to the covariate value x and simply write $\tau^*_\alpha(x)$ as $\tau(\mathbf{w})$ and $\hat{\tau}_\alpha(x)$ as $\hat{\tau}(\mathbf{w})$. We say that a vector of weights is consistent if the discrete versions of the weighted and the unweighted conditional expected shortfall coincide, that is, if $\sum_{i=1}^{I} w_i Q_i(x) = I^{-1}\sum_{i=1}^{I} Q_i(x)$. Given I, the set of consistent weights is the following subset of the I-dimensional simplex

$$\mathcal{W} = \left\{\mathbf{w}\colon \mathbf{w} \geq 0,\ \mathbf{w}^\top \boldsymbol{\imath} = 1,\ \tau(\mathbf{w}) - \tau_0 = 0\right\},$$

where $\boldsymbol{\imath}$ is the I-dimensional column vector of ones and $\tau_0 = I^{-1}\sum_{i=1}^{I} Q_i(x)$. Dropping the requirements that the weights are nonnegative gives the larger set

$$\overline{\mathcal{W}} = \left\{\mathbf{w}\colon \mathbf{w}^\top \boldsymbol{\imath} = 1,\ \tau(\mathbf{w}) - \tau_0 = 0\right\}.$$

Two different asymptotically efficient WICQF estimators may be defined, depending on which set of consistent weights is considered, $\mathcal{W}$ or $\overline{\mathcal{W}}$.

Definition 2. *A WICQF estimator is asymptotically efficient if it is based on a vector of weights* $\mathbf{w}^*$ *that solves the problem*

$$\min_{\mathbf{w}\in\mathcal{W}} \mathrm{AV}(\hat{\tau}(\mathbf{w})). \tag{5.1}$$

A WICQF estimator is unconstrained asymptotically efficient if it is based on a vector of weights $\mathbf{w}^*$ *that solves the problem*

$$\min_{\mathbf{w}\in\overline{\mathcal{W}}} \mathrm{AV}(\hat{\tau}(\mathbf{w})). \tag{5.2}$$

The objective functions in (5.1) and (5.2) are convex in the vector $\mathbf{w}$, while the equality and inequality constraints that characterize $\mathcal{W}$ and $\overline{\mathcal{W}}$ are linear. Thus, the efficient estimators are solutions to standard convex optimization problems. We can therefore take advantage of the following convenient properties: (i) any local optimum is necessarily a global optimum, (ii) duality theory can be used to detect unfeasibility, hence algorithms are easy to initialize, and (iii) efficient numerical solution methods are available.

Notice that the solution to (5.1) need not be unique. Distance from the uniform weights $w_i = 1/I$ is a possible criterion for uniquely choosing among the optima. The solution to (5.2) is instead unique because of strict convexity of the asymptotic variance. In this case, however, $W(F)$ is no longer a probability distribution since the weight function is allowed to be a signed density. One advantage of allowing the weights to take negative values is a larger efficiency gain. This point will be discussed in more detail in the next section. The other advantage is computational. Because the optimal weights depend on both α and x, the burden associated with computing the

weights for each value of α and x may be reduced if a closed-form expression is available. As the next result shows, the minimization problem (5.2) allows us to do so.

Theorem 4. *Let the conditional distribution of Y_t given X_t be the same for all t and let $\hat{Q}(p\,|\,x)$ satisfy the conditions of Theorem 1. The weights $\overline{\mathbf{w}}^* \in \overline{\mathcal{W}}$ that define the unconstrained asymptotically efficient estimator satisfy the equation system*

$$\begin{pmatrix} \overline{\mathbf{w}}^* \\ \boldsymbol{\lambda} \end{pmatrix} = \mathbb{C}^{-1}\boldsymbol{\tau}, \tag{5.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$ is the vector of Lagrange multipliers associated with the two linear constraints that define $\overline{\mathcal{W}}$, $\boldsymbol{\tau} = (0, \ldots, 0, -1, -\tau_0)^\top$, and $\mathbb{C}$ is the $(I+2)$-dimensional square matrix

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{V} & -\mathbf{R} \\ -\mathbf{R}^\top & \mathbf{0} \end{pmatrix}, \tag{5.4}$$

with $\mathbf{R} = (\boldsymbol{\imath}, \mathbf{Q})$ and $\mathbf{Q} = (Q_1(x), \ldots, Q_I(x))^\top$.

By using the inversion formula for partitioned matrices, it is easy to show that

$$\overline{\mathbf{w}}^* = \boldsymbol{V}^{-1}\mathbf{R}(\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R})^{-1}\,(1, \tau_0)^\top\,.$$

Inserting this expression into the formula for the asymptotic variance gives

$$\mathrm{AV}(\hat{\tau}(\overline{\mathbf{w}}^*)) = \overline{\mathbf{w}}^{*\top}\boldsymbol{V}\overline{\mathbf{w}}^* = (1, \tau_0)(\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R})^{-1}\,(1, \tau_0)^\top\,.$$

Because the optimal weights depend on the distribution of the returns, in practice they must be estimated. A straightforward way of estimating the weights $\overline{\mathbf{w}}^*$ is to replace $\mathbf{R}$, $\boldsymbol{V}$ and τ_0 with consistent estimates.

Alternatively, consistent estimates of the optimal weights may be constructed by considering the empirical versions of problems (5.2) and (5.2). Given a decreasing sequence $\{r_T\}$ of positive real numbers tending to zero as $T \to \infty$, the empirical version of (5.2) consist of minimizing the estimated asymptotic variance $\widehat{\mathrm{AV}}(\hat{\tau}(\mathbf{w}))$ over the set of weights

$$\widehat{\mathcal{W}} = \left\{\mathbf{w}\colon \mathbf{w} \geq 0,\ \mathbf{w}^\top\boldsymbol{\imath} = 1,\ |\hat{\tau}(\mathbf{w}) - \hat{\tau}_0| \leq r_T\right\},$$

where $\hat{\tau}_0 = I^{-1}\sum_{i=1}^{I}\hat{Q}_i(x)$ is the unweighted ICQF estimator. The empirical version of (5.2) replaces $\widehat{\mathcal{W}}$ by

$$\widehat{\overline{\mathcal{W}}} = \left\{\mathbf{w}\colon\ \mathbf{w}^\top\boldsymbol{\imath} = 1,\ |\hat{\tau}(\mathbf{w}) - \hat{\tau}_0| \leq r_T\right\}.$$

Let $\widehat{\mathbf{w}}^*$ and $\widehat{\overline{\mathbf{w}}}^*$ be the solutions to the empirical versions of problems (5.2) and (5.2) respectively. The following result guarantees that these weights are consistent for the optimal weights $\mathbf{w}^*$ and $\overline{\mathbf{w}}^*$ under appropriate conditions on the rate of convergence of r_T to zero.

Theorem 5. *Suppose that the conditions of Theorem 1 and Corollary 1 both hold. Further assume that*

$$\lim_{T\to\infty} r_T = 0, \qquad \lim_{T\to\infty} r_T\sqrt{T} = \infty. \tag{5.5}$$

Then, as $T \to \infty$*,*

(i) If the vector $\mathbf{w}^*$ *is the unique minimizer of* $\mathrm{AV}(\hat{\tau}(\mathbf{w}))$ *in* $\overline{\mathcal{W}}$*, then*

$$\|\widehat{\mathbf{w}}^* - \mathbf{w}^*\| = \sup_{1\le i\le I} |\hat{w}_i^* - w_i^*| = o_P(1). \tag{5.6}$$

(ii) $\|\widehat{\overline{\mathbf{w}}}^* - \overline{\mathbf{w}}^*\| = o_P(1)$.

