SUPPLEMENTAL APPENDIX TO "INFERENCE ON MODEL PARAMETERS WITH MANY L-MOMENTS"

LUIS A. F. ALVAREZ, CHANG CHIANN, AND PEDRO A. MORETTIN

Abstract. This Supplemental Appendix presents the proofs of the main results in the paper, as well as details on the methods of selection of L-moments, the extensions to "residual analysis" and conditional models, and additional information on the Monte Carlo exercises.

Contents

Alvarez: Department of Economics, University of São Paulo.
Chiann: Department of Statistics, University of São Paulo.
Morettin: Department of Statistics, University of São Paulo.
E-mail addresses: luis.alvarez@usp.br, chang@ime.usp.br, pam@ime.usp.br.

Appendix A. Proof of main results in the text

A.1. Proof of Proposition 1.

Proof. For $\theta \in \Theta$, define:

$$M(\theta) \coloneqq \left[\int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u)' du\right] (W^R) \left[\int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u) du\right]$$

$$M_0(\theta) \coloneqq \left[\int_{\underline{p}}^{\bar{p}} \left(Q_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u)' du\right] (\Omega^R) \left[\int_{\underline{p}}^{\bar{p}} \left(Q_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u) du\right]$$

$$h^R(\theta) \coloneqq \int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u) du$$

$$h_0^R(\theta) \coloneqq \int_{\underline{p}}^{\bar{p}} \left(Q_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u) du\,.$$

Then, proceeding similarly to Theorem 2.6 of Newey and McFadden (1994), we have that, by application of the Cauchy-Schwarz inequality and the properties of the spectral norm:

$$\begin{aligned} |M(\theta) - M_0(\theta)| &\leq |(h^R(\theta) - h_0^R(\theta))' W^R (h^R(\theta) - h_0^R(\theta))| + \\ |h_0^R(\theta)' (W^R + W^{R'}) (h^R(\theta) - h_0^R(\theta))| &+ |h_0^R(\theta)' (W^R - \Omega^R) h_0^R(\theta)| \leq \\ \leq \|W^R\|_2 \|h^R(\theta) - h_0^R(\theta)\|_2^2 &+ 2\|W^R\|_2 \|h_0^R(\theta) - h^R(\theta)\|_2 \|h_0^R(\theta)\|_2 + \|W^R - \Omega^R\|_2 \|h_0^R(\theta)\|_2^2\,. \end{aligned}$$

We analyse the behaviour of each term separately. First, note that, by Bessel's inequality and Assumption 1:

$$\|h^R(\theta) - h_0^R(\theta)\|_2^2 = \sum_{l=1}^{R} \left[\int_{\underline{p}}^{\bar{p}} [\hat{Q}_Y(u) - Q_Y(u)] P_l(u) du\right]^2 \leq \|(\hat{Q}_Y(\cdot) - Q_Y(\cdot)) \mathbb{1}_{[\underline{p},\bar{p}]}\|_{L^2[0,1]}^2 = o_{p^*}(1)\,,$$

where the upper bound does not depend on θ. Next, we have:

$$\|h_0^R(\theta)\|_2^2 \leq \|(Q_Y(\cdot) - Q_Y(\cdot|\theta)) \mathbb{1}_{[\underline{p},\bar{p}]}\|_{L^2[0,1]}^2 \leq 2 \sup_{\Delta \in \Theta} \|Q_Y(\cdot|\Delta) \mathbb{1}_{[\underline{p},\bar{p}]}\|_{L^2[0,1]} < \infty\,,$$

where we use Bessel's inequality (Kreyszig, 1989, page 157) and the last part of Assumption 4. Combining these facts with Assumption 3, we obtain:

$$\sup_{\theta \in \Theta} |M(\theta) - M_0(\theta)| \xrightarrow{P^*} 0\,.$$

Finally we verify the unique identifiability condition of Pötscher and Prucha (1997, Definition 3.1). Since $M_0(\theta_0) = 0$, the condition subsumes to verifying that, for each $\epsilon > 0$:

$$\liminf_{T,R \to \infty} \inf_{\theta \in \Theta : \|\theta - \theta_0\|_2 \geq \epsilon} M_0(\theta) > 0\,.$$

This condition is clearly implied by Assumption 3. Applying Lemma 3.1. of Pötscher and Prucha (1997), we conclude that $\hat{\theta} \xrightarrow{P^*} \theta_0$, as desired. □

A.2. **Proof of Proposition 2.**

Proof. Following the usual argument in the Generalised Method of Moments literature (Newey and McFadden, 1994), we first show that $\hat{\theta}_T$ satisfies a first order condition with high probability. Indeed, under Assumption 5, straightforward application of the dominated convergence theorem reveals that $h^R(\theta) = \int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^R(u)du$ is differentiable on $\mathcal{O}$, with derivative given by differentiation under the integral sign. Moreover, since $\theta_0 \in \mathcal{O}$ and $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\theta} \in \mathcal{O}$ with probability approaching one (wpa 1). It thus follows that

$$\nabla_{\theta'} h^R(\hat{\theta})' W^R h^R(\hat{\theta}) = 0\,, \tag{1}$$

holds wpa1, where $\nabla_{\theta'} h^R(\tilde{\theta})$ denotes the Jacobian of h^R with respect to θ, evaluated at $\tilde{\theta}$.

Next, since, for each $u \in [\underline{p}, \bar{p}]$, $\theta \mapsto Q_Y(u|\theta)$ is continuously differentiable on $\mathcal{O}$, a mean-value-expansion yields that, with probability approaching one:

$$h^R(\hat{\theta}) = h^R(\theta_0) + \nabla_{\theta'} \widetilde{h^R(\theta)}(\hat{\theta} - \theta)\,,$$

where $\nabla_{\theta'} \widetilde{h^R(\theta)}$ is the $R \times d$ matrix where each line l is equal to $-\int_{\underline{p}}^{\overline{p}} \nabla_{\theta'} Q_Y(u|\tilde{\theta}(u)) P_l(u)du$, and $\tilde{\theta}(u)$ is a u-specific element in the line segment between $\hat{\theta}$ and θ_0. Rearranging terms, adding and subtracting Ω^R yields:

$$\begin{aligned} &\nabla_{\theta'} h^R(\hat{\theta})' \Omega^R h^R(\theta_0) + \nabla_{\theta'} h^R(\hat{\theta})' (W^R - \Omega^R) h^R(\theta_0) \\ &= -\nabla_{\theta'} h^R(\hat{\theta})' \Omega^R \nabla_{\theta'} \widetilde{h^R(\theta)}(\hat{\theta} - \theta_0) - \nabla_{\theta'} h^R(\hat{\theta})' (W^R - \Omega^R) \nabla_{\theta'} \widetilde{h^R(\theta)}(\hat{\theta} - \theta_0)\,. \end{aligned}$$

The crucial step now is to work out asymptotic tightness of a normalization of $h^R(\theta_0)$. Observe that Assumption 6 entails that:

$$\left\| \sqrt{T} h^R(\theta_0) \right\|_2^2 \leq \sum_{l=1}^{\infty} \left| \int_{\underline{p}}^{\bar{p}} \sqrt{T}(\hat{Q}_Y(u) - Q_Y(u)) P_l(u)du \right|^2 \leq \| \sqrt{T}(\hat{Q}_Y(\cdot) - Q_Y(\cdot)) \mathbb{1}_{[\underline{p},\bar{p}]} \|^2_{L^2[0,1]} = O_{p^*}(1)\,.$$

The next step in the proof concerns the approximation of $\nabla_{\theta'} h^R(\hat{\theta})$ to $\nabla_{\theta'} h^R(\theta_0)$ in the spectral norm. Notice that, by the properties of the spectral norm and Bessel's inequality:

$$\| \nabla_{\theta'} h^R(\hat{\theta}) - \nabla_{\theta'} h^R(\theta_0) \|_2^2 \leq \sum_{s=1}^{d} \| [\partial_{\theta_s} Q_Y(\cdot|\hat{\theta}) - \partial_{\theta_s} Q_Y(\cdot|\theta_0)] \mathbb{1}_{[\underline{p},\bar{p}]} \|^2_{L^2[0,1]}\,.$$

We claim that, $\hat{\theta} \xrightarrow{P^*} \theta_0$, together with Assumption 5, is sufficient to ensure the upper bound above is $o_{p^*}(1)$. Since d is fixed, we may consider the argument for a fixed $s = 1, 2, \ldots d$. Fix $\eta, \epsilon > 0$. Since, by assumption, $\partial_s Q_Y(u|\theta)$ is continuous at θ_0, uniformly in u; there exists $\delta > 0$ such that:

$$\|\theta - \theta_0\|_2 \leq \delta \implies |\partial_s Q_Y(u|\theta) - \partial_s Q_Y(u|\theta_0)| \leq \frac{\sqrt{\epsilon}}{\bar{p} - \underline{p}} \quad \forall u \in [\underline{p}, \bar{p}]\,.$$

Now, since $\hat{\theta} \xrightarrow{P^*} \theta_0$, there exists $N \in \mathbb{N}$ such that, for all $T \geq N$:

$$P^*(\|\hat{\theta} - \theta_0\|_2 \leq \delta) \geq 1 - \eta\,,$$

implying that, by monotonicity of the outer probability, for $T \geq N$:

$$P^*(\|[\partial_{\theta_s} Q_Y(\cdot|\hat{\theta}) - \partial_{\theta_s} Q_Y(\cdot|\theta_0)]\mathbb{1}_{[\underline{p},\bar{p}]}\|^2_{L^2[0,1]} \leq \epsilon) \geq 1 - \eta\,.$$

Since the choice of η and ϵ is arbitrary, we obtain that:

$$\|[\partial_{\theta_s} Q_Y(\cdot|\hat{\theta}) - \partial_{\theta_s} Q_Y(\cdot|\theta_0)]\mathbb{1}_{[\underline{p},\bar{p}]}\|^2_{L^2[0,1]} = o_{p^*}(1)\,,$$

and since d is fixed, we conclude that:

$$\|\nabla_{\theta'} h^R(\hat{\theta}) - \nabla_{\theta'} h^R(\theta_0)\|_2^2 = o_{p^*}(1)\,.$$

Next, we would like to similarly argue that $\|\nabla_{\theta'} \widetilde{h^R(\theta)} - \nabla_{\theta'} h^R(\theta_0)\|_2^2 = o_{P^*}(1)$. The difficulty here is that each u possesses its u-specific $\tilde{\theta}(u)$. Note, however, that by Bessel's inequality:

$$\|\nabla_{\theta'} \widetilde{h^R(\theta)} - \nabla_{\theta'} h^R(\theta_0)\|_2^2 \leq \sum_{s=1}^{d} \left[\int_{\underline{p}}^{\bar{p}} \left(\partial_{\theta_s} Q_Y(u|\tilde{\theta}(u)) - \partial_{\theta_s} Q_Y(u|\theta_0)\right) P_l(u) du\right]^2.$$

Under Assumption 7, a mean-value expansion of the right-hand side above, followed by using Hölder's inequality, $\|P_l\|_{L^2[0,1]} = 1$, the Cauchy-Schwarz inequality, and that $\|\tilde{\theta} - \theta_0\|_2 \leq \|\hat{\theta} - \theta_0\|_2$ for any $\tilde{\theta}$ in the line segment between $\hat{\theta}$ and θ_0; yields

$$\left[\int_{\underline{p}}^{\bar{p}} \left(\partial_{\theta_s} Q_Y(u|\tilde{\theta}(u)) - \partial_{\theta_s} Q_Y(u|\theta_0)\right) P_l(u) du\right]^2 = \left[\int_{\underline{p}}^{\bar{p}} \nabla_\theta \partial_{\theta_s} Q_Y(u|\check{\theta}(u))'(\tilde{\theta}(u) - \theta_0) P_l(u) du\right]^2 \leq$$
$$\leq \int_{\underline{p}}^{\bar{p}} \left[\nabla_\theta \partial_{\theta_s} Q_Y(u|\check{\theta}(u))'(\tilde{\theta}(u) - \theta_0)\right]^2 du \leq \left(\int_{\underline{p}}^{\bar{p}} \|\nabla_\theta \partial_{\theta_s} Q_Y(u|\check{\theta}(u))\|_2^2 du\right) \cdot \|\hat{\theta} - \theta_0\|_2^2 = o_p(1)\,,$$

as desired.

Next, using that $\|\nabla_{\theta'} h^R(\theta_0)\|_2^2 = O(1)$ (which follows from Bessel's inequality and the last part of Assumption 5) and the previous results, we arrive at:

$$(\nabla_{\theta'} h^R(\theta_0)' \Omega^R \nabla_{\theta'} h^R(\theta_0) + r^{TR})\sqrt{T}(\hat{\theta} - \theta_0) = -\nabla_{\theta'} h^R(\theta_0)' \Omega^R (\sqrt{T} h^R(\theta_0)) + o_{P^*}(1)\,,$$

where the remainder r^{TR} satisfies $\|r^{TR}\|_2^2 = o_{P^*}(1)$. To proceed, we need to ensure that the matrix $(\nabla_{\theta'} h^R(\theta_0)' \Omega^R \nabla_{\theta'} h^R(\theta_0) + r^{TR})$ is invertible with high probability. Notice that, under the condition in Assumption 8, we have, using the Bauer-Fike theorem (Bhatia, 1997, Theorem VIII.3.1), that, wpa 1,

$$\lambda_{\min}\left(\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0)+r^{TR}\right)>0\,,$$

and further using that, for invertible matrices A_0 and A

$$\|A^{-1}-A_0^{-1}\|_2=\|A^{-1}(A_0-A)A_0^{-1}\|_2\leq\|A^{-1}\|_2\|(A_0-A)\|_2\|A_0^{-1}\|_2\,,$$
$$\|A^{-1}\|_2=\|A^{-1}(A-A_0)A_0^{-1}-A_0^{-1}\|_2\leq\|A_0^{-1}\|_2+\|A^{-1}\|_2\|(A_0-A)\|_2\|A_0^{-1}\|_2\,,$$

We obtain that:

$$\|A^{-1}-A_0^{-1}\|_2\leq\|A_0^{-1}\|_2\frac{\|A_0-A\|_2}{\|A_0^{-1}\|_2^{-1}-\|A-A_0\|_2}\,.$$

Taking $A_0=\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0)$ and $A=\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0)+r^{TR}$, and using that Assumption 8 implies $\|\left(\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0)\right)^{-1}\|_2$ is bounded above uniformly in R,[1] we conclude that:

$$\|(\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0)+r^{RT})^{-1}-(\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0))^{-1}\|_2=o_{P^*}(1)\,.$$

From which we conclude that, wpa 1:

$$\sqrt{T}(\hat{\theta}-\theta_0)=-(\nabla_{\theta'}h^R(\theta_0)'\Omega^R\nabla_{\theta'}h^R(\theta_0))^{-1}\nabla_{\theta'}h^R(\theta_0)'\Omega^R(\sqrt{T}h^R(\theta_0))+o_{P^*}(1)\,,$$

proving the desired asymptotic linear representation. □

Appendix B. Verification of the boundedness condition in Assumption 4 for the GEV and GPD distributions

B.1. **Generalized Extreme Value.** Consider a Generalized Extreme Value with location parameter $\xi\in\mathbb{R}$, scale parameter $\alpha>0$ and shape parameter $k\in\mathbb{R}$. Denoting by $\theta=(\xi,\alpha,k)'$ the vector of parameters, we have that the quantile function is given by (Hosking, 1986, page 70):

$$Q(u|\theta)=\begin{cases}\xi+\alpha(1-(-\log(u))^k)/k & k\neq 0\\ \xi-\alpha\log(-\log(u)) & k=0\end{cases}$$

In this case, for $k>-1/2$, it follows that (Singh, 1998, page 178):

$$\int_0^1 Q(u|\theta)^2du=\begin{cases}\left(\xi+\frac{\alpha}{k}[1-\Gamma(1+k)]\right)^2+\frac{\alpha^2}{k^2}\left(\Gamma(1+2k)-\Gamma(1+k)^2\right)\,, & k\neq 0\\ \left(\xi-\alpha\Gamma'(1)\right)^2+\alpha^2\left(\Gamma''(1)-(\Gamma'(1))^2\right)\,, & k=0\end{cases},$$

where Γ denotes the Gamma function; and the integral under $k=0$ coincides with the limit when $k\to 0$ (Kotz and Nadarajah, 2000, page 12). By continuity of the Gamma function on $\mathbb{R}_{++}$, it follows that $\theta\mapsto\int_0^1 Q(u|\theta)^2du$ is continuous on $\mathbb{R}\times\mathbb{R}_{++}\times(-1/2,\infty)$. Consequently, for any compact parameter space $\Theta\subseteq\mathbb{R}\times\mathbb{R}_{++}\times(-1/2,\infty)$, the uniform boundedness condition in

[1] For a positive (semi)definite symmetric matrix, eigenvalues and singular values coincide. Consequently, $\|A_0^{-1}\|=\frac{1}{\lambda_{\min}(A_0)}$, which is bounded above by Assumption 8.

Assumption 4 will be satisfied with $0 = \underline{p} < \overline{p} = 1$. In addition, if we choose trimming constants $0 < \underline{p}$ and $\overline{p} < 1$, it is possible to consider a compact parameter space $\Theta \subseteq \mathbb{R} \times \mathbb{R}_{++} \times \mathbb{R}$, since, in this case, for any $\theta \in \Theta$:

$$\int_{\underline{p}}^{\overline{p}} Q(u|\theta)^2 du \leq \left(\overline{p} - \underline{p}\right) \left(Q(\underline{p}|\theta)^2 \vee Q(\overline{p}|\theta)^2\right) ,$$

with both $Q(\overline{p}|\theta)$ and $Q(\overline{p}|\theta)$ continuous in θ.

B.2. **Generalized Pareto.** Consider a Generalized Pareto Distribution with location parameter $\xi \in \mathbb{R}$, scale parameter $\alpha > 0$ and shape parameter $k \in \mathbb{R}$. Denoting by $\theta = (\xi, \alpha, k)'$ the vector of parameters, we have that the quantile function is given by (Hosking, 1986, page 67):

$$Q(u|\theta) = \begin{cases} \xi + \alpha(1 - (1-u)^k)/k\,, & k \neq 0 \\ \xi - \alpha \log(1-u)\,, & k = 0 \end{cases} ,$$

If $k > -1/2$, we have that (Hosking and Wallis, 1987):

$$\int_0^1 Q(u|\theta)^2 du = \left(\xi + \frac{\alpha}{(1+k)}\right)^2 + \frac{\alpha^2}{(1+k)^2(1+2k)} ,$$

from which it follows that the boundedness condition in Assumption 4 is satisfied with $0 = \underline{p} < \overline{p} = 1$ for a parameter space $\Theta \subseteq \mathbb{R} \times \mathbb{R}_{++} \times (-1/2, \infty)$ such that $\Pi_{1,2}\Theta = \{(\theta_1, \theta_2) : \theta \in \Theta\}$ is compact and $\inf\{\theta_3 : \theta \in \Theta\} > -1/2$. Moreover, and similarly to the GEV case, if one considers a trimming constant $\overline{p} < 1$, then it is possible to consider parameter spaces $\Theta \subseteq \mathbb{R} \times \mathbb{R}_{++} \times \mathbb{R}$ such that $\{(\theta_1, \theta_2) : \theta \in \Theta\}$ is compact and $\Pi_3\Theta = \{\theta_3 : \theta \in \Theta\}$ is bounded below, since, in this case, $\int_0^{\overline{p}} Q(u|\theta)^2 du \leq \overline{p}\left(Q(\overline{p}|\theta)^2 \vee Q(0|\theta)^2\right) = \overline{p}\left(Q(\overline{p}|\theta)^2 \vee \xi^2\right)$, with $Q(\overline{p}|\theta)^2 \leq \left(\xi + \frac{\alpha}{k}\right)^2$ if $k > 0$.

Appendix C. Relation between Assumptions and different notions of identification

C.1. **Relation between the strong identifiability part of Assumption 4 and the usual notion of identifiability in parametric families.** In what follows, consider the population version of the objective function stated in Assumption 4 in the main text, with the choice $0 = \underline{p} < \overline{p} = 1$.

$$V_R^*(\theta) := \left[\int_0^1 \left(Q_Y(u|\theta) - Q_Y(u|\theta_0)\right) \mathbf{P}^R(u)' du\right] \Omega^R \left[\int_0^1 \left(Q_Y(u|\theta) - Q_Y(u|\theta_0)\right) \mathbf{P}^R(u) du\right] .$$

Proposition C.1. *Suppose that Θ is compact, that $\theta \mapsto \int_0^1 Q(u|\theta)^2 du$ is bounded and $(\theta', \theta'') \mapsto \int_0^1 (Q(u|\theta') - Q(u|\theta''))^2 du$ is continuous, that the $\{P_l\}_l$ form an orthonormal basis in $L^2[0,1]$, and that the smallest eigenvalue of Ω_R is bounded away from zero, uniformly in R. Then the parametric family θ_0 is identified in the usual sense (meaning $\theta \neq \theta_0 \implies F_\theta \neq F_{\theta_0}$) if, and only if, for every $\epsilon > 0$:*

$$\liminf_{R \to \infty} \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} V_R^*(\theta) > 0$$

Proof. Suppose that θ_0 is not identified in the usual sense. Then there exists $\tilde{\theta} \in \Theta$ such that $\tilde{\theta} \neq \theta$ and $F_{\tilde{\theta}} = F_{\theta_0}$. Consequently, $Q(\cdot|\theta_0) = Q(\cdot|\tilde{\theta})$, and, taking $\epsilon^* = \|\tilde{\theta} - \theta_0\| > 0$, we have:

$$\inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon^*} V_R^*(\theta) = 0\,, \quad \forall R \in \mathbb{N}\,.$$

In the other direction, suppose that θ_0 is identified in the usual sense. Fix $\epsilon > 0$. Since $V^*(\theta) := \int_0^1 (Q(u|\theta) - Q(u|\theta_0))^2 du$ is continuous and bounded and $\{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}$ is compact, identifiability in the usual sense implies that $\inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} V^*(\theta) > 0$.[2] Moreover, since Ω_R is symmetric and real, it admits an eigendecomposition $D_R \Lambda_R D_R'$, where Λ_R a diagonal matrix with the eigenvalues of Ω_R, and $D_R' D_R = D_R D_R' = \mathbb{I}_R$. This implies that, denoting by $D_{l,R}$ the l-th column of D_R and defining $v_R(\theta) = \int_0^1 \left(Q_Y(u|\theta) - Q_Y(u|\theta_0)\right) \mathbf{P}^R(u) du$:

$$V_R^*(\theta) = \sum_{l=1}^R \lambda_{l,R} (D_{l,R}' v_R(\theta))^2 \geq \underline{\lambda} \sum_{l=1}^R (D_{l,R}' v_R(\theta))^2 = \underline{\lambda} v_R(\theta)' \left(\sum_{l=1}^R D_{l,R} D_{l,R}' \right) v_R(\theta) = \underline{\lambda} \|v_R(\theta)\|^2\,,$$

where $\underline{\lambda} > 0$ is the uniform lower bound on the eigenvalues of the Ω_R. Now, since the $\{P_l\}_{l \in \mathbb{N}}$ form an orthonormal basis in $L^2[0,1]$, it follows from Parseval identity (Kreyszig, 1989, page 170) that, for every $\theta \in \Theta$, as $R \to \infty$, $\|v_R(\theta)\|^2 \uparrow V^*(\theta)$. Moreover, notice that, by the mean-value theorem, for every $\theta', \theta'' \in \Theta$:

$$\begin{aligned} |\|v_R(\theta')\|^2 - \|v_R(\theta'')\|^2| \leq 2\tilde{C}_{\theta',\theta''} |\|v_R(\theta')\| - \|v_R(\theta'')\|| \leq 2C^* \|v_R(\theta') - v_R(\theta'')\| \leq \\ 2C^* \sqrt{\int_0^1 (Q(u|\theta') - Q(u|\theta''))^2 du}\,, \end{aligned} \tag{2}$$

where $\tilde{C}_{\theta',\theta''} \in [\|v_R(\theta')\|, \|v_R(\theta'')\|]$, with $\|v_R(\theta')\| \vee \|v_R(\theta'')\| \leq \sup_{\theta \in \Theta} \sqrt{\int_0^1 Q(u|\theta)^2 du} =: C^*$ by Bessel inequality; and where the last inequality in (2) follows again by Bessel inequality. From (2), we conclude that the functions $\theta \mapsto \|v_R(\theta)\|^2$, $R \in \mathbb{N}$, are uniformly equicontinuous. Moreover, this sequence of functions is also uniformly bounded, as they converge pontwise monotonically to the bounded function V^*. Consequently, it follows by the Arzelà-Ascoli theorem that these functions converge uniformly to V^*, implying that:

$$\lim_{R \to \infty} \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} \|v_R(\theta)\|^2 = \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} V^*(\theta) > 0\,,$$

thus yielding that $\liminf_{R \to \infty} \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} V_R^*(\theta) > 0$.

□

C.2. **Relation between eigenvalue assumption and identification.** The goal of this section is to show how Assumption 8 is related to identification. We consider a stronger version of Assumption 4 as follows:

[2] If not, there would exist $\tilde{\theta} \neq \theta_0$ such that $V^*(\theta) = 0 \implies F_{\tilde{\theta}} = F_{\theta_0}$.

Assumption C.1. There exists $C > 0$ and $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that, for every $R \in \mathbb{N}$ and $\epsilon > 0$:

$$\inf_{\theta \in \Theta : \|\theta - \theta_0\|_2 \geq \epsilon} \left[\int_{\underline{p}}^{\bar{p}} \left(Q_Y(u|\theta) - Q_Y(u|\theta_0) \right) \mathbf{P}^R(u)' du \right] \Omega^R \left[\int_{\underline{p}}^{\bar{p}} \left(Q_Y(u|\theta) - Q_Y(u|\theta_0) \right) \mathbf{P}^R(u) du \right] \geq C h(\epsilon) \,,$$

where $h(x) > 0$ for all $x > 0$ and $\lim_{x \to 0} \frac{h(x)}{x^2} = 1$.

It is clear that Assumption C.1 implies Assumption 4. Perhaps less obviously, Assumption C.1 implies Assumption 8 under conditions that allow differentiability under the integral sign (Assumption 5).

Proposition C.2. *Suppose Assumption 5 holds. Then Assumption C.1 implies Assumption 8.*

Proof. Suppose Assumption C.1 holds. Fix $\iota \in \mathbb{R}^d$, $\|\iota\|_2 = 1$. We then have that, by Assumption C.1:

$$\left[\frac{1}{\epsilon} \int_{\underline{p}}^{\bar{p}} \left(Q_Y(u|\theta_0 + \epsilon\iota) - Q_Y(u|\theta_0) \right) \mathbf{P}^R(u)' du \right] \Omega^R \left[\frac{1}{\epsilon} \int_{\underline{p}}^{\bar{p}} \left(Q_Y(u|\theta_0 + \epsilon\iota) - Q_Y(u|\theta_0) \right) \mathbf{P}^R(u) du \right] > C \frac{h(\epsilon)}{\epsilon^2} \,.$$

Taking limits yields that:

$$\iota' \nabla_{\theta'} h^R(\theta_0)' \Omega^R \nabla_{\theta'} h^R(\theta_0) \iota \geq C \,.$$

Now, since $\nabla_{\theta'} h^R(\theta_0)' \Omega^R \nabla_{\theta'} h^R(\theta_0)$ is symmetric and real, it admits an eigendecomposition $P_R \Lambda_R P_R'$, where $P_R P_R' = P_R' P_R = \mathbb{I}_d$ and $\Lambda_R = \text{diag}(\lambda_{1R}, \lambda_{2R} \ldots \lambda_{dR})$, with $\lambda_{1R} \leq \lambda_{2R} \ldots \leq \lambda_{dR}$ being the eigenvalues of $\nabla_{\theta'} h^R(\theta_0)' \Omega^R \nabla_{\theta'} h^R(\theta_0)$. This in turn implies that:

$$\lambda_{1R} = \min_{x : \|x\|_2 = 1} x' \Lambda_R x = \min_{u : \|u\|_2 = 1} (P_R' u)' \Lambda_R (P_R' u) \geq C > 0 \,,$$

which proves the desired result. □

Appendix D. Comparison with series-IV estimator of Donald et al. (2003)

In this Appendix, we compare our proposed generalised L-moment estimator with the series-IV estimator introduced by Donald et al. (2003) in the context of inference based on conditional moment restrictions. Specifically, consider a scalar parameter $\beta_0 \in \mathbb{R}$ that is identified through a conditional moment restriction of the form:

$$\mathbb{E}[Z(\beta_0)|X] = 0 \,,$$

and suppose one has access to a random sample $\{(Z_t(\cdot), X_t)\}_{t=1}^T$ from $(Z(\cdot), X)$, with $Z(b)$ having finite second moment for every $b \in \mathbb{R}$, and $\mathbb{V}[Z(b)|X] \leq C_b$ for some $C_b \in \mathbb{R}$ and every $b \in \mathbb{R}$. Let $\{p_l(\cdot)\}_{l \in \mathbb{N}}$ be a sequence of series transformations that is able to approximate $\mathbb{E}[Z(b)|X]$ in mean-squared error, for any $b \in \mathbb{R}$. For $R \in \mathbb{N}$, we denote by $P_R(X_i) = (p_1(X_i), p_2(X_i), \ldots, p_R(X_i))'$ and $\boldsymbol{P}_R = \begin{bmatrix} P_R(X_1) & P_R(X_2) & \ldots & P_R(X_T) \end{bmatrix}'$. We assume that the basis functions are "orthogonalised", in the sense that $\boldsymbol{P}_R' \boldsymbol{P}_R = \mathbb{I}_R$. Do further define $\boldsymbol{z}(b) = (Z_1(b), \ldots Z_T(b))'$, and

$\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \dots X_T \end{bmatrix}'$. Donald et al.'s series-IV approach to estimating β_0 consists in minimizing the following criterion function:

$$\hat{S}(b) = \sum_{l=1}^{R} \left(\sum_{t=1}^{T} Z_t(b) p_l(X_t) \right)^2 = \boldsymbol{z}(b)' \boldsymbol{P}_R \boldsymbol{P}_R' \boldsymbol{z}(b) \,.$$

D.1. **Comparison between consistency arguments.** The consistency argument in Theorem 5.1. of Donald et al. relies on showing that, as $T, R \to \infty$:

$$\hat{s}(b) \coloneqq \frac{\hat{S}(b)}{T} \xrightarrow{p} \mathbb{E}[(\mathbb{E}[Z(b)|X])^2] =: s_0(b) \,.$$

The error of estimating $s_0(b)$ by $\hat{s}(b)$ can be decomposed into three terms:

$$s_0(b) - \hat{s}(b) = \left(\mathbb{E}[\mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}]^2] - \frac{1}{T} \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}]' \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] \right) + \frac{1}{T} \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}]' (\mathbb{I}_T - \boldsymbol{P}_R \boldsymbol{P}_R') \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] + \frac{1}{T} \left(\mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}]' \boldsymbol{P}_R \boldsymbol{P}_R' \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] - \boldsymbol{z}(b)' \boldsymbol{P}_R \boldsymbol{P}_R' \boldsymbol{z}(b) \right) .$$

The above error consists of three parts. The first term is $o_p(1)$ by the law of large numbers. The second term may be seen as an "approximation bias" component and is $o_p(1)$ by the series transformation approximation property.[3] The third term may be seen as a "variance component", whose proper control imposes restrictions on the rate of growth of R. Indeed, control of this term depends crucially on showing that:

$$\tilde{C} \coloneqq \frac{1}{T} \left((\mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] - \boldsymbol{z}(b))' \boldsymbol{P}_R \boldsymbol{P}_R' (\mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] - \boldsymbol{z}(b)) \right)$$

is $o_P(1)$. By the cyclic invariance of the trace operator, we have:

$$\mathbb{E}\left[\tilde{C}|\boldsymbol{X}\right] = \operatorname{tr}\left(\frac{1}{T} \left(\mathbb{E}\left[(\mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] - \boldsymbol{z}(b)) (\mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] - \boldsymbol{z}(b))' \middle| \boldsymbol{X} \right] \boldsymbol{P}_R \boldsymbol{P}_R' \right) \right) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{V}[Z_t(b)|X_t] (\boldsymbol{P}_R \boldsymbol{P}_R')_{tt} \leq \frac{C_b}{T} \sum_{t=1}^{T} (\boldsymbol{P}_R \boldsymbol{P}_R')_{tt} = \frac{C_b K}{T} \,,$$

which shows that the rate condition $\frac{R}{T} \to 0$ is sufficient to ensure that $\mathbb{E}\left[\tilde{C}\right] = o(1)$. Moreover, if $\mathbb{V}[Z(b)|X]$ is bounded *below* by a constant $c_b > 0$, then the rate condition is also necessary for $\mathbb{E}\left[\tilde{C}\right] = o(1)$.

[3] Indeed, since, by assumption, there exists a sequence $\gamma_K \in \mathbb{R}^K$, $K \in \mathbb{N}$, such that $\lim_{K \to \infty} \mathbb{E}[|Z(b) - \gamma_K' P_K(X)|^2] = 0$, and given that $\mathbb{I}_T - \boldsymbol{P}_R \boldsymbol{P}_R'$ is a residual-maker matrix, it follows that:

$$\mathbb{E}\left[\left| \frac{1}{T} \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}]' (\mathbb{I}_T - \boldsymbol{P}_R \boldsymbol{P}_R') \mathbb{E}[\boldsymbol{z}(b)|\boldsymbol{X}] \right| \right] \leq \mathbb{E}[|Z(b) - \gamma_R' P_R(X)|^2] = o(1)$$

By Markov inequality, we conclude that $\frac{R}{T} \to 0$ is a sufficient condition to ensure that $\tilde{C} = o_p(1)$. In addition, if $Z(b)$ has finite fourth moment and $\mathbb{V}[Z(b)|X]$ is bounded below by a constant $c_b > 0$, the rate requirement is also necessary for $\tilde{C} = o_p(1)$, since, in this case, there exists a constant $\Lambda > 0$ not depending on T or R such that $\mathbb{E}[\tilde{C}^2] \leq \Lambda$. This implies that $\tilde{C}$ is uniformly integrable (Durrett, 2019, Theorem 4.6.2) and hence, if $\tilde{C} = o_p(1)$, one has that $\mathbb{E}[\tilde{C}] = o(1)$ (Durrett, 2019, Theorem 4.6.3), thus implying that $\frac{R}{T} \to 0$ under the lower bound in the conditional variance.

The previous discussion evidences that, in the context of the series-IV estimator of Donald et al. (2003), the rate condition $R/T \to 0$ is essential for consistency. Why does our proposed generalised L-moment estimator not require such rate restriction? An inspection of the proof of Proposition 1 presented in Appendix A.1 reveals that the "analogous" term to $\tilde{C}$ in this case is

$$\tilde{D} \coloneqq \left\| \int_{\underline{p}}^{\bar{p}} \left[\left(\hat{Q}_Y(u) - Q_Y(u|\theta) \right) - \left(Q_Y(u) - Q_Y(u|\theta) \right) \right] \mathbf{P}^R(u) \, du \right\|_2^2 ,$$

where we have assumed an identity weighting matrix for the sake of clarity and comparability with the series-IV estimator. Now, by Bessel's inequality, we have that:

$$\tilde{D} \leq \| (\hat{Q}_Y(\cdot) - Q_Y(\cdot)) \mathbb{1}_{[\underline{p},\bar{p}]} \|^2_{L^2[0,1]} ,$$

where crucially the upper bound does not depend on R. Consequently, if $\| (\hat{Q}_Y(\cdot) - Q_Y(\cdot)) \mathbb{1}_{[\underline{p},\bar{p}]} \|^2_{L^2[0,1]} \xrightarrow{p} 0$ (which is implied by uniform consistency of sample quantiles on $[\underline{p}, \bar{p}]$), then $\tilde{D} = o_p(1)$. Therefore, it is the special structure of L-moments that enables us to dispense with rate requirements.

D.2. **Comparison between linearisation arguments.** The linearisation argument underlying the proof of Theorem 5.2 of Donald et al. (2003) assumes the rate restriction $R/T^2 \to 0$. Inspection of their argument reveals that this restriction is crucially used in order to ensure that term:

$$\tilde{E} \coloneqq \frac{1}{\sqrt{T}} \nabla_b \boldsymbol{z}(\beta_0)' \boldsymbol{P}_R \boldsymbol{P}_R' \boldsymbol{z}(\beta_0) ,$$

satisfies:

$$\tilde{E} = \frac{1}{\sqrt{T}} \mathbb{E}[\nabla_b \boldsymbol{z}(\beta_0) | \boldsymbol{X}]' \boldsymbol{P}_R \boldsymbol{P}_R' \boldsymbol{z}(\beta_0) + o_P(1) .$$

In other words, the rate restriction $R/T^2 \to 0$ is used in order to ensure that the "bias term" $\frac{1}{T} \mathbb{E} \left[\nabla_b \boldsymbol{z}(\beta_0)' \boldsymbol{P}_R \boldsymbol{P}_R' \boldsymbol{z}(\beta_0) \right] = \mathbb{E}[\nabla_b Z(\beta_0) \left(P_R(X)' P_R(X) \right) Z(\beta_0)] = O(R)$ is $o(T^{-1/2})$, and thus that $\tilde{E} = O_P(1)$ by application of a central limit theorem.[4] In contrast, inspection of the proof of Proposition 2 provided in Appendix A.2 shows that the analogous term is (again assuming identity weights for clarity):

[4]It has long been recognised in the literature (Newey, 1990; Donald and Newey, 2001) that, in GMM estimation, controlling this form of "own-observation-bias" stemming from correlation between the "individual gradient" at the truth $\nabla_b Z_t(\beta_0)$ and the"individual moment" at the truth $Z_t(\beta_0)$ typically requires, for asymptotic normality, stronger restrictions on the growth rate of R than consistency does.

$$\tilde{F} = -\sqrt{T}\left(\int_{\underline{p}}^{\bar{p}} \nabla_\theta Q_Y(u|\theta_0)\boldsymbol{P}_R(u)'du\right)\int_{\underline{p}}^{\bar{p}}\left(\hat{Q}_Y(u) - Q_Y(u)\right)\boldsymbol{P}_R(u)du\,.$$

Two important distinctions arise with respect to the term $\tilde{E}$. First, due to the additive separability between $\hat{Q}_Y(\cdot)$ and $Q_Y(\cdot|\theta)$ in the difference between theoretical and empirical L-moments entering our estimator, the gradient present in $\tilde{F}$ is not affected by estimation error in $\hat{Q}_Y$. This contrasts with the term $\tilde{E}$, whose bias is precisely due to correlation between the gradient and the sample moments.[5] Secondly, the special structure of L-moments enables us to straightforwardly apply Bessel's inequality to show that:

$$\|\tilde{F}\|_2^2 \leq \left(\sum_{j=1}^{d}\|\partial_{\theta_j}Q_Y(\cdot|\theta_0)\mathbb{1}_{[\underline{p},\bar{p}]}\|^2_{L^2[0,1]}\right)\|\sqrt{T}(\hat{Q}_Y(\cdot) - Q_Y(\cdot))\|^2_{L^2[0,1]}\,,$$

which ensures that $\tilde{F}$ is bounded in probability if $\|\sqrt{T}(\hat{Q}_Y(\cdot) - Q_Y(\cdot))\|^2_{L^2[0,1]}$ is. Again, the special structure of L-moments allowed us to bound a crucial term without resort to rate requirements.

Appendix E. Calculations for optimal weighting matrix in the iid case

Consider the optimal weighting matrix as in equation (16) of the main text. We focus on the case where $0 = \underline{p} < \bar{p} = 1$ and the data is iid. Note that we may write:

$$\Omega^R = \mathbb{E}\left[\frac{B_T(U)}{f_{\theta_0}(Q_y(U))}\mathbf{P}^R(U)\frac{B_T(V)}{f_{\theta_0}(Q_y(V))}\mathbf{P}^R(V)\right]^-\,,$$

where U and V are independent random variables, independent from the Brownian bridge B_T. By Foubini's theorem, we have:

$$\Omega^R = \mathbb{E}\left[\frac{(U\wedge V - UV)}{f_{\theta_0}(Q_y(U))f_{\theta_0}(Q_y(V))}\mathbf{P}^R(U)\mathbf{P}^R(V)\right]^-\,.$$

Since standard L-moments consist of a choice of weighting functions $\mathbf{P}^R$ where each entry is a linear combination of polynomials, it suffices, for the purposes of numerical computation, to analyse the formula for polynomials U^r and V^s. In particular, we can estimate:

$$\Pi_{r,s} = \int_0^1\int_0^1 \frac{(U\wedge V - UV)}{f_{\theta_0}(Q_y(U))f_{\theta_0}(Q_y(V))}U^rV^s dUdV\,,$$

using a first step consistent estimator $\tilde{\theta}$ of θ_0, the empirical quantile function, and numerical integration as follows:

$$\hat{\Pi}_{r,s} = \frac{1}{H^2}\sum_{i=1}^{H}\sum_{j=1}^{H}\frac{\left[\left(\frac{i-0.5}{H}\right)\wedge\left(\frac{j-0.5}{H}\right) - \left(\frac{i-0.5}{H}\right)\left(\frac{j-0.5}{H}\right)\right]}{f_{\tilde{\theta}}\left(Q_Y\left(\left(\frac{i-0.5}{H}\right)\middle|\tilde{\theta}\right)\right)f_{\tilde{\theta}}\left(Q_Y\left(\left(\frac{j-0.5}{H}\right)\middle|\tilde{\theta}\right)\right)}\left(\frac{i-0.5}{H}\right)^r\left(\frac{j-0.5}{H}\right)^s\,,$$

[5] The absence of correlation between the Jacobian at the truth and the moments at the truth is a more general feature of minimum-distance-style estimators (Newey and Smith, 2004).

where H is the number of grid points. Alternatively, we may use a nonparametric estimator for the quantile derivative $Q'_Y(u) = \frac{1}{f_Y(Q_Y(u))}$.

Appendix F. Test statistic for overidentifying restrictions

As noted in Remark 6 in the main text, the strong approximation discussed in Section 3.4 motivates a test statistic for overidentifying restrictions. Suppose $R > d$. Denoting by $M(\cdot)$ the objective function of the estimator, we consider the test-statistic:

$$J := T \cdot M(\hat{\theta}_T) \,.$$

Under the null that the model is correctly specified, i.e. that there exists $\theta \in \Theta$ such that $Q_Y(\cdot) = Q_Y(\cdot|\theta)$, the results in Section 3.4 of the main text reveal that:

$$J = T \cdot M(\hat{\theta}_T) = \left[\int_{\underline{p}}^{\bar{p}} \sqrt{T}\left[(\hat{Q}_Y(u) - Q_Y(u)) - \nabla_{\theta'} Q_Y(u|\theta_0)(\hat{\theta} - \theta_0)\right] \mathbf{P}^R(u) du\right]' \Omega^R$$
$$\left[\int_{\underline{p}}^{\bar{p}} \sqrt{T}\left[(\hat{Q}_Y(u) - Q_Y(u)) - \nabla_{\theta'} Q_Y(u|\theta_0)(\hat{\theta} - \theta_0)\right] \mathbf{P}^R(u) du\right] + o_{P^*}(1) =$$
$$\|(\Omega^R)^{1/2}\left(\mathbb{I}_{R\times R} - \nabla_{\theta'} h^R(\theta_0)(\nabla_{\theta'} h^R(\theta_0)'\Omega^R \nabla_{\theta'} h^R(\theta_0))^{-1} \nabla_{\theta'} h^R(\theta_0)'\Omega^R\right)\sqrt{T} h_R(\theta_0)\|_2 + o_p(1) \,.$$

This approximation can be used, along with the Gaussian strong approximations discussed in this section, to approximate the distribution of the statistic under the null. Specifically, when the optimal weighting scheme (16) is used, it follows from the properties of idempotent matrices that the distribution of the test statistic may be approximated by a chi-squared distribution with $R-d$ degrees of freedom. To show that this approximation indeed conduces to valid inference, let $\tilde{J}$ denote the distribution of the leading term in the representation above, with $\sqrt{T}h_R(\theta_0)$ replaced by the Gaussian approximating random variable. Let $e_{TR} = J - \tilde{J}$ be the approximation error. It follows by Lemma S.14 of Fan et al. (2023) that, for any $\epsilon > 0$:

$$\sup_{c\in\mathbb{R}} |\mathbb{P}[J \le c] - \mathbb{P}[\tilde{J} \le c]| \le \sup_{c\in\mathbb{R}} \mathbb{P}[c \le \tilde{J} \le c + \epsilon] + \mathbb{P}[|e_{TR}| > \epsilon] \,.$$

In addition, by Theorem 2.7 in Götze et al. (2019), we have that, for some constant $C > 0$:

$$\sup_{c\in\mathbb{R}} \mathbb{P}[c \le \tilde{J} \le c + \epsilon] \le \frac{C\epsilon}{\sqrt{(R-d)}} \,.$$

Consequently:

$$\sup_{c\in\mathbb{R}} |\mathbb{P}[J \le c] - \mathbb{P}[\tilde{J} \le c]| \le \frac{C\epsilon}{\sqrt{(R-d)}} + \epsilon] + \mathbb{P}[|e_{TR}| > \epsilon] \,,$$

The approximation error e_{TR} consists of two parts: the error due to linearisation, and the error due to approximating $\sqrt{T}h_R(\theta_0)$ by a Gaussian random variable. In light of our discussion in the

main text, both errors are $o_{P^*}(1)$. As a consequence, for an arbitrary choice of $\epsilon > 0$, and as $T, R \to \infty$, we obtain that:

$$\lim_{T,R\to\infty} \sup_{c\in\mathbb{R}} |\mathbb{P}[J \leq c] - \mathbb{P}[\tilde{J} \leq c]| = 0\,,$$

justifying the validity of the chi-squared approximation when R increases with the sample size. In contrast, if R is held fixed, we have to explicitly take the rate of the error e_{TR} into account. As we have discussed, this error can be decomposed into two parts: a linearization error, and a strong approximation error. In Appendix K, we provide conditions that ensure the linearization error is $O_P(T^{-1/2})$. The rate of the Gaussian approximation error depends on the dependence between observations and the assumptions on the distribution: Theorems 1 and 2 in the main text provide rates in the iid and strongly mixing settings. Let b_T denote the rate of the second type of error, and $c_T := T^{-1/2} \vee b_T$. In this case, by taking $\epsilon = (c_T)^\alpha$ for some $0 < \alpha < 1$, we are able to establish validity of the approximation with fixed R.

Appendix G. Bootstrap-based inference

In this Appendix, we show how one can leverage the Gaussian strong approximation result presented in the main text to perform bootstrap-based inference. We focus on the iid setting. Consider the asymptotic linear representation (9). In the main text, we have shown that, under a Gaussian approximation, the term $\sqrt{T}h^R(\theta_0)$ can be approximated by the integral of a Brownian bridge. Consider, now, the alternative process:

$$A_T = -(\nabla_{\theta'} h^R(\theta_0)'\Omega^R \nabla_{\theta'} h^R(\theta_0))^{-1} \nabla_{\theta'} h^R(\theta_0)'\Omega^R \left[\int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^R(u)du\right],$$

where $\check{Q}_Y(u)$ is the quantile function associated with distribution function $\check{F}_Y(y) = \sum_{t=1}^T \Delta_t \mathbb{1}\{Y_t \leq y\}$, where $\Delta_t = \frac{Z_i}{\sum_{t=1}^T Z_t}$, and the Z_t are iid random variables, independent from the data, with $\mathbb{E}Z_t = 1$, $\mathbb{V}Z_t = 1$, and a moment generating function (MGF) that exists on a neighborhood of zero. The distribution $\check{F}_Y(y)$ constructed in such way is known as a weighted bootstrap estimator of the empirical distribution $\hat{F}_Y$. The weighted bootstrap is quite general and encompasses, among others, the Bayesian bootstrap (Rubin, 1981).

If, in addition to the conditions in Theorem 1 of the main text, we assume $\sup_{y\in(a,b)} |f'_Y(y)| < \infty$ and $A = \lim_{y\downarrow a} f_Y(y) < \infty$, $B = \lim_{y\uparrow a} f_Y(y) < \infty$ with $\min\{A, B\} > 0$, then Theorem 7 in Alvarez-Andrade and Bouzebda (2013) indicates that $\left[\int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^R(u)du\right]$ is strongly approximated by the integral of a Gaussian process that is **identically distributed** to the strong approximation of the term $\sqrt{T}h^R(\theta_0)$ obtained in the main text. Specifically, inspection of the proof in Alvarez-Andrade and Bouzebda (2013) reveals that there exists a sequence of Brownian bridges $\{\tilde{B}_n\}_{n\in\mathbb{N}}$, where each B_n is *independent* from $Y_1, Y_2, \ldots, Y_n$, such that, as $T, R \to \infty$:

$$\left\| \int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^R(u)du - \int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^R(u)\frac{1}{f_Y(Q_Y(u))}\tilde{B}_n du \right\|_2 \overset{\text{a.s.}}{=} O(\log n/\sqrt{n}) .$$

It thus follows by Markov inequality that:

$$\mathbb{P}\left[\left\| \int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^R(u)du - \int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^R(u)\frac{1}{f_Y(Q_Y(u))}\tilde{B}_n du \right\|_2 > \epsilon \middle| Y_1, \ldots, Y_T \right] = o_P(1) ,$$

for every $\epsilon > 0$. Since the Brownian bridges are **independent** from the data, the conditional convergence justifies the use of the weighted bootstrap to approximate the distribution of A_T. Indeed, given consistent estimators of θ_0 an Ω_R, we can approximate the distribution of A_T by generating a large number of simulations of the $\{Z_t\}_{t=1}^T$ and computing, for each simulation, the quantile function of the resulting weighted cdf. The distribution across simulations can then be used to approximate the distribution of the A_T.

Appendix H. Inference based on Bahadur-Kiefer representation

In this Appendix, we discuss how we can conduct inference by relying on a Bahadur-Kiefer representation. We first state the result of Kiefer, in the iid context, as extended by Csorgo and Revesz (1978).

Theorem H.1 (Bahadur-Kiefer, Csorgo and Revesz (1978))**.** *Let $Y_1, Y_2 \ldots Y_T$ be an iid sequence of random variables with a continuous distribution function F which is also twice differentiable on (a, b), where $-\infty \leq a = \sup\{z : F(z) = 0\}$ and $b = \inf\{z : F(z) = 1\} \leq \infty$. Suppose that $F'(z) = f(z) > 0$ for $z \in (a, b)$. Assume that, for $\gamma > 0$:*

$$\sup_{a<x<b} F(x)(1 - F(x)) \left| \frac{f'(x)}{f^2(x)} \right| \leq \gamma ,$$

where f denotes the density of F. Moreover, assume that f is nondecreasing (nonincreasing) on an interval to the right of a (to the left of b). We then have that

$$\begin{aligned} \sup_{0<u<1} |f(Q_Y(u))\sqrt{T}(\hat{Q}_Y(u) - Q_Y(u)) - \sqrt{T}(\hat{F}_Y(Q_Y(u)) - F(Q_Y(u))| &\overset{a.s.}{=} \\ &\overset{a.s.}{=} O(T^{-1/4}(\log T)^{1/2}(\log\log T)^{\frac{1}{4}}) . \end{aligned} \tag{3}$$

The result above could be used as the basis for an inferential procedure – as well as for the computation of the optimal weights Ω^R. Indeed, we note that, under the assumptions on the theorem above, $F(Q_Y(u)) = u$ and $\hat{F}_Y(Q_Y(u)) = \frac{1}{T}\sum_{t=1}^T \mathbb{1}\{U_t \leq u\}$, where the $U_t := F(Y_T)$ are iid uniform random variables. Suppose that $\int_{\underline{p}}^{\bar{p}} \frac{1}{f_Y(Q_Y(u))^2} du < \infty$. Then, using the representation of the theorem above in the asymptotic linear representation (9) in the main text and applying Bessel's inequality, we get:

$$\sqrt{T}(\hat{\theta}-\theta_0) =$$

$$= -(\nabla_{\theta'} h^R(\theta_0)' \Omega^R \nabla_{\theta'} h^R(\theta_0))^{-1} \nabla_{\theta'} h^R(\theta_0)' \Omega^R \left[\int_{\underline{p}}^{\bar{p}} \frac{\sqrt{T}(\hat{F}_Y(Q_Y(u)) - F_Y(Q_Y(u)))}{f_Y(Q_Y(u))} \mathbf{P}^R(u) du \right] + o_{P^*}(1) , \tag{4}$$

where the distribution of the leading term is known (it could be simulated by drawing T independent Uniform[0,1] random variables many times) up to θ_0.

There is a sizeable literature on Bahadur-Kiefer representations in the context of dependent observations (see Kulik (2007) and references therein). Nonetheless, in the context of dependent observations, it would be more difficult to use (4) as a basis for an inferential procedure, as in this case there would be dependence between the $U_t \coloneqq F_Y(Y_t)$ uniform random variables entering the empirical cdf. For that reason, our focus in this section is on the iid case.

Finally, to show the validity of our approach to inference based on drawing uniform random variables, we note that, under a Bahadur-Kiefer approximation, we have that:

$$V_{T,R}^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_t^{R,T} + o_p(1) ,$$

where $V_{T,R}$ is the variance of the leading term of the Bahadur-Kiefer representation, $\mathbb{E}[\xi_t^{R,T}] = 0$ and $\mathbb{V}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_t^{R,T}\right] = 1$. In the iid context, it is immediate that the conditions of Lindeberg's CLT for triangular arrays (Durrett, 2019, Theorem 3.4.10) are satisfied, from which it follows that $V_{T,R}^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \mathbb{I}_d)$.[6] Observe that as a byproduct of such convergence, we obtain that the Kolmogorov distance between the distribution of $\sqrt{T}(\hat{\theta_T} - \theta_0)$ and that of the leading term of the representation (3) goes to zero, analogously to the result in equation (15) in the main text. This result justifies our approach to inference.

We collect the discussion of this section in the next corollary.

Corollary H.1. *Suppose Assumptions 1-8 hold. Moreover, suppose a Bahadur-Kiefer representation such as (3) is valid; and that* $\int_{\underline{p}}^{\bar{p}} \frac{1}{f_Y(Q_Y(u))^2} du < \infty$. *Then, as* $T, R \to \infty$, *the approximation (4) holds. In addition, under the conditions of Theorem H.1,* $F_Y(Q_Y(u)) = u$ *and* $\hat{F}_Y(Q_Y(u)) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{U_t \leq u\}$, *where the* $\{U_t\}_{t=1}^{T}$ *are iid Uniform[0,1] random variables. Moreover, under the previous assumptions, and as* $T, R \to \infty$, $V_{T,R}^{-1/2} \sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \mathbb{I}_d)$, *where* $V_{T,R}$ *is the variance of the leading term in (4); and a bound analogous to equation (15) in the main text holds.*

[6] In the dependent case, even though it is not feasible to leverage the Bahadur-Kiefer representation directly for inference, it is possible to adopt it to establish, under (possibly) additional assumptions, weak convergence of the estimator, by verifying the conditions of a CLT for triangular arrays under dependent data. For example, in the stationary mixing case, one could verify if the conditions of Theorem 4.4 in Rio (2017) hold.

Remark H.1 (Optimal choice of weighting matrix under Bahadur-Kiefer approximation)**.** It should be noted that the optimal choice of weighting matrix under the Bahadur-Kiefer representation **coincides** with (16) in the main text. This is due to the fact that both the Brownian bridge and empirical distribution process share the same covariance kernel.

Remark H.2 (Distribution of the overidentifying test statistic in Remark 6)**.** Note that we could use the distributional results in this section to compute the distribution of the test statistic in Remark 6 in the main text under the null.

Appendix I. Asymptotic efficiency

In this Appendix, we analyse whether our L-moment estimator is asymptotically efficient. We consider the case where $0 = \underline{p} < \overline{p} = 1$, since in this case all information on the curve is used; for simplicity, we also focus on the iid case. In this setting, we will say our L-moment estimator is *asymptotically efficient* if its asymptotic variance coincides with the inverse of the Fisher information matrix of the parametric model. Unless stated otherwise, we work under Assumptions 1-8 and those of Corollary H.1. To proceed with the analysis, we introduce the alternative estimator:

$$\tilde{\theta}_T \in \operatorname{argmin}_{\theta\in\Theta} \sum_{i\in\mathcal{G}_T}\sum_{j\in\mathcal{G}_T} (\hat{Q}_Y(i) - Q_Y(i|\theta))\kappa_{i,j}(\hat{Q}_Y(j) - Q_Y(j|\theta))\,, \tag{5}$$

for a grid of G_T points $\mathcal{G}_T = \{g_1, g_2, \ldots, g_{G_T}\} \subseteq (0,1)$ and weights $\kappa_{i,j}$, $i,j \in \mathcal{G}_T$. This is a weighted version of a "percentile-based estimator", which is used in contexts where it is difficult to maximise the likelihood (Gupta and Kundu, 2001). It amounts to choosing θ so as to match a weighted combination of the order statistics in the sample.

Under regularity conditions similar to the ones in the main text,[7] the estimator in (5) admits the following asymptotic linear representation as $T \to \infty$ and $G_T \to \infty$ at a rate:

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = -(\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\partial Q_{G_T})^{-1}\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\sqrt{T}Q_{G_T} + o_p(1)\,, \tag{6}$$

where $Q_{G_T} = Q_{G_T}(\theta_0) = (\hat{Q}_Y(g_1) - Q_Y(g_1|\theta_0), \ldots, \hat{Q}_Y(g_T) - Q_Y(g_T|\theta_0))'$; ∂Q_{G_T} is the Jacobian matrix of $Q_{G_T}(\theta)$ evaluated at θ_0; and $\boldsymbol{\kappa}_{G_T}$ is the matrix containing the $\kappa_{i,j}$. Using the Bahadur-Kiefer representation (assumed by Corollary H.1), we arrive at:

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = -(\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\partial Q_{G_T})^{-1}\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\left[\boldsymbol{f}^{-1} * \sqrt{T}F_{G_T}\right] + o_p(1)\,, \tag{7}$$

where we define $F_{G_T} = (\hat{F}_Y(Q_Y(g_1)) - F_Y(Q_Y(g_1)), \ldots, \hat{F}_Y(Q_Y(g_{G_T})) - F_Y(Q_Y(g_{G_T})))'$; $\boldsymbol{f}^{-1} = (1/f_Y(Q_Y(g_1)), \ldots, 1/f_Y(Q_Y(g_T)))'$; and $*$ denotes entry-by-entry multiplication.

For a given $\mathcal{G}_T$, representation (7) yields the following choice of optimal weighting matrix, $\boldsymbol{\kappa}^* = \mathbb{V}[\boldsymbol{f}^{-1} * \sqrt{T}F_{G_T}]^{-1}$; and this implies that the variance of the leading term of (7) under such

[7] We omit these conditions for brevity, but we note that, using the notation in (6), since we assume $\|\sqrt{T}Q_{G_T}\|_\infty = O_p(1)$ (implied by Assumption 6), it is crucial that $\|\partial Q'_{G_T}\boldsymbol{\kappa}\|_\infty = O_p(1)$, where $\|\cdot\|_\infty$ is the operator norm induced by the vector norm. This condition can be shown to hold for the optimal choice of weights described below under some conditions. We also require a restriction on the growth rate of G_T so as to control the error of a mean-value expansion of increasing dimension.

choice is $\mathbb{V}^* = (\partial Q'_{G_T} \boldsymbol{\kappa}_{G_T} \partial Q_{G_T})^{-1}$. But, if we take the grid $\mathcal{G}_T$ as $\left\{\frac{1}{G_T+1}, \frac{2}{G_T+1}, \ldots, \frac{G_T}{G_T+1}\right\}$, it follows from Lemma C.1. in Firpo et al. (2022) that:

$$\mathbb{V}^* = ((\partial Q_{G_T} * (\mathbf{1}'_d \otimes \boldsymbol{f}))' \Sigma_{G_T}^{-1} (\partial Q_{G_T} * (\mathbf{1}'_d \otimes \boldsymbol{f}))^{-1} ,$$

where

$$(\Sigma_{G_T}^{-1})_{g_i,g_j} = \mathbb{1}_{\{g_i=g_j\}} 2(G_T+1) - (\mathbb{1}_{\{g_i=g_{j+1}\}} + \mathbb{1}_{\{g_i=g_{j-1}\}})(G_T+1) .$$

It then follows that, for $d_1, d_2 \in \{1, 2 \ldots, d\}$:

$$\begin{aligned}
\left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = \\
(G_T+1) \sum_{i=2}^{G_T} f_Y(Q_Y(g_i)) \partial_{d_1} Q_Y(g_i|\theta_0) \left[f_Y(Q_Y(g_i) \partial_{d_2} Q_Y(g_i|\theta_0) - f_Y(Q_Y(g_{i-1}) \partial_{d_2} Q_Y(g_i|\theta_0)\right] \\
-(G_T+1) \sum_{i=1}^{G_T-1} f_Y(Q_Y(g_i)) \partial_{d_1} Q_Y(g_i|\theta_0) \left[f_Y(Q_Y(g_{i+1}) \partial_{d_2} Q_Y(g_i|\theta_0) - f_Y(Q_Y(g_i) \partial_{d_2} Q_Y(g_i|\theta_0)\right] \\
+(G_T+1)(f_Y(Q_Y(g_1)))^2 \partial_{d_1} Q_Y(g_1|\theta_0) \partial_{d_2} Q_Y(g_1|\theta_0) \\
+(G_T+1)(f_Y(Q_Y(g_{G_T})))^2 \partial_{d_1} Q_Y(g_{G_T}|\theta_0) \partial_{d_2} Q_Y(g_{G_T}|\theta_0) .
\end{aligned} \tag{8}$$

Assuming the tail condition:[8]

$$\lim_{u \to 0} \frac{(f_Y(Q_Y(u)))^2 \partial_{d_1} Q_Y(u|\theta_0) \partial_{d_2} Q_Y(u|\theta_0) + (f_Y(Q_Y(1-u)))^2 \partial_{d_1} Q_Y(1-u|\theta_0) \partial_{d_2} Q_Y(1-u|\theta_0)}{u} = 0 , \tag{9}$$

leads to the last two terms of (8) being asymptotically negligible as $T \to \infty$. If we further assume the $u \mapsto f_Y(Q(u)) \partial_{d_1} Q_Y(u|\theta_0)$ are differentiable uniformly on $(0,1)$, it follows from Riemann integration that:

$$\lim_{T \to \infty} \left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = \int_0^1 \left. \frac{d\left[f_Y(Q_Y(v)) \partial_{d_1} Q_Y(v|\theta_0)\right]}{dv} \right|_{v=u} \left. \frac{d\left[f_Y(Q_Y(v)) \partial_{d_2} Q_Y(v|\theta_0)\right]}{dv} \right|_{v=u} du .$$

But then, from the relation:

$$F_Y(Q_Y(u|\theta)|\theta) = u \implies f_Y(Q_Y(u|\theta)) \partial_d Q_Y(u|\theta) = -\partial_d F_Y(Q_Y(u)|\theta) ,$$

it follows, by exchanging the order of differentiation:

$$\begin{aligned}
\left. \frac{d\left[f_Y(Q_Y(v)) \partial_{d_1} Q_Y(v|\theta_0)\right]}{dv} \right|_{v=u} = -\partial_d \left[f_Y(Q_Y(u)|\theta) \cdot \left. \frac{dQ_Y(v)}{dv} \right|_{v=u} \right] \Bigg|_{\theta=\theta_0} = \\
-\partial_d f_Y(Q_Y(u)|\theta_0) \cdot \frac{1}{f_Y(Q_Y(u))} ,
\end{aligned}$$

and, using the quantile representation of a random variable, we conclude that:

$$\lim_{T \to \infty} \left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = (I(\theta_0))_{d_1,d_2} ,$$

[8] A similar tail condition is considered in a working paper version of Firpo et al. (2022).

where $I(\theta) = \mathbb{E}[\nabla_\theta \log(f(Y|\theta))\nabla_{\theta'} \log(f(Y|\theta))]$ is the Fisher information matrix. We have thus shown that, under the proposed grid and optimal weights, the asymptotic variance of the leading term of the first order representation (7) of estimator (5) converges to the inverse Fisher information. We summarise this point in the lemma below.

