Inference in parametric models with many L-moments

Luis Antonio Fantozzi Alvarez

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Doctor of Science

Program: Statistics
Advisor: Prof. Dr. Pedro Morettin
Coadvisor: Prof. Dr. Chang Chiann

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da Capes e do CNPq.

São Paulo
September, 2022

Inference in parametric models with many L-moments

Luis Antonio Fantozzi Alvarez

This version of the thesis includes the corrections and modifications suggested by the Examining Committee during the defense of the original version of the work, which took place on September 29, 2022.

A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Examining Committee:

Prof. Dr. Chang Chiann (co-advisor) – IME-USP
Prof. Dr. Silvia Ferrari – IME-USP
Prof. Dr. Cristine Pinto – Insper
Prof. Dr. Marcelo Fernandes – FGV-SP
Prof. Dr. Ricardo Masini – Princeton University

Ficha catalográfica elaborada com dados inseridos pelo(a) autor(a)
Biblioteca Carlos Benjamin de Lyra
Instituto de Matemática e Estatística
Universidade de São Paulo

Alvarez, Luis Antonio Fantozzi
Inference in Parametric Models with Many L-moments / Luis Antonio Fantozzi Alvarez; orientador, Pedro Alberto Morettin; coorientadora, Chang Chiann. - São Paulo, 2022.
74 p.: il.

Tese (Doutorado) - Programa de Pós-Graduação em Estatística / Instituto de Matemática e Estatística / Universidade de São Paulo.
Bibliografia
Versão corrigida

1. L-momentos. 2. Método dos momentos generalizados. 3. Métodos de seleção de hiperparâmetros. 4. Modelos semiparamétricos. I. Morettin, Pedro Alberto. II. Título.

Bibliotecárias do Serviço de Informação e Biblioteca
Carlos Benjamin de Lyra do IME-USP, responsáveis pela
estrutura de catalogação da publicação de acordo com a AACR2:
Maria Lúcia Ribeiro CRB-8/2766; Stela do Nascimento Madruga CRB 8/7534.

Agradecimentos

Agradeço meus orientadores, professores Chang Chiann e Pedro Morettin, pela orientação cuidadosa e diligente. A professora Chang inicialmente aceitou orientar-me nesse tópico a despeito da distância deste do cotidiano da Estatística; e sempre se dispôs a ajudar-me em minhas escolhas acadêmicas. A ela sou imensamente grato por isso. Agradeço também o professor Morettin por ter aceitado participar dessa orientação quando o trabalho ainda se resumia a um projeto de pesquisa, e por sua disposição em auxiliar-me sempre que necessário.

Aos membros da banca, professores Sílvia Ferrari, Cristine Pinto, Marcelo Fernandes e Ricardo Masini, agradeço-lhes a leitura minuciosa e os comentários pertinentes realizados durante a defesa. As críticas e sugestões certamente melhoraram a exposição atual e orientarão a agenda de pesquisa decorrente deste trabalho. Como é de praxe, enfatizo que quaisquer erros porventura encontrados nesta tese são de minha inteira responsabilidade.

Foi na disciplina ministrada pela professora Silvia Ferrari que aprendi sