Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models

Stéphane Girard, Gilles Stupfler, Antoine Usseglio-Carleve

Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models

Stéphane Girard[1], **Gilles Stupfler**[2,†] **and Antoine Usseglio-Carleve**[1,2,*]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, e-mail:* Stephane.Girard@inria.fr; [*] Antoine.Usseglio-Carleve@ensai.fr

[2] *Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France, e-mail:* [†] gilles.stupfler@ensai.fr

Abstract: Expectiles define a least squares analogue of quantiles. They have been the focus of a substantial quantity of research in the context of actuarial and financial risk assessment over the last decade. The behaviour and estimation of unconditional extreme expectiles using independent and identically distributed heavy-tailed observations has been investigated in a recent series of papers. We build here a general theory for the estimation of extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise; our approach is supported by general results of independent interest on residual-based extreme value estimators in heavy-tailed regression models, and is intended to cope with covariates having a large but fixed dimension. We demonstrate how our results can be applied to a wide class of important examples, among which linear models, single-index models as well as ARMA and GARCH time series models. Our estimators are showcased on a numerical simulation study and on real sets of actuarial and financial data.

MSC 2010 subject classifications: Primary 62G32; secondary 62G08, 62G20, 62G30.
Keywords and phrases: Expectiles, extreme value analysis, heavy-tailed distribution, heteroscedasticity, regression models, residual-based estimators, single-index model, tail empirical process of residuals.

1. Introduction

1.1. Motivation

A traditional way of considering extreme events is to estimate extreme quantiles of a random variable $Y \in \mathbb{R}$, such as the negative daily log-return of a stock market index in finance, so that large values of Y correspond to extreme losses on the market, or the magnitude of a claim in insurance. A better understanding of the extremes of Y can often be achieved by inferring the conditional extremes of Y given a covariate $\boldsymbol{X}$. Recent examples include the analysis of high healthcare costs in [51] and large insurance claims in [42]. We focus on the case when Y given $\boldsymbol{X}$ is heavy-tailed (*i.e.* Paretian-tailed); this assumption underpins the aforementioned papers and is generally appropriate to the modelling of actuarial and financial data. Under no further assumptions on the structure of $(\boldsymbol{X}, Y)$, nonparametric smoothing methods such as those of [7, 18] can be used. Those techniques suffer from the curse of dimensionality, compounded in conditional extreme value statistics by the necessity to select only the few high observations relevant to the analysis. Early attempts at tackling the low-dimensional restriction, such as [12], were built on parametric models. Later attempts have mostly used quantile regression: a seminal paper is [6], developed further by [23, 51, 52]. An approach based on Tail Dimension Reduction was adopted by [17].

These techniques, and more generally the current state of art in conditional extreme value analysis, rely on quantiles, which only use the information on the frequency of tail events and not on their actual magnitudes. This is an issue in risk assessment, where knowing the magnitude of typical extreme losses is important. One way of tackling this problem is to work with expectiles, introduced in [39]. The τth regression expectile of Y given $\boldsymbol{X}$ is obtained from the τth regression quantile by replacing absolute deviations by squared deviations:

$$\xi_\tau(Y|\boldsymbol{x}) = \underset{\theta\in\mathbb{R}}{\arg\min}\, \mathbb{E}\left(\left[\eta_\tau(Y-\theta) - \eta_\tau(Y)\right]|\boldsymbol{X}=\boldsymbol{x}\right),$$

where $\eta_\tau(y) = |\tau - \mathbb{1}\{y \leq 0\}|\, y^2$ is the expectile check function and $\mathbb{1}\{\cdot\}$ is the indicator function. Expectiles are well-defined and unique when the underlying distribution has a finite first moment

(see [1] and Theorem 1 in [39]). Unlike quantiles, expectiles depend on both the probability of tail values and their realisations (see [32]). In addition, expectiles induce the only coherent, law-invariant and elicitable risk measure (see [59]) and therefore benefit from the existence of a natural backtesting methodology. Expectiles are thus a sensible risk management tool to use, as a complement or an alternative to quantiles.

The literature has essentially focused on estimating expectiles with a fixed level τ (see *e.g.* [27, 31]). The estimation of extreme expectiles, where $\tau = \tau_n \to 1$ as the sample size n tends to infinity, remains largely unexplored; it was initiated by [9, 11] in the unconditional heavy-tailed case. Our focus is to provide and discuss the theory of estimators of extreme conditional expectiles, in models that may cope with a large but fixed dimension of the covariate $\boldsymbol{X}$. In doing so we shall develop a novel theory of independent interest for the asymptotic analysis of residual-based extreme value estimators in heavy-tailed regression models.

1.2. Expectiles and regression models

We outline our general idea in the location-scale shift linear regression model. Let $(\boldsymbol{X}_i, Y_i)$, $1 \leq i \leq n$ be a sample from a random pair $(\boldsymbol{X}, Y)$ such that $Y = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X} + (1 + \boldsymbol{\theta}^\top \boldsymbol{X})\varepsilon$. The parameters $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ are unknown, and so are the distributions of the covariate $\boldsymbol{X} \in \mathbb{R}^d$ and the unobserved noise variable $\varepsilon \in \mathbb{R}$. We also suppose that $\boldsymbol{X}$ is independent of ε, and has a support K such that $1 + \boldsymbol{\theta}^\top \boldsymbol{x} > 0$ for all $\boldsymbol{x} \in K$. In this model, by location equivariance and positive homogeneity of expectiles (Theorem 1(iii) in [39]), we may write $\xi_\tau(Y|\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x} + (1 + \boldsymbol{\theta}^\top \boldsymbol{x})\xi_\tau(\varepsilon)$. A natural idea to estimate the extreme conditional expectile $\xi_{\tau_n}(Y|\boldsymbol{x})$, where $\tau = \tau_n \to 1$ as $n \to \infty$, is to first construct estimators $\widehat{\alpha}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ of the model parameters using a weighted least squares method and then construct residuals which can be used, instead of the unobservable errors, to estimate extreme expectiles of ε. This expectile estimator can be adapted from, for instance, an empirical asymmetric least squares method (see [9, 11]). If ε has a finite second moment, the weighted least squares approach produces $\sqrt{n}-$consistent estimators, and it is reasonable to expect that the asymptotic normality properties of the estimators of [9, 11] carry over to their residual-based versions. An estimator of the extreme conditional expectile $\xi_{\tau_n}(Y|\boldsymbol{x})$ is then readily obtained as $\widehat{\xi}_{\tau_n}(Y|\boldsymbol{x}) = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} + (1 + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{x})\widehat{\xi}_{\tau_n}(\varepsilon)$.

Our main objective in this paper is to generalise this construction in heteroscedastic regression models of the form $Y = g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon$ where g and $\sigma > 0$ are two measurable functions of $\boldsymbol{X}$, so that $\xi_{\tau_n}(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_n}(\varepsilon)$. If ε is centred and has unit variance, this model can essentially be viewed as $\mathbb{E}(Y|\boldsymbol{X}) = g(\boldsymbol{X})$ and $\mathrm{Var}(Y|\boldsymbol{X}) = \sigma^2(\boldsymbol{X})$, and is called *location-dispersion regression model* in [49]. Even though our theory will be valid for arbitrary regression models of this form, one should keep in mind models adapted to the consideration of a large dimension d, where the estimation of g and σ will not suffer from the curse of dimensionality and thus reasonable rates of convergence can be achieved. The viewpoint we deliberately adopt is that the estimation of g and σ is the "easy" part of the estimation of $\xi_{\tau_n}(Y|\boldsymbol{x})$ because, depending on the model, it can be tackled by known parametric or semiparametric techniques that are easy to implement and converge faster than the extreme value procedure for the estimation of $\xi_{\tau_n}(\varepsilon)$. This converts the problem of conditional extreme value estimation into the question of being able to carry out extreme value inference based on residuals rather than the unobserved noise variables, which is nonetheless a difficult question because residuals are neither independent nor identically distributed.

In Section 2, given that residuals of the model are available, we provide high-level, fairly easy to check and reasonable sufficient conditions under which the asymptotics of residual-based estimators of $\xi_{\tau_n}(\varepsilon)$ are those of their unfeasible, unobserved error-based counterparts. Several of our results are of independent interest: in particular, we prove in Section 2.2 a non-trivial result on Gaussian approximations of the tail empirical process of the residuals, which is an important step in proving asymptotic theory for extreme value estimators in general regression models. In Section 3, we shall then consider five fully worked-out examples. We start with the location-scale shift linear regression model in Section 3.1, a heteroscedastic single-index model in Section 3.2, and a heteroscedastic, Tobit-type left-censored model in Section 3.3. The latter example allows us to show how our method adapts to a situation

where the model $Y = g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon$ is valid in the right tail rather than globally. Aside from these three examples, we study the two general ARMA and GARCH time series models in Section 3.4 as a way to illustrate how our results may be used to tackle the problem of dynamic extreme conditional expectile estimation. Section 4 examines the behaviour of our estimators on simulated and real data, and Section 5 discusses our findings and research perspectives. All the necessary mathematical proofs, as well as further details and results related to our finite-sample studies, are deferred to the Supplementary Material.

2. General theoretical toolbox for extreme expectile estimation in heavy-tailed regression models

Our general framework is the following. Let $(\boldsymbol{X}_i, Y_i)$, $1 \leq i \leq n$ be part of a (strictly) stationary sequence of copies of the random pair $(\boldsymbol{X}, Y)$, with $Y \in \mathbb{R}$, such that

$$Y = g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon \tag{1}$$

where g and $\sigma > 0$ are two measurable functions of $\boldsymbol{X}$. The unobserved noise variable $\varepsilon \in \mathbb{R}$ is centred and independent of $\boldsymbol{X}$; in other words, for each i, $\boldsymbol{X}_i$ is independent of ε_i, although we do not assume independence between the pairs $(\boldsymbol{X}_i, \varepsilon_i)$. In addition, we suppose throughout that the $\varepsilon_i = (Y_i - g(\boldsymbol{X}_i))/\sigma(\boldsymbol{X}_i)$ are independent.

It follows from this assumption that a conditional expectile $\xi_{\tau_n}(Y|\boldsymbol{x})$ can be written as $\xi_{\tau_n}(Y|\boldsymbol{X} = \boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_n}(\varepsilon|\boldsymbol{X} = \boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_n}(\varepsilon)$, where we used the location equivariance and positive homogeneity to obtain the first identity, and the independence between $\boldsymbol{X}$ and ε to get the second identity. We assume throughout Section 2 that g and σ have been estimated, and we concentrate on estimating the extreme expectile $\xi_{\tau_n}(\varepsilon)$, with the objective of ultimately constructing an estimator of $\xi_{\tau_n}(Y|\boldsymbol{x})$. Denoting by $\tau \mapsto q_\tau(\varepsilon)$ the quantile function of ε, we work under the following first-order Pareto-type condition:

$\mathcal{C}_1(\gamma)$ The tail quantile function of ε, defined by $U(t) = q_{1-t^{-1}}(\varepsilon)$ for $t > 1$, is regularly varying with index $\gamma > 0$: $U(tx)/U(t) \to x^\gamma$ as $t \to \infty$ for any $x > 0$.

Condition $\mathcal{C}_1(\gamma)$ is equivalent to assuming that the survival function of ε, denoted hereafter by $\overline{F} : x \mapsto \mathbb{P}(\varepsilon > x)$, is regularly varying with index $-1/\gamma < 0$ (see [13], Proposition B.1.9). Together with condition $\mathbb{E}|\varepsilon_-| < \infty$, where $\varepsilon_- = \min(\varepsilon, 0)$, the assumption $\gamma < 1$ ensures that the first moment of ε exists, which entails that expectiles of ε of any order are well-defined. Both of these conditions shall be part of our minimal assumptions throughout.

The essential difficulty to overcome in our setup is that the ε_i are unobserved. However, because g and σ have been estimated, by $\overline{g}$ and $\overline{\sigma}$ say, we have access to residuals $\widehat{\varepsilon}_i^{(n)} = (Y_i - \overline{g}(\boldsymbol{X}_i))/\overline{\sigma}(\boldsymbol{X}_i)$ constructed from the regression model (1). Our idea in this section will be to construct estimators of extreme expectiles based on the observable $\widehat{\varepsilon}_i^{(n)}$, and study their theoretical properties when they are in some sense "close" to the true, unobserved ε_i.

We start by the case of an *intermediate level* τ_n, meaning that $\tau_n \to 1$ and $n(1-\tau_n) \to \infty$. Section 2.1 below focuses on a residual-based Least Asymmetrically Weighted Squares (LAWS) estimator. Section 2.2 then introduces a competitor based on the connection between (theoretical) extreme expectiles and quantiles and new general results on tail empirical processes of residuals in heavy-tailed models. Section 2.3 extrapolates these estimators to properly extreme levels τ_n' using a Weissman-type construction warranted by the heavy-tailed assumption (see [53]), and combines these extrapolated devices with the estimators of g and σ to finally obtain an estimator of the extreme conditional expectile $\xi_{\tau_n'}(Y|\boldsymbol{x})$.

2.1. Intermediate step, direct construction: residual-based LAWS

Assume that τ_n is an intermediate sequence, *i.e.* $\tau_n \to 1$ and $n(1-\tau_n) \to \infty$. If the errors ε_i were available, we could estimate $\xi_{\tau_n}(\varepsilon)$ by $\widehat{\xi}_{\tau_n}(\varepsilon)$ minimising $\sum_{i=1}^n \eta_{\tau_n}(\varepsilon_i - u)$ with respect to u. We replace

the unobserved ε_i by the observed residuals $\widehat{\varepsilon}_i^{(n)}$, resulting in the LAWS estimator

$$\widehat{\xi}_{\tau_n}(\varepsilon) = \arg\min_{u\in\mathbb{R}} \sum_{i=1}^{n} \eta_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - u).$$

Our first main theorem is a flexible result stating that $\widehat{\xi}_{\tau_n}(\varepsilon)$ is a $\sqrt{n(1-\tau_n)}-$relatively asymptotically normal estimator of the high, intermediate expectile $\xi_{\tau_n}(\varepsilon)$ provided the gap between residuals and unobservable errors is not too large. For technical extensions to the case of a random sample size or independent arrays, see Lemmas C.5 and C.8.

Theorem 2.1. *Assume that there is $\delta > 0$ such that $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$, that ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0 < \gamma < 1/2$ and $\tau_n \uparrow 1$ is such that $n(1-\tau_n) \to \infty$. Suppose moreover that the array of random variables $\widehat{\varepsilon}_i^{(n)}$, $1 \le i \le n$, satisfies*

$$\sqrt{n(1-\tau_n)} \max_{1\le i\le n} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|}{1+|\varepsilon_i|} \xrightarrow{\mathbb{P}} 0. \tag{2}$$

Then we have $\sqrt{n(1-\tau_n)}\left(\dfrac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \dfrac{2\gamma^3}{1-2\gamma}\right).$

Remark 1. Theorem 2.1 is a non-trivial extension of Theorem 2 in [9] to the case when the ε_i are unobserved. The difference lies in the fact that the estimator $\widehat{\xi}_{\tau_n}(\varepsilon)$ is much more difficult to handle directly; Condition (2), on the weighted distance between the ε_i and the $\widehat{\varepsilon}_i^{(n)}$, allows for a control of the gap between $\widehat{\xi}_{\tau_n}(\varepsilon)$ and the unfeasible $\check{\xi}_{\tau_n}(\varepsilon)$, with the presence of the denominator $1+|\varepsilon_i|$ making it possible to deal with heteroscedasticity in practice. We shall use this key condition again in our results in Section 2.2. It will be satisfied when the structure of the model $Y = g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon$ is estimated at a faster rate than the $\sqrt{n(1-\tau_n)}-$rate of convergence of intermediate expectile estimators. The proof is based on rigorously establishing that $\widehat{\xi}_{\tau_n}(\varepsilon)$ and $\check{\xi}_{\tau_n}(\varepsilon)$ have the same asymptotic distribution; the striking fact is that this only requires stationarity of the ε_i, with independence only being used to conclude that $\check{\xi}_{\tau_n}(\varepsilon)$ is asymptotically Gaussian by Theorem 2 in [9] and therefore $\widehat{\xi}_{\tau_n}(\varepsilon)$ must be so. Theorem 2.1 can then be expected to have analogues when the ε_i are stationary but weakly dependent, thus covering (for example) regression models with time series errors as in [48], as long as one can prove the $\sqrt{n(1-\tau_n)}-$asymptotic normality of $\check{\xi}_{\tau_n}(\varepsilon)$. An example of such a result for stationary and mixing ε_i has been investigated in [10].

This estimator is purely nonparametric. We focus now on an alternative estimator based on a connection between extreme expectiles and quantiles, warranted by the heavy-tailed context.

2.2. Intermediate step, indirect construction

We start by recalling, as shown in Proposition 2.3 of [2], that the heavy-tailed condition on $t \mapsto U(t) = q_{1-t^{-1}}(\varepsilon)$ entails

$$\lim_{\tau\uparrow 1} \frac{\xi_\tau(\varepsilon)}{q_\tau(\varepsilon)} = (\gamma^{-1} - 1)^{-\gamma}.$$

Therefore, if $\overline{\gamma}$ is a consistent estimator of γ, and $\overline{q}_{\tau_n}(\varepsilon)$ is a consistent estimator of $q_{\tau_n}(\varepsilon)$, we can estimate the intermediate expectile $\xi_{\tau_n}(\varepsilon)$ by the so-called indirect estimator

$$\widetilde{\xi}_{\tau_n}(\varepsilon) = (\overline{\gamma}^{-1} - 1)^{-\overline{\gamma}}\, \overline{q}_{\tau_n}(\varepsilon).$$

An extension of Theorem 1 in [9] (see Proposition A.1 in Appendix A) shows that under the following classical second-order refinement of condition $\mathcal{C}_1(\gamma)$, the asymptotic distribution of the estimator $\widetilde{\xi}_{\tau_n}(\varepsilon)$ is determined under high-level conditions on $(\overline{\gamma}, \overline{q}_{\tau_n}(\varepsilon))$.

$\mathcal{C}_2(\gamma, \rho, A)$ For all $x > 0$,

$$\lim_{t\to\infty} \frac{1}{A(t)}\left[\frac{U(tx)}{U(t)} - x^\gamma\right] = x^\gamma \frac{x^\rho - 1}{\rho},$$

where A is a function converging to 0 at infinity and having constant sign, and $\rho \leq 0$. Here and in what follows, $(x^\rho - 1)/\rho$ is to be read as $\log x$ when $\rho = 0$.

We now explain how one may construct and study residual-based estimators $\overline{\gamma}$ and $\overline{q}_{\tau_n}(\varepsilon)$. Let $z_{1,n} \leq z_{2,n} \leq \cdots \leq z_{n,n}$ be the ordered $n-$tuple associated with an $n-$tuple $(z_1, z_2, \ldots, z_n)$. A number of estimators of γ can be adapted to our case and written as a functional of the tail empirical quantile process of the residuals, among which the popular Hill estimator ([24]):

$$\widehat{\gamma}_k = \frac{1}{k}\sum_{i=1}^{k} \log \frac{\widehat{\varepsilon}^{(n)}_{n-i+1,n}}{\widehat{\varepsilon}^{(n)}_{n-k,n}} = \int_0^1 \log\left(\frac{\widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}}{\widehat{\varepsilon}^{(n)}_{n-k,n}}\right) ds.$$

Here $\lfloor\cdot\rfloor$ denotes the floor function. We may also adapt in the same way the moment-type statistics which intervene in the construction of the moment estimator of [14], and the general class of estimators studied by [43]. These estimators depend on the choice of an effective sample size $k = k(n) \to \infty$ and $k/n \to 0$; it is useful to think of k as being $k = \lfloor n(1-\tau_n)\rfloor$.

It is therefore worthwhile to study the asymptotic behaviour of the tail empirical quantile process $s \mapsto \widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}$ of residuals, and of its log-counterpart. This is of course a difficult task, because the array of residuals is not made of independent random variables. To tackle this problem, we first recall that under condition $\mathcal{C}_2(\gamma,\rho,A)$, one can write a weighted uniform Gaussian approximation of the tail empirical quantile process of the (unobserved) ε_i:

$$\frac{\varepsilon_{n-\lfloor ks\rfloor,n}}{q_{1-k/n}(\varepsilon)} = s^{-\gamma} + \frac{1}{\sqrt{k}}\left(\gamma s^{-\gamma-1}W_n(s) + \sqrt{k}A(n/k)s^{-\gamma}\frac{s^{-\rho}-1}{\rho} + s^{-\gamma-1/2-\delta}\,\mathrm{o}_{\mathbb{P}}(1)\right)$$

uniformly in $s \in (0,1]$, where W_n is a sequence of standard Brownian motions and $\delta > 0$ is arbitrarily small (see Theorem 2.4.8 in [13]), provided $k = k(n) \to \infty$, $k/n \to 0$, and $\sqrt{k}A(n/k) = \mathrm{O}(1)$. For certain results which require the study of the log-spacings $\log \varepsilon_{n-\lfloor ks\rfloor,n} - \log \varepsilon_{n-k,n}$, such as the convergence of the Hill estimator, an approximation of the log-tail empirical quantile process is sometimes preferred: uniformly in $s \in (0,1]$,

$$\frac{1}{\gamma}\log\left(\frac{\varepsilon_{n-\lfloor ks\rfloor,n}}{q_{1-k/n}(\varepsilon)}\right) = \log\frac{1}{s} + \frac{1}{\sqrt{k}}\left(s^{-1}W_n(s) + \sqrt{k}A(n/k)\frac{1}{\gamma}\frac{s^{-\rho}-1}{\rho} + s^{-1/2-\delta}\,\mathrm{o}_{\mathbb{P}}(1)\right).$$

Our next result is that, if the error made in the construction of the residuals is not too large, then these approximations hold for the tail empirical quantile process of residuals as well.

Theorem 2.2. *Assume that condition $\mathcal{C}_2(\gamma,\rho,A)$ holds. Let $k = k(n) = \lfloor n(1-\tau_n)\rfloor$ where $\tau_n \uparrow 1$, $n(1-\tau_n) \to \infty$ and $\sqrt{n(1-\tau_n)}A((1-\tau_n)^{-1}) = \mathrm{O}(1)$. Suppose that the array of random variables $\widehat{\varepsilon}^{(n)}_i$, $1 \leq i \leq n$, satisfies (2). Then there exists a sequence W_n of standard Brownian motions such that, for any $\delta > 0$ sufficiently small: uniformly in $s \in (0,1]$,*

$$\frac{\widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}}{q_{1-k/n}(\varepsilon)} = s^{-\gamma} + \frac{1}{\sqrt{k}}\left(\gamma s^{-\gamma-1}W_n(s) + \sqrt{k}A(n/k)s^{-\gamma}\frac{s^{-\rho}-1}{\rho} + s^{-\gamma-1/2-\delta}\,\mathrm{o}_{\mathbb{P}}(1)\right)$$

and

$$\frac{1}{\gamma}\log\left(\frac{\widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}}{q_{1-k/n}(\varepsilon)}\right) = \log\frac{1}{s} + \frac{1}{\sqrt{k}}\left(s^{-1}W_n(s) + \sqrt{k}A(n/k)\frac{1}{\gamma}\frac{s^{-\rho}-1}{\rho} + s^{-1/2-\delta}\,\mathrm{o}_{\mathbb{P}}(1)\right).$$

Theorem 2.2 is the second main contribution of this paper. It is a non-trivial asymptotic result, because there is no guarantee that ranks of the original error sequence are preserved in the residual sequence, and it therefore is not obvious at first sight that Condition (2) on the gap between errors and their corresponding residuals is in fact sufficient to ensure that the tail empirical quantile process based on residuals has similar properties to its unobserved errors-based analogue. As an illustration, we work out the asymptotic properties of the residual-based, Hill-type estimator of the extreme value index γ of the errors, as well as the asymptotic behaviour of the related indirect intermediate expectile estimator in Corollary 2.1 below.

Corollary 2.1. *Assume that condition $\mathcal{C}_2(\gamma,\rho,A)$ holds. Let $\tau_n \uparrow 1$ satisfy $n(1-\tau_n)\to\infty$ and $\sqrt{n(1-\tau_n)}A((1-\tau_n)^{-1})\to\lambda\in\mathbb{R}$. Suppose that the array of random variables $\widehat{\varepsilon}_i^{(n)}$, $1\le i\le n$, satisfies (2). If $\overline{\gamma}=\widehat{\gamma}_{\lfloor n(1-\tau_n)\rfloor}$ and $\overline{q}_{\tau_n}(\varepsilon)=\widehat{\varepsilon}^{(n)}_{n-\lfloor n(1-\tau_n)\rfloor,n}$, then*

$$\sqrt{n(1-\tau_n)}\left(\overline{\gamma}-\gamma,\frac{\overline{q}_{\tau_n}(\varepsilon)}{q_{\tau_n}(\varepsilon)}-1\right)\xrightarrow{d}(\Gamma,\Theta),$$

where $\Gamma\sim\mathcal{N}\left(\lambda/(1-\rho),\gamma^2\right)$ and $\Theta\sim\mathcal{N}\left(0,\gamma^2\right)$ are independent. As a consequence, if moreover $\mathbb{E}|\varepsilon_-|<\infty$, $0<\gamma<1$, $\mathbb{E}(\varepsilon)=0$ and $\sqrt{n(1-\tau_n)}/q_{\tau_n}(\varepsilon)=\mathrm{O}(1)$, one has

$$\sqrt{n(1-\tau_n)}\left(\frac{\widetilde{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)\xrightarrow{d}\mathcal{N}\left(\lambda\left[\frac{m(\gamma)}{1-\rho}-b(\gamma,\rho)\right],\gamma^2\left[1+[m(\gamma)]^2\right]\right),$$

with $m(\gamma)=(1-\gamma)^{-1}-\log(\gamma^{-1}-1)$ and $b(\gamma,\rho)=\dfrac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho}+\dfrac{(\gamma^{-1}-1)^{-\rho}-1}{\rho}$.

This result is our third main contribution. Such results on residual-based extreme value estimators are, to the best of our knowledge, quite scarce in the literature: see Section 2 in [52] and Section 3 in [51] in linear quantile regression models, and Section 3 in [50] in a nonparametric homoscedastic quantile regression model. Our result relaxes the linear or homoscedastic assumptions, and provides a reasonable general theoretical framework for the estimation of the extreme value index and intermediate quantile via residuals of a regression model. Similarly to Theorem 2.2, this may be of wider interest in general extreme value regression problems with heavy-tailed random errors.

Like for fully observed data, the indirect expectile estimator tends to have a lower asymptotic variance than the direct LAWS estimator, although it is typically biased. A comparison of these two estimators in the case of completely observed data is provided in Section 3 of [9].

Remark 2. When, with probability 1, $g(\boldsymbol{X})$ is bounded and $\sigma(\boldsymbol{X})$ is positive and bounded (this is the setup of our simulation study for linear and single-index models, see Section 4.1), one could estimate γ using the $Y_i=g(\boldsymbol{X}_i)+\sigma(\boldsymbol{X}_i)\varepsilon_i$ directly, because then the Y_i all have extreme value index γ (see Lemma A.4 in Appendix A). A competitor to the estimator $\widehat{\gamma}_k$ is thus $\check{\gamma}_k=k^{-1}\sum_{i=1}^k\log(Y_{n-i+1,n}/Y_{n-k,n})$. A numerical comparison of the estimators $\widehat{\gamma}_k$ and $\check{\gamma}_k$ (which we do not report to save space) shows, however, that the residual-based estimator $\widehat{\gamma}_k$ has by far the best finite-sample performance. The idea is that the presence of the shift $g(\boldsymbol{X}_i)$ and scaling $\sigma(\boldsymbol{X}_i)$ in the Y_i introduces a large amount of bias in the estimation of γ by $\check{\gamma}_k$; removing these two components in the calculation of the residuals substantially improves finite-sample results. A related point is made in [13] (p.83).

Remark 3. The earlier work of [25] provides general tools to obtain the asymptotic normality of the Hill estimator based on a filtered process. The essential difference with our approach is that we put our assumptions directly on the gap between the residuals and the unobserved noise variables; by contrast, the methodology of [25] essentially assumes that the residuals are obtained through a parametric filter, and makes technical assumptions on the regularity of the parametric model and the gap between the estimated parameter and its true value. The latter approach is very powerful when working with time series models, as typical such models (ARMA, GARCH, ARMA-GARCH) have a parametric formulation. By contrast, we avoid the parametric specification and therefore can handle a large class of possibly semiparametric regression models (such as heteroscedastic single-index models, see Section 3.2), while still providing useful results for time series models (see Section 3.4).

The theory in [25] allows for non-independent errors in autoregressive time series, see Section 3.2 therein. This corresponds to when the filter does not correctly describe the underlying structure of the time series, and can be used in misspecified models. Our results use the independence of the errors, but may also be extended to the stationary weakly dependent case: our argument for the proof of Theorem 2.2 (and hence for Corollary 2.1) relies on, first, quantifying the gap between the tail empirical quantile process based on the unobserved errors and its version based on the residuals (see Lemma A.3), and then on a Gaussian approximation of the tail empirical quantile process for independent heavy-tailed variables. Inspecting the proofs reveals that both of these steps can in fact

be carried out when the ε_i are only stationary, $\beta-$mixing and satisfy certain anti-clustering conditions, because a Gaussian approximation of the tail empirical quantile process also holds then, see for instance Theorem 2.1 in [15].

2.3. Extrapolation for extreme conditional expectile estimation

We finally develop high-level results for the estimation of properly extreme conditional expectiles whose level $\tau_n' \to 1$ can converge to 1 at an arbitrarily fast rate. One would typically choose $\tau_n' = 1 - p_n$ for an exceedance probability p_n not greater than $1/n$, see *e.g.* Chapter 4 of [13] in the context of extreme quantile estimation. Following [53], intermediate quantiles of order τ_n can be extrapolated to the extreme level τ_n', using the heavy-tailed assumption. This idea successfully carries over to expectile estimation because of the asymptotic proportionality relationship $\xi_\tau(\varepsilon)/q_\tau(\varepsilon) \to (\gamma^{-1}-1)^{-\gamma}$ as $\tau \uparrow 1$, resulting in the following class of estimators of $\xi_{\tau_n'}(\varepsilon)$:

$$\overline{\xi}^{\star}_{\tau_n'}(\varepsilon) = \left(\frac{1-\tau_n'}{1-\tau_n}\right)^{-\overline{\gamma}} \overline{\xi}_{\tau_n}(\varepsilon),$$

where $\overline{\gamma}$ and $\overline{\xi}_{\tau_n}(\varepsilon)$ are consistent estimators of γ and of the intermediate expectile $\xi_{\tau_n}(\varepsilon)$. In the context of a regression model of the form (1), these would be based on the residuals obtained via estimators $\overline{g}(\boldsymbol{x})$ and $\overline{\sigma}(\boldsymbol{x})$ of $g(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$. One can then estimate $\xi_{\tau_n'}(Y|\boldsymbol{x})$ in model (1) by $\overline{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x}) = \overline{g}(\boldsymbol{x}) + \overline{\sigma}(\boldsymbol{x})\overline{\xi}^{\star}_{\tau_n'}(\varepsilon)$. We examine the convergence of this estimator.

Theorem 2.3. *Assume that $\mathbb{E}|\varepsilon_-| < \infty$ and condition $\mathcal{C}_2(\gamma,\rho,A)$ holds with $0 < \gamma < 1$ and $\rho < 0$. Assume further that $\mathbb{E}(\varepsilon) = 0$ and $\tau_n, \tau_n' \uparrow 1$ satisfy*

$$n(1-\tau_n) \to \infty,\ \frac{1-\tau_n'}{1-\tau_n} \to 0,\ \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]} \to \infty, \tag{3}$$

$$\sqrt{n(1-\tau_n)}A((1-\tau_n)^{-1}) \to \lambda \in \mathbb{R} \quad \text{and} \quad \frac{\sqrt{n(1-\tau_n)}}{q_{\tau_n}(\varepsilon)} = \mathrm{O}(1). \tag{4}$$

Suppose also that $\sqrt{n(1-\tau_n)}(\overline{\xi}_{\tau_n}(\varepsilon)/\xi_{\tau_n}(\varepsilon) - 1) = \mathrm{O}_{\mathbb{P}}(1)$ and $\sqrt{n(1-\tau_n)}(\overline{\gamma}-\gamma) \xrightarrow{d} \Gamma$, where Γ is nondegenerate. Then

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]}\left(\frac{\overline{\xi}^{\star}_{\tau_n'}(\varepsilon)}{\xi_{\tau_n'}(\varepsilon)} - 1\right) \xrightarrow{d} \Gamma.$$

Finally, if model (1) holds (with $\boldsymbol{X}$ independent of ε) and, at a given point $\boldsymbol{x}$, the estimators $\overline{g}(\boldsymbol{x})$ and $\overline{s}(\boldsymbol{x})$ satisfy $\overline{g}(\boldsymbol{x}) - g(\boldsymbol{x}) = \mathrm{O}_{\mathbb{P}}(1)$ and $\sqrt{n(1-\tau_n)}(\overline{\sigma}(\boldsymbol{x}) - \sigma(\boldsymbol{x})) = \mathrm{O}_{\mathbb{P}}(1)$, then

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]}\left(\frac{\overline{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x})}{\xi_{\tau_n'}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \Gamma.$$

Remark 4. This result applies to the residual-based direct LAWS estimator and indirect quantile-based estimator under the conditions that ensure their $\sqrt{n(1-\tau_n)}-$consistency. These conditions essentially amount to assuming that the structure of the model is estimated at a rate faster than $\sqrt{n(1-\tau_n)}$, see Theorem 2.1, the related Remark 1, and Corollary 2.1.

The goal of the next section is to work out several examples where Condition (2) on the rate of estimation of the structure of the model can be satisfied. In each example, we explain our estimation method, and show how this yields estimators of extreme conditional expectiles whose asymptotic behaviour we analyse, using the high-level results we have just developed.

3. Applications of our theoretical results

3.1. Location-scale shift linear regression model

We concentrate here on applications in the popular example of location-scale shift linear regression model, which we recall below.

Model (M_1) The random pair $(\boldsymbol{X}, Y)$ is such that $Y = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X} + (1 + \boldsymbol{\theta}^\top \boldsymbol{X})\varepsilon$. Here the random covariate $\boldsymbol{X}$ is independent of the centred noise variable ε, and has a density function $f_{\boldsymbol{X}}$ on $\mathbb{R}^d$ whose support is a compact set K such that $1 + \boldsymbol{\theta}^\top \boldsymbol{x} > 0$ for all $\boldsymbol{x} \in K$.

Model (M_1) features heteroscedasticity. It is well-known that in this model, traditional methods such as ordinary least squares are consistent but inefficient. A particular concern in our case is also to find accurate estimators of the heteroscedasticity parameter $\boldsymbol{\theta}$; indeed,

$$\xi_{\tau_n}(Y|\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x} + (1 + \boldsymbol{\theta}^\top \boldsymbol{x})\xi_{\tau_n}(\varepsilon) \text{ with } \xi_{\tau_n}(\varepsilon) \to \infty \text{ as } n \to \infty,$$

so that, when n is large, even a moderately large error in the estimation of $\boldsymbol{\theta}$ can result in a substantial error in the estimation of the extreme conditional expectile $\xi_{\tau_n}(Y|\boldsymbol{x})$. We suggest a two-stage procedure to estimate $(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$, based on independent data points $(\boldsymbol{X}_i, Y_i)_{1 \le i \le n}$.

1. (Preliminary step) Compute the ordinary least squares estimators $\widetilde{\alpha}$ and $\widetilde{\boldsymbol{\beta}}$ of α and $\boldsymbol{\beta}$, and then the ordinary least squares estimator of $\boldsymbol{\theta}$ based on the absolute residuals $\widetilde{Z}_i = |Y_i - (\widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)|$, that is, $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\nu}}/\widetilde{\mu}$ and

$$(\widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}) = \operatorname*{arg\,min}_{(a,\boldsymbol{b})} \sum_{i=1}^n (Y_i - a - \boldsymbol{b}^\top \boldsymbol{X}_i)^2, \quad (\widetilde{\mu}, \widetilde{\boldsymbol{\nu}}) = \operatorname*{arg\,min}_{(c,\boldsymbol{d})} \sum_{i=1}^n (\widetilde{Z}_i - c - \boldsymbol{d}^\top \boldsymbol{X}_i)^2.$$

2. (Weighted step) Compute the least squares estimators $\widehat{\alpha}$ and $\widehat{\boldsymbol{\beta}}$ of α and $\boldsymbol{\beta}$, weighted using estimated standard deviations obtained via $\widetilde{\boldsymbol{\theta}}$, and then the weighted least squares estimator of $\boldsymbol{\theta}$ based on the absolute residuals $\widehat{Z}_i = |Y_i - (\widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)|$, *i.e.* $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\nu}}/\widehat{\mu}$ and

$$(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = \operatorname*{arg\,min}_{(a,\boldsymbol{b})} \sum_{i=1}^n \left(\frac{Y_i - a - \boldsymbol{b}^\top \boldsymbol{X}_i}{1 + \widetilde{\boldsymbol{\theta}}^\top \boldsymbol{X}_i} \right)^2, \quad (\widehat{\mu}, \widehat{\boldsymbol{\nu}}) = \operatorname*{arg\,min}_{(c,\boldsymbol{d})} \sum_{i=1}^n \left(\frac{\widehat{Z}_i - c - \boldsymbol{d}^\top \boldsymbol{X}_i}{1 + \widetilde{\boldsymbol{\theta}}^\top \boldsymbol{X}_i} \right)^2.$$

Remark 5. This is a one-iteration version of a general weighted least squares procedure where estimates obtained at a given step are fed back into the next iteration to update weights, this procedure being repeated n_0 times. Simulation results seem to indicate that iterating the procedure further does not improve the accuracy of the estimators in practice.

Once these estimates have been obtained, we can construct the sample of (weighted) residuals $\widehat{\varepsilon}_i^{(n)} = (Y_i - (\widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i))/(1 + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{X}_i)$. One can then estimate $\xi_{\tau_n}(\varepsilon)$ by the direct LAWS estimator $\widehat{\xi}_{\tau_n}(\varepsilon)$ described in Section 2.1. The weighted least squares estimators of α, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are $\sqrt{n}-$consistent (for the sake of completeness, we state this result as Lemma C.1 in Appendix C). The consistency and asymptotic normality of $\widehat{\xi}_{\tau_n}(\varepsilon)$ are therefore a corollary of Theorem 2.1, and this in turn yields the asymptotic behaviour of the estimators

$$\widehat{\xi}_{\tau_n}(Y|\boldsymbol{x}) = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} + (1 + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{x})\widehat{\xi}_{\tau_n}(\varepsilon) \quad \text{(intermediate level)}$$

$$\text{and} \quad \widehat{\xi}^\star_{\tau'_n}(Y|\boldsymbol{x}) = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} + (1 + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{x}) \left(\frac{1 - \tau'_n}{1 - \tau_n} \right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_n}(\varepsilon) \quad \text{(extreme level)}$$

where $\overline{\gamma}$ is a consistent estimator of γ constructed on the residuals.

Corollary 3.1. *Assume that the setup is that of the heteroscedastic linear model* (M_1)*. Suppose also that* $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$ *for some* $\delta > 0$*, and that* $\tau_n \uparrow 1$ *with* $n(1 - \tau_n) \to \infty$*.*

(i) Assume that ε *satisfies condition* $\mathcal{C}_1(\gamma)$ *with* $0 < \gamma < 1/2$*. Then for any* $\boldsymbol{x} \in K$*,*

$$\sqrt{n(1 - \tau_n)} \left(\frac{\widehat{\xi}_{\tau_n}(Y|\boldsymbol{x})}{\xi_{\tau_n}(Y|\boldsymbol{x})} - 1 \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\gamma^3}{1 - 2\gamma}\right).$$

(ii) Assume further that ε *satisfies condition* $\mathcal{C}_2(\gamma, \rho, A)$ *with* $\rho < 0$*. Suppose also that* $\tau_n, \tau'_n \uparrow 1$ *satisfy* (3) *and* (4)*. If there is a nondegenerate limiting random variable* Γ *such that* $\sqrt{n(1 - \tau_n)}(\overline{\gamma} - \gamma) \xrightarrow{d} \Gamma$*, then*

$$\frac{\sqrt{n(1 - \tau_n)}}{\log[(1 - \tau_n)/(1 - \tau'_n)]} \left(\frac{\widehat{\xi}^\star_{\tau'_n}(Y|\boldsymbol{x})}{\xi_{\tau'_n}(Y|\boldsymbol{x})} - 1 \right) \xrightarrow{d} \Gamma.$$

We may similarly obtain the asymptotic normality of the indirect estimators $\widetilde{\xi}_{\tau_n}(Y|\boldsymbol{x})$ and $\widetilde{\xi}^{\star}_{\tau'_n}(Y|\boldsymbol{x})$ of the intermediate and extreme expectiles $\xi_{\tau_n}(Y|\boldsymbol{x})$ and $\xi_{\tau'_n}(Y|\boldsymbol{x})$, defined as

$$\widetilde{\xi}_{\tau_n}(Y|\boldsymbol{x}) = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} + (1+\widehat{\boldsymbol{\theta}}^\top \boldsymbol{x})(\overline{\gamma}^{-1}-1)^{-\overline{\gamma}} \widehat{\varepsilon}^{(n)}_{n-\lfloor n(1-\tau_n)\rfloor,n}$$
$$\text{and } \widetilde{\xi}^{\star}_{\tau'_n}(Y|\boldsymbol{x}) = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} + (1+\widehat{\boldsymbol{\theta}}^\top \boldsymbol{x}) \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\overline{\gamma}} (\overline{\gamma}^{-1}-1)^{-\overline{\gamma}} \widehat{\varepsilon}^{(n)}_{n-\lfloor n(1-\tau_n)\rfloor,n}.$$

Here $\overline{\gamma}$ is the residual-based Hill estimator of γ; the asymptotic properties of the estimators are obtained using Corollary 2.1 and Theorem 2.3. See Corollary E.1 in Appendix E.

Remark 6. Corollary 3.1 requires a second moment of the noise variable ε because of the use of the weighted least squares method and the residual-based LAWS estimator of intermediate expectiles. The R package `CASdatasets` contains numerous examples of real actuarial data sets for which the assumption of a finite variance is perfectly sensible. When this assumption is violated, the alternative is to use a more robust method for the estimation of the model structure and then use the indirect expectile estimator of Section 2.2. A more robust method for the estimation of α and $\boldsymbol{\beta}$ is, for instance, the one-step estimator of [40]. Such methods typically require some regularity on the joint distribution of $(\boldsymbol{X}, \varepsilon)$, but avoid moment assumptions. The convergence of the indirect expectile-based estimator built on the residuals will then only require a finite first moment, see Corollary 2.1 and Theorem 2.3.

3.2. Heteroscedastic single-index model

The linear structure of the location-scale shift regression model of Section 3.1 has its limitations in that it fails to model more involved regression relationships between Y and $\boldsymbol{X}$. A model with greater flexibility is the heteroscedastic single-index model; this allows to handle complicated regression equations, including when the dimension d is large, in a satisfactory way thanks to the single-index structure.

Model (M_2) The random pair $(\boldsymbol{X}, Y)$ is such that $Y = g(\boldsymbol{\beta}^\top \boldsymbol{X}) + \sigma(\boldsymbol{\beta}^\top \boldsymbol{X})\varepsilon$. Here g and $\sigma > 0$ are measurable functions. The random covariate $\boldsymbol{X}$ is independent of the noise variable ε, and has a density function $f_{\boldsymbol{X}}$ on $\mathbb{R}^d$ whose support is a compact and convex set K with nonempty interior K^{o}. Besides, the variable ε is centred and such that $\mathbb{E}|\varepsilon| = 1$.