Corollary 2. *Under the conditions of Theorem 5,* $|\hat{\tau}(\widehat{\mathbf{w}}^*) - \tau_0| = o_P(1)$ *and* $\left|\hat{\tau}(\widehat{\overline{\mathbf{w}}}^*) - \tau_0\right| = o_P(1)$.

For $\hat{\tau}(\widehat{\mathbf{w}}^*)$, the proof of the corollary is an immediate consequence of the triangle inequality $|\hat{\tau}(\widehat{\mathbf{w}}^*) - \tau_0| \le |\hat{\tau}(\widehat{\mathbf{w}}^*) - \hat{\tau}(\mathbf{w}^*)| + |\hat{\tau}(\mathbf{w}^*) - \tau(\mathbf{w})| + |\tau(\mathbf{w}) - \tau_0|$. Similarly for $\hat{\tau}(\widehat{\overline{\mathbf{w}}}^*)$.

6. Asymptotic efficiency gains of WICQF estimators

This section studies the gains in asymptotic efficiency of WICQF estimators by providing analytical results for a number of distributions with characteristics that are typical of returns on financial assets, namely asymmetry, skewness or heavy tails (see e.g. McNeil and Frey [19]). Our results show that asymptotic efficiency gains are substantial, especially in the case of distributions with heavy tails.

We define the asymptotic efficiency gain of a WICQF estimator $\hat{\tau}(\mathbf{w}) = \sum_{i=1}^{I} w_i \hat{Q}_i(x)$, relative to the unweighted ICQF estimator $\hat{\tau}_0 = I^{-1}\sum_{i=1}^{I} \hat{Q}_i(x)$, as

$$\mathrm{eff}(\mathbf{w}) = 1 - \frac{\mathrm{AV}(\hat{\tau}(\mathbf{w}))}{\mathrm{AV}(\hat{\tau}_0))}.$$

Notice that $\mathrm{eff}(\mathbf{w})) = 1 - \mathrm{ARE}(\mathbf{w})^{-1}$, where $\mathrm{ARE}(\mathbf{w}) = \mathrm{AV}(\hat{\tau}_0)/\,\mathrm{AV}(\hat{\tau}(\mathbf{w}))$ is the asymptotic efficiency of $\hat{\tau}(\mathbf{w})$ relative to $\hat{\tau}_0$, namely the ratio of the sample sizes that are approximately needed for the two estimators to attain the same variance.

The distributions that we consider include mixtures of normals, Student's t, exponential, logistic, Gumbel and generalized Pareto (GP) distributions. Mixture of normals can approximate arbitrarily well any continuous distribution (see e.g. McLachlan and Peel [18]). We consider mixtures of a standard normal and a general $\mathcal{N}(\mu, \sigma^2)$ distribution with mixing coefficient π, where the parameters μ, σ and π are chosen to generate distributions with a substantial left tail. The

contaminating distribution has mean $\mu = x$, where x is one of the extreme left percentiles (the 1st, 2nd and 3rd) of the $\mathcal{N}(0,1)$ distribution, and constant standard deviation σ equal to 0.2 or 0.3. The mixing coefficient is set equal to .95. As for the t distributions, we limit ourselves to cases when the number of degrees of freedom is low (3 and 4).

The exponential, logistic and Gumbel distributions have quantile functions of the form

$$Q(p) = \mu + \sigma\zeta(p),$$

for some location parameter $\mu \in \mathbb{R}$ and some scale parameter $\sigma > 0$, where $\zeta(p)$ is a continuous and strictly increasing quantile function. In this case, it is easy to see that

$$\tau(\mathbf{w}) - \tau_0 = \sigma \sum_{i=1}^{I} \zeta(p_i) \left(w_i - \frac{1}{I} \right).$$

Moreover, for all $\mathbf{w}$, the formula for the asymptotic variance of $\hat{\tau}(\mathbf{w})$ satisfies $\mathrm{AV}_{(\mu,\sigma)}(\hat{\tau}(\mathbf{w})) = \sigma^2 \mathrm{AV}_{(0,1)}(\hat{\tau}(\mathbf{w}))$, where the subscript (μ, σ) refers to the location and scale parameters of the distribution. As a consequence, we have

$$\mathrm{eff}_{(\mu,\sigma)}(\hat{\tau}(\mathbf{w})) = \mathrm{eff}_{(0,1)}(\hat{\tau}(\mathbf{w})).$$

Thus, for these distributions, there is no loss of generality in confining attention to the standardized values of the location and scale parameters, that is, $\mu = 0$ and $\sigma = 1$. It is clear that, although the above identity holds for the asymptotic efficiency gain, the optimal weights $\mathbf{w}^*$ and the asymptotic distribution of $\hat{\tau}(\mathbf{w}^*)$ will generally depend on the location and scale parameters.

Finally, the GP distribution has distribution function $F(y) = [1 - \xi(y-\mu)/\sigma]^{-1/\xi}$, where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter, and $\xi \in \mathbb{R}$ is the shape parameter. The quantile function of the GP is equal to $Q(p) = \mu + \sigma(1 - p^{-\xi})/\xi$, so the asymptotic efficiency gain only depends on the shape parameter ξ. In this case, we set $\mu = 0$ and $\sigma = 1$, and let ξ take the values 0.1, 0.2 and 0.3.

As for the other parameters in our study, we set $\alpha = .10$ and $I = 25$. For the standard normal, the t, the exponential and the GP distributions we also carry out a sensitivity analysis by varying the number of grid-points I between 2 and 200.

We consider both unconstrained and constrained asymptotically efficient estimators. For the latter, we impose nonnegativity of the weights by using a penalty function containing a term that diverges to infinity as any of the weights becomes negative. Finally, we consider three increasingly restrictive specifications of the set of consistent weights, namely: (i) nonparametric without

nonnegativity constraints ($\overline{\mathcal{W}}$), (ii) nonparametric with nonnegativity constraints ($\mathcal{W}$), and (iii) parameterized as a mixture of beta distributions restricted to the interval $[0, \alpha]$. The last case corresponds to a weight function of the form $w(p) = \pi w_1(p) + (1 - \pi)w_2(p)$, with

$$w_j(p) = \frac{1}{\alpha\, b(c_j, d_j)} \left(\frac{p}{\alpha}\right)^{c_j - 1} \left(1 - \frac{p}{\alpha}\right)^{d_j - 1}, \qquad j = 1, 2,$$

where $0 \leq p \leq \alpha$, $\pi \in [0, 1/2]$, $b(c_j, d_j) = \Gamma(c_j)\Gamma(d_j)/\Gamma(c_j + d_j)$, $(c_j, d_j) \in \Psi$, and Ψ a compact subset of $\mathbb{R}^2$ that contains the point $(1, 1)$ in its interior. The relationship between the three sets of weights implies that the asymptotic efficiency gain for the nonparametric case without nonnegativity constraints (the "unconstrained case") is greater or equal to that for the nonparametric case with negativity constraints (the "constrained case"), which is itself greater or equal to that for the parametric case.

Table 1 shows our results. The table consists of two panels. The first is for normal mixtures and t distributions. The second is for distributions with invariance of the asymptotic efficiency gain to location and scale, namely the normal, the logistic, the exponential, the Gumbel and the GP distributions. For each distribution, moving from the left to the right column, we increase the probability mass in the tail. This is achieved by decreasing the mean of the contaminating distribution for the normal mixture, by decreasing the number of degrees of freedom for the t distribution, and by increasing ξ, the shape parameter for the GP distribution.

For the normal mixtures, the asymptotic efficiency gain is higher the more extreme is the contamination of the original distribution (the lower σ and the more negative x). The gain ranges from 4% for $\sigma = 0.3$ and $x = -1.881$, to values around 30% for the parametric specification, to more than 50% for the nonparametric specifications with values corresponding to $\sigma = 0.3$ and $x = -2.326$. For t distributions, the asymptotic efficiency gains are 2% for the t with 4 degrees of freedom and more than 5% for the t with 3 degrees of freedom. For the normal distribution, the efficiency gain is very small (around 1%) for all specifications. It is also very small (gain less than 3%) for the logistic, the exponential and the Gumbel. For the GP distribution, the efficiency gain increases as ξ increases, reaching 6% when $\xi = 0.3$.

Figure 1 plots the asymptotic variance of both the unweighted ICQF and the unconstrained asymptotically efficient WICQF estimator for different values of the parameter I and different choices of the conditional distribution of Y_t. For both types of estimators, the asymptotic variance increases rapidly up to around $I = 50$ and then tends to flatten out. The difference in asymptotic variance between fat tailed and non-fat tailed distributions is remarkable, and so is the difference in asymptotic variance between the unweighted and the weighted estimators but only for fat tailed

distributions.

Figure 2 plots the asymptotic efficiency gains of the optimal WICQF relative to the unweighted ICQF. In the case of the standard normal and the exponential distributions, eff is very small and appears to be single-peaked in I, with a maximum around $I = 10$. For the t and GP distributions, instead, eff is large and appears to be nondrecreasing and concave in I These results suggest that asymptotic efficiency gains are especially influenced by heaviness of the tail of the parent distribution in a neighborhood of the quantile of interest. This may explain why, in the mixture of normals example, we have a combination of low asymptotic variance and high asymptotic efficiency gains, while for the logistic and the exponential it is the other way round. Instead, for the t distribution, we have both high asymptotic variance and high asymptotic efficiency gains.

Figure 3 shows the vectors of optimal weights $\mathbf{w}^*$ and $\overline{\mathbf{w}}^*$. Two factors govern the behavior of these weights. One is the precision of the quantile estimator, which increases as we approach α. The second is the consistency constraint that $\tau(\mathbf{w}) = \tau_0$. Because of this constraint, overweighting of the quantiles close to α must be compensated somewhere else. Where the compensation occurs, depends on the conditional distribution of Y_t. In the unconstrained case, for the fat tailed distributions, the two factors entail overweighting of the bottom quantiles.

Even if not illustrated here, the behavior of the optimal weights is similar for the exponential and the other non-fat tailed distributions in our study (the standard normal and the Gumbel), and is also similar for the fat-tailed t and GP distributions.

7. Empirical application

This section considers an application to daily data on the returns on 6 European stock indexes. For each stock index, we compare the results obtained for the unweighted ICQF estimator $\hat{\tau}_0$ and the optimal (unconstrained asymptotically efficient) WICQF estimator $\hat{\tau}^* = \hat{\tau}(\widehat{\overline{\mathbf{w}}}^*)$, both at level $\alpha = .10$.

Raw daily data range from December 30, 1994, to December 31, 2007. The outcome variable Y_t is the daily return on a stock index, defined as the logarithmic daily difference in the stock index. The stock indexes considered are: Xetra Dax 30 (Frankfurt), CAC 40 (Paris), S&P MIB 30 (Milan), IBEX 35 (Madrid) and AEX (Amsterdam). The vector of covariates X_t includes the dividend yield on European equities, the rates of change of future prices of oil and non-energy commodity prices, the rate of change of the euro-dollar exchange rate, and measures of risk spread (the difference between the 10-year German government bond yield and the short-term interest

rate) and term spread (the difference between the 10- and the 5-year German government bond yield). Table 2 provides details on data transformation.

Covariates enter the linear models for the regression quantiles with a one-period lag. Conditional quantiles are estimated using rolling samples of size $T_0 = 500$. For each $t = T_0, \ldots, T$, the estimated conditional quantiles evaluated at the current value of the covariates are used to predict the conditional expected shortfall at time $t + 1$.

In order to construct the optimal weights for the WICQF estimator, we need an estimate of the asymptotic variance for each given set of weights. To simplify the calculations, we make the crude assumption that the conditional distribution of Y_t given X_t is GP with parameters $(0, 1, \xi)$, where the parameter ξ is estimated by maximum likelihood. Given the estimate of the asymptotic variance, the optimal weights are then derived, for $I = 20$ and each rolling window sample of size $T = 500$, by solving the empirical version of equation system (5.3).

Following McNeil and Frey [19], we compare the predictive ability of the two estimators by looking at the distribution of their out-of-sample forecast error for all quantile violation events, that is, cases when thes return is lower than the correspondent predicted α-level quantile (the VaR of level $1 - \alpha$). The forecast error is defined as the difference between the observed return and the predicted expected shortfall, conditional on the covariates and the occurrence of a quantile violation event.

Tables 3 and 4 report summary statistics of the empirical distribution of the two estimators and the associated forecast error over 2,147 rolling windows. We show the mean, the standard deviation and the difference between the 99th and the 1st percentiles, all expressed in percentage points. Quantile violation events occur in approximatively 12% of cases. Summary statistics are similar for the two estimators and, in fact, the optimal weights are close to uniform. The standard deviation of the empirical distribution of the two estimators varies in a tight range between 0.9% and 1.1%. The quantile difference is always above 4%, but is smaller for the optimal WICQF estimator. As for the forecast error (Table 4), its mean value is always negative for both estimators. Moreover, the mean forecast error tends to be somewhat smaller (in absolute value) for the optimal WICQF than for the unweighted ICQF estimator. Overall, the optimal WICQF estimator appears to behave slightly better, in terms of forecast error, than the unweighted ICQF estimator.

8. Conclusions

In this paper we generalize a class of estimators of the α-level conditional expected shortfall obtained by integrating the estimated conditional quantile regression function over a possibly

data-dependent interval. Our general class of estimators assigns different weights to the different quantiles, thus attaining higher asymptotic efficiency relative to the case when no weighting is used. We provide asymptotic results that open the way to inference. These results rely on the critical assumption that the extreme regression quantiles can be ignored. Additional research is needed to relax this strong assumption.

Our study of the asymptotic efficiency gains associated with the proposed estimators suggests that these gains are substantial, especially in the case of distributions with heavy tails. Finally, in our empirical application to daily financial data, the optimal WICQF estimator appears to behave slightly better, in terms of out-of-sample forecast error, than the unweighted ICQF estimator.

References

[1] Acerbi, C., Tasche, D., 2002. On the coherence of expected shortfall. *Journal of Banking and Finance* 26, 1487–1503.

[2] Acerbi, C., 2002. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance* 26, 1505–1518.

[3] Angrist, J., Chernozhukov, V., Fernández-Val, I., 2006. Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* 74, 539–563.

[4] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. *Mathematical Finance* 9, 203–228.

[5] Bassett, G.W., Koenker, R., Kordas, G., 2004. Pessimistic portfolio allocation and Choquet expected utility. *Journal of Financial Econometrics* 2, 477–492.

[6] Bertsimas, D., Lauprete, G.J., Samarov, A., 2004. Shortfall as a risk measure: Properties, optimization and applications. *Journal of Economic Dynamics & Control* 28, 1353–1381.

[7] Cai Z., Wang X., 2008. Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics* 2008, 120–130.

[8] Chernozhukov, V., 2005. Extremal quantile regression. *Annals of Statistics*, 33, 806–839.

[9] Chernozhukov, V., Fernadez-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. *Econometrica*, 1093–1125.

[10] Csörgö S., Haeusler E., Mason, D.M., 1991. The asymptotic distribution of extreme sums. *Annals of Probability* 19, 783–811.

[11] Delbaen, F., 2000. Coherent risk measures on general probability spaces. ETH Zürich, mimeo.

[12] Fan, J., Gijbels, I., 1996. *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

[13] Fitzenberger, B., 1998. The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics* 82, 235–287.

[14] Hall, P., Wolff, R.C.L., Yao, Q., 1999. Methods for estimating a conditional distribution function, *Journal of the American Statistical Association* 94, 154–163.

[15] Koenker, R., Bassett G., 1978. Regression quantiles. *Econometrica* 46, 33–50.

[16] Koenker, R., 2005. *Quantile Regression*. Cambridge University Press: New York.

[17] Mason, D.M., Shorack, G.R., 1992. Necessary and sufficient conditions for asymptotic normality of L-statistics. *Annals of Probability* 20, 1779–1804.

[18] McLachlan, G., Peel, D., 2000. *Finite Mixture Models*. Wiley: New York.

[19] McNeil, A.J., Frey, R., 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance* 7, 271–300.

[20] Peracchi, F., Tanase, A.V., 2008. On estimating the conditional expected shortfall. *Applied Stochastic Models in Business and Industry* 24, 471–493.

[21] Prelec, D., 1998. The probability weighting function. *Econometrica* 66, 497–527.

[22] van der Vaart, A.W., Wellner, J.A., 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer: Berlin.

[23] Yaari, M.E., 1987. The dual theory of choice under risk. *Econometrica* 55, 95–115.

Table 1: Asymptotic variance AV of the asymptotically efficient WICQF estimator and asymptotic efficiency gain eff relative to the unweighted ICQF estimator. For each distributional assumption, results are grouped according to three different specifications of the set of consistent weights: (i) nonparametric without nonnegativity constraints, (ii) nonparametric with nonnegativity constraint, and (iii) parameterized as a mixture of beta distributions. The level α is set to 10%.

		$w(\cdot)$	AV	eff(%)	AV	eff(%)	AV	eff(%)
					$\pi\mathcal{N}(0,1)+(1-\pi)\mathcal{N}(x,\sigma^2)$			
π	σ		$x=-1.881$		$x=-2.054$		$x=-2.326$	
.95	.2	(i)	2.147	25.7	2.570	30.1	3.995	57.3
		(ii)		20.7		20.1		28.7
		(iii)		16.7		15.7		16.7
.95	.3	(i)	2.583	4.3	2.979	5.9	4.313	25.8
		(ii)		4.2		4.9		12.6
		(iii)		3.6		3.2		10.0
					$t[r]$			
			$r=4$		$r=3$		$r=2$	
		(i)	17.319	2.7	31.497	6.6	113.249	17.5
		(ii)		2.5		5.7		13.3
		(iii)		2.4		5.3		11.9
					$\mathrm{GP}(0,1,\xi)$			
			$\xi=.1$		$\xi=.2$		$\xi=.3$	
		(i)	35.943	.3	75.817	2.2	164.282	5.8
		(ii)		.3		1.9		4.9
		(iii)		.3		1.8		4.5
			$\mathcal{N}(0,1)$		Logistic$(0,1)$		Gumbel$(0,1)$	
		(i)	3.601	1.4	18.997	.1	1.608	2.9
		(ii)		1.4		.1		2.9
		(iiii)		1.4		.1		2.9
					Exponential(1)			
		(i)	17.474	.2				
		(ii)		.2				
		(iii)		.2				

Table 2: Transformations and summary statistics of the composite indexes and the covariates. The transformed data range from January 3, 1995 to December 27, 2007, and values are expressed in percentage points. The total number of observations is 2,646.

Variable	Description	Mean	SD	$Q_{99} - Q_{01}$
DAX	Xetra Dax 30 return	$2.6 \cdot 10^{-2}$	1.4	7.8
CAC40	CAC 40 return	$3.5 \cdot 10^{-2}$	1.3	7.4
MIB30	S&P MIB 30 return	$3.5 \cdot 10^{-2}$	1.3	6.8
IBEX35	IBEX 35return	$5.8 \cdot 10^{-2}$	1.3	6.9
AEX	AEX return	$1.9 \cdot 10^{-2}$	1.3	7.6
STOXX50	DJ Euro Stoxx 50 return	$3.0 \cdot 10^{-2}$	1.3	7.4
ECOMM	Commodity price log diff	$2.3 \cdot 10^{-2}$	0.7	3.4
EDY	Dividend yield	79	27	101
EFX	EUR/USD log diff	$-0.5 \cdot 10^{-2}$	0.6	3.1
EOIL	Oil price log diff	$7.5 \cdot 10^{-2}$	2.2	11.0
ERSP	Risk spread	115	45	191
ESP	Term Spread	146	90	375

Table 3: Summary statistics of the empirical distribution over 2,147 rolling windows of the one-step ahead predicted shortfall (expressed in percentage points) for the unweighted ICQF estimator $\hat{\tau}_0$ and the unconstrained asymptotically efficient WCQF estimator $\hat{\tau}^* = \hat{\tau}(\widehat{\mathbf{w}}^*)$. The level α is equal to .10.

Estimator	Mean	SD	$Q_{.99}$-$Q_{.01}$	Mean	SD	$Q_{.99}$-$Q_{.01}$
	Xetra Dax 30			CAC 40		
$\hat{\tau}_0$	-2.365	1.079	4.864	-2.157	0.896	4.178
$\hat{\tau}^*$	-2.359	1.072	4.833	-2.164	0.899	4.189
	S&P MIB 30			IBEX 35		
$\hat{\tau}_0$	-2.090	1.081	4.653	-2.078	1.019	4.613
$\hat{\tau}^*$	-2.090	1.082	4.666	-2.082	1.021	4.596
	AEX			DJ Euro Stoxx		
$\hat{\tau}_0$	-2.135	1.001	5.057	-2.125	0.934	4.468
$\hat{\tau}^*$	-2.136	1.000	5.043	-2.128	0.936	4.474

Table 4: Summary statistics of the empirical distribution over 2,147 rolling windows of the one-step ahead forecast error (expressed in percentage points) in quantile violation cases for the unweighted ICQF estimator $\hat{\tau}_0$ and the unconstrained asymptotically efficient WCQF estimator $\hat{\tau}^* = \hat{\tau}(\widehat{\mathbf{w}}^*)$. The level α is equal to .10.

Estimator	Mean	SD	$Q_{.99}$-$Q_{.01}$	Mean	SD	$Q_{.99}$-$Q_{.01}$
	Xetra Dax 30			CAC 40		
$\hat{\tau}_0$	-0.067	0.753	3.460	-0.052	0.721	3.409
$\hat{\tau}^*$	-0.072	0.754	3.479	-0.045	0.722	3.444
	S&P MIB 30			IBEX 35		
$\hat{\tau}_0$	-0.072	0.706	4.147	-0.058	0.783	4.330
$\hat{\tau}^*$	-0.071	0.706	4.219	-0.052	0.783	4.323
	AEX			DJ Euro Stoxx		
$\hat{\tau}_0$	-0.166	0.769	3.898	-0.099	0.692	3.387
$\hat{\tau}^*$	-0.163	0.769	3.956	-0.093	0.695	3.416

Figure 1: Asymptotic variance of the unweighted ICQF and the unconstrained asymptotically efficient WICQF estimator for $\alpha = .10$, different values of I and different choices of the conditional distribution of Y_t given X_t.

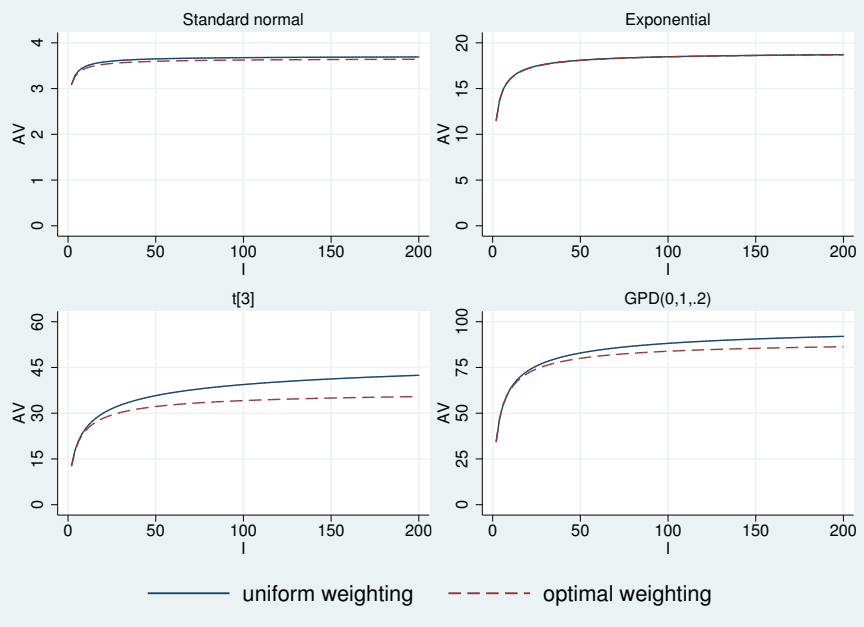


Figure 2: Efficiency gain (in percentage points) of the unconstrained asymptotically efficient WICQF estimator relative to the unweighted ICQF estimator for $\alpha = .10$, different values of I and different choices of the conditional distribution of Y_t given X_t.

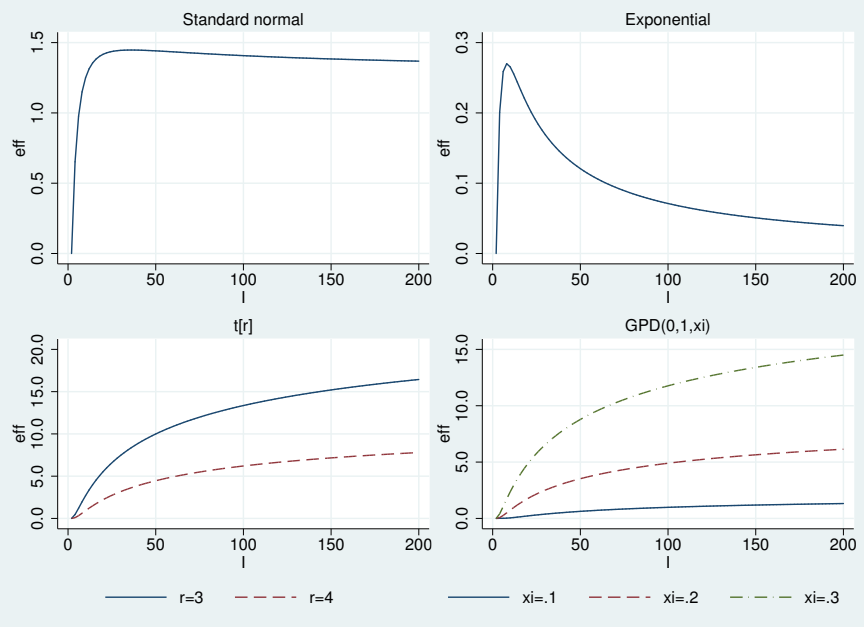
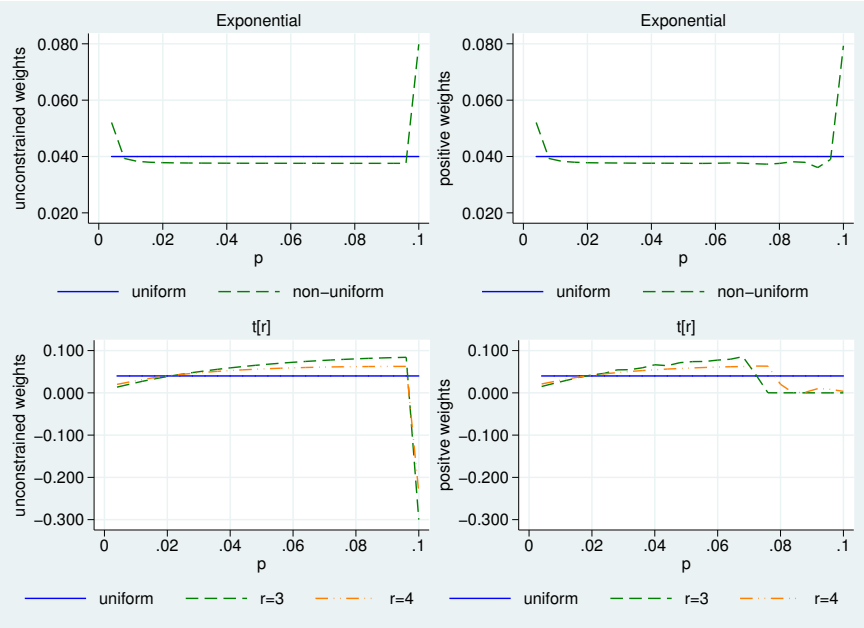


Figure 3: The optimal constrained $\mathbf{w}^*$ and unconstrained weights $\overline{\mathbf{w}}^*$ of the WICQF estimator for $\alpha = .10$, $I = 25$ and different choices of the conditional distribution of Y_t given X_t.



Appendix A.

Proof of Theorem 2

Under the assumptions (i)–(iv), we have consistency and asymptotic normality of the quantile regression estimator (see Koenker [16]) More precisely,

$$\left(\mathbf{J}(p)\sqrt{T}[\hat{\beta}(p)-\beta(p)],\mathbf{J}(s)\sqrt{T}[\hat{\beta}(s)-\beta(s)]\right)\xrightarrow{d}(Z(p),Z(s)) \tag{A.1}$$

where $\mathbf{J}(p)=\mathbb{E}\left[f(\beta(p)^\top X_t|X_t)X_tX_t^\top\right]$ is positive definite for all $p\in(0,1)$ and $Z(p)$ is a zero mean Gaussian random variable with $\mathrm{cov}(Z(p_i),Z(p_j))=\Sigma_{ij}$.

Now let $\tilde{\beta}(\alpha)=\sum_{i=1}^I w_i\hat{\beta}(p_i)$ be a linear combination of the I estimates $\hat{\beta}(p_1)\dots\hat{\beta}(p_I)$ of the true population quantile coefficients $\beta(p_1)\dots\beta(p_I)$. The weights vector $\mathbf{w}=(w_1,\dots w_I)^\top$ is deterministic. Therefore, as matrix $\mathbf{J}(p)$ is positive definite for any $p\in(0,1)$, from (A.1) we have that the vector $\{\sqrt{T}(\hat{Q}(p_i)-\bar{Q}(p_i)),\ i=1,\dots,I\}$ is asymptotically normal with covariance function given by $V(p_i,p_j)=x^\top\mathbf{J}_i^{-1}\Sigma_{i,j}\mathbf{J}_j^{-1}x$ and the result immediately follows.

Proof of Theorem 3

Using (4.1), we have that

$$\begin{aligned}\sqrt{T}\left[\sum_{i=1}^I w_i(\hat{Q}_i(x)-Q_i(x))\right]&=\sqrt{T}\left[\int_{\alpha_0}^{\alpha}w(p)\hat{Q}(p\,|\,x)\mathrm{d}p-\int_{\alpha_0}^{\alpha}w(p)Q(p\,|\,x)\mathrm{d}p\right]\\&+\sqrt{T}\sum_{i=1}^I\left(\int_{p_{i-1}}^{p_i}Q(p\,|\,x)w(p)dp-\frac{Q_i(x)+Q_{i-1}(x)}{2}w_i\right)\end{aligned}$$

Then, in view of the weak convergence assumptions, it is enough to prove that the second term goes to zero.

We first consider the unweighted version: $w_i=(p_i-p_{i-1})/(\alpha-\alpha_0)=1/I$. In this case, the discrete approximation $\sum_{i=1}^I(Q_i(x)+Q_{i-1}(x))/2I$ is obtained by integrating the spline function:

$$Q(p\,|\,x)=Q_{i-1}(x)+I(p-p_{i-1})(Q_i(x)-Q_{i-1}(x)),$$

for $p\in(p_{i-1},p_i]$.

Then, we perform a Taylor expansion for $Q(p\,|\,x)$ around p_{i-1} up to the second term:

$$\begin{aligned}&\int_{p_{i-1}}^{p_i}[Q(p\,|\,x)-Q_{i-1}(x)+I(p-p_{i-1})(Q_i(x)-Q_{i-1}(x))]dp\\&\quad=\int_{p_{i-1}}^{p_i}\left[q(p_{i-1}\,|\,x)(p-p_{i-1})+O((p-p_{i-1})^2)-I(p-p_{i-1})(Q_i(x)-Q_{i-1}(x))\right]dp\\&\quad=\left(q(p_{i-1}\,|\,x)-\frac{(Q_i(x)-Q_{i-1}(x))}{1/I}\right)\frac{(p_i-p_{i-1})^2}{2}+O(I^{-3})=O(I^{-3}).\end{aligned}$$

Then, under the assumption $|w(p)| < \infty$ for all $(0, \alpha]$,

$$\begin{aligned}
&\sqrt{T}\left|\sum_{i=1}^{I}\left(\int_{p_{i-1}}^{p_i} Q(p\,|\,x)w(p)dp - \frac{Q_i(x)+Q_{i-1}(x)}{2}w_i\right)\right| \\
&\qquad = \sqrt{T}\left|\sum_{i=1}^{I}\left(\int_{p_{i-1}}^{p_i} Q(p\,|\,x)w(p)dp - \frac{Q_i(x)+Q_{i-1}(x)}{2}\int_{p_{i-1}}^{p_i} w(p)dp\right)\right| \\
&\qquad \leq \sqrt{T}\sup_{p\in(\alpha_0,\alpha]}|w(p)|\sum_{i=1}^{I}\left|\int_{p_{i-1}}^{p_i} Q(p\,|\,x)dp - \frac{Q_i(x)+Q_{i-1}(x)}{2I}\right| \\
&\qquad \leq \sqrt{T}\sup_{p\in(\alpha_0,\alpha]}|w(p)|O(I^{-2}) \to 0.
\end{aligned}$$

Proof of Theorem 4

The Lagrangian function is

$$\mathcal{L}_u^Q = \mathbf{w}^\top \boldsymbol{V}\mathbf{w} + \lambda_1(1 - \mathbf{w}^\top \boldsymbol{\imath}) + \lambda_2(\tau - \mathbf{w}^\top \mathbf{Q}). \tag{A.2}$$

One obtains the result (5.3), by solving the system of linear equations:

$$\begin{array}{l} \frac{\partial \mathcal{L}_u}{\partial \mathbf{w}} = 0 \\ \frac{\partial \mathcal{L}_u}{\partial \lambda_i} = 0 \quad i = 1, 2. \end{array}$$

The fact that $\boldsymbol{V}$ is positive semidefinite ensures that $\tau(\overline{\mathbf{w}}^*)$ is the minimum over $\mathcal{W}_I$.

Proof of Theorem 5

(a.i) Let $\mathbb{M} : \mathcal{W} \mapsto \mathbb{R}^+$ equal to $\mathbb{M} = AV\left(\hat{\tau}(\mathbf{w})\right)$ and $\mathbb{M}_T = \widehat{AV}\left(\hat{\tau}(\mathbf{w})\right)$.

Let also $\tilde{\mathbf{w}}^* = \arg\min_{\hat{\mathcal{W}}} \mathbb{M}$. We immediately have that

$$\|\hat{\mathbf{w}}^* - \mathbf{w}^*\| \leq \|\tilde{\mathbf{w}}^* - \hat{\mathbf{w}}^*\| + \|\tilde{\mathbf{w}}^* - \mathbf{w}^*\|. \tag{A.3}$$

The limit (5.6) then follows if both the terms in the right hand side of (A.3) tend to zero in probability.

For the first term, we consider the fact that, by definition, $\mathbb{M}(\tilde{\mathbf{w}}^*) \geq \mathbb{M}(\mathbf{w})$ and $\mathbb{M}_T(\hat{\mathbf{w}}^*) \geq \mathbb{M}_T(\mathbf{w})$ for all $\mathbf{w}$ in $\widehat{\mathcal{W}}$. Moreover, from the consistency of the estimator $\widehat{\mathrm{AV}}(\hat{\tau}(\mathbf{w}))$ for all $\mathbf{w}$ (Corollary 1), it follows that $|\mathbb{M}_T(\mathbf{w}) - \mathbb{M}(\mathbf{w})| \to 0$ in probability. We need to prove that the above convergence is true uniformly in $\mathcal{W} \cup \widehat{\mathcal{W}}$. In fact, let

$$\mathcal{W}_1 = \left\{\mathbf{w} :\ \mathbf{w} \geq 0,\ \mathbf{w}^\top \boldsymbol{\imath} = 1\right\},$$

so that $\mathcal{W}_1 \supseteq \mathcal{W} \cup \widehat{\mathcal{W}}$. The family $\mathcal{W}_1$ is univocally determined by the family of monotone non-decreasing step-functions:

$$\mathcal{W}_1 \Leftrightarrow \left\{ W : W(x) = \sum_{i \leq j} w_i,\ p_j \leq x < p_{j+1},\ W(x) \geq W(y),\ \alpha \geq x > y \geq 0 \right\}. \qquad \text{(A.4)}$$

¿From Theorem 2.7.5 in [22] we have that $\log N_{[]}(\varepsilon, \mathcal{W}_1, L_2(R)) \leq K/\varepsilon$, for every probability measure R and for some constant K independent on ε. This implies that for every $\varepsilon > 0$ we can find $m(\varepsilon) < e^{K'/\varepsilon}$ couples of bounded step functions $\{(f_j^L, f_j^U), j = 1, \ldots, m(\varepsilon)\}$ such that $f_L \leq W \leq f_j^U$ and $\|f_j^L - f_j^U\|_2 < \varepsilon$. As a consequence, setting $\mathbf{f}_j^L = (f_j^L(p_1), \ldots, f_j^L(p_I))$

$$\begin{aligned} \mathbb{M}_T(\mathbf{w}) - \mathbb{M}(\mathbf{w}) &\leq \mathbb{M}_T(\mathbf{f}_j^L) - \mathbb{M}(\mathbf{f}_j^L) + \mathbb{M}(\mathbf{f}_j^L) - \mathbb{M}(\mathbf{w}) \\ &\leq \mathbb{M}_T(\mathbf{f}_j^L) - \mathbb{M}(\mathbf{f}_j^L) - C\varepsilon^2 \end{aligned}$$

where the constant C depends on the elements of $\boldsymbol{V}$ only. Letting $\varepsilon \to 0$, and repeating the argument changing the sign and using f_j^U one obtains:

$$\sup_{\mathbf{w} \in \overline{\mathcal{W}}} |\mathbb{M}_T(\mathbf{w}) - \mathbb{M}(\mathbf{w})| = o_P(1).$$

Now that we have proved $\|\mathbb{M}_T - \mathbb{M}\| = o_P(1)$, a slight modification of Corollary 3.2.3 of [22] yields $\|\tilde{\mathbf{w}}^* - \hat{\mathbf{w}}^*\| = o_P(1)$.

It now remains to prove that also $\|\tilde{\mathbf{w}}^* - \mathbf{w}^*\| = o_P(1)$. It is easy to show that, from $\sqrt{T}-$consistency of $\hat{\tau}(\mathbf{w})$ and $\hat{\tau}_0$,

$$\widehat{\mathcal{W}} \approx \left\{ \mathbf{w} : \mathbf{w}^\top \boldsymbol{\imath} = 1, \mathbf{w} \geq 0,\ |\tau(\mathbf{w}) - \tau_0| < r_T + O(c \cdot T^{-1/2}) \right\} \qquad \text{(A.5)}$$

with $c \neq 0$. The right-had-side of (A.5) describes a monotonically decreasing sequence of subsets of W_1, whose limit in probability is, because of (5.5), equal to $\mathcal{W}$. Then,

$$\mathbb{M}(\tilde{\mathbf{w}}^*) = \inf_{\mathbf{w} \in \widehat{\mathcal{W}}} \mathbb{M}(\mathbf{w}) \leq \inf_{\mathbf{w} \in \mathcal{W}} \mathbb{M}(\mathbf{w}) + o_P(1) \approx \mathbb{M}(\mathbf{w}^*)$$

and by continuity of $\mathbb{M}$ with respect to $\mathbf{w}$,

$$\mathbb{M}(\lim_{T \to \infty} \tilde{\mathbf{w}}^*) = \lim_{T \to \infty} \mathbb{M}(\tilde{\mathbf{w}}^*) = \lim_{T \to \infty} \inf_{\mathbf{w} \in \widehat{W}} \mathbb{M}(\mathbf{w}) = \inf_{\mathcal{W}} \mathbb{M}(\mathbf{w}) = \mathbb{M}(\mathbf{w}^*),$$

that, together with the uniqueness of $\mathbf{w}^*$, implies $\|\tilde{\mathbf{w}}^* - \mathbf{w}^*\| = o_P(1)$.

(ii) Let Δ be a countable dense subset of the interval $[-r_T, r_T]$. Then, for very $\delta \in \Delta$, let

$$\widehat{\overline{\mathcal{W}}}^{\delta} = \left\{ \mathbf{w} : \mathbf{w}^{\top} \boldsymbol{\imath} = 1, \hat{\tau}(\mathbf{w}) = \hat{\tau}_0 + \delta \right\}.$$

Then we have $\widehat{\overline{\mathcal{W}}} \supseteq \cup_{\delta \in \Delta} \widehat{\overline{\mathcal{W}}}^{\delta}$. We can limit ourselves at considering the sets indexed by Δ. For every fixed $\delta \in \Delta$, it follows from Theorem 4 that

$$\begin{pmatrix} \hat{\overline{\mathbf{w}}}^{*}_{\delta} \\ \boldsymbol{\lambda}_{\delta} \end{pmatrix} = \arg\min_{\widehat{\overline{\mathcal{W}}}^{\delta}} \mathbb{M}_T = \hat{\mathbb{C}}_T^{-1} \hat{\boldsymbol{\tau}}_{\delta}$$

where $\hat{\mathbb{C}} = \begin{pmatrix} \hat{\boldsymbol{V}} & -\hat{\mathbf{R}} \\ -\hat{\mathbf{R}}^{\top} & \mathbf{0} \end{pmatrix}$, $\hat{\mathbf{R}} = \left(\imath, \hat{\mathbf{Q}} \right)$, $\hat{\mathbf{Q}}$ the $I-$dimensional vector of estimates for $\mathbf{Q}$, and where $\hat{\boldsymbol{\tau}}_{\delta}^{\top} = (0, \ldots, 0, -1, -(\hat{\tau}_0 - \delta))$. ¿From the block matrix inversion formula,

$$\overline{\mathbf{w}}^{*} = \boldsymbol{V}^{-1}\mathbf{R}(\mathbf{R}^{\top}\boldsymbol{V}^{-1}\mathbf{R})^{-1}\bar{\boldsymbol{\tau}} \quad \text{and} \quad \hat{\overline{\mathbf{w}}}^{*}_{\delta} = \hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1}\bar{\boldsymbol{\tau}}_{\delta},$$

where $\bar{\boldsymbol{\tau}}_{\delta}^{\top} = (-1, -(\hat{\tau}_0 + \delta))$ and $\bar{\boldsymbol{\tau}}^{\top} = (-1, -\tau_0)$.

Then, for every $\delta \in \Delta$,

$$\begin{aligned} \|\hat{\overline{\mathbf{w}}}^{*}_{\delta} - \overline{\mathbf{w}}^{*}\| &= \left\| \boldsymbol{V}^{-1}\mathbf{R}(\mathbf{R}^{\top}\boldsymbol{V}^{-1}\mathbf{R})^{-1}\bar{\boldsymbol{\tau}} - \hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1}\bar{\boldsymbol{\tau}}_{\delta} \right\| \\ &\leq \left\| \hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} \left(\bar{\boldsymbol{\tau}}_{\delta} - \bar{\boldsymbol{\tau}} \right) \right\| \\ &\quad + \left\| \left(\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - \boldsymbol{V}^{-1}\mathbf{R}(\mathbf{R}^{\top}\boldsymbol{V}^{-1}\mathbf{R})^{-1} \right) \bar{\boldsymbol{\tau}} \right\|. \end{aligned} \tag{A.6}$$

The first term satisfies

$$\begin{aligned} \left\| \hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} \left(\bar{\boldsymbol{\tau}}_{\delta} - \bar{\boldsymbol{\tau}} \right) \right\| &\leq \left\| \hat{\boldsymbol{V}} \right\|^{-1/2} \left\| \hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}} \right\|^{-1/2} O_P\left(\delta + T^{-1/2} \right) \\ &= O_P\left(\delta + T^{-1/2} \right) = o_P(1). \end{aligned}$$

The quantity $\|\boldsymbol{V}\|$ is the algebraic norm of a $m \times n-$dimensional matrix $\boldsymbol{V}$, defined as

$$\|\boldsymbol{V}\| = \max \left\{ \frac{\|\boldsymbol{V}x\|_2}{\|x\|_2}, x \in \mathbb{R}^n, x \neq 0 \right\}$$

and, if $\boldsymbol{V}$ is a square positive semidefinite matrix, $\|\boldsymbol{V}\| = \lambda_{\max}(\boldsymbol{V})$, the maximum eigenvalue of $\boldsymbol{V}$.

For the second term of (A.6), we have

$$\begin{aligned} &\left\| \left(\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}_{Q}^{-1}\hat{\mathbf{R}})^{-1} \pm \boldsymbol{V}^{-1}\mathbf{R}(\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - \boldsymbol{V}^{-1}\mathbf{R}(\mathbf{R}^{\top}\boldsymbol{V}^{-1}\mathbf{R})^{-1} \right) \bar{\boldsymbol{\tau}} \right\| \\ &\leq \left\| (\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}) \right\|^{-1} \left\| \hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}} - \boldsymbol{V}^{-1}\mathbf{R} \right\| \|\bar{\boldsymbol{\tau}}\| + \left\| \boldsymbol{V}^{-1}\mathbf{R} \right\| \left\| (\hat{\mathbf{R}}^{\top}\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - (\mathbf{R}^{\top}\boldsymbol{V}^{-1}\mathbf{R})^{-1} \right\| \|\bar{\boldsymbol{\tau}}\|. \end{aligned}$$

Then, we need to show that the right hand side of the above inequality is $o_P(1)$.

It is enough to show that $\left\|\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}} - \boldsymbol{V}^{-1}\mathbf{R}\right\| = o_P(1)$ and also $\left\|(\hat{\mathbf{R}}^\top\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - (\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R})^{-1}\right\| = o_P(1)$, because the other terms, for fixed α and I, are $O_P(1)$.

First of all,

$$\left\|\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}} - \boldsymbol{V}^{-1}\mathbf{R} \pm \hat{\boldsymbol{V}}^{-1}\mathbf{R}\right\| \leq \left\|\hat{\boldsymbol{V}}^{-1}\right\| \left\|\hat{\mathbf{R}} - \mathbf{R}\right\| + \|\mathbf{R}\| \left\|\hat{\boldsymbol{V}}^{-1} - \boldsymbol{V}^{-1}\right\|.$$

It is easy to see that, under the assumptions, $\left\|\hat{\mathbf{R}} - \mathbf{R}\right\| = O_P(T^{-1/2})$ because $\hat{Q}_i(x)$ is consistent for $Q_i(x)$, $i = 1, \ldots, I$. Moreover,

$$\left\|\hat{\boldsymbol{V}}^{-1} - \boldsymbol{V}^{-1}\right\| \leq \left\|\boldsymbol{V}^{-1}\right\| \left\|I - \boldsymbol{V}^{-1/2}\hat{\boldsymbol{V}}^{-1}\boldsymbol{V}^{-1/2}\right\| = O_P(T^{-1/2}) \tag{A.7}$$

(see for instance the Proof of Proposition 4 in Broniatowski and Leorato). For the other term, we observe that

$$\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R} = \begin{pmatrix} \mathsf{a} & \mathsf{b} \\ \mathsf{b} & \mathsf{c} \end{pmatrix} \qquad \hat{\mathbf{R}}^\top\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}} = \begin{pmatrix} \hat{\mathsf{a}} & \hat{\mathsf{b}} \\ \hat{\mathsf{b}} & \hat{\mathsf{c}} \end{pmatrix}$$

where $\mathsf{a} = \sum_{i,j}\boldsymbol{V}^-(i,j)$, $\mathsf{b} = \sum_i Q_i(x)\sum_j \boldsymbol{V}^-(i,j)$, $\mathsf{c} = \sum_{i,j}\boldsymbol{V}^-(i,j)Q_i(x)Q_j(x)$ and $\boldsymbol{V}^-(i,j)$ is the (i,j)th element of the matrix $\boldsymbol{V}^{-1}$. Moreover, $\hat{\mathsf{a}}$, $\hat{\mathsf{b}}$ and $\hat{\mathsf{c}}$ are the empirical counerparts.

Therefore,

$$\left(\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R}\right)^{-1} = \frac{1}{\mathsf{ac} - \mathsf{b}^2}\begin{pmatrix} \mathsf{c} & -\mathsf{b} \\ -\mathsf{b} & \mathsf{a} \end{pmatrix} \qquad \left(\hat{\mathbf{R}}^\top\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}\right)^{-1} = \frac{1}{\hat{\mathsf{a}}\hat{\mathsf{c}} - \hat{\mathsf{b}}^2}\begin{pmatrix} \hat{\mathsf{c}} & -\hat{\mathsf{b}} \\ -\hat{\mathsf{b}} & \hat{\mathsf{a}} \end{pmatrix}$$

and

$$\left\|(\hat{\mathbf{R}}^\top\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - (\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R})^{-1}\right\| = \lambda_{\max}\left((\hat{\mathbf{R}}^\top\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - (\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R})^{-1}\right). \tag{A.8}$$

In order to compute the maximum eigenvalue in (A.8), we consider

$$\det\left((\hat{\mathbf{R}}^\top\hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}})^{-1} - (\mathbf{R}^\top\boldsymbol{V}^{-1}\mathbf{R})^{-1} - \lambda I\right) = 0$$

that gives

$$\lambda = \frac{1}{2}\left[\frac{(\mathsf{a}+\mathsf{c})(\hat{\mathsf{a}}\hat{\mathsf{c}} - \hat{\mathsf{b}}^2) - (\hat{\mathsf{a}}+\hat{\mathsf{c}})(\mathsf{ac} - \mathsf{b}^2)}{(\mathsf{ac} - \mathsf{b}^2)(\hat{\mathsf{a}}\hat{\mathsf{c}} - \hat{\mathsf{b}}^2)} \pm \sqrt{\Delta_\lambda}\right]$$

where for the determinant it holds

$$\Delta_\lambda = \left[\frac{\mathsf{a}+\mathsf{c}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{a}}+\hat{\mathsf{c}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right]^2 - 4\left[\frac{\mathsf{b}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{b}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right] \leq \left[\left|\frac{\mathsf{a}+\mathsf{c}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{a}}+\hat{\mathsf{c}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right| + 2\left|\frac{\mathsf{b}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{b}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right|\right]^2.$$

Thus, the maximum eigenvalue is

$$\begin{aligned}\lambda_{\max} &\leq \left[\frac{\mathsf{a}+\mathsf{c}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{a}}+\hat{\mathsf{c}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right] + \left[\left|\frac{\mathsf{a}+\mathsf{c}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{a}}+\hat{\mathsf{c}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right| + 2\left|\frac{\mathsf{b}}{\mathsf{ac}-\mathsf{b}^2} - \frac{\hat{\mathsf{b}}}{(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)}\right|\right] \\ &\leq \frac{2\max\{\mathsf{ac}-\mathsf{b}^2, \hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2\}\left[|\hat{\mathsf{a}}-\mathsf{a}|+\left|\hat{\mathsf{b}}-\mathsf{b}\right|+|\hat{\mathsf{c}}-\mathsf{c}|\right]}{(\mathsf{ac}-\mathsf{b}^2)(\hat{\mathsf{a}}\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)} \\ &= \frac{2\left[|\hat{\mathsf{a}}-\mathsf{a}|+\left|\hat{\mathsf{b}}-\mathsf{b}\right|+|\hat{\mathsf{c}}-\mathsf{c}|\right]}{\min\left\{\det\left(\mathbf{R}^\top \boldsymbol{V}^{-1}\mathbf{R}\right), \det\left(\hat{\mathbf{R}}^\top \hat{\boldsymbol{V}}^{-1}\hat{\mathbf{R}}\right)\right\}}.\end{aligned}$$

The term in the denominator is strictly positive. The term $\|\hat{\mathsf{b}}-\mathsf{b}\| = o_P(1)$ because of the $\sqrt{T}-$consistency of $\hat{Q}_i(x)$, for all $i = 1\ldots, I$ and because of the consistency of $\hat{\boldsymbol{V}}$ for $\boldsymbol{V}$, that also implies $\|\hat{\mathsf{a}}-\mathsf{a}\| = o_P(1)$. Finally,

$$\begin{aligned}\|\hat{\mathsf{c}}-\mathsf{c}\| \quad &\leq \textstyle\sum_{i,j}\left|\hat{Q}_i(x)\hat{Q}_j(x) - Q_i(x)Q_j(x)\right|\boldsymbol{V}^-(i,j) \\ &\quad + \textstyle\sum_{i,j}\left(\hat{Q}_i(x)\hat{Q}_j(x)\right)\left|\hat{\boldsymbol{V}^-}(i,j) - \boldsymbol{V}^-(i,j)\right| = o_P(1).\end{aligned}$$

So far, we have proved that, for every $\delta \in \Delta$, $\left\|\hat{\overline{\mathbf{w}}}^*_\delta - \overline{\mathbf{w}}^*\right\| = O_P(\delta) + o_P(1)$ where the o_P term is independent on δ. Then, in view of $\hat{\overline{\mathbf{w}}}^* = \arg\min_{\delta\in\Delta}\mathbb{M}_T\left(\hat{\overline{\mathbf{w}}}^*_\delta\right)$ it immediately follows that

$$\left\|\hat{\overline{\mathbf{w}}}^* - \overline{\mathbf{w}}^*\right\| \leq \sup_{\delta\in\Delta}\left\|\hat{\overline{\mathbf{w}}}^*_\delta - \overline{\mathbf{w}}^*\right\| \leq O_P(r_T) + o_P(1).$$