Lemma I.1. *Consider the estimator (5). Assume that representation (7) holds. If, in addition, the tail condition (9) and the uniform differentiability condition in the text holds; then the variance of the leading term of representation (7) converges to the inverse Fisher information.*

How does the previous estimator relate to our L-moment estimator? Notice that, if the $\{P_l\}_{l\in\mathbb{N}}$ form an orthonormal **basis** on $L^2[0,1]$, then, for $X \in L^2[0,1]$:

$$X(u) = \sum_{l=1}^{\infty} \left(\int_0^1 X(s) P_l(s) ds \right) P_l(u) \ .$$

Therefore, since $\{Q_Y(\cdot|\theta) : \theta \in \Theta_0\} \subseteq L^2[0,1]$,[9] we have:

$$\sum_{i\in\mathcal{G}_T} \sum_{j\in\mathcal{G}_T} (\hat{Q}_Y(i) - Q_Y(i|\theta)) \kappa^*_{i,j} (\hat{Q}_Y(j) - Q_Y(j|\theta)) =$$

$$\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \left[\int_0^1 (\hat{Q}_Y(u) - Q_Y(u|\theta)) P_k(u) du \right] \tilde{\kappa}_{k,l} \left[\int_0^1 (\hat{Q}_Y(u) - Q_Y(u|\theta)) P_l(u) du \right] =: \tilde{A}^{\infty}_T(\theta) \ ,$$

which shows that the optimal estimator we described is a generalised L-moment estimator which uses infinitely many L-moments and suitable weights $\tilde{\kappa}_{k,l} = \sum_{i\in\mathcal{G}_T} \sum_{j\in\mathcal{G}_T} P_k(i) \kappa^*_{i,j} P_l(j)$. Consider an alternative L-moment estimator that uses only the first R L-moments and weights $\tilde{\boldsymbol{\kappa}}_R = (\tilde{\kappa}_{i,j})_{i,j=1,\dots R}$. Denote the estimator by $\check{\theta}_T$, and its objective function by $A^R_T(\theta)$. It can be shown that, for an identifiable parametric family and $\|\tilde{\boldsymbol{\kappa}}\|_2 = O(1)$, $\nabla_\theta A^R_T(\check{\theta}_T) = \nabla_\theta A^{\infty}_T(\check{\theta}_T) + o_{p^*}(T^{-1/2})$. This shows that the estimator admits the same first order representation as (7); and from the previous lemma we know the variance of the leading term of this representation converges to $I(\theta_0)^{-1}$. It thus follows that $\check{\theta}_T$ is asymptotically efficient.[10] But then, since the optimal weights (16) minimise the variance of the leading term in (4) (recall Remark H.1), we conclude that they too must, asymptotically, yield a variance equal to $I(\theta_0)^{-1}$. This shows that the generalised L-moment estimator is efficient, in the sense that its asymptotic variance coincides with $I(\theta_0)^{-1}$.

We collect the discussion of this section in the corollary below:

Corollary I.1. *Suppose the conditions of the previous lemma hold. Suppose the $\{P_l\}_{l\in\mathbb{N}}$ constitute an orthonormal **basis**. Consider the estimator $\check{\theta}_T$ defined in the main text. Suppose that Assumptions 1-8 and those of Corollary H.1 hold with $W^R = \Omega^R = \boldsymbol{\kappa}_R$. We then have that, for*

[9] This is implied by Assumption 4.

[10] Let Ψ_T denote the variance of the leading term of representation (4) of the estimator $\check{\theta}$. By weak convergence (the last part of Corollary H.1) and Fatou's lemma, it follows that $\liminf_{T\to\infty} \xi'(\Psi_T^{-1/2} M_T \Psi_T^{-1/2} - \mathbb{I}_d)\xi \geq 0$ for any $\xi \in \mathbb{R}^d$, where $\lim_T M_T = I(\theta_0)^{-1}$. It then follows that $\lim_{T\to\infty} \Psi_T = I(\theta_0)^{-1}$, from which we conclude that $\check{\theta}$ is asymptotically efficient.

an identifiable parametric family:

$$\lim_{T,R\to\infty} V^*_{T,R} = I(\theta_0)^{-1} \ ,$$

where $V^*_{T,R}$ *is the variance of the leading term of* (4) *under the optimal choice of weights, i.e.* $V^*_{T,R} = (\nabla_{\theta'} h^R(\theta_0)' \Omega^*_R \nabla_{\theta'} h^R(\theta_0))^{-1}$*, where* Ω^*_R *is given by* (16).

Remark I.1 (Related estimators)**.** Similarly to (5), we can show that estimators based on minimising the objective functions:

$$\begin{aligned} W^1(\theta) &:= \int_0^1 w(u)(\hat{Q}_Y(u) - Q_Y(u|\theta))^2 du \ , \\ W^2(\theta) &:= \int_0^1 \int_0^1 (\hat{Q}_Y(u) - Q_Y(u|\theta)) w(u,v) (\hat{Q}_Y(v) - Q_Y(v|\theta)) dv du \ , \end{aligned} \tag{10}$$

are also L-moment-based estimators which use infinitely many L-moments. A similar argument as the one in this section then shows that our generalised method of L-moments estimator under optimal weights will be at least as efficient as estimators based on minimising (10). However, given that we are able to control the number of L-moments used in estimation in finite samples, it is expected that our method will lead to nonasymptotic performance gains.

Appendix J. Monte Carlo exercise: additional results

J.1. **Results for linear combinations of parameters.** In this Appendix, we revisit the data-generating processes of the Monte Carlo exercises in Section 4 of the main text, but now consider linear combinations $\delta'\theta_0$, $\delta \in \mathbb{R}^d$, as the target parameters. Specifically, for a given L-moment-based estimator $\hat{\theta}_R$ and linear combination $\delta \in \mathbb{R}^d$, the relative RMSE, vis-à-vis the MLE, is given by:

$$\text{RRMSE}(\delta, \hat{\theta}_R) := \sqrt{\frac{\mathbb{E}[(\delta'\hat{\theta}_R - \delta'\theta_0)^2]}{\mathbb{E}[(\delta'\hat{\theta}_{\text{MLE}} - \delta'\theta_0)^2]}} = \sqrt{\frac{\delta'(\mathbb{E}[(\hat{\theta}_R - \theta_0)\hat{\theta}_R - \theta_0)']\delta}{\delta'\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta_0)(\hat{\theta}_{\text{MLE}} - \theta_0)'])\delta}} \ ,$$

Since we have no direct interest in any particular linear combination δ, we consider the relative RMSE under the most and least favourable values (directions) of δ. These are defined, respectively, as:

$$\underline{\text{RRMSE}(\hat{\theta}_R)} := \min_{\delta\in\mathbb{R}^d:\delta\neq 0} \text{RRMSE}(\delta, \hat{\theta}_R) =$$

$$\sqrt{\lambda_{\min}(\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta_0)(\hat{\theta}_{\text{MLE}} - \theta_0)']^{-1/2}\mathbb{E}[(\hat{\theta}_R - \theta_0)(\hat{\theta}_R - \theta_0)']\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta_0)(\hat{\theta}_{\text{MLE}} - \theta_0)']^{-1/2})}$$

$$\overline{\text{RRMSE}(\hat{\theta}_R)} := \max_{\delta\in\mathbb{R}^d:\delta\neq 0} \text{RRMSE}(\delta, \hat{\theta}_R) =$$

$$\sqrt{\lambda_{\max}(\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta_0)(\hat{\theta}_{\text{MLE}} - \theta_0)']^{-1/2}\mathbb{E}[(\hat{\theta}_R - \theta_0)(\hat{\theta}_R - \theta_0)']\mathbb{E}[(\hat{\theta}_{\text{MLE}} - \theta_0)(\hat{\theta}_{\text{MLE}} - \theta_0)']^{-1/2})}$$

Tables J.1 and J.2 report the relative RMSE of the four estimators considered in the main text, under the most and least favourable directions δ, with the choice of R that minimizes the most (least) favourable-RMSE (this choice is reported in parentheses), respectively in the GEV and GPD exercises. Values above 1 indicate that the MLE outperforms the L-moment estimator, under the stated choice of R, in the corresponding (most or least favourable) direction. In the GEV design, two-step L-moment estimators are able to offer improvements over the MLE in the most favourable directions in smaller samples sizes (with RMSE improvements of nearly 8%), while severely mitigating losses of first-step estimators in the least favourable directions. Indeed, first-step estimators in the GEV exercise perform quite poorly in the least-favourable direction, especially in the largest sample size, with root-mean-squared errors over 25% larger than the MLE. In contrast,the càglàd two-step estimators has RMSE only 1.3% larger than the MLE in the least favourable direction and largest sample size.

Similarly to the main text, in the GPD design, first-step and two-step estimators perform well relatively to the MLE, in both the most and least favourable directions, even in the largest sample size. Gains of the L-moment approach can reach 16% in the smallest sample size and most-favourable direction (4% in the smallest sample size and least favourable direction). Finally, we observe that, consistent with our theoretical results, the MSE-minimising number of L-moments for two-step estimator increases with sample size when we consider the least favourable direction, in both designs.

TABLE J.1. GEV : maximal and minimal relative RMSE for linear combintations of parameters (MSE-minimising choice of R)

	Most favourable δ			Least favourable δ		
	T= 50	T= 100	T= 500	T= 50	T= 100	T= 500
Càglàd FS	0.957	0.993	1.000	1.077	1.142	1.253
	(3)	(3)	(5)	(3)	(3)	(3)
Càglàd TS	0.922	0.966	0.999	1.032	1.016	1.013
	(11)	(11)	(3)	(4)	(5)	(30)
Unbiased FS	0.929	0.970	0.995	1.113	1.172	1.259
	(3)	(5)	(5)	(3)	(3)	(3)
Unbiased TS	0.928	0.969	0.994	1.037	1.013	1.018
	(3)	(3)	(5)	(5)	(7)	(20)

J.2. **Comparison with trimming approaches.** In this Appendix, we compare our L-moment-based approach with maximum likelihood estimators that attempt to control the influence of extreme observations. We consider two approaches. In one of the approaches, we first estimate the model parameters via MLE, here denoted by $\tilde{\theta}$. We then discard (trim) those observations Y_t such that $Y_t > Q(1-\epsilon|\tilde{\theta})$, where ϵ is a trimming proportion parameter. We then reestimate the model via MLE in the restricted set. We label this approach "trimmed MLE".[11] We also consider

[11]Our trimmed MLE approach may be seen as a one-step approximation to more complex trimming approaches where the indices of the discarded observations and the model parameters θ are simultaneously estimated (e.g. Hadi and Luceño, 1997; Awasthi et al., 2022).

TABLE J.2. GPD : maximal and minimal relative RMSE for linear combintations of parameters (MSE-minimising choice of R)

	Most favourable δ			Least favourable δ		
	T= 50	T= 100	T= 500	T= 50	T= 100	T= 500
Càglàd FS	0.895	0.931	0.988	1.004	1.004	1.004
	(6)	(4)	(4)	(3)	(3)	(6)
Càglàd TS	0.874	0.912	0.977	0.996	1.001	1.002
	(3)	(3)	(3)	(3)	(3)	(7)
Unbiased FS	0.838	0.902	0.980	0.960	0.982	1.000
	(6)	(4)	(3)	(3)	(3)	(7)
Unbiased TS	0.838	0.878	0.962	0.960	0.977	0.997
	(3)	(5)	(34)	(2)	(10)	(38)

an alternative, "tilted MLE" approach (Choi et al., 2000), that seeks to find θ by maximizing the following quantity over Θ:

$$\sup_{(p)_i \in \mathcal{P}_\epsilon} \sum_{i=1}^{T} p_i \log(f(Y_i|\theta)) \,,$$

where $\mathcal{P}_\epsilon$ is the subset of the simplex Δ^{T-1} such that $D_{\mathrm{KL}}((p_i)_i||(1/n)_i) = -\log(1-\epsilon)$, with $D_{\mathrm{KL}}((p_i)_i||(1/n)_i)$ denoting the Kullback-Leibler divergence of the uniform distribution on the indices $\{1, 2 \ldots, T\}$ from the distribution $(p_i)_{i=1}^T$ over $\{1, 2 \ldots, T\}$. The estimator amounts to running the MLE in a "reweighted" dataset, where the weights are chosen in order to minimize the KL divergence of the model from the (reweighted) data, subject to the constraint that weights are not far astray from the untilted empirical distribution. As argued by Choi et al. (2000), in the context of a data contamination model, the parameter $\epsilon \in [0, 1)$ amounts to the proportion of observations allowed to be corrupted, i.e. that do not follow the parameteric model f_θ of interest.

Tables J.3 and J.4 replicate the results for the Càglàd TS estimator under the MSE-minimizing choice of R presented in Tables 1 and 2 in the main text, and compare it with the trimmed and tilted MLE approaches. For the trimming/tilting proportion ϵ, we consider values $\epsilon \in \{0.1, 0.01, 0.001\}$. For the tilted MLE, we also report in parantheses the percentage of cases where the method for finding the estimator proposed in Choi et al. (2000) did *not* converge. We observe that the trimming approach never compares fabourably to the Càglàd TS estimator under the MSE-minimizing choice of R. As for the tilted MLE, it is able to compete with the Càglàd TS estimator for some combinations of tail quantiles and sample sizes, under a suitable choice of ϵ. However, it is important to note that the competitiviness and overall performance of this method is **very** dependent on the trimming fraction ϵ. For example, in the GEV design with $\tau = 0.999$ and $T = 500$, the relative RMSE changes from 1.6% with $\epsilon = 0.1\%$ to 76.9% when $\epsilon = 1\%$. Such sensitivity limits applicability of this approach in these designs, especially since, to the best of our knowledge, there does not exist any method to tune the tilting proportion ϵ with an aim to obtain RMSE gains over the MLE. [12]

[12] Choi et al. (2000) propose a heuristic to select ϵ which consists in computing the QQ-plot that compares the reweighted empirical quantiles with the parametric quantiles obtained from the trimmed MLE estimator. The

TABLE J.3. GEV : comparison with Trimmed and Tilted MLE methods

	$T = 50$				$T = 100$				$T = 500$			
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$
Càglàd TS	1.005	0.960	0.818	0.692	1.003	0.981	0.910	0.840	1.004	0.998	0.990	0.979
	(12)	(3)	(5)	(5)	(11)	(3)	(5)	(5)	(30)	(4)	(90)	(90)
Trimming 10%	1.193	2.177	1.505	0.886	1.374	3.033	2.367	1.619	2.266	6.677	5.847	4.548
Trimming 1%	1.005	1.093	1.020	0.950	1.006	1.169	1.075	0.931	1.013	1.723	1.813	1.645
Trimming 0.1%	1.000	1.006	1.003	1.000	1.001	1.019	1.015	1.004	1.002	1.053	1.063	1.037
Tilting 10%	1.846	2.258	1.223	0.736	2.428	3.296	2.000	1.273	4.886	7.573	5.224	3.799
	(38%)	(38%)	(38%)	(38%)	(31%)	(31%)	(31%)	(31%)	(28%)	(28%)	(28%)	(28%)
Tilting 1%	1.110	1.136	0.818	0.687	1.242	1.466	1.036	0.792	1.922	2.936	2.231	1.769
	(5%)	(5%)	(5%)	(5%)	(1%)	(1%)	(1%)	(1%)	(0%)	(0%)	(0%)	(0%)
Tilting 0.1%	1.005	0.961	0.880	0.844	1.026	1.018	0.905	0.852	1.133	1.335	1.137	1.016
	(4%)	(4%)	(4%)	(4%)	(1%)	(1%)	(1%)	(1%)	(0%)	(0%)	(0%)	(0%)

TABLE J.4. GPD : comparison with Trimmed and Tilted MLE methods

	$T = 50$				$T = 100$				$T = 500$			
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$
Càglàd TS	0.959	0.981	0.822	0.649	0.978	0.987	0.899	0.837	0.995	0.997	0.980	0.969
	(3)	(2)	(3)	(2)	(3)	(3)	(3)	(3)	(5)	(100)	(3)	(100)
Trimming 10%	1.081	2.162	1.558	0.767	1.172	2.972	2.364	1.475	1.651	6.471	5.614	3.984
Trimming 1%	1.031	1.092	1.053	0.949	1.029	1.167	1.138	0.929	1.047	1.716	2.002	1.751
Trimming 0.1%	1.000	1.001	0.999	0.998	1.008	1.010	1.016	1.000	1.018	1.048	1.101	1.063
Tilting 10%	2.171	2.497	1.215	1.054	3.088	3.517	1.841	1.037	6.902	7.843	4.497	2.851
	(12%)	(12%)	(12%)	(12%)	(6%)	(6%)	(6%)	(6%)	(2%)	(2%)	(2%)	(2%)
Tilting 1%	1.054	1.204	0.835	0.725	1.333	1.505	0.981	0.762	2.554	2.985	1.875	1.323
	(1%)	(1%)	(1%)	(1%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
Tilting 0.1%	0.907	0.984	0.910	0.910	0.988	1.029	0.917	0.875	1.226	1.352	1.076	0.954
	(3%)	(3%)	(3%)	(3%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)	(0%)

J.3. **Results on confidence interval coverage and length.** In this section, we study the coverage and length properties of confidence based on the normal approximations derived in the main text. We work in the same setting of the Monte Carlo exercise of Section 4 in the main text, where the goal was quantile estimation. We focus on the behaviour of the Càglàd two-step estimator.

J.3.1. *GEV design.* To begin understanding the quality of the normal approximations derived in the main text, we analyse the coverage and length properties of confidence intervals based on normal critical values and the *true* sampling variance of the estimators. Specifically, we study confidence intervals of the form $Q(\tau|\hat{\theta}) \pm \sqrt{\mathbb{V}[Q(\tau|\hat{\theta})}q_Z(1-(1-\beta)/2)$, where β is the nominal coverage level, $q_Z(u)$ is the u-quantile of a standard-normal distribution, and $\mathbb{V}[Q(\tau|\hat{\theta})]$ is the true sampling-variance of the plug-in quantile estimator, which we recover from the Monte Carlo draws. We consider the case $\beta = 0.95$.

tilting proportion ϵ should then be chosen so as to make the QQ-plot "close" to the 45 degrees line. Their heuristic is motivated by data corruption concerns, though, and not RMSE reductions. Indeed, notice that, inasmuch as parametric tail quantile estimators offer MSE improvements over nonparametric empirical quantiles, one would expect differences between these estimators.

Figure J.1 reports, in blue, the coverage of the confidence intervals based on the Càglád two-step estimator, for different values of R. In red, we also report the coverage of MLE-based CIs that use the true sampling variance and normal critical values. The nominal level $\beta = 0.95$ is presented as a black horizontal line. As one can observe, coverage of the L-moment-based CIs is close to the nominal level for every combination of sample size and target quantile, suggesting that the strong approximations derived in the main text offer a good approximation to the designs at had. Moreover, and consistent with our theoretical results that do not impose any restrictions on the growth rate of R in the derivation of the normal approximations, coverage is constant across the values of R. The MLE-based CIs also have coverage very close to the nominal in all sample sizes and target quantiles.

Figure J.2 reports how the length of the CIs based on the true sampling variance changes with different values of R. Length is reported as a proportion of the length of the MLE-based CIs, meaning that values above one indicate that the L-moment-based CIs are larger than those based on the MLE. As expected from our efficiency results, the length-minimising choice of R is always competitive with the MLE, offering substantial improvements in length for smaller sample sizes/more extreme quantiles, and working as well as the MLE in the largest sample size.

Next, we analyse the behaviour of feasible versions of the above CIs that rely on estimators of the asymptotic variance. For the MLE, we rely on a delta-method approximation combined with an estimator of the asymptotic variance of $\hat{\theta}_{\text{MLE}}$ based on the Hessian of the objective function. For the two-step Càglád estimator, we rely on the delta-method-type result in 2 to estimate the variance $\mathbb{V}[Q(\tau|\hat{\theta}_R)]$, as:

$$\widehat{\mathbb{V}[Q(\tau|\hat{\theta}_R)]} = \frac{\nabla_\theta Q(\tau|\hat{\theta}_R)'\hat{V}\nabla_\theta Q(\tau|\hat{\theta}_R)}{T},$$

where $\hat{V}$ is an estimator of the asymptotic variance of the optimally-weighted estimator $\hat{\theta}_R$, which is given by:

$$\hat{V} = \widehat{\mathbb{V}[\hat{\theta}_R]} = \left(\left(\int_0^1 \nabla_\theta Q(u|\hat{\theta}_R)\boldsymbol{P}_R(u)'du\right)\hat{\Omega}_R\left(\int_0^1 \boldsymbol{P}_R(u)\nabla_\theta Q(u|\hat{\theta}_R)'du\right)\right)^{-1},$$

with $\hat{\Omega}_R$ the estimator of the optimal weighting matrix used in the minimization of $\hat{\theta}_R$.

Figure J.3 presents the coverage of the feasible CIs that rely on estimators of the asymptotic variance. We see that, in the largest sample size ($T = 500$), coverage is close to the nominal level for all target quantiles. Coverage is also quite close to the nominal level in sample sizes $T = 50$ and $T = 100$ at the median ($\tau = 0.5$). However, in samples $T = 50$ and $T = 100$, as we move further into the tails, coverage tends to deteriorate for both the MLE-based as well as the L-moment-based CIs. In these settings, the L-moment-based CIs undercover more than the MLE-based CIs, in some cases by a small margin, but with especially large differences at the two tailmost quantiles ($\tau = 0.99$ and $\tau = 0.999$) in the smallest sample size ($T = 50$). Finally, we note that, for a given T and τ, coverage is insensitive to the choice of R.

In order to better understand the drivers of undercoverage at the tails in smaller sample sizes, we report, in Figure J.4, the median-length of the feasible CIs. We normalize these lengths by the length of the *unfeasible MLE-based CI that relies on the true sampling-variance*. By observing the red lines, we note that the feasible MLE-based CIs understate the true sampling variance at the tails in smaller sample sizes, as the red line in these cases can be substantially below one. Similarly, by comparing the blue lines in Figure J.4 with the corresponding blue lines in Figure J.2, we see that the feasible CIs based on the Càglád estimator share a similar pattern, understating the true length of the unfeasible Càglàd-based CIs in the extreme quantiles of smaller sample sizes.

To understand how much the understatement of the correct sampling variance contributes to undercoverage of the feasible confidence intervals in smaller sample sizes and extreme quantiles, we present, in Figure J.5, the coverage of *unfeasible* confidence intervals that, while still relying on estimators for the variance, rescale these by the amount of underestimation verified in the previous discussion.[13] We focus on sample sizes $T < 500$ and quantiles $\tau > 0.5$. We note that, except for the smallest sample size and more extreme quantile, rescaling the variance estimator so that, on average, it correctly assesses sampling uncertainty does not seem to substantially improve the coverage of confidence intervals. This suggests that other elements are at play in driving the undercoverage. One obvious candidate is correlation between the variance estimator and the estimator for the target quantile, which, in smaller samples and more extreme quantiles, may generate a non-normal reference distribution for the test that is inverted to construct the confidence interval.

To remove the effect of this correlation and improve coverage of the Càglád-based CI, we suggest a simple procedure based on our strong approximations. First, we note that the feasible CI is based on "inversion" of:

$$\mathbf{t} = \frac{Q(\tau|\hat{\theta}_R) - \hat{Q}(\tau|\theta_0)}{\sqrt{\widehat{\mathbb{V}[Q(\tau|\hat{\theta}_R)]}}} = \sqrt{T}\frac{Q(\tau|\hat{\theta}_R) - \hat{Q}(\tau|\theta_0)}{\sqrt{\nabla_\theta Q(\tau|\hat{\theta}_R)'\hat{V}\nabla_\theta Q(\tau|\hat{\theta}_R)}}.$$

Now, performing a first-order Taylor-expansion on $\mathbf{t}$ in terms of $\hat{\theta}_R$ separately in the numerator and in the variance estimator that enters the denominator, and leveraging the strong approximation of $\sqrt{T}(\hat{\theta}_R - \theta_0)$ in the main text suggests the following distributional approximation:[14]

$$\widetilde{\mathbf{t}} = \frac{\sqrt{T}\nabla_\theta Q(\tau|\theta_0)Z_T}{\sqrt{\nabla_\theta Q(\tau|\theta_0)'V\nabla_\theta Q(\tau|\theta_0) + 2\nabla_\theta Q(\tau|\theta_0)'V\nabla_{\theta\theta'}Q(\tau|\theta_0)Z_T/T}}$$

where $Z_T \sim N(0, V)$, with $V = \mathbb{V}[\hat{\theta}_R]$. Notice that this approximation captures correlation between the numerator and denominator, as Z_T appears in both terms. By replacing θ_0 and V

[13] Such unfeasible rescaling may be seen as "a best case" scenario for feasible strategies that attempt to "bias-correct" the standard error estimator, e.g. Welch-style corrections (e.g. Welch, 1951; Belloni et al., 2012; Imbens and Kolesár, 2016), or variance estimators based in higher-order expansions such as those presented in Appendix K.1 for the generalised L-moment estimator.

[14] In our expansion, we explicitly do not expand $\hat{V}$ in terms of $\hat{\theta}_R$, because estimation error in $\hat{V}$ does not seem to be at the source of undercoverage in smaller sample sizes, given that the feasible CIs perform well at the median even in the smallest sample size.

with estimators and simulating Z_T, one may estimate the quantiles of $\widetilde{\mathbf{t}}$, which can then be used to construct confidence intervals of the form:

$$\left[Q(\tau|\hat{\theta}_R) - q_{\widetilde{\mathbf{t}}}(1-\widehat{(1-\beta)}/2)\sqrt{\widehat{\mathbb{V}[Q(\tau|\hat{\theta}_R)]}}, Q(\tau|\hat{\theta}_R) - q_{\widetilde{\mathbf{t}}}(\widehat{(1-\beta)}/2)\sqrt{\widehat{\mathbb{V}[Q(\tau|\hat{\theta}_R)]}}\right].$$

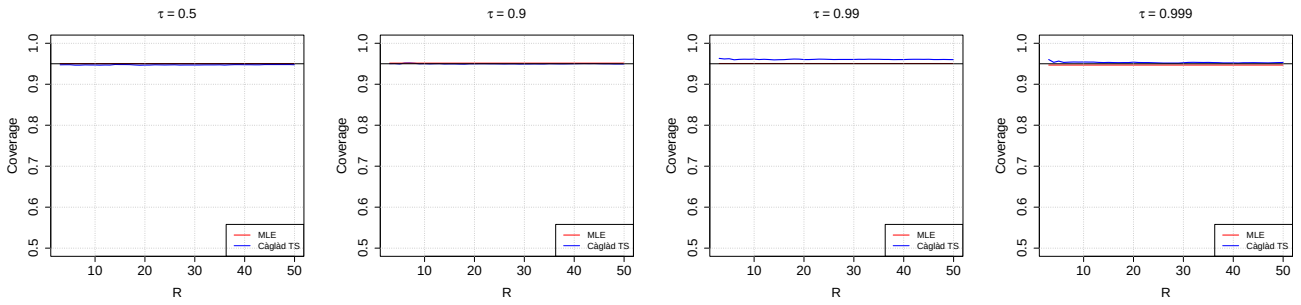
Figure J.6 reports the coverage of our feasible corrected L-moment confidence intervals in blue. We compare these with the coverage of the feasible *uncorrected* MLE confidence interval (in red). We note that our correction substantially improves coverage in those settings wherein the uncorrected CIs would most undercover. Except at $\tau = 0.999$ and $T = 50$, coverage of the resulting CI is quite close to the nominal level. For the case $\tau = 0.999$ and $T = 50$, coverage of the corrected CI becomes quite close to the uncorrected MLE, whereas previously the uncorrected Càglád CI undercovered by a larger margin. The remaining undercoverage in this case can be removed by "bias-correcting" the Càglád-variance estimator – e.g. by estimating the higher-order variance of our estimators by leveraging the higher-order approximations in Appendix K (implemented, for example, via Algorithm K.1)–, as the remaining amount of undercoverage corresponds to the improvement of rescaling over the feasible uncorrected CIs reported in Figures J.5.

To understand how much is the increase in length imparted by our correction, the blue lines in Figure J.7 report the median relative length (vis-à-vis the unfeasible MLE-based CI that relies on the true sampling variance) of our corrected CIs. We also report the relative length of the uncorrected MLE-based CIs (red solid lines) and, for comparison, the relative length of an unfeasibly corrected MLE-based CI that uses the true quantiles of the sampling distribution of the t-statistic inverted in the construction of the confidence intervals (red dotted lines). The latter may be seen as a best-case scenario for the length of feasible corrections to the MLE that attempt to estimate the approximation to the distribution of the t-statistic. The benefits of our approach are clear: in four out of the five settings where our corrected L-moment-based CIs display coverage close to the nominal level, length is always below the length of the unfeasibly corrected MLE. The exception is the case $T = 100$ and $\tau = 0.99$, where length is on average 5 relative percentage points above the unfeasibly corrected MLE. Note, however, that one would expect such differences to vanish, or even revert, once we consider feasible corrections for the MLE, since this would introduce estimation error into the quantiles.

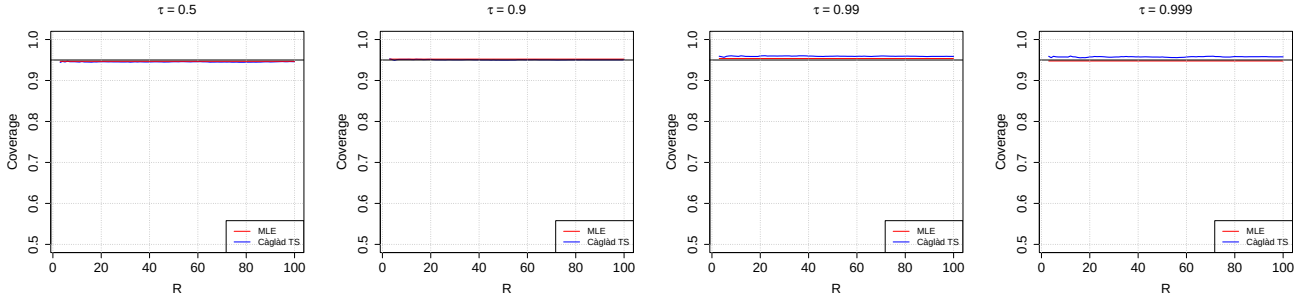
In the case where coverage of our correction is close to the uncorrected MLE ($\tau = 0.999$ and $T = 50$), length of our corrected L-moment-based CIs is below **both** the uncorrected and feasibly uncorrected MLEs. This is also true at $\tau = 0.9$ for the length-minimising choice of R.

J.3.2. *GPD design.* Figures J.8 to J.14 report the preceding analyses in the GPD design. Overall patterns are similar to the GEV design. There are two important differences, though. First, in those settings where the feasible (uncorrected) MLE and L-moment-based CIs undercover, they undercover by a similar margin, even in the smallest sample size and tailmost quantile (see Figure J.10). This is different from the GEV design, where the L-moment CI undercovers by more than the MLE, especially in the smallest sample size and tailmost quantile. Secondly, we note that,

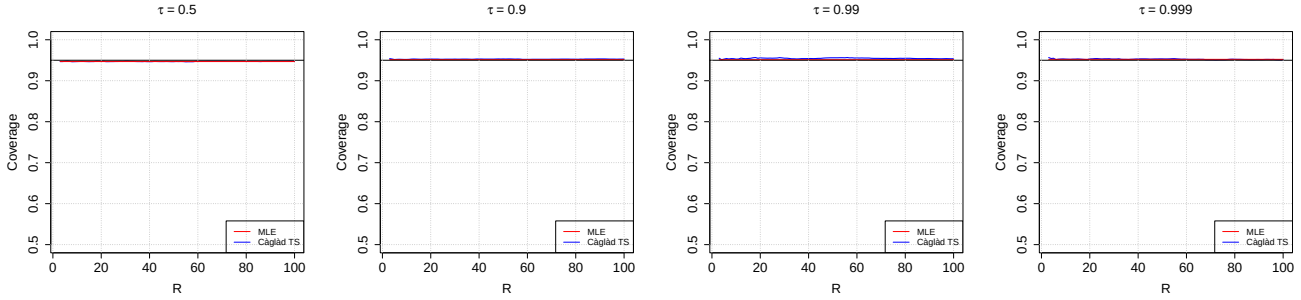
in Figure J.14, our proposed feasible correction to the L-moment CIs never underperforms the unfeasibly corrected MLE, with the blue solid line always below the red dotted line. In contrast, this happens in one of the six cases considered in the GEV design, though, as we discussed in the preceding section, we expect these differences to disappear once we consider a feasible correction to the MLE CI.



(A) $T = 50$

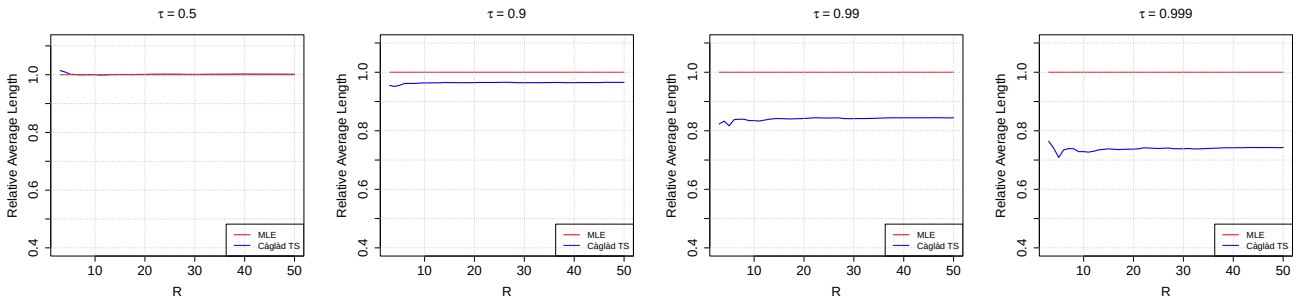


(B) $T = 100$

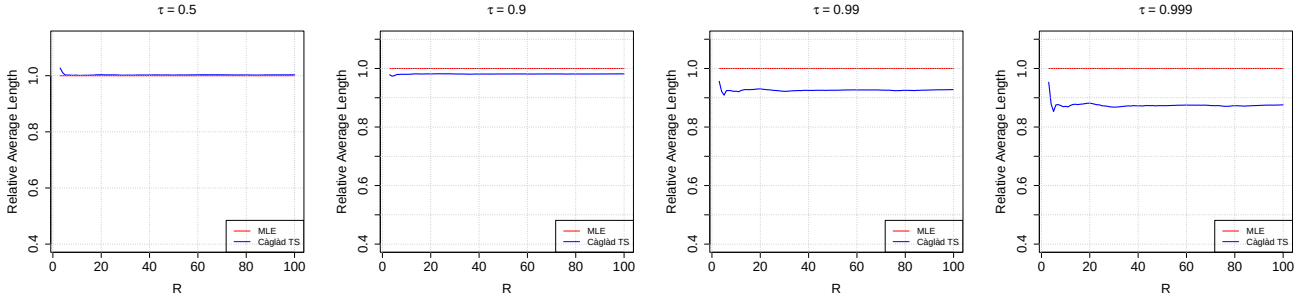


(C) $T = 500$

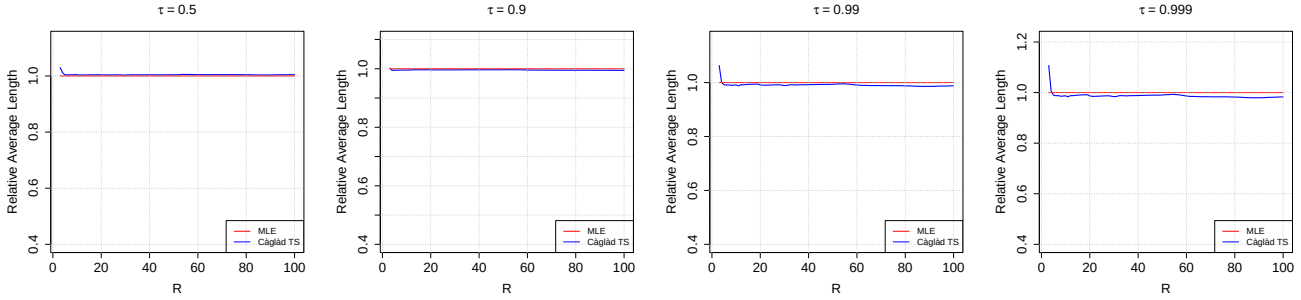
FIGURE J.1. GEV: coverage of confidence intervals based on the true sampling variance.



(A) $T = 50$

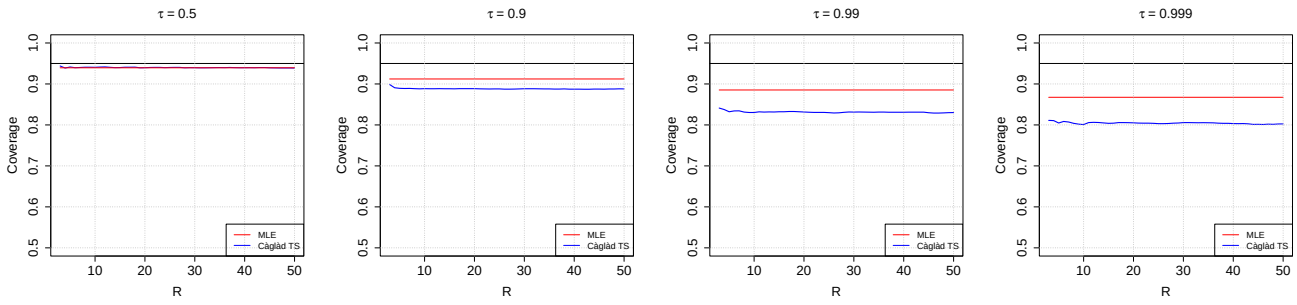


(B) $T = 100$

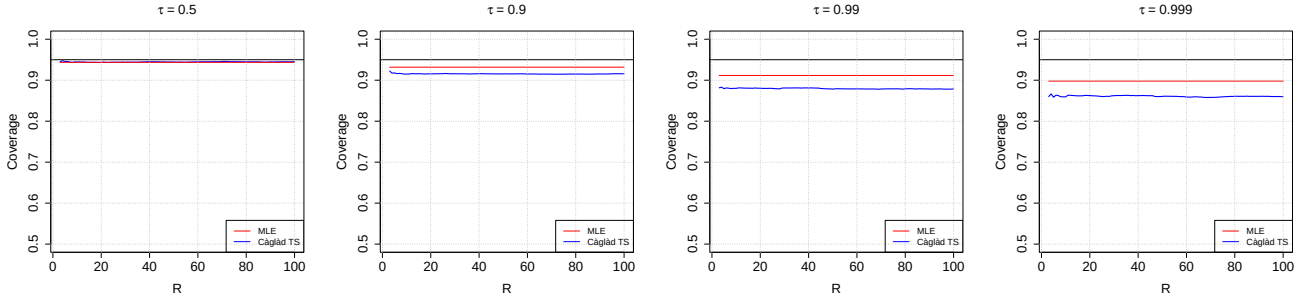


(C) $T = 500$

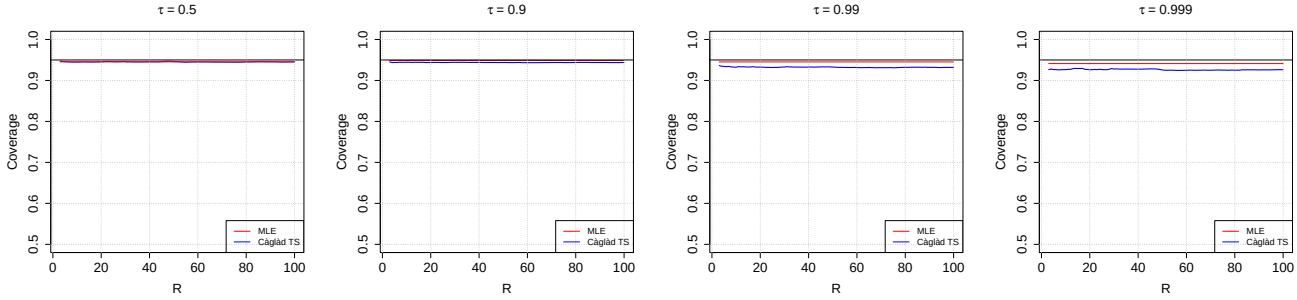
FIGURE J.2. GEV: relative length of confidence (vis-à-vis the unfeasible MLE-based CI) intervals based on the true sampling variance.



(A) $T = 50$

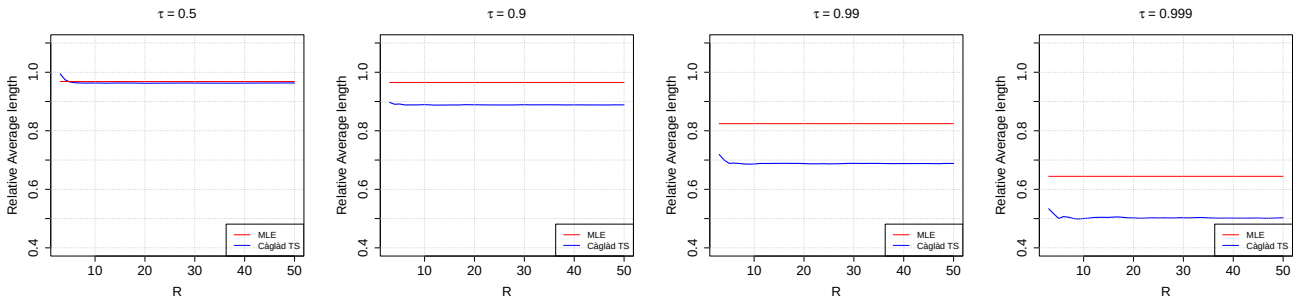


(B) $T = 100$

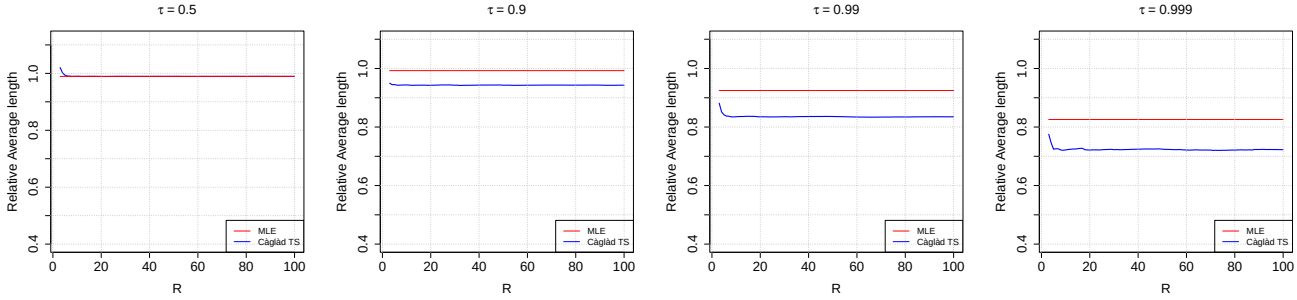


(C) $T = 500$

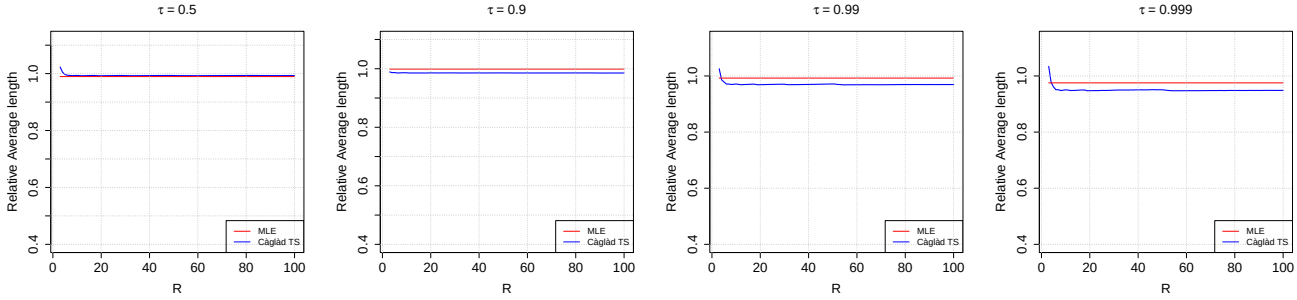
FIGURE J.3. GEV: coverage of confidence intervals that rely on an estimator of the asymptotic variance.



(A) $T = 50$

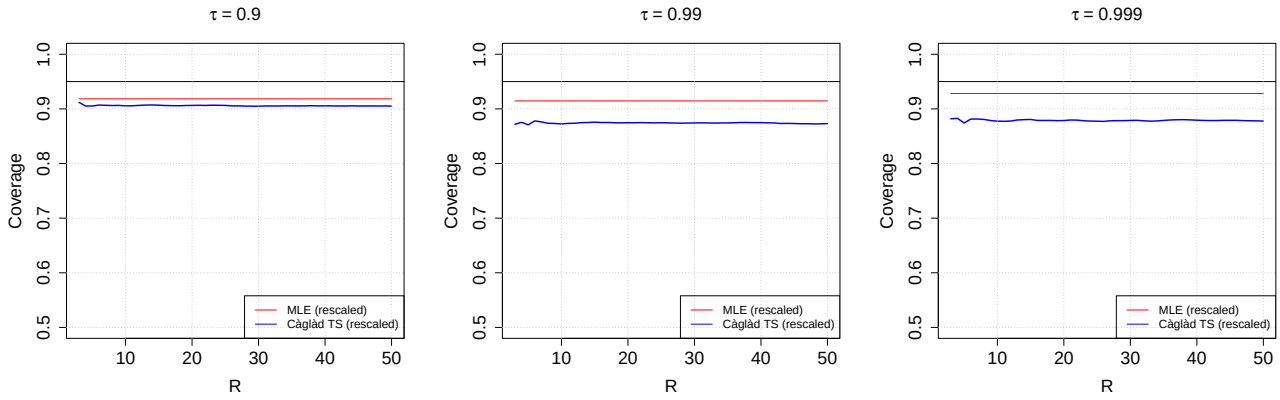


(B) $T = 100$

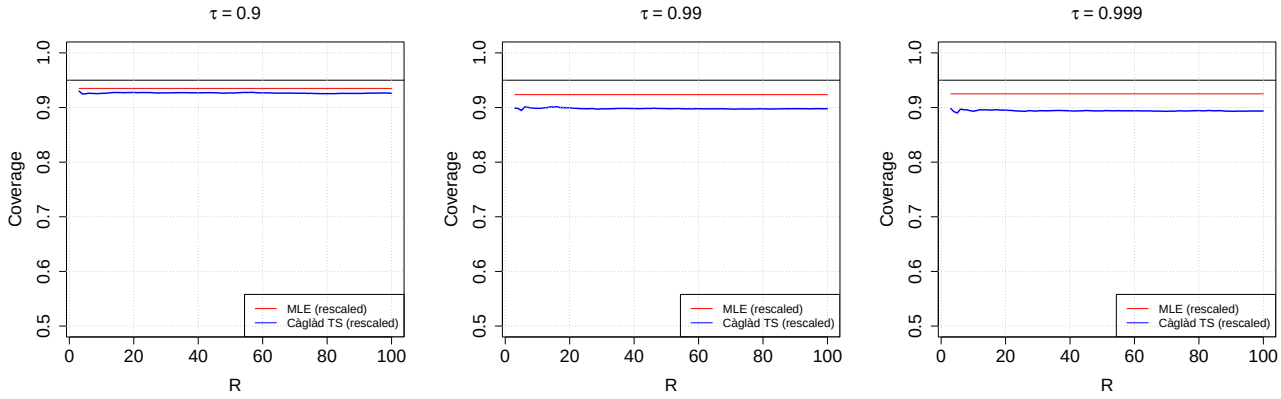


(C) $T = 500$

FIGURE J.4. GEV: relative length (vis-à-vis the unfeasible MLE-based CI) of confidence intervals that rely on an estimator of the asymptotic variance.

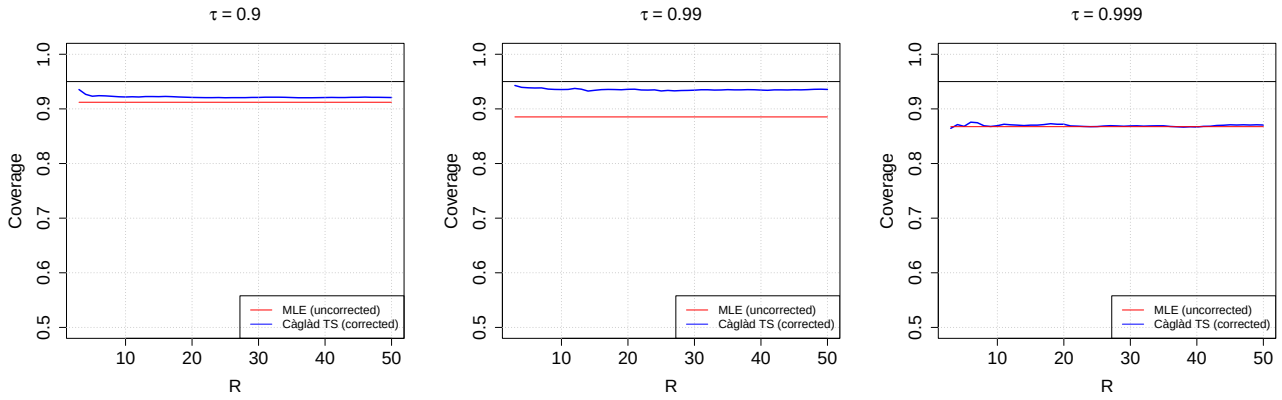


(A) $T = 50$

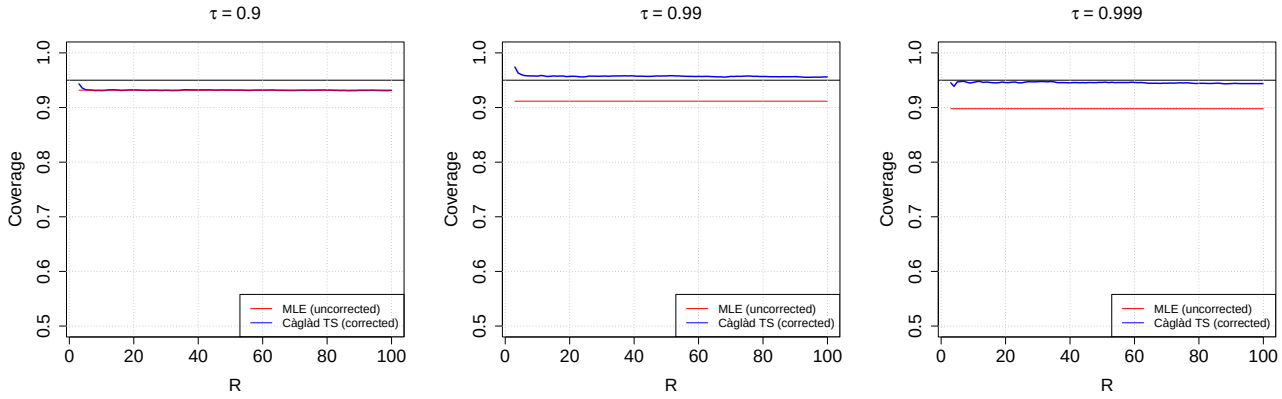


(B) $T = 100$

FIGURE J.5. GEV: coverage of unfeasible confidence intervals that rely on rescaled estimators of the asymptotic variance.

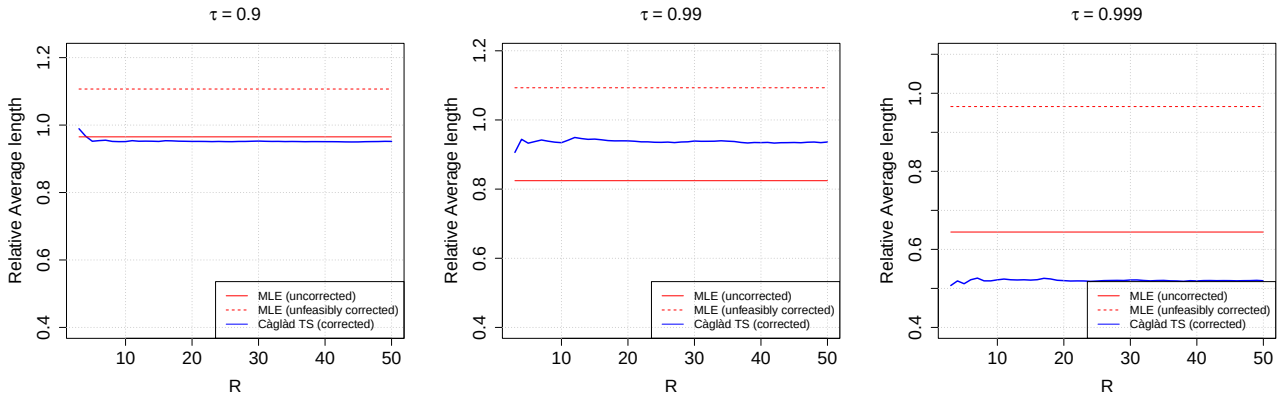


(A) $T = 50$

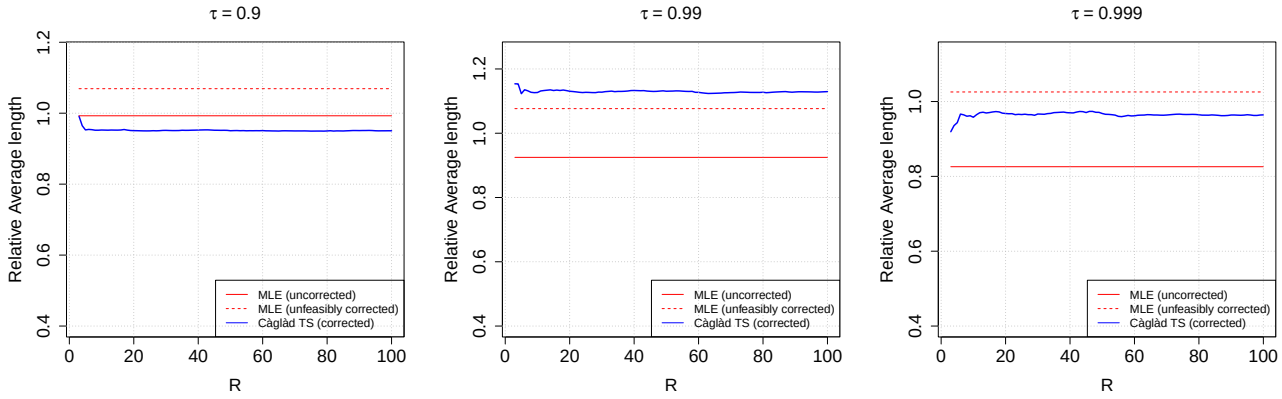


(B) $T = 100$

FIGURE J.6. GEV: coverage of feasible confidence intervals that rely on corrected quantiles.

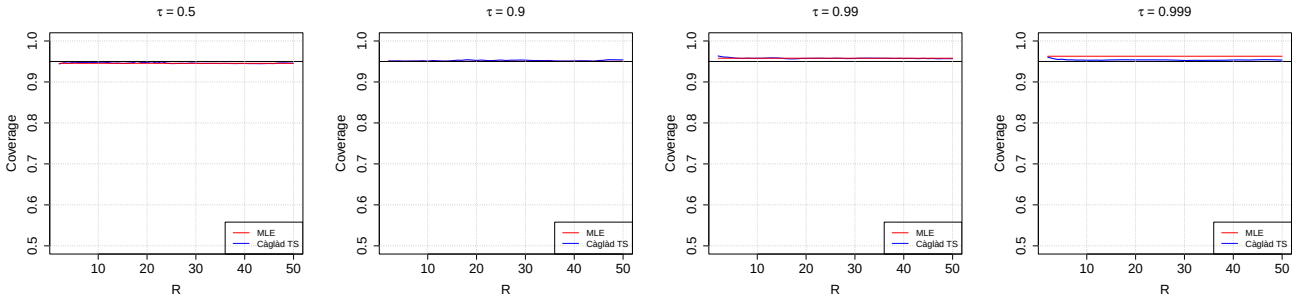


(A) $T = 50$

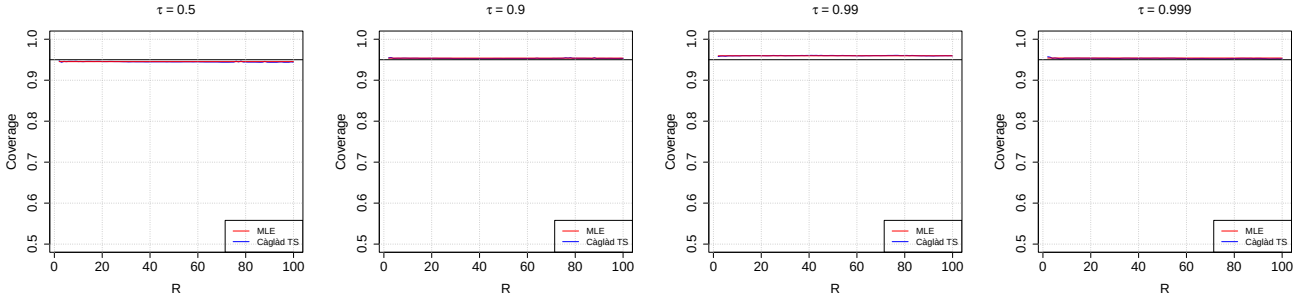


(B) $T = 100$

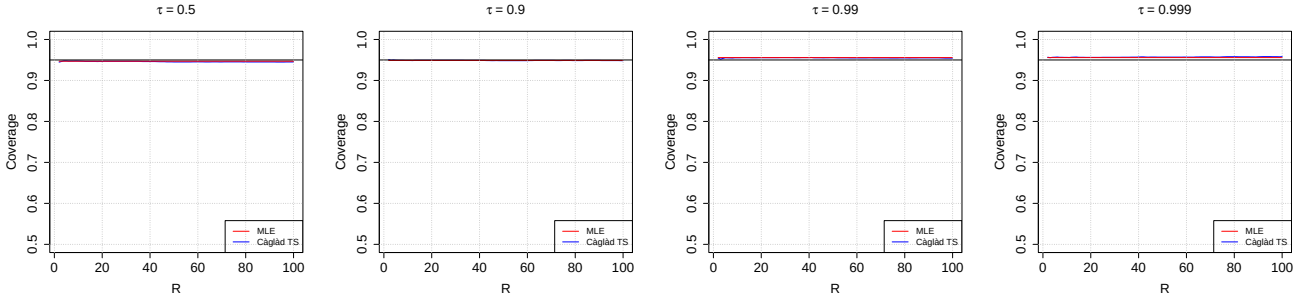
FIGURE J.7. GEV: relative length (vis-à-vis the unfeasible MLE-based CI that relies on the true sampling variance) of feasible confidence intervals that rely on corrected quantiles.