For identifiability purposes, we will assume that g is continuously differentiable, $\|\boldsymbol{\beta}\| = 1$ (where $\|\cdot\|$ denotes the Euclidean norm) and that the first non-zero component of $\boldsymbol{\beta}$ is positive. This guarantees that $\boldsymbol{\beta}$ is identifiable. Other sets of identifiability conditions are possible, see *e.g.* [28]. In this regression model, the conditional mean and variance have the same single-index structure. There are analogue models where the direction of projection in σ is a vector $\boldsymbol{\theta}$ possibly different from $\boldsymbol{\beta}$ (see *e.g.* [58]). In practice, model (M_2) is already very flexible, and for the sake of simplicity we therefore ignore this more general case; in the latter, the direction in the variance component can be estimated at the $\sqrt{n}-$rate (see Theorem 1 in [58]), and it is readily checked that our methodology below extends to this case.

In model (M_2), $\xi_{\tau_n}(Y|\boldsymbol{x}) = g(\boldsymbol{\beta}^\top \boldsymbol{x}) + \sigma(\boldsymbol{\beta}^\top \boldsymbol{x})\xi_{\tau_n}(\varepsilon)$. There are numerous $\sqrt{n}-$consistent estimators of $\boldsymbol{\beta}$ (see *e.g.* Chapter 2 of [28]). We thus assume that such an estimator $\widehat{\boldsymbol{\beta}}$ has been constructed, *i.e.* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathrm{O}_{\mathbb{P}}(1)$. Estimate now g with

$$\widehat{g}_{h_n,t_n}(z) = \sum_{i=1}^n Y_i \mathbb{1}\{|Y_i| \le t_n\} L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right) \Bigg/ \sum_{i=1}^n L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right).$$

Here L is a probability density function on $\mathbb{R}$, $h_n \to 0$ is a bandwidth sequence and $t_n \to \infty$ is a positive truncating sequence. This is inspired by an estimator of [21]; truncating helps in dealing with heavy tails. Besides, analogously to what we observed in model (M_1), $\sigma(\boldsymbol{\beta}^\top \boldsymbol{X})$ is the conditional first moment of $|Y - g(\boldsymbol{\beta}^\top \boldsymbol{X})|$. Introduce then absolute residuals $\widehat{Z}_{i,h_n,t_n} = |Y_i - \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)|$ and consider a Nadaraya-Watson-type estimator:

$$\widehat{\sigma}_{h_n,t_n}(z) = \sum_{i=1}^n \widehat{Z}_{i,h_n,t_n} \mathbb{1}\left\{\widehat{Z}_{i,h_n,t_n} \le t_n\right\} L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right) \Bigg/ \sum_{i=1}^n L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right).$$

In Proposition C.1 (see Appendix C) we show that, under conditions tailored to our framework, both of these estimators converge uniformly on any compact subset K_0 of the interior of the support of $\boldsymbol{X}$ at the rate $n^{2/5}/\sqrt{\log n}$ under the condition $nh_n^5 \to c \in (0,\infty)$. Similar results, mostly on the estimation of the link function g, are available in the literature; see for example [33] for an estimator based on smoothing splines, as well as references therein.

The residuals are then $\widehat{\varepsilon}_i^{(n)} = (Y_i - \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i))/\widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)$. Translated in terms of these residuals, Proposition C.1 reads

$$\frac{n^{2/5}}{\sqrt{\log n}} \max_{1\le i\le n} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|}{1+|\varepsilon_i|} \mathbb{1}\{\boldsymbol{X}_i \in K_0\} = \mathrm{O}_{\mathbb{P}}(1),$$

for any compact subset K_0 of the interior of the support of $\boldsymbol{X}$. The restriction to such a compact subset makes sense since kernel regression estimators strongly suffer from boundary effects (see, among many others, [34]). This restriction is not important in practice since one would only trust the estimates of g and σ on a sub-domain of the support where sufficiently many observations from $\boldsymbol{X}$ have been recorded. It implies, however, that the residuals $\widehat{\varepsilon}_i^{(n)}$ that can be used for the estimation of the high conditional expectile are those for which $\boldsymbol{X}_i \in K_0$. More precisely, let $\widehat{\varepsilon}_{1,K_0}^{(n)}, \ldots, \widehat{\varepsilon}_{N,K_0}^{(n)}$ be those residuals whose corresponding covariate vectors $\boldsymbol{X}_i \in K_0$ and $N = N(K_0,n) = \sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in K_0\}$ be their total number. Define

$$\widehat{\xi}_{\tau_N}(\varepsilon) = \operatorname*{arg\,min}_{u\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^N \eta_{\tau_N}(\widehat{\varepsilon}_{i,K_0}^{(n)} - u),$$

with $\tau_N = \tau_m$ when $N = m > 0$. This yields the estimators

$$\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x}) = \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}) + \widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x})\widehat{\xi}_{\tau_N}(\varepsilon) \text{ (intermediate level)}$$

$$\text{and } \widehat{\xi}^\star_{\tau'_N}(Y|\boldsymbol{x}) = \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}) + \widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}) \left(\frac{1-\tau'_N}{1-\tau_N}\right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_N}(\varepsilon) \text{ (extreme level).}$$

Again, the estimator $\overline{\gamma}$ is typically calculated using high order statistics of the residuals $\widehat{\varepsilon}_{i,K_0}^{(n)}$; for example, this can be the Hill estimator taking into account the top $\lfloor N(1-\tau_N)\rfloor$ order statistics of these residuals (see Lemma C.6(ii) for the asymptotic properties of this estimator). The next result focuses on the estimators $\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x})$ and $\widehat{\xi}^\star_{\tau'_N}(Y|\boldsymbol{x})$.

Theorem 3.1. *Work in model (M_2). Assume that ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0<\gamma<1/2$ and the conditions of Proposition C.1 in Appendix C hold. Let $\tau_n = 1-n^{-a}$ with $a \in (1/5,1)$, K_0 be a compact subset of K° such that $\mathbb{P}(\boldsymbol{X}\in K_0)>0$, and $N = N(K_0,n)$.*

(i) We have, for any $\boldsymbol{x}\in K_0$, $\sqrt{N(1-\tau_N)}\left(\dfrac{\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x})}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \dfrac{2\gamma^3}{1-2\gamma}\right)$.

(ii) Assume moreover that ε satisfies condition $\mathcal{C}_2(\gamma,\rho,A)$ with $\rho<0$. Suppose also that $\tau_n, \tau'_n \uparrow 1$ satisfy (3) and (4). If there is a nondegenerate limiting random variable Γ such that $\sqrt{N(1-\tau_N)}(\overline{\gamma}-\gamma) \xrightarrow{d} \Gamma$, then for any $\boldsymbol{x}\in K_0$,

$$\frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau'_N)]}\left(\frac{\widehat{\xi}^\star_{\tau'_N}(Y|\boldsymbol{x})}{\xi_{\tau'_N}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \Gamma.$$

Remark 7. Compared to Corollary 3.1, Theorem 3.1 features the additional restriction $\tau_n = 1-n^{-a}$ with $a\in(1/5,1)$. This means that the intermediate expectile to be estimated has to be high enough so that the rate of (semiparametric) estimation of the structure of the model is faster than that of the intermediate expectile and the extreme value index γ.

Remark 8. In Theorem 3.1, the order of the conditional expectile to be estimated and rates of convergence are random and dictated by the number $N = N(K_0,n)$ of covariates $\boldsymbol{X}_i \in K_0$ (where model structure can be estimated at the rate $n^{2/5}/\sqrt{\log n}$). Random convergence rates are not unusual in situations where the effective sample size is random: see, for example, Corollary 1.1 in [45] and Theorem 3 in [54] in the context of randomly truncated observations. The random rate of convergence

$\sqrt{N(1-\tau_N)}$ in Theorem 3.1 can nonetheless be replaced by a nonrandom rate because, with the notation of Theorem 3.1 and if $p_0 = \mathbb{P}(\boldsymbol{X} \in K_0)$, $\sqrt{N(1-\tau_N)} = [np_0]^{(1-a)/2}(1+\mathrm{o}_{\mathbb{P}}(1))$ by the law of large numbers. Similarly, in convergence (ii) and if $\tau'_n = 1 - n^{-b}$ with $b > a$, one can replace $1-\tau'_N$ by the nonrandom sequence $1-\tau'_{np_0} = (np_0)^{-b}$ and the rate of convergence in (ii) can be substituted with the nonrandom rate of convergence $[np_0]^{(1-a)/2}/[(b-a)\log(np_0)]$.

Let us finally mention that, if $\overline{\gamma}$ is the residual-based Hill estimator, an analogous result (Theorem E.1 in Appendix E) holds for the indirect extreme conditional expectile estimator

$$\widetilde{\xi}^{\star}_{\tau'_N}(Y|\boldsymbol{x}) = \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}) + \widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}) \left(\frac{1-\tau'_N}{1-\tau_N}\right)^{-\overline{\gamma}} (\overline{\gamma}^{-1}-1)^{-\overline{\gamma}} \widehat{\varepsilon}^{(n)}_{N-\lfloor N(1-\tau_N)\rfloor,N,K_0}.$$

Again, its asymptotic distribution is controlled by that of $\overline{\gamma}$.

3.3. Heteroscedastic left-censored (Tobit) regression model

We briefly discuss how the assumption that our model describes globally the structure of $(\boldsymbol{X}, Y)$ can be relaxed, through the example of the left-censored regression model below.

Model (M_3) The random pair $(\boldsymbol{X}, Y)$ satisfies $Y = g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon$ when $g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon > y_0$, and $Y = y_0$ otherwise. Here y_0 is known and g and $\sigma > 0$ are measurable functions. The random covariate $\boldsymbol{X} \in \mathbb{R}^d$ is independent of the centred noise variable ε such that $\mathbb{E}|\varepsilon| = 1$. On the support of $\boldsymbol{X}$, the functions g and σ are bounded and σ is bounded away from 0.

When g is linear and σ is constant, this is the Tobit model of [47] with non-Gaussian errors. The heteroscedastic case is considered in *e.g.* [35, 41], where it is shown how a linear g can be estimated at the $\sqrt{n}-$rate, with standard nonparametric rates obtained under no assumption on g. Such models are important in economics (see [47]) and insurance (to model a net loss, *i.e.* claim amount minus deductible when the former exceeds the latter, and 0 otherwise).

Here, if ε is heavy-tailed, there is $\tau_c \in (0,1)$ such that for $\tau \in [\tau_c, 1]$, the conditional quantile function of Y given $\boldsymbol{X}$ satisfies $q_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})q_\tau(\varepsilon)$ (see Lemma C.7(i)), linking model (M_3) to the tail regression models of [50, 52]. We do not have an analogue formula for expectiles because they are not equivariant by taking increasing transformations, but

$$\xi_\tau(Y|\boldsymbol{x}) \approx \xi_\tau(g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon|\boldsymbol{X} = \boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_\tau(\varepsilon) \text{ as } \tau \uparrow 1 \text{ (see Lemma C.7(ii))}$$

which is much weaker than the relationship $\xi_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_\tau(\varepsilon)$ true when the regression model is valid globally. It is also weaker than a specification of the form $\xi_\tau(Y|\boldsymbol{x}) = r(\boldsymbol{x}) + \xi_\tau(\varepsilon)$ for $\tau \in [\tau_c, 1]$, which would be an expectile-based version of the model of [50].

Assume that there are estimators $\widehat{g}$ of g and $\widehat{\sigma}$ of σ which are v_n-uniformly consistent (for some $v_n \to \infty$) on a measurable subset K_0 of the support of $\boldsymbol{X}$ such that $\mathbb{P}(\boldsymbol{X} \in K_0) > 0$. Let $(\mathcal{X}_i, \mathcal{Y}_i, e_i)$ stand for all those N vectors (where N is random) relative to noncensored observations with covariate vectors in K_0, *i.e.* $\mathcal{Y}_i = g(\mathcal{X}_i) + \sigma(\mathcal{X}_i)e_i$ and $\mathcal{X}_i \in K_0$ for $1 \le i \le N$. Construct residuals as $\widehat{e}_i^{(N)} = (\mathcal{Y}_i - \widehat{g}(\mathcal{X}_i))/\widehat{\sigma}(\mathcal{X}_i)$. These approximate unobservable e_i that, given $N = m > 0$, are m i.i.d. copies of a random variable e such that $\mathbb{P}(e > t) = p^{-1}\,\mathbb{P}(\varepsilon > t)$ for t large enough, where $p = \mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}) \mid \boldsymbol{X} \in K_0) > 0$ (see Lemma C.7(iii)). In particular one easily shows that if ε has extreme value index γ, then e has too, and $\xi_\tau(\varepsilon)/\xi_\tau(e) \to p^\gamma$ as $\tau \uparrow 1$ (see Lemma C.7(iv)). Let $N_0 = \sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in K_0\}$. The fact that N/N_0 is a $\sqrt{n}-$consistent estimator of p motivates the estimators

$$\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x}) = \widehat{g}(\boldsymbol{x}) + \widehat{\sigma}(\boldsymbol{x}) \left(\frac{N}{N_0}\right)^{\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}} \widehat{\xi}_{\tau_N}(e) \text{ (intermediate level)}$$

$$\text{and } \widehat{\xi}^{\star}_{\tau_N}(Y|\boldsymbol{x}) = \widehat{g}(\boldsymbol{x}) + \widehat{\sigma}(\boldsymbol{x}) \left(\frac{N}{N_0}\right)^{\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}} \left(\frac{1-\tau'_N}{1-\tau_N}\right)^{-\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}} \widehat{\xi}_{\tau_N}(e) \text{ (extreme level)}$$

where $\widehat{\xi}_{\tau_N}(e)$ is the LAWS estimator of the expectile of e at level τ_N, based on the residuals $\widehat{e}_i^{(N)}$, and $\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}$ is the Hill estimator based on the top $\lfloor N(1-\tau_N)\rfloor$ elements of these same residuals. We examine the convergence of the above estimators next.

Theorem 3.2. *Work in model* (M_3). *Assume that* $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$ *for some* $\delta > 0$, *and suppose that* $\widehat{g}$ *and* $\widehat{\sigma}$ *are* v_n-*uniformly consistent estimators (here* $v_n \to \infty$*) of* g *and* σ *on* K_0 *with* $\mathbb{P}(\boldsymbol{X} \in K_0) > 0$. *Let* $\tau_n = 1 - n^{-a}$ *with* $a \in (0,1)$ *and assume that* $n^{1-a}/v_n^2 \to 0$.

(i) Assume that ε *satisfies condition* $\mathcal{C}_2(\gamma, \rho, A)$ *with* $0 < \gamma < 1/2$. *If* $\sqrt{n(1-\tau_n)}A((1-\tau_n)^{-1}) \to \lambda \in \mathbb{R}$ *and* $\sqrt{n(1-\tau_n)}/q_{\tau_n}(\varepsilon) \to \mu \in \mathbb{R}$ *then, for any* $\boldsymbol{x} \in K_0$,

$$\sqrt{N(1-\tau_N)}\left(\frac{\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x})}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(\mathfrak{b}(\gamma,\rho,p,\boldsymbol{x}), \mathfrak{v}(\gamma,p)\right) \quad \textit{with}$$

$$\begin{aligned}
&\mathfrak{b}(\gamma,\rho,p,\boldsymbol{x}) \\
&= \gamma(\gamma^{-1}-1)^{\gamma}\left(p^{\gamma}\mathbb{E}\left[\varepsilon \,\middle|\, \varepsilon > \frac{y_0 - g(\boldsymbol{X})}{\sigma(\boldsymbol{X})}, \boldsymbol{X} \in K_0\right] - \mathbb{E}\left[\max\left(\varepsilon, \frac{y_0 - g(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)\right]\right)\mu \\
&+ \left\{\frac{p^{-\rho}\log p}{1-\rho} + \frac{p^{-\rho}-1}{\rho}\left(1 + \rho\left[\frac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho} + \frac{(\gamma^{-1}-1)^{-\rho}-1}{\rho}\right]\right)\right\}\lambda \quad \textit{and} \\
&\mathfrak{v}(\gamma,p) = \frac{2\gamma^3}{1-2\gamma} + 2\log p\frac{\gamma^3(\gamma^{-1}-1)^{\gamma}}{(1-\gamma)^2} + (\log p)^2\gamma^2.
\end{aligned}$$

(ii) Assume moreover that $\rho < 0$ *and* $\tau_n, \tau_n' \uparrow 1$ *satisfy* (3) *and* (4). *Then, for any* $\boldsymbol{x} \in K_0$,

$$\frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau_N')]}\left(\frac{\widehat{\xi}^{\star}_{\tau_N'}(Y|\boldsymbol{x})}{\xi_{\tau_N'}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(p^{-\rho}\frac{\lambda}{1-\rho}, \gamma^2\right).$$

Note that when observations with $\boldsymbol{X} \in K_0$ are never censored, we find $p = 1$, $N = \sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in K_0\}$, $\mathfrak{b}(\gamma,\rho,1,\boldsymbol{x}) = 0$ (because $\mathbb{E}(\varepsilon) = 0$) and $\mathfrak{v}(\gamma,1) = 2\gamma^3/(1-2\gamma)$, which then makes convergence (i) above analogous to Theorem 3.1(i). As expected, the asymptotic distribution in (ii) is identical to that of the classical Weissman-Hill estimator when $p = 1$.

3.4. Time series models

Expectiles can be interpreted in terms of the gain-loss ratio. This is a popular performance measure in portfolio management, well-known in the literature on no good deal valuation in incomplete markets (see [2] and references therein). Financial applications typically require working with stationary but dependent time series data. We present here, in two such time series contexts, applications of our results to the dynamic prediction of extreme expectiles given past observations. We only focus on LAWS estimators; extensions of our theory to indirect expectile estimation can be found in Appendix E.

3.4.1. The ARMA model

We start with the following general ARMA(p,q) model.

Model (T_1) The stationary time series $(Y_t)_{t\in\mathbb{Z}}$ satisfies $Y_t = \sum_{j=1}^p \phi_j Y_{t-j} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$ where $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \in \mathbb{R}$ are unknown coefficients. The polynomials $P(z) = 1 - \sum_{j=1}^p \phi_j z^j$ and $Q(z) = 1 + \sum_{j=1}^q \theta_j z^j$ have no common root, and no root inside the unit disk of the complex plane. Finally, (ε_t) is an i.i.d. sequence of copies of ε such that $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon^2) < \infty$, and $\mathbb{P}(\varepsilon > x)/\mathbb{P}(|\varepsilon| > x) \to \ell \in (0,1]$ as $x \to \infty$.

In model (T_1), the process (Y_t) is causal and invertible, and so can be represented as a linear time series in the ε_{t-j}, $j \geq 0$, by Theorem 3.1.1 in [4]. A conditional one-step ahead expectile based on data up to time n is then $\xi_{\tau_n}(Y_{n+1} \,|\, \mathcal{F}_n) = \sum_{j=1}^p \phi_j Y_{n+1-j} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \xi_{\tau_n}(\varepsilon)$ where $\mathcal{F}_n = \sigma(Y_n, Y_{n-1}, \ldots)$ is the past $\sigma-$field at time n. In general, $\sum_{j=1}^p \phi_j Y_{n+1-j} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j}$ is calculated upon the unobservable $\varepsilon_n, \ldots, \varepsilon_{n+1-q}$, which are all linear functions of $(Y_{n+1-j})_{j\geq 0}$ since (Y_t) is an invertible ARMA process. This is why the dynamic expectile $\xi_{\tau_n}(Y_{n+1} \,|\, \mathcal{F}_n)$ to be estimated is conditional upon the whole past $\mathcal{F}_n$ of the process; in the AR(p) case when $q = 0$, this becomes the simpler conditional expectile $\xi_{\tau_n}(Y_{n+1} \,|\, Y_n, Y_{n-1}, \ldots, Y_{n-p+1})$, determined by the past p values only.

Among others, the Gaussian maximum likelihood estimator and the ordinary least squares estimator of the ϕ_j and θ_j are $\sqrt{n}-$asymptotically normal because $\mathbb{E}(\varepsilon^2) < \infty$ (see Theorem 10.8.2 in [4]). We then assume that the estimators $\widehat{\phi}_{1,n}, \ldots, \widehat{\phi}_{p,n}, \widehat{\theta}_{1,n}, \ldots, \widehat{\theta}_{q,n}$ are such that $\widehat{\phi}_{j,n} = \phi_j + \mathrm{O}_{\mathbb{P}}(n^{-1/2})$ and $\widehat{\theta}_{j,n} = \theta_j + \mathrm{O}_{\mathbb{P}}(n^{-1/2})$. To construct residuals, set $\widehat{\varepsilon}^{(n)}_{\max(p,q)-q+1} = \cdots = \widehat{\varepsilon}^{(n)}_{\max(p,q)} = 0$ and define $\widehat{\varepsilon}^{(n)}_t = Y_t - \sum_{j=1}^p \widehat{\phi}_{j,n} Y_{t-j} - \sum_{j=1}^q \widehat{\theta}_{j,n} \widehat{\varepsilon}^{(n)}_{t-j}$, for $\max(p,q)+1 \leq t \leq n$. We consider the asymptotic behaviour of the estimators

$$\widehat{\xi}_{\tau_n}(Y_{n+1} \,|\, \mathcal{F}_n) = \sum_{j=1}^p \widehat{\phi}_{j,n} Y_{n+1-j} + \sum_{j=1}^q \widehat{\theta}_{j,n} \widehat{\varepsilon}^{(n)}_{n+1-j} + \widehat{\xi}_{\tau_n}(\varepsilon) \quad (\tau_n \text{ intermediate}),$$

$$\widehat{\xi}^{\star}_{\tau'_n}(Y_{n+1} \,|\, \mathcal{F}_n) = \sum_{j=1}^p \widehat{\phi}_{j,n} Y_{n+1-j} + \sum_{j=1}^q \widehat{\theta}_{j,n} \widehat{\varepsilon}^{(n)}_{n+1-j} + \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_n}(\varepsilon) \quad (\tau'_n \text{ extreme}),$$

where $\widehat{\xi}_{\tau_n}(\varepsilon)$ is the LAWS estimator of $\xi_{\tau_n}(\varepsilon)$ and $\overline{\gamma}$ is a consistent estimator of γ, both constructed on the residuals $\widehat{\varepsilon}^{(n)}_t$ for $t_n \leq t \leq n$ only, where $t_n / \log n \to \infty$ and $t_n/n \to 0$. This condition on t_n ensures that the influence of the incorrect starting values for the residuals has vanished; in the autoregressive case $q = 0$, one can use all the $\widehat{\varepsilon}^{(n)}_t$ for $p+1 \leq t \leq n$.

Theorem 3.3. *Work in the ARMA model (T_1). Suppose also that there is $\delta > 0$ such that $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$, and that $\tau_n \uparrow 1$ is such that $n(1-\tau_n) \to \infty$.*

(i) Assume that ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0 < \gamma < 1/2$, and that $n^{2\gamma+\iota}(1-\tau_n) \to 0$ for some $\iota > 0$. Then

$$\sqrt{n(1-\tau_n)} \left(\frac{\widehat{\xi}_{\tau_n}(Y_{n+1} \,|\, \mathcal{F}_n)}{\xi_{\tau_n}(Y_{n+1} \,|\, \mathcal{F}_n)} - 1 \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\gamma^3}{1-2\gamma}\right).$$

(ii) Assume further that ε satisfies condition $\mathcal{C}_2(\gamma, \rho, A)$ with $\rho < 0$. Suppose also that $\tau_n, \tau'_n \uparrow 1$ satisfy (3) and (4) (in addition to $n^{2\gamma+\iota}(1-\tau_n) \to 0$). If there is a nondegenerate limiting random variable Γ such that $\sqrt{n(1-\tau_n)}(\overline{\gamma} - \gamma) \xrightarrow{d} \Gamma$, then

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]} \left(\frac{\widehat{\xi}^{\star}_{\tau'_n}(Y_{n+1} \,|\, \mathcal{F}_n)}{\xi_{\tau'_n}(Y_{n+1} \,|\, \mathcal{F}_n)} - 1 \right) \xrightarrow{d} \Gamma.$$

3.4.2. The GARCH model

ARMA models are widely applicable but well-known for failing to replicate the time-varying volatility typically displayed by financial time series. Our next focus is on general GARCH(p,q) models, which arguably constitute the best-known and most employed class of heteroscedastic time series models.

Model (T_2) The stationary time series $(Y_t)_{t\in\mathbb{Z}}$ satisfies $Y_t = \sigma_t \varepsilon_t$, with $\sigma_t > 0$ such that $\sigma_t^2 = \omega + \sum_{j=1}^p \beta_j \sigma^2_{t-j} + \sum_{j=1}^q \alpha_j Y^2_{t-j}$ and $\omega, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p > 0$ are unknown coefficients, and (ε_t) is an i.i.d. sequence of copies of ε such that $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon^2) = 1$ and $\mathbb{P}(\varepsilon^2 = 1) < 1$. Suppose also that the sequence of matrices

$$A_t = \begin{pmatrix}
\alpha_1\varepsilon_t^2 & \cdots & \cdots & \cdots & \alpha_q\varepsilon_t^2 & \beta_1\varepsilon_t^2 & \cdots & \cdots & \cdots & \beta_p\varepsilon_t^2 \\
1 & 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\
\alpha_1 & \cdots & \cdots & \cdots & \alpha_q & \beta_1 & \cdots & \cdots & \cdots & \beta_p \\
0 & \cdots & \cdots & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 0 & 1 & 0
\end{pmatrix}$$

has a negative top Lyapunov exponent, *i.e.* $\lim_{t\to\infty} t^{-1}\mathbb{E}(\log\|A_tA_{t-1}\cdots A_1\|) < 0$ with probability 1 (where $\|\cdot\|$ is an arbitrary matrix norm).

The above condition on (A_t) is necessary and sufficient for the existence of a stationary, nonanticipative solution, see Theorem 2.4 p.30 of [16]. Condition $\mathbb{P}(\varepsilon^2 = 1) < 1$ ensures identifiability. In pure ARCH models ($p = 0$), one can estimate the model with weighted least squares regression of Y_t^2 on its past. This estimator is $\sqrt{n}-$asymptotically normal if $\mathbb{E}(Y_t^4) < \infty$ (see Theorem 6.3 p.132 in [16]). Under further conditions on model coefficients (see p.41 of [16]), this may reduce to $\mathbb{E}(\varepsilon^4) < \infty$, but this is still a substantial restriction in our context of heavy-tailed ε. An alternative is the weighted L^1-regression estimator of [29], whose $\sqrt{n}-$asymptotic normality requires some regularity on the distribution of ε rather than finite moments. In GARCH models, the self-weighted quasi-maximum exponential likelihood estimator of [57] is $\sqrt{n}-$asymptotically normal for square-integrable innovations.

Take $\sqrt{n}-$consistent estimators $\widehat{\omega}_n$, $\widehat{\alpha}_{j,n}$ and $\widehat{\beta}_{j,n}$. To construct residuals, set $\widehat{\sigma}^{(n)}_{\max(p,q)-p+1} = \cdots = \widehat{\sigma}^{(n)}_{\max(p,q)} = \widehat{\omega}_n$, and then define $(\widehat{\sigma}^{(n)}_t)^2 = \widehat{\omega}_n + \sum_{j=1}^p \widehat{\beta}_{j,n}(\widehat{\sigma}^{(n)}_{t-j})^2 + \sum_{j=1}^q \widehat{\alpha}_{j,n}Y^2_{t-j}$ and $\widehat{\varepsilon}^{(n)}_t = Y_t/\widehat{\sigma}^{(n)}_t$, for $\max(p,q)+1 \le t \le n$. Denoting again by $\mathcal{F}_n$ the past $\sigma-$field and letting $\widehat{\sigma}^2_{n+1} = \widehat{\omega}_n + \sum_{j=1}^p \widehat{\beta}_{j,n}\widehat{\sigma}^2_{n+1-j} + \sum_{j=1}^q \widehat{\alpha}_{j,n}Y^2_{n+1-j}$ be the predicted volatility at time $n+1$, one-step ahead estimators of intermediate and extreme conditional expectiles are

$$\widehat{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n) = \widehat{\sigma}_{n+1}\widehat{\xi}_{\tau_n}(\varepsilon) \;\text{ and }\; \widehat{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n) = \widehat{\sigma}_{n+1} \times \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_n}(\varepsilon)$$

respectively, where $\widehat{\xi}_{\tau_n}(\varepsilon)$ is the LAWS estimator of $\xi_{\tau_n}(\varepsilon)$ and $\overline{\gamma}$ is a consistent estimator of γ, both constructed on the residuals $\widehat{\varepsilon}^{(n)}_t$ for $t_n \le t \le n$ only, where $t_n/\log n \to \infty$ and $t_n/n \to 0$ (for pure ARCH models when $p = 0$, all residuals for $q+1 \le t \le n$ may be used).

Theorem 3.4. *Work in the GARCH model (T_2). Suppose also that there is $\delta > 0$ such that $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$, and that $\tau_n = 1 - n^{-a}$ for some $a \in (0,1)$.*

(i) If ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0 < \gamma < 1/2$, then

$$\sqrt{n(1-\tau_n)}\left(\frac{\widehat{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\gamma^3}{1-2\gamma}\right).$$

(ii) Assume further that ε satisfies condition $\mathcal{C}_2(\gamma,\rho,A)$ with $\rho < 0$. Suppose also that $\tau_n, \tau'_n \uparrow 1$ satisfy (3) and (4). If there is a nondegenerate limiting random variable Γ such that $\sqrt{n(1-\tau_n)}(\overline{\gamma}-\gamma) \xrightarrow{d} \Gamma$, then

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\left(\frac{\widehat{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)} - 1\right) \xrightarrow{d} \Gamma.$$

4. Finite-sample study

We showcase our estimators on simulated and real data. Sections 4.1 and 4.2 contain simulation studies, first in the regression models of Sections 3.1 and 3.2, and then in the time series models examined in Section 3.4. Sections 4.3 and 4.4 apply our methods to, respectively, a set of insurance data and a financial time series.

In this section we use, as a way to estimate the extreme value index, the bias-reduced version of the Hill estimator of [19]: this estimator $\widehat{\gamma}^{\mathrm{RB}}_k$ is obtained from the Hill estimator $\widehat{\gamma}_k$ by

$$\widehat{\gamma}^{\mathrm{RB}}_k = \widehat{\gamma}_k\left(1 - \frac{\widehat{b}}{1-\widehat{\rho}}\left(\frac{n}{k}\right)^{\widehat{\rho}}\right),$$

where throughout, $k = \lfloor n(1-\tau_n)\rfloor$, and $\widehat{b}$ and $\widehat{\rho}$ are consistent estimators of the quantities b and ρ under condition $\mathcal{C}_2(\gamma,\rho,A)$ and the additional assumption that $A(t) = b\gamma t^{\rho}$. The estimators $\widehat{b}$ and $\widehat{\rho}$

may be found in [19] and are available from the R function `mop` in the R package `evt0`; of course, we shall use here their residual-based versions. We also consider the following bias-reduced version of the family of direct extreme expectile estimators of ε:

$$\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(\varepsilon) = \widehat{\xi}^{\star}_{\tau'_n}(\varepsilon)\left(1 - \widehat{\gamma}^{\mathrm{RB}}_k \frac{\left((\widehat{\gamma}^{\mathrm{RB}}_k)^{-1} - 1\right)^{-\widehat{\rho}}\left(1-\widehat{\gamma}^{\mathrm{RB}}_k\right)\widehat{b}}{\widehat{\rho}\left(1-\widehat{\gamma}^{\mathrm{RB}}_k - \widehat{\rho}\right)}\left(\frac{n}{k}\right)^{\widehat{\rho}} + 2\widehat{\gamma}^{\mathrm{RB}}_k \frac{k}{n}\right).$$

This expression is motivated by the proof of Proposition 1 and Corollary 1 in [9]. We similarly consider the following bias-reduced version of the family of indirect estimators:

$$\widetilde{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(\varepsilon) = \widetilde{\xi}^{\star}_{\tau'_n}(\varepsilon)\left(1 - \widehat{\gamma}^{\mathrm{RB}}_k \frac{\widehat{b}}{\widehat{\rho}}\left(\frac{n}{k}\right)^{\widehat{\rho}}\right).$$

These procedures improve the accuracy of our estimators, without affecting their asymptotic properties (see [19]). They naturally give rise to extreme conditional expectile estimators $\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(Y|\boldsymbol{x})$ and $\widetilde{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(Y|\boldsymbol{x})$, to which we refer in the present section.

4.1. Simulation study: linear and single-index models

We simulate $N = 1{,}000$ samples of $n = 1{,}000$ observations $(\boldsymbol{X}_i, Y_i)$, $1 \le i \le n$. Here $\boldsymbol{X} \in \mathbb{R}^4$, with independent components, the first three being uniformly distributed on $(0,1)$, and the fourth following a Beta$(2,1)$ distribution. We then simulate from two different models on $(\boldsymbol{X}, Y)$:

(G1) $Y = 1 + \boldsymbol{\beta}^\top \boldsymbol{X} + \left(1/2 + \boldsymbol{\beta}^\top \boldsymbol{X}\right)\varepsilon$.
(G2) $Y = 1 + \exp\left(\boldsymbol{\beta}^\top \boldsymbol{X} - 2\right) + \left(3/2 + \exp\left(\boldsymbol{\beta}^\top \boldsymbol{X} - 2\right)\right)\varepsilon$.

Model (G1) is a location-scale shift linear regression model, while model (G2) is a heteroscedastic single-index model. In both cases, the coefficient vector $\boldsymbol{\beta} = (1,1,1,1)$ and ε is a noise variable, independent of $\boldsymbol{X}$, with a normalised symmetric Burr distribution, that is, $\varepsilon = -\rho\varepsilon_0/B((\gamma-1)/\rho, (\rho-\gamma)/\rho)$, where B is the Beta function and ε_0 has density

$$f_0(x) = \frac{1}{2\gamma}\frac{|x|^{-\rho/\gamma-1}}{\left(1+|x|^{-\rho/\gamma}\right)^{1-1/\rho}} \quad (x \in \mathbb{R}). \tag{5}$$

We consider the cases $\gamma \in \{0.1, 0.2, 0.3, 0.4\}$ and the second-order parameter $\rho = -1$.

Our aim is to estimate extreme expectiles $\xi_{\tau'_n}(Y|\boldsymbol{x})$, in both of these models. We compare the performances of several procedures, constructed using the following four strategies:

(S1) We assume that Y is linked to $\boldsymbol{X}$ by a location-scale shift linear regression model, *i.e.* $Y = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X} + \left(1 + \boldsymbol{\theta}^\top \boldsymbol{X}\right)\varepsilon$. The methodology used for the estimation of $\xi_{\tau'_n}(Y|\boldsymbol{x})$ is outlined in Section 3.1, and the bias-reduced direct estimator is used.
(S1i) Identical to (S1), but the bias-reduced indirect estimator is used instead.
(S2) We assume that Y is linked to $\boldsymbol{X}$ by the heteroscedastic single-index model $Y = g\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) + \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right)\varepsilon$. The vector $\boldsymbol{\beta}$ is estimated using the algorithm of [58] (see 1.(a)–(c) on page 1240 therein), with g and σ estimated using the procedure described in Section 3.2, with $h_n = 0.3$ and $t_n = n^{2/5} \approx 15.85$. The bias-reduced direct estimator is used.
(S2i) Identical to (S2), but the bias-reduced indirect estimator is used instead.

These procedures are compared with the following eight benchmarks:

(B1) We assume no specific structure on $(\boldsymbol{X}, Y)$ and, at $\boldsymbol{X} = \boldsymbol{x}$, we use a local bias-reduced direct estimator relying on those Y_i whose $\boldsymbol{X}_i$ are the 100 nearest neighbours of $\boldsymbol{x}$. In this procedure we use $k = 20$, *i.e.* $\tau_n = 0.8$ for the extrapolation step.
(B1i) Identical to (B1), but the bias-reduced indirect estimator is used instead.
(B2) We assume the homoscedastic single-index model $Y = g\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) + \varepsilon$ with known $\boldsymbol{\beta} = (1,1,1,1)$. The function g is estimated through the Nadaraya-Watson estimator, with a bandwidth chosen using the R package `np`. The bias-reduced direct estimator is used.

(B3) Identical to (S2), although $\boldsymbol{\beta}$ is assumed to be known and equal to $(1,1,1,1)$.
(B4) We assume that the structure of the model linking Y to $\boldsymbol{X}$ is fully known, *i.e.* we know $\boldsymbol{\beta}$ and the location and scale functions, and we use the direct estimator (no bias reduction).
(B4i) Identical to (B4), although the indirect estimator is used instead (no bias reduction).
(B5) Identical to (B4), although the bias-reduced direct estimator is used instead.
(B5i) Identical to (B5), although the bias-reduced indirect estimator is used instead.

In each procedure except (B1) and (B1i), the intermediate expectile level used as an anchor in the extreme value index and extreme expectile estimators is fixed at $\tau_n = 0.9$, corresponding to $k = \lfloor n(1-\tau_n) \rfloor = 100$; in (S2), (S2i), (B2) and (B3), we use the Epanechnikov kernel in the estimation of the link functions g and σ. To assess the performance of our methods, we compute, for a given estimator $\overline{\xi}^{\star}_{\tau'_n}(Y|\boldsymbol{x})$, the Relative Mean Absolute Deviation (RMAD)

$$\overline{\mathrm{RMAD}} = \underset{1\leq m\leq N}{\mathrm{median}} \left| \frac{\overline{\xi}^{\star,(m)}_{\tau'_n}(Y|\boldsymbol{x})}{\xi_{\tau'_n}(Y|\boldsymbol{x})} - 1 \right|,$$

where $\boldsymbol{x}^\top = (1/2, 1/2, 1/2, 1/3)$. The quantity $\overline{\xi}^{\star,(m)}_{\tau'_n}(Y|\boldsymbol{x})$ denotes the estimator calculated on the mth replication, at the level $\tau'_n = 1 - 5/n = 0.995$. The error $\overline{\mathrm{RMAD}}$ gives an idea of the uncertainty on extreme conditional expectiles at a typical data point in the centre of the data cloud. Finally, for all $\alpha \in (0,1)$, the true expectiles $\xi_\alpha(Y|\boldsymbol{x})$ are deduced from $\xi_\alpha(\varepsilon_0)$, obtained by solving the equation $\psi(y)/(2\psi(y)+y) = 1-\alpha$ via the R function `uniroot`, where $\psi(y) = \int_y^\infty \mathbb{P}(\varepsilon_0 > t)\, dt$ is computed with the R function `integrate`.

Results are reported in Table F.1 in Appendix F. In the linear model (G1), methods (S1) and (S1i) are clearly the best, and single-index based methods (S2) and (S2i) perform reasonably well. In fact, for the heaviest tail, methods (S2) and (S2i) slightly outperform (S1) and (S1i) because they are more robust to the highest values in the sample. In the single-index model (G2), methods (S2) and (S2i) perform best, and method (S2) is quite close to the unrealistic benchmark (B3); methods (S1) and (S1i) are heavily penalised by the misspecification of the conditional mean and variance. The nonparametric benchmarks (B1) and (B1i) are surprisingly competitive, perhaps because they benefit from a degree of robustness against heteroscedasticity. Not accounting for heteroscedasticity is indeed very detrimental, as a comparison of method (S2) and benchmarks (B2), (B3) shows, even with the unrealistic advantage of a correct pre-specification of the direction $\boldsymbol{\beta}$. Finally, a comparison of benchmarks (B4) and (B5) shows that even though an unrealistic correct pre-specification of the model structure is obviously beneficial, getting the extreme value step right is very important: in the linear model (G1), method (S1) outperforms benchmark (B4) for $\gamma \in \{0.1, 0.2\}$, and is competitive otherwise, because it features a bias-reduction scheme at the extreme value step.

It appears that while knowing model structure is an advantage for lighter-tailed models, this advantage disappears when the noise variable has a heavier tail, thus illustrating that the extreme value step, rather than model estimation, is indeed the major contributor to estimation error. For instance, when $\gamma = 0.2$, the RMAD of benchmark (B5) is only 5% smaller than the RMAD of method (S2) in the single-index model (G2), and method (S2) is even slightly more accurate when γ is larger. The difference when $\gamma = 0.1$ makes sense: in this setup where extreme expectiles are comparatively smaller, an error on the conditional mean or variance will have more consequences. Let us conclude that while we used the intermediate level $k_n = 100$ for the sake of computational efficiency, in practice one may want to use a data-driven criterion for the choice of k_n. In Appendix F.1, we suggest an adaptation of an Asymptotic Mean-Squared Error (AMSE) minimisation criterion; we repeated this simulation exercise with this choice of k_n and observed that there is no obvious advantage in the data-driven choice although results are competitive. Full results are reported in Table F.2.

4.2. Simulation study: time series models

We simulate $N = 1{,}000$ replications of time series of size $n+1 = 1{,}001$ from two different models:

(T1) An ARMA$(1,1)$ model $Y_t = \phi Y_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t$, where the parameters ϕ and θ are estimated using default settings of the R function `arma` from package `tseries`.
(T2) A GARCH$(1,1)$ model $Y_t = (\omega + \alpha Y_{t-1}^2 + \beta\sigma_{t-1}^2)^{1/2}\,\varepsilon_t$, where ω, α and β are estimated using default settings of the R function `garch` from package `tseries`.

The ε_t are i.i.d. with common density f_0 as in (5) and $\rho = -1$; in the GARCH$(1,1)$ model, these innovations are rescaled by $\sqrt{\Gamma(1-2\gamma)\Gamma(1+2\gamma)}$ to guarantee that $\mathbb{E}[\varepsilon^2] = 1$.

We estimate a one-step ahead extreme expectile $\xi_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)$, where $\mathcal{F}_n$ denotes the past $\sigma-$field at time n. We then compute, on the mth sample, the target value $\xi^{(m)}_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)$, its direct estimate $\widehat{\xi}^{\star,\mathrm{RB},(m)}_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)$ and its indirect counterpart $\widetilde{\xi}^{\star,\mathrm{RB},(m)}_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)$, where $\tau_n' = 1 - 5/n = 0.995$ and $k_n = n(1-\tau_n) = 100$. We calculate their RMAD

$$\mathrm{RMAD} = \underset{1\le m\le N}{\mathrm{median}} \left| \frac{\overline{\xi}^{\star,(m)}_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi^{(m)}_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)} - 1 \right|, \quad \text{with } \overline{\xi}^{\star,(m)}_{\tau_n'} = \widehat{\xi}^{\star,\mathrm{RB},(m)}_{\tau_n'} \text{ or } \widetilde{\xi}^{\star,\mathrm{RB},(m)}_{\tau_n'}.$$

In the ARMA model, we take $\phi, \theta \in \{0.1, 0.5\}$; in the GARCH model, we fix $\omega = 0.1$ and take $(\alpha,\beta) \in \{(0.1,0.1),(0.1,0.45),(0.45,0.1),(0.1,0.85)\}$. In each model, we take $\gamma \in \{0.1, 0.2, 0.3, 0.4\}$. Note that the GARCH model is second-order stationary only if $\alpha+\beta < 1$ (see Theorem 2.5 in [16]). Our methods are compared with the (unrealistic) benchmarks generated from knowing model coefficients (and therefore observing the innovations).

Results are reported in Table F.3. In the ARMA model, the RMAD does not seem overly sensitive to the parameters ϕ and θ, but increases with the tail index γ. In the GARCH model, errors seem to be sensitive to whether the model is close to second-order stationarity (note the slightly different errors in the case $(\alpha,\beta) = (0.1, 0.85)$ and $\gamma \in \{0.1, 0.2\}$). In both models, the indirect estimator has an advantage over the direct estimator, which gets smaller as the tail gets heavier. Knowing the true values of the coefficients does not bring a large improvement, except maybe for the lightest tails; this again underlines that most of the estimation error, and hence of the uncertainty on the estimates, originates from the extreme value step, rather than model estimation. With our data-driven choice of k_n, the indirect estimator typically stays the best, although the direct estimator improves substantially, see Table F.4.

4.3. Real data analysis: Vehicle insurance data

We consider the Vehicle Insurance Customer Data[1], made of $n = 9{,}134$ total (*i.e.* cumulative over the duration of the contract) claim amounts Y of insurance policyholders according to their lifetime value X_1 (in USD), income X_2 (in USD), number X_3 of months since last claim and number X_4 of months since policy inception. We follow the methodology of Section 3.2. A cross-validation procedure using the R function `npindexbw` (from the package `np`) gives a selected bandwidth $h^* \approx 0.1$ (for covariates standardised by their respective maxima). We also choose $t^* = \infty$. We obtain $\widehat{\boldsymbol{\beta}} \simeq (-0.923, 0.386, -0.001, -0.002)$, which seems to indicate that only lifetime value X_1 and income X_2 play a role in the prediction of Y. The estimated functions $\widehat{g}$ and $\widehat{\sigma}$ are depicted in the top left panel of Figure 1 (the kernel function L is the Epanechnikov kernel).

We now estimate an extreme conditional expectile $\xi_{\tau_n'}(Y|\boldsymbol{x})$ at level $\tau_n' = 1 - 1/(nh^*) \approx 0.999$. The top right panel of Figure 1 shows the direct extreme conditional expectile estimator for $k^* = 200$ and $\tau^* = 1 - k^*/n$ (the bottom right panel of Figure 1 shows that the heavy-tailed assumption on the noise is reasonable). The heteroscedastic single-index model captures the variation in the shape of the data cloud fairly well, and the extreme conditional expectile curve gives a reasonable idea of the conditional extremes of the data. Interpreting an expectile curve, meanwhile, is not always straightforward. However, in this insurance example, the expectile $\xi_{\tau_n'}(Y|\boldsymbol{x})$ satisfies the following

[1] Available at `https://www.kaggle.com/ranja7/vehicle-insurance-customer-data` and from the authors upon request.

gain-loss ratio criterion (see [2]):

$$1-\tau_n' \approx \frac{1-\tau_n'}{\tau_n'} = \frac{\mathbb{E}((Y-\xi_{\tau_n'}(Y|\boldsymbol{x}))\mathbb{1}\{Y>\xi_{\tau_n'}(Y|\boldsymbol{x})\}|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{E}((\xi_{\tau_n'}(Y|\boldsymbol{x})-Y)\mathbb{1}\{Y<\xi_{\tau_n'}(Y|\boldsymbol{x})\}|\boldsymbol{X}=\boldsymbol{x})} \approx \frac{\mathbb{E}((Y-\xi_{\tau_n'}(Y|\boldsymbol{x}))\mathbb{1}\{Y>\xi_{\tau_n'}(Y|\boldsymbol{x})\}|\boldsymbol{X}=\boldsymbol{x})}{\xi_{\tau_n'}(Y|\boldsymbol{x})-\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})}.$$

In other words, $\xi_{\tau_n'}(Y|\boldsymbol{x})$ is the aggregate premium to be collected over the lifetime of the contract so that, for customers having the list of characteristics $\boldsymbol{x}$, the ratio between average losses exclusively incurred by claims made by such customers above that level and net average profit is approximately the small quantity $1-\tau_n'$. This value $\xi_{\tau_n'}(Y|\boldsymbol{x})$ can be thus interpreted as a high safety margin for the insurer, and has an even clearer meaning to reinsurers, who only face a loss when the claim exceeds a certain high threshold.

We compare extreme conditional expectile and quantile estimates at the same level τ_n', the latter being obtained by combining the standard Weissman-type estimate of an extreme quantile of the noise with our estimates $\widehat{g}$ and $\widehat{\sigma}$. It can be seen in Figure 1 that the extreme conditional quantile estimate is outside a pointwise 95% bootstrap confidence interval for the extreme conditional expectile (constructed using an adapted methodology called *semiparametric Pareto tail bootstrap*, see Appendix F.2). This may be relevant to insurance companies, for whom lower (*i.e.* more optimistic) assessments of risk translate into marketable contracts with lower premiums and hence improved competitivity, while policymakers and regulators would favour the higher (*i.e.* more pessimistic) quantile estimates to hedge better against systemic risk. Interestingly, the regression median is below the regression mean, so there is a qualitative difference between central and extreme assessments of risk using expectiles and quantiles: a risk assessment based on the regression mean (*i.e.* a central conditional expectile) is more conservative than if it were based on the regression median (*i.e.* a central conditional quantile), but extreme conditional expectile risk measurements are less conservative than those made with extreme conditional quantiles.

4.4. Real data analysis: Australian dollar exchange rates

The analysis of exchange rate risk is a very difficult but key question in economics. Links between exchange rates and fundamental economic principles have been established; among others, in the long term and all other things being equal, a rise in a country's price level is correlated with depreciation of its currency, while an increased demand for exports (resp. imports) is correlated with appreciation (resp. depreciation) of its currency, see *e.g.* p. 201 of [22]. An accurate analysis of exchange rate risk therefore informs strategic decisions made by firms, such as the extent to which they import and export and whether or not they should invest in foreign markets, which have consequences on their competitiveness on the global marketplace. We study the daily log-returns of two exchange rates linking developed, geographically distant economies: the Australian Dollar/Swiss Franc (AUD/CHF) and Australian Dollar/Swedish Krona (AUD/SEK) exchange rates from 1st March 2015 to 28th February 2019, represented in the left panels of Figure 2 (sample size $n = 1{,}043$). The literature has suggested that expectiles can be fruitfully used to estimate quantiles (see *e.g.* [2, 46]). Our goal is then to estimate the (dynamic) extreme conditional quantile $q_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)$ of level $\tau_n' = 0.995 \approx 1-5/n$ on the final day. We consider a GARCH$(1,1)$ model, motivated by the finding of [38] that GARCH models fit past Australian exchange rates well; the R function `garch` (in the package `tseries`) returns, with the notation of Section 3.4.2, $(\widehat{\omega}_n, \widehat{\alpha}_n, \widehat{\beta}_n) = (4.20\times10^{-7}, 0.943, 0.0465)$ for AUD/CHF and $(1.21\times10^{-5}, 0.576, 0.119)$ for AUD/SEK. Based on Section 2.2, we construct the quantile estimator

$$\widehat{q}^{\star,\mathrm{RB}}_{\tau_n'}(\varepsilon) = \big((\widehat{\gamma}^{\mathrm{RB}}_k)^{-1}-1\big)^{\widehat{\gamma}^{\mathrm{RB}}_k}\,\widehat{\xi}^{\star,\mathrm{RB}}_{\tau_n'}(\varepsilon).$$

With $k^*=50$ and $\tau_n = 1-k^*/(n-1)$, we get $\widehat{\gamma}^{\mathrm{RB}}_k = 0.189$ for AUD/CHF (resp. 0.211 for AUD/SEK) and $\widehat{q}^{\star,\mathrm{RB}}_{\tau_n'}(\varepsilon) = 2.40$ (resp. 2.58) (graphical evidence of a heavy right tail of ε is given on the right panels of Figure 2). To check that our estimates make sense, we recall the characterisation of $q_{\tau_n'}(\varepsilon)$

as $0.995 = \tau_n' = \mathbb{E}(\mathbb{1}\{\varepsilon \leq q_{\tau_n'}(\varepsilon)\})$ and compare that with

$$\frac{1}{n-1}\sum_{i=2}^{n} \mathbb{1}\left\{\widehat{\varepsilon}_i^{(n)} < \widehat{q}^{\star,\mathrm{RB}}_{\tau_n'}(\varepsilon)\right\} \approx 0.99424 \text{ for AUD/CHF (resp. 0.99520 for AUD/SEK).}$$

This is indeed very close to the expected value $\tau_n' = 0.995$. Our estimate can be compared with a bias-reduced version $\widetilde{q}^{\star,\mathrm{RB}}_{\tau_n'}(\varepsilon)$ of the classical extrapolated estimate of [53]:

$$\widetilde{q}^{\star,\mathrm{RB}}_{\tau_n'}(\varepsilon) = \widetilde{q}^{\star}_{\tau_n'}(\varepsilon)\left(1 - \widehat{\gamma}^{\mathrm{RB}}_{k}\frac{\widehat{b}}{\widehat{\rho}}\left(\frac{n}{k}\right)^{\widehat{\rho}}\right),$$

where $\widetilde{q}^{\star}_{\tau_n'}(\varepsilon)$ is the residual-based Weissman quantile estimator using $\widehat{\gamma}^{\mathrm{RB}}_{k^\star}$ in its extrapolation step. This estimate is 2.48 for AUD/CHF (resp. 2.64 for AUD/SEK). Our expectile-based estimate of 2.40 (resp. 2.58) is slightly lower; this makes sense, as the estimated value of γ is lower than 1/4, and extreme expectile-based estimates can be thought to reflect this rather light tail by producing lower point estimates than their quantile counterparts (and when $\gamma > 1/4$, expectile-based quantile estimates seem to be higher than traditional estimates, see *e.g.* Section 7.1 in [9]). This lower assessment of risk may be interesting to financial companies, as opposed to regulators who may prefer classical quantile-based estimates. Finally, the predicted quantile estimate of $q_{\tau_n'}(Y_{n+1} \,|\, \mathcal{F}_n)$ on 1st March 2019 is 0.0138 with a asymptotic Gaussian 95% confidence interval (see Appendix F.2) of $[0.0122, 0.0154]$ for AUD/CHF (resp. 0.0156, confidence interval $[0.0136, 0.0178]$ for AUD/SEK). This amounts to a daily variation of 1.4% of the AUD/CHF exchange rate (resp. 1.6% for AUD/SEK).

5. Discussion and perspectives

We provide a general toolbox for the estimation of extreme conditional expectiles, by showing how a simple assumption on the residuals of the model makes it possible to obtain the convergence of residual-based estimators of the extremes of the noise. By applying our results in examples not limited to low dimensions, we contribute to the broader question of how to model extremes with a large number of covariates. The works of [17, 23, 51, 52] introduce dedicated modelling assumptions on the tail conditional quantiles of Y. The tail linear quantile regression model of [52] is not straightforward to interpret: even when the conditional quantile is in fact linear in $\boldsymbol{x}$ (for any τ), this model is the arguably complicated linear model linking Y to $\boldsymbol{X}$ with random coefficients (see p.808 of [6]). Our generic model provides a straightforward way of seeing the effect $\boldsymbol{X}$ has on Y and avoids the crossing problem (unlike the method of [52]), since the structure of the model is estimated only once. The nonparametric model of [17], meanwhile, rests upon the estimation of a Tail Dimension Reduction subspace, which can only be done using the pairs $(\boldsymbol{X}_i, Y_i)$ such that Y_i is large. This entails a potentially substantial loss of modelling strength compared to our approach. Besides, the aforementioned papers focus on the case of i.i.d. data $(\boldsymbol{X}_i, Y_i)$; our method allows us to consider popular time series examples.

Among future research perspectives, it would be nice to extend our results for ARMA and GARCH models in the ARMA-GARCH model, to allow for heteroscedasticity in time series not having mean 0. Besides, the basic principle of our approach relies on location equivariance and positive homogeneity, which are true for numerous interesting functionals, *e.g.* coherent spectral risk measures, including the very recent concept of extremiles ([8]). Adapting our approach to other risk measures constitutes an interesting avenue for further work. Another perspective is to relax the heavy-tailed assumption, to extend the applicability of our method. As far as we know, even in the simple unconditional i.i.d. case, there are currently no estimation procedures available for extreme expectiles of either light-tailed or short-tailed distributions, which are the other setups one would consider in an extreme value framework. The investigation of such procedures and their extension to regression models are beyond the scope of this paper. Finally, an approach that fully accounts for joint uncertainty between model estimation and extreme value estimation would be an important next step in order to handle the strongest possible forms of heteroscedasticity. This will at least require uniform weighted Gaussian approximations of the tail empirical residual-based quantile process; this very difficult question needs

to be solved on a case-by-case basis, because the structure of residuals is completely controlled by the structure of the model. In linear regression, the current state of the art seems to be uniform non-weighted approximations on the real line (see [5], especially Section 6 therein). The absence of weighting, and hence of a meaningful sense of the gap between the tail empirical process and the tail of the true quantile function, makes it impossible to use such results for extreme value inference. We are not aware of such results in single-index models, not even non-weighted and in the homoscedastic case. This is a very substantial research project in itself which we defer to future work.

Acknowledgements

The authors acknowledge an anonymous Associate Editor and three anonymous reviewers for their very helpful comments that led to a greatly improved version of this paper. This research was supported by the French National Research Agency under the grant ANR-19-CE40-0013/ExtremReg project. S. Girard gratefully acknowledges the support of the Chair Stress Test, led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas, and the support of the French National Research Agency in the framework of the Investissements d'Avenir program (ANR-15-IDEX-02).

References

[1] Abdous, B. and Remillard, B. (1995). Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics* **47** 371–384.
[2] Bellini, F. and Di Bernardino, E. (2017). Risk management with expectile. *The European Journal of Finance* **23** 487–506.
[3] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Theory of Probability and its Applications* **10** 323–331.
[4] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods (second edition).* Springer.
[5] Bücher, A., Segers, J. and Volgushev, S. (2014). When uniform weak convergence fails: empirical processes for dependence functions and residuals via epi- and hypographs. *Annals of Statistics* **42** 1598–1634.
[6] Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics* **33** 806–839.
[7] Daouia, A., Gardes, L., Girard, S. and Lekina, A. (2011). Kernel estimators of extreme level curves. *TEST* **20** 311–333.
[8] Daouia, A., Gijbels, I. and Stupfler, G. (2019). Extremiles: A new perspective on asymmetric least squares. *Journal of the American Statistical Association* **114** 1366–1381.
[9] Daouia, A., Girard, S. and Stupfler, G. (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B* **80** 263–292.
[10] Daouia, A., Girard, S. and Stupfler, G. (2019). Extreme M-quantiles as risk measures: From L^1 to L^p optimization. *Bernoulli* **25** 264–309.
[11] Daouia, A., Girard, S. and Stupfler, G. (2020). Tail expectile process and risk assessment. *Bernoulli* **26** 531–556.
[12] Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B* **52** 393–425.
[13] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction.* Springer.
[14] Dekkers, A. L. M., Einmahl, J. H. J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics* **17** 1833–1855.
[15] Drees, H. (2003). Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* **9** 617–657.
[16] Francq, C. and Zakoïan, J. M. (2010). *GARCH Models: Structure, Statistical Inference and Financial Applications.* John Wiley & Sons.
[17] Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes* **21** 57–95.

[18] Gardes, L. and Stupfler, G. (2014). Estimation of the conditional tail index using a smoothed local Hill estimator. *Extremes* **17** 45–75.
[19] Gomes, M. I., Brilhante, M. F. and Pestana, D. (2016). New reduced-bias estimators of a positive extreme value index. *Communications in Statistics - Simulation and Computation* **45** 833–862.
[20] Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748.
[21] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* **84** 986–995.
[22] Harrison, B., Smith, C. and Davies, B. (1992). *Introductory Economics*. Macmillan Press.
[23] He, F., Cheng, Y. and Tong, T. (2016). Estimation of high conditional quantiles using the Hill estimator of the tail index. *Journal of Statistical Planning and Inference* **176** 64–77.
[24] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3** 1163–1174.
[25] Hill, J. B. (2015). Tail index estimation for a filtered dependent time series. *Statistica Sinica* **25** 609–629.
[26] Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes.
[27] Holzmann, H. and Klar, B. (2016). Expectile asymptotics. *Electronic Journal of Statistics* **10** 2355–2371.
[28] Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer.
[29] Horváth, L. and Liese, F. (2004). L_p-estimators in ARCH models. *Journal of Statistical Planning and Inference* **119** 277–309.
[30] Knight, K. (1999). Epi-convergence in distribution and stochastic equi-semicontinuity Technical Report, University of Toronto.
[31] Krätschmer, V. and Zähle, H. (2017). Statistical inference for expectile-based risk measures. *Scandinavian Journal of Statistics* **44** 425–454.
[32] Kuan, C. M., Yeh, J. H. and Hsu, Y. C. (2009). Assessing value at risk with CARE, the Conditional Autoregressive Expectile models. *Journal of Econometrics* **150** 261–270.
[33] Kuchibhotla, A. K. and Patra, R. K. (2020). Efficient estimation in single index models through smoothing splines. *Bernoulli* **26** 1587–1618.
[34] Kyung-Joon, C. and Schucany, W. R. (1998). Nonparametric kernel regression estimation near endpoints. *Journal of Statistical Planning and Inference* **66** 289–304.
[35] Lewbel, A. and Linton, O. (2002). Nonparametric censored and truncated regression. *Econometrica* **70** 765–779.
[36] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics* **21** 255–285.
[37] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis* **17** 571–599.
[38] McKenzie, M. D. (1997). ARCH modelling of Australian bilateral exchange rate data. *Applied Financial Economics* **7** 147–164.
[39] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847.
[40] Newey, W. K. and Powell, J. L. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* **6** 295–317.
[41] Powell, J. L. (1986). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* **54** 1435–1460.
[42] Rohrbeck, C., Eastoe, E. F., Frigessi, A. and Tawn, J. A. (2018). Extreme value modelling of water-related insurance claims. *Annals of Applied Statistics* **12** 246–282.
[43] Segers, J. (2001). Residual estimators. *Journal of Statistical Planning and Inference* **98** 15–27.
[44] Stupfler, G. (2019). On a relationship between randomly and non-randomly thresholded empirical average excesses for heavy tails. *Extremes* **22** 749–769.
[45] Stute, W. and Wang, J. L. (2008). The central limit theorem under random truncation.

Bernoulli **14** 604–622.

[46] Taylor, J. W. (2008). Estimating Value at Risk and Expected Shortfall using expectiles. *Journal of Financial Econometrics* **6** 231–252.

[47] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36.

[48] Tsay, R. S. (1984). Regression models with time Series errors. *Journal of the American Statistical Association* **79** 118–124.

[49] Van Keilegom, I. and Wang, L. (2010). Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics* **4** 133–160.

[50] Velthoen, J., Cai, J. J., Jongbloed, G. and Schmeits, M. (2019). Improving precipitation forecasts using extreme quantile regression. *Extremes* **22** 599–622.

[51] Wang, H. J. and Li, D. (2013). Estimation of extreme conditional quantiles through power transformation. *Journal of American Statistical Association* **108** 1062–1074.

[52] Wang, H. J., Li, D. and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of American Statistical Association* **107** 1453–1464.

[53] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association* **73** 812–815.

[54] Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics* **13** 163–177.

[55] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statistics* **14** 1261–1350.

[56] Zelterman, D. (1993). A semiparametric bootstrap technique for simulating extreme order statistics. *Journal of the American Statistical Association* **88** 477–485.

[57] Zhu, K. and Ling, S. (2011). Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMAGARCH/IGARCH models. *Annals of Statistics* **39** 2131–2163.

[58] Zhu, L., Dong, Y. and Li, R. (2013). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statistica Sinica* **23** 1215–1235.

[59] Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance* **26** 901–918.

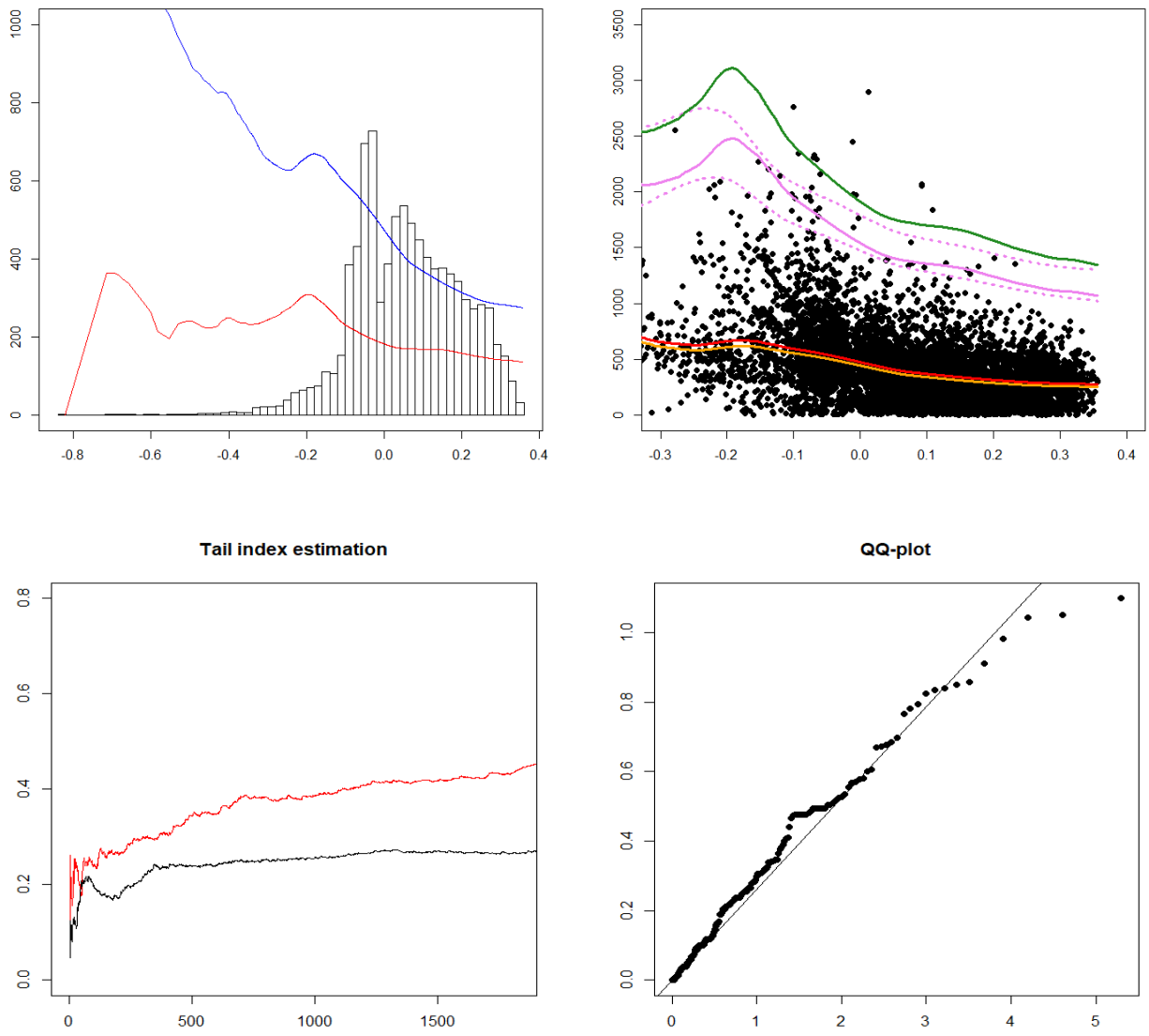

FIG 1. *Vehicle Insurance Customer data. Top left: estimates of g (red curve) and σ (blue curve) with a histogram of the $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i$. Top right: estimates of the regression mean (red line) and median (orange line) and of the estimated conditional expectile (solid purple line; dotted lines represent bootstrap pointwise 95% confidence intervals) and quantile (green line) at level $\tau_n' = 1 - 1/(nh^*) \approx 0.999$ in the $(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}, y)$ plane. Bottom left: curves $k \mapsto \widehat{\gamma}_k^{\mathrm{RB}}$ on the non-filtered data Y_i (black curve) and residuals (red curve). Bottom right: Exponential QQ-plot of the log-spacings $\log(\widehat{\varepsilon}_{n-i+1,n}^{(n)}/\widehat{\varepsilon}_{n-k^*,n}^{(n)})$, $1 \le i \le k^* = 200$. The straight line has slope $\widehat{\gamma}_{k^*}^{\mathrm{RB}} = 0.263$.*

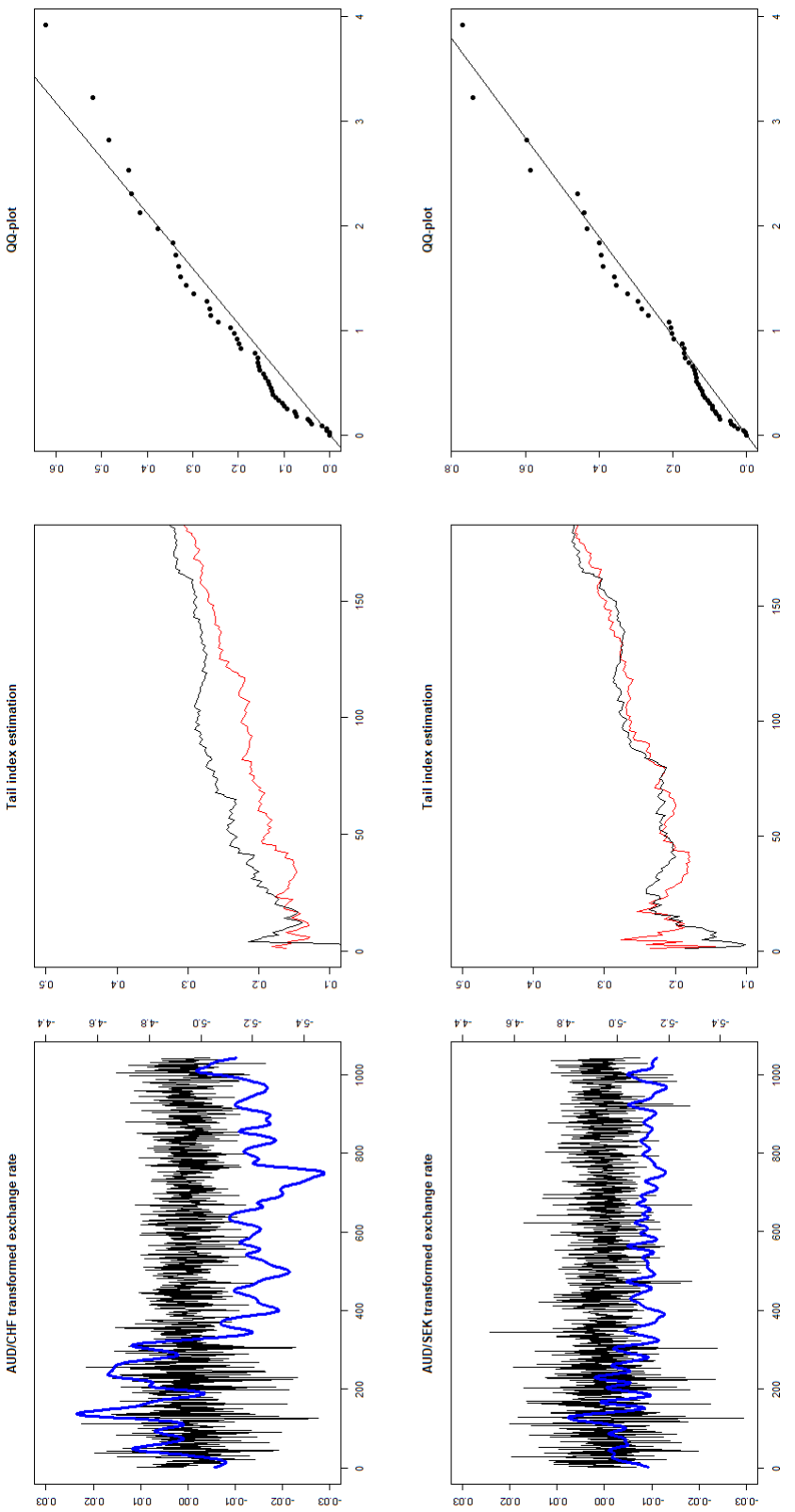

FIG 2. *Australian Dollar exchange rate data. Left panel: daily log-returns (black line) and estimated daily volatility (blue bold line, smoothed using the R* `smooth.spline` *function with parameter* 0.35*) from 1st March 2015 to 28th February 2019. Middle panel: Bias-reduced Hill curves* $k \mapsto \widehat{\gamma}_k^{\mathrm{RB}}$ *on the non-filtered data* Y_i *(black curve) and residuals (red curve). Right panel: Exponential QQ-plot of the log-spacings* $\log(\widehat{\varepsilon}_{n-i+1,n}^{(n)}/\widehat{\varepsilon}_{n-k^*,n}^{(n)})$*,* $1 \le i \le k^* = 50$*. The straight line has slope* $\widehat{\gamma}_{k^*}^{\mathrm{RB}}$*. Top panels: AUD/CHF data, bottom panels: AUD/SEK data.*

Supplementary Material

This supplementary material document contains the proofs of all theoretical results in the main paper, preceded by auxiliary results and their proofs (Sections A and B for the main results, and Sections C and D for the worked-out examples). It also provides further theoretical results related to indirect estimators in Section E, and further details about our finite-sample procedures and studies in Section F.

Appendix A: Theoretical toolbox: Auxiliary results and their proofs

Lemma A.1 below is a result on the mean excess function of a sample of heavy-tailed random variables, used in the proof of Theorem 2.1.

Lemma A.1. *Assume that ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0 < \gamma < 1/2$ and $\tau_n \uparrow 1$ is such that $n(1-\tau_n) \to \infty$. Let moreover $t_n \to \infty$ be a nonrandom sequence such that $\overline{F}(t_n)/(1-\tau_n) \to c \in (0,\infty)$. Then*

$$\frac{1}{nt_n(1-\tau_n)} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}\{\varepsilon_i > t_n\} \xrightarrow{\mathbb{P}} \frac{c}{1-\gamma}.$$

Proof. Write first

$$\frac{1}{nt_n(1-\tau_n)} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}\{\varepsilon_i > t_n\} = \frac{c+\mathrm{o}(1)}{nt_n\overline{F}(t_n)} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}\{\varepsilon_i > t_n\}.$$

The idea is now to split the sum on the right-hand side as follows:

$$\frac{1}{nt_n\overline{F}(t_n)} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}\{\varepsilon_i > t_n\} = \frac{1}{n\overline{F}(t_n)} \sum_{i=1}^{n} \mathbb{1}\{\varepsilon_i > t_n\} + \frac{1}{nt_n\overline{F}(t_n)} \sum_{i=1}^{n} (\varepsilon_i - t_n) \mathbb{1}\{\varepsilon_i > t_n\}.$$

Straightforward expectation and variance calculations yield

$$\mathbb{E}\left(\frac{1}{n\overline{F}(t_n)} \sum_{i=1}^{n} \mathbb{1}\{\varepsilon_i > t_n\}\right) = 1,$$

$$\mathrm{Var}\left(\frac{1}{n\overline{F}(t_n)} \sum_{i=1}^{n} \mathbb{1}\{\varepsilon_i > t_n\}\right) = \mathrm{O}\left(\frac{1}{n\overline{F}(t_n)}\right) = \mathrm{O}\left(\frac{1}{n(1-\tau_n)}\right) \to 0,$$

$$\mathbb{E}\left(\frac{1}{nt_n\overline{F}(t_n)} \sum_{i=1}^{n} (\varepsilon_i - t_n) \mathbb{1}\{\varepsilon_i > t_n\}\right) = \frac{1}{t_n} \int_{t_n}^{\infty} \frac{\overline{F}(x)}{\overline{F}(t_n)} dx \to \frac{\gamma}{1-\gamma},$$

$$\text{and } \mathrm{Var}\left(\frac{1}{nt_n\overline{F}(t_n)} \sum_{i=1}^{n} (\varepsilon_i - t_n) \mathbb{1}\{\varepsilon_i > t_n\}\right) = \mathrm{O}\left(\frac{1}{n\overline{F}(t_n)}\right) = \mathrm{O}\left(\frac{1}{n(1-\tau_n)}\right) \to 0.$$

Therefore

$$\frac{1}{nt_n(1-\tau_n)} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}\{\varepsilon_i > t_n\} \xrightarrow{\mathbb{P}} c\left(1 + \frac{\gamma}{1-\gamma}\right) = \frac{c}{1-\gamma}$$

as announced. □

The next auxiliary result is an extension of Theorem 1 in [9]. It drops the assumption of an independent sequence and of an increasing underlying distribution function. We note that the bias term $b(\gamma,\rho)$ of our result below is simpler than the corresponding bias term of Theorem 1 in [9], due to the assumption of a centred noise variable.

Proposition A.1. *Assume that $\mathbb{E}|\varepsilon_-| < \infty$, that condition $\mathcal{C}_2(\gamma,\rho,A)$ holds with $0 < \gamma < 1$, and that $\mathbb{E}(\varepsilon) = 0$. Let $\tau_n \uparrow 1$ be such that $n(1-\tau_n) \to \infty$, $\sqrt{n(1-\tau_n)}A((1-\tau_n)^{-1}) \to \lambda \in \mathbb{R}$ and $\sqrt{n(1-\tau_n)}/q_{\tau_n}(\varepsilon) = \mathrm{O}(1)$. Then, if*

$$\sqrt{n(1-\tau_n)}\left(\overline{\gamma} - \gamma, \frac{\overline{q}_{\tau_n}(\varepsilon)}{q_{\tau_n}(\varepsilon)} - 1\right) \xrightarrow{d} (\Gamma, \Theta),$$

we have

$$\sqrt{n(1-\tau_n)}\left(\frac{\widetilde{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)\xrightarrow{d} m(\gamma)\Gamma+\Theta-\lambda b(\gamma,\rho)$$

with $m(\gamma)=(1-\gamma)^{-1}-\log(\gamma^{-1}-1)$ *and*

$$b(\gamma,\rho)=\frac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho}+\frac{(\gamma^{-1}-1)^{-\rho}-1}{\rho}.$$

Proof. Note that $(\overline{\gamma}^{-1}-1)^{-\overline{\gamma}}\xrightarrow{\mathbb{P}}(\gamma^{-1}-1)^{-\gamma}$ and $\overline{q}_{\tau_n}(\varepsilon)/q_{\tau_n}(\varepsilon)-1\xrightarrow{\mathbb{P}}0$, so that linearising leads to

$$\begin{aligned}\frac{\widetilde{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1&=\left(\frac{(\overline{\gamma}^{-1}-1)^{-\overline{\gamma}}}{(\gamma^{-1}-1)^{-\gamma}}-1\right)+\left(\frac{\overline{q}_{\tau_n}(\varepsilon)}{q_{\tau_n}(\varepsilon)}-1\right)(1+\mathrm{o}_{\mathbb{P}}(1))\\&\quad+\left(\frac{(\gamma^{-1}-1)^{-\gamma}q_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)(1+\mathrm{o}_{\mathbb{P}}(1)).\end{aligned}\tag{6}$$

To control the bias term, use Proposition 1 in [11], of which a consequence is, for the centred variable ε,

$$\begin{aligned}\sqrt{n(1-\tau_n)}\left(\frac{(\gamma^{-1}-1)^{-\gamma}q_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)&=-\lambda\left[\frac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho}+\frac{(\gamma^{-1}-1)^{-\rho}-1}{\rho}\right]+\mathrm{o}(1)\\&=-\lambda b(\gamma,\rho)+\mathrm{o}(1).\end{aligned}$$

Reporting this in (6) and using the delta-method, we obtain

$$\sqrt{n(1-\tau_n)}\left(\frac{\widetilde{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)\xrightarrow{d} m(\gamma)\Gamma+\Theta-\lambda b(\gamma,\rho).$$

This is precisely the required result. □

The following rearrangement lemma is an extension of Lemma 1 in [18], which we use in the proof of Lemma A.3 below.

Lemma A.2. *Let $n\geq 2$ and $(a_1,\ldots,a_n)$ and $(b_1,\ldots,b_n)$ be two $n-$tuples of real numbers such that for all $i\in\{1,\ldots,n\}$, $a_i\leq b_i$. Then for all $i\in\{1,\ldots,n\}$, $a_{i,n}\leq b_{i,n}$.*

Proof. See the proof of Lemma 1 in [18], which, although the original result was stated for $n-$tuples featuring no ties, carries over to this more general case with no modification. □

The following lemma is the key to the proof of Theorem 2.2. In our context, its interpretation is that the gap between the tail empirical quantile process of the residuals and the analogue process based on the unobserved errors is bounded above by the gap between errors and their corresponding residuals; this will be used to give an approximation of the tail empirical quantile process of the errors by the tail empirical quantile process of the residuals.

Lemma A.3. *Let $k=k(n)\to\infty$ be a sequence of integers with $k/n\to 0$. Assume that ε has an infinite right endpoint. Suppose further that the ε_i are independent copies of ε and that the array of random variables $\widehat{\varepsilon}_i^{(n)}$, $1\leq i\leq n$, satisfies*

$$R_n:=\max_{1\leq i\leq n}\frac{|\widehat{\varepsilon}_i^{(n)}-\varepsilon_i|}{1+|\varepsilon_i|}\xrightarrow{\mathbb{P}}0.$$

Then we have both

$$\sup_{0<s\leq 1}\left|\frac{\widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}}{\varepsilon_{n-\lfloor ks\rfloor,n}}-1\right|=\mathrm{O}_{\mathbb{P}}(R_n)\quad\textit{and}\quad\sup_{0<s\leq 1}\left|\log\left(\frac{\widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}}{\varepsilon_{n-\lfloor ks\rfloor,n}}\right)\right|=\mathrm{O}_{\mathbb{P}}(R_n).$$

Proof. Clearly:

$$\forall i \in \{1,\dots,n\},\ \varepsilon_i - R_n(1+|\varepsilon_i|) =: \xi_i \le \widehat{\varepsilon}_i^{(n)} \le \zeta_i := \varepsilon_i + R_n(1+|\varepsilon_i|).$$

It then follows from Lemma A.2 that

$$\forall i \in \{1,\dots,n\},\ \xi_{i,n} \le \widehat{\varepsilon}_{i,n}^{(n)} \le \zeta_{i,n}.$$

Note that for any $r \in (-1,1)$, the function $x \mapsto x + r(1+|x|)$ is increasing. Therefore, on the event $\{R_n \le 1/4\}$, whose probability gets arbitrarily high as n increases, we have:

$$\forall i \in \{1,\dots,n\},\ \varepsilon_{i,n} - R_n(1+|\varepsilon_{i,n}|) = \xi_{i,n} \le \widehat{\varepsilon}_{i,n}^{(n)} \le \zeta_{i,n} = \varepsilon_{i,n} + R_n(1+|\varepsilon_{i,n}|).$$

Now, by Lemma 3.2.1 in [13] together with the equality $\varepsilon \overset{d}{=} U(Z)$ where Z has a unit Pareto distribution, we get $\varepsilon_{n-k,n} \xrightarrow{\mathbb{P}} +\infty$. On the event $A_n := \{R_n \le 1/4\} \cap \{\varepsilon_{n-k,n} \ge 1\}$, which likewise has probability arbitrarily large, we obtain

$$\forall i \ge n-k,\ (1-R_n)\varepsilon_{i,n} - R_n \le \widehat{\varepsilon}_{i,n}^{(n)} \le (1+R_n)\varepsilon_{i,n} + R_n.$$

In other words, on A_n, and for any $s \in (0,1]$,

$$-2R_n \le -R_n\left(1 + \frac{1}{\varepsilon_{n-\lfloor ks\rfloor,n}}\right) \le \frac{\widehat{\varepsilon}_{n-\lfloor ks\rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks\rfloor,n}} - 1 \le R_n\left(1 + \frac{1}{\varepsilon_{n-\lfloor ks\rfloor,n}}\right) \le 2R_n.$$

This shows that

$$\sup_{0<s\le 1}\left|\frac{\widehat{\varepsilon}_{n-\lfloor ks\rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks\rfloor,n}} - 1\right| = \mathrm{O}_{\mathbb{P}}(R_n).$$

Note further that, on A_n,

$$\forall s \in (0,1],\ \log(1-2R_n) \le \log\left(\frac{\widehat{\varepsilon}_{n-\lfloor ks\rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks\rfloor,n}}\right) \le \log(1+2R_n).$$

Since $\log(1+x) \le x$ and $\log(1-x) \ge -2x$ for all $x \in [0,1/2]$, this yields, on A_n,

$$\forall s \in (0,1],\ \left|\log\left(\frac{\widehat{\varepsilon}_{n-\lfloor ks\rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks\rfloor,n}}\right)\right| \le 4R_n.$$

As a consequence,

$$\sup_{0<s\le 1}\left|\log\left(\frac{\widehat{\varepsilon}_{n-\lfloor ks\rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks\rfloor,n}}\right)\right| = \mathrm{O}_{\mathbb{P}}(R_n).$$

This concludes the proof. □

The final auxiliary result of this section is used as part of Remark 2. It can be seen as a Breiman-type result, see [3] for the original Breiman lemma.

Lemma A.4. *Suppose that the random variable Y can be written $Y = Z_1 + Z_2\,\varepsilon$, where*

- *Z_1 is a bounded random variable,*
- *Z_2 is a (strictly) positive and bounded random variable,*
- *ε satisfies condition $\mathcal{C}_1(\gamma)$,*
- *Z_2 is independent of ε.*

Then Y satisfies condition $\mathcal{C}_1(\gamma)$.

Proof. We prove that for all $x > 0$, $\mathbb{P}(Y > tx)/\mathbb{P}(Y > t) \to x^{-1/\gamma}$ as $t \to \infty$. Note that if a_1, b_1 are such that $Z_1 \in [a_1, b_1]$ with probability 1,

$$\frac{\mathbb{P}(Z_2\,\varepsilon > tx - a_1)}{\mathbb{P}(Z_2\,\varepsilon > t - b_1)} \leq \frac{\mathbb{P}(Y > tx)}{\mathbb{P}(Y > t)} \leq \frac{\mathbb{P}(Z_2\,\varepsilon > tx - b_1)}{\mathbb{P}(Z_2\,\varepsilon > t - a_1)}.$$

This entails, for any fixed $\varepsilon \in (0,1)$, that for t large enough,

$$\frac{\mathbb{P}(Z_2\,\varepsilon > t(x+\varepsilon))}{\mathbb{P}(Z_2\,\varepsilon > t(1-\varepsilon))} \leq \frac{\mathbb{P}(Y > tx)}{\mathbb{P}(Y > t)} \leq \frac{\mathbb{P}(Z_2\,\varepsilon > t(x-\varepsilon))}{\mathbb{P}(Z_2\,\varepsilon > t(1+\varepsilon))}.$$

Let $b_2 > 0$ be such that $Z_2 \in (0, b_2]$ with probability 1. Since Z_2 is independent of ε, we have for any $t > 0$

$$\frac{\mathbb{P}(Z_2\,\varepsilon > t)}{\mathbb{P}(\varepsilon > t)} = \int_0^{b_2} \frac{\mathbb{P}(\varepsilon > t/z)}{\mathbb{P}(\varepsilon > t)} \mathbb{P}_{Z_2}(dz).$$

Use now Potter bounds (see *e.g.* Proposition B.1.9.5 in [13]) and the dominated convergence theorem to obtain

$$\frac{\mathbb{P}(Z_2\,\varepsilon > t)}{\mathbb{P}(\varepsilon > t)} \to \int_0^{b_2} z^\gamma \mathbb{P}_{Z_2}(dz) = \mathbb{E}(Z_2^\gamma) \in (0, \infty).$$

This implies that $Z_2\,\varepsilon$ is, like ε, heavy-tailed with extreme value index γ. In particular

$$\frac{\mathbb{P}(Z_2\,\varepsilon > t(x \mp \varepsilon))}{\mathbb{P}(Z_2\,\varepsilon > t(1 \pm \varepsilon))} = \frac{\mathbb{P}(Z_2\,\varepsilon > t(x \mp \varepsilon))}{\mathbb{P}(Z_2\,\varepsilon > t)} \frac{\mathbb{P}(Z_2\,\varepsilon > t)}{\mathbb{P}(Z_2\,\varepsilon > t(1 \pm \varepsilon))} \to (1 \pm \varepsilon)^{1/\gamma}(x \mp \varepsilon)^{-1/\gamma}$$

as $t \to \infty$. Conclude that

$$(1-\varepsilon)^{1/\gamma}(x+\varepsilon)^{-1/\gamma} \leq \liminf_{t\to\infty} \frac{\mathbb{P}(Y > tx)}{\mathbb{P}(Y > t)} \leq \limsup_{t\to\infty} \frac{\mathbb{P}(Y > tx)}{\mathbb{P}(Y > t)} \leq (1+\varepsilon)^{1/\gamma}(x-\varepsilon)^{-1/\gamma}$$

for any $\varepsilon > 0$, and let $\varepsilon \downarrow 0$ to complete the proof. □

Appendix B: Theoretical toolbox: Proofs of the main results

Proof of Theorem 2.1. Note that

$$\sqrt{n(1-\tau_n)}\left(\frac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)} - 1\right) = \arg\min_{u\in\mathbb{R}} \chi_n(u)$$

$$\text{with } \chi_n(u) := \frac{1}{2\xi^2_{\tau_n}(\varepsilon)} \sum_{i=1}^n \left[\eta_{\tau_n}\left(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon) - \frac{u\xi_{\tau_n}(\varepsilon)}{\sqrt{n(1-\tau_n)}}\right) - \eta_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon))\right].$$

Define

$$\psi_n(u) := \frac{1}{2\xi^2_{\tau_n}(\varepsilon)} \sum_{i=1}^n \left[\eta_{\tau_n}\left(\varepsilon_i - \xi_{\tau_n}(\varepsilon) - \frac{u\xi_{\tau_n}(\varepsilon)}{\sqrt{n(1-\tau_n)}}\right) - \eta_{\tau_n}(\varepsilon_i - \xi_{\tau_n}(\varepsilon))\right].$$

In other words, $\psi_n(u)$ is the counterpart of $\chi_n(u)$ based on the true, unobservable errors ε_i. Note that for any n, $u \mapsto \psi_n(u)$ is a continuously differentiable convex function. We shall prove that, pointwise in u, $\chi_n(u) - \psi_n(u) \xrightarrow{\mathbb{P}} 0$. The result will then be a straightforward consequence of a convexity lemma stated as Theorem 5 in [30] together with the convergence

$$\psi_n(u) \xrightarrow{d} -uZ\sqrt{\frac{2\gamma}{1-2\gamma}} + \frac{u^2}{2\gamma} \quad \text{as} \quad n \to \infty$$

(in the sense of finite-dimensional convergence, with Z being standard Gaussian) shown in the proof of Theorem 2 in [9].

We start by recalling that

$$\frac{1}{2}(\eta_\tau(x-y)-\eta_\tau(x)) = -\int_0^y \varphi_\tau(x-t)dt$$

where $\varphi_\tau(y) = |\tau - \mathbb{1}\{y \le 0\}|y$ (see Lemma 2 in [9]). Therefore

$$\begin{aligned}&\chi_n(u) - \psi_n(u)\\ &= -\frac{1}{\xi^2_{\tau_n}(\varepsilon)}\sum_{i=1}^n \int_0^{u\xi_{\tau_n}(\varepsilon)/\sqrt{n(1-\tau_n)}} [\varphi_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon) - t) - \varphi_{\tau_n}(\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t)]dt.\end{aligned}$$

Set $I_n(u) = [0, |u|\xi_{\tau_n}(\varepsilon)/\sqrt{n(1-\tau_n)}]$. Since

$$\begin{aligned}&|\chi_n(u) - \psi_n(u)|\\ &\le \frac{|u|}{\xi_{\tau_n}(\varepsilon)\sqrt{n(1-\tau_n)}}\sum_{i=1}^n \sup_{|t|\in I_n(u)} |\varphi_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon) - t) - \varphi_{\tau_n}(\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t)|,\end{aligned}$$

it is enough to show that

$$\begin{aligned}T_n(u) &:= \frac{1}{\xi_{\tau_n}(\varepsilon)\sqrt{n(1-\tau_n)}}\sum_{i=1}^n \sup_{|t|\in I_n(u)} |\varphi_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon) - t) - \varphi_{\tau_n}(\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t)|\\ &\xrightarrow{\mathbb{P}} 0.\end{aligned} \tag{7}$$

We now apply Lemma 3 in [9], which gives, for any $x, h \in \mathbb{R}$,

$$|\varphi_\tau(x-h) - \varphi_\tau(x)| \le |h|(1 - \tau + 2\mathbb{1}\{x > \min(h, 0)\}).$$

This translates into

$$\begin{aligned}&|\varphi_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon) - t) - \varphi_{\tau_n}(\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t)|\\ &\le |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|(1 - \tau_n + 2\mathbb{1}\{\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t > \min(\varepsilon_i - \widehat{\varepsilon}_i^{(n)}, 0)\}).\end{aligned}$$

Hence the inequality

$$T_n(u) \le T_{1,n} + T_{2,n}(u) \tag{8}$$

with

$$T_{1,n} := \frac{\sqrt{1-\tau_n}}{\xi_{\tau_n}(\varepsilon)\sqrt{n}}\sum_{i=1}^n |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i| \quad \text{and}$$

$$T_{2,n}(u) := \frac{2}{\xi_{\tau_n}(\varepsilon)\sqrt{n(1-\tau_n)}}\sum_{i=1}^n \sup_{|t|\in I_n(u)} |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|\mathbb{1}\{\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t > \min(\varepsilon_i - \widehat{\varepsilon}_i^{(n)}, 0)\}.$$

We first focus on $T_{1,n}$. Define $R_{n,i} := |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|/(1 + |\varepsilon_i|)$ and $R_n = \max_{1\le i\le n} R_{n,i}$. We have

$$T_{1,n} \le \left[\frac{\sqrt{n(1-\tau_n)}}{\xi_{\tau_n}(\varepsilon)}R_n\right] \times \frac{1}{n}\sum_{i=1}^n (1 + |\varepsilon_i|) = \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{n(1-\tau_n)}}{\xi_{\tau_n}(\varepsilon)}R_n\right)$$

by the law of large numbers. Note now that $\xi_{\tau_n}(\varepsilon) \to \infty$ and thus

$$T_{1,n} = \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{n(1-\tau_n)}}{q_{\tau_n}(\varepsilon)}R_n\right) = \mathrm{o}_{\mathbb{P}}\left(\sqrt{n(1-\tau_n)}R_n\right) \xrightarrow{\mathbb{P}} 0 \tag{9}$$

by assumption. We now turn to the control of $T_{2,n}(u)$, for which we write, for any t,

$$\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t > \min(\varepsilon_i - \widehat{\varepsilon}_i^{(n)}, 0) \Rightarrow \varepsilon_i - \xi_{\tau_n}(\varepsilon) - t > 0 \ \text{ or } \ \widehat{\varepsilon}_i^{(n)} - \xi_{\tau_n}(\varepsilon) - t > 0.$$

It follows that, for n large enough, we have, for any t such that $|t| \in I_n(u)$,

$$\varepsilon_i - \xi_{\tau_n}(\varepsilon) - t > \min(\varepsilon_i - \widehat{\varepsilon}_i^{(n)}, 0) \Rightarrow \varepsilon_i > \frac{\xi_{\tau_n}(\varepsilon)}{2} \text{ or } \widehat{\varepsilon}_i^{(n)} > \frac{\xi_{\tau_n}(\varepsilon)}{2}. \tag{10}$$

Now, for n large enough and with arbitrarily large probability as $n \to \infty$, $|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i| \leq (1 + |\varepsilon_i|)/2$ for any $i \in \{1, \ldots, n\}$, so that after some algebra,

$$\widehat{\varepsilon}_i^{(n)} > \frac{\xi_{\tau_n}(\varepsilon)}{2} \Rightarrow \varepsilon_i + \frac{1}{2}|\varepsilon_i| > \frac{1}{2}(\xi_{\tau_n}(\varepsilon) - 1) \Rightarrow \varepsilon_i + \frac{1}{2}|\varepsilon_i| > \frac{1}{4}\xi_{\tau_n}(\varepsilon)$$

because $\xi_{\tau_n}(\varepsilon) \to \infty$. Since the quantity $x + |x|/2$ can only be positive if $x > 0$, it follows that, with arbitrarily large probability,

$$\widehat{\varepsilon}_i^{(n)} > \frac{\xi_{\tau_n}(\varepsilon)}{2} \Rightarrow \varepsilon_i > \frac{1}{6}\xi_{\tau_n}(\varepsilon). \tag{11}$$

Combining (10) and (11) results in the following bound, valid with arbitrarily large probability as $n \to \infty$:

$$T_{2,n}(u) \leq \frac{2}{\xi_{\tau_n}(\varepsilon)\sqrt{n(1-\tau_n)}} \sum_{i=1}^{n} |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i| \mathbb{1}\left\{\varepsilon_i > \frac{1}{6}\xi_{\tau_n}(\varepsilon)\right\}.$$

By assumption on $|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|$, this leads to

$$T_{2,n}(u) \leq 4\left[\frac{\sqrt{n(1-\tau_n)}}{\xi_{\tau_n}(\varepsilon)} R_n\right] \times \frac{1}{n(1-\tau_n)} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}\left\{\varepsilon_i > \frac{1}{6}\xi_{\tau_n}(\varepsilon)\right\}.$$

Finally, the regular variation property of $\overline{F}$ and the asymptotic proportionality relationship between $\xi_{\tau_n}(\varepsilon)$ and $q_{\tau_n}(\varepsilon)$ ensure that

$$\lim_{n\to\infty} \frac{\overline{F}\left(\xi_{\tau_n}(\varepsilon)/6\right)}{1-\tau_n} \text{ exists, is positive and finite.}$$

Lemma A.1 then entails

$$T_{2,n}(u) = \mathrm{O}_{\mathbb{P}}\left(\sqrt{n(1-\tau_n)} R_n\right) \xrightarrow{\mathbb{P}} 0 \tag{12}$$

by assumption. Combining (7), (8), (9) and (12) completes the proof. □

Proof of Theorem 2.2. To prove the first expansion, write

$$\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{q_{1-k/n}(\varepsilon)} - s^{-\gamma} = \frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks \rfloor,n}} \left(\frac{\varepsilon_{n-\lfloor ks \rfloor,n}}{q_{1-k/n}(\varepsilon)} - s^{-\gamma}\right) + s^{-\gamma}\left(\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks \rfloor,n}} - 1\right).$$

Use Lemma A.3 and Theorem 2.4.8 in [13] to get

$$\begin{aligned}
&\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks \rfloor,n}} \left(\frac{\varepsilon_{n-\lfloor ks \rfloor,n}}{q_{1-k/n}(\varepsilon)} - s^{-\gamma}\right) \\
&= \frac{1}{\sqrt{k}}\left[\gamma s^{-\gamma-1} W_n(s) + \sqrt{k} A(n/k) s^{-\gamma} \frac{s^{-\rho}-1}{\rho} + s^{-\gamma-1/2-\delta}\, \mathrm{o}_{\mathbb{P}}(1)\right]
\end{aligned} \tag{13}$$

uniformly in $s \in (0, 1]$. Applying Lemma A.3 again gives

$$s^{-\gamma}\left|\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks \rfloor,n}} - 1\right| \leq s^{-\gamma-1/2-\delta}\left|\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks \rfloor,n}} - 1\right| = \frac{s^{-\gamma-1/2-\delta}}{\sqrt{k}}\, \mathrm{o}_{\mathbb{P}}(1) \tag{14}$$

uniformly in $s \in (0, 1]$. Combine (13) and (14) to complete the proof of the first expansion. The proof of the second expansion is based on the equality

$$\log\left(\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{q_{1-k/n}(\varepsilon)}\right) = \log\left(\frac{\varepsilon_{n-\lfloor ks \rfloor,n}}{q_{1-k/n}(\varepsilon)}\right) + \log\left(\frac{\widehat{\varepsilon}_{n-\lfloor ks \rfloor,n}^{(n)}}{\varepsilon_{n-\lfloor ks \rfloor,n}}\right)$$

and follows exactly the same ideas. □

Proof of Corollary 2.1. Notice that, by Theorem 2.2, there is a sequence W_n of standard Brownian motions such that, for any $\delta > 0$ sufficiently small:

$$\begin{aligned}\widehat{\gamma}_k &= \int_0^1 \log\left(\frac{\widehat{\varepsilon}^{(n)}_{n-\lfloor ks\rfloor,n}}{\widehat{\varepsilon}^{(n)}_{n-k,n}}\right) ds \\ &= \int_0^1 \left\{\gamma \log\frac{1}{s} + \frac{\gamma}{\sqrt{k}}\left[s^{-1}W_n(s) - W_n(1)\right] + A\left(\frac{n}{k}\right)\left[\frac{s^{-\rho}-1}{\rho} + s^{-1/2-\delta}\,\mathrm{o}_{\mathbb{P}}(1)\right]\right\} ds.\end{aligned}$$

We then obtain that $\widehat{\gamma}_k$ can be written

$$\sqrt{k}(\widehat{\gamma}_k - \gamma) = \frac{\lambda}{1-\rho} + \gamma\int_0^1 \left[s^{-1}W_n(s) - W_n(1)\right] + \mathrm{o}_{\mathbb{P}}(1).$$

Similarly,

$$\sqrt{k}\left(\frac{\widehat{\overline{\varepsilon}}^{(n)}_{n-k,n}}{q_{1-k/n}(\varepsilon)} - 1\right) = \gamma W_n(1) + \mathrm{o}_{\mathbb{P}}(1).$$

Noting that the Gaussian terms in these two asymptotic expansions are independent completes the proof. □

Proof of Theorem 2.3. The key is to note that

$$\begin{aligned}\frac{\overline{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x})}{\xi_{\tau_n'}(Y|\boldsymbol{x})} - 1 &= \left(1 + \frac{g(\boldsymbol{x})}{\sigma(\boldsymbol{x})\xi_{\tau_n'}(\varepsilon)}\right)^{-1}\left(\frac{\overline{\xi}^{\star}_{\tau_n'}(\varepsilon)}{\xi_{\tau_n'}(\varepsilon)} - 1\right) \\ &\quad + \frac{\overline{g}(\boldsymbol{x}) - g(\boldsymbol{x})}{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_n'}(\varepsilon)} + \left(1 + \frac{g(\boldsymbol{x})}{\sigma(\boldsymbol{x})\xi_{\tau_n'}(\varepsilon)}\right)^{-1}\frac{\overline{\sigma}(\boldsymbol{x}) - \sigma(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\frac{\overline{\xi}^{\star}_{\tau_n'}(\varepsilon)}{\xi_{\tau_n'}(\varepsilon)}.\end{aligned}$$

Using the convergence $\xi_\tau(\varepsilon)/q_\tau(\varepsilon) \to (\gamma^{-1}-1)^{-\gamma}$ as $\tau \uparrow 1$ and the heavy-tailed condition, we find $1/\xi_{\tau_n'}(\varepsilon) = \mathrm{o}(1/\xi_{\tau_n}(\varepsilon)) = \mathrm{o}(1/q_{\tau_n}(\varepsilon))$. Our assumptions show that this is a $\mathrm{o}(1/\sqrt{n(1-\tau_n)})$ and therefore

$$\begin{aligned}&\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]}\left(\frac{\overline{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x})}{\xi_{\tau_n'}(Y|\boldsymbol{x})} - 1\right) \\ &= \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]}\left(\frac{\overline{\xi}^{\star}_{\tau_n'}(\varepsilon)}{\xi_{\tau_n'}(\varepsilon)} - 1\right)(1 + \mathrm{o}_{\mathbb{P}}(1)) + \mathrm{o}_{\mathbb{P}}(1).\end{aligned}$$

Our result is then shown by adapting the proof of Theorem 5 of [11], with the condition $\rho < 0$ being used exclusively to control the bias term appearing naturally because of the extrapolation procedure applied to the heavy-tailed random variable ε. We omit the details. □

Appendix C: Worked-out examples: Auxiliary results and their proofs

Lemma C.1 gives the rate of convergence of the weighted least squares estimators in model (M_1). Here and throughout all $\mathrm{O}_{\mathbb{P}}(1)$ statements are meant componentwise.

Lemma C.1. *Assume that $(\boldsymbol{X}_i, Y_i)_{i\geq 1}$ are independent random pairs generated from model (M_1). Suppose further that $\mathbb{E}(\varepsilon^2) < \infty$. Then we have*

$$\sqrt{n}(\widehat{\alpha} - \alpha) = \mathrm{O}_{\mathbb{P}}(1),\ \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathrm{O}_{\mathbb{P}}(1) \quad \textit{and} \quad \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathrm{O}_{\mathbb{P}}(1).$$

Proof. We introduce the notation

$$\mathfrak{X} = \begin{pmatrix} 1 & \boldsymbol{X}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_n^\top \end{pmatrix},\ \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ and } \boldsymbol{\Omega} = \mathrm{diag}([1 + \boldsymbol{\theta}^\top \boldsymbol{X}_1]^2, \ldots, [1 + \boldsymbol{\theta}^\top \boldsymbol{X}_n]^2).$$

A preliminary step is to remark that for any $\boldsymbol{a} = (a_0, a_1, \ldots, a_d)^\top \in \mathbb{R}^{d+1}$,

$$\boldsymbol{a}^\top \mathfrak{X}^\top \mathfrak{X} \boldsymbol{a} = \sum_{i=1}^{n} [a_0 + (a_1, \ldots, a_d)\boldsymbol{X}_i]^2 > 0$$

$$\text{and } \boldsymbol{a}^\top \mathfrak{X}^\top \boldsymbol{\Omega}^{-1} \mathfrak{X} \boldsymbol{a} = \sum_{i=1}^{n} \left[1 + \boldsymbol{\theta}^\top \boldsymbol{X}_i\right]^{-2} [a_0 + (a_1, \ldots, a_d)\boldsymbol{X}_i]^2 > 0$$

with probability 1, because $\boldsymbol{X}$ has a continuous distribution (and as such, does not put mass on affine hyperplanes of $\mathbb{R}^d$). The symmetric matrices $\mathfrak{X}^\top \mathfrak{X}$ and $\mathfrak{X}^\top \boldsymbol{\Omega}^{-1} \mathfrak{X}$ therefore have full rank with probability 1. Since, by the law of large numbers,

$$\frac{1}{n}\left[\mathfrak{X}^\top \mathfrak{X}\right]_{i+1,j+1} \xrightarrow{\mathbb{P}} \mathbb{E}\left(X_i X_j\right) \text{ and } \frac{1}{n}\left[\mathfrak{X}^\top \boldsymbol{\Omega}^{-1} \mathfrak{X}\right]_{i+1,j+1} \xrightarrow{\mathbb{P}} \mathbb{E}\left(\left[1 + \boldsymbol{\theta}^\top \boldsymbol{X}\right]^{-2} X_i X_j\right)$$

(where $X_0 = 1$ for notational convenience), the same argument shows that $\mathfrak{X}^\top \mathfrak{X}/n$ and $\mathfrak{X}^\top \boldsymbol{\Omega}^{-1} \mathfrak{X}/n$ converge in probability to symmetric positive definite matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ say.

Our first step is to show that the preliminary estimators $\widetilde{\alpha}$, $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\theta}}$ are $\sqrt{n}-$consistent. Rewrite model (M_1) for the available data as

$$\boldsymbol{Y} = \mathfrak{X} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \left[\mathfrak{X} \begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right] \circ \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}^\top = (\varepsilon_1, \ldots, \varepsilon_n)$ and $\circ$ denotes the Hadamard (entrywise) product of matrices. By standard least squares theory,

$$\begin{pmatrix} \widetilde{\alpha} \\ \widetilde{\boldsymbol{\beta}} \end{pmatrix} = \left(\mathfrak{X}^\top \mathfrak{X}\right)^{-1} \mathfrak{X}^\top \boldsymbol{Y}.$$

A direct calculation then yields

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\widetilde{\alpha} - \alpha) \\ \sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \end{pmatrix} &= n\left(\mathfrak{X}^\top \mathfrak{X}\right)^{-1} \times \frac{1}{\sqrt{n}} \mathfrak{X}^\top \left\{\left[\mathfrak{X} \begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right] \circ \boldsymbol{\varepsilon}\right\} \\ &= n\left(\mathfrak{X}^\top \mathfrak{X}\right)^{-1} \times \begin{pmatrix} n^{-1/2} \sum_{i=1}^{n} \left[1 + \boldsymbol{\theta}^\top \boldsymbol{X}_i\right] \varepsilon_i \\ n^{-1/2} \sum_{i=1}^{n} \left[1 + \boldsymbol{\theta}^\top \boldsymbol{X}_i\right] X_{i1} \varepsilon_i \\ \vdots \\ n^{-1/2} \sum_{i=1}^{n} \left[1 + \boldsymbol{\theta}^\top \boldsymbol{X}_i\right] X_{id} \varepsilon_i \end{pmatrix}. \end{aligned}$$

Set for notational convenience $X_{i0} = 1$. Since, for any $m \in \{0, 1 \ldots, d\}$, the random variables $\left[1 + \boldsymbol{\theta}^\top \boldsymbol{X}_i\right] X_{im} \varepsilon_i$, $1 \leq i \leq n$, are independent, centred and square-integrable, the standard multivariate central limit theorem combined with the convergence $n\left(\mathfrak{X}^\top \mathfrak{X}\right)^{-1} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}_1^{-1}$ yields

$$\sqrt{n}(\widetilde{\alpha} - \alpha) = \mathrm{O}_{\mathbb{P}}(1) \text{ and } \sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathrm{O}_{\mathbb{P}}(1). \tag{15}$$

We then prove that $\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathrm{O}_{\mathbb{P}}(1)$. Recalling that

$$\widetilde{\boldsymbol{\theta}} = \frac{\widetilde{\boldsymbol{\nu}}}{\widetilde{\mu}} \text{ and } \boldsymbol{\theta} = \frac{\boldsymbol{\nu}}{\mu}$$

where $\mu = \mathbb{E}|\varepsilon| > 0$ and $\boldsymbol{\nu} = \mu\boldsymbol{\theta}$, it is enough to show that $\sqrt{n}(\widetilde{\mu} - \mu) = \mathrm{O}_{\mathbb{P}}(1)$ and $\sqrt{n}(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}) = \mathrm{O}_{\mathbb{P}}(1)$. Defining

$$\boldsymbol{Z} = \begin{pmatrix} |Y_1 - (\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_1)| \\ \vdots \\ |Y_n - (\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_n)| \end{pmatrix} = \mathfrak{X} \begin{pmatrix} \mu \\ \boldsymbol{\nu} \end{pmatrix} + \left[\mathfrak{X} \begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right] \circ \boldsymbol{e},$$

where $\boldsymbol{e}^\top = (|\varepsilon_1| - \mathbb{E}|\varepsilon|, \ldots, |\varepsilon_n| - \mathbb{E}|\varepsilon|)$, and defining then $\widetilde{\boldsymbol{Z}}$ in the obvious way, we have

$$\begin{pmatrix} \widetilde{\mu} \\ \widetilde{\boldsymbol{\nu}} \end{pmatrix} = \left(\mathfrak{X}^\top \mathfrak{X}\right)^{-1} \mathfrak{X}^\top \widetilde{\boldsymbol{Z}}.$$

We therefore obtain

$$\begin{pmatrix} \sqrt{n}(\widetilde{\mu}-\mu) \\ \sqrt{n}(\widetilde{\boldsymbol{\nu}}-\boldsymbol{\nu}) \end{pmatrix} = n\left(\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\mathfrak{X}}\right)^{-1} \times \frac{1}{\sqrt{n}}\boldsymbol{\mathfrak{X}}^\top\left\{\left[\boldsymbol{\mathfrak{X}}\begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right]\circ \boldsymbol{e}\right\} + n\left(\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\mathfrak{X}}\right)^{-1} \times \boldsymbol{\mathfrak{X}}^\top\left(\frac{1}{\sqrt{n}}\left[\widetilde{\boldsymbol{Z}}-\boldsymbol{Z}\right]\right). \tag{16}$$

Since $e = |\varepsilon| - \mathbb{E}|\varepsilon|$ is independent of $\boldsymbol{X}$ and has a finite variance, repeating the proof of (15) gives

$$n\left(\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\mathfrak{X}}\right)^{-1} \times \frac{1}{\sqrt{n}}\boldsymbol{\mathfrak{X}}^\top\left\{\left[\boldsymbol{\mathfrak{X}}\begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right]\circ \boldsymbol{e}\right\} = \mathrm{O}_{\mathbb{P}}(1). \tag{17}$$

Furthermore,

$$\boldsymbol{\mathfrak{X}}^\top\left(\frac{1}{\sqrt{n}}\left[\widetilde{\boldsymbol{Z}}-\boldsymbol{Z}\right]\right) = \begin{pmatrix} n^{-1/2}\sum_{i=1}^n \left[\widetilde{Z}_i - Z_i\right] \\ n^{-1/2}\sum_{i=1}^n X_{i1}\left[\widetilde{Z}_i - Z_i\right] \\ \vdots \\ n^{-1/2}\sum_{i=1}^n X_{id}\left[\widetilde{Z}_i - Z_i\right] \end{pmatrix}.$$

Recalling that $\boldsymbol{X}$ lies in a compact set, we find that for any $m \in \{0, 1, \ldots, d\}$,

$$\left|n^{-1/2}\sum_{i=1}^n X_{im}\left[\widetilde{Z}_i - Z_i\right]\right| = \mathrm{O}_{\mathbb{P}}\left(\sqrt{n}\max_{1\le i\le n}|(\widetilde{\alpha}-\alpha) + (\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}_i|\right) = \mathrm{O}_{\mathbb{P}}(1)$$

by (15). Combining this with (16), (17) and the convergence $n\left(\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\mathfrak{X}}\right)^{-1} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}_1^{-1}$, we get indeed $\sqrt{n}(\widetilde{\mu}-\mu) = \mathrm{O}_{\mathbb{P}}(1)$ and $\sqrt{n}(\widetilde{\boldsymbol{\nu}}-\boldsymbol{\nu}) = \mathrm{O}_{\mathbb{P}}(1)$ and thus

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}) = \mathrm{O}_{\mathbb{P}}(1). \tag{18}$$

We are now ready to prove the convergence of the weighted estimators $\widehat{\alpha}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$. By standard weighted least squares theory,

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} = \left(\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\mathfrak{X}}\right)^{-1}\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{Y}.$$

It follows that

$$\begin{pmatrix} \sqrt{n}(\widehat{\alpha}-\alpha) \\ \sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \end{pmatrix} = n\left(\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\mathfrak{X}}\right)^{-1} \times \frac{1}{\sqrt{n}}\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\left\{\left[\boldsymbol{\mathfrak{X}}\begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right]\circ \boldsymbol{\varepsilon}\right\} \tag{19}$$

where $\widetilde{\boldsymbol{\Omega}}$ is obtained from $\boldsymbol{\Omega}$ in the obvious manner. Note that for any $i, j \in \{0, \ldots, d\}$,

$$\frac{1}{n}\left[\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\mathfrak{X}}\right]_{i+1,j+1} = \frac{1}{n}\sum_{k=1}^n \frac{X_{ki}X_{kj}}{\left[1+\boldsymbol{\theta}^\top\boldsymbol{X}_k\right]^2} \text{ and}$$

$$\frac{1}{n}\left[\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\mathfrak{X}}\right]_{i+1,j+1} = \frac{1}{n}\sum_{k=1}^n \frac{X_{ki}X_{kj}}{\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_k\right]^2}.$$

Recalling once again that $\boldsymbol{X}$ lies in a compact set, that $1+\boldsymbol{\theta}^\top\boldsymbol{X}$ is bounded from below by a positive constant, and (18), we find, by the law of large numbers,

$$\frac{1}{n}\left[\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\mathfrak{X}}\right] - \frac{1}{n}\left[\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\mathfrak{X}}\right] \xrightarrow{\mathbb{P}} 0 \text{ and thus } n\left(\boldsymbol{\mathfrak{X}}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\mathfrak{X}}\right)^{-1} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}_2^{-1}. \tag{20}$$

Besides, for any $m \in \{0, 1, \ldots, d\}$,

$$\left[\frac{1}{\sqrt{n}}\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\} - \frac{1}{\sqrt{n}}\mathfrak{X}^\top\mathbf{\Omega}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\}\right]_{m+1}$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[1+\boldsymbol{\theta}^\top\boldsymbol{X}_i\right]X_{im}\varepsilon_i\left[\frac{1}{\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i\right]^2} - \frac{1}{\left[1+\boldsymbol{\theta}^\top\boldsymbol{X}_i\right]^2}\right]$$
$$= -\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^\top\left\{\frac{1}{n}\sum_{i=1}^{n}X_{im}\varepsilon_i\frac{2+\boldsymbol{\theta}^\top\boldsymbol{X}_i+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i}{\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i\right]^2\left[1+\boldsymbol{\theta}^\top\boldsymbol{X}_i\right]}\boldsymbol{X}_i\right\}.$$

Using again the properties of $\boldsymbol{X}$ and (18), some straightforward algebra yields that

$$R_n := \sqrt{n}\max_{1\le i\le n}\left|\frac{2+\boldsymbol{\theta}^\top\boldsymbol{X}_i+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i}{\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i\right]^2} - \frac{2}{1+\boldsymbol{\theta}^\top\boldsymbol{X}_i}\right| = \mathrm{O}_{\mathbb{P}}(1).$$

Conclude that

$$\left[\frac{1}{\sqrt{n}}\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\} - \frac{1}{\sqrt{n}}\mathfrak{X}^\top\mathbf{\Omega}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\}\right]_{m+1}$$
$$= -2\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^\top\left\{\frac{1}{n}\sum_{i=1}^{n}X_{im}\varepsilon_i\left[\frac{2}{[1+\boldsymbol{\theta}^\top\boldsymbol{X}_i]^2}+\mathrm{O}_{\mathbb{P}}\left(\frac{R_n}{\sqrt{n}}\right)\right]\boldsymbol{X}_i\right\}.$$

Since ε is centred and independent of $\boldsymbol{X}$, we may combine the properties of $\boldsymbol{X}$ and (18) with the law of large numbers to get

$$\frac{1}{\sqrt{n}}\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\} - \frac{1}{\sqrt{n}}\mathfrak{X}^\top\mathbf{\Omega}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\} = \mathrm{o}_{\mathbb{P}}(1). \tag{21}$$

Now clearly

$$\left[\frac{1}{\sqrt{n}}\mathfrak{X}^\top\mathbf{\Omega}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\}\right]_{m+1} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_{im}}{1+\boldsymbol{\theta}^\top\boldsymbol{X}_i}\varepsilon_i$$

so that, by the standard multivariate central limit theorem,

$$\frac{1}{\sqrt{n}}\mathfrak{X}^\top\mathbf{\Omega}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{\varepsilon}\right\} = \mathrm{O}_{\mathbb{P}}(1). \tag{22}$$

Combining (19), (20), (21) and (22) results in

$$\sqrt{n}(\widehat{\alpha}-\alpha) = \mathrm{O}_{\mathbb{P}}(1) \quad \text{and} \quad \sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \mathrm{O}_{\mathbb{P}}(1).$$

We complete the proof by showing that $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = \mathrm{O}_{\mathbb{P}}(1)$. It is again enough to show that $\sqrt{n}(\widehat{\mu}-\mu) = \mathrm{O}_{\mathbb{P}}(1)$ and $\sqrt{n}(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}) = \mathrm{O}_{\mathbb{P}}(1)$. Write

$$\begin{pmatrix}\sqrt{n}(\widehat{\mu}-\mu)\\ \sqrt{n}(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu})\end{pmatrix} = n\left(\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\mathfrak{X}\right)^{-1}\times\frac{1}{\sqrt{n}}\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\left\{\left[\mathfrak{X}\begin{pmatrix}1\\ \boldsymbol{\theta}\end{pmatrix}\right]\circ\boldsymbol{e}\right\}$$
$$+ n\left(\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\mathfrak{X}\right)^{-1}\times\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\left(\frac{1}{\sqrt{n}}\left[\widehat{\boldsymbol{Z}}-\boldsymbol{Z}\right]\right).$$

Furthermore,

$$\mathfrak{X}^\top\widetilde{\mathbf{\Omega}}^{-1}\left(\frac{1}{\sqrt{n}}\left[\widehat{\boldsymbol{Z}}-\boldsymbol{Z}\right]\right) = \begin{pmatrix} n^{-1/2}\sum_{i=1}^{n}\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i\right]^{-2}\left[\widehat{Z}_i-Z_i\right]\\ n^{-1/2}\sum_{i=1}^{n}X_{i1}\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i\right]^{-2}\left[\widehat{Z}_i-Z_i\right]\\ \vdots\\ n^{-1/2}\sum_{i=1}^{n}X_{id}\left[1+\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{X}_i\right]^{-2}\left[\widehat{Z}_i-Z_i\right]\end{pmatrix}.$$

Recalling the properties of $\boldsymbol{X}$ and the $\sqrt{n}-$convergence of $\widehat{\alpha}$, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\theta}}$, we find that for any $m \in \{0, 1, \ldots, d\}$,

$$\left| n^{-1/2} \sum_{i=1}^{n} X_{im} \left[1 + \widetilde{\boldsymbol{\theta}}^{\top} \boldsymbol{X}_i \right]^{-2} \left[\widehat{Z}_i - Z_i \right] \right| = \mathrm{O}_{\mathbb{P}} \left(\sqrt{n} \max_{1 \leq i \leq n} |(\widehat{\alpha} - \alpha) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\top} \boldsymbol{X}_i| \right) = \mathrm{O}_{\mathbb{P}}(1).$$

Combining this with (20) and straightforward adaptations of (21) and (22) with $\boldsymbol{e}$ in place of $\boldsymbol{\varepsilon}$, we find $\sqrt{n}(\widehat{\mu} - \mu) = \mathrm{O}_{\mathbb{P}}(1)$ and $\sqrt{n}(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}) = \mathrm{O}_{\mathbb{P}}(1)$ as required. □

Lemma C.2 is a general uniform consistency result which is useful for the analysis of the single-index model (M_2).

Lemma C.2. *Assume that $(\mathcal{X}_i, \mathcal{Y}_i)_{i \geq 1}$ are independent copies of a bivariate random pair $(\mathcal{X}, \mathcal{Y})$ such that:*

- *$\mathcal{X}$ has support $[a, b]$, with $a < b$, and a density function $f_{\mathcal{X}}$ which is uniformly bounded on compact sub-intervals of (a, b).*
- *There exists $\delta > 0$ such that $\mathbb{E}|\mathcal{Y}|^{2+\delta} < \infty$ and the conditional moment function $z \mapsto \mathbb{E}\left[|\mathcal{Y}|^{2+\delta} | \mathcal{X} = z\right]$ is uniformly bounded on compact sub-intervals of (a, b).*

Let further:

- *$(\mathcal{V}_i)$ be a sequence of independent copies of a bounded random variable $\mathcal{V}$.*
- *L be a Lipschitz continuous function with support contained in $[-1, 1]$.*

Assume finally that $nh_n^5 \to c \in (0, \infty)$, and $t_n = n^t$ with $2/(5+\delta) < t < 2/5$. Then for any $a_1, b_1 \in [a, b]$ with $a < a_1 < b_1 < b$,

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \leq z \leq b_1} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \mathcal{Y}_i \mathbb{1}\{|\mathcal{Y}_i| \leq t_n\} \mathcal{V}_i L\left(\frac{z - \mathcal{X}_i}{h_n} \right) - \frac{1}{h_n} \mathbb{E}\left[\mathcal{Y} \mathcal{V} L\left(\frac{z - \mathcal{X}}{h_n} \right) \right] \right| = \mathrm{O}_{\mathbb{P}}(1).$$

We note that, as a consequence, we have a similar uniform consistency result for the non-truncated version of the smoothed empirical moment, that is

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \leq z \leq b_1} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \mathcal{Y}_i \mathcal{V}_i L\left(\frac{z - \mathcal{X}_i}{h_n} \right) - \frac{1}{h_n} \mathbb{E}\left[\mathcal{Y} \mathcal{V} L\left(\frac{z - \mathcal{X}}{h_n} \right) \right] \right| = \mathrm{O}_{\mathbb{P}}(1)$$

under the further assumption $\mathbb{E}|\mathcal{Y}|^{5/2+\delta} < \infty$. This follows from noting that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} \{|\mathcal{Y}_i| > t_n\} \right) \leq n\mathbb{P}\left(|\mathcal{Y}| > t_n\right) = \mathrm{O}\left(\frac{n}{t_n^{5/2+\delta}} \right) = \mathrm{O}\left(n^{1-(5+2\delta)/(5+\delta)} \right) = \mathrm{o}(1)$$

by Markov's inequality. The stronger moment assumption $\mathbb{E}|\mathcal{Y}|^{5/2+\delta} < \infty$ already appears in [37] in the context of local polynomial estimation.

Proof. The basic idea is to control the oscillation of the random function

$$z \mapsto \frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \mathcal{Y}_i \mathbb{1}\{|\mathcal{Y}_i| \leq t_n\} \mathcal{V}_i L\left(\frac{z - \mathcal{X}_i}{h_n} \right) - \frac{1}{h_n} \mathbb{E}\left[\mathcal{Y} \mathcal{V} L\left(\frac{z - \mathcal{X}}{h_n} \right) \right] \right|$$

and then use this control to prove that it is sufficient to show uniform consistency over a fine grid instead, which can be done by using Bernstein's exponential inequality. Our proof adapts the method of [20] (proof of Theorem 2).

Define $\mathrm{Y}_i^{(n)} := \mathcal{Y}_i\,\mathcal{V}_i \mathbb{1}\{|\mathcal{Y}_i| \le t_n\}$ and $\mathrm{Y}^{(n)} := \mathcal{Y}\,\mathcal{V} \mathbb{1}\{|\mathcal{Y}| \le t_n\}$. Then

$$\begin{aligned}
&\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \le z \le b_1} \left| \frac{1}{nh_n} \sum_{i=1}^n \mathcal{Y}_i \mathbb{1}\{|\mathcal{Y}_i| \le t_n\}\, \mathcal{V}_i\, L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathcal{Y}\,\mathcal{V}\, L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right| \\
&\le \frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \le z \le b_1} \left| \frac{1}{nh_n} \sum_{i=1}^n \mathrm{Y}_i^{(n)}\, L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)}\, L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right| \\
&+ \frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \le z \le b_1} \frac{1}{h_n}\mathbb{E}\left[|\mathcal{Y}|\,|\mathcal{V}| \mathbb{1}\{|\mathcal{Y}| > t_n\} \left| L\left(\frac{z-\mathcal{X}}{h_n}\right)\right|\right]. \qquad (23)
\end{aligned}$$

The second term on the right-hand side of (23) is controlled by noting that, thanks to a change of variables,

$$\begin{aligned}
&\frac{1}{h_n}\mathbb{E}\left[|\mathcal{Y}|\,|\mathcal{V}| \mathbb{1}\{|\mathcal{Y}| > t_n\} \left| L\left(\frac{z-\mathcal{X}}{h_n}\right)\right|\right] \\
&= \mathrm{O}\left(\int_{-1}^1 \mathbb{E}\left[|\mathcal{Y}| \mathbb{1}\{|\mathcal{Y}| > t_n\} | \mathcal{X} = z - h_n u\right] |L(u)|\, f_{\mathcal{X}}(z - h_n u)\, du \right) \\
&= \mathrm{O}\left(t_n^{-1-\delta} \int_{-1}^1 \mathbb{E}\left[|\mathcal{Y}|^{2+\delta} | \mathcal{X} = z - h_n u\right] |L(u)|\, f_{\mathcal{X}}(z - h_n u)\, du \right) = \mathrm{O}(t_n^{-1-\delta})
\end{aligned}$$

uniformly in $z \in [a_1, b_1]$. Here the boundedness of $\mathcal{V}$, the integrability of $|L|$ and the assumption that the $(2+\delta)-$conditional moment of $\mathcal{Y}$ and the density function $f_{\mathcal{X}}$ are uniformly bounded on compact sub-intervals of (a, b) were all used. Finally

$$t_n^{-1-\delta} = n^{-(1+\delta)t} = \mathrm{o}(n^{-2/5}) = \mathrm{o}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)$$

so that

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \le z \le b_1} \frac{1}{h_n}\mathbb{E}\left[|\mathcal{Y}|\,|\mathcal{V}| \mathbb{1}\{|\mathcal{Y}| > t_n\} \left| L\left(\frac{z-\mathcal{X}}{h_n}\right)\right|\right] = \mathrm{o}(1). \qquad (24)$$

Combining (23) and (24), we find that it is sufficient to show that

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \le z \le b_1} \left| \frac{1}{nh_n} \sum_{i=1}^n \mathrm{Y}_i^{(n)}\, L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)}\, L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right| = \mathrm{O}_{\mathbb{P}}(1). \qquad (25)$$

We now replace the supremum in (25) by a supremum over a grid by focusing on the oscillation of the left-hand side. For a given $z \in \mathbb{R}$, let

$$A_n(z) := \left\{ z' \in [a_1, b_1] \;\middle|\; |z' - z| \le h_n \frac{\sqrt{\log n}}{n^{2/5}} \right\}.$$

Then $[a_1, b_1]$ is covered by the $A_n(z_{n,j})$, with

$$z_{n,j} = a_1 + j\, h_n \frac{\sqrt{\log n}}{n^{2/5}},\ j = 1, \ldots, \left\lfloor \frac{b_1 - a_1}{h_n \frac{\sqrt{\log n}}{n^{2/5}}} \right\rfloor =: N_n,$$

where $\lfloor \cdot \rfloor$ denotes the floor function. Besides, writing $|L(z') - L(z)| \le C_L |z' - z|$ by Lipschitz continuity of L, we also find

$$|z' - z| \le 1 \Rightarrow |L(z') - L(z)| \le |z' - z| \mathcal{L}(z) \ \text{ with } \mathcal{L}(z) := C_L \mathbb{1}\{|z| \le 2\}.$$

Let $z_{n,j}$ be a grid point and $z \in A_n(z_{n,j})$. By construction $|z - z_{n,j}|/h_n \le \sqrt{\log(n)}/n^{2/5}$ which converges to 0, so that, for n large enough,

$$\forall i \in \{1, \ldots, n\},\ \left| L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - L\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right)\right| \le \frac{\sqrt{\log n}}{n^{2/5}} \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right).$$

Then

$$\begin{aligned}
&\frac{n^{2/5}}{\sqrt{\log n}} \sup_{z\in A_n(z_{n,j})} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right]\right| \\
&\leq \frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| \\
&+ \frac{1}{nh_n}\sum_{i=1}^n |\mathrm{Y}_i^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) + \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right] \\
&\leq \frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| \\
&+ \left| \frac{1}{nh_n}\sum_{i=1}^n |\mathrm{Y}_i^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| \\
&+ 2\times \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right].
\end{aligned}$$

By the boundedness of $\mathcal{V}$, of $f_{\mathcal{X}}$ and of $z\mapsto \mathbb{E}\left[|\mathcal{Y}|\,|\mathcal{X}=z\right]$ over compact sub-intervals of (a,b), we find, for n large enough,

$$\sup_{a_1\leq z\leq b_1} \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \leq C_0$$

where C_0 is a finite constant. Consequently, for any constant $C > 2C_0$,

$$\begin{aligned}
&\frac{n^{2/5}}{\sqrt{\log n}} \sup_{z\in A_n(z_{n,j})} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right]\right| \\
&\leq \frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| \\
&+ \frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n |\mathrm{Y}_i^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| + C
\end{aligned}$$

where the (crude) inequality $n^{2/5}/\sqrt{\log n}\geq 1$, for n large enough, was used. Conclude, by writing $[a_1,b_1]\subset \cup_{1\leq j\leq N_n} A_n(z_{n,j})$, that

$$\begin{aligned}
&\mathbb{P}\left(\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1\leq z\leq b_1} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right]\right| > 3C\right) \\
&\leq N_n \max_{1\leq j\leq N_n} \mathbb{P}\left(\frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| > C\right) \\
&+ N_n \max_{1\leq j\leq N_n} \mathbb{P}\left(\frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n |\mathrm{Y}_i^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z_{n,j}-\mathcal{X}}{h_n}\right)\right]\right| > C\right).
\end{aligned}$$

We finish the proof by showing

$$\mathbb{P}\left(\frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n \mathrm{Y}_i^{(n)} L\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right]\right| > C\right) = \mathrm{O}\left(\frac{1}{n}\right) \tag{26}$$

and

$$\mathbb{P}\left(\frac{n^{2/5}}{\sqrt{\log n}} \left| \frac{1}{nh_n}\sum_{i=1}^n |\mathrm{Y}_i^{(n)}|\, \mathcal{L}\left(\frac{z-\mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[|\mathrm{Y}^{(n)}|\, \mathcal{L}\left(\frac{z-\mathcal{X}}{h_n}\right)\right]\right| > C\right) = \mathrm{O}\left(\frac{1}{n}\right) \tag{27}$$

for C large enough, uniformly in $z \in [a_1, b_1]$. Since N_n is of order $n^{2/5}/(h_n\sqrt{\log(n)}) \approx n^{3/5}/\sqrt{\log(n)} = \mathrm{o}(n)$, this will entail

$$\mathbb{P}\left(\frac{n^{2/5}}{\sqrt{\log n}} \sup_{a_1 \le z \le b_1} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \mathrm{Y}_i^{(n)} L\left(\frac{z - \mathcal{X}_i}{h_n}\right) - \frac{1}{h_n}\mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right| > 3C \right) = \mathrm{o}(1)$$

for C large enough, which is sufficient for our purposes. We only show (26) uniformly in $z \in [a_1, b_1]$; the proof of (27) is identical. Rewrite the left-hand side of (26) as

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} \left\{ \mathrm{Y}_i^{(n)} L\left(\frac{z - \mathcal{X}_i}{h_n}\right) - \mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right\} \right| > Cu_n \right),$$

with $u_n := n^{3/5} h_n \sqrt{\log n}$. Let v be a constant such that $|\mathcal{V}| \le v$ with probability 1. Note that for any i we have the crude bound

$$\left| \mathrm{Y}_i^{(n)} L\left(\frac{z - \mathcal{X}_i}{h_n}\right) - \mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right| \le 2vt_n \max_{-1 \le u \le 1} |L(u)|.$$

Remark also that, for n large enough,

$$\mathrm{Var}\left(\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right) \le v^2 \mathbb{E}\left[\mathcal{Y}^2 L^2\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \le Dh_n$$

for some finite constant D, by uniform boundedness of $f_{\mathcal{X}}$ and $z \mapsto \mathbb{E}\left[\mathcal{Y}^2 \,|\mathcal{X} = z\right]$ over compact sub-intervals of (a, b). By the Bernstein exponential inequality we get

$$\begin{aligned} &\mathbb{P}\left(\left| \sum_{i=1}^{n} \left\{ \mathrm{Y}_i^{(n)} L\left(\frac{z - \mathcal{X}_i}{h_n}\right) - \mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right\} \right| > Cu_n \right) \\ &\le 2\exp\left(-\frac{C^2 u_n^2/2}{Dnh_n + 2Cvt_n u_n \max_{[-1,1]} |L|/3} \right). \end{aligned}$$

Recalling that $t_n = n^t$ with $2/(5+\delta) < t < 2/5$, $u_n = n^{3/5} h_n \sqrt{\log n}$ and $nh_n^5 \to c \in (0, \infty)$, one finds

$$\frac{1}{\log n} \times \frac{C^2 u_n^2/2}{Dnh_n + 2Cvt_n u_n \max_{[-1,1]} |L|/3} \to \frac{c^{1/5} C^2}{2D} \quad \text{as} \quad n \to \infty$$

and therefore there is a constant $C' > 0$, independent of C, such that for n large enough

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} \left\{ \mathrm{Y}_i^{(n)} L\left(\frac{z - \mathcal{X}_i}{h_n}\right) - \mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right\} \right| > Cu_n \right) \le 2\exp\left(-C' C^2 \log n\right)$$

uniformly in $z \in [a_1, b_1]$. For C large enough, this yields

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} \left\{ \mathrm{Y}_i^{(n)} L\left(\frac{z - \mathcal{X}_i}{h_n}\right) - \mathbb{E}\left[\mathrm{Y}^{(n)} L\left(\frac{z-\mathcal{X}}{h_n}\right)\right] \right\} \right| > Cu_n \right) = \mathrm{O}\left(\frac{1}{n}\right)$$

which is equivalent to (26). This completes the proof. □

Lemma C.3 provides a uniform control, tailored to the assumptions of Proposition C.1, of the gap between smoothed moments and their asymptotic equivalents.

Lemma C.3. *Assume that the bivariate random pair $(\mathcal{X}, \mathcal{Y})$ is such that:*

- *$\mathcal{X}$ has support $[a, b]$, with $a < b$, and a density function $f_{\mathcal{X}}$ which has a continuous derivative on (a, b).*
- *The conditional moment function $m_{\mathcal{Y}|\mathcal{X}} : z \mapsto \mathbb{E}(\mathcal{Y}|\mathcal{X} = z)$ is well-defined and has a continuous derivative on (a, b).*

- *L is a bounded measurable function with support contained in $[-1,1]$.*

Then, as $h \to 0$:

(i) For any $a_1, b_1 \in [a,b]$ with $a < a_1 < b_1 < b$, we have, uniformly in $z \in [a_1, b_1]$,

$$\frac{1}{h}\mathbb{E}\left[\mathcal{Y}\, L\left(\frac{z-\mathcal{X}}{h}\right)\right] = m_{\mathcal{Y}|\mathcal{X}}(z) f_{\mathcal{X}}(z) \int_{-1}^{1} L(u)du - h\{m'_{\mathcal{Y}|\mathcal{X}}(z) f_{\mathcal{X}}(z) + m_{\mathcal{Y}|\mathcal{X}}(z) f'_{\mathcal{X}}(z)\} \int_{-1}^{1} uL(u)du + \mathrm{o}(h).$$

(ii) If moreover $f_{\mathcal{X}}$ and $m_{\mathcal{Y}|\mathcal{X}}$ are twice continuously differentiable on (a,b) then, uniformly in $z \in [a_1, b_1]$,

$$\begin{aligned}&\frac{1}{h}\mathbb{E}\left[\mathcal{Y}\, L\left(\frac{z-\mathcal{X}}{h}\right)\right] \\ &= m_{\mathcal{Y}|\mathcal{X}}(z) f_{\mathcal{X}}(z) \int_{-1}^{1} L(u)du - h\{m'_{\mathcal{Y}|\mathcal{X}}(z) f_{\mathcal{X}}(z) + m_{\mathcal{Y}|\mathcal{X}}(z) f'_{\mathcal{X}}(z)\} \int_{-1}^{1} uL(u)du \\ &+ \frac{h^2}{2}\{m''_{\mathcal{Y}|\mathcal{X}}(z) f_{\mathcal{X}}(z) + 2m'_{\mathcal{Y}|\mathcal{X}}(z) f'_{\mathcal{X}}(z) + m_{\mathcal{Y}|\mathcal{X}}(z) f''_{\mathcal{X}}(z)\} \int_{-1}^{1} u^2 L(u)du + \mathrm{o}(h^2).\end{aligned}$$

Proof. Note that

$$\frac{1}{h}\mathbb{E}\left[\mathcal{Y}\, L\left(\frac{z-\mathcal{X}}{h}\right)\right] = \int_{-1}^{1} m_{\mathcal{Y}|\mathcal{X}}(z-hu) f_{\mathcal{X}}(z-hu) L(u)\, du.$$

Parts (i) and (ii) are obtained by using the following Taylor formulae with integral remainder:

$$\varphi(z+\delta) = \varphi(z) + \delta\varphi'(z) + \int_z^{z+\delta} [\varphi'(t) - \varphi'(z)]\, dt$$

and

$$\varphi(z+\delta) = \varphi(z) + \delta\varphi'(z) + \frac{\delta^2}{2}\varphi''(z) + \int_z^{z+\delta} (z+\delta-t)[\varphi''(t) - \varphi''(z)]\, dt$$

applied to the function $\varphi : z \mapsto m_{\mathcal{Y}|\mathcal{X}}(z) f_{\mathcal{X}}(z)$. To get a uniform control of the remainders, use the fact that this function has uniformly continuous derivatives on any compact sub-interval of $[a,b]$, by Heine's theorem. □

Our next auxiliary result is the uniform consistency (with rate) of the estimators of g and σ in the heteroscedastic single-index model of Section 3.2.

Proposition C.1. *Assume that $(\boldsymbol{X}_i, Y_i)_{i\geq 1}$ are independent random pairs generated from the single-index model (M_2). Assume further that:*

- *The functions g and $\sigma > 0$ are continuous on $K_{\boldsymbol{\beta}}$ and twice continuously differentiable on the interior $K_{\boldsymbol{\beta}}^{\circ}$ of $K_{\boldsymbol{\beta}}$.*
- *The projection $\boldsymbol{\beta}^\top \boldsymbol{X}$ has a density function $f_{\boldsymbol{\beta}^\top \boldsymbol{X}}$ which is twice continuously differentiable and positive on $K_{\boldsymbol{\beta}}^{\circ}$.*
- *Each of the conditional moment functions $z \mapsto \mathbb{E}(X_j | \boldsymbol{\beta}^\top \boldsymbol{X} = z)$, $j \in \{1, \ldots, d\}$ is continuously differentiable on $K_{\boldsymbol{\beta}}^{\circ}$.*
- *There is $\delta > 0$ such that $\mathbb{E}|\varepsilon|^{2+\delta} < \infty$.*
- *L is a twice continuously differentiable and symmetric probability density function with support contained in $[-1,1]$.*

Assume also that $nh_n^5 \to c \in (0,\infty)$, *and* $t_n = n^t$ *with* $2/(5+\delta) < t < 2/5$. *Then, for any compact subset* K_0 *of* K^{o} *and any estimator* $\widehat{\boldsymbol{\beta}}$ *such that* $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \mathrm{O}_{\mathbb{P}}(1)$, *we have*

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x}\in K_0} \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)\right| = \mathrm{O}_{\mathbb{P}}(1)$$

$$\textit{and} \quad \frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x}\in K_0} \left| \widehat{\sigma}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)\right| = \mathrm{O}_{\mathbb{P}}(1).$$

Before proving this result, note that when K is convex, its projection $K_{\boldsymbol{\beta}} := \{\boldsymbol{\beta}^\top \boldsymbol{x},\, \boldsymbol{x} \in K\}$, which is also the support of $\boldsymbol{\beta}^\top \boldsymbol{X}$, is a compact interval containing at least two points (because K has a nonempty interior). Note also that Proposition C.1 is tailored to our framework in the sense that the assumption $\mathbb{E}|\varepsilon|^{2+\delta} < \infty$, which puts a constraint on the tail heaviness of the noise variable, is intuitively close to minimal for the estimation of g and σ by estimators of Nadaraya-Watson type. An inspection of the proof reveals that a similar theorem holds if $\widehat{g}_{h_n,t_n}$ and $\widehat{\sigma}_{h_n,t_n}$ are replaced by non-truncated versions, under the stronger moment assumption $\mathbb{E}|\varepsilon|^{5/2+\delta} < \infty$; see the comment below the statement of Lemma C.2. The regularity assumption on $z \mapsto \mathbb{E}(X_j|\boldsymbol{\beta}^\top \boldsymbol{X} = z)$ is a technical requirement, which is for instance satisfied if the density function $f_{\boldsymbol{X}}$ is continuously differentiable and positive on K°.

Proof. We start by proving the assertion on $\widehat{g}_{h_n,t_n}$. Define a truncated pseudo-Nadaraya-Watson estimator by

$$\widetilde{g}_{h_n,t_n}(z) = \sum_{i=1}^n Y_i \mathbb{1}\{|Y_i| \le t_n\}\, L\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right) \Bigg/ \sum_{i=1}^n L\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right).$$

The idea is to write

$$\begin{aligned}\left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)\right| \le{}& \left| g\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)\right| \\ &+ \left| \widetilde{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right)\right| \\ &+ \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \widetilde{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right)\right| \end{aligned} \tag{28}$$

and control each term on the right-hand side of (28) separately. To control the first term, we first apply the mean value theorem:

$$\left| g\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)\right| \le \left| \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}\right| \times \sup_{\lambda\in[0,1]} \left| g'\left(\boldsymbol{\beta}^\top \boldsymbol{x} + \lambda\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \boldsymbol{x}\right)\right|.$$

Since $K_0 \subset K^{\mathrm{o}}$, the distance between the compact set K_0 and the (compact) topological boundary of K is positive, *i.e.* $\rho := \inf\{\|x - y\|,\, x \in K_0,\, y \in K \setminus K^{\mathrm{o}}\} > 0$. It is then straightforward to show that, letting $K_{\boldsymbol{\beta}} = [u, v]$, we have $\boldsymbol{\beta}^\top \boldsymbol{x} \in [u + \rho/2, v - \rho/2]$ for any $\boldsymbol{x} \in K_0$. Since $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, we obtain that, with arbitrarily large probability as $n \to \infty$,

$$\forall \lambda \in [0,1],\ \forall \boldsymbol{x} \in K_0,\ \boldsymbol{\beta}^\top \boldsymbol{x} + \lambda\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \boldsymbol{x} \in [u + \rho/4, v - \rho/4]. \tag{29}$$

Because g' is continuous and therefore bounded on compact intervals contained in (u, v), this gives

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x}\in K_0} \left| g\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)\right| = \mathrm{O}_{\mathbb{P}}\left(\frac{n^{2/5}}{\sqrt{\log n}} \times \frac{1}{\sqrt{n}}\right) = \mathrm{o}_{\mathbb{P}}(1). \tag{30}$$

To control the second term, we show the uniform consistency of the regression pseudo-estimator $\widetilde{g}_{h_n,t_n}$. The assumptions of Lemma C.2 are fulfilled for $(\mathcal{X}, \mathcal{Y}, \mathcal{V}) = (\boldsymbol{\beta}^\top \boldsymbol{X}, Y, 1) = (\boldsymbol{\beta}^\top \boldsymbol{X}, g\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) +$

$\sigma\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right)\varepsilon, 1)$ and $(\mathcal{X}, \mathcal{Y}, \mathcal{V}) = (\boldsymbol{\beta}^\top \boldsymbol{X}, 1, 1)$. Recalling that ε is independent of $\boldsymbol{X}$ and centred, Lemma C.2 then provides

$$\widetilde{g}_{h_n,t_n}(z) = \frac{\frac{1}{h_n}\mathbb{E}\left[Y\, L\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}}{h_n}\right)\right] + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)}{\frac{1}{h_n}\mathbb{E}\left[L\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}}{h_n}\right)\right] + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)}$$

uniformly on any (fixed) compact subset of $K_{\boldsymbol{\beta}}^{\mathrm{o}} = (u, v)$. Noting that $h_n \sim (c/n)^{1/5}$ and $\int_{-1}^{1} uL(u)du = 0$ (because L is symmetric), Lemma C.3(ii) therefore entails

$$\widetilde{g}_{h_n,t_n}(z) = \frac{f_{\boldsymbol{\beta}^\top \boldsymbol{X}}(z)g(z) + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)}{f_{\boldsymbol{\beta}^\top \boldsymbol{X}}(z) + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)} = g(z) + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)$$

uniformly on any compact subset of (u, v), the last equality being correct because $f_{\boldsymbol{\beta}^\top \boldsymbol{X}}$ is bounded from below by a positive constant on such sets. Together with (29) for $\lambda = 1$, this yields

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x}\in K_0} \left|\widetilde{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - g\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right)\right| = \mathrm{O}_{\mathbb{P}}(1). \tag{31}$$

We conclude by controlling the third term in the right-hand side of (28). The idea is to define $\mathfrak{Y}_i^{(n)} := Y_i \mathbb{1}\{|Y_i| \le t_n\}$ and, for any z and $p = 0, 1$,

$$\widehat{m}_n^{(p)}(z) := \frac{1}{nh_n}\sum_{i=1}^{n}\left[\mathfrak{Y}_i^{(n)}\right]^p L\left(\frac{z-\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right)$$
$$\text{and } \widetilde{m}_n^{(p)}(z) := \frac{1}{nh_n}\sum_{i=1}^{n}\left[\mathfrak{Y}_i^{(n)}\right]^p L\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right).$$

With this notation,

$$\begin{aligned}\widehat{g}_{h_n,t_n}(z) - \widetilde{g}_{h_n,t_n}(z) &= \frac{\widehat{m}_n^{(1)}(z)}{\widehat{m}_n^{(0)}(z)} - \frac{\widetilde{m}_n^{(1)}(z)}{\widetilde{m}_n^{(0)}(z)} \\ &= \frac{[\widehat{m}_n^{(1)}(z) - \widetilde{m}_n^{(1)}(z)]\widetilde{m}_n^{(0)}(z) - [\widehat{m}_n^{(0)}(z) - \widetilde{m}_n^{(0)}(z)]\widetilde{m}_n^{(1)}(z)}{(\widetilde{m}_n^{(0)}(z) + [\widehat{m}_n^{(0)}(z) - \widetilde{m}_n^{(0)}(z)])\widetilde{m}_n^{(0)}(z)}.\end{aligned} \tag{32}$$

Since

$$\left|\widetilde{m}_n^{(0)}(z) - f_{\mathcal{X}}(z)\right| = \mathrm{o}_{\mathbb{P}}(1) \text{ and } \left|\widetilde{m}_n^{(1)}(z) - f_{\mathcal{X}}(z)g(z)\right| = \mathrm{o}_{\mathbb{P}}(1) \tag{33}$$

uniformly on any compact subset of (u, v) by Lemmas C.2 and C.3(ii), we concentrate on differences of the form

$$\widehat{m}_n^{(p)}(z) - \widetilde{m}_n^{(p)}(z) = \frac{1}{nh_n}\sum_{i=1}^{n}\left[\mathfrak{Y}_i^{(n)}\right]^p \left\{L\left(\frac{z-\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right) - L\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right)\right\}.$$

By Taylor's theorem with integral remainder applied to the function L, we find

$$\begin{aligned}&\widehat{m}_n^{(p)}(z) - \widetilde{m}_n^{(p)}(z) \\ &= -\frac{1}{nh_n}\sum_{i=1}^{n}\left[\mathfrak{Y}_i^{(n)}\right]^p \times \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}_i}{h_n} L'\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right) \\ &+ \frac{1}{nh_n}\sum_{i=1}^{n}\left[\mathfrak{Y}_i^{(n)}\right]^p \times \frac{1}{2}\left\{\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}_i}{h_n}\right\}^2 L''\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right) \\ &+ \frac{1}{nh_n}\sum_{i=1}^{n}\left[\mathfrak{Y}_i^{(n)}\right]^p \times \int_{(z-\boldsymbol{\beta}^\top \boldsymbol{X}_i)/h_n}^{(z-\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)/h_n}\left(\frac{z-\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n} - s\right)\left\{L''(s) - L''\left(\frac{z-\boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right)\right\} ds \\ &=: T_{1,n}(z) + T_{2,n}(z) + T_{3,n}(z).\end{aligned} \tag{34}$$

We handle these three terms separately.

Control of $T_{1,n}(z)$: Note that

$$T_{1,n}(z) = -\frac{1}{h_n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \left[\frac{1}{nh_n}\sum_{i=1}^n \left[\mathfrak{Y}_i^{(n)}\right]^p L'\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right)\boldsymbol{X}_i\right].$$

Recall that $\boldsymbol{X}$ has compact support; Lemma C.2 (choosing $\mathcal{V} = X_j$, $1 \le j \le d$) then yields

$$T_{1,n}(z) = -\frac{1}{h_n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \left[\frac{1}{h_n}\mathbb{E}\left(Y^p L'\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}}{h_n}\right)\boldsymbol{X}\right) + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)\right]$$

uniformly on any compact subset of (u, v). Because for any $j \in \{1, \ldots, d\}$,

$$\mathbb{E}(Y^p X_j | \boldsymbol{\beta}^\top \boldsymbol{X} = z) = [\mathbb{1}\{p = 0\} + g(z)\mathbb{1}\{p = 1\}]\,\mathbb{E}(X_j | \boldsymbol{\beta}^\top \boldsymbol{X} = z),$$

the conditional moment function $z \mapsto \mathbb{E}(Y^p X_j | \boldsymbol{\beta}^\top \boldsymbol{X} = z)$ satisfies the regularity requirements of Lemma C.3(i). By Lemma C.3(i) and the symmetry of L,

$$\frac{1}{h_n}\mathbb{E}\left(Y^p L'\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}}{h_n}\right)\boldsymbol{X}\right) = \mathrm{O}(h_n)$$

uniformly on any compact subset of (u, v). Since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})$, this yields

$$T_{1,n}(z) = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) = \mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right) \quad \text{uniformly on any compact subset of } (u, v). \tag{35}$$

Control of $T_{2,n}(z)$: Recall that $\boldsymbol{X}$ has compact support, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})$, and L'' is bounded to obtain, using the law of large numbers,

$$\sup_{z \in \mathbb{R}} |T_{2,n}(z)| = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{nh_n^3} \times \frac{1}{n}\sum_{i=1}^n |Y_i|^p\right) = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{nh_n^3}\right) = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{n^{2/5}}\right) = \mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right). \tag{36}$$

Control of $T_{3,n}(z)$: Use a change of variables to rewrite the integral term in $T_{3,n}(z)$ as

$$\begin{aligned}
&\int_{(z - \boldsymbol{\beta}^\top \boldsymbol{X}_i)/h_n}^{(z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)/h_n} \left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n} - s\right)\left\{L''(s) - L''\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right)\right\} ds \\
&= \int_0^{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i / h_n} \left(\frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i}{h_n} - u\right)\left\{L''\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n} + u\right) - L''\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right)\right\} du.
\end{aligned}$$

Since $\boldsymbol{X}$ has compact support and $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})$ we have

$$\max_{1 \le i \le n} \left|\frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i}{h_n}\right| = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{h_n\sqrt{n}}\right) = \mathrm{o}_{\mathbb{P}}(1).$$

By uniform continuity of the continuous and compactly supported function L'', it follows that

$$\max_{1 \le i \le n} \sup_{z \in \mathbb{R}} \sup_{|u| \le |(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i| / h_n} \left|L''\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n} + u\right) - L''\left(\frac{z - \boldsymbol{\beta}^\top \boldsymbol{X}_i}{h_n}\right)\right| = \mathrm{o}_{\mathbb{P}}(1).$$

We then get

$$\begin{aligned}
\sup_{z \in \mathbb{R}} |T_{3,n}(z)| &= \mathrm{o}_{\mathbb{P}}\left(\frac{1}{nh_n}\sum_{i=1}^n |Y_i|^p \left|\int_0^{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i / h_n} \left|\frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i}{h_n} - u\right| du\right|\right) \\
&= \mathrm{o}_{\mathbb{P}}\left(\frac{1}{nh_n}\sum_{i=1}^n |Y_i|^p \left[\frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i}{h_n}\right]^2\right) \\
&= \mathrm{o}_{\mathbb{P}}\left(\frac{1}{nh_n^3} \times \frac{1}{n}\sum_{i=1}^n |Y_i|^p\right) = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{nh_n^3}\right) = \mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right).
\end{aligned} \tag{37}$$

Combine (32), (33), (34), (35), (36) and (37) to obtain

$$\widehat{g}_{h_n,t_n}(z) - \widetilde{g}_{h_n,t_n}(z) = \frac{\mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)}{f_{\boldsymbol{\beta}^\top \boldsymbol{X}}(z) + \mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)} = \mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right)$$

uniformly on any compact subset of (u, v). Using (29) again with $\lambda = 1$, we get

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x} \in K_0} \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \widetilde{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) \right| = \mathrm{o}_{\mathbb{P}}(1). \tag{38}$$

Combining (28), (30), (31) and (38) concludes the proof of the assertion on $\widehat{g}_{h_n,t_n}$.

We turn to the control of $\widehat{\sigma}_{h_n,t_n}$, where the added difficulty is that the computation of the estimator is based on the absolute residuals $\widehat{Z}_{i,h_n,t_n} = \left| Y_i - \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i\right) \right|$ rather than on the "true values" $Z_i := \left| Y_i - g\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right) \right|$. We thus introduce its pseudo-estimator analogue based on the Z_i,

$$\overline{\sigma}_{h_n,t_n}(z) := \sum_{i=1}^n Z_i \mathbb{1}\left\{Z_i \le t_n\right\} L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right) \Bigg/ \sum_{i=1}^n L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right)$$

and we seek to control $|\widehat{\sigma}_{h_n,t_n}(z) - \overline{\sigma}_{h_n,t_n}(z)|$, for $z = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, uniformly in $\boldsymbol{x} \in K_0$. Write

$$\begin{aligned}
&\widehat{\sigma}_{h_n,t_n}(z) - \overline{\sigma}_{h_n,t_n}(z) \\
&= \sum_{i=1}^n \left[\widehat{Z}_{i,h_n,t_n} \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \le t_n \right\} - Z_i \mathbb{1}\left\{Z_i \le t_n\right\} \right] L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right) \Bigg/ \sum_{i=1}^n L\left(\frac{z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i}{h_n}\right).
\end{aligned}$$

Note that the only pairs $(\boldsymbol{X}_i, Y_i)$ making a nonzero contribution to this difference are those for which $|z - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \le h_n$. For $\boldsymbol{x} \in K_0$, we thus focus on controlling

$$\sup_{\boldsymbol{x} \in K_0} \left| \widehat{Z}_{i,h_n,t_n} \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \le t_n \right\} - Z_i \mathbb{1}\left\{Z_i \le t_n\right\} \right| \mathbb{1}\left\{ \left| \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i \right| \le h_n \right\}.$$

Since $\left| \widehat{Z}_{i,h_n,t_n} - Z_i \right| \le \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right) \right|$, the triangle inequality yields

$$\begin{aligned}
&\sup_{\boldsymbol{x} \in K_0} \left| \widehat{Z}_{i,h_n,t_n} \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \le t_n \right\} - Z_i \mathbb{1}\left\{Z_i \le t_n\right\} \right| \mathbb{1}\left\{ \left| \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i \right| \le h_n \right\} \\
&\le \sup_{\boldsymbol{x} \in K_0} \max_{i : |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \le h_n} \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right) \right| \qquad (39) \\
&+ \sup_{\boldsymbol{x} \in K_0} Z_i \left| \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \le t_n \right\} - \mathbb{1}\left\{Z_i \le t_n\right\} \right| \mathbb{1}\left\{ \left| \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i \right| \le h_n \right\}. \qquad (40)
\end{aligned}$$

We focus on (39) first, where the idea is to use our uniform convergence result on $\widehat{g}_{h_n,t_n}$. Write

$$|\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \le h_n \Rightarrow |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{X}_i| \le h_n + |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}_i| = h_n + \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$$

irrespective of the index i and $\boldsymbol{x} \in K_0$, so that, with arbitrarily large probability as $n \to \infty$,

$$\forall i \in \{1, \dots, n\},\ \forall \boldsymbol{x} \in K_0,\ |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \le h_n \Rightarrow |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{X}_i| \le 2h_n.$$

Recall that, by (29), $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} \in [u + \rho/4, v - \rho/4]$ with arbitrarily large probability as $n \to \infty$, irrespective of $\boldsymbol{x} \in K_0$. Since $h_n \to 0$, this yields, with arbitrarily large probability as $n \to \infty$,

$$\forall i \in \{1, \dots, n\},\ \forall \boldsymbol{x} \in K_0,\ |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \le h_n \Rightarrow \boldsymbol{\beta}^\top \boldsymbol{X}_i \in [u + \rho/8, v - \rho/8].$$

In other words, for such indices i, $\boldsymbol{X}_i$ belongs to the intersection of K and the inverse image of the closed interval $[u+\rho/8, v-\rho/8]$ by the (continuous) projection mapping $\boldsymbol{x} \mapsto \boldsymbol{\beta}^\top \boldsymbol{x}$. This intersection is itself a compact set K_1, say, and therefore, with arbitrarily large probability as $n \to \infty$,

$$\forall i \in \{1, \ldots, n\},\ \forall \boldsymbol{x} \in K_0,\ |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \leq h_n \Rightarrow \boldsymbol{X}_i \in K_1.$$

Note also that $K_1 \subset K^\circ$ since K_1 is contained in the (open) inverse image of the open interval $(u+\rho/16, v-\rho/16)$ by the same projection mapping. It then follows from our uniform convergence result on $\widehat{g}_{h_n,t_n}$ that

$$\sup_{\boldsymbol{x} \in K_0} \max_{i:\, |\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i| \leq h_n} \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right) \right| = \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right). \tag{41}$$

We can now control (40). Clearly

$$\begin{aligned}
&\left| \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \leq t_n \right\} - \mathbb{1}\left\{ Z_i \leq t_n \right\} \right| \\
&= \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \leq t_n,\, Z_i > t_n \right\} + \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} > t_n,\, Z_i \leq t_n \right\}.
\end{aligned}$$

Recall that $\left| \widehat{Z}_{i,h_n,t_n} - Z_i \right| \leq \left| \widehat{g}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i\right) - g\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right) \right|$ and use (41) together with the assumption $t_n \to \infty$ to find that, with arbitrarily large probability as $n \to \infty$,

$$\begin{aligned}
\forall i \in \{1, \ldots, n\},\ &\sup_{\boldsymbol{x} \in K_0} Z_i \left| \mathbb{1}\left\{ \widehat{Z}_{i,h_n,t_n} \leq t_n \right\} - \mathbb{1}\left\{ Z_i \leq t_n \right\} \right| \mathbb{1}\left\{ \left| \widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i \right| \leq h_n \right\} \\
&\leq Z_i \mathbb{1}\left\{ Z_i \leq 2t_n,\, Z_i > t_n \right\} + Z_i \mathbb{1}\left\{ Z_i > t_n/2,\, Z_i \leq t_n \right\} \\
&\leq Z_i \mathbb{1}\left\{ t_n/2 < Z_i \leq 2t_n \right\}.
\end{aligned} \tag{42}$$

Combine (41) and (42) to obtain, with arbitrarily large probability as $n \to \infty$,

$$\begin{aligned}
&\sup_{\boldsymbol{x} \in K_0} \left| \widehat{\sigma}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \overline{\sigma}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) \right| \\
&\leq \sup_{\boldsymbol{x} \in K_0} \left[\overline{\sigma}_{h_n,2t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \overline{\sigma}_{h_n,t_n/2}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) \right] + \mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{\log n}}{n^{2/5}}\right).
\end{aligned} \tag{43}$$

To conclude, note that since $\mathbb{E}|\varepsilon| = 1$,

$$Z := \left| Y - g\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) \right| = \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) + \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right)\left(|\varepsilon| - \mathbb{E}|\varepsilon|\right).$$

This single-index model linking Z to $\boldsymbol{X}$ has the same structure as model (M_2) and satisfies our assumptions, with g replaced by σ and ε replaced by $|\varepsilon| - \mathbb{E}|\varepsilon|$. Since for this model $\overline{\sigma}_{h_n,t_n}$ plays the role of $\widehat{g}_{h_n,t_n}$, we can use the first part of the Proposition to get

$$\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x} \in K_0} \left| \overline{\sigma}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right) \right| = \mathrm{O}_{\mathbb{P}}(1). \tag{44}$$

The result then follows by using (43) to write

$$\begin{aligned}
\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x} \in K_0} \left| \widehat{\sigma}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right) \right| \leq\ &\frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x} \in K_0} \left| \overline{\sigma}_{h_n,2t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right) \right| \\
&+ \frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x} \in K_0} \left| \overline{\sigma}_{h_n,t_n/2}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right) \right| \\
&+ \frac{n^{2/5}}{\sqrt{\log n}} \sup_{\boldsymbol{x} \in K_0} \left| \overline{\sigma}_{h_n,t_n}\left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}\right) - \sigma\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right) \right| \\
&+ \mathrm{O}_{\mathbb{P}}(1)
\end{aligned}$$

and then by using (44) as well as its analogues with t_n replaced by $t_n/2$ and $2t_n$. □

The following de-conditioning lemma is a stronger version of Lemma 8 in [44].

Lemma C.4. *Let $N = N(n) \xrightarrow{\mathbb{P}} \infty$ be a random sequence of integers that, for each n, takes its values in $\{0, 1, \ldots, n\}$. Suppose that (G_n) and (H_m) are sequences of random elements taking values in a metric space S endowed with its Borel $\sigma-$field. Assume that*

$$\forall n \geq 1,\ \forall m \in \{1, \ldots, n\},\ G_n \,|\, \{N(n) = m\} \stackrel{d}{=} H_m.$$

Then:

(i) If $H_m \xrightarrow{d} H$ as $m \to \infty$, we have $G_n \xrightarrow{d} H$ as $n \to \infty$.

If moreover S is a linear space endowed with a norm $\|\cdot\|$, then:

(ii) If $\|H_m\| = \mathrm{O}_{\mathbb{P}}(1)$, we have $\|G_n\| = \mathrm{O}_{\mathbb{P}}(1)$.

Finally, in the case $S = \mathbb{R}$:

(iii) If $H_m \xrightarrow{\mathbb{P}} +\infty$ as $m \to \infty$, we have $G_n \xrightarrow{\mathbb{P}} +\infty$ as $n \to \infty$.

Proof. Use the law of total probability to write, for any positive integer m_0 and any Borel subset A of S,

$$\begin{aligned}\mathbb{P}(G_n \in A) &= \mathbb{P}(G_n \in A, N(n) \leq m_0) + \sum_{m=m_0+1}^{n} \mathbb{P}(G_n \in A \,|\, N(n) = m)\mathbb{P}(N(n) = m) \\ &= \mathbb{P}(G_n \in A, N(n) \leq m_0) + \sum_{m=m_0+1}^{n} \mathbb{P}(H_m \in A)\mathbb{P}(N(n) = m). \qquad (45)\end{aligned}$$

To show (i), let A be a continuity set of H (in the sense that $\mathbb{P}(H \in \partial A) = 0$, where ∂A is the topological boundary of A). By the Portmanteau theorem, there is an integer m_0 such that for $m > m_0$, $|\mathbb{P}(H_m \in A) - \mathbb{P}(H \in A)| \leq \varepsilon/3$. With this choice of m_0 we have, for n large enough,

$$\begin{aligned}&|\mathbb{P}(G_n \in A) - \mathbb{P}(H \in A)| \\ &\leq \mathbb{P}(G_n \in A, N(n) \leq m_0) + \mathbb{P}(H \in A)\mathbb{P}(N(n) \leq m_0) + \frac{\varepsilon}{3} \sum_{m=m_0+1}^{n} \mathbb{P}(N(n) = m) \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.\end{aligned}$$

This proves (i). To show statements (ii) and (iii), deduce from (45) that for any m_0,

$$\mathbb{P}(G_n \in A) \leq \sup_{m > m_0} \mathbb{P}(H_m \in A) + \mathrm{o}(1) \text{ as } n \to \infty.$$

Fix $\varepsilon > 0$. To prove (ii), let $C > 0$ and m_0 be such that $\mathbb{P}(\|H_m\| > C) \leq \varepsilon/2$ for any $m > m_0$, and apply the above inequality with A being the complement of the closed ball with centre the origin and radius C along with this choice of m_0 to get $\mathbb{P}(\|G_n\| > C) \leq \varepsilon$ for n large enough, which is the desired result. Finally, to prove (iii), pick an arbitrary t and set $A = A_t = (-\infty, t]$. There is an integer m_0 such that $\mathbb{P}(H_m \in A_t) \leq \varepsilon/2$ for $m > m_0$; applying the above inequality with this choice of m_0 yields $\mathbb{P}(G_n \in A_t) \leq \varepsilon$ for n large enough, which is (iii). □

Our next result is a technical extension of Theorem 2.1 to the case when the sample size n is random. This will be key to the proof of our main theorems in Sections 3.2 and 3.3, where one has to work with a selected subset of observations whose size N is indeed random.

Lemma C.5. *Assume that there is $\delta > 0$ such that $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$, that ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0 < \gamma < 1/2$ and $\tau_n \uparrow 1$ is such that $n(1 - \tau_n) \to \infty$. Let $N = N(n) \xrightarrow{\mathbb{P}} \infty$ be a random sequence of integers that, for each n, takes its values in $\{0, 1, \ldots, n\}$. Suppose that, for any n and on the event $\{N > 0\}$, $\widehat{\varepsilon}_i^{(n)}$ and $\varepsilon_i^{(n)}$, $1 \leq i \leq N$ are given such that*

- *For any $n \geq 1$ and any $m \in \{1, \ldots, n\}$, the distribution of $(\varepsilon_1^{(n)}, \ldots, \varepsilon_N^{(n)})$ given $N = m$ is the distribution of m independent copies of ε,*
- *We have*

$$\sqrt{N(1-\tau_N)} \max_{1 \leq i \leq N} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i^{(n)}|}{1 + |\varepsilon_i^{(n)}|} \xrightarrow{\mathbb{P}} 0.$$

Let finally $\widehat{\xi}_{\tau_N}(\varepsilon) = \arg\min_{u \in \mathbb{R}} \sum_{i=1}^N \eta_{\tau_N}(\widehat{\varepsilon}_i^{(n)} - u)$ on $\{N > 0\}$ and 0 otherwise, as well as

$$\psi_N(u) = \frac{1}{2\xi_{\tau_N}^2(\varepsilon)} \sum_{i=1}^N \left[\eta_{\tau_N}\left(\varepsilon_i^{(n)} - \xi_{\tau_N}(\varepsilon) - \frac{u\xi_{\tau_N}(\varepsilon)}{\sqrt{N(1-\tau_N)}} \right) - \eta_{\tau_N}(\varepsilon_i^{(n)} - \xi_{\tau_N}(\varepsilon)) \right]$$

$$\text{and } \chi_N(u) = \frac{1}{2\xi_{\tau_N}^2(\varepsilon)} \sum_{i=1}^N \left[\eta_{\tau_N}\left(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_N}(\varepsilon) - \frac{u\xi_{\tau_N}(\varepsilon)}{\sqrt{N(1-\tau_N)}} \right) - \eta_{\tau_N}(\widehat{\varepsilon}_i^{(n)} - \xi_{\tau_N}(\varepsilon)) \right]$$

on $\{N > 0\}$, and 0 otherwise. Then we have $\chi_N(u) - \psi_N(u) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ and

$$\sqrt{N(1-\tau_N)} \left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1 \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\gamma^3}{1-2\gamma}\right).$$

Proof. To show that $\chi_N(u) - \psi_N(u) \xrightarrow{\mathbb{P}} 0$, following the ideas of the proof of Theorem 2.1, it is enough to prove that

$$T_{1,N} = \frac{\sqrt{1-\tau_N}}{\xi_{\tau_N}(\varepsilon)\sqrt{N}} \sum_{i=1}^N |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i^{(n)}| \xrightarrow{\mathbb{P}} 0 \tag{46}$$

and that, if $I_N(u) = [0, |u|\xi_{\tau_N}(\varepsilon)/\sqrt{N(1-\tau_N)}]$,

$$\begin{aligned} T_{2,N}(u) &= \frac{2}{\xi_{\tau_N}(\varepsilon)\sqrt{N(1-\tau_N)}} \\ &\quad \times \sum_{i=1}^N \sup_{|t| \in I_N(u)} |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i^{(n)}| \mathbb{1}\{\varepsilon_i^{(n)} - \xi_{\tau_N}(\varepsilon) - t > \min(\varepsilon_i^{(n)} - \widehat{\varepsilon}_i^{(n)}, 0)\} \\ &\quad \xrightarrow{\mathbb{P}} 0. \end{aligned} \tag{47}$$

Clearly, since $N = N(n) \xrightarrow{\mathbb{P}} \infty$ and in particular $N > 0$ with arbitrarily large probability,

$$T_{1,N} = \mathrm{o}_{\mathbb{P}}\left(\frac{1}{N} \sum_{i=1}^N (1 + |\varepsilon_i^{(n)}|) \right) = \mathrm{o}_{\mathbb{P}}(1)$$

where the law of large numbers is combined with the de-conditioning Lemma C.4(i), to show that $N^{-1}\sum_{i=1}^N (1 + |\varepsilon_i^{(n)}|) \xrightarrow{\mathbb{P}} 1 + \mathbb{E}|\varepsilon| < \infty$. This proves (46). We now turn to the control of $T_{2,N}(u)$. Use that $N = N(n) \xrightarrow{\mathbb{P}} \infty$ and follow the ideas leading to (11) in the proof of Theorem 2.1 to find, for n large enough,

$$\varepsilon_i^{(n)} - \xi_{\tau_N}(\varepsilon) - t > \min(\varepsilon_i^{(n)} - \widehat{\varepsilon}_i^{(n)}, 0) \Rightarrow \varepsilon_i^{(n)} > \frac{1}{6}\xi_{\tau_N}(\varepsilon)$$

with arbitrarily large probability, irrespective of $i \in \{1, \ldots, N\}$ and t such that $|t| \in I_N(u)$. Therefore, with arbitrarily large probability as $n \to \infty$:

$$\begin{aligned} T_{2,N}(u) &\leq \frac{2}{\xi_{\tau_N}(\varepsilon)\sqrt{N(1-\tau_N)}} \sum_{i=1}^N |\widehat{\varepsilon}_i^{(n)} - \varepsilon_i^{(n)}| \mathbb{1}\left\{ \varepsilon_i^{(n)} > \frac{1}{6}\xi_{\tau_N}(\varepsilon) \right\} \\ &= \mathrm{o}_{\mathbb{P}}\left(\frac{1}{N\xi_{\tau_N}(\varepsilon)(1-\tau_N)} \sum_{i=1}^N \varepsilon_i^{(n)} \mathbb{1}\left\{ \varepsilon_i^{(n)} > \frac{1}{6}\xi_{\tau_N}(\varepsilon) \right\} \right). \end{aligned}$$

Combine Lemma A.1 with the de-conditioning Lemma C.4(i) to get

$$T_{2,N}(u) = \mathrm{o}_{\mathbb{P}}(1).$$

This is (47). Combine (46) and (47) to get $\chi_N(u) - \psi_N(u) \xrightarrow{\mathbb{P}} 0$. Now a combination of the conclusion of the proof of Theorem 2 in [9] and the de-conditioning Lemma C.4(i) yields

$$\chi_N(u) = \psi_N(u) + \mathrm{o}_{\mathbb{P}}(1) \xrightarrow{d} -uZ\sqrt{\frac{2\gamma}{1-2\gamma}} + \frac{u^2}{2\gamma} \text{ as } n \to \infty$$

in the sense of finite-dimensional convergence, with Z being standard Gaussian. Since $\chi_N(u)$ is convex in u, the conclusion follows using the convexity lemma stated as Theorem 5 in [30]. □

Lemma C.6(i) below is a technical extension of Lemma A.3 to the case of a random sample size. It is essential in, among others, proving that the Hill estimator based on a random number of residuals is asymptotically Gaussian, which is stated below as Lemma C.6(ii); this will be used extensively in Sections 3.2 and 3.3.

Lemma C.6. *Let $k = k(n) \to \infty$ be a sequence of integers with $k/n \to 0$. Assume that ε has an infinite right endpoint. Let $N = N(n) \xrightarrow{\mathbb{P}} \infty$ be a random sequence of integers that, for each n, takes its values in $\{0, 1, \ldots, n\}$. Suppose that, for any n and on the event $\{N > 0\}$, $\widehat{\varepsilon}_i^{(n)}$ and $\varepsilon_i^{(n)}$, $1 \le i \le N$ are given such that*

- *For any $n \ge 1$ and any $m \in \{1, \ldots, n\}$, the distribution of $(\varepsilon_1^{(n)}, \ldots, \varepsilon_N^{(n)})$ given $N = m$ is the distribution of m independent copies of ε,*
- *We have*

$$R_N := \max_{1 \le i \le N} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i^{(n)}|}{1 + |\varepsilon_i^{(n)}|} \xrightarrow{\mathbb{P}} 0.$$

(i) Then we have both

$$\sup_{0<s\le 1} \left| \frac{\widehat{\varepsilon}^{(n)}_{N-\lfloor k(N)s \rfloor, N}}{\varepsilon^{(n)}_{N-\lfloor k(N)s \rfloor, N}} - 1 \right| = \mathrm{O}_{\mathbb{P}}(R_N) \quad \textit{and} \quad \sup_{0<s\le 1} \left| \log \left(\frac{\widehat{\varepsilon}^{(n)}_{N-\lfloor k(N)s \rfloor, N}}{\varepsilon^{(n)}_{N-\lfloor k(N)s \rfloor, N}} \right) \right| = \mathrm{O}_{\mathbb{P}}(R_N).$$

Here by convention $\widehat{\varepsilon}^{(n)}_{N-\lfloor k(N)s \rfloor, N}$ and $\varepsilon^{(n)}_{N-\lfloor k(N)s \rfloor, N}$ are equal to 1 on the event $\{N = 0\}$.

(ii) If moreover ε satisfies condition $\mathcal{C}_2(\gamma, \rho, A)$ and $\tau_n \uparrow 1$ is such that $n(1-\tau_n) \to \infty$, $\sqrt{n(1-\tau_n)}A((1-\tau_n)^{-1}) \to \lambda \in \mathbb{R}$ and $\sqrt{N(1-\tau_N)}R_N \xrightarrow{\mathbb{P}} 0$, then the Hill estimator

$$\widehat{\gamma}_{\lfloor N(1-\tau_N) \rfloor} = \frac{1}{\lfloor N(1-\tau_N) \rfloor} \sum_{i=1}^{\lfloor N(1-\tau_N) \rfloor} \log \frac{\widehat{\varepsilon}^{(n)}_{N-i+1,N}}{\widehat{\varepsilon}^{(n)}_{N-\lfloor N(1-\tau_N) \rfloor, N}}$$

is such that $\sqrt{N(1-\tau_N)}(\widehat{\gamma}_{\lfloor N(1-\tau_N) \rfloor} - \gamma) \xrightarrow{d} \mathcal{N}(\lambda/(1-\rho), \gamma^2)$.

Proof. We follow the proof of Lemma A.3. On the event $\{N > 0\} \cap \{R_N \le 1/4\}$, having arbitrarily high probability, we may write

$$\forall i \in \{1, \ldots, N\},\ \varepsilon^{(n)}_{i,N} - R_N(1 + |\varepsilon^{(n)}_{i,N}|) \le \widehat{\varepsilon}^{(n)}_{i,N} \le \varepsilon^{(n)}_{i,N} + R_N(1 + |\varepsilon^{(n)}_{i,N}|).$$

Given $N = m$, the random variable $\varepsilon^{(n)}_{N-k(N),N}$ has the same distribution as $\varepsilon_{m-k(m),m}$, the $(m-k(m))$th order statistic of a sample of m independent copies of ε. Since $\varepsilon_{m-k(m),m} \xrightarrow{\mathbb{P}} +\infty$ as $m \to \infty$, we obtain likewise $\varepsilon^{(n)}_{N-k(N),N} \xrightarrow{\mathbb{P}} +\infty$ by the de-conditioning Lemma C.4(iii). On the event $A_n := \{N > 0\} \cap \{R_N \le 1/4\} \cap \{\varepsilon_{N-k(N),N} \ge 1\}$, whose probability tends to 1, we have

$$\forall i \ge N - k(N),\ (1 - R_N)\varepsilon^{(n)}_{i,N} - R_N \le \widehat{\varepsilon}^{(n)}_{i,N} \le (1 + R_N)\varepsilon^{(n)}_{i,N} + R_N.$$

Therefore, on A_N,

$$\forall s \in (0,1],\ -2R_N \leq \frac{\widehat{\varepsilon}^{(n)}_{N-\lfloor k(N)s\rfloor,N}}{\varepsilon^{(n)}_{N-\lfloor k(N)s\rfloor,N}} - 1 \leq 2R_N.$$

Mimic then the final stages of the proof of Lemma A.3 to conclude the proof of (i).

(ii) Define

$$\widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor} = \frac{1}{\lfloor N(1-\tau_N)\rfloor} \sum_{i=1}^{\lfloor N(1-\tau_N)\rfloor} \log \frac{\varepsilon^{(n)}_{N-i+1,N}}{\varepsilon^{(n)}_{N-\lfloor N(1-\tau_N)\rfloor,N}}.$$

By (i) and the assumption $\sqrt{N(1-\tau_N)}R_N \xrightarrow{\mathbb{P}} 0$,

$$\sqrt{N(1-\tau_N)}(\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) = \sqrt{N(1-\tau_N)}(\widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) + \mathrm{o}_{\mathbb{P}}(1).$$

Combine Lemma C.4(i) and Theorem 3.2.5 in [13] to conclude the proof of (ii). □

Lemma C.7 contains the crucial arguments behind our construction in Section 3.3.

Lemma C.7. *Work in model (M_3). Assume that ε satisfies condition $\mathcal{C}_1(\gamma)$ and that K_0 is a measurable subset of the support of $\boldsymbol{X}$ such that $\mathbb{P}(\boldsymbol{X} \in K_0) > 0$.*

(i) There exists $\tau_c \in (0,1)$ such that $q_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})q_\tau(\varepsilon)$ for any $\tau \in [\tau_c, 1]$ and any $\boldsymbol{x}$ in the support of $\boldsymbol{X}$.

(ii) If $\mathbb{E}|\varepsilon_-| < \infty$ and $0 < \gamma < 1$, one has

$$\xi_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_\tau(\max(\varepsilon, (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x}))).$$

In particular the expectile $\xi_\tau(Y|\boldsymbol{X} = \boldsymbol{x})$ is asymptotically equivalent to $\xi_\tau(g(\boldsymbol{X})+\sigma(\boldsymbol{X})\varepsilon|\boldsymbol{X} = \boldsymbol{x})$ as $\tau \uparrow 1$.

(iii) The probability $\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}), \boldsymbol{X} \in K_0)$ is not zero. Let e have the same distribution as $(Y - g(\boldsymbol{X}))/\sigma(\boldsymbol{X})$ given that $g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon > y_0$ and $\boldsymbol{X} \in K_0$. Then for t so large that $(y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}) \leq t$ with probability 1,

$$\mathbb{P}(e > t) = \frac{\mathbb{P}(\varepsilon > t)}{\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}) \,|\, \boldsymbol{X} \in K_0)}.$$

In particular, e satisfies condition $\mathcal{C}_1(\gamma)$.

(iv) Let $p = \mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}) \,|\, \boldsymbol{X} \in K_0)$. Then $q_\tau(\varepsilon)/q_\tau(e) \to p^\gamma$ as $\tau \uparrow 1$. If moreover $\mathbb{E}|\varepsilon_-| < \infty$ and $0 < \gamma < 1$, then $\xi_\tau(\varepsilon)/\xi_\tau(e) \to p^\gamma$ as $\tau \uparrow 1$.

(v) If, in addition to $\mathbb{E}|\varepsilon_-| < \infty$ and $0 < \gamma < 1$, the random variable ε satisfies condition $\mathcal{C}_2(\gamma, \rho, A)$, then e satisfies condition $\mathcal{C}_2(\gamma, \rho, p^{-\rho}A)$ and, as $\tau \uparrow 1$,

$$\begin{aligned} p^\gamma \frac{\xi_\tau(e)}{\xi_\tau(\varepsilon)} &= 1 + p^\gamma \frac{\gamma(\gamma^{-1}-1)^\gamma}{q_\tau(\varepsilon)} \left(\mathbb{E}\left[\varepsilon \,\middle|\, \varepsilon > \frac{y_0 - g(\boldsymbol{X})}{\sigma(\boldsymbol{X})},\ \boldsymbol{X} \in K_0 \right] + \mathrm{o}(1) \right) \\ &\quad + \frac{p^{-\rho} - 1}{\rho} \left(1 + \rho \left[\frac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho} + \frac{(\gamma^{-1}-1)^{-\rho} - 1}{\rho} \right] + \mathrm{o}(1) \right) A((1-\tau)^{-1}). \end{aligned}$$

(vi) Under the assumptions of (v), as $\tau \uparrow 1$,

$$\begin{aligned} &\frac{\xi_\tau(Y|\boldsymbol{x})}{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_\tau(\varepsilon)} \\ &= 1 + \frac{\gamma(\gamma^{-1}-1)^\gamma}{q_\tau(\varepsilon)} \left(\mathbb{E}\left[\max\left(\varepsilon, \frac{y_0 - g(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \right) \right] + \mathrm{o}(1) \right) + \mathrm{o}(|A((1-\tau)^{-1})|). \end{aligned}$$

Proof. The key point is to remark that $Y = \max(g(\boldsymbol{X})+\sigma(\boldsymbol{X})\varepsilon, y_0)$. By independence between $\boldsymbol{X}$ and ε, the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ is then the distribution of $\max(g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\varepsilon, y_0) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \max(\varepsilon, (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x}))$.

(i) The τth conditional quantile of Y given $\boldsymbol{X} = \boldsymbol{x}$ is

$$q_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \max(q_\tau(\varepsilon), (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x})).$$

Since g and $1/\sigma$ are bounded on the support of $\boldsymbol{X}$ and $q_\tau(\varepsilon) \to \infty$ as $\tau \uparrow 1$, one has $q_\tau(\varepsilon) > (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x})$ for τ large enough, irrespective of $\boldsymbol{x}$. Conclude that there is $\tau_c \in (0,1)$ with $q_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x}) q_\tau(\varepsilon)$ for any $\tau \in [\tau_c, 1]$ and any $\boldsymbol{x}$ in the support of $\boldsymbol{X}$, as required.

(ii) By location equivariance and positive homogeneity of expectiles, the τth conditional expectile of Y given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\xi_\tau(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \xi_\tau(\max(\varepsilon, (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x}))).$$

To conclude, it is sufficient to show that for any t_0, the extreme expectiles of ε and $\max(\varepsilon, t_0)$ are asymptotically equivalent. To do so we note that the definition of the τth unconditional expectile $\xi_\tau(\varepsilon)$ of ε as

$$\xi_\tau(\varepsilon) = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}(\eta_\tau(\varepsilon - \theta) - \eta_\tau(\varepsilon))$$

can equivalently be obtained as the τth quantile associated to the distribution function E defined as

$$1 - E(y) = \frac{\mathbb{E}\left[(\varepsilon - y)\mathbb{1}_{\{\varepsilon > y\}}\right]}{2\,\mathbb{E}\left[(\varepsilon - y)\mathbb{1}_{\{\varepsilon > y\}}\right] + y - \mathbb{E}[\varepsilon]}.$$

See *e.g.* the final paragraph of p.373 in [1]. Similarly the τth expectile $\xi_\tau(\max(\varepsilon, t_0))$ of $\max(\varepsilon, t_0)$ is obtained as the τth quantile associated to the distribution function E_0 defined as

$$1 - E_0(y) = \frac{\mathbb{E}\left[(\max(\varepsilon, t_0) - y)\mathbb{1}_{\{\max(\varepsilon, t_0) > y\}}\right]}{2\,\mathbb{E}\left[(\max(\varepsilon, t_0) - y)\mathbb{1}_{\{\max(\varepsilon, t_0) > y\}}\right] + y - \mathbb{E}[\max(\varepsilon, t_0)]}.$$

It is straightforward to check that for $y > t_0$

$$1 - E_0(y) = \frac{\mathbb{E}\left[(\varepsilon - y)\mathbb{1}_{\{\varepsilon > y\}}\right]}{2\,\mathbb{E}\left[(\varepsilon - y)\mathbb{1}_{\{\varepsilon > y\}}\right] + y - \mathbb{E}[\max(\varepsilon, t_0)]}.$$

Lemma 3(i) in [44] (with f therein chosen as the identity function and $a = 1$) entails that $y \mapsto 1/(1 - E(y))$ and $y \mapsto 1/(1 - E_0(y))$ are asymptotically equivalent as $y \to \infty$ and regularly varying with positive index. Let U and U_0 denote the pertaining tail quantile functions, *i.e.* the left-continuous inverses of $1/(1-E)$ and $1/(1-E_0)$; these are also regularly varying, and we will conclude by proving that U and U_0 are asymptotically equivalent. A combination of Equations (1.2.26) and (1.2.28) in [13] and the regular variation property of U entails

$$\begin{aligned}
&\lim_{t\to\infty} \frac{t^{-1}}{(1-E)(U(t))} = \lim_{t\to\infty} \frac{t^{-1}}{(1-E_0)(U(t))} = \lim_{t\to\infty} \frac{t^{-1}}{(1-E_0)(U_0(t))} = 1,\\
&\lim_{t\to\infty} t^{-1} U(1/(1-E)(t)) = \lim_{t\to\infty} t^{-1} U(1/(1-E_0)(t)) = \lim_{t\to\infty} t^{-1} U_0(1/(1-E_0)(t)) = 1.
\end{aligned}$$

Apply Proposition B.1.9.10 in [13] to obtain that U and U_0 are indeed asymptotically equivalent, thus completing the proof of (ii).

(iii) First of all, if $\mathbb{P}_{\boldsymbol{X}}$ denotes the distribution of $\boldsymbol{X}$,

$$\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}), \boldsymbol{X} \in K_0) = \int_{K_0} \mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x}))\, \mathbb{P}_{\boldsymbol{X}}(d\boldsymbol{x}) > 0$$

because $\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x})) > 0$ for any $\boldsymbol{x}$ (since ε is heavy-tailed) and $\mathbb{P}(\boldsymbol{X} \in K_0) > 0$. Write then

$$\begin{aligned}
\mathbb{P}(e > t) &= \mathbb{P}(\varepsilon > t \,|\, g(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon > y_0, \boldsymbol{X} \in K_0)\\
&= \frac{\mathbb{P}(\varepsilon > t, \varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}), \boldsymbol{X} \in K_0)}{\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}), \boldsymbol{X} \in K_0)}.
\end{aligned}$$

It is indeed possible to take t so large that $(y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}) \leq t$ with probability 1 since g and $1/\sigma$ are bounded on the support of $\boldsymbol{X}$. For such t,

$$\mathbb{P}(e > t) = \frac{\mathbb{P}(\varepsilon > t, \boldsymbol{X} \in K_0)}{\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}), \boldsymbol{X} \in K_0)} = \frac{\mathbb{P}(\varepsilon > t)}{\mathbb{P}(\varepsilon > (y_0 - g(\boldsymbol{X}))/\sigma(\boldsymbol{X}) \,|\, \boldsymbol{X} \in K_0)}$$

by independence between $\boldsymbol{X}$ and ε, which is the required result.

(iv) That $q_\tau(\varepsilon)/q_\tau(e) \to p^\gamma$ as $\tau \uparrow 1$ directly follows from the identity $\mathbb{P}(e > t) = p^{-1}\mathbb{P}(\varepsilon > t)$ for t large enough, and therefore $q_\tau(e) = q_{1-p(1-\tau)}(\varepsilon)$ for τ close enough to 1, combined with the regular variation property of $t \mapsto U(t) = q_{1-t^{-1}}(\varepsilon)$. The convergence $\xi_\tau(\varepsilon)/\xi_\tau(e) \to p^\gamma$ as $\tau \uparrow 1$ follows from the asymptotic proportionality relationship between extreme quantiles and expectiles applied to both e and ε (which have the same extreme value index).

(v) Recall from the proof of (iv) that for τ close enough to 1, $q_\tau(e) = q_{1-p(1-\tau)}(\varepsilon)$. Set $V(t) = q_{1-t^{-1}}(e)$ and pick $x > 0$. For t large enough, we find

$$\begin{aligned}\frac{V(tx)}{V(t)} = \frac{U(p^{-1}tx)}{U(p^{-1}t)} &= x^\gamma + A(p^{-1}t)\left(x^\gamma \frac{x^\rho - 1}{\rho} + \mathrm{o}(1)\right) \\ &= x^\gamma + p^{-\rho}A(t)\left(x^\gamma \frac{x^\rho - 1}{\rho} + \mathrm{o}(1)\right)\end{aligned}$$

by assumption $\mathcal{C}_2(\gamma, \rho, A)$ on ε and regular variation of $|A|$ with index ρ (see Section 2.3 in [13]). This exactly means that e satisfies condition $\mathcal{C}_2(\gamma, \rho, p^{-\rho}A)$. Write then

$$p^\gamma \frac{\xi_\tau(e)}{\xi_\tau(\varepsilon)} = p^\gamma \frac{q_\tau(e)}{q_\tau(\varepsilon)} \times (\gamma^{-1} - 1)^\gamma \frac{\xi_\tau(e)}{q_\tau(e)} \times (\gamma^{-1} - 1)^{-\gamma} \frac{q_\tau(\varepsilon)}{\xi_\tau(\varepsilon)}. \tag{48}$$

Use again the identity $q_\tau(e) = q_{1-p(1-\tau)}(\varepsilon)$ for τ close enough to 1 to get

$$p^\gamma \frac{q_\tau(e)}{q_\tau(\varepsilon)} = p^\gamma \frac{U(p^{-1}(1-\tau)^{-1})}{U((1-\tau)^{-1})} = 1 + \left(\frac{p^{-\rho} - 1}{\rho} + \mathrm{o}(1)\right) A((1-\tau)^{-1}). \tag{49}$$

Proposition 1(i) in [11] applied to the random variable ε (having expectation 0) entails

$$\begin{aligned}&(\gamma^{-1} - 1)^{-\gamma} \frac{q_\tau(\varepsilon)}{\xi_\tau(\varepsilon)} \\ &= 1 - \left(\frac{(\gamma^{-1} - 1)^{-\rho}}{1 - \gamma - \rho} + \frac{(\gamma^{-1} - 1)^{-\rho} - 1}{\rho} + \mathrm{o}(1)\right) A((1-\tau)^{-1}) + \mathrm{o}\left(\frac{1}{q_\tau(\varepsilon)}\right).\end{aligned} \tag{50}$$

This same result applied to the random variable e, which satisfies condition $\mathcal{C}_2(\gamma, \rho, p^{-\rho}A)$, gives

$$\begin{aligned}(\gamma^{-1} - 1)^\gamma \frac{\xi_\tau(e)}{q_\tau(e)} &= 1 + \frac{\gamma(\gamma^{-1} - 1)^\gamma}{q_\tau(e)}(\mathbb{E}(e) + \mathrm{o}(1)) \\ &\quad + \left(\frac{(\gamma^{-1} - 1)^{-\rho}}{1 - \gamma - \rho} + \frac{(\gamma^{-1} - 1)^{-\rho} - 1}{\rho} + \mathrm{o}(1)\right) p^{-\rho} A((1-\tau)^{-1}) \\ &= 1 + p^\gamma \frac{\gamma(\gamma^{-1} - 1)^\gamma}{q_\tau(\varepsilon)} \left(\mathbb{E}\left[\varepsilon \,\middle|\, \varepsilon > \frac{y_0 - g(\boldsymbol{X})}{\sigma(\boldsymbol{X})}, \boldsymbol{X} \in K_0\right] + \mathrm{o}(1)\right) \\ &\quad + p^{-\rho}\left(\frac{(\gamma^{-1} - 1)^{-\rho}}{1 - \gamma - \rho} + \frac{(\gamma^{-1} - 1)^{-\rho} - 1}{\rho} + \mathrm{o}(1)\right) A((1-\tau)^{-1}).\end{aligned} \tag{51}$$

Combine (48), (49), (50) and (51) to get (v).

(vi) From (ii),

$$\begin{aligned}\frac{\xi_\tau(Y|\boldsymbol{x})}{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_\tau(\varepsilon)} - 1 &= \frac{\sigma(\boldsymbol{x})[\xi_\tau(\max(\varepsilon, (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x}))) - \xi_\tau(\varepsilon)]}{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_\tau(\varepsilon)} \\ &= \left(\frac{\xi_\tau(\max(\varepsilon, (y_0 - g(\boldsymbol{x}))/\sigma(\boldsymbol{x})))}{\xi_\tau(\varepsilon)} - 1\right)(1 + \mathrm{o}(1))\end{aligned}$$

because $\xi_\tau(\varepsilon) \to \infty$ as $\tau \uparrow 1$. To complete the proof we show that for any t_0,

$$\frac{\xi_\tau(\max(\varepsilon, t_0))}{\xi_\tau(\varepsilon)} = 1 + \frac{\gamma(\gamma^{-1}-1)^\gamma}{q_\tau(\varepsilon)} \left(\mathbb{E}[\max(\varepsilon, t_0)] + \mathrm{o}(1)\right) + \mathrm{o}(|A((1-\tau)^{-1})|)$$

as $\tau \uparrow 1$. This is done by, first, writing

$$\frac{\xi_\tau(\max(\varepsilon, t_0))}{\xi_\tau(\varepsilon)} = \frac{\xi_\tau(\max(\varepsilon, t_0))}{q_\tau(\max(\varepsilon, t_0))} \times \frac{q_\tau(\max(\varepsilon, t_0))}{q_\tau(\varepsilon)} \times \frac{q_\tau(\varepsilon)}{\xi_\tau(\varepsilon)} = \frac{\xi_\tau(\max(\varepsilon, t_0))}{q_\tau(\max(\varepsilon, t_0))} \times \frac{q_\tau(\varepsilon)}{\xi_\tau(\varepsilon)}$$

for τ close enough to 1. Then, using the fact that $\max(\varepsilon, t_0)$ and ε have the same quantile function for τ large enough, we obtain, by Proposition 1(i) in [11],

$$\begin{aligned}(\gamma^{-1}-1)^\gamma \frac{\xi_\tau(\max(\varepsilon, t_0))}{q_\tau(\max(\varepsilon, t_0))} &= 1 + \frac{\gamma(\gamma^{-1}-1)^\gamma}{q_\tau(\varepsilon)} (\mathbb{E}[\max(\varepsilon, t_0)] + \mathrm{o}(1)) \\ &\quad + \left(\frac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho} + \frac{(\gamma^{-1}-1)^{-\rho} - 1}{\rho} + \mathrm{o}(1) \right) A((1-\tau)^{-1}).\end{aligned}$$

Combining this with (50) completes the proof. □

Our final auxiliary result is a direct extension of Theorem 2.1 to the case when the residuals $\widehat{\varepsilon}_i^{(n)}$ approximate an array $\varepsilon_i^{(n)}$, with $1 \le i \le s_n \to \infty$. This will be useful to deal with the case of ARMA and GARCH models.

Lemma C.8. *Let (s_n) be a positive sequence of integers tending to infinity. Assume that, for any n, the $\varepsilon_i^{(n)}$, $1 \le i \le s_n$, are independent copies of a random variable ε such that there is $\delta > 0$ with $\mathbb{E}|\varepsilon_-|^{2+\delta} < \infty$ and ε satisfies condition $\mathcal{C}_1(\gamma)$ with $0 < \gamma < 1/2$. Let $\tau_n \uparrow 1$ be such that $s_n(1-\tau_n) \to \infty$. Suppose moreover that the array of random variables $\widehat{\varepsilon}_i^{(n)}$, $1 \le i \le s_n$, satisfies*

$$\sqrt{s_n(1-\tau_n)} \max_{1 \le i \le s_n} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i^{(n)}|}{1 + |\varepsilon_i^{(n)}|} \xrightarrow{\mathbb{P}} 0.$$

Define

$$\widehat{\xi}_{\tau_n}(\varepsilon) = \arg\min_{u \in \mathbb{R}} \sum_{i=1}^{s_n} \eta_{\tau_n}(\widehat{\varepsilon}_i^{(n)} - u).$$

Then we have $\sqrt{s_n(1-\tau_n)} \left(\dfrac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)} - 1 \right) \xrightarrow{d} \mathcal{N}\left(0, \dfrac{2\gamma^3}{1-2\gamma}\right)$.

Appendix D: Worked-out examples: Proofs of the main results

Proof of Corollary 3.1. (i) The key is to write

$$\begin{aligned}&\sqrt{n(1-\tau_n)} \left(\frac{\widehat{\xi}_{\tau_n}(Y|\boldsymbol{x})}{\xi_{\tau_n}(Y|\boldsymbol{x})} - 1 \right) \\ &= \frac{(1+\boldsymbol{\theta}^\top \boldsymbol{x})\xi_{\tau_n}(\varepsilon)}{\alpha + \boldsymbol{\beta}^\top \boldsymbol{x} + (1+\boldsymbol{\theta}^\top \boldsymbol{x})\xi_{\tau_n}(\varepsilon)} \times \sqrt{n(1-\tau_n)} \left(\frac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)} - 1 \right) \\ &+ \frac{\sqrt{1-\tau_n}}{\alpha + \boldsymbol{\beta}^\top \boldsymbol{x} + (1+\boldsymbol{\theta}^\top \boldsymbol{x})\xi_{\tau_n}(\varepsilon)} \times \sqrt{n} \left[\widehat{\alpha} - \alpha + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \boldsymbol{x} \right] \\ &+ \frac{\sqrt{1-\tau_n}\, \widehat{\xi}_{\tau_n}(\varepsilon)}{\alpha + \boldsymbol{\beta}^\top \boldsymbol{x} + (1+\boldsymbol{\theta}^\top \boldsymbol{x})\xi_{\tau_n}(\varepsilon)} \times \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{x}.\end{aligned}$$

Now

$$\widehat{\varepsilon}_i^{(n)} - \varepsilon_i = \frac{\alpha - \widehat{\alpha} + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_i}{1 + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{X}_i} + \frac{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{X}_i}{1 + \widehat{\boldsymbol{\theta}}^\top \boldsymbol{X}_i}\varepsilon_i.$$

Then clearly, by Lemma C.1 and since $\boldsymbol{X}$ has a compact support,

$$\sqrt{n} \max_{1 \le i \le n} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|}{1 + |\varepsilon_i|} = \mathrm{O}_{\mathbb{P}}(1).$$

We conclude by combining Lemma C.1, Theorem 2.1 and the convergence $\xi_{\tau_n}(\varepsilon) \to \infty$.

(ii) Combine (i) with the second convergence in Theorem 2.3. □

Proof of Theorem 3.1. (i) We first show

$$\sqrt{N(1-\tau_N)} \left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1 \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{2\gamma^3}{1-2\gamma}\right). \tag{52}$$

Let $\varepsilon_{1,K_0}, \ldots, \varepsilon_{N,K_0}$ be those noise variables whose corresponding covariates $\boldsymbol{X}_i \in K_0$, and note that given $N = m > 0$, $(\varepsilon_{1,K_0}, \ldots, \varepsilon_{N,K_0}) \overset{d}{=} (\varepsilon_1, \ldots, \varepsilon_m)$. Besides, $N = N(K_0, n)$ is a binomial random variable with parameters n and $\mathbb{P}(\boldsymbol{X} \in K_0)$, so that $N/n \xrightarrow{\mathbb{P}} \mathbb{P}(\boldsymbol{X} \in K_0) > 0$. Since $\tau_n = 1 - n^{-a}$ with $a \in (1/5, 1)$,

$$\sqrt{N(1-\tau_N)} = N^{(1-a)/2} = \mathrm{O}_{\mathbb{P}}(n^{(1-a)/2}) = \mathrm{o}_{\mathbb{P}}(n^{2/5}/\sqrt{\log n})$$

so that

$$\sqrt{N(1-\tau_N)} \max_{1 \le i \le N} \frac{|\widehat{\varepsilon}_{i,K_0}^{(n)} - \varepsilon_{i,K_0}|}{1 + |\varepsilon_{i,K_0}|} = \mathrm{o}_{\mathbb{P}}\left(\frac{n^{2/5}}{\sqrt{\log n}} \max_{1 \le i \le n} \frac{|\widehat{\varepsilon}_i^{(n)} - \varepsilon_i|}{1 + |\varepsilon_i|} \mathbb{1}\{\boldsymbol{X}_i \in K_0\} \right) = \mathrm{o}_{\mathbb{P}}(1).$$

Apply then Lemma C.5 to get (52). Statement (i) then follows in a straightforward way from Proposition C.1 and the representation

$$\begin{aligned} \frac{\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x})}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1 = {} & \frac{\widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}) - g(\boldsymbol{\beta}^\top \boldsymbol{x})}{g(\boldsymbol{\beta}^\top \boldsymbol{x}) + \sigma(\boldsymbol{\beta}^\top \boldsymbol{x})\xi_{\tau_N}(\varepsilon)} + \frac{\widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}) - \sigma(\boldsymbol{\beta}^\top \boldsymbol{x})}{g(\boldsymbol{\beta}^\top \boldsymbol{x}) + \sigma(\boldsymbol{\beta}^\top \boldsymbol{x})\xi_{\tau_N}(\varepsilon)} \widehat{\xi}_{\tau_N}(\varepsilon) \\ & + \frac{\sigma(\boldsymbol{\beta}^\top \boldsymbol{x})}{\sigma(\boldsymbol{\beta}^\top \boldsymbol{x}) + g(\boldsymbol{\beta}^\top \boldsymbol{x})/\xi_{\tau_N}(\varepsilon)} \left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1 \right). \end{aligned}$$

(ii) Set $\widehat{\xi}^\star_{\tau'_N}(\varepsilon) = \left(\frac{1-\tau'_N}{1-\tau_N} \right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_N}(\varepsilon)$. Use the ideas of the proof of Theorem 2.3 to find that

$$\frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau'_N)]} \left(\frac{\widehat{\xi}^\star_{\tau'_N}(Y|\boldsymbol{x})}{\xi_{\tau'_N}(Y|\boldsymbol{x})} - 1 \right) \quad \text{and} \quad \frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau'_N)]} \left(\frac{\widehat{\xi}^\star_{\tau'_N}(\varepsilon)}{\xi_{\tau'_N}(\varepsilon)} - 1 \right)$$

have the same asymptotic distribution. Our result is then shown by using the assumption $\sqrt{N(1-\tau_N)}(\overline{\gamma} - \gamma) \xrightarrow{d} \Gamma$, as well as convergence (52) and by adapting directly the proof of Theorem 5 of [11] to obtain

$$\frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau'_N)]} \left(\frac{\widehat{\xi}^\star_{\tau'_N}(\varepsilon)}{\xi_{\tau'_N}(\varepsilon)} - 1 \right) \xrightarrow{d} \Gamma.$$

We omit the details. □

Proof of Theorem 3.2. First of all, define

$$\widehat{\xi}_{\tau_N}(\varepsilon) := \left(\frac{N}{N_0}\right)^{\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}} \widehat{\xi}_{\tau_N}(e)$$

so that $\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x}) = \widehat{g}(\boldsymbol{x}) + \widehat{\sigma}(\boldsymbol{x})\widehat{\xi}_{\tau_N}(\varepsilon)$. Then

$$\begin{aligned}
&\frac{\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x})}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1 \\
&= \left(\frac{\widehat{g}(\boldsymbol{x}) + \widehat{\sigma}(\boldsymbol{x})\widehat{\xi}_{\tau_N}(\varepsilon)}{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_N}(\varepsilon)} - 1\right) \frac{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(Y|\boldsymbol{x})} \\
&+ \left(\frac{g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1\right) \\
&= \left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1\right)(1 + \mathrm{o}_{\mathbb{P}}(1)) + \mathrm{o}_{\mathbb{P}}(|\widehat{g}(\boldsymbol{x}) - g(\boldsymbol{x})|) + \mathrm{O}_{\mathbb{P}}\left(\left|\frac{\widehat{\sigma}(\boldsymbol{x})}{\sigma(\boldsymbol{x})} - 1\right|\right) \\
&- \frac{\gamma(\gamma^{-1}-1)^{\gamma}}{q_{\tau_N}(\varepsilon)}\left(\mathbb{E}\left[\max\left(\varepsilon, \frac{y_0 - g(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)\right] + \mathrm{o}_{\mathbb{P}}(1)\right) + \mathrm{o}_{\mathbb{P}}(|A((1-\tau_N)^{-1})|)
\end{aligned}$$

by Lemma C.7(vi), the consistency assumption on $\widehat{g}$ and $\widehat{\sigma}$, and $N = N(n) \xrightarrow{\mathbb{P}} \infty$. Now $1/v_n = \mathrm{o}_{\mathbb{P}}(1/\sqrt{N(1-\tau_N)})$, because $n^{1-a}/v_n^2 \to 0$ and $N(1-\tau_N) = N^{1-a} \leq n^{1-a}$. The v_n-consistency of $\widehat{g}$ and $\widehat{\sigma}$ then entails

$$\begin{aligned}
\sqrt{N(1-\tau_N)}\left(\frac{\widehat{\xi}_{\tau_N}(Y|\boldsymbol{x})}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1\right) &= \sqrt{N(1-\tau_N)}\left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1\right)(1 + \mathrm{o}_{\mathbb{P}}(1)) \\
&- \gamma(\gamma^{-1}-1)^{\gamma}\mathbb{E}\left[\max\left(\varepsilon, \frac{y_0 - g(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)\right]\mu + \mathrm{o}_{\mathbb{P}}(1). \qquad (53)
\end{aligned}$$

It is therefore sufficient to consider the convergence of $\widehat{\xi}_{\tau_N}(\varepsilon)$. Write

$$\begin{aligned}
\log\left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)}\right) &= (\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma)\log\left(\frac{N}{N_0}\right) + \gamma\left[\log\left(\frac{N}{N_0}\right) - \log p\right] \\
&+ \log\left(\frac{\widehat{\xi}_{\tau_N}(e)}{\xi_{\tau_N}(e)}\right) + \log\left(p^{\gamma}\frac{\xi_{\tau_N}(e)}{\xi_{\tau_N}(\varepsilon)}\right).
\end{aligned}$$

The quantity N/N_0 is a $\sqrt{n}-$consistent estimator of $p > 0$, thus making the second term a $\mathrm{O}_{\mathbb{P}}(1/\sqrt{n}) = \mathrm{o}_{\mathbb{P}}(1/\sqrt{N(1-\tau_N)})$, and the fourth term is controlled with Lemma C.7(v) and a Taylor expansion. Therefore

$$\begin{aligned}
\log\left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)}\right) &= [\log p + \mathrm{o}_{\mathbb{P}}(1)](\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) + \log\left(\frac{\widehat{\xi}_{\tau_N}(e)}{\xi_{\tau_N}(e)}\right) \\
&+ p^{\gamma}\gamma(\gamma^{-1}-1)^{\gamma}\left(\mathbb{E}\left[\varepsilon \,\middle|\, \varepsilon > \frac{y_0 - g(\boldsymbol{X})}{\sigma(\boldsymbol{X})}, \boldsymbol{X} \in K_0\right]\right)\frac{\mu}{\sqrt{N(1-\tau_N)}} \\
&+ \frac{p^{-\rho}-1}{\rho}\left(1 + \rho\left[\frac{(\gamma^{-1}-1)^{-\rho}}{1-\gamma-\rho} + \frac{(\gamma^{-1}-1)^{-\rho}-1}{\rho}\right]\right)\frac{\lambda}{\sqrt{N(1-\tau_N)}} \\
&+ \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{N(1-\tau_N)}}\right). \qquad (54)
\end{aligned}$$

It remains to analyse the joint convergence of $\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}$ and $\widehat{\xi}_{\tau_N}(e)$. First, clearly

$$\max_{1\leq i\leq N}\frac{|\widehat{e}_i^{(n)} - e_i|}{1 + |e_i|} = \mathrm{O}_{\mathbb{P}}(1/v_n) = \mathrm{o}_{\mathbb{P}}(1/\sqrt{N(1-\tau_N)}).$$

Here the v_n-uniform consistency of $\widehat{g}$ and $\widehat{\sigma}$ on K_0 and boundedness of $1/\sigma$ on the support of $\boldsymbol{X}$ were used, along with again $n^{1-a}/v_n^2 \to 0$, and the identity $N(1-\tau_N) = N^{1-a} \leq n^{1-a}$. Set then

$$\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} = \frac{1}{\lfloor N(1-\tau_N)\rfloor} \sum_{i=1}^{\lfloor N(1-\tau_N)\rfloor} \log \frac{\widehat{e}^{(n)}_{N-i+1,N}}{\widehat{e}^{(n)}_{N-\lfloor N(1-\tau_N)\rfloor,N}}$$
$$\text{and } \widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor} = \frac{1}{\lfloor N(1-\tau_N)\rfloor} \sum_{i=1}^{\lfloor N(1-\tau_N)\rfloor} \log \frac{e_{N-i+1,N}}{e_{N-\lfloor N(1-\tau_N)\rfloor,N}}.$$

By Lemma C.6(i),

$$\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} = \widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor} + \mathrm{o}_{\mathbb{P}}(1/\sqrt{N(1-\tau_N)})$$

and therefore

$$\sqrt{N(1-\tau_N)}(\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) = \sqrt{N(1-\tau_N)}(\widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) + \mathrm{o}_{\mathbb{P}}(1). \tag{55}$$

Let further

$$\widehat{\xi}_{\tau_N}(e) = \arg\min_{u\in\mathbb{R}} \sum_{i=1}^{N} \eta_{\tau_N}(\widehat{e}^{(n)}_i - u) \text{ and } \widetilde{\xi}_{\tau_N}(e) = \arg\min_{u\in\mathbb{R}} \sum_{i=1}^{N} \eta_{\tau_N}(e_i - u)$$

along with

$$\psi_N(u) = \frac{1}{2\xi^2_{\tau_N}(e)} \sum_{i=1}^{N} \left[\eta_{\tau_N}\left(e_i - \xi_{\tau_N}(e) - \frac{u\xi_{\tau_N}(e)}{\sqrt{N(1-\tau_N)}} \right) - \eta_{\tau_N}(e_i - \xi_{\tau_N}(e)) \right]$$
$$\text{and } \chi_N(u) = \frac{1}{2\xi^2_{\tau_N}(e)} \sum_{i=1}^{N} \left[\eta_{\tau_N}\left(\widehat{e}^{(n)}_i - \xi_{\tau_N}(e) - \frac{u\xi_{\tau_N}(e)}{\sqrt{N(1-\tau_N)}} \right) - \eta_{\tau_N}(\widehat{e}^{(n)}_i - \xi_{\tau_N}(e)) \right].$$

Lemma C.5 entails $\chi_N(u) = \psi_N(u) + \mathrm{o}_{\mathbb{P}}(1)$. Recall the notation $\varphi_\tau(y) = |\tau - \mathbb{1}\{y \leq 0\}|y$ and write, as in the proof of Theorem 2 in [9], $\psi_N(u) = -u\mathcal{T}_{1,N} + \mathcal{T}_{2,N}(u)$ with

$$\mathcal{T}_{1,N} = \frac{1}{\sqrt{N(1-\tau_N)}} \sum_{i=1}^{N} \frac{1}{\xi_{\tau_N}(e)} \varphi_{\tau_N}(e_i - \xi_{\tau_N}(e))$$

and

$$\mathcal{T}_{2,N}(u)$$
$$= -\frac{1}{\xi^2_{\tau_N}(e)} \sum_{i=1}^{N} \int_0^{u\xi_{\tau_N}(e)/\sqrt{N(1-\tau_N)}} (\varphi_{\tau_N}(e_i - \xi_{\tau_N}(e) - z) - \varphi_{\tau_N}(e_i - \xi_{\tau_N}(e)))\, dz.$$

The distribution of the e_i, $1 \leq i \leq N$, given $N = m$, is the distribution of m independent copies of e. Using the arguments of the proof of Theorem 2 in [9] and Lemma C.4(i) and (ii), we obtain $\mathcal{T}_{1,N} = \mathrm{O}_{\mathbb{P}}(1)$ and $\mathcal{T}_{2,N}(u) \xrightarrow{\mathbb{P}} u^2/2\gamma$. It follows that

$$\chi_N(u) = \psi_N(u) + \mathrm{o}_{\mathbb{P}}(1) = \frac{u^2}{2\gamma} - u\mathcal{T}_{1,N} + \mathrm{o}_{\mathbb{P}}(1).$$

Conclude, by the basic corollary on p.2 in [26], that the minimisers of χ_N and ψ_N are both only a $\mathrm{o}_{\mathbb{P}}(1)$ away from the minimiser of the right-hand side, and thus only a $\mathrm{o}_{\mathbb{P}}(1)$ away from each other. This can be rephrased as

$$\sqrt{N(1-\tau_N)} \left(\frac{\widehat{\xi}_{\tau_N}(e)}{\xi_{\tau_N}(e)} - 1 \right) = \sqrt{N(1-\tau_N)} \left(\frac{\widetilde{\xi}_{\tau_N}(e)}{\xi_{\tau_N}(e)} - 1 \right) + \mathrm{o}_{\mathbb{P}}(1). \tag{56}$$

Finally, the distribution of the pair $(\widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor}, \widetilde{\xi}_{\tau_N}(e))$ given $N = m$ is equal to the distribution of their counterparts $(\breve{\gamma}_{\lfloor m(1-\tau_m)\rfloor}, \breve{\xi}_{\tau_m}(e))$ based on m independent copies of e. Combine then Theorem 3 in [11], which provides the bivariate asymptotic distribution of $(\breve{\gamma}_{\lfloor m(1-\tau_m)\rfloor}, \breve{\xi}_{\tau_m}(e))$, with Lemma C.4(i) to get

$$\sqrt{N(1-\tau_N)}\left(\widetilde{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma, \frac{\widetilde{\xi}_{\tau_N}(e)}{\xi_{\tau_N}(e)} - 1\right) \xrightarrow{d} \mathcal{N}(\mathcal{B}(\rho, p), \mathcal{V}(\gamma)) \tag{57}$$

with $\mathcal{B}(\rho, p) = (p^{-\rho}\lambda/(1-\rho), 0)$ (recall that e satisfies condition $\mathcal{C}_2(\gamma, \rho, p^{-\rho}A)$) and

$$\mathcal{V}(\gamma) = \begin{pmatrix} \gamma^2 & \dfrac{\gamma^3(\gamma^{-1}-1)^\gamma}{(1-\gamma)^2} \\ \dfrac{\gamma^3(\gamma^{-1}-1)^\gamma}{(1-\gamma)^2} & \dfrac{2\gamma^3}{1-2\gamma} \end{pmatrix}.$$

Combining (53), (54), (55), (56), (57) with the delta method completes the proof of (i).

(ii) Define

$$\widehat{\xi}^{\star}_{\tau'_N}(\varepsilon) := \left(\frac{1-\tau'_N}{1-\tau_N}\right)^{-\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}} \left(\frac{N}{N_0}\right)^{\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}} \widehat{\xi}_{\tau_N}(e)$$

so that $\widehat{\xi}^{\star}_{\tau'_N}(Y|\boldsymbol{x}) = \widehat{g}(\boldsymbol{x}) + \widehat{\sigma}(\boldsymbol{x})\widehat{\xi}^{\star}_{\tau'_N}(\varepsilon)$. Then

$$\begin{aligned}\frac{\widehat{\xi}^{\star}_{\tau'_N}(Y|\boldsymbol{x})}{\xi_{\tau'_N}(Y|\boldsymbol{x})} - 1 = &\left(\frac{\widehat{\xi}^{\star}_{\tau'_N}(\varepsilon)}{\xi_{\tau'_N}(\varepsilon)} - 1\right)(1 + \mathrm{o}_{\mathbb{P}}(1)) + \mathrm{o}_{\mathbb{P}}(|\widehat{g}(\boldsymbol{x}) - g(\boldsymbol{x})|) + \mathrm{O}_{\mathbb{P}}\left(\left|\frac{\widehat{\sigma}(\boldsymbol{x})}{\sigma(\boldsymbol{x})} - 1\right|\right) \\ &+ \mathrm{O}_{\mathbb{P}}(1/q_{\tau'_N}(\varepsilon)) + \mathrm{o}_{\mathbb{P}}(|A((1-\tau'_N)^{-1})|)\end{aligned}$$

by Lemma C.7(vi), the consistency assumption on $\widehat{g}$ and $\widehat{\sigma}$, and $N = N(n) \xrightarrow{\mathbb{P}} \infty$. Our bias conditions combined with the regular variation properties of $t \mapsto q_{1-t^{-1}}(\varepsilon)$ and $t \mapsto |A(t)|$ and the v_n-uniform consistency of $\widehat{g}$ and $\widehat{\sigma}$ on K_0 yield

$$\frac{\widehat{\xi}^{\star}_{\tau'_N}(Y|\boldsymbol{x})}{\xi_{\tau'_N}(Y|\boldsymbol{x})} - 1 = \left(\frac{\widehat{\xi}^{\star}_{\tau'_N}(\varepsilon)}{\xi_{\tau'_N}(\varepsilon)} - 1\right)(1 + \mathrm{o}_{\mathbb{P}}(1)) + \mathrm{o}_{\mathbb{P}}(1/\sqrt{N(1-\tau_N)}).$$

Since, from the proof of (i),

$$\begin{aligned}&\sqrt{N(1-\tau_N)}(\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) \xrightarrow{d} \mathcal{N}(p^{-\rho}\lambda/(1-\rho), \gamma^2) \\ \text{and } &\sqrt{N(1-\tau_N)}\left(\frac{\widehat{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1\right) = \mathrm{O}_{\mathbb{P}}(1),\end{aligned}$$

a direct adaptation of the proof of Theorem 5 of [11] produces

$$\begin{aligned}\frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau'_N)]}\left(\frac{\widehat{\xi}^{\star}_{\tau'_N}(\varepsilon)}{\xi_{\tau'_N}(\varepsilon)} - 1\right) &= \sqrt{N(1-\tau_N)}(\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor} - \gamma) + \mathrm{o}_{\mathbb{P}}(1) \\ &\xrightarrow{d} \mathcal{N}\left(p^{-\rho}\frac{\lambda}{1-\rho}, \gamma^2\right).\end{aligned}$$

We omit the details. □

Proof of Theorem 3.3. (i) Write first

$$\begin{aligned}
&\sqrt{n(1-\tau_n)}\left(\frac{\widehat{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right)\\
&=\frac{\xi_{\tau_n}(\varepsilon)}{\sum_{j=1}^p\phi_jY_{n+1-j}+\sum_{j=1}^q\theta_j\varepsilon_{n+1-j}+\xi_{\tau_n}(\varepsilon)}\times\sqrt{n(1-\tau_n)}\left(\frac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)\\
&+\sqrt{n(1-\tau_n)}\frac{\sum_{j=1}^p(\widehat{\phi}_{j,n}-\phi_j)Y_{n+1-j}}{\sum_{j=1}^p\phi_jY_{n+1-j}+\sum_{j=1}^q\theta_j\varepsilon_{n+1-j}+\xi_{\tau_n}(\varepsilon)}\\
&+\sqrt{n(1-\tau_n)}\frac{\sum_{j=1}^q(\widehat{\theta}_{j,n}-\theta_j)\varepsilon_{n+1-j}}{\sum_{j=1}^p\phi_jY_{n+1-j}+\sum_{j=1}^q\theta_j\varepsilon_{n+1-j}+\xi_{\tau_n}(\varepsilon)}\\
&+\sqrt{n(1-\tau_n)}\frac{\sum_{j=1}^q\widehat{\theta}_{j,n}(\widehat{\varepsilon}^{(n)}_{n+1-j}-\varepsilon_{n+1-j})}{\sum_{j=1}^p\phi_jY_{n+1-j}+\sum_{j=1}^q\theta_j\varepsilon_{n+1-j}+\xi_{\tau_n}(\varepsilon)}.
\end{aligned}$$

To control the gap between residuals and unobserved innovations we rewrite the ARMA model in vector form, namely as $\boldsymbol{Y}_{t,p}=\boldsymbol{A}\boldsymbol{Y}_{t-1,p}-\boldsymbol{B}\boldsymbol{\varepsilon}_{t-1,q}+\boldsymbol{\varepsilon}_{t,q}$ with

$$\boldsymbol{Y}_{t,p}=\begin{pmatrix}Y_t\\Y_{t-1}\\\vdots\\Y_{t-p+1}\end{pmatrix}\text{ and }\boldsymbol{A}=\begin{pmatrix}\phi_1&\cdots&\cdots&\cdots&\phi_p\\1&0&\cdots&\cdots&0\\0&1&\cdots&\cdots&0\\\vdots&\ddots&\ddots&\ddots&\vdots\\0&\cdots&\cdots&1&0\end{pmatrix},$$

$$\boldsymbol{\varepsilon}_{t,q}=\begin{pmatrix}\varepsilon_t\\\varepsilon_{t-1}\\\vdots\\\varepsilon_{t-q+1}\end{pmatrix}\text{ and }\boldsymbol{B}=\begin{pmatrix}-\theta_1&\cdots&\cdots&\cdots&-\theta_q\\1&0&\cdots&\cdots&0\\0&1&\cdots&\cdots&0\\\vdots&\ddots&\ddots&\ddots&\vdots\\0&\cdots&\cdots&1&0\end{pmatrix}.$$

Set $r=\max(p,q)$. Since $\widehat{\varepsilon}^{(n)}_t=Y_t-\sum_{j=1}^p\widehat{\phi}_{j,n}Y_{t-j}-\sum_{j=1}^q\widehat{\theta}_{j,n}\widehat{\varepsilon}^{(n)}_{t-j}$ for $r+1\le t\le n$, we have $\boldsymbol{Y}_{t,p}=\widehat{\boldsymbol{A}}_n\boldsymbol{Y}_{t-1,p}-\widehat{\boldsymbol{B}}_n\widehat{\boldsymbol{\varepsilon}}^{(n)}_{t-1,q}+\widehat{\boldsymbol{\varepsilon}}^{(n)}_{t,q}$, where the notation is defined by replacing the ε_t, ϕ_j and θ_j by the $\widehat{\varepsilon}^{(n)}_t$, $\widehat{\phi}_{j,n}$ and $\widehat{\theta}_{j,n}$. It follows that for such t

$$\begin{aligned}
\widehat{\boldsymbol{\varepsilon}}^{(n)}_{t,q}-\boldsymbol{\varepsilon}_{t,q}&=(\boldsymbol{A}-\widehat{\boldsymbol{A}}_n)\boldsymbol{Y}_{t-1,p}-(\boldsymbol{B}-\widehat{\boldsymbol{B}}_n)\widehat{\boldsymbol{\varepsilon}}^{(n)}_{t-1,q}+\boldsymbol{B}(\widehat{\boldsymbol{\varepsilon}}^{(n)}_{t-1,q}-\boldsymbol{\varepsilon}_{t-1,q})\\
&=\sum_{j=1}^{t-r}\boldsymbol{B}^{j-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}}_n)\boldsymbol{Y}_{t-j,p}-\sum_{j=1}^{t-r}\boldsymbol{B}^{j-1}(\boldsymbol{B}-\widehat{\boldsymbol{B}}_n)\boldsymbol{\varepsilon}_{t-j,q}\\
&-\sum_{j=1}^{t-r}\boldsymbol{B}^{j-1}(\boldsymbol{B}-\widehat{\boldsymbol{B}}_n)(\widehat{\boldsymbol{\varepsilon}}^{(n)}_{t-j,q}-\boldsymbol{\varepsilon}_{t-j,q})-\boldsymbol{B}^{t-r}\boldsymbol{\varepsilon}_{r,q}
\end{aligned}\tag{58}$$

because $\widehat{\boldsymbol{\varepsilon}}^{(n)}_{r,q}=0$. Observe now that by causality of $(Y_t)_{t\in\mathbb{Z}}$, the Y_t have the linear representation $Y_t=\sum_{j=0}^\infty\psi_j\varepsilon_{t-j}$, and it is a consequence of the arguments in the proof of Theorem 3.1.1 in [4] that the ψ_j define a summable series and decay geometrically fast, *i.e.* $|\psi_j|\le C\,R^j$ for real constants $C>0$ and $R\in(0,1)$. Write, for $1\le t\le n$,

$$|Y_t|\le\sum_{j=0}^{t-1}|\psi_j||\varepsilon_{t-j}|+\sum_{j=t}^{\infty}|\psi_j||\varepsilon_{t-j}|\le\left(\sum_{j=0}^{\infty}|\psi_j|\right)\max_{1\le t\le n}|\varepsilon_t|+CR\sum_{l=0}^{\infty}R^l|\varepsilon_{-l}|.$$

The last sum on the right-hand side is finite with probability 1 because ε has a finite first moment. Conclude that $\max_{1\le t\le n}|Y_t|=\mathrm{O}_{\mathbb{P}}(1+\max_{1\le t\le n}|\varepsilon_t|)$. Since the ε_t are independent and satisfy $\mathcal{C}_1(\gamma)$, we find

$$\max_{1\le t\le n}|\varepsilon_t|=\mathrm{O}_{\mathbb{P}}(n^{\gamma+\iota})\ \text{ and then }\ \max_{1\le t\le n}|Y_t|=\mathrm{O}_{\mathbb{P}}(n^{\gamma+\iota})\ \text{ for any }\iota>0,\tag{59}$$

by condition $\mathbb{P}(\varepsilon > x)/\mathbb{P}(|\varepsilon| > x) \to \ell \in (0,1]$ as $x \to \infty$, combined with Theorem 1.1.6 and Lemma 1.2.9 in [13], and Potter bounds (see *e.g.* Proposition B.1.9.5 in [13]). Notice now that $\boldsymbol{B}$ is essentially the companion matrix of the polynomial $Q(z) = 1 + \sum_{j=1}^{q} \theta_j z^j$. It is a standard exercise in linear algebra to show that $\boldsymbol{B}$ has characteristic polynomial

$$\det(\lambda \mathrm{I}_p - \boldsymbol{B}) = \lambda^q + \sum_{j=1}^{q} \theta_j \lambda^{q-j} = \lambda^q Q(1/\lambda).$$

Since Q has no root z such that $|z| \leq 1$, all eigenvalues of $\boldsymbol{B}$ must then have a modulus smaller than 1, *i.e.* its spectral radius $\rho(\boldsymbol{B})$ is smaller than 1. Let $\|\cdot\|$ denote indifferently the supremum norm on $\mathbb{R}^d$ spaces and the induced operator norm on square matrices, and recall that $\|\boldsymbol{B}^j\|^{1/j} \to \rho(\boldsymbol{B})$ as $j \to \infty$ (this is in fact true for any operator norm), which means in particular that the series $\sum_{j\geq 0} \|\boldsymbol{B}^j\|$ is summable. Defining $\widehat{\varepsilon}_1^{(n)} = \cdots = \widehat{\varepsilon}_{r-q}^{(n)} = 0$ for the sake of convenience, we obtain

$$\begin{aligned}
\max_{1\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t| &\leq \max_{r+1\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t| + \max_{1\leq t\leq r} |\varepsilon_t| \\
&\leq \|\boldsymbol{A} - \widehat{\boldsymbol{A}}_n\| \sum_{j=0}^{\infty} \|\boldsymbol{B}^j\| \max_{1\leq t\leq n} |Y_t| + \|\boldsymbol{B} - \widehat{\boldsymbol{B}}_n\| \sum_{j=0}^{\infty} \|\boldsymbol{B}^j\| \max_{1\leq t\leq n} |\varepsilon_t| \\
&\quad + \|\boldsymbol{B} - \widehat{\boldsymbol{B}}_n\| \sum_{j=0}^{\infty} \|\boldsymbol{B}^j\| \max_{1\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t| + \left(1 + \sup_{j\geq 0} \|\boldsymbol{B}^j\|\right) \max_{1\leq t\leq r} |\varepsilon_t|.
\end{aligned}$$

By $\sqrt{n}-$consistency of the $\widehat{\phi}_{j,n}$ and $\widehat{\theta}_{j,n}$, $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}_n\| = \mathrm{O}_{\mathbb{P}}(n^{-1/2})$ and $\|\boldsymbol{B} - \widehat{\boldsymbol{B}}_n\| = \mathrm{O}_{\mathbb{P}}(n^{-1/2})$. Isolate then $\max_{1\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t|$ to conclude that

$$\max_{1\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t| = \mathrm{O}_{\mathbb{P}}\left(1 + n^{-1/2}\left[\max_{1\leq t\leq n} |Y_t| + \max_{1\leq t\leq n} |\varepsilon_t|\right]\right) = \mathrm{O}_{\mathbb{P}}(1)$$

by (59) and the assumption $\gamma < 1/2$. We now use (58) again, this time to control $\max_{t_n\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t|$ to apply Theorem 2.1 (for the sample size $n - t_n + 1 = n(1 + \mathrm{o}(1))$, since the estimator $\widehat{\xi}_{\tau_n}(\varepsilon)$ is based upon the last $n - t_n + 1$ residuals). For $t \geq t_n \to \infty$, $\|\boldsymbol{B}^{t-r}\boldsymbol{\varepsilon}_{r,q}\| \leq \|\boldsymbol{B}^{t-r}\|\|\boldsymbol{\varepsilon}_{r,q}\|$ and $t - r \geq t_n/2$ for n large enough; hence, by (58), the bound

$$\max_{t_n\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t| = \mathrm{O}_{\mathbb{P}}\left(n^{-1/2}\left[1 + \max_{1\leq t\leq n} |Y_t| + \max_{1\leq t\leq n} |\varepsilon_t|\right] + \sup_{j\geq t_n/2} \|\boldsymbol{B}^j\|\right).$$

Since $\|\boldsymbol{B}^j\|^{1/j} \to \rho(\boldsymbol{B}) \in [0,1)$ as $j \to \infty$, we have for n large enough

$$\max_{t_n\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t| = \mathrm{O}_{\mathbb{P}}\left(n^{-1/2}\left[1 + \max_{1\leq t\leq n} |Y_t| + \max_{1\leq t\leq n} |\varepsilon_t|\right] + (1-\kappa)^{t_n}\right)$$

for some $\kappa \in (0,1)$. We have $\sqrt{n}(1-\kappa)^{t_n} \to 0$ because $t_n/\log n \to \infty$. Conclude that

$$\max_{t_n\leq t\leq n} \frac{|\widehat{\varepsilon}_t^{(n)} - \varepsilon_t|}{1 + |\varepsilon_t|} = \mathrm{O}_{\mathbb{P}}\left(\max_{t_n\leq t\leq n} |\widehat{\varepsilon}_t^{(n)} - \varepsilon_t|\right) = \mathrm{O}_{\mathbb{P}}(n^{\gamma - 1/2 + \iota}) \quad \text{for all } \iota > 0$$

and therefore

$$\sqrt{n(1-\tau_n)} \max_{t_n\leq t\leq n} \frac{|\widehat{\varepsilon}_t^{(n)} - \varepsilon_t|}{1 + |\varepsilon_t|} = \mathrm{o}_{\mathbb{P}}(1).$$

Complete the proof by combining the $\sqrt{n}-$consistency of the estimators $\widehat{\phi}_{j,n}$ and $\widehat{\theta}_{j,n}$, Lemma C.8 (an extension of Theorem 2.1 necessary here since at each step, the indices of the relevant ε_i may not be contained in those relevant to the previous step and thus, strictly speaking, we do not work with a single i.i.d. sequence) and the convergence $\xi_{\tau_n}(\varepsilon) \to \infty$.

(ii) Set

$$\widehat{\xi}^{\star}_{\tau'_n}(\varepsilon) = \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_n}(\varepsilon)$$

and write

$$\begin{aligned}
&\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\left(\frac{\widehat{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right)\\
&= \frac{\xi_{\tau'_n}(\varepsilon)}{\sum_{j=1}^p \phi_j Y_{n+1-j} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \xi_{\tau'_n}(\varepsilon)} \times \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\left(\frac{\widehat{\xi}^{\star}_{\tau'_n}(\varepsilon)}{\xi_{\tau'_n}(\varepsilon)}-1\right)\\
&+ \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]} \times \frac{\sum_{j=1}^p (\widehat{\phi}_{j,n}-\phi_j) Y_{n+1-j}}{\sum_{j=1}^p \phi_j Y_{n+1-j} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \xi_{\tau'_n}(\varepsilon)}\\
&+ \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]} \times \frac{\sum_{j=1}^q (\widehat{\theta}_{j,n}-\theta_j) \varepsilon_{n+1-j}}{\sum_{j=1}^p \phi_j Y_{n+1-j} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \xi_{\tau'_n}(\varepsilon)}\\
&+ \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]} \times \frac{\sum_{j=1}^q \widehat{\theta}_{j,n} (\widehat{\varepsilon}^{(n)}_{n+1-j}-\varepsilon_{n+1-j})}{\sum_{j=1}^p \phi_j Y_{n+1-j} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \xi_{\tau'_n}(\varepsilon)}.
\end{aligned}$$

Combine then what was obtained in (i) with the first convergence in Theorem 2.3. □

Proof of Theorem 3.4. (i) Recall that $\widehat{\omega}_n$, the $\widehat{\alpha}_{j,n}$ and the $\widehat{\beta}_{j,n}$ are consistent estimators of (strictly) positive parameters, and thus are positive with arbitrarily high probability as $n \to \infty$. In what follows we implicitly work on this high probability event. Write

$$\begin{aligned}
\sqrt{n(1-\tau_n)}\left(\frac{\widehat{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right) &= \sqrt{n(1-\tau_n)}\left(\frac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}-1\right)\\
&\quad + \sqrt{n(1-\tau_n)}\left(\frac{\widehat{\sigma}_{n+1}}{\sigma_{n+1}}-1\right)\frac{\widehat{\xi}_{\tau_n}(\varepsilon)}{\xi_{\tau_n}(\varepsilon)}.
\end{aligned}$$

Define $r = \max(p,q)$. For any t with $r+1 \le t \le n$,

$$\frac{|\widehat{\varepsilon}^{(n)}_t - \varepsilon_t|}{1+|\varepsilon_t|} \le \left|\frac{\sigma_t}{\widehat{\sigma}^{(n)}_t}-1\right| = \left|\frac{\sigma_t^2-(\widehat{\sigma}^{(n)}_t)^2}{\widehat{\sigma}^{(n)}_t(\sigma_t+\widehat{\sigma}^{(n)}_t)}\right| \le \left|\frac{\sigma_t^2}{(\widehat{\sigma}^{(n)}_t)^2}-1\right|. \tag{60}$$

We focus on $|(\widehat{\sigma}^{(n)}_t)^2 - \sigma_t^2|$. Note that $\boldsymbol{v}_{t,p} = \boldsymbol{Z}_{t,q} + \boldsymbol{B}\boldsymbol{v}_{t-1,p}$ with

$$\boldsymbol{v}_{t,p} = \begin{pmatrix} \sigma_t^2 \\ \sigma_{t-1}^2 \\ \vdots \\ \sigma_{t-p+1}^2 \end{pmatrix}, \quad \boldsymbol{Z}_{t,q} = \begin{pmatrix} \omega + \sum_{j=1}^q \alpha_j Y_{t-j}^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{pmatrix} \beta_1 & \cdots & \cdots & \cdots & \beta_p \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}.$$

Similarly $\widehat{\boldsymbol{v}}^{(n)}_{t,p} = \widehat{\boldsymbol{Z}}^{(n)}_{t,q} + \widehat{\boldsymbol{B}}_n \widehat{\boldsymbol{v}}^{(n)}_{t-1,p}$ where the notation is defined by replacing the σ_t^2, ω, the α_j and β_j by the $(\widehat{\sigma}^{(n)}_t)^2$, $\widehat{\omega}_n$, the $\widehat{\alpha}_{j,n}$ and $\widehat{\beta}_{j,n}$. For $r+1 \le t \le n$ then,

$$\boldsymbol{v}_{t,p} = \sum_{j=0}^{t-r-1} \boldsymbol{B}^j \boldsymbol{Z}_{t-j,q} + \boldsymbol{B}^{t-r}\boldsymbol{v}_{r,p}, \quad \widehat{\boldsymbol{v}}^{(n)}_{t,p} = \sum_{j=0}^{t-r-1} \widehat{\boldsymbol{B}}^j_n \widehat{\boldsymbol{Z}}^{(n)}_{t-j,q} + \widehat{\boldsymbol{B}}^{t-r}_n \widehat{\boldsymbol{v}}^{(n)}_{r,p}$$

and therefore

$$\begin{aligned}
&\widehat{\boldsymbol{v}}^{(n)}_{t,p} - \boldsymbol{v}_{t,p}\\
&= \sum_{j=0}^{t-r-1} \widehat{\boldsymbol{B}}^j_n (\widehat{\boldsymbol{Z}}^{(n)}_{t-j,q} - \boldsymbol{Z}_{t-j,q}) + \sum_{j=0}^{t-r-1} (\widehat{\boldsymbol{B}}^j_n - \boldsymbol{B}^j)\boldsymbol{Z}_{t-j,q} + \widehat{\boldsymbol{B}}^{t-r}_n \widehat{\boldsymbol{v}}^{(n)}_{r,p} - \boldsymbol{B}^{t-r}\boldsymbol{v}_{r,p}.
\end{aligned}$$

This readily provides

$$(\widehat{\sigma}_t^{(n)})^2 = \sum_{j=0}^{t-r-1} \widehat{\boldsymbol{B}}_n^j(1,1) \left(\widehat{\omega}_n + \sum_{i=1}^{q} \widehat{\alpha}_{i,n} Y_{t-j-i}^2 \right) + (\widehat{\boldsymbol{B}}_n^{t-r} \widehat{\boldsymbol{v}}_{r,p}^{(n)})(1)$$

where $\boldsymbol{u}(1)$ denotes the first element of a vector $\boldsymbol{u}$ and $\boldsymbol{A}(1,1)$ the top left element of a matrix $\boldsymbol{A}$, and similarly

$$\begin{aligned}(\widehat{\sigma}_t^{(n)})^2 - \sigma_t^2 = & \sum_{j=0}^{t-r-1} \widehat{\boldsymbol{B}}_n^j(1,1) \left(\widehat{\omega}_n - \omega + \sum_{i=1}^{q} (\widehat{\alpha}_{i,n} - \alpha_i) Y_{t-j-i}^2 \right) \\ & + \sum_{j=0}^{t-r-1} (\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)) \left(\omega + \sum_{i=1}^{q} \alpha_i Y_{t-j-i}^2 \right) \\ & + (\widehat{\boldsymbol{B}}_n^{t-r} \widehat{\boldsymbol{v}}_{r,p}^{(n)} - \boldsymbol{B}^{t-r} \boldsymbol{v}_{r,p})(1). \end{aligned} \tag{61}$$

We compare each term in (61) to $(\widehat{\sigma}_t^{(n)})^2$. First of all

$$\begin{aligned} & \frac{1}{(\widehat{\sigma}_t^{(n)})^2} \left| \sum_{j=0}^{t-r-1} \widehat{\boldsymbol{B}}_n^j(1,1) \left(\widehat{\omega}_n - \omega + \sum_{i=1}^{q} (\widehat{\alpha}_{i,n} - \alpha_i) Y_{t-j-i}^2 \right) \right| \\ & \leq \left| \frac{\widehat{\omega}_n - \omega}{\widehat{\omega}_n} \right| + \sum_{i=1}^{q} \left| \frac{\widehat{\alpha}_{i,n} - \alpha_i}{\widehat{\alpha}_{i,n}} \right| = \mathrm{O}_{\mathbb{P}}(n^{-1/2}). \end{aligned} \tag{62}$$

Now $\boldsymbol{B}$ and $\widehat{\boldsymbol{B}}_n$ are positive matrices, so that if $\kappa_n := \max_{1 \leq i \leq p} |\widehat{\beta}_{i,n} - \beta_i|$, clearly $\widehat{\boldsymbol{B}}_n \leq (1 + \kappa_n)\boldsymbol{B}$ elementwise and thus $\widehat{\boldsymbol{B}}_n^j \leq (1 + \kappa_n)^j \boldsymbol{B}^j$ elementwise for any j. In particular $\widehat{\boldsymbol{B}}_n^j(1,1) \leq (1 + \kappa_n)^j \boldsymbol{B}^j(1,1)$ and likewise $\boldsymbol{B}^j(1,1) \leq (1 + \kappa_n)^j \widehat{\boldsymbol{B}}_n^j(1,1)$. Hence the bound

$$\begin{aligned} |\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)| & \leq [(1 + \kappa_n)^j - 1] \max(\boldsymbol{B}^j(1,1), \widehat{\boldsymbol{B}}_n^j(1,1)) \\ & \leq j\kappa_n (1 + \kappa_n)^{j-1} \max(\boldsymbol{B}^j(1,1), \widehat{\boldsymbol{B}}_n^j(1,1)) \\ & \leq j\kappa_n (1 + \kappa_n)^{2j-1} \boldsymbol{B}^j(1,1). \end{aligned} \tag{63}$$

Like in the proof of Theorem 3.3, let $\|\cdot\|$ denote indifferently the supremum norm on $\mathbb{R}^d$ spaces and the induced operator norm on square matrices. Notice that $|\boldsymbol{B}^j(1,1)| \leq \|\boldsymbol{B}^j\|$; since $\|\boldsymbol{B}^j\|^{1/j} \to \rho(\boldsymbol{B}) \in [0,1)$ as $j \to \infty$ (to check that indeed the spectral radius $\rho(\boldsymbol{B}) \in [0,1)$, use Corollary 2.2 in [16]) and $\kappa_n = \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})$, we have

$$\sum_{j=0}^{\infty} |\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)| = \mathrm{O}_{\mathbb{P}}(\kappa_n) = \mathrm{O}_{\mathbb{P}}(n^{-1/2}).$$

Recalling that $(\widehat{\sigma}_t^{(n)})^2 \geq \widehat{\omega}_n \xrightarrow{\mathbb{P}} \omega > 0$, we therefore obtain

$$\max_{r+1 \leq t \leq n} \frac{1}{(\widehat{\sigma}_t^{(n)})^2} \left| \sum_{j=0}^{t-r-1} (\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)) \right| = \mathrm{O}_{\mathbb{P}}(n^{-1/2}). \tag{64}$$

Next we write, for any $i \in \{1, \ldots, q\}$,

$$\frac{1}{(\widehat{\sigma}_t^{(n)})^2} \left| \sum_{j=0}^{t-r-1} (\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)) \alpha_i Y_{t-j-i}^2 \right| \leq \sum_{j=0}^{t-r-1} \frac{|\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)| \alpha_i Y_{t-j-i}^2}{\widehat{\omega}_n + \widehat{\boldsymbol{B}}_n^j(1,1) \widehat{\alpha}_{i,n} Y_{t-j-i}^2}.$$

Similarly to (63), $|\widehat{\boldsymbol{B}}_n^j(1,1) - \boldsymbol{B}^j(1,1)| \le j\kappa_n(1+\kappa_n)^{2j-1}\widehat{\boldsymbol{B}}_n^j(1,1)$. Thus, for any $s>0$,

$$\begin{aligned}
&\frac{1}{(\widehat{\sigma}_t^{(n)})^2}\left|\sum_{j=0}^{t-r-1}(\widehat{\boldsymbol{B}}_n^j(1,1)-\boldsymbol{B}^j(1,1))\alpha_i Y_{t-j-i}^2\right| \\
&\le \kappa_n\frac{\alpha_i}{\widehat{\alpha}_{i,n}}\sum_{j=0}^{t-r-1} j(1+\kappa_n)^{2j-1}\frac{\widehat{\boldsymbol{B}}_n^j(1,1)\widehat{\alpha}_{i,n}Y_{t-j-i}^2/\widehat{\omega}_n}{1+\widehat{\boldsymbol{B}}_n^j(1,1)\widehat{\alpha}_{i,n}Y_{t-j-i}^2/\widehat{\omega}_n} \\
&\le \kappa_n\frac{\alpha_i}{\widehat{\alpha}_{i,n}}\sum_{j=0}^{t-r-1} j(1+\kappa_n)^{2j-1}\left(\frac{\widehat{\boldsymbol{B}}_n^j(1,1)\widehat{\alpha}_{i,n}Y_{t-j-i}^2}{\widehat{\omega}_n}\right)^s \\
&\le \kappa_n\times\frac{\alpha_i}{\widehat{\alpha}_{i,n}}\left(\frac{\widehat{\alpha}_{i,n}}{\widehat{\omega}_n}\right)^s\times\sum_{j=0}^{t-r-1} j(1+\kappa_n)^{2j-1}\left((1+\kappa_n)^j\boldsymbol{B}^j(1,1)\right)^s Y_{t-j-i}^{2s}
\end{aligned}$$

where the inequality $x/(1+x)\le x^s$, valid for any s and $x>0$, was used. Because $|\boldsymbol{B}^j(1,1)|\le\|\boldsymbol{B}^j\|$ and $\|\boldsymbol{B}^j\|^{1/j}\to\rho(\boldsymbol{B})\in[0,1)$ as $j\to\infty$, as well as $\kappa_n\to 0$ in probability, we have

$$\sum_{j=0}^{\infty} j(1+\kappa_n)^{2j-1}((1+\kappa_n)^j\boldsymbol{B}^j(1,1))^s<\infty$$

with arbitrarily high probability as $n\to\infty$. Hence the bound

$$\max_{r+1\le t\le n}\frac{1}{(\widehat{\sigma}_t^{(n)})^2}\left|\sum_{j=0}^{t-r-1}(\widehat{\boldsymbol{B}}_n^j(1,1)-\boldsymbol{B}^j(1,1))\alpha_i Y_{t-j-i}^2\right| = \mathrm{O}_{\mathbb{P}}\left(n^{-1/2}\max_{1\le t\le n}Y_t^{2s}\right)$$

valid for any $s>0$. Recall that there is $s_0>0$ such that $\mathbb{E}(Y_1^{2s_0})<\infty$ (see Corollary 2.3 p.36 in [16]). Using the identity

$$\mathbb{E}(Y_1^{2s_0}) = \int_0^\infty \mathbb{P}(Y_1^{2s_0}>y)\,dy<\infty$$

and noting that the function $y\mapsto\mathbb{P}(Y_1^{2s_0}>y)$ is nonnegative and nonincreasing, it is a standard exercise to show that $\mathbb{P}(Y_1^{2s_0}>y)=\mathrm{o}(y^{-1})$ as $y\to\infty$. Conclude that, for any $s\le s_0$, $n\,\mathbb{P}(Y_1^{2s}>n^{s/s_0}) = n\,\mathbb{P}(Y_1^{2s_0}>n)=\mathrm{o}(1)$, and then that

$$\mathbb{P}\left(\max_{1\le t\le n}Y_t^{2s}>n^{s/s_0}\right)\le n\,\mathbb{P}(Y_1^{2s}>n^{s/s_0})=\mathrm{o}(1),$$

of which a consequence is that $\max_{1\le t\le n}Y_t^{2s}=\mathrm{O}_{\mathbb{P}}(n^{s/s_0})$ for any $s\le s_0$. In particular, since s can be chosen arbitrarily small,

$$\max_{r+1\le t\le n}\frac{1}{(\widehat{\sigma}_t^{(n)})^2}\left|\sum_{j=0}^{t-r-1}(\widehat{\boldsymbol{B}}_n^j(1,1)-\boldsymbol{B}^j(1,1))\alpha_i Y_{t-j-i}^2\right| = \mathrm{O}_{\mathbb{P}}(n^{\iota-1/2})\ \text{ for all } \iota>0. \tag{65}$$

Finally, for $t\ge t_n$ and n large enough,

$$\begin{aligned}
&\max_{t_n\le t\le n}\frac{1}{(\widehat{\sigma}_t^{(n)})^2}|(\widehat{\boldsymbol{B}}_n^{t-r}\widehat{\boldsymbol{v}}_{r,p}^{(n)}-\boldsymbol{B}^{t-r}\boldsymbol{v}_{r,p})(1)| \\
&\le\frac{1}{\widehat{\omega}_n}\sup_{j\ge t_n/2}\left\{\|\widehat{\boldsymbol{B}}_n^j\|+\|\boldsymbol{B}^j\|\right\}\max_{r-p+1\le t\le r}\left\{\widehat{\sigma}_t^{(n)}+\sigma_t\right\} = \mathrm{O}_{\mathbb{P}}\left(\sup_{j\ge t_n/2}\left\{\|\widehat{\boldsymbol{B}}_n^j\|+\|\boldsymbol{B}^j\|\right\}\right)
\end{aligned}$$

by consistency of $\widehat{\omega}_n$, definition of $\widehat{\sigma}_{r-p+1}^{(n)},\dots,\widehat{\sigma}_r^{(n)}$ and finiteness of at least a fractional moment of the σ_t (and hence finiteness of the σ_t with probability 1; see Corollary 2.3 p.36 in [16]). Besides, it

is a simple exercise in linear algebra to show that for a $d \times d$ matrix with nonnegative elements, $\|A\| = \max_{1\le i\le d} \sum_{j=1}^{d} A(i,j)$; consequently

$$\max_{t_n\le t\le n} \frac{1}{(\widehat{\sigma}_t^{(n)})^2} |(\widehat{\boldsymbol{B}}_n^{t-r}\widehat{\boldsymbol{v}}_{r,p}^{(n)} - \boldsymbol{B}^{t-r}\boldsymbol{v}_{r,p})(1)| = \mathrm{O}_{\mathbb{P}}\left(\sup_{j\ge t_n/2} \left\{ (1+\kappa_n)^j \|\boldsymbol{B}^j\| \right\} \right).$$

Recall that $\|\boldsymbol{B}^j\|^{1/j} \to \rho(\boldsymbol{B}) \in [0,1)$ as $j \to \infty$ and $\kappa_n \to 0$ in probability, so that

$$\max_{t_n\le t\le n} \frac{1}{(\widehat{\sigma}_t^{(n)})^2} |(\widehat{\boldsymbol{B}}_n^{t-r}\widehat{\boldsymbol{v}}_{r,p}^{(n)} - \boldsymbol{B}^{t-r}\boldsymbol{v}_{r,p})(1)| = \mathrm{O}_{\mathbb{P}}(n^{-1/2}) \tag{66}$$

because $t_n/\log n \to \infty$. Combine (60), (61), (62), (64), (65), (66) and recall that $\tau_n = 1 - n^{-a}$ to find

$$\sqrt{n(1-\tau_n)} \max_{t_n\le t\le n} \left| \frac{\sigma_t^2}{(\widehat{\sigma}_t^{(n)})^2} - 1 \right| \xrightarrow{\mathbb{P}} 0 \ \text{ and then } \ \sqrt{n(1-\tau_n)} \max_{t_n\le t\le n} \frac{|\widehat{\varepsilon}_t^{(n)} - \varepsilon_t|}{1+|\varepsilon_t|} \xrightarrow{\mathbb{P}} 0$$

by (60). The inequality $|\widehat{\sigma}_{n+1}/\sigma_{n+1} - 1| \le |\widehat{\sigma}_{n+1}^2/\sigma_{n+1}^2 - 1|$ and a similar argument yield

$$\sqrt{n(1-\tau_n)} \left| \frac{\widehat{\sigma}_{n+1}}{\sigma_{n+1}} - 1 \right| \xrightarrow{\mathbb{P}} 0.$$

Conclude by applying Lemma C.8 (for the sample size $s_n = n - t_n + 1 = n(1+\mathrm{o}(1))$, since the estimator $\widehat{\xi}_{\tau_n}(\varepsilon)$ is based upon the last $n - t_n + 1$ residuals; this array version of Theorem A.1 is necessary once again here).

(ii) Set

$$\widehat{\xi}^{\star}_{\tau_n'}(\varepsilon) = \left(\frac{1-\tau_n'}{1-\tau_n} \right)^{-\overline{\gamma}} \widehat{\xi}_{\tau_n}(\varepsilon)$$

and write

$$\begin{aligned}
&\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]} \left(\frac{\widehat{\xi}^{\star}_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau_n'}(Y_{n+1}\,|\,\mathcal{F}_n)} - 1 \right) \\
&= \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]} \left(\frac{\widehat{\xi}^{\star}_{\tau_n'}(\varepsilon)}{\xi_{\tau_n'}(\varepsilon)} - 1 \right) \\
&+ \frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]} \left(\frac{\widehat{\sigma}_{n+1}}{\sigma_{n+1}} - 1 \right) \frac{\widehat{\xi}^{\star}_{\tau_n'}(\varepsilon)}{\xi_{\tau_n'}(\varepsilon)}.
\end{aligned}$$

To conclude, combine (i) with the relationship $\widehat{\sigma}_{n+1}/\sigma_{n+1} = 1 + \mathrm{O}_{\mathbb{P}}(1/\sqrt{n(1-\tau_n)})$ and the first convergence in Theorem 2.3. □

Appendix E: Additional results on indirect estimators and their proofs

This section focuses on the indirect versions of our extreme expectile estimators. The first result is an analogue of Corollary 3.1 in the heteroscedastic linear regression model (M_1), for the indirect estimators $\widetilde{\xi}_{\tau_n}(Y|\boldsymbol{x})$ and $\widetilde{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x})$ defined as

$$\begin{aligned}
\widetilde{\xi}_{\tau_n}(Y|\boldsymbol{x}) &= \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{x} + (1 + \widehat{\boldsymbol{\theta}}^{\top}\boldsymbol{x})(\overline{\gamma}^{-1} - 1)^{-\overline{\gamma}} \widehat{\varepsilon}^{(n)}_{n-\lfloor n(1-\tau_n)\rfloor,n} \\
\text{and } \widetilde{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x}) &= \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{x} + (1 + \widehat{\boldsymbol{\theta}}^{\top}\boldsymbol{x}) \left(\frac{1-\tau_n'}{1-\tau_n} \right)^{-\overline{\gamma}} (\overline{\gamma}^{-1} - 1)^{-\overline{\gamma}} \widehat{\varepsilon}^{(n)}_{n-\lfloor n(1-\tau_n)\rfloor,n}.
\end{aligned}$$

Here $\overline{\gamma} = \widehat{\gamma}_{\lfloor n(1-\tau_n)\rfloor}$ is assumed to be the Hill estimator based on residuals, as in Section 2.2.

Corollary E.1. *Assume that the setup is that of the heteroscedastic linear model* (M_1)*. Suppose that* $\mathbb{E}|\varepsilon_-|^2 < \infty$*. Assume further that* ε *satisfies condition* $\mathcal{C}_2(\gamma, \rho, A)$ *with* $0 < \gamma < 1/2$*,* $\rho < 0$*, and that* $\tau_n, \tau_n' \uparrow 1$ *satisfy* (3) *and* (4)*. Then for any* $\boldsymbol{x} \in K$*,*

$$\sqrt{n(1-\tau_n)}\left(\frac{\widetilde{\xi}_{\tau_n}(Y|\boldsymbol{x})}{\xi_{\tau_n}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(\lambda\left[\frac{m(\gamma)}{1-\rho} - b(\gamma,\rho)\right], \gamma^2\left[1 + [m(\gamma)]^2\right]\right),$$

with the notation of Corollary 2.1, and

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]}\left(\frac{\widetilde{\xi}^{\star}_{\tau_n'}(Y|\boldsymbol{x})}{\xi_{\tau_n'}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right).$$

Proof of Corollary E.1. To obtain the first convergence, repeat the proof of Corollary 3.1, with $\widehat{\xi}_{\tau_n}(\varepsilon)$ replaced by $\widetilde{\xi}_{\tau_n}(\varepsilon) = (\widehat{\gamma}^{-1}_{\lfloor n(1-\tau_n)\rfloor} - 1)^{-\widehat{\gamma}_{\lfloor n(1-\tau_n)\rfloor}}\,\widehat{\varepsilon}^{(n)}_{n-\lfloor n(1-\tau_n)\rfloor,n}$, and apply Corollary 2.1 rather than Theorem 2.1. The second convergence is obtained by combining the first convergence with Theorem 2.3. □

The second result considers, in the context of the heteroscedastic single-index model (M_2), the indirect estimators $\widetilde{\xi}_{\tau_N}(Y|\boldsymbol{x})$ and $\widetilde{\xi}^{\star}_{\tau_N'}(Y|\boldsymbol{x})$ defined, for an $\boldsymbol{x} \in K_0$, as

$$\widetilde{\xi}_{\tau_N}(Y|\boldsymbol{x}) = \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{x}) + \widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{x})(\overline{\gamma}^{-1} - 1)^{-\overline{\gamma}}\widehat{\varepsilon}^{(n)}_{N-\lfloor N(1-\tau_N)\rfloor,N,K_0}$$

at the intermediate level, and

$$\widetilde{\xi}^{\star}_{\tau_N'}(Y|\boldsymbol{x}) = \widehat{g}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{x}) + \widehat{\sigma}_{h_n,t_n}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{x})\left(\frac{1-\tau_N'}{1-\tau_N}\right)^{-\overline{\gamma}}(\overline{\gamma}^{-1} - 1)^{-\overline{\gamma}}\widehat{\varepsilon}^{(n)}_{N-\lfloor N(1-\tau_N)\rfloor,N,K_0}$$

at the extreme level. Here $\overline{\gamma} = \widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}$ is assumed to be the Hill estimator based on the random number of residuals $\lfloor N(1-\tau_N)\rfloor$ where $N = \sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in K_0\}$.

Theorem E.1. *Work in model* (M_2)*. Assume that* ε *satisfies condition* $\mathcal{C}_2(\gamma, \rho, A)$ *with* $0 < \gamma < 1/2$ *and* $\rho < 0$ *and that the conditions of Proposition C.1 in Appendix C hold. Let* K_0 *be a compact subset of* K° *such that* $\mathbb{P}(\boldsymbol{X} \in K_0) > 0$*, and* $N = N(K_0, n)$*. In addition, suppose that the sequences* $\tau_n = 1 - n^{-a}$ *with* $a \in (1/5, 1)$ *and* $\tau_n' \uparrow 1$ *satisfy* (3) *and* (4)*. Then, for any* $\boldsymbol{x} \in K_0$*,*

$$\sqrt{N(1-\tau_N)}\left(\frac{\widetilde{\xi}_{\tau_N}(Y|\boldsymbol{x})}{\xi_{\tau_N}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(\lambda\left[\frac{m(\gamma)}{1-\rho} - b(\gamma,\rho)\right], \gamma^2\left[1 + [m(\gamma)]^2\right]\right),$$

with the notation of Corollary 2.1, and

$$\frac{\sqrt{N(1-\tau_N)}}{\log[(1-\tau_N)/(1-\tau_N')]}\left(\frac{\widetilde{\xi}^{\star}_{\tau_N'}(Y|\boldsymbol{x})}{\xi_{\tau_N'}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right).$$

Proof of Theorem E.1. Combine Corollary 2.1 with the de-conditioning Lemma C.4(i) to obtain

$$\sqrt{N(1-\tau_N)}\left(\frac{\widetilde{\xi}_{\tau_N}(\varepsilon)}{\xi_{\tau_N}(\varepsilon)} - 1\right) \xrightarrow{d} \mathcal{N}\left(\lambda\left[\frac{m(\gamma)}{1-\rho} - b(\gamma,\rho)\right], \gamma^2\left[1 + [m(\gamma)]^2\right]\right),$$

where $\widetilde{\xi}_{\tau_N}(\varepsilon) = (\widehat{\gamma}^{-1}_{\lfloor N(1-\tau_N)\rfloor} - 1)^{-\widehat{\gamma}_{\lfloor N(1-\tau_N)\rfloor}}\,\widehat{\varepsilon}^{(n)}_{N-\lfloor N(1-\tau_N)\rfloor,N,K_0}$. Complete the proof by following the final four lines of the proof of Theorem 3.1(i) (this crucially relies on the assumptions of Proposition C.1) and the proof of Theorem 3.1(ii). □

The third result focuses on the indirect estimators

$$\widetilde{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n) = \sum_{j=1}^{p}\widehat{\phi}_{j,n}Y_{n+1-j} + \sum_{j=1}^{q}\widehat{\theta}_{j,n}\widehat{\varepsilon}^{(n)}_{n+1-j} + (\overline{\gamma}^{-1}-1)^{-\overline{\gamma}}\overline{q}_{\tau_n}(\varepsilon)$$
$$\text{and } \widetilde{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n) = \sum_{j=1}^{p}\widehat{\phi}_{j,n}Y_{n+1-j} + \sum_{j=1}^{q}\widehat{\theta}_{j,n}\widehat{\varepsilon}^{(n)}_{n+1-j} + \left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\overline{\gamma}}(\overline{\gamma}^{-1}-1)^{-\overline{\gamma}}\overline{q}_{\tau_n}(\varepsilon)$$

in the ARMA(p,q) model (T_1). Here $\overline{q}_{\tau_n}(\varepsilon) = \widehat{\varepsilon}^{(n)}_{n-t_n+1-\lfloor(n-t_n+1)(1-\tau_n)\rfloor,n-t_n+1}$ is a top order statistic of the last $n-t_n+1$ residuals $\widehat{\varepsilon}^{(n)}_{t_n},\widehat{\varepsilon}^{(n)}_{t_n+1},\ldots,\widehat{\varepsilon}^{(n)}_{n}$, with $t_n/\log n\to\infty$ and $t_n/n\to 0$, and $\overline{\gamma}$ is assumed to be the Hill estimator based on these residuals.

Theorem E.2. *Work in model (T_1). Assume further that ε satisfies condition $\mathcal{C}_2(\gamma,\rho,A)$ with $0<\gamma<1/2$ and $\rho<0$, and that $\tau_n,\tau'_n\uparrow 1$ satisfy (3) and (4). If moreover $n^{2\gamma+\iota}(1-\tau_n)\to 0$ for some $\iota>0$, then*

$$\sqrt{n(1-\tau_n)}\left(\frac{\widetilde{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right)\xrightarrow{d}\mathcal{N}\left(\lambda\left[\frac{m(\gamma)}{1-\rho}-b(\gamma,\rho)\right],\gamma^2\left[1+[m(\gamma)]^2\right]\right),$$

with the notation of Corollary 2.1, and

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\left(\frac{\widetilde{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right)\xrightarrow{d}\mathcal{N}\left(\frac{\lambda}{1-\rho},\gamma^2\right).$$

Proof of Theorem E.2. Mimic the proof of Theorem 3.3, applying (an array version of) Corollary 2.1 rather than Lemma C.8. □

The fourth and final result gives the asymptotic properties of the indirect estimators

$$\widetilde{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n) = \widehat{\sigma}_{n+1}(\overline{\gamma}^{-1}-1)^{-\overline{\gamma}}\overline{q}_{\tau_n}(\varepsilon)$$
$$\text{and } \widetilde{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n) = \widehat{\sigma}_{n+1}\times\left(\frac{1-\tau'_n}{1-\tau_n}\right)^{-\overline{\gamma}}(\overline{\gamma}^{-1}-1)^{-\overline{\gamma}}\overline{q}_{\tau_n}(\varepsilon)$$

in the GARCH(p,q) model (T_2), where again $\overline{q}_{\tau_n}(\varepsilon) = \widehat{\varepsilon}^{(n)}_{n-t_n+1-\lfloor(n-t_n+1)(1-\tau_n)\rfloor,n-t_n+1}$ is a top order statistic of the last $n-t_n+1$ residuals $\widehat{\varepsilon}^{(n)}_{t_n},\widehat{\varepsilon}^{(n)}_{t_n+1},\ldots,\widehat{\varepsilon}^{(n)}_{n}$, with $t_n/\log n\to\infty$ and $t_n/n\to 0$, and $\overline{\gamma}$ is assumed to be the Hill estimator based on these residuals.

Theorem E.3. *Work in model (T_2). Assume further that ε satisfies condition $\mathcal{C}_2(\gamma,\rho,A)$ with $0<\gamma<1/2$ and $\rho<0$. Suppose also that $\tau_n,\tau'_n\uparrow 1$ satisfy (3) and (4) with $\tau_n=1-n^{-a}$ for $a\in(0,1)$. Then*

$$\sqrt{n(1-\tau_n)}\left(\frac{\widetilde{\xi}_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right)\xrightarrow{d}\mathcal{N}\left(\lambda\left[\frac{m(\gamma)}{1-\rho}-b(\gamma,\rho)\right],\gamma^2\left[1+[m(\gamma)]^2\right]\right),$$

with the notation of Corollary 2.1, and

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\left(\frac{\widetilde{\xi}^{\star}_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}{\xi_{\tau'_n}(Y_{n+1}\,|\,\mathcal{F}_n)}-1\right)\xrightarrow{d}\mathcal{N}\left(\frac{\lambda}{1-\rho},\gamma^2\right).$$

Proof of Theorem E.3. Mimic the proof of Theorem 3.4, applying (an array version of) Corollary 2.1 rather than Lemma C.8. □

Appendix F: Finite-sample study: Details on computational procedures and further finite-sample results

F.1. Optimal choice of the intermediate level τ_n

In the calculation of our extreme value estimates, the intermediate level τ_n is a tuning parameter that has to be chosen. This is of course essentially equivalent to choosing the parameter $k_n = \lfloor n(1-\tau_n) \rfloor$ representing the effective sample size in the Hill estimator used for the extrapolation. There are various ways of choosing k_n; we briefly discuss here a procedure based on an asymptotic mean-squared error minimisation criterion. As highlighted in Equation (3.2.13) p.77 in [13], the asymptotic mean-squared error of the Hill estimator under $\mathcal{C}_2(\gamma, \rho, A)$ is:

$$\mathrm{AMSE}(k_n) := \frac{1}{(1-\rho)^2}\left[A\left(\frac{n}{k_n}\right)\right]^2 + \frac{\gamma^2}{k_n}.$$

Let us consider the typical case of an auxiliary function $A(t) = ct^\rho$, $c \neq 0$, $\rho < 0$, as done by [13] on p.78 therein. Minimising the AMSE with respect to k_n yields an optimal value k_n^* satisfying

$$k_n^* = \left\lfloor \left(\frac{\gamma^2(1-\rho)^2}{-2\rho c^2}\right)^{1/(1-2\rho)} n^{-2\rho/(1-2\rho)} \right\rfloor.$$

This optimal value of k_n fulfills the well-known bias-variance trade-off in extreme value analysis, by balancing in an optimal way the variance increasing with low k_n and the bias increasing with high k_n. In practice, this value of k_n^* is of course unavailable because it depends on the unknown values of γ, ρ and A (through the parameter c). In our simulation study where a sample of $n = 1{,}000$ data points is available, we therefore suggest to use the sample counterpart $\widehat{k}_n^*$ of k_n^* obtained through plugging in a prior estimate of γ calculated using the bias-reduced Hill estimator with $k_n = n/10 = 100$, along with estimates of c and ρ obtained using the function `mop` from the R package `evt0`, all based of course on residuals of the model rather than the unobservable noise variables.

To check the quality of the estimation with this choice $\widehat{k}_n^*$ of k_n, we repeated our simulation studies in Sections 4.1 and 4.2, with the same parameters but with $\widehat{k}_n^*$ in place of $k_n = 100$. Results are reported in Tables F.2 and F.4. It is readily seen there that there is no obvious advantage in using a data-driven criterion for the choice of k_n, except in the direct estimator in our time series models. This is most likely because a data-driven choice of k_n is itself random and therefore may contribute to estimation uncertainty.

F.2. Pointwise confidence interval construction

We have explained, following our simulation studies in Sections 4.1 and 4.2, that most of the uncertainty in the problem of estimating extreme conditional expectiles appears indeed to come from the extreme value step. This seems to be particularly the case as soon as $\gamma \geq 0.2$. One may then use the asymptotic results developed in this paper to carry out pointwise inference about extreme conditional quantiles. Indeed, in typical cases the limit law in Theorem 2.3 is standard, and in fact is even Gaussian, because it is the limiting distribution of the extreme value index estimator $\overline{\gamma}$; under their respective suitable conditions, all common extreme value index estimators are asymptotically Gaussian. This is the case for the Hill estimator, of course, as we state in our Corollary 2.1, but also for, among others, the Pickands estimator, the Maximum Likelihood for the Generalised Pareto approximation, the moment estimator of [14] and probability weighted moment estimators (see respectively Theorems 3.3.5, 3.4.2, 3.5.4 and 3.6.1 in [13]). Asymptotic bias terms depend on γ, the second-order parameter ρ and the auxiliary function A, while asymptotic variances are functions of γ only. For instance, if $\overline{\gamma}$ is the Hill estimator $\widehat{\gamma}_{\lfloor n(1-\tau_n)\rfloor}$ as in Corollary 2.1, Theorem 2.3 reads, in model (1),

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau_n')]}\left(\frac{\overline{\xi}^\star_{\tau_n'}(Y|\boldsymbol{x})}{\xi_{\tau_n'}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right).$$

Model	Procedure	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$
Linear (G1)	(S1)	$2.29 \cdot 10^{-2}$	$3.56 \cdot 10^{-2}$	$6.46 \cdot 10^{-2}$	$1.13 \cdot 10^{-1}$
	(S1i)	$1.37 \cdot 10^{-2}$	$3.14 \cdot 10^{-2}$	$6.51 \cdot 10^{-2}$	$1.21 \cdot 10^{-1}$
	(S2)	$2.74 \cdot 10^{-2}$	$3.76 \cdot 10^{-2}$	$6.17 \cdot 10^{-2}$	$9.86 \cdot 10^{-2}$
	(S2i)	$3.11 \cdot 10^{-2}$	$3.57 \cdot 10^{-2}$	$5.93 \cdot 10^{-2}$	$1.05 \cdot 10^{-1}$
	(B1)	$1.26 \cdot 10^{-1}$	$8.06 \cdot 10^{-2}$	$9.89 \cdot 10^{-2}$	$1.93 \cdot 10^{-1}$
	(B1i)	$1.58 \cdot 10^{-1}$	$7.85 \cdot 10^{-2}$	$9.75 \cdot 10^{-2}$	$1.96 \cdot 10^{-1}$
	(B2)	$1.22 \cdot 10^{-1}$	$1.09 \cdot 10^{-1}$	$9.90 \cdot 10^{-2}$	$1.08 \cdot 10^{-1}$
	(B3)	$2.52 \cdot 10^{-2}$	$3.93 \cdot 10^{-2}$	$6.78 \cdot 10^{-2}$	$1.16 \cdot 10^{-1}$
	(B4)	$4.82 \cdot 10^{-2}$	$4.13 \cdot 10^{-2}$	$6.34 \cdot 10^{-2}$	$1.04 \cdot 10^{-1}$
	(B4i)	$8.15 \cdot 10^{-3}$	$2.73 \cdot 10^{-2}$	$6.23 \cdot 10^{-2}$	$1.18 \cdot 10^{-1}$
	(B5)	$2.26 \cdot 10^{-2}$	$3.53 \cdot 10^{-2}$	$6.23 \cdot 10^{-2}$	$1.06 \cdot 10^{-1}$
	(B5i)	$9.47 \cdot 10^{-3}$	$3.09 \cdot 10^{-2}$	$6.38 \cdot 10^{-2}$	$1.12 \cdot 10^{-1}$
Single index (G2)	(S1)	$1.83 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	$8.13 \cdot 10^{-2}$	$1.09 \cdot 10^{-1}$
	(S1i)	$1.96 \cdot 10^{-1}$	$1.18 \cdot 10^{-1}$	$6.97 \cdot 10^{-2}$	$1.01 \cdot 10^{-1}$
	(S2)	$3.90 \cdot 10^{-2}$	$4.38 \cdot 10^{-2}$	$6.89 \cdot 10^{-2}$	$1.08 \cdot 10^{-1}$
	(S2i)	$5.75 \cdot 10^{-2}$	$4.27 \cdot 10^{-2}$	$6.53 \cdot 10^{-2}$	$1.08 \cdot 10^{-1}$
	(B1)	$1.43 \cdot 10^{-1}$	$8.89 \cdot 10^{-2}$	$1.18 \cdot 10^{-1}$	$2.06 \cdot 10^{-1}$
	(B1i)	$1.74 \cdot 10^{-1}$	$7.64 \cdot 10^{-2}$	$1.14 \cdot 10^{-1}$	$2.05 \cdot 10^{-1}$
	(B2)	$3.46 \cdot 10^{-1}$	$2.79 \cdot 10^{-1}$	$2.37 \cdot 10^{-1}$	$1.95 \cdot 10^{-1}$
	(B3)	$2.97 \cdot 10^{-2}$	$4.20 \cdot 10^{-2}$	$7.20 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$
	(B4)	$5.84 \cdot 10^{-2}$	$4.82 \cdot 10^{-2}$	$7.14 \cdot 10^{-2}$	$1.13 \cdot 10^{-1}$
	(B4i)	$9.86 \cdot 10^{-3}$	$3.19 \cdot 10^{-2}$	$7.01 \cdot 10^{-2}$	$1.28 \cdot 10^{-1}$
	(B5)	$2.73 \cdot 10^{-2}$	$4.18 \cdot 10^{-2}$	$7.01 \cdot 10^{-2}$	$1.15 \cdot 10^{-1}$
	(B5i)	$1.15 \cdot 10^{-2}$	$3.61 \cdot 10^{-2}$	$7.18 \cdot 10^{-2}$	$1.22 \cdot 10^{-1}$

TABLE F.1

RMAD of methods (S1), (S2), (S1i) and (S2i), and of benchmarks (B1)–(B5i), in models (G1)–(G2). Estimators based on the fixed intermediate level $k_n = n/10 = 100$.

Consistent estimators of ρ and A are available from the work of [19], adapted here by using residuals instead of the unobserved errors. In each case the asymptotic bias and variance terms can then be estimated, and carrying out inference on the extreme conditional expectile of interest is, in principle, straightforward.

For consistency with our finite-sample studies and especially our real data analyses, we discuss the implementation of such confidence intervals based on the bias-reduced estimators $\widehat{\gamma}_k^{\mathrm{RB}}$, obtained by a bias reduction of the Hill estimator $\widehat{\gamma}_k$ (where throughout $k = \lfloor n(1-\tau_n) \rfloor$) and $\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(\varepsilon)$, obtained by a bias reduction of the direct extrapolated estimator $\widehat{\xi}^{\star}_{\tau'_n}(\varepsilon)$, whose expression can be found at the beginning of Section 4. Combined with appropriate model structure estimators converging quickly enough, these naturally give rise to an estimator $\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(Y|\boldsymbol{x})$ which, by Theorem 2.3, should satisfy

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\left(\frac{\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(Y|\boldsymbol{x})}{\xi_{\tau'_n}(Y|\boldsymbol{x})} - 1\right) \xrightarrow{d} \mathcal{N}(0,\gamma^2).$$

In line with standard practice in extreme value analysis for heavy tails, we consider instead the equivalent version

$$\frac{\sqrt{n(1-\tau_n)}}{\log[(1-\tau_n)/(1-\tau'_n)]}\log\left(\frac{\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(Y|\boldsymbol{x})}{\xi_{\tau'_n}(Y|\boldsymbol{x})}\right) \xrightarrow{d} \mathcal{N}(0,\gamma^2)$$

obtained via the delta-method, as this has been observed several times to yield more reasonable confidence intervals when using Weissman-type extrapolated estimators (see *e.g.* [15] in the context of extreme quantile estimation). This immediately provides an asymptotic pointwise 95% confidence interval for $\xi_{\tau'_n}(Y|\boldsymbol{x})$ as

$$\widehat{I}^{(1)}_{\tau'_n}(\boldsymbol{x}) = \left[\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(Y|\boldsymbol{x})\exp\left(\pm 1.96\frac{\log[(1-\tau_n)/(1-\tau'_n)]}{\sqrt{n(1-\tau_n)}}\widehat{\gamma}^{\mathrm{RB}}_{\lfloor n(1-\tau_n)\rfloor}\right)\right].$$

A slightly different construction, also motivated by Theorem 2.3, is possible by building the confidence interval directly on the estimator $\widehat{\xi}^{\star,\mathrm{RB}}_{\tau'_n}(\varepsilon)$ first and combining with location and scale afterwards. In

Model	Procedure	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$
Linear (G1)	(S1)	$1.48 \cdot 10^{-2}$	$3.33 \cdot 10^{-2}$	$6.86 \cdot 10^{-2}$	$1.25 \cdot 10^{-1}$
	(S1i)	$1.49 \cdot 10^{-2}$	$2.95 \cdot 10^{-2}$	$6.63 \cdot 10^{-2}$	$1.24 \cdot 10^{-1}$
	(S2)	$3.50 \cdot 10^{-2}$	$4.38 \cdot 10^{-2}$	$6.84 \cdot 10^{-2}$	$1.15 \cdot 10^{-1}$
	(S2i)	$3.61 \cdot 10^{-2}$	$4.05 \cdot 10^{-2}$	$6.68 \cdot 10^{-2}$	$1.11 \cdot 10^{-1}$
	(B1)	$1.26 \cdot 10^{-1}$	$8.06 \cdot 10^{-2}$	$9.89 \cdot 10^{-2}$	$1.93 \cdot 10^{-1}$
	(B1i)	$1.58 \cdot 10^{-1}$	$7.85 \cdot 10^{-2}$	$9.75 \cdot 10^{-2}$	$1.96 \cdot 10^{-1}$
	(B2)	$1.48 \cdot 10^{-1}$	$1.26 \cdot 10^{-1}$	$1.17 \cdot 10^{-1}$	$1.31 \cdot 10^{-1}$
	(B3)	$2.25 \cdot 10^{-2}$	$3.98 \cdot 10^{-2}$	$7.26 \cdot 10^{-2}$	$1.18 \cdot 10^{-1}$
	(B4)	$1.41 \cdot 10^{-2}$	$3.56 \cdot 10^{-2}$	$6.28 \cdot 10^{-2}$	$1.15 \cdot 10^{-1}$
	(B4i)	$8.79 \cdot 10^{-3}$	$2.95 \cdot 10^{-2}$	$6.86 \cdot 10^{-2}$	$1.26 \cdot 10^{-1}$
	(B5)	$1.14 \cdot 10^{-2}$	$3.30 \cdot 10^{-2}$	$6.66 \cdot 10^{-2}$	$1.15 \cdot 10^{-1}$
	(B5i)	$8.64 \cdot 10^{-3}$	$2.84 \cdot 10^{-2}$	$6.43 \cdot 10^{-2}$	$1.12 \cdot 10^{-1}$
Single index (G2)	(S1)	$1.85 \cdot 10^{-1}$	$1.23 \cdot 10^{-1}$	$1.01 \cdot 10^{-1}$	$1.30 \cdot 10^{-1}$
	(S1i)	$2.03 \cdot 10^{-1}$	$1.32 \cdot 10^{-1}$	$9.27 \cdot 10^{-2}$	$1.23 \cdot 10^{-1}$
	(S2)	$5.96 \cdot 10^{-2}$	$5.25 \cdot 10^{-2}$	$8.13 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$
	(S2i)	$6.68 \cdot 10^{-2}$	$5.04 \cdot 10^{-2}$	$7.32 \cdot 10^{-2}$	$1.14 \cdot 10^{-1}$
	(B1)	$1.43 \cdot 10^{-1}$	$8.89 \cdot 10^{-2}$	$1.18 \cdot 10^{-1}$	$2.06 \cdot 10^{-1}$
	(B1i)	$1.74 \cdot 10^{-1}$	$7.64 \cdot 10^{-2}$	$1.14 \cdot 10^{-1}$	$2.05 \cdot 10^{-1}$
	(B2)	$3.48 \cdot 10^{-1}$	$2.77 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$	$2.07 \cdot 10^{-1}$
	(B3)	$2.99 \cdot 10^{-2}$	$4.67 \cdot 10^{-2}$	$8.11 \cdot 10^{-2}$	$1.28 \cdot 10^{-1}$
	(B4)	$1.71 \cdot 10^{-2}$	$4.15 \cdot 10^{-2}$	$7.06 \cdot 10^{-2}$	$1.25 \cdot 10^{-1}$
	(B4i)	$1.06 \cdot 10^{-2}$	$3.44 \cdot 10^{-2}$	$7.71 \cdot 10^{-2}$	$1.37 \cdot 10^{-1}$
	(B5)	$1.38 \cdot 10^{-2}$	$3.85 \cdot 10^{-2}$	$7.49 \cdot 10^{-2}$	$1.25 \cdot 10^{-1}$
	(B5i)	$1.05 \cdot 10^{-2}$	$3.32 \cdot 10^{-2}$	$7.23 \cdot 10^{-2}$	$1.22 \cdot 10^{-1}$

TABLE F.2

RMAD of methods (S1), (S2), (S1i) and (S2i), and of benchmarks (B1)–(B5i), in models (G1)–(G2). Estimators based on the data-driven intermediate level $\widehat{k}_n^*$.

this case, an asymptotic pointwise 95% confidence interval for $\xi_{\tau_n'}(\varepsilon)$ is

$$\left[\widehat{\xi}_{\tau_n'}^{\star,\mathrm{RB}}(\varepsilon) \exp\left(\pm 1.96 \frac{\log[(1-\tau_n)/(1-\tau_n')]}{\sqrt{n(1-\tau_n)}} \widehat{\gamma}^{\mathrm{RB}}_{\lfloor n(1-\tau_n)\rfloor}\right)\right].$$

In the class of regression models (1) where $\xi_{\tau_n'}(Y|\boldsymbol{x}) = g(\boldsymbol{x}) + \sigma(\boldsymbol{x})\xi_{\tau_n'}(\varepsilon)$, this yields an alternative asymptotic pointwise 95% confidence interval for $\xi_{\tau_n'}(Y|\boldsymbol{x})$ as

$$\widehat{I}^{(2)}_{\tau_n'}(\boldsymbol{x}) = \left[\overline{g}(\boldsymbol{x}) + \overline{\sigma}(\boldsymbol{x})\widehat{\xi}_{\tau_n'}^{\star,\mathrm{RB}}(\varepsilon) \exp\left(\pm 1.96 \frac{\log[(1-\tau_n)/(1-\tau_n')]}{\sqrt{n(1-\tau_n)}} \widehat{\gamma}^{\mathrm{RB}}_{\lfloor n(1-\tau_n)\rfloor}\right)\right]$$

if g and σ are estimated by $\overline{g}$ and $\overline{\sigma}$ sufficiently fast that the asymptotic behaviour of $\widehat{\xi}_{\tau_n'}^{\star,\mathrm{RB}}(\varepsilon)$ dominates. In a model where the conditional mean is assumed to be 0 (for example GARCH models), the intervals $\widehat{I}^{(1)}_{\tau_n'}(\boldsymbol{x})$ and $\widehat{I}^{(2)}_{\tau_n'}(\boldsymbol{x})$ coincide. We chose to work with $\widehat{I}^{(1)}_{\tau_n'}(\boldsymbol{x})$, which compared to $\widehat{I}^{(2)}_{\tau_n'}(\boldsymbol{x})$ is typically slightly more conservative. We illustrate this construction in the top left panel of Figure F.1 below, on the example of the Vehicle Insurance Customer data of Section 4.3.

There are situations however where trusting these confidence intervals might be difficult. For instance, in regression models featuring the estimation of a nonparametric component (such as the heteroscedastic single-index model in Section 3.2, used for the analysis of the Vehicle Insurance Customer data) whose rate of convergence may be close to the rate of convergence of the extreme value estimator, disregarding the uncertainty incurred at the model estimation stage may be problematic, especially in regions where data is relatively sparse. It may then be more prudent to move away from the asymptotic approximation and use instead an approach that fully takes into account the uncertainty in the estimation. We propose and contrast here a couple of alternatives based on regression bootstrap methods. We develop our ideas in the example of the heteroscedastic single-index model of Section 3.2. Suppose that from a data set $(\boldsymbol{X}_i, Y_i)_{1\leq i\leq n}$, we have estimated a direction vector $\widehat{\boldsymbol{\beta}}$ along with mean and standard deviation functions $\widehat{g}$ and $\widehat{\sigma}$. One possibility to describe the uncertainty in the estimation of $\xi_{\tau_n'}(Y|\boldsymbol{x})$ is to use the wild bootstrap, widespread in the heteroscedastic regression literature and

Model	Parameters	Estimator	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$
ARMA	$(\phi, \theta) = (0.1, 0.1)$	Direct	$7.25 \cdot 10^{-2}$	$7.82 \cdot 10^{-2}$	$1.09 \cdot 10^{-1}$	$1.51 \cdot 10^{-1}$
	(estimated)	Indirect	$2.81 \cdot 10^{-2}$	$5.14 \cdot 10^{-2}$	$8.96 \cdot 10^{-2}$	$1.45 \cdot 10^{-1}$
	$(\phi, \theta) = (0.1, 0.1)$	Direct	$7.16 \cdot 10^{-2}$	$7.47 \cdot 10^{-2}$	$1.06 \cdot 10^{-1}$	$1.47 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.66 \cdot 10^{-2}$	$4.86 \cdot 10^{-2}$	$9.06 \cdot 10^{-2}$	$1.43 \cdot 10^{-1}$
	$(\phi, \theta) = (0.1, 0.5)$	Direct	$7.10 \cdot 10^{-2}$	$7.54 \cdot 10^{-2}$	$1.09 \cdot 10^{-1}$	$1.54 \cdot 10^{-1}$
	(estimated)	Indirect	$2.92 \cdot 10^{-2}$	$5.07 \cdot 10^{-2}$	$9.33 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$
	$(\phi, \theta) = (0.1, 0.5)$	Direct	$7.14 \cdot 10^{-2}$	$7.74 \cdot 10^{-2}$	$1.06 \cdot 10^{-1}$	$1.49 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.75 \cdot 10^{-2}$	$5.03 \cdot 10^{-2}$	$9.21 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.1)$	Direct	$7.24 \cdot 10^{-2}$	$8.07 \cdot 10^{-2}$	$1.12 \cdot 10^{-1}$	$1.53 \cdot 10^{-1}$
	(estimated)	Indirect	$2.98 \cdot 10^{-2}$	$5.26 \cdot 10^{-2}$	$9.51 \cdot 10^{-2}$	$1.44 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.1)$	Direct	$7.09 \cdot 10^{-2}$	$7.41 \cdot 10^{-2}$	$1.09 \cdot 10^{-1}$	$1.48 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.78 \cdot 10^{-2}$	$5.05 \cdot 10^{-2}$	$9.13 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.5)$	Direct	$6.56 \cdot 10^{-2}$	$8.16 \cdot 10^{-2}$	$1.13 \cdot 10^{-1}$	$1.57 \cdot 10^{-1}$
	(estimated)	Indirect	$2.90 \cdot 10^{-2}$	$5.52 \cdot 10^{-2}$	$9.67 \cdot 10^{-2}$	$1.51 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.5)$	Direct	$7.02 \cdot 10^{-2}$	$7.69 \cdot 10^{-2}$	$1.11 \cdot 10^{-1}$	$1.52 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.67 \cdot 10^{-2}$	$5.03 \cdot 10^{-2}$	$9.19 \cdot 10^{-2}$	$1.43 \cdot 10^{-1}$
GARCH	$(\alpha, \beta) = (0.1, 0.1)$	Direct	$7.00 \cdot 10^{-2}$	$7.40 \cdot 10^{-2}$	$1.05 \cdot 10^{-1}$	$1.45 \cdot 10^{-1}$
	(estimated)	Indirect	$1.60 \cdot 10^{-2}$	$4.78 \cdot 10^{-2}$	$8.85 \cdot 10^{-2}$	$1.35 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.1)$	Direct	$7.04 \cdot 10^{-2}$	$7.38 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.59 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$	$9.09 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.45)$	Direct	$7.02 \cdot 10^{-2}$	$7.47 \cdot 10^{-2}$	$1.04 \cdot 10^{-1}$	$1.38 \cdot 10^{-1}$
	(estimated)	Indirect	$1.66 \cdot 10^{-2}$	$4.79 \cdot 10^{-2}$	$8.70 \cdot 10^{-2}$	$1.29 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.45)$	Direct	$7.04 \cdot 10^{-2}$	$7.38 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.59 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$	$9.09 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.45, 0.1)$	Direct	$7.14 \cdot 10^{-2}$	$7.59 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$
	(estimated)	Indirect	$1.65 \cdot 10^{-2}$	$4.89 \cdot 10^{-2}$	$9.10 \cdot 10^{-2}$	$1.33 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.45, 0.1)$	Direct	$7.04 \cdot 10^{-2}$	$7.38 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.59 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$	$9.09 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.85)$	Direct	$6.51 \cdot 10^{-2}$	$8.33 \cdot 10^{-2}$	$1.06 \cdot 10^{-1}$	$1.22 \cdot 10^{-1}$
	(estimated)	Indirect	$2.39 \cdot 10^{-2}$	$6.57 \cdot 10^{-2}$	$9.26 \cdot 10^{-2}$	$1.13 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.85)$	Direct	$7.04 \cdot 10^{-2}$	$7.38 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.59 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$	$9.09 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$

TABLE F.3

RMAD of the (bias-reduced) direct and indirect extreme conditional expectile estimators in ARMA and GARCH models. Estimators based on the fixed intermediate level $k_n = n/10 = 100$.

whose origins can be traced back to [55]. This consists in resampling $(\boldsymbol{X}_i, Y_i^*)_{1 \le i \le n}$ as follows:

$$Y_i^* = \widehat{g}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i) + (Y_i - \widehat{g}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i))\varepsilon_i^*,$$

where $(\varepsilon_i^*)_{1 \le i \le n}$ are i.i.d. copies of a random variable ε^* having mean 0 and variance 1. A natural, possible choice for ε^* is the standard normal distribution. We illustrate this methodology on the example of the Vehicle Insurance Customer data of Section 4.3. We simulated $N = 5{,}000$ such bootstrap samples $(\boldsymbol{X}_i, Y_i^*)_{1 \le i \le n}$; in each sample, we kept the direction vector $\widehat{\boldsymbol{\beta}}$ fixed and equal to its estimated value based on the original sample, and we estimated the functions g and σ using the same method as in the real data analysis in Section 4.3. This is sensible because the estimator $\widehat{\boldsymbol{\beta}}$ converges much faster than the nonparametric estimators of g and σ, and therefore keeping the direction fixed is very unlikely to be incorrect as far as uncertainty quantification is concerned. Using residuals and the direct, bias-reduced extreme conditional expectile estimator results in an estimate of $\xi_{\tau_n'}(Y|\boldsymbol{x})$ which, for the jth bootstrap sample, we denote by $\widehat{\xi}^{\star,\mathrm{RB},(j)}_{\tau_n'}(Y|\boldsymbol{x})$. We finally build, for a fixed $\boldsymbol{x}$, pointwise 95% bootstrap confidence intervals calculated by taking the empirical quantiles at levels 2.5% and 97.5% of the $\widehat{\xi}^{\star,\mathrm{RB},(j)}_{\tau_n'}(Y|\boldsymbol{x})$, $1 \le j \le N$. These are reported in the top right panel of Figure F.1. At extreme levels (say here $\tau_n' = 1 - 1/(nh^*)$, with $h^* = 0.1$) the confidence intervals look reasonable on the right half of the graph. However, they seem to very substantially overestimate the uncertainty in the left half, where data is sparser; this is especially clear around $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x} = -0.2$, where the estimated extreme conditional expectile curve already extrapolates far beyond the observations locally relevant, which suggests that the upper bound of the associated confidence interval should be relatively close to the point estimate, but this is not the case. Moreover, the wild bootstrap method appears to be very sensitive to the choice of distribution of ε^* (alternative choices include the Rademacher distribution

Model	Parameters	Estimator	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$
ARMA	$(\phi, \theta) = (0.1, 0.1)$	Direct	$3.49 \cdot 10^{-2}$	$5.64 \cdot 10^{-2}$	$9.55 \cdot 10^{-2}$	$1.48 \cdot 10^{-1}$
	(estimated)	Indirect	$3.11 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$	$9.63 \cdot 10^{-2}$	$1.51 \cdot 10^{-1}$
	$(\phi, \theta) = (0.1, 0.1)$	Direct	$2.50 \cdot 10^{-2}$	$5.60 \cdot 10^{-2}$	$9.80 \cdot 10^{-2}$	$1.45 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.85 \cdot 10^{-2}$	$4.62 \cdot 10^{-2}$	$9.31 \cdot 10^{-2}$	$1.45 \cdot 10^{-1}$
	$(\phi, \theta) = (0.1, 0.5)$	Direct	$3.86 \cdot 10^{-2}$	$5.65 \cdot 10^{-2}$	$9.86 \cdot 10^{-2}$	$1.48 \cdot 10^{-1}$
	(estimated)	Indirect	$3.16 \cdot 10^{-2}$	$5.08 \cdot 10^{-2}$	$9.35 \cdot 10^{-2}$	$1.48 \cdot 10^{-1}$
	$(\phi, \theta) = (0.1, 0.5)$	Direct	$2.63 \cdot 10^{-2}$	$5.59 \cdot 10^{-2}$	$9.91 \cdot 10^{-2}$	$1.46 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.92 \cdot 10^{-2}$	$4.70 \cdot 10^{-2}$	$9.46 \cdot 10^{-2}$	$1.46 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.1)$	Direct	$3.64 \cdot 10^{-2}$	$5.91 \cdot 10^{-2}$	$9.79 \cdot 10^{-2}$	$1.45 \cdot 10^{-1}$
	(estimated)	Indirect	$3.21 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	$9.61 \cdot 10^{-2}$	$1.54 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.1)$	Direct	$2.62 \cdot 10^{-2}$	$5.66 \cdot 10^{-2}$	$1.02 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.93 \cdot 10^{-2}$	$4.69 \cdot 10^{-2}$	$9.62 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.5)$	Direct	$3.50 \cdot 10^{-2}$	$6.35 \cdot 10^{-2}$	$9.72 \cdot 10^{-2}$	$1.53 \cdot 10^{-1}$
	(estimated)	Indirect	$3.17 \cdot 10^{-2}$	$5.31 \cdot 10^{-2}$	$9.73 \cdot 10^{-2}$	$1.52 \cdot 10^{-1}$
	$(\phi, \theta) = (0.5, 0.5)$	Direct	$2.62 \cdot 10^{-2}$	$6.12 \cdot 10^{-2}$	$1.03 \cdot 10^{-1}$	$1.54 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.87 \cdot 10^{-2}$	$5.02 \cdot 10^{-2}$	$9.78 \cdot 10^{-2}$	$1.51 \cdot 10^{-1}$
GARCH	$(\alpha, \beta) = (0.1, 0.1)$	Direct	$2.54 \cdot 10^{-2}$	$5.52 \cdot 10^{-2}$	$9.97 \cdot 10^{-2}$	$1.41 \cdot 10^{-1}$
	(estimated)	Indirect	$1.85 \cdot 10^{-2}$	$4.51 \cdot 10^{-2}$	$9.26 \cdot 10^{-2}$	$1.45 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.1)$	Direct	$2.43 \cdot 10^{-2}$	$5.51 \cdot 10^{-2}$	$9.94 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.82 \cdot 10^{-2}$	$4.58 \cdot 10^{-2}$	$9.33 \cdot 10^{-2}$	$1.49 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.45)$	Direct	$2.40 \cdot 10^{-2}$	$5.41 \cdot 10^{-2}$	$9.64 \cdot 10^{-2}$	$1.43 \cdot 10^{-1}$
	(estimated)	Indirect	$1.80 \cdot 10^{-2}$	$4.45 \cdot 10^{-2}$	$9.16 \cdot 10^{-2}$	$1.42 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.45)$	Direct	$2.43 \cdot 10^{-2}$	$5.51 \cdot 10^{-2}$	$9.94 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.82 \cdot 10^{-2}$	$4.58 \cdot 10^{-2}$	$9.33 \cdot 10^{-2}$	$1.49 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.45, 0.1)$	Direct	$2.48 \cdot 10^{-2}$	$5.33 \cdot 10^{-2}$	$1.04 \cdot 10^{-1}$	$1.49 \cdot 10^{-1}$
	(estimated)	Indirect	$1.86 \cdot 10^{-2}$	$4.76 \cdot 10^{-2}$	$9.71 \cdot 10^{-2}$	$1.52 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.45, 0.1)$	Direct	$2.43 \cdot 10^{-2}$	$5.51 \cdot 10^{-2}$	$9.94 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.82 \cdot 10^{-2}$	$4.58 \cdot 10^{-2}$	$9.33 \cdot 10^{-2}$	$1.49 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.85)$	Direct	$3.16 \cdot 10^{-2}$	$7.53 \cdot 10^{-2}$	$1.10 \cdot 10^{-1}$	$1.41 \cdot 10^{-1}$
	(estimated)	Indirect	$2.79 \cdot 10^{-2}$	$6.66 \cdot 10^{-2}$	$1.04 \cdot 10^{-1}$	$1.35 \cdot 10^{-1}$
	$(\alpha, \beta) = (0.1, 0.85)$	Direct	$2.43 \cdot 10^{-2}$	$5.51 \cdot 10^{-2}$	$9.94 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$
	(known, benchmark)	Indirect	$1.82 \cdot 10^{-2}$	$4.58 \cdot 10^{-2}$	$9.33 \cdot 10^{-2}$	$1.49 \cdot 10^{-1}$

TABLE F.4

RMAD of the (bias-reduced) direct and indirect extreme conditional expectile estimators in ARMA and GARCH models. Estimators based on the data-driven intermediate level $\widehat{k}_n^*$.

or asymmetric two-point distributions such as the one on p.257 of [36]). Our interpretation is that the wild bootstrap is too conservative here because it fails to get a good idea of the right tail behaviour in the data.

To remedy this problem we suggest a second, semiparametric bootstrap method. This time, the Y_i^*, $1 \leq i \leq n$, are simulated as

$$Y_i^* = \widehat{g}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i) + \widehat{\sigma}(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}_i)\varepsilon_i^*,$$

where the ε_i^* are obtained by

1. Simulating u_i from the standard uniform distribution on $[0, 1]$,
2. If $u_i \in [p, 1-p]$, for a fixed $p \in (0, 1)$, taking $\varepsilon_i^* = \widehat{F}^{-1}(u_i)$, where $\widehat{F}$ is the empirical distribution function of the residuals $\widehat{\varepsilon}_i$,
3. If $u_i > 1 - p$, taking $\varepsilon_i^* = ((1 - u_i)/p)^{-\widehat{\gamma}}\widehat{F}^{-1}(1 - p)$, where $\widehat{\gamma} = \widehat{\gamma}^{\mathrm{RB}}$ is the bias-reduced Hill estimator (with $k_n = 200$ as in Section 4.3) based on the residuals $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$,
4. If $u_i < p$, taking $\varepsilon_i^* = (u_i/p)^{-\widehat{\gamma}_\ell}\widehat{F}^{-1}(p)$, where $\widehat{\gamma}_\ell = \widehat{\gamma}_\ell^{\mathrm{RB}}$ is the bias-reduced Hill estimator (with $k_n = 200$) based on the negative residuals $-\widehat{\varepsilon}_1, \ldots, -\widehat{\varepsilon}_n$.

We chose $p = 0.001$; further investigations, which we do not report here, suggest that results are not too sensitive to the choice of p as long as $p \in [0.001, 0.01]$. The idea of steps 3 and 4 above is to allow the resampling algorithm to give a faithful idea of the right and left tails of the data through the use of the Pareto approximations of these tails. We call this algorithm the *semiparametric Pareto tail bootstrap*. Somewhat similar ideas have appeared before in the literature, see *e.g.* [56] whose aim was to approximate the distribution of extreme order statistics.

We illustrate this methodology again on the example of the Vehicle Insurance Customer data of Sec-

tion 4.3. We simulate $N = 5{,}000$ bootstrap samples $(\boldsymbol{X}_i, Y_i^*)_{1\leq i\leq n}$ and, like previously, we keep the direction vector $\widehat{\boldsymbol{\beta}}$ fixed and estimate the functions g and σ using the same method as in Section 4.3. This yields extrapolated direct bias-reduced estimates of $\xi_{\tau_n'}(Y|\boldsymbol{x})$ in each sample and therefore pointwise 95% bootstrap confidence intervals calculated by taking the empirical quantiles at levels 2.5% and 97.5% of these estimates. These intervals are reported in the bottom left panel of Figure F.1; all three intervals are compared to each other on the bottom right panel of this Figure. All intervals are roughly similar on the right part of the graph, but on the left part where data is more sparse, the semiparametric Pareto tail bootstrap intervals appear to give a much better idea of the type of tail the data exhibits. Our recommended strategy for inference about extreme conditional expectiles is thus the following:

- If the structure of the regression model under consideration can be estimated at the parametric rate $\sqrt{n}$, then use the pointwise Gaussian asymptotic confidence interval $\widehat{I}^{(1)}_{\tau_n'}$ motivated by the asymptotic theory,
- If the structure of the model contains non- or semiparametric components, use pointwise confidence intervals provided by the semiparametric Pareto tail bootstrap.

The pointwise 95% confidence intervals provided to support our real data analyses in Sections 4.3 and 4.4 are obtained following this strategy.

FIG F.1. *Vehicle Insurance Customer data, pointwise* 95% *confidence intervals. Top left panel: asymptotic Gaussian confidence intervals, top right: wild bootstrap confidence intervals, bottom left: semiparametric Pareto tail bootstrap confidence intervals, bottom right: all three intervals. In all panels, the red line is the regression mean, the solid purple line is the (direct, bias-reduced) extreme conditional expectile estimate at level* $\tau_n' = 1 - 1/(nh^*)$ *and the dashed, dashed-dotted and dotted lines represent the* 95% *confidence intervals, in the* $(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{x}, y)$ *plane.*