(A) $T = 50$

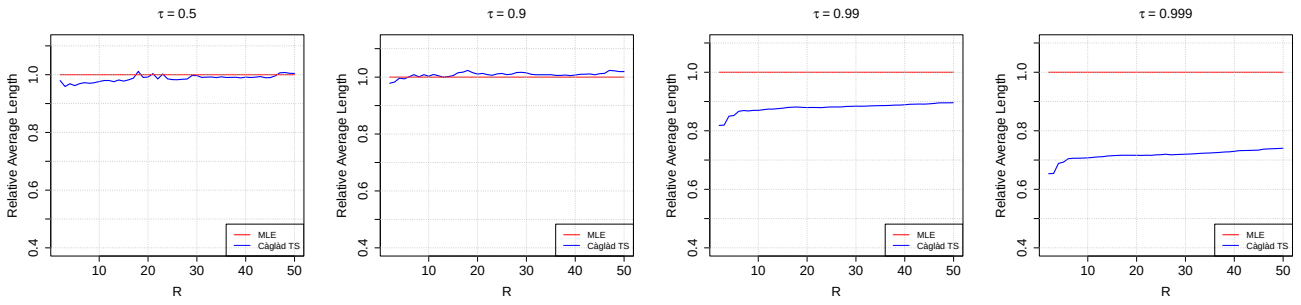


(B) $T = 100$

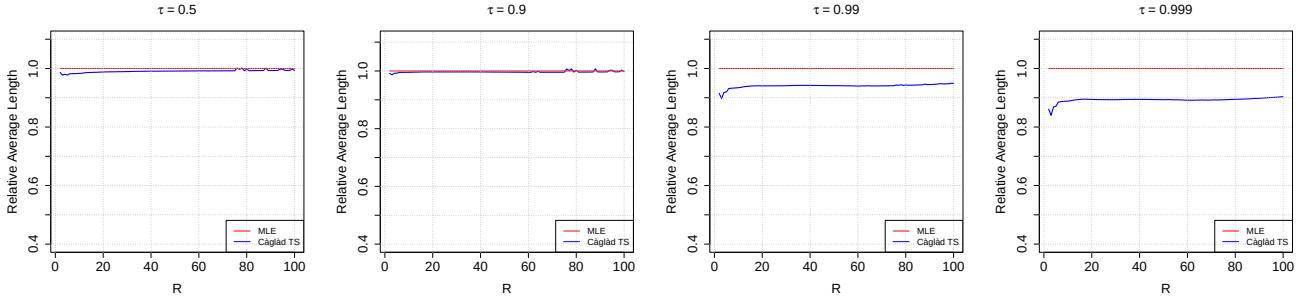


(C) $T = 500$

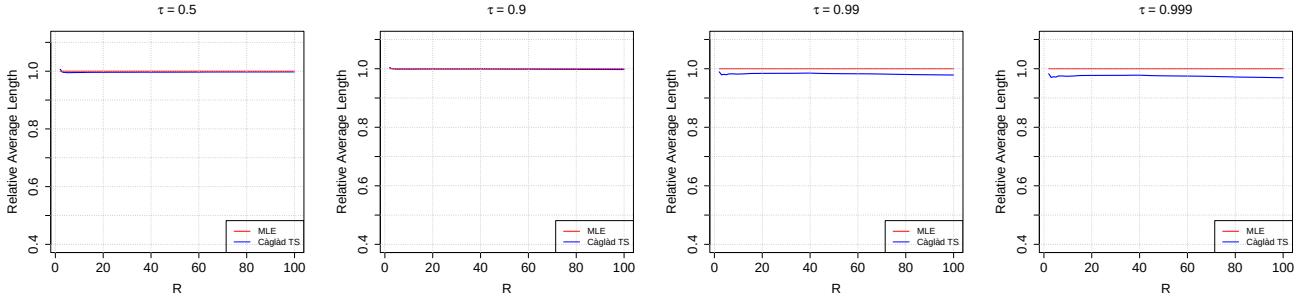
FIGURE J.8. GPD: coverage of confidence intervals based on the true sampling variance.



(A) $T = 50$

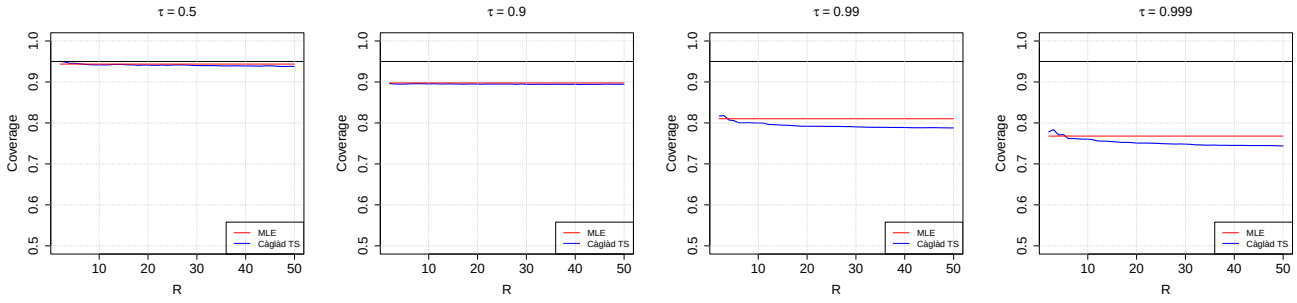


(B) $T = 100$

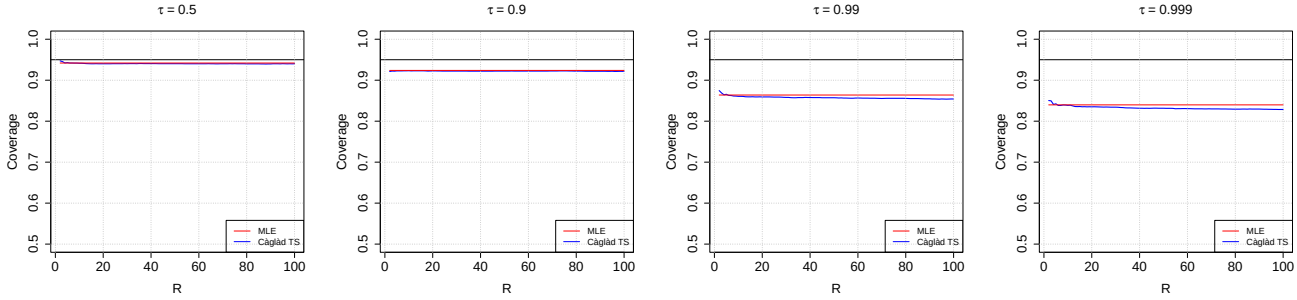


(C) $T = 500$

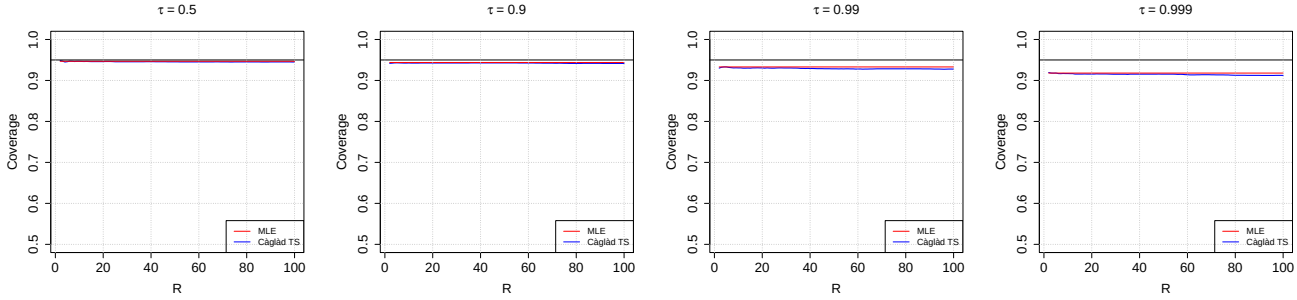
FIGURE J.9. GPD: relative length of confidence (vis-à-vis the unfeasible MLE-based CI) intervals based on the true sampling variance.



(A) $T = 50$

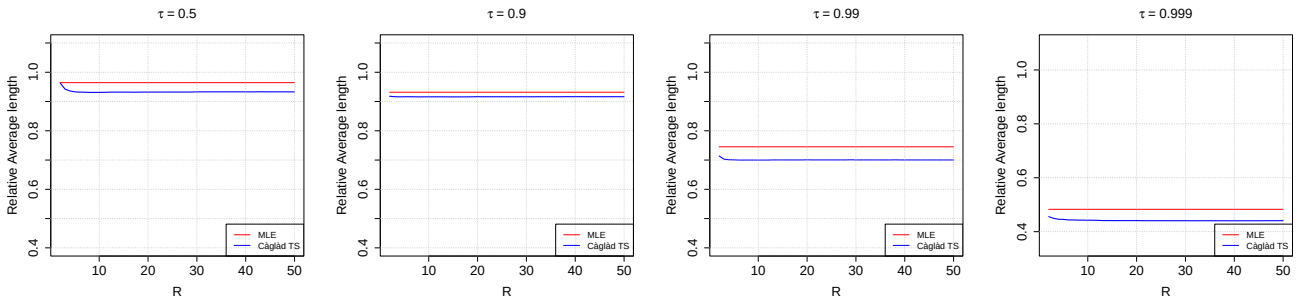


(B) $T = 100$

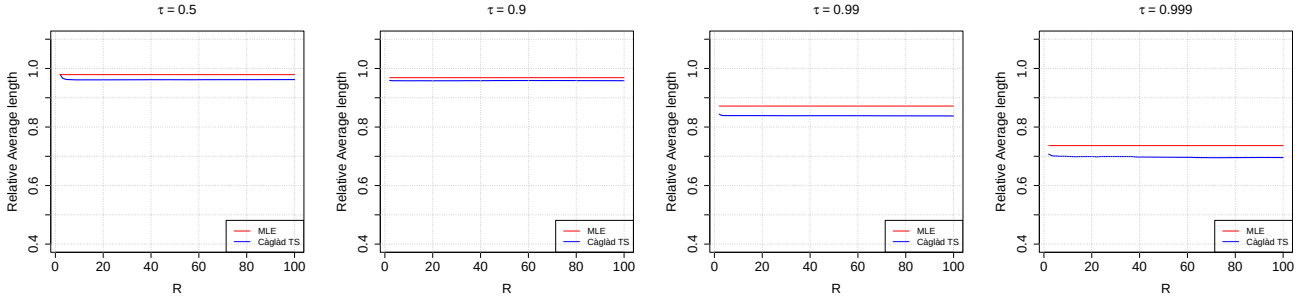


(C) $T = 500$

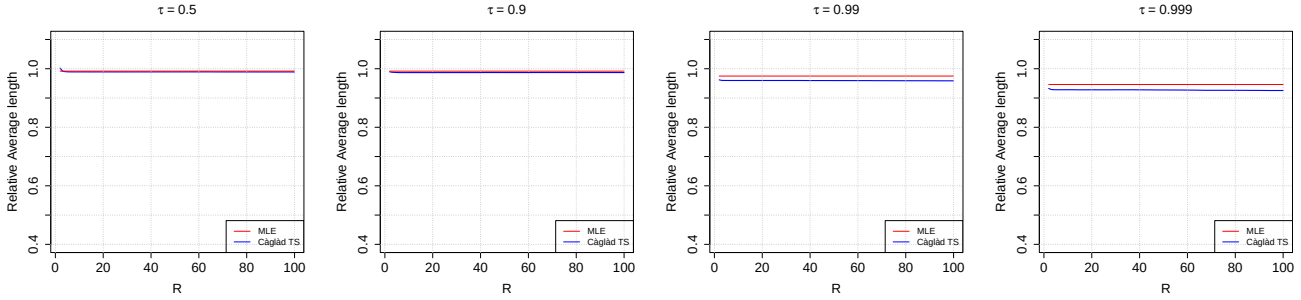
FIGURE J.10. GPD: coverage of confidence intervals that rely on an estimator of the asymptotic variance.



(A) $T = 50$

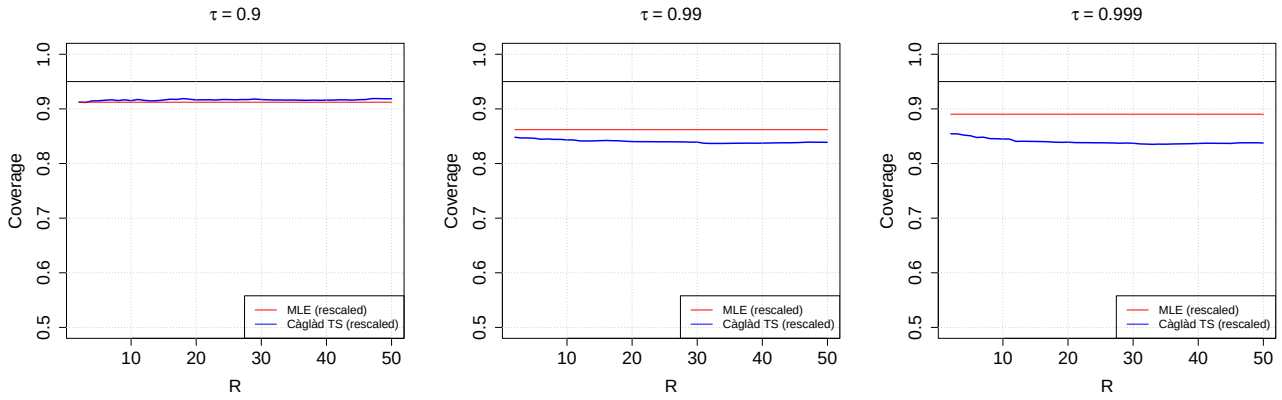


(B) $T = 100$

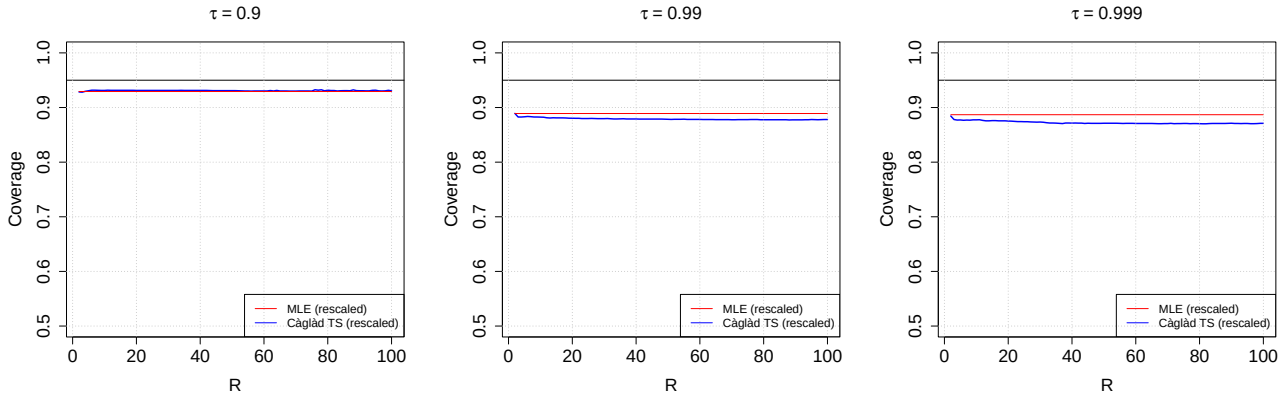


(C) $T = 500$

FIGURE J.11. GPD: relative length (vis-à-vis the unfeasible MLE-based CI) of confidence intervals that rely on an estimator of the asymptotic variance.

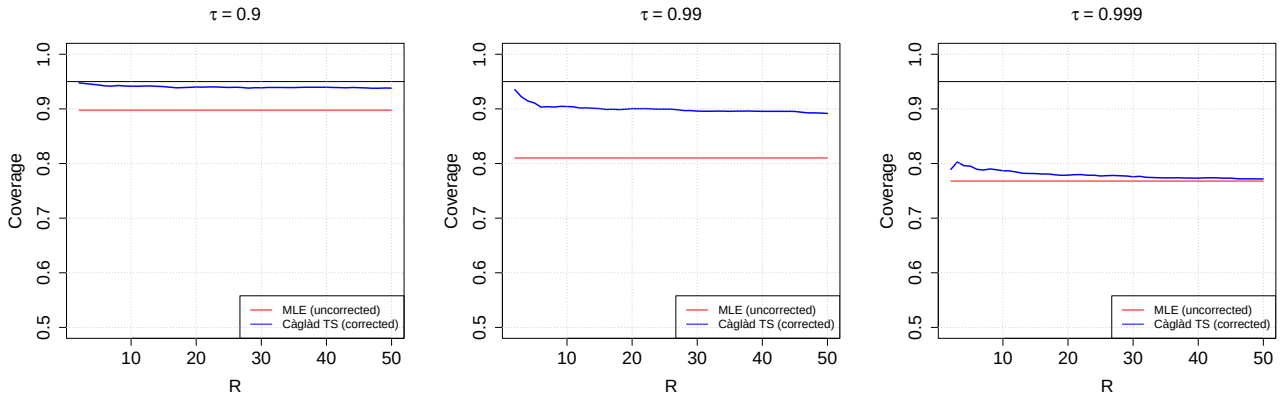


(A) $T = 50$

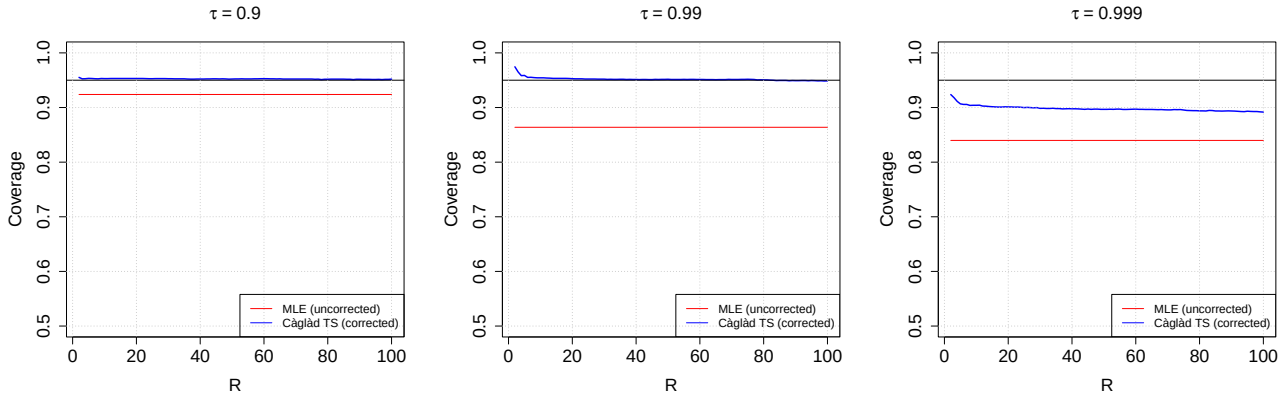


(B) $T = 100$

FIGURE J.12. GPD: coverage of unfeasible confidence intervals that rely on rescaled estimators of the asymptotic variance.

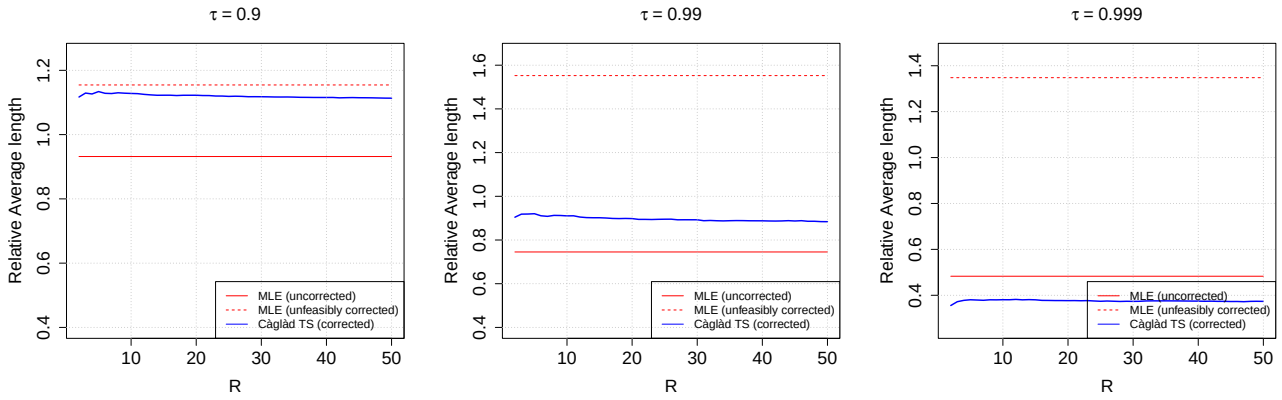


(A) $T = 50$

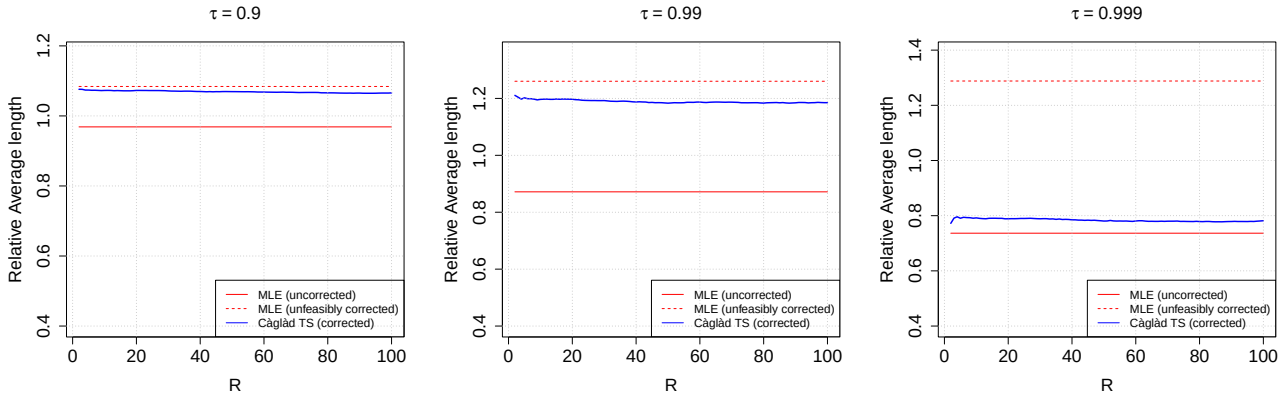


(B) $T = 100$

FIGURE J.13. GPD: coverage of feasible confidence intervals that rely on corrected quantiles.



(A) $T = 50$



(B) $T = 100$

FIGURE J.14. GPD: relative length (vis-à-vis the unfeasible MLE-based CI that relies on the true sampling variance) of feasible confidence intervals that rely on corrected quantiles.

APPENDIX K. DETAILS ON SELECTION METHODS

K.1. **Higher order expansion of the generalised L-moment estimator.** In this section, we derive a higher order expression for the L-moment estimator in Section 3 of the main text. Our goal is to derive a representation of the estimator as follows:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = \Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T} + O_P(T^{-3/2}), \tag{11}$$

for tight sequences of random variables Θ_1^T, Θ_2^T, Θ_3^T. Under uniform integrability conditions on Θ_1^T, Θ_2^T, Θ_3^T and the remainder, representation (11) allows us to write:

$$\mathbb{E}[T(\hat{\theta}_T - \theta_0)(\hat{\theta}_T - \theta_0)'] = \mathbb{E}\left[\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)'\right] + O(T^{-3/2}), \tag{12}$$

which may be used as a basis for a method of selecting R, provided the expectation on the right-hand side is estimable.[15] The idea would be to choose R so as to minimise a linear combination of an estimator of the expectation on the right-hand side. Alternatively, if the goal is to estimate a scalar function of the true parameter, $g(\theta_0)$, one could use the higher-order expansion (11) to construct the higher-order MSE of the estimator $g(\hat{\theta}_T)$. We return to this point in a remark by the end of this section.

To derive representation (11) for the L-moment estimator, we assume, in addition to Assumptions 1-8 in the main text, the following conditions.

Assumption K.1. *As $T, R \to \infty$, $\mathbb{P}[W_R^{-1}$ and Ω_R^{-1} exist$] \to 1$. We also assume that, as $T, R \to \infty$, $W_R^{-1} = \Omega_R^{-1} + O_p(T^{-1/2})$.*

Assumption K.2. *$Q_Y(u|\theta)$ is five times continuously differentiable on $\mathcal{O}$,for each $u \in [\underline{p}, \bar{p}]$. The partial derivatives of $Q_Y(u|\theta)$ with respect to θ, up to the fourth order, are square integrable on $[\underline{p}, \bar{p}]$, for each $\theta \in \mathcal{O}$. For ach $i, j, k, l, m \in \{1, 2, \ldots, p\}$, the partial derivatives satisfy $\sup_{\theta \in \mathcal{O}} \sup_{u \in [\underline{p}, \bar{p}]} |\frac{\partial^5 Q_Y(u|\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k \partial\theta_l \partial\theta_m}| < \infty$.*

In the next proposition, we use Assumptions 1-8, K.1 and K.2 to provide a higher order expansion of the L-moment estimator. Our proof strategy mimics that used in Newey and Smith (2004) to derive a higher order expression for a GMM estimator with a fixed number of moments, but with additional care to take into account that $R \to \infty$ and the L-moment structure in our setting.[16]

Proposition K.1. *Suppose Assumptions 1-8, K.1 and K.2 are satisfied. Then (11) holds for $\hat{\beta} = \begin{pmatrix} \hat{\theta}' & -h^R(\hat{\theta})' W^R \end{pmatrix}'$ with the objects as follows:*

$$\Theta_1^T = -M_0^{-1}\sqrt{T}m(\beta_0)\,,$$

$$\Theta_2^T = M_0^{-1}\sqrt{T}(M - M_0)M_0^{-1}\sqrt{T}m(\beta_0) - \frac{M_0^{-1}}{2}\sum_j \left(M_0^{-1}\sqrt{T}m(\beta_0)\right)_j \partial_j M M_0^{-1}\sqrt{T}m(\beta_0)\,,$$

$$\Theta_3^T = -M_0^{-1}\sqrt{T}(M - M_0)M_0^{-1}\Theta_2^T - \frac{M_0^{-1}}{2}\sum_j (\Theta_T^1)_j \partial_j M \Theta_T^2 - \frac{M_0^{-1}}{2}\sum_j (\Theta_T^2)_j \partial_j M \Theta_T^1 +$$
$$+\frac{1}{6}\sum_{i,j}(M_0^{-1}\sqrt{T}m(\beta_0))_i (M_0^{-1}\sqrt{T}m(\beta_0))_j \partial_{i,j} M (M_0^{-1}\sqrt{T}m(\beta_0))\,.$$

where M_0 and m are defined in the proof of the theorem.

Proof. See Appendix K.4. □

[15] Even if the uniform integrability conditions that allow us to write (12) from (11) do not hold, we can posit that our goal is to minimise the MSE of the leading term in (11). This is the Nagar (1959) style approach of Rothenberg (1984) and Donald and Newey (2001). We return to this point later on.

[16] Donald et al. (2009) consider the higher order expansion of a GMM-type estimator with an increasing number of moment conditions, but their results hold for a special type of moment conditions, which inhibits direct application of their results to our L-moment setting.

The previous proposition yields a higher-order expansion of the L-moment estimator. Nonetheless, this expansion depends on two quantities whose moments may not be immediately computed: (i) the estimation error of the inverse of the weighting matrix, $(W^R)^{-1} - (\Omega^R)^{-1}$; (ii) moments of the (recentered) L-moment vector, $\sqrt{T}h^R(\theta_0)$. We deal with each term separately.

With regards to the estimation error of the inverse, it is possible to derive an $O_{P^*}(T^{-1})$ expansion of $\sqrt{T}((W^R)^{-1} - (\Omega^R)^{-1})$, which can then be plugged onto (11) to obtain an $O_{P^*}(T^{-3/2})$ expansion in terms of quantities whose moments may be estimated. In particular, if $(\Omega^R)^{-1}$ may be written as a function $M^R(\theta_0)$ (e.g. the optimal weights under a Gaussian approximation) and $(W^R)^{-1} = M^R(\tilde{\theta}_T)$ for a preliminary estimator with representation $\sqrt{T}(\tilde{\theta}_T - \theta_0) = \Pi_T^1 + \frac{\Pi_T^2}{\sqrt{T}} + O_p(T^{-1})$ (e.g. the L-moment estimator with identity weights or the MLE estimator), then the result can be obtained under uniform differentiability conditions on $M^R(\cdot)$. We state these below:

Lemma K.1. *Suppose $(\Omega^R)^{-1} = M^R(\theta_0)$ and that we estimate it by $(W^R)^{-1} = M^R(\tilde{\theta}_T)$, where $\tilde{\theta}_T$ is a preliminary estimator with representation $\sqrt{T}(\tilde{\theta}_T - \theta_0) = \Pi_T^1 + \frac{\Pi_T^2}{\sqrt{T}} + O_p(T^{-1})$. Suppose that the entries in M^R are three times continuously differentiable on $\mathcal{O}$. Let $\partial_{s_1,\dots,s_k} M^R(\theta)$ be the $R \times R$ matrix with entry (i,j) corresponding to the partial derivative $\partial_{s_1,\dots,s_k}(M^R(\theta))_{i,j}$. Suppose that $\sum_{i=1}^d \|\partial_i M^R(\theta_0)\|_2^2 = O(1)$, $\sum_{i=1}^d \sum_{j=1}^d \|\partial_{i,j} M^R(\theta_0)\|_2^2 = O(1)$ and $\sup_{\theta \in \mathcal{O}} \sum_{i=1}^d \sum_{j=1}^d \|\partial_{i,j} M^R(\theta)\|_2^2 = O(1)$. Then the estimator satisfies:*

$$\sqrt{T}((W^R)^{-1} - (\Omega^R)^{-1}) = \Xi_T^1 + \frac{\Xi_T^2}{\sqrt{T}} + O_p(T^{-1}),$$

where

$$\Xi_T^1 = \sum_{i=1}^d \partial_i M^R(\theta_0)(\Pi_1^T)_i,$$

$$\Xi_T^2 = \sum_{i=1}^d \partial_i M^R(\theta_0)(\Pi_2^T)_i + \sum_{i=1}^d \sum_{j=1}^d \partial_{ij} M^R(\theta_0) \left[(\Pi_1^T)_i (\Pi_1^T)_j\right].$$

Proof. The proof follows by performing a third order mean value expansion and using the assumptions to show the third derivative term is $O_P(T^{-1})$. □

As for computing moments of $\sqrt{T}h^R(\theta_0)$, one may be tempted to use the strong approximations considered in Section 3 to obtain estimates of these. Nonetheless, we argue this approximation may not be desirable: in particular, it would imply that there is no bias in the estimation of L-moments, whereas it is known that the latter constitutes a large part of the mean squared error of quantile estimators (Franguridi et al., 2022). To better formalise this notion, we follow Donald and Newey (2001), Donald et al. (2009) and Okui (2009) in defining a Nagar (1959) style approximation to the MSE $M_T := \mathbb{E}\left[\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)'\right]$ as the sum $\hat{V}_T + \hat{H}_T$, where $\hat{V}_T$ is the first-order variance of the estimator, $\hat{H}_T$ are higher-order terms, and the approximation errors $\hat{E}_T := M_T - (\hat{V}_T + \hat{H}_T)$ and $F_T := T(\hat{\theta}_T - \theta_0)(\hat{\theta} - \theta_0) - \left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)'$ satisfy

$$\frac{\|\hat{E}_T + F_T\|_2}{\|\hat{H}_T\|_2} = o_p(1)\,. \tag{13}$$

Clearly, the Gaussian approximation to the moments of $\sqrt{T}h^R(\theta_0)$ is not "Nagar", as quantile estimators are generally second-order biased.

In a fully parametric setting and when the data is iid, we may use a parametric bootstrap approach to directly estimate M_T. Indeed, given a preliminary estimator $\tilde{\theta}$ of θ_0, we may draw random samples with T observations from $F_{\tilde{\theta}}$ and use these to simulate $\sqrt{T}h^R(\theta_0)$, which can then be used to approximate Θ_1, Θ_2 and Θ_3. One then averages the simulated quadratic form over simulations to estimate M_T. Differently from a Gaussian approximation, this approach immediately incorporates higher-order biases, as the simulated approximation to $\sqrt{T}h^R(\theta_0)$ will generally have non-zero mean. Importantly, however, such approach is limited to the iid setting (or, more generally, settings where the sampling mechanism is known). It is not immediately extended to semiparametric settings either, as in these cases the distribution of the data is not fully specified.[17] Given these limitations, it is important to consider alternative approaches to approximating moments of $\sqrt{T}h^R(\theta_0)$.

To circumvent the limitations in the previous paragraph, one could follow the approach used in the weighting matrix and try to find an $O_p(T^{-3/2})$ expansion of $\sqrt{T}h^R(\theta_0)$ in terms of estimable terms, which could then be plugged onto (11) to obtain a "feasible" $O_p(T^{-3/2})$ expansion. Higher-order expansions of quantile estimators have long been hindered by the "non-smoothness" of the estimation procedure. Recently, Franguridi et al. (2022) have been able to solve this problem by casting quantile estimation as a particular case of quantile regression (Koenker and Bassett, 1978) and relying on empirical process machinery. However, their approach entails a higher-order expansion only up to order $O_P(T^{-3/4})$. If one wishes to compute additional higher-order terms, one could alternatively use the results in Lee et al. (2017, 2018), who rely on Phillips (1991)'s heuristic – whereby a nonsmooth estimator is assumed to satisfy a first-order condition in terms of a (nonsmooth) subgradient, which then allows a Taylor expansion in terms of the Dirac delta function –, to obtain higher-order expansions of quantile estimators. However, as pointed out by Franguridi et al. (2022), Phillips's heuristic does not account for $O_P(T^{-1/2})$ terms stemming from the nonunicity of quantile estimators, and these terms can have nonnegligible effects on the higher-order bias of the estimator.

Remark K.1 (Higher-order expansions for other scalar quantities)**.** Suppose we want to estimate a scalar function $g_T(\theta_0)$ of the true parameter, where we allow the function to vary with sample size. Suppose each g_t, $t \in \mathbb{N}$, is four times continuously differentiable; and define the rate

[17] See Appendix L.1 and Alvarez and Biderman (2024) for extensions of the L-moment approach to semiparametric settings.

$\sup_{\theta\in\Theta}\sum_{i=1}^{p}\sum_{j=1}^{p}\|\partial_{i,j}\nabla_{\theta\theta'}g_t(\theta)\|_2 = O(\xi_T)$. A fourth order mean-value expansion then yields:

$$\begin{aligned}\sqrt{T}(g_T(\hat{\theta}_T) - g_T(\theta_0)) = \nabla_{\theta'}g_T(\theta_0)\sqrt{T}(\hat{\theta}_T - \theta_0) + \sqrt{T}(\hat{\theta}_T - \theta_0)'\nabla_{\theta\theta'}g_T(\theta_0)(\hat{\theta}_T - \theta_0) + \\ \sum_{i=1}^{p}\sqrt{T}(\hat{\theta}_{i,T} - \theta_{i,0})(\hat{\theta}_T - \theta_0)\partial_i\nabla_{\theta\theta'}g_T(\theta_0)(\hat{\theta}_T - \theta_0) + \\ \sum_{i=1}^{p}\sum_{j=1}^{p}\sqrt{T}(\hat{\theta}_{i,T} - \theta_{i,0})(\hat{\theta}_{j,T} - \theta_{j,0})(\hat{\theta}_T - \theta_0)\partial_{ij}\nabla_{\theta\theta'}g_T(\tilde{\theta}_T)(\hat{\theta}_T - \theta_0)\,,\end{aligned} \tag{14}$$

where $\tilde{\theta}_T$ lies in the line segment between θ_0 and $\hat{\theta}_T$. Under uniform integrability of $\sqrt{T}(\hat{\theta}_T - \theta_0)$, it follows that:

$$\begin{aligned}\mathbb{E}[T(g_T(\hat{\theta}_T) - g_T(\theta_0))^2] = \mathbb{E}\Big[\Big(\nabla_{\theta'}g_T(\theta_0)\sqrt{T}(\hat{\theta}_T - \theta_0) + \sqrt{T}(\hat{\theta}_T - \theta_0)'\nabla_{\theta\theta'}g_T(\theta_0)(\hat{\theta}_T - \theta_0) + \\ \sum_{i=1}^{p}\sqrt{T}(\hat{\theta}_{i,T} - \theta_{i,0})(\hat{\theta}_T - \theta_0)\partial_i\nabla_{\theta\theta'}g_T(\theta_0)(\hat{\theta}_T - \theta_0)\Big)^2\Big] + O\left((\psi_T)\cdot(\xi_T T^{-3/2})\right)\,,\end{aligned}$$

where ψ_T is the order of the sum of the first three terms in the mean-value expansion (14). We can then plug (12) on the above to obtain an expansion in terms of estimable terms. Useful sequences of g_T would be $g_T(\theta) = Q_Y(u_T|\theta)$ (in quantile estimation) or $g_T(\theta) = F_\theta(p_T)$ (in probability estimation).

K.2. **A Lasso-based alternative.** In this section, we briefly review the selection method proposed by Luo et al. (2015) in the GMM context. As discussed in the main text, our generalised L-moment estimator can be seen as combining the R moments used in estimation into d linear restrictions via the mapping:

$$\hat{A}_R h_R(\hat{\theta}) = 0\,, \tag{15}$$

where the combination matrix is estimated as:

$$\hat{A}_R = \nabla_{\theta'}h^R(\hat{\theta})'W^R\,. \tag{16}$$

The poor behaviour of the L-moment estimator with large R may be partly attributed to the estimation of (16). Indeed, we note that the term $\Theta_T^2/\sqrt{T}$ in the higher order expansion of Proposition K.1 is closely related to the estimation error of Ω^R and $\nabla_{\theta'}h^R(\theta_0)$; and correlation between the estimation error of these quantities with $h^R(\theta_0)$ affects the bias due to this term.[18] Suppose $\Xi = (W^R)^{-1}$ exists (as in an estimator of the optimal weighting matrix). Instead of estimating $A_R = \nabla_{\theta'}h^R(\theta_0)'\Omega^R$ by $\hat{A}_R$, Luo et al. (2015) propose to estimate the j-th row of A_l as (adapting their program to our context):

[18]Notice that correlation between $\nabla_{\theta'}h^R(\hat{\theta})$ and $h^R(\theta_0)$ is due solely to estimation error of $\hat{\theta}$, since the Jacobian $\nabla_{\theta'}h^R(\theta_0)$ is nonstochastic in our setting. This is a more general feature of minimum-distance-style estimators, and stands in contrast with GMM estimators where the Jacobian of the empirical moment condition at the truth is random, which possess an additional bias term due to this additional source of correlation (Newey and Smith, 2004).

$$\tilde{\lambda}_j \in \text{argmin}_{\lambda \in \mathbb{R}^R} \frac{1}{2}\lambda' \Xi \lambda - \lambda' \nabla_{\theta'} h^R(\tilde{\theta}) e_j + \frac{k}{T} \sum_{l=1}^{R} \nu_l^j \cdot |\lambda_l| \,, \tag{17}$$

for penalties $k \geq 0, \nu_l^j \geq 0$, $l = 1, \ldots R$; and where $\tilde{\theta}$ is a preliminary estimator and e_j is a $d \times 1$ vector with one in the j-th entry and zero elsewhere. Observe that, when the penalties are set to zero, the solution is $\hat{\lambda}_j = \Xi^{-1} \nabla_{\theta'} h^R(\tilde{\theta}) e_j$, which coincides with the j-th row of $\hat{A}_R$. In general, however, the penalties will induce sparsity on the estimated rows, so only a few entries are selected. Importantly, (17) can be efficiently estimated by quadratic programming algorithms.[19] The problem is also well-defined even if Ξ, but not Ξ^{-1}, exists.

Once the d rows of A_R are estimated, we can stack them onto $\tilde{A}_R = \begin{bmatrix} \hat{\lambda}_1 & \hat{\lambda}_2 & \ldots & \hat{\lambda}_d \end{bmatrix}'$ and estimate θ_0 by solving:

$$\tilde{A}_R h^R(\check{\theta}) = 0 \,. \tag{18}$$

Alternatively, we may adopt a "post-Lasso" procedure, which is known to reduce regularisation bias (Belloni et al., 2012). In our setting, this amounts to running our two-step L-moment estimator using as targets those moments selected by matrix $\tilde{A}_R$, i.e. we use the moments given by the indices $\mathcal{I}_S = \{l \in \{1, 2, \ldots, R\} : \hat{\lambda}_{ij} = 0 \text{ for some } j = 1 \ldots d\}$.

In their paper, Luo et al. (2015) provide theoretical guarantees that, in the GMM context with iid data, if Ξ is the inverse of the optimal weighting matrix and the true combination matrix A_R is *approximately sparse* – in the sense that it is well approximated by a sparse matrix at a rate –, then the estimator based on (18) is asymptotically efficient under some additional conditions and as $T, R \to \infty$. Their result can be adapted to our L-moment context – an extension we pursue in Supplemental Appendix Appendix K.5. In what follows, we contrast the Lasso approach with the higher-order MSE method in a Monte Carlo exercise.

K.3. **Monte Carlo Exercise.** We return to the Monte Carlo exercise in Section 3. We consider the behaviour of four estimators: (i) the L-moment estimator with identity weights and $R = d$ (**FS**); (ii) the two-step generalised L-moment estimator with R selected in order to minimise the approximate RMSE of the quantile one wishes to estimate (**TS RMSE**); (iii) the generalised L-moment estimator with optimal weights and Lasso selection (**TS Lasso**); and (iv) the post-Lasso estimator that runs the two-step generalised estimator using only the moments selected in the Lasso procedure (**TS Post-Lasso**). For conciseness, we only consider estimators based on the "càglàd" L-moment estimator (3). As in Section 3 of the main text, we compare the root-mean-squared error (RMSE) of each approach with that obtained from a MLE plug-in.

A few details with regards to the methods used are in order. First, in method (ii), the approximate RMSE of the target quantile is computed using a parametric bootstrap and the expansion in Remark K.1, **up to second order**. We do so by considering a third order mean-value expansion of the target quantile (i.e. we discard the third order terms in (14) when estimating the RMSE), and by working with the expansion (11) of the L-moment estimator up to order T^{-1}. See

[19] For example, the `quadprog` package in `R` (Turlach and Weingessel, 2011).

Algorithm K.1 for the pseudo-code of our resulting approach and the script `selection.R` in the accompanying Github reposition for a computational implementation in `R`. We discard third-order terms from the expansion for computational reasons – doing so allows us to compute the required derivatives inexpensively using closed-form expressions for Jacobians and Hessians involved in these expressions.[20,21] As can be seen from the formulae in Remark K.1, if $\hat{\theta} - \theta_0$ is (approximately) symmetrically distributed, then dropping third order terms should not affect the bias term that composes the MSE estimate, though it could change the estimated RMSE by changing higher-order variance-related terms. In such settings, one would thus expect the RMSE formula that ignores third order terms to skew selection towards lower bias-inducing choices of R. Given the nonlinear nature of the problem, one expects such choices to lead to smaller values of R than including further higher-order terms would. In future research, it would be interesting to assess whether there are gains in including third-order terms in the estimated higher-order MSE. In our Monte Carlo simulations, we run our selection approach with R ranging from d to $T \wedge 100$.

As for the Lasso-based approaches, we follow the recommendations in Belloni et al. (2012) and Luo et al. (2015) when setting the penalty k and the moment-specific loadings v_l^j. Specifically the penalty k follows the rule in Luo et al. (2015). In setting the penalty-specific loading, we observe that Assumption K.5 in Appendix K.5 requires that, for each program $j = 1 \dots d$ and variable $l = 1, \dots R$, the penalty $v_l^j k/T$ dominate, with high probability, the derivative with respect to the l-th variable in the optimisation, evaluated at the target sparse approximation. Following Luo et al. (2015), we ensure this by setting v_l^j equal to an estimate of an upper bound to the standard error of the gradient at the sparse approximation. We estimate this bound by first using the Delta-method to compute an approximate variance to the entries of the matrices Ξ and $\nabla_{\theta'} h^R(\tilde{\theta})$ that enter the quadratic program in equation (17).[22] We then use these variance estimates in the binomial search Algorithm 1 of Luo et al. (2015) to find an upper bound to the standard error of the gradient at the sparse approximation. Our penalty is "coarse", in the sense that we do not refine the upper bound by using the results of a previous Lasso estimator and then iterating this formula. Given the findings in Belloni et al. (2012) and Luo et al. (2015), one would expect that such refinements would lead to less stringent regularisation, though we leave the design of a proper refinement algorithm for future research. Importantly, for each program j, we modify the loadings v_l^j to be equal to zero for $l = 1, 2, \dots, d$, i.e. we do not regularise the first d L-moments,

[20]For those distributions for which an applied user does not have the required Jacobians and Hessians in closed form, our computational implementation computes the required derivatives using differentiation routines available in the R package `autodiffr`, which serves as a wrapper to Julia routines that compute Jacobians and Hessians efficiently using automatic differentiation.

[21]While it is certainly possible to use symbolic differentiation (e.g. Mathematica routines) to obtain closed-form expressions for third-order derivatives for the GEV and GPD families, the evaluation of such derivatives, as well as the computation of the cross-products involved in the higher-order term Θ_T^3, start to become computationally prohibitive as we consider larger candidate values of R.

[22]Notice that, per the discussion in Section K.2, Ξ is an estimator of the matrix whose (generalised) inverse corresponds to the optimal weighting scheme. In our application, we compute this estimator and the gradient $\nabla_{\theta'} h^R(\tilde{\theta})$ by taking $\tilde{\theta}$ as the FS estimator.

so they are effectively "always included" in the second step estimation. For the Lasso estimators, we consider a maximum number of allowed L-moments (the number R in (17)) as $2(T \wedge 100)$.

Tables K.1 and K.2 replicate the results from Tables 3 and 4 in the main text, but including the TS Lasso procedure. We note that the TS Lasso tends to perform poorly. This is due to the large regularisation bias imparted by the "coarse" penalty, which tends to dominate the RMSE. The Post-Lasso approach is able to attenuate such bias and produce estimators with desirable properties; still, in smaller samples, the relative performance of this approach vis-à-vis TS RMSE can be unfavourable.

Algorithm K.1 Pseudo-code for higher-order mean-squared error calculation

Require: Maximum number of L-moments $\overline{R}$, number of bootstrap simulations B. Target quantile τ.
Ensure: Higher-order MSE estimates of quantile estimators based on the Càglád two-step estimators $\hat{\theta}_R$, $R \in \{1, \dots \bar{R}\}$, and a MSE-minimizing choice R^*.

1: Estimate θ_0 using the Càglád estimator with $R = d$, denoted by $\tilde{\theta}$.
2: Compute and store the derivatives pertaining to expansion of the target quantile: $\nabla_\theta Q(\tau|\tilde{\theta})$, $\nabla_{\theta\theta'} Q(\tau|\tilde{\theta})$,
3: Compute and store the derivatives pertaining to estimation of the Jacobian of the objective function: $\{\int_0^1 \nabla_\theta Q(u|\tilde{\theta}) P_r(u)du, \int_0^1 \nabla_{\theta\theta'} Q(u|\tilde{\theta}) P_r(u)du\}$, $r = 1, \dots, \overline{R}$.
4: Compute and store the evaluation of the following functions and its partial derivatives, with respect to θ at $\tilde{\theta}$: $\int_0^1 \int_0^1 \frac{(U \wedge V - UV)}{f_\theta(Q(U|\theta)) f_\theta(Q(V|\theta))} U^r V^s dU dV$, for $(r, s) \in \{1, \dots, R\}^2$. These quantities pertain to estimation of the optimal weights.
5: Compute and store the vector of L-moments $\hat{L} \leftarrow Q(u|\tilde{\theta}) \boldsymbol{P}_{\overline{R}}(u)du$.
6: **for** b=1 to B **do**
7: Generate a sample of size T from $F_{\tilde{\theta}}$. Denote it by $\mathcal{Z}_b$.
8: Calculate the empirical quantiles of $\mathcal{Z}_b$, $\hat{Q}_b$, and compute and store $\tilde{h}_b \leftarrow \sqrt{T}(\int_0^1 \hat{Q}_b(u) \boldsymbol{P}_{\overline{R}}(u)du - \hat{L})$.
9: **end for**
10: **for** R=1 to $\overline{R}$ **do**
11: **for** b=1 to $\overline{B}$ **do**
12: Compute the quantities $\tilde{\Theta}^1_T$ and $\tilde{\Theta}^2_T$ of the higher-order expansion (11) using the first R entries of $\tilde{h}_b$ and the derivatives and estimated optimal weights of the first R L-moments. Store these quantities as $\tilde{\Theta}^1_{R,b}$ and $\tilde{\Theta}^2_{R,b}$.
13: **end for**
14: Estimate and store the MSE of $Q(\tau|\hat{\theta}_R)$ as:

$$\widehat{\text{MSE}(Q(\tau|\hat{\theta}_R))} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \left(\nabla_{\theta'} Q(\tau|\tilde{\theta})(\tilde{\Theta}^1_{R,b} + T^{-\frac{1}{2}} \tilde{\Theta}^2_{R,b}) + (\tilde{\Theta}^1_{R,b} + T^{-\frac{1}{2}} \tilde{\Theta}^2_{R,b})' \nabla_{\theta\theta'} Q(\tau|\tilde{\theta})(\tilde{\Theta}^1_{R,b} + T^{-\frac{1}{2}} \tilde{\Theta}^2_{R,b}) \right)^2$$

15: **end for**
16: Set $R^* \leftarrow \text{argmin}_{R \in \{1, \dots, \overline{R}\}} \widehat{\text{MSE}(Q(\tau|\hat{\theta}_R))}$.

TABLE K.1. GEV : relative RMSE under different selection procedures

	$T = 50$				$T = 100$				$T = 500$			
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$
FS	1.026	0.962	0.821	0.737	1.031	0.983	0.950	0.928	1.028	1.004	1.061	1.095
	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
TS RMSE	1.008	0.964	0.794	0.674	1.004	0.987	0.923	0.865	1.005	0.999	1.006	1.009
	(16.81)	(3.66)	(3.3)	(3.41)	(33.02)	(4.17)	(4.32)	(4.57)	(20.29)	(35.51)	(40.91)	(43.17)
TS Lasso	4.983	1.897	1.803	> 10	8.188	3.059	2.924	> 10	> 10	4.352	2.481	3.630
	(7.99)	(7.99)	(7.99)	(7.99)	(9.24)	(9.24)	(9.24)	(9.24)	(9.95)	(9.95)	(9.95)	(9.95)
TS Post-Lasso	1.017	0.975	0.857	0.781	1.010	0.988	0.928	0.866	1.006	0.999	0.999	0.993
	(7.99)	(7.99)	(7.99)	(7.99)	(9.24)	(9.24)	(9.24)	(9.24)	(9.95)	(9.95)	(9.95)	(9.95)

TABLE K.2. GPD : relative RMSE under different selection procedures

	$T = 50$				$T = 100$				$T = 500$			
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$
FS	0.984	0.981	0.824	0.648	0.988	0.993	0.917	0.856	1.007	1.005	0.990	0.982
	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)
TS RMSE	0.964	0.994	0.817	0.640	0.980	0.990	0.896	0.828	0.997	0.999	0.978	0.970
	(2.86)	(4.43)	(2.61)	(2.96)	(3.59)	(4.31)	(3.02)	(3.09)	(5.34)	(48.42)	(31.11)	(29.71)
TS Lasso	> 10	> 10	> 10	> 10	> 10	> 10	> 10	> 10	> 10	> 10	> 10	> 10
	(3.52)	(3.52)	(3.52)	(3.52)	(3.7)	(3.7)	(3.7)	(3.7)	(3.78)	(3.78)	(3.78)	(3.78)
TS Post-Lasso	0.995	0.999	0.891	0.741	0.992	0.999	0.950	0.905	0.998	1.000	0.985	0.977
	(3.52)	(3.52)	(3.52)	(3.52)	(3.7)	(3.7)	(3.7)	(3.7)	(3.78)	(3.78)	(3.78)	(3.78)

K.4. **Proof of Proposition K.1.**

Proof. On $\hat{\theta}_T \in \mathcal{O}$ and existence of $(W^R)^{-1}$ and $(\Omega^R)^{-1}$, the estimator satisfies the following first order condition:

$$\nabla_{\theta'} h^R(\hat{\theta})' W^R h^R(\hat{\theta}) = 0\,,$$

which may be written as (Newey and Smith, 2004):

$$\begin{pmatrix} -\nabla_{\theta'} h^R(\hat{\theta})'\hat{\lambda} \\ -h^R(\hat{\theta}) - (W^R)^{-1}\hat{\lambda} \end{pmatrix} = 0\,,$$

where, by Proposition 2, $\hat{\theta} - \theta_0 = O_p(T^{-1/2})$ and:

$$\|\hat{\lambda}\|_2 \leq \|W^R\|_2 \|(\hat{Q}_Y(\cdot) - Q_Y(\cdot|\hat{\theta}))\mathbb{1}_{[\underline{p},\bar{p}]}\|_{L^2[0,1]}\,,$$

implying that $\|\hat{\lambda}\|_2 = O_p(T^{-1/2})$.

Put $\beta \coloneqq (\theta', \lambda')'$. Let:

$$m(\beta) \coloneqq \begin{pmatrix} -\nabla_{\theta'} h^R(\theta)'\lambda \\ -h^R(\theta) - (W^R)^{-1}\lambda \end{pmatrix}.$$

The estimator solves $m(\hat{\beta}) = 0$. Let $\lambda_0 \coloneqq 0_{R\times 1}$ and $\beta_0 \coloneqq (\theta_0', \lambda_0')'$. On $\hat{\theta}_T \in \mathcal{O}$ and existence of $(W^R)^{-1}$ and $(\Omega^R)^{-1}$, a fourth order mean-value expansion of $\hat{\beta}$ around β_0 yields:

$$0 = m(\hat{\beta}) = m(\beta_0) + M(\hat{\beta} - \beta_0) + \frac{1}{2}\sum_j (\hat{\beta}_j - \beta_{j0})\partial_j M(\hat{\beta} - \beta_0) +$$

$$+\frac{1}{6}\sum_{i,j}(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\partial_{i,j}M(\hat{\beta} - \beta_0) + \frac{1}{24}\sum_{g,i,j}(\hat{\beta}_g - \beta_{g0})(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\widetilde{\partial_{g,i,j}M}(\hat{\beta} - \beta_0)\,,$$

where $M = \nabla_{\beta'} m(\beta_0)$ and $\partial_j M$ is the $(d+R)\times(d+R)$ matrix with entry (l,k) equal to $\frac{\partial m^l(\beta_0)}{\partial \beta_j \partial \beta_k}$. Similarly, $\partial_{i,j}M$ is a $(d+R)\times(d+R)$ matrix with entry (l,k) equal to $\frac{\partial m^l(\theta_0)}{\partial \beta_i \beta_j \partial \beta_k}$; and $\widetilde{\partial_{g,i,j}M}$ is a $(d+R)\times(d+R)$ matrix with the fourth order partial derivatives evaluated at u-specific $\tilde{\theta}(u)$ in the line segment between hat θ_0 and $\hat{\theta}$.

Next, we observe that:

$$M = \begin{pmatrix} 0 & -\nabla_{\theta'} h^R(\theta_0)' \\ -\nabla_{\theta'} h^R(\theta_0) & -(W^R)^{-1} \end{pmatrix}.$$

Letting:

$$M_0 \coloneqq \begin{pmatrix} 0 & -\nabla_{\theta'} h^R(\theta_0)' \\ -\nabla_{\theta'} h^R(\theta_0) & -(\Omega^R)^{-1} \end{pmatrix}.$$

Then by Assumption K.1, $\|M - M_0\|_2 = O_p(T^{-1/2})$. Moreover, note that:

$$M_0^{-1} = \begin{pmatrix} \Omega_R^* & -\Omega_R^* \nabla_{\theta'} h^R(\theta_0)' \Omega_R \\ -\Omega_R \nabla_{\theta'} h^R(\theta_0) \Omega_R^* & -\Omega_R + \Omega_R \nabla_{\theta'} h^R(\theta_0) \Omega_R^* \nabla_{\theta'} h^R(\theta_0)' \Omega_R \end{pmatrix},$$

where $\Omega_R^* = (\nabla_{\theta'} h^R(\theta_0)' \Omega_R \nabla_{\theta'} h^R(\theta_0))^{-1}$, which exists by Assumption 8. Rearranging, we have:

$$\begin{aligned}(\hat{\beta} - \beta_0) = -M_0^{-1} m(\beta_0) - M_0^{-1}(M - M_0)(\hat{\beta} - \beta_0) - \frac{M_0^{-1}}{2}\sum_j (\hat{\beta}_j - \beta_{j0})\partial_j M(\hat{\beta} - \beta_0) - \\ - \frac{M_0^{-1}}{6}\sum_{i,j}(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\partial_{i,j}M(\hat{\beta} - \beta_0) - \\ - \frac{M_0^{-1}}{24}\sum_{g,i,j}(\hat{\beta}_g - \beta_{g0})(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\partial_{g,i,j}M(\hat{\beta} - \beta_0) - \\ - \frac{M_0^{-1}}{24}\sum_{g,i,j}(\hat{\beta}_g - \beta_{g0})(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\left[\widetilde{\partial_{g,i,j}M} - \partial_{g,i,j}M\right](\hat{\beta} - \beta_0)\,.\end{aligned} \tag{19}$$

Our first goal is to show that, on $\hat{\theta}_T \in \mathcal{O}$ and existence of $(W^R)^{-1}$ and $(\Omega^R)^{-1}$, $(\hat{\beta} - \beta_0) = -M_0^{-1} m(\beta_0) + O_p(T^{-1})$. We split the proof in several steps. First, note that $\|M_0^{-1}\|_2 = O(1)$. Indeed, for a block-matrix, it follows by the properties of the operator norm that:

$$\left\| \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\|_2 \leq \|A\|_2 + \|B\|_2 + \|C\|_2 + \|D\|_2\,.$$

Then, since $\|\Omega_R^*\|_2 = O(1)$ (Assumption 8), $\Omega_R = O(1)$ (Assumption 3), $\|\nabla_{\theta'} h^R(\theta_0)\|_2^2 \leq \text{tr}(\nabla_{\theta'} h^R(\theta_0)' \nabla_{\theta'} h^R(\theta_0)) \leq \sum_{s=1}^{p} \int_{\underline{p}}^{\bar{p}} |\partial_{\theta_s} Q_Y(u|\theta_0)|^2 du < \infty$ (Assumption 5), it follows that $\|M_0^{-1}\|_2 = O(1)$.

Next, we claim that, except for the first term, all terms on the right-hand side of (19) are $O_p(T^{-1})$. Clearly, $\|(M - M_0)(\hat{\beta} - \beta_0)\| = O_p(T^{-1})$ As for the third term, one needs to characterize $\partial_j M$. For $j \leq d$, we get:

$$\partial_j M = \begin{pmatrix} 0 & -\partial_j \nabla_{\theta'} h^R(\theta_0)' \\ -\partial_j \nabla_{\theta'} h^R(\theta_0) & 0 \end{pmatrix},$$

whereas, for $j \geq d+1$:

$$\partial_j M = \begin{pmatrix} -\nabla_{\theta\theta'} h_{j-d}(\theta_0) & 0 \\ 0 & 0 \end{pmatrix}.$$

This implies that:

$$\|\sum_j (\hat{\beta}_j - \beta_{j0}) \partial_j M (\hat{\beta} - \beta_0)\|_2 \leq$$

$$\leq \|\hat{\beta} - \beta_0\|_2 \cdot \sum_{j=1}^{d+R} |\hat{\beta}_j - \beta_{j0}| \cdot \|\partial_j M_j\|_2 \leq \|\hat{\beta} - \beta_0\|_2^2 \sqrt{\sum_{j=1}^{d+R} \|\partial_j M_j\|_2^2} = O_p(T^{-1}),$$

where we used that $\sum_{j=1}^{p+R} \|\partial_j M_j\|_2^2 = O(1)$, which follows from Bessel's inequality and Assumption Assumption K.2.

Next, we characterize $\partial_{ij} M$. For $1 \leq i, j \leq d$:

$$\partial_{ij} M = \begin{pmatrix} 0 & -\partial_{ij} \nabla_{\theta'} h^R(\theta_0)' \\ -\partial_{ij} \nabla_{\theta'} h^R(\theta_0) & 0 \end{pmatrix},$$

whilst, for $i \leq d$ and $j \geq d+1$:

$$\partial_{ij} M = \begin{pmatrix} -\partial_i \nabla_{\theta\theta'} h_{j-d}(\theta_0) & 0 \\ 0 & 0 \end{pmatrix}$$

and, finally, for $i, j \geq d+1$:

$$\partial_{ij} M = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

which, when using Bessel's inequality and Assumption K.1, implies that:

$$\|\sum_{i,j} (\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0}) \partial_{i,j} M (\hat{\beta} - \beta_0)\|_2 \leq \|\hat{\beta} - \beta_0\|_2^3 \cdot \sqrt{\sum_{i,j} \|\partial_{i,j} M\|_2^2} = O_p(T^{-3/2}).$$

By a similar argument, we can show that the fifth term, which involves fourth order derivatives, is $O_p(T^{-2})$. Finally, we can use the last part of Assumption K.2 in a similar way as in the proof of Proposition 2 to show that the last term is $O_p(T^{-2})$.

Next, using the $O_p(T^{-1})$ representation of $\hat{\beta} - \beta_0$, we get that, on $\hat{\theta}_T \in \mathcal{O}$ and existence of $(W^R)^{-1}$ and $(\Omega^R)^{-1}$:

$$\hat{\beta}-\beta_0 = -M_0^{-1}m(\beta_0)+M_0^{-1}(M-M_0)M_0^{-1}m(\beta_0)-\frac{M_0^{-1}}{2}\sum_j \left(M_0^{-1}m(\beta_0)\right)_j \partial_j M M_0^{-1}m(\beta_0)+O_p(T^{-3/2})\,.$$

Plugging this expression back onto (19) and disconsidering terms that are $O_p(T^{-2})$ allows us to define Θ_1^T, Θ_2^T and Θ_3^T as in (11). To conclude, we must show that focusing on the event that the inverse exists and $\hat{\theta}_T \in \mathcal{O}$ does not change the rates we have obtained. In particular, note that we have already shown that:

$$\sqrt{T}(\hat{\theta} - \theta_0) = \mathbb{1}_{I_T}\left[\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T} + O_{P^*}(T^{-3/2})\right] + \mathbb{1}_{I_T^\complement}\sqrt{T}(\hat{\theta}_T - \theta_0)\,,$$

where I_T is the event that $\hat{\theta}_T \in \mathcal{O}$ and that $(W^R)^{-1}$ and $(\Omega^R)^{-1}$ exist. By Assumption K.1 and $\theta_0 \in \mathcal{O}$, we know that $\mathbb{1}_{I_T} \xrightarrow{p} 1$ and $\mathbb{1}_{I_T^\complement}\sqrt{T}(\hat{\theta}_T - \theta_0) = o_p(1)$.[23] To show the rates derived from (19) are not affected, we show that $\mathbb{1}_{I_T^\complement} = o_p(T^{-3/2})$. Indeed, fix $\epsilon > 0$ and note that there exists $T^* \in \mathbb{N}$ such that, for $T \geq T^*$:

$$T^{3/2}\mathbb{1}_{I_T^\complement} > \epsilon \iff \mathbb{1}_{I_T^\complement} = 1\,,$$

but $\mathbb{P}[I_T^\complement] \to 0$, which proves the desired result. Using that $\mathbb{1}_{I_T^\complement} = o_p(T^{-3/2})$, we can write

$$\sqrt{T}(\hat{\theta} - \theta_0) = \Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T} + O_{P^*}(T^{-3/2}) +$$
$$\mathbb{1}_{I_T^\complement}\sqrt{T}(\hat{\theta}_T - \theta_0) - \mathbb{1}_{I_T^\complement}\left[\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T} + O_{P^*}(T^{-3/2})\right] = \Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T} + O_{P^*}(T^{-3/2})\,,$$

which proves the result. □

K.5. **Conditions for validity of Lasso approach.** This appendix presents sufficient conditions for the validity of the Lasso approach outlined in the main text and detailed in Appendix K.2. We adapt the conditions in Luo et al. (2015) to our setting. In addition to the assumptions in the main text, we require sparse eigenvalue conditions that enable invertibility (and bounded spectral norm) of "small" submatrices of Ξ; and approximate sparsity of the combination matrix A_R. We state these assumptions below. In what follows, define, for $v \in \mathbb{R}^n$, $\|v\|_0 \coloneqq \#\{j : v_j \neq 0\}$.

Assumption K.3 (Approximate sparsity of combination matrix)**.** *For $j = 1, \ldots, d$, let $\lambda_j^* \coloneqq A_R' e_j$. We assume that, for each j, there exist constants K_j^l, K_j^u and a sequence of vectors $\bar{\lambda}_d \in \mathbb{R}^R$ such that, as $T, R \to \infty$:*

(1) $\|\bar{\lambda}_j\|_0 = s_T$

[23] In a similar vein, we have implicitly used that $\mathbb{1}_{\hat{\theta}_T \in \mathcal{O}} \xrightarrow{p} 1$ and $\mathbb{1}_{\hat{\theta}_T \notin \mathcal{O}}\sqrt{T}(\hat{\theta}_T - \theta_0) = o_p(1)$ in the proof of the linear representation of Proposition 2.

(2) $\|\bar{A}_R - A_R\|_2 = o(1)$, *where* $\bar{A}_R = \begin{bmatrix} \bar{\lambda}_1 & \bar{\lambda}_2^* & \dots & \bar{\lambda}_d^* \end{bmatrix}'$.

Assumption K.4 (Sparse eigenvalue and spectral norm condition)**.** *Let:*

$$\kappa(s, \Xi) = \min_{\delta \in \mathbb{R}^R : \|\delta\|_0 \leq s, \|\delta\|_2 = 1} \delta' \Xi \delta \,,$$
$$\phi(s, \Xi) = \max_{\delta \in \mathbb{R}^R : \|\delta\|_0 \leq s, \|\delta\|_2 = 1} \delta' \Xi \delta \,.$$

We assume there exist constants $0 < \kappa_1 \leq \kappa_2$ *such that:*

$$\lim_{T,R \to \infty} \mathbb{P}[\kappa_1 \leq \kappa(s_T \log(T), \Xi) \leq \phi(s_T \log(T), \Xi) \leq \kappa_2] = 1 \,.$$

The last assumption restricts the penalties k and ν_i^j, $i = 1, \dots R$, $j = 1, \dots, d$. In particular, we require these penalties to be sufficiently harsh so as to dominate the gradient $\hat{S}_j(\lambda) = \Xi\lambda - \nabla_{\theta'} h^R(\tilde{\theta}) e_j$ of the unpenalised objective function, evaluated at the sparse approximation $\bar{\lambda}_j$.

Assumption K.5 (Penalties)**.** *The penalties satisfy:*

(1) For a sequence α_T *converging to zero such that* $\alpha_T R \to \infty$*:*

$$\mathbb{P}\left[\max_{j=1,\dots,d} \max_{i=1,\dots R} \left| (S_j(\bar{\lambda}_j))_i / \nu_i^j \right| \leq \frac{k}{T} \right] \geq 1 - \alpha_T \,,$$

where $k = (1+\epsilon)\sqrt{T \Phi^{-1}(1 - \frac{\alpha_T}{4Rd})}$ *for some* $\epsilon > 0$*, and where* Φ *denotes the cdf of a normal distribution.*

(2) There exist constants $a > 0$ *and* $b < \infty$*, such that:*

$$\lim \mathbb{P}\left[a \leq \min_{j=1,\dots d} \min_{i=1,\dots,R} \nu_i^j \leq \max_{j=1,\dots d} \max_{i=1,\dots,R} \nu_i^j \leq b \right] = 1 \,.$$

Under Assumptions K.3-K.5, it follows, by application of Lemma 26 in Luo et al. (2015), that there exists a constant K_λ and a sequence of ϵ_T converging to zero, such that, with probability at least $1 - \epsilon_T$

$$\max_{j=1,\dots,d} \|\bar{\lambda}_j - \hat{\lambda}_j\|_1 \leq K_\lambda \sqrt{\frac{s_T^2 \log(\frac{Rd}{\alpha_T})}{T}} \,, \tag{20}$$

where $\hat{\lambda}_j$ denotes the solution to program (17).

The bound in (20), together with Assumption 12, implies that:

$$\|\tilde{A}_R - A_R\|_2 \leq \|\tilde{A}_R - A_R\|_F + \|\bar{A}_R - A_R\|_2 = O_p\left(\sqrt{\frac{s_T^2 \log(\frac{Rd}{\alpha_T})}{T}} \right) \,.$$

If we assume that $\frac{s_T^2 \log(R)}{T} \to 0$, then $\|\tilde{A}_R - A_R\|_2 = o_p(1)$, and the Lasso approach consistently estimates the combination matrix. We can then derive the properties of the Lasso-based estimator in a similar vein as to Propositions 1 and 2 in Section 3. To see this, we observe that the estimator $\hat{\theta}^{\text{selected}}$ solves $S(\hat{\theta}^{\text{selected}}) = 0$, where:

$$S(\theta) = \tilde{A}_S h^R(\theta)\,.$$

If we define the population objective as $S_0(\theta) = A_R[\int_{\underline{p}}^{\overline{p}}(\hat{Q}_Y(u) - Q_Y(u|\theta)\boldsymbol{P}^R(u)]du$, then we can proceed as in the consistency proof of Proposition 1. Similarly, we can proceed as in the proof of Proposition 2 to obtain an asymptotic linear representation of the estimator.

Appendix L. Details on extensions

L.1. **"Residual" analysis of semi- and nonparametric models.** Consider the model for a scalar outcome Y outlined in the main text:

$$\begin{aligned} Y &= h(\epsilon, X; \gamma_0), \quad \gamma \in \Gamma \subseteq \mathcal{B}\,, \\ \epsilon &\sim F_{\theta_0}, \quad \theta_0 \in \Theta \subseteq \mathbb{R}^d\,, \end{aligned} \tag{21}$$

where we assume that the researcher has access to a first-step nonparametric estimator of γ_0; and that the scalar ϵ follows a continuous distribution function with true density f_ϵ. We assume that Γ is a subset of the Banach space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$. We consider the case where the data consists of independent copies of (Y, X).

Following Ichimura and Newey (2022), let $\gamma(F)$ denote the probability limit of the first-step estimator when the distribution of $Z := (X, Y)$ is F. Denote by F_0 the true distribution of Z, thus $\gamma_0 = \gamma(F_0)$. We seek to find a function $\boldsymbol{b}(u; Z; \gamma(F), \psi(F))$ such that $\mathbb{E}_F[\boldsymbol{b}(u; Z; \gamma(F), \psi(F))] = 0$, $\mathbb{E}_F[(\boldsymbol{b}(u; Z; \gamma, \psi(F))^2] < \infty$, and for every distribution function H (restricted except for regularity conditions) and $l \in \mathbb{N}$:

$$\frac{\partial}{\partial \tau}\int_{\underline{p}}^{\overline{p}} Q_{h^{-1}(Y,X;\gamma_\tau(H|F))}(u)P_l(u)du\Big|_{\tau=0} = \int_{\underline{p}}^{\overline{p}}\int \boldsymbol{b}(u; z; \gamma(F), \psi(F))P_l(u)H(dz)du\,, \tag{22}$$

where $\psi(F)$ denotes a set of nuisance parameters, and, for $\tau \in [0, 1]$, $\gamma_\tau(H|F) = \gamma(F + \tau(H - F)))$. The function $\boldsymbol{b}$ is known as a *first-step influence function*, quantifying the impact of estimating $\gamma(F)$ on the estimator of the L-moments.

Following Chernozhukov et al. (2022), we will focus on correcting L-moments by a sample average of an estimated version of the first-step influence function. To see why, take F equal to the empirical distribution function of Z, and $H = F_0$. We may then consider the following *distributional Taylor expansion* (Kennedy, 2023) around F:

$$\begin{aligned} &\sqrt{T}\left(\int_{\underline{p}}^{\overline{p}} Q_{h^{-1}(Y,X;\hat{\gamma})}(u)P_l(u)du - \int_{\underline{p}}^{\overline{p}} Q_{h^{-1}(Y,X;\gamma_0)}(u)P_l(u)du\right) = \\ &-\sqrt{T}\frac{\partial}{\partial \tau}\int_{\underline{p}}^{\overline{p}} Q_{h^{-1}(Y,X;\gamma_\tau(F_0|F))}(u)P_l(u)du\Big|_{\tau=0} + \sqrt{T}\int_{\underline{p}}^{\overline{p}} E_{\text{lin}}(u)P_l(u)du = \\ &-\int_{\underline{p}}^{\overline{p}}\int \boldsymbol{b}(u; z; \hat{\gamma}, \hat{\psi})P_l(u)F_0(dz)du + \sqrt{T}\int_{\underline{p}}^{\overline{p}} E_{\text{lin}}(u)P_l(u)du,\,, \end{aligned} \tag{23}$$

where the second equality follows from (22). By Bessel's inequality, $T\|\int_{\underline{p}}^{\overline{p}} E_{\text{lin}}(u)\boldsymbol{P}^R(u)du\|^2 \leq T\int_{\underline{p}}^{\overline{p}} E_{\text{lin}}(u)^2du$. Under smoothness conditions on the model, the linearisation error can be further shown to be bounded above by $T\int_{\underline{p}}^{\overline{p}} E_{\text{lin}}(u)^2du \leq CT\|\hat{\gamma}-\gamma_0\|_{\mathcal{B}}^4$ (see the discussion on Assumption 3.4 in Chernozhukov et al. (2018); and the examples in Section 4.3 of Kennedy (2023)). Therefore, in these settings, if $\|\hat{\gamma}-\gamma_0\|_{\mathcal{B}} = o_{F_0}(T^{-1/4})$, one would have that $T\|\int_{\underline{p}}^{\overline{p}} E_{\text{lin}}(u)\boldsymbol{P}^R(u)du\|^2 = o_{F_0}(1)$, and (23) would imply that correcting by an average of the first-step influence function may be employed to remove the effect of first-step estimation.

In the following examples, we provide calculations for the function

$$\boldsymbol{B}_l(Z;\gamma_0,\psi_0) := \int_{\underline{p}}^{\overline{p}} \boldsymbol{b}(u;Z;\gamma_0;\psi_0)P_l(u)du$$

in three models. For that, we rely on the following observation:

$$\begin{gathered} F_{h^{-1}(Y,X;\gamma_\tau)}(Q_{h^{-1}(Y,X;\gamma_\tau)}(u)) = u \quad \forall u\in(0,1), \tau\in(0,1) \implies \\ \frac{\partial}{\partial\tau}Q_{h^{-1}(Y,X;\gamma_\tau)}(u)\Big|_{\tau=0} = -\frac{1}{f_\epsilon(Q_\epsilon(u))}\frac{\partial}{\partial\tau}\mathbb{P}_{F_0}[Y\leq h(Q_\epsilon(u),X;\gamma_\tau)]\Big|_{\tau=0}. \end{gathered} \tag{24}$$

Whenever the order of differentiation and integration can be exchanged in (22), representation (24) will allow us to apply the results of Ichimura and Newey (2022) in our context.

Example 1 (Semiparametric conditional mean model)**.** The model is given by:

$$\begin{gathered} Y = \gamma_0(X) + \epsilon\,, \\ \epsilon \sim F_{\theta_0}, \quad \theta_0 \in \Theta \subseteq \mathbb{R}^d\,. \end{gathered}$$

where for every distribution H, the probability limit of the preliminary consistent estimator of γ_0 is given by:

$$\gamma(H) \in \text{argmin}_{w\in\mathcal{W}}\,\mathbb{E}_H[(Y-w(X))^2]\,,$$

with $\mathcal{W}$ ia closed linear subspace of $L_2(F_{0,X})$ that does not depend on H. Notice that $\frac{\partial}{\partial\tau}\mathbb{P}_{F_0}[Y\leq Q_\epsilon(u)+\gamma_\tau(X)]\Big|_{\tau=0} = \frac{\partial}{\partial\tau}\mathbb{E}_{F_0}[\mathbb{P}_{F_0}[Y\leq Q_\epsilon(u)+\gamma_\tau(X)|X]]\Big|_{\tau=0}$. If the order of differentiation and integration can be exchanged, Proposition 1 of Ichimura and Newey (2022) yields:

$$\boldsymbol{B}_l(Z;\gamma,\psi) = -\int_{\underline{p}}^{\overline{p}} \psi(u|X)(Y-\gamma(X))P_l(u)du\,,$$

where

$$\psi_0(u|X) \in \text{argmin}_{w\in\mathcal{W}}\,\mathbb{E}_{F_0}[(\omega_0(u|X)-w(X))^2]\,,$$

with $\omega_0(u|X) = \frac{1}{f_\epsilon(Q_\epsilon(u))}f_{\epsilon|X}(Q_\epsilon(u)|X)$.

Example 2 (Semiparametric conditional quantile model)**.** The model is given by

$$\begin{gathered} Y = \gamma_0(\epsilon|X)\,, \\ \epsilon|X \sim \text{Uniform}[0,1]\,, \end{gathered}$$

where, for each $\tau \in (0,1)$ and distribution H:

$$\gamma(H)(\tau|X) \in \text{argmin}_{w\in\mathcal{W}} \mathbb{E}_H[|Y - w(X)|(\tau \mathbf{1}\{Y - w(X) > 0\} + (1-\tau)\mathbf{1}\{Y - w(X) < 0\})]\,,$$

with the class $\mathcal{W}$ defined as in the previous example. In this model, the L-moment approach may be used as a specification testing tool, since the model implies $\epsilon \sim \text{Uniform}[0,1]$.

In this setting, if $\mathcal{W}$ contains constant functions, Proposition 1 of Ichimura and Newey (2022) yields:

$$\boldsymbol{B}_l(Z;\gamma,\psi) = -\int_{\underline{p}}^{\overline{p}} (u - \mathbf{1}\{Y < \gamma(u|X)\}) P_l(u) du$$

Example 3 (Linear-nonparametric instrumental variable model)**.** We consider the linear-nonparametric model of Athey et al. (2019):

$$\begin{aligned} Y &= \alpha_0(X) + \beta_0(X)W + \epsilon \\ \epsilon &\sim F_{\theta_0}, \quad \theta_0 \in \Theta \subseteq \mathbb{R}^d\,, \end{aligned}$$

where $\mathbb{E}_{F_0}[\epsilon|X] = 0$, but W is believed to be endogenous. We assume that the researcher has access to a scalar instrumental variable S satisfying $\mathbb{E}_{F_0}[\epsilon|X,S] = 0$. Following Athey et al. (2019), we consider consistent preliminary estimators of α and β whose probability limits are characterized by, for each H:

$$\begin{aligned} \mathbb{E}_H[(Y - \alpha(H)(X) - \beta(H)(X)W)|X] &= 0\,, \\ \mathbb{E}_H[S(Y - \alpha(H)(X) - \beta(H)(X)W)|X] &= 0\,. \end{aligned} \tag{25}$$

In this case, if $\mathbb{E}_{F_0}[S|X,W] = a_0(X) + b_0(X)W$ with $\mathbb{P}_{F_0}[b_0(X) \neq 0] = 1$ and $\mathbb{E}_{F_0}[W|X,S] = c_0(X) + d_0(X)S$ with $\mathbb{P}_{F_0}[d_0(X) \neq 0] = 1$, it follows by Proposition 3 of Ichimura and Newey (2022) that:

$$\boldsymbol{B}_l(Z;\gamma,\psi) = -\int_{\underline{p}}^{\overline{p}} \left(g(u|X)\frac{(S - a(X))}{b(X)} + h(u|X) \right) (Y - \alpha(X) - \beta(X)W) P_l(u) du\,,$$

where:

$$(g_0(u|X), h_0(u|X)) \in \text{argmin}_{o,p\in L^2(F_{0,X})} \mathbb{E}_{F_0}\left[(\omega_0(u|X,W) - o(X) - p(X)W)^2\right]\,,$$

with $\omega_0(u|X,S) = \frac{1}{f_\epsilon(Q_\epsilon(u))} f_{\epsilon|X,W}(Q_\epsilon(u)|X,W)$.

Once the form of correction is known, we follow the double machine learning literature (Chernozhukov et al., 2018, 2022; Kennedy, 2023) and propose a sample-split bias-corrected version of the L-moment estimator. Fix $\kappa \in (0,1)$. Partition the sample into two blocks, with $T_1 = \lfloor T\kappa \rfloor$ and $T_2 = T - \lfloor T\kappa \rfloor$ observations. Let $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the set of indices of each partition. We propose to estimate θ_0 through the following steps:

(1) Using $\mathcal{I}_1$, estimate γ_0. Denote by $\hat{\gamma}$ this first-step estimator.

(2) Using $\mathcal{I}_2$, compute, for each $u \in (0,1)$, $\hat{Q}_{\hat{\epsilon}}(u)$ as the u-th empirical quantile of $\{h^{-1}(Y_i, X_i; \hat{\gamma}) : i \in \mathcal{I}_2\}$.

(3) Using $\mathcal{I}_1$, estimate ψ_0. Denote by $\hat{\psi}$ this estimator.

(4) Compute the influence function adjustment:

$$\widehat{\boldsymbol{A}} = -\frac{1}{T_2} \sum_{i \in \mathcal{I}_2} \begin{pmatrix} \boldsymbol{B}_1(Z_i; \hat{\gamma}, \hat{\psi}) \\ \boldsymbol{B}_2(Z_i; \hat{\gamma}, \hat{\psi}) \\ \vdots \\ \boldsymbol{B}_R(Z_i; \hat{\gamma}, \hat{\psi}) \end{pmatrix} .$$

(5) Estimate the model by:

$$\hat{\theta}_{\mathcal{I}_2} \in \arg\inf_{\theta \in \Theta} \left[\int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_{\hat{\epsilon}}(u) - Q_{\epsilon}(u|\theta) \right) \mathbf{P}^R(u)' du - \hat{\boldsymbol{A}} \right] W^R \left[\int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_{\hat{\epsilon}}(u) - Q_Y(u|\theta) \right) \mathbf{P}^R(u) du - \hat{\boldsymbol{A}} \right] ,$$

where W^R is a possibly estimated weighting matrix.

For sample-splitting not to result in efficiency losses, one may adopt *cross-fitting*, i.e. one swaps the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ in the above sequence, and compute the final estimator as $\hat{\theta} = \kappa \hat{\theta}_{\mathcal{I}_1} + (1-\kappa)\hat{\theta}_{\mathcal{I}_2}$ and .

Consistency and asymptotic linearity of $\hat{\theta}_{\mathcal{I}_2}$ follows by standard uniform differentiability conditions, with proofs similar to those of Propositions A.1 and A.2, if the crucial conditions:

$$T \left\| \int_{\underline{p}}^{\bar{p}} E_{\text{lin}}(u) \boldsymbol{P}^R(u) du \right\|^2 = o_{F_0}(1) ,$$

and

$$\sqrt{T} \left\| \hat{\boldsymbol{A}} - \boldsymbol{A} - \tilde{\boldsymbol{A}} \right\|_2 = o_{F_0}(1) ,$$

hold. Here, $\boldsymbol{A}$ is given by:

$$\boldsymbol{A} = -\frac{1}{T_2} \sum_{i \in \mathcal{I}_2} \begin{pmatrix} \boldsymbol{B}_1(Z_i; \gamma_0, \psi_0) \\ \boldsymbol{B}_2(Z_i; \gamma_0, \psi_0) \\ \vdots \\ \boldsymbol{B}_R(Z_i; \gamma_0, \psi_0) \end{pmatrix} ,$$

and $\tilde{\boldsymbol{A}}$ is given by:

$$\tilde{\boldsymbol{A}} = - \begin{pmatrix} \int \boldsymbol{B}_1(z; \hat{\gamma}, \hat{\psi}) F_0(dz) \\ \vdots \\ \int \boldsymbol{B}_R(z; \hat{\gamma}, \hat{\psi}) F_0(dz) \end{pmatrix}$$

In our three examples, that $\sqrt{T}\|\hat{\boldsymbol{A}} - \boldsymbol{A} - \tilde{\boldsymbol{A}}\|_2 = o_{F_0}(1)$ holds follows by, first, applying Bessel's inequality to show that

$$T\|\hat{\boldsymbol{A}}-\boldsymbol{A}-\tilde{\boldsymbol{A}}\|_2^2 \leq \int_{\underline{p}}^{\overline{p}} T\left(\frac{1}{T_2}\sum_{i\in\mathcal{I}_2}\left[\boldsymbol{b}(u,Z_i;\gamma_0,\psi_0)-\boldsymbol{b}(u,Z_i;\hat{\gamma},\hat{\psi})-\int \boldsymbol{b}(u,z;\hat{\gamma},\hat{\psi})F_0(dz)\right]\right)^2 du\,.$$

That the right-hand side converges in probability to zero can then be established under weak consistency requirements on $\hat{\gamma}$ and $\hat{\psi}$. Importantly, one does not have to resort to Donsker conditions, due to the special structure provided by sample-splitting (see Kennedy, 2023, Lemma 1 and the accompanying discussion).

Under the above conditions, the estimator admits the following asymptotic linear representation:

$$\sqrt{T_2}(\hat{\theta}_{\mathcal{I}_2}-\theta_0) = -(J^R(\theta_0)'\Omega^R J^R(\theta_0))^{-1}J^R(\theta_0)'\Omega^R\left(\sqrt{T_2}\int_{\underline{p}}^{\overline{p}}(\hat{Q}_{\hat{\epsilon}}(u)-Q_{h^{-1}(Y,X;\hat{\gamma})}(u))\mathbf{P}^R(u)du-\sqrt{T_2}\boldsymbol{A}\right)+o_{F_0}(1)\,,$$

where $J^R(\theta) = -\int_{\underline{p}}^{\bar{p}}\nabla_{\theta'}Q_\epsilon(u|\theta)\mathbf{P}^R(u)du$, and, for each $\gamma\in\Gamma$, $Q_{h^{-1}(Y,X;\gamma)}(u)$ is the population quantile function of $h^{-1}(Y,X;\gamma)$.

If we assume that the conditions on the distribution of $h^{-1}(Y,X;\gamma)$ ensuring that a Bahadur-Kiefer approximation as discussed in Appendix H hold uniformly over γ in a neighbourhood of γ_0, then we may use these results as a tool for inference. For example, if the conditions in the statement of Theorem H.1 hold in such way, then we are able to show that:

$$\sqrt{T_2}(\hat{\theta}_{\mathcal{I}_2}-\theta_0) = -(J^R(\theta_0)'\Omega^R J^R(\theta_0))^{-1}J^R(\theta_0)'\Omega^R\left(\sqrt{T_2}(\boldsymbol{C}-\boldsymbol{A})\right)+o_{F_0}(1)\,,\,,$$

with where,

$$\boldsymbol{C} = \frac{1}{T_2}\sum_{i\in\mathcal{I}_2}\int_{\underline{p}}^{\overline{p}}\frac{1}{f_{h^{-1}(Y,X;\hat{\gamma})}(Q_{h^{-1}(Y,X;\hat{\gamma})}(u))}(\mathbf{1}\{h^{-1}(Y_i,X_i;\hat{\gamma})\leq Q_{h^{-1}(Y,X;\hat{\gamma})}(u)\}-u)\boldsymbol{P}_R(u)du\,.$$

Notice that, due to the nature of sample-splitting, conditional on $\hat{\gamma}$, both $\boldsymbol{C}$ and $\boldsymbol{A}$ are zero-mean random-variables whose variances and covariance can be estimated by sample-analogs. This result serves as a basis for inference in this setting. Moreover, the conditional variance of $\sqrt{T_2}\,[\boldsymbol{C}-\boldsymbol{A}]$ can be used to compute the optimal weighting scheme to be used in step (5) of our estimation.

L.2. **Details on prediction intervals.**

L.2.1. *Asymptotic validity.* We consider a setting that nests our application in the main text as a particular case. Specifically, we consider that the researcher posits the following potential outcome model for the untreated outcome that is observed absent a policy intervention in a panel of n units, indexed by i, over T periods, indexed by t:

$$\begin{aligned} Y_{it}(0) = h_t(\epsilon_{it},X_{it};\gamma_0), \quad \gamma\in\Gamma\subseteq\mathcal{B}\,,\\ \epsilon_{i,t}\sim F_{\theta_0}, \quad \theta_0\in\Theta\subseteq\mathbb{R}^d\,, \end{aligned} \tag{26}$$

where $e\mapsto h_t(e,x;\gamma_0)$ is strictly increasing, for every x in the support of X_{it}; and the distribution of $Y_{i,t}(0)$ has no point masses. The treatment intervention date t^* takes values in a set $\mathcal{T}\subseteq\mathbb{N}\cup\{\infty\}$,

which may include $t^* = \infty$ if, for some realisation of the random variables, the intervention would never take place. As in the main text, we assume that t^* is independent of ϵ_{i,t^*}. Fix a significance level $\alpha \in (0,1)$. Let $\hat{\gamma}_{n,\tau}$ and $\hat{\theta}_{n,\tau}$ denote the (generally unfeasible) estimators of γ_0 and θ_0 computed using the *untreated* potential outcomes of the n individuals in the panel and the first τ periods. The **actual** estimators are given by $\hat{\gamma}_{n,t^*-1}$ and $\hat{\theta}_{n,t^*-1}$ We assume that, for every $\tau < \sup \mathcal{T} \wedge T$, $\text{plim}_{n\to\infty} \sup_{x \in \text{supp}\, X_{i,\tau+1}} |h_{\tau+1}(Q_{\hat{\theta}_{n,\tau}}(1-\alpha), x; \hat{\gamma}_{n,\tau}) - h_{\tau+1}(Q_{\theta_0}(1-\alpha), x; \gamma_0)| = 0$. In the setting of the main text, this assumption requires that there always be, over alternative realisations of sampling uncertainty, sufficient pre-treatment periods to estimate the outcome model $Y_{it}(0)$ using lags as instruments.

Our proposed lower confidence region for a unit i is given by:

$$\hat{\mathcal{I}}_{i,1-\alpha} = [Y_{i,t^*} - h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\alpha), X_{i,t^*}; \hat{\gamma}_{n,t^*}), \infty)\,,$$

and, we may assume, without loss, that, in the event that $t^* > T$, the confidence region is estimated without uncertainty, as, in practice, nothing is reported in this case by the researcher and thus no "mistakes" are made.

The asymptotic validity of the confidence region can be shown by relying on a similar argument to the one in Appendix A of Alvarez and Ferman (2024). Specifically, we observe that, under the stated assumptions, $h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\alpha), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) \xrightarrow{p} h_{t^*}(Q_{\theta_0}(1-\alpha), X_{i,t^*}; \gamma_0)$, since, for any tolerance $\nu > 0$:

$$\lim_{n\to\infty} \mathbb{P}[|h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\alpha), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) - h_{t^*}(Q_{\theta_0}(1-\alpha), X_{i,t^*}; \gamma_0)| > \nu] \leq$$

$$\lim_{n\to\infty} \sum_{\tau\in\mathcal{T}} \mathbb{P}\left[\left\{\sup_{x\in\text{supp}\, X_{i,\tau}} |h_\tau(Q_{\hat{\theta}_{n,\tau-1}}(1-\alpha), x; \hat{\gamma}_{n,\tau-1}) - h_\tau(Q_{\theta_0}(1-\alpha), x; \gamma_0)| > \nu\right\} \cap \{t^* = \tau\}\right] =$$

$$\sum_{\tau\in\mathcal{T}} \lim_{n\to\infty} \mathbb{P}\left[\left\{\sup_{x\in\text{supp}\, X_{i,\tau}} |h_\tau(Q_{\hat{\theta}_{n,\tau-1}}(1-\alpha), x; \hat{\gamma}_{n,\tau-1}) - h_\tau(Q_{\theta_0}(1-\alpha), x; \gamma_0)| > \nu\right\} \cap \{t^* = \tau\}\right] = 0$$

(we pass the limit under the sum because at most the $T < \infty$ first terms in the sum are different than zero). Consequently, given that the distribution of $Y_{i,t}(0)$ has no point masses, it follows by the continuous mapping theorem that:

$$\mathbb{1}\{Y_{it^*}(1) - Y_{it^*}(0) \in \mathcal{I}_{i,1-\alpha}\} = \mathbb{1}\{Y_{it^*}(0) \in (-\infty, h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\alpha), X_{i,t^*}; \hat{\gamma}_{n,t^*-1})]\} \xrightarrow{p}$$
$$\mathbb{1}\{Y_{it^*}(0) \in (-\infty, h_{t^*}(Q_{\theta_0}(1-\alpha), X_{i,t^*}; \gamma_0)]\} = \mathbb{1}\{\epsilon_{it^*} \leq Q_{\theta_0}(1-\alpha)\}\,,$$

where the last passage followed by the map h_t being strictly increasing in the first entry. Asymptotic coverage then follows by the bounded convergence theorem and the independence between t^* and ϵ_{t^*}, by noting that:

$$\lim_{n\to\infty} \mathbb{P}[Y_{it^*}(1) - Y_{it^*}(0) \in \mathcal{I}_{i,1-\alpha}] = \mathbb{P}[\epsilon_{it^*} \leq Q_{\theta_0}(1-\alpha)] = \sum_{\tau\in\mathcal{T}} \mathbb{P}[t^* = \tau]\mathbb{P}[\epsilon_{i,\tau} \leq Q_{\theta_0}(1-\alpha)|t^* = \tau] = \sum_{\tau\in\mathcal{T}} \mathbb{P}[t^* = \tau]\mathbb{P}[\epsilon_{i,\tau} \leq Q_{\theta_0}(1-\alpha)] = 1-\alpha$$

L.2.2. *Bonferrroni correction.* The prediction interval presented in the previous section has an asymptotic justification. However, in finite samples, estimation error of θ_0 and γ_0 may induce miscoverage. Following an idea similar to Cattaneo et al. (2021), we consider two simple Bonferroni-style adjustments that may improve upon coverage. The first adjustment relies on the assumption that ϵ_{i,t^*} is independent of the data used to estimate θ_0 and γ_0. In this case, for $\beta \in (0,1)$ and $\phi \geq 0$, a simple calculation reveals that coverage probability of the modified interval $\hat{I}_{i,\beta} - \phi$ can be bounded below by:

$$\mathbb{P}[Y_{it^*}(1) - Y_{it^*}(0) \in \hat{I}_{i,\beta} - \phi] = \mathbb{P}[Y_{it^*}(0) \leq h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\beta), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) + \phi] \geq$$
$$\mathbb{P}[\{\epsilon_{it^*} \leq Q_{\theta_0}(1-\beta)\} \cap \{h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\beta), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) - h_{t^*}(Q_{\theta_0}(1-\beta), X_{i,t^*}; \gamma_0) \geq -\phi\}] = .$$
$$(1-\beta)\mathbb{P}[\ \{h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\beta), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) - h_{t^*}(Q_{\theta_0}(1-\beta), X_{i,t^*}; \gamma_0) \geq -\phi\}]$$

Now, if $h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\beta), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) - h_{t^*}(Q_{\theta_0}(1-\beta), X_{i,t^*}; \gamma_0)$ is approximately normally distributed around zero in larger samples, which can be the case if a central limit theorem applies to this term, and if the variance of this term can be estimated, then one can calibrate ϕ and β based on the normal distribution as to ensure coverage at the $(1-\alpha)$ level. Notice that there is some flexibility in the choice of ϕ and β in this case, which can be chosen so as to maximizie the low endpoint of the interval.

The previous correction relied on independence between ϵ_{i,t^*} and the data used in the estimation. If that is not the case, than a simpler Bonferroni bound may be obtained by applying the union bound to the probability of *miscoverage* of $\hat{I}_{i,\beta} - \phi$. In this case, one calibrates β and ϕ such that $\beta + \mathbb{P}[\ \{h_{t^*}(Q_{\hat{\theta}_{n,t^*-1}}(1-\beta), X_{i,t^*}; \hat{\gamma}_{n,t^*-1}) - h_{t^*}(Q_{\theta_0}(1-\beta), X_{i,t^*}; \gamma_0) \leq -\phi\}] \leq \alpha$. Notice that this adjustment is necessarily more conservative than the previous one, though the difference can be small.

L.3. **Conditional models.** Suppose the researcher postulates a conditional model $\{Q_{Y|X}(|X;\theta) : \theta \in \Theta\}$, where Y is a scalar outcome of interest, X are a set of controls, and $\Theta \subseteq \mathbb{R}^d$. We consider the estimator $\hat{\theta}$ of the true parameter θ_0:

$$\hat{\theta} \in \text{argmin}_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \left\| \Omega(X_t)^{1/2} \left(\int_{\underline{p}}^{\overline{p}} \left(\hat{Q}_{Y|X}(u|X_t) - Q_{Y|X}(u|X_t;\theta) \right) \boldsymbol{P}_R(u) du \right) \right\|_2^2, \tag{27}$$

where Ω is a $R \times R$ symmetric positive semidefinite weighting function of the controls X, and $\hat{Q}_{Y|X}$ is a preliminary nonparametric estimator of the conditional quantile process $(u,x) \mapsto Q_{Y|X}(u|x)$. The formulation may be seen as an extension of Ai and Chen (2003)'s approach to models defined by conditional moments to a conditional L-moment setting.

Suppose that we rely on the nonparametric quantile series regression estimator of Belloni et al. (2019), and that the conditions underlying their Comment 3 and Theorem 2 are satisfied. In this case, under identifiability and uniform differentiability conditions similar to those used in Propositions 1 and 2, it is possible to show that the estimator admits the asymptotic linear representation.

$$\sqrt{T}(\hat{\theta}-\theta_0) = -\left(\frac{1}{T}\sum_{t=1}^{T}\left(\int_{\underline{p}}^{\overline{p}}\partial_{\theta'}Q_{Y|X}(u|X_t;\theta_0)\boldsymbol{P}_R(u)du\right)'\Omega(X_t)\left(\int_{\underline{p}}^{\overline{p}}\partial_{\theta'}Q_{Y|X}(u|X_t;\theta_0)\boldsymbol{P}_R(u)du\right)\right)^{-1}\times \left(\frac{1}{T}\sum_{t=1}^{T}\left(\int_{\underline{p}}^{\overline{p}}\partial_{\theta'}Q_{Y|X}(u|X_t;\theta_0)\boldsymbol{P}_R(u)du\right)'\Omega(X_t)\left(\int_{\underline{p}}^{\overline{p}}\sum_{s=1}^{T}\frac{1}{\sqrt{T}}Z_t'J_T^{-1}(u)Z_s(u-\mathbf{1}\{U_s\leq u\})\boldsymbol{P}_R(u)du\right)\right) +o_P(1)\,, \tag{28}$$

where Z_t is the vector of transformations of X_t used in the series estimator;

$$J_T(u)=\frac{1}{T}\sum_{t=1}^{T}f_{Y|X}(Q_Y(u|X_t)|X_t)Z_tZ_t',$$

and $\{U_t\}_{t=1}^T$ are independent uniform random variables, independent from $\{X_t\}_{t=1}^T$. It then follows that the optimal weighting scheme is given by:

$$\Omega(X_t)^*=\mathbb{V}\left[\left(\int_{\underline{p}}^{\overline{p}}\sum_{s=1}^{T}\frac{1}{\sqrt{T}}Z_t'J_T^{-1}(u)Z_s(u-\mathbf{1}\{U_s\leq u\})\boldsymbol{P}_R(u)du\right)\Big|X_1,\ldots,X_T\right]^{-}.$$

This optimal weighting scheme can be estimated by relying on an estimator of $u\mapsto J_T(u)$ (Belloni et al. (2019) discuss nonparametric estimators of this function; a semiparametric version of this quantity that relies on a preliminary estimator of θ_0 can also be used); and on simulation from uniform random variables.

Inference using normal critical values can be performed under the assumptions underlying Theorem 5 of Belloni et al. (2019), which ensures a strong approximation of the series estimator to a Gaussian process. A weighted bootstrap approximation can also be used, if the assumptions underlying Theorem 6 of Belloni et al. (2019) hold.

Next, we note that, by considering u-specific orthogonalizations of the Z_t when estimating a quantile function $x\mapsto Q_{Y|X}(u|x)$, it is without loss to assume that, for every u:

$$\sqrt{f_{Y|X}(Q_Y(u|X_t)|X_t)f_{Y|X}(Q_Y(u|X_s)|X_s)}Z_t'Z_s=\mathbf{1}\{t=s\}\,.$$

Using this fact, we are able to show that the variance of the leading term of the linear representation of the optimally weighted generalized L-moment estimator is:

$$\left(\left(\frac{1}{T}\sum_{t=1}^{T}\int_{\underline{p}}^{\overline{p}}\partial_{\theta'}Q_{Y|X}(u|X_t;\theta_0)\boldsymbol{P}_R(u)du\right)'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{V}\left[\int_{\underline{p}}^{\overline{p}}\frac{(u-\mathbf{1}\{U_t\leq u\})}{f_{Y|X}(u|X_t)}\boldsymbol{P}_R(u)du\Big|X_t\right]\right)^{-}\left(\frac{1}{T}\sum_{t=1}^{T}\int_{\underline{p}}^{\overline{p}}\partial_{\theta'}Q_{Y|X}(u|X_t;\theta_0)\boldsymbol{P}_R(u)du\right)\right)^{-1},$$

which closely resembles the variance of the optimally weighted L-moment estimator in the *unconditional* case. Indeed, the above expression differs from the unconditional version in that it averages the relevant matrices across the sample support points of X. Proceeding by analogy to Appendix I, we are then able to show that, when we rely on an orthonormal *basis* $\{P_l\}_l$ and $0 \leq \underline{p} < \overline{p} \leq 1$, our proposed estimator is efficient, in the sense that its asymptotic variance coincides with the inverse of the expected conditional Fisher information matrix.

Appendix M. Analytical expressions for the Generalized Extreme Value and Generalized Pareto Distributions

This Appendix contains analytical expressions of the theoretical L-moments, as well as the gradients and Hessians used in computing the L-moment estimator and its higher-order expansion for both the GEV and GPD families of distributions. The file `gev_npd.nb` provides a Mathematica notebook that analytically derives some of these expressions.

M.1. **Generalized Extreme Value distribution.**

M.1.1. *Theoretical L-moments and its derivatives.* Following Hosking (1986), the l-th probability-weighted moment of a GEV distribution with location parameter m, scale parameter r and shape parameter k is given by:

$$\int_0^1 Q_\theta(u)u^l du = \begin{cases} \frac{m}{l+1} + \frac{r\left(1-(l+1)^{-k}\Gamma(k+1)\right)}{k(l+1)} & \text{if } k \neq 0 \\ \frac{m}{l+1} + \frac{r\log(l+1)+\gamma r}{l+1} & \text{if } k = 0 \end{cases},$$

where Γ denotes the Gamma function and γ the Euler-Mascheroni constant.

The gradient of the probability-weighted moment is given by:

$$\begin{bmatrix} \frac{1}{l+1} \\ \frac{1-(l+1)^{-k}\Gamma(k+1)}{k(l+1)} \\ \frac{r\left((l+1)^{-k}\Gamma(k+1)\log(l+1)-(l+1)^{-k}\Gamma(k+1)\psi^{(0)}(k+1)\right)}{k(l+1)} - \frac{r\left(1-(l+1)^{-k}\Gamma(k+1)\right)}{k^2(l+1)} \end{bmatrix},$$

at $k \neq 0$, with $\psi^{(j)}(x)$ denoting the j-th derivative of $x \mapsto \log(\Gamma(x))$ (the polygamma function of order j) and:

$$\begin{bmatrix} \frac{1}{l+1} \\ \frac{\log(l+1)+\gamma}{l+1} \\ -\frac{r\left(6\log^2(l+1)+12\gamma\log(l+1)+6\gamma^2+\pi^2\right)}{12(l+1)} \end{bmatrix}$$

at $k = 0$. Finally, the Hessian of the probability-weighted moment is given by:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{(l+1)^{-k}\Gamma(k+1)\log(l+1)-(l+1)^{-k}\Gamma(k+1)\psi^{(0)}(k+1)}{k(l+1)} - \frac{1-(l+1)^{-k}\Gamma(k+1)}{k^2(l+1)} \\ 0 & \frac{(l+1)^{-k}\Gamma(k+1)\log(l+1)-(l+1)^{-k}\Gamma(k+1)\psi^{(0)}(k+1)}{k(l+1)} - \frac{1-(l+1)^{-k}\Gamma(k+1)}{k^2(l+1)} & \frac{2r\left(1-(l+1)^{-k}\Gamma(k+1)\right)}{k^3(l+1)} - \frac{2r\left((l+1)^{-k}\Gamma(k+1)\log(l+1)-(l+1)^{-k}\Gamma(k+1)\psi^{(0)}(k+1)\right)}{k^2(l+1)} + \frac{r\left(-(l+1)^{-k}\Gamma(k+1)\log^2(l+1)+(l+1)^{-k}(-\Gamma(k+1))\psi^{(0)}(k+1)^2-(l+1)^{-k}\Gamma(k+1)\psi^{(1)}(k+1)+2(l+1)^{-k}\Gamma(k+1)\psi^{(0)}(k+1)\log(l+1)\right)}{k(l+1)} \end{bmatrix}$$

at $k \neq 0$ and:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{6\log^2(l+1)+12\gamma\log(l+1)+6\gamma^2+\pi^2}{12(l+1)} \\ 0 & -\frac{6\log^2(l+1)+12\gamma\log(l+1)+6\gamma^2+\pi^2}{12(l+1)} & \frac{r\left(2\log^3(l+1)+6\gamma\log^2(l+1)+\left(6\gamma^2+\pi^2\right)\log(l+1)+2\gamma^3+\gamma\pi^2-2\psi^{(2)}(1)\right)}{6(l+1)} \end{bmatrix}$$

at $k = 0$.

M.1.2. *Quantile function and its derivatives.* Again following Hosking (1986), the u-th quantile of a GEV distribution with location parameter m, scale parameter r and shape parameter k is given by:

$$Q_\theta(u) = \begin{cases} m + \frac{r\left(1-(-\log(u))^k\right)}{k} & \text{if } k \neq 0 \\ m - r\log(-\log(u)) & \text{if } k = 0 \end{cases}$$

The gradient function of the u-th quantile is given by:

$$\begin{bmatrix} 1 \\ \frac{1-(-\log(u))^k}{k} \\ -\frac{r\left(1-(-\log(u))^k\right)}{k^2} - \frac{r\log(-\log(u))(-\log(u))^k}{k} \end{bmatrix}$$

at $k \neq 0$ and:

$$\begin{bmatrix} 1 \\ -\log(-\log(u)) \\ -\frac{1}{2}r\log^2(-\log(u)) \end{bmatrix}$$

at $k = 0$. The Hessian is given by:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1-(-\log(u))^k}{k^2} - \frac{\log(-\log(u))(-\log(u))^k}{k} \\ 0 & -\frac{1-(-\log(u))^k}{k^2} - \frac{\log(-\log(u))(-\log(u))^k}{k} & \frac{2r(1-(-\log(u))^k)}{k^3} + \frac{2r\log(-\log(u))(-\log(u))^k}{k^2} - \frac{r\log^2(-\log(u))(-\log(u))^k}{k} \end{bmatrix}$$

at $k \neq 0$, and:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2}\log^2(-\log(u)) \\ 0 & -\frac{1}{2}\log^2(-\log(u)) & -\frac{1}{3}r\log^3(-\log(u)) \end{bmatrix}$$

at $k = 0$. Finally, for estimating the higher-order terms pertaining to estimation of the optimal weighting function, we require to compute the gradient, with respect to the model parameters, of the quantile density function $Q'(u|\theta) = \frac{1}{f(Q(u|\theta)|\theta))}$. The gradient is given by:

$$0 \quad \frac{(-\log(u))^{\text{shape}-1}}{u} \quad \frac{\text{scale}\log(-\log(u))(-\log(u))^{\text{shape}-1}}{u}$$

at $k \neq 0$ and:

$$\begin{bmatrix} 0 \\ \frac{(-\log(u))^{k-1}}{u} \\ \frac{r\log(-\log(u))(-\log(u))^{k-1}}{u} \end{bmatrix}$$

and

$$\begin{bmatrix} 0 \\ -\frac{1}{u\log(u)} \\ -\frac{r\log(-\log(u))}{u\log(u)} \end{bmatrix}$$

at $k = 0$.

M.2. **Generalized Pareto Distribution.**

M.2.1. *Theoretical L-moments and its derivatives.* Following Hosking (1986), the l-th probability-weighted moment of a GPD distribution with location parameter m, scale parameter r and shape parameter k is given by:

$$\int_0^1 Q_\theta(u)u^l du = \begin{cases} \frac{m}{l+1} + \frac{r\left(1-\frac{\Gamma(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}\right)}{k(l+1)} & \text{if } k \neq 0 \\ \frac{m}{l+1} + \frac{r\Gamma(l+1)(\psi^{(0)}(l+2)+\gamma)}{\Gamma(l+2)} & \text{if } k = 0 \end{cases}.$$

The gradient of the l-th probability-weighted moment is given by:

$$\begin{bmatrix} \frac{1}{l+1} \\ \frac{1-\frac{\Gamma(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}}{k(l+1)} \\ \frac{r\left(\frac{\Gamma(k+1)\Gamma(l+2)\psi^{(0)}(k+l+2)}{\Gamma(k+l+2)} - \frac{\Gamma(k+1)\psi^{(0)}(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}\right)}{k(l+1)} - \frac{r\left(1-\frac{\Gamma(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}\right)}{k^2(l+1)} \end{bmatrix}$$

at $k \neq 0$ and:

$$\begin{bmatrix} \frac{1}{l+1} \\ \frac{\psi^{(0)}(l+2)+\gamma}{l+1} \\ -\frac{r\Gamma(l+1)\left(6\psi^{(0)}(l+2)^2 + 12\gamma\psi^{(0)}(l+2) - 6\psi^{(1)}(l+2) + 6\gamma^2 + \pi^2\right)}{12\Gamma(l+2)} \end{bmatrix}$$

at $k = 0$. The Hessian is given by:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{\frac{\Gamma(k+1)\Gamma(l+2)\psi^{(0)}(k+l+2)}{\Gamma(k+l+2)} - \frac{\Gamma(k+1)\psi^{(0)}(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}}{k(l+1)} - \frac{1-\frac{\Gamma(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}}{k^2(l+1)} \\ 0 & \frac{\frac{\Gamma(k+1)\Gamma(l+2)\psi^{(0)}(k+l+2)}{\Gamma(k+l+2)} - \frac{\Gamma(k+1)\psi^{(0)}(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}}{k(l+1)} - \frac{1-\frac{\Gamma(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}}{k^2(l+1)} & \frac{2r\left(1-\frac{\Gamma(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}\right)}{k^3(l+1)} - \frac{2r\left(\frac{\Gamma(k+1)\Gamma(l+2)\psi^{(0)}(k+l+2)}{\Gamma(k+l+2)} - \frac{\Gamma(k+1)\psi^{(0)}(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}\right)}{k^2(l+1)} \\ & & + \frac{r}{k(l+1)}\left(-\frac{\Gamma(k+1)\psi^{(0)}(k+1)^2\Gamma(l+2)}{\Gamma(k+l+2)} + \frac{2\Gamma(k+1)\psi^{(0)}(k+1)\Gamma(l+2)\psi^{(0)}(k+l+2)}{\Gamma(k+l+2)}\right. \\ & & \left. + \frac{\Gamma(k+1)\Gamma(l+2)\psi^{(1)}(k+l+2)}{\Gamma(k+l+2)} - \frac{\Gamma(k+1)\Gamma(l+2)\psi^{(0)}(k+l+2)^2}{\Gamma(k+l+2)} - \frac{\Gamma(k+1)\psi^{(1)}(k+1)\Gamma(l+2)}{\Gamma(k+l+2)}\right) \end{bmatrix}$$

at $k \neq 0$, and:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{\Gamma(1+l)\left(6\gamma^2+\pi^2+12\gamma\psi^{(0)}(2+l)+6\psi^{(0)}(2+l)^2-6\psi^{(1)}(2+l)\right)}{12\Gamma(2+l)} \\ 0 & -\frac{\Gamma(1+l)\left(6\gamma^2+\pi^2+12\gamma\psi^{(0)}(2+l)+6\psi^{(0)}(2+l)^2-6\psi^{(1)}(2+l)\right)}{12\Gamma(2+l)} & \frac{r}{6}\left(\frac{1}{\Gamma(2+l)\Gamma(1+l)}\left(2\gamma^3+\gamma\pi^2+6\gamma\psi^{(0)}(2+l)^2\right.\right. \\ & & \left. + \psi^{(0)}(2+l)\left(6\gamma^2+\pi^2-12\psi^{(1)}(2+l)\right)-6\gamma\psi^{(1)}(2+l)-2\psi^{(2)}(1)\right) \\ & & \left. + \frac{2\left(\psi^{(0)}(2+l)^3+3\psi^{(0)}(2+l)\psi^{(1)}(2+l)+\psi^{(2)}(2+l)\right)}{1+l}\right) \end{bmatrix}$$

at $k = 0$.

M.2.2. *Quantile function and its derivatives.* The quantile function is given by:

$$Q_\theta(u) = \begin{cases} m + \frac{r\left(1-(1-u)^k\right)}{k} & \text{if } k \neq 0 \\ m + \frac{r\left(1-(1-u)^k\right)}{k} & \text{if } k = 0 \end{cases}.$$

The gradient with respect to the model parameters is given by:

$$\begin{bmatrix} 1 \\ \frac{1-(1-u)^k}{k} \\ -\frac{r(1-(1-u)^k)}{k^2} - \frac{r(1-u)^k\log(1-u)}{k} \end{bmatrix}$$

at $k \neq 0$ and:

$$\begin{bmatrix} 1 \\ -\log(1-u) \\ -\frac{1}{2}r\log^2(1-u) \end{bmatrix}$$

at $k = 0$. The Hessian is given by:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1-(1-u)^k}{k^2} - \frac{(1-u)^k\log(1-u)}{k} \\ 0 & -\frac{1-(1-u)^k}{k^2} - \frac{(1-u)^k\log(1-u)}{k} & \frac{2r(1-(1-u)^k)}{k^3} + \frac{2r(1-u)^k\log(1-u)}{k^2} - \frac{r(1-u)^k\log^2(1-u)}{k} \end{bmatrix}$$

at $k \neq 0$ and:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2}\log^2(1-u) \\ 0 & -\frac{1}{2}\log^2(1-u) & -\frac{1}{3}r\log^3(1-u) \end{bmatrix}$$

at $k = 0$. Finally, the gradient of $Q'_\theta(u)$ with respect to θ is:

$$\begin{bmatrix} 0 \\ (1-u)^{k-1} \\ r(1-u)^{k-1}\log(1-u) \end{bmatrix}$$

at $k \neq 0$ and:

$$\begin{bmatrix} 0 \\ \frac{1}{1-u} \\ \frac{r\log(1-u)}{1-u} \end{bmatrix}$$

at $k = 0$.

Appendix N. Simple sufficient conditions for $L^2(0,1)$ consistency of empirical quantiles

The following is a useful lemma for establishing $L^2(0,1)$ convergence of empirical quantiles.

Lemma N.1. *Let $Y_1, \ldots, Y_T$ be a random sample from a distribution with finite $(2+\delta)$-moment that admits Lebesgue density f such that $u \mapsto f(Q_Y(u))$ is continuous, where Q_Y is the quantile function of this distribution. Then, denoting by $\hat{Q}_Y$ the empirical quantiles from $Y_1, \ldots, Y_T$, we have that, as $T \to \infty$:*

$$\int_0^1 (\hat{Q}_Y(u) - Q_Y(u))^2 du \xrightarrow{P} 0 \ , .$$

Proof. Note that, by Fubini theorem, we can always write:

$$\mathbb{E}\left[\int_0^1 (\hat{Q}_Y(u) - Q_Y(u))^2 du\right] = \int_0^1 \mathbb{E}[(\hat{Q}_Y(u) - Q_Y(u))^2] du = \int_0^1 g_T(u) du \,,$$

where $g_T(u) = \mathbb{E}[(\hat{Q}_Y(u) - Q_Y(u))^2]$. Now, given that the distribution has a finite moment, and under the stated smoothness assumptions on $f \circ Q_Y$, $\lim_{T\to\infty} g_T(u) = 0$ for every $u \in (0,1)$ by Proposition 1 of Mason (1984). But then, we observe that, taking $\rho = \frac{2+\delta}{2} > 1$:

$$\int_0^1 g_T(u)^\rho du \leq \int_0^1 \mathbb{E}[|\hat{Q}_Y(u) - Q_Y(u)|^{2+\delta}]du \leq$$
$$2^{2+\delta}\left(\mathbb{E}\left[\int_0^1 |\hat{Q}_Y(v)|^{2+\delta}dv\right] + \mathbb{E}\left[\int_0^1 |Q_Y(v)|^{2+\delta}dv\right]\right) =$$
$$2^{2+\delta}\left(\mathbb{E}\left[\frac{\sum_{t=1}^T |Y_t|^{2+\delta}}{T}\right] + \mathbb{E}[|Y_1|^{2+\delta}]\right) = 2^{2+1+\delta}\mathbb{E}[|Y_1|^{2+\delta}] < \infty\,,$$

where the first inequality follows from Lyapunov inequality, and Fubini theorem is used in the second inequality. Since the quantity $2^{2+1+\delta}\mathbb{E}[|Y_1|^{2+\delta}]$ does not depend on T, it follows from Theorem 4.6.2 in Durrett that the sequence $\{g_t\}_t$ is uniformly integrable. Consequently, by Theorem 4.6.3 of Durrett (2019), $\mathbb{E}\left[\int_0^1 (\hat{Q}_Y(u) - Q_Y(u))^2 du\right] \to 0$, and a final application of Markov inequality yields that $\int_0^1 (\hat{Q}_Y(u) - Q_Y(u))^2 du \xrightarrow{P} 0$, as desired. □

In the previous proof, we relied on a finite $(2+\delta)$-moment to establish uniform integrability of g_T. If one replaces this moment condition with the assumption that there exist real constants C, k_1, k_2 such that $f(Q_Y(u))^{-1} \leq Cu^{k_1}(1-u)^{k_2}, \forall u \in (0,1)$, then Mason (1984, pages 248-249) shows that the distribution has a moment (and consequently, $g_T(u) \to 0$ for every $u \in (0,1)$ by his Proposition 1) and that the $\{g_t\}_t$ are uniformly integrable. Consequently, $\int_0^1 g_T(u)du \to 0$, and consistency in $L^2(0,1)$ holds. We state this alternative result below:

Lemma N.2. *Let $Y_1, \ldots, Y_T$ be a random sample from a distribution that admits Lebesgue density f such that $u \mapsto f(Q_Y(u))$ is continuous, where Q_Y is the quantile function of this distribution. If there exist real constants C, k_1, k_2 such that $f(Q_Y(u))^{-1} \leq Cu^{k_1}(1-u)^{k_2}, \forall u \in (0,1)$, then, denoting by $\hat{Q}_Y$ the empirical quantiles from $Y_1, \ldots, Y_T$, we have that, as $T \to \infty$:*

$$\int_0^1 (\hat{Q}_Y(u) - Q_Y(u))^2 du \xrightarrow{P} 0\ ,.$$

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Alvarez, L. and C. Biderman (2024). The learning effects of subsidies to bundled goods: a semi-parametric approach. *arXiv:2311.01217*.

Alvarez, L. A. F. and B. Ferman (2024). On "imputation of counterfactual outcomes when the errors are predictable": Discussions on misspecification and suggestions of sensitivity analyses. *Journal of Business & Economic Statistics 42*(4), 1123–1127.

Alvarez-Andrade, S. and S. Bouzebda (2013). Strong approximations for weighted bootstrap of empirical and quantile processes with applications. *Statistical Methodology 11*, 36–52.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148 – 1178.

Awasthi, P., A. Das, W. Kong, and R. Sen (2022). Trimmed maximum likelihood estimation for robust learning in generalized linear models.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(1), 4 – 29. Annals: In Honor of Roger Koenker.

Bhatia, R. (1997). *Matrix Analysis*. Springer New York.

Cattaneo, M. D., Y. Feng, and R. T. and (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association 116*(536), 1865–1880. PMID: 35756161.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica 90*(4), 1501–1535.

Choi, E., P. Hall, and B. Presnell (2000, 06). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika 87*(2), 453–465.

Csorgo, M. and P. Revesz (1978, 07). Strong approximations of the quantile process. *Annals of Statistics 6*(4), 882–894.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics 117*(1), 55–93.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics 152*(1), 28–36.

Donald, S. G. and W. K. Newey (2001). Choosing the number of instruments. *Econometrica 69*(5), 1161–1191.

Durrett, R. (2019). *Probability: Theory and Examples* (5 ed.). Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Fan, J., R. P. Masini, and M. C. Medeiros (2023). Bridging factor and sparse models. *The Annals of Statistics 51*(4), 1692 – 1717.

Firpo, S., A. F. Galvao, C. Pinto, A. Poirier, and G. Sanroman (2022). Gmm quantile regression. *Journal of Econometrics 230*(2), 432–452.

Franguridi, G., B. Gafarov, and K. Wuthrich (2022). Bias correction for quantile regression estimators.

Götze, F., A. Naumov, V. Spokoiny, and V. Ulyanov (2019). Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli 25*(4A), 2538 – 2563.

Gupta, R. D. and D. Kundu (2001). Generalized exponential distribution: Different method of estimations. *Journal of Statistical Computation and Simulation 69*(4), 315–337.

Hadi, A. S. and A. Luceño (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis 25*(3), 251–272.

Hosking, J. R. (1986). *The theory of probability weighted moments*. IBM Research Division, TJ Watson Research Center New York, USA.

Hosking, J. R. M. and J. R. Wallis (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics 29*(3), 339–349.

Ichimura, H. and W. K. Newey (2022). The influence function of semiparametric estimators. *Quantitative Economics 13*(1), 29–61.

Imbens, G. W. and M. Kolesár (2016, 10). Robust Standard Errors in Small Samples: Some Practical Advice. *The Review of Economics and Statistics 98*(4), 701–712.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: a review.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Kotz, S. and S. Nadarajah (2000). *Extreme value distributions: theory and applications*. Imperial College Press.

Kreyszig, E. (1989). *Introductory Functional Analysis with Applications*. Wiley.

Kulik, R. (2007, 11). Bahadur–kiefer theory for sample quantiles of weakly dependent linear processes. *Bernoulli 13*(4), 1071–1090.

Lee, T.-H., A. Ullah, and H. Wang (2017). The second-order bias and mse of quantile estimators. *Unpublished manuscript*.

Lee, T.-H., A. Ullah, and H. Wang (2018). The second-order bias of quantile estimators. *Economics Letters 173*, 143–147.

Luo, Y. et al. (2015). *High-dimensional econometrics and model selection*. Ph. D. thesis, Massachusetts Institute of Technology.

Mason, D. M. (1984, 02). Weak convergence of the weighted empirical quantile process in $l^2(0,1)$. *Annals of Probability 12*(1), 243–255.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica 27*(4), 575–595.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*(4), 809–837.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Volume 4, pp. 2111 – 2245. Elsevier.

Newey, W. K. and R. J. Smith (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Okui, R. (2009). The optimal choice of moments in dynamic panel data models. *Journal of Econometrics 151*(1), 1–16.

Phillips, P. C. B. (1991). A shortcut to lad estimator asymptotics. *Econometric Theory 7*(4), 450–463.

Pötscher, B. M. and I. R. Prucha (1997). *Dynamic Nonlinear Econometric Models*. Springer Berlin Heidelberg.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*. Springer Berlin Heidelberg.

Rothenberg, T. J. (1984). Chapter 15 approximating the distributions of econometric estimators and test statistics. In *Handbook of Econometrics*, Volume 2, pp. 881–935. Elsevier.

Rubin, D. B. (1981). The Bayesian Bootstrap. *Annals of Statistics 9*(1), 130 – 134.

Singh, V. (1998). *Entropy-based parameter estimation in hydrology*, Volume 30. Springer Science & Business Media.

Turlach, B. A. and A. Weingessel (2011). *quadprog: Functions to solve Quadratic Programming Problems.* R package version 1.5-4.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38*(3/4), 330–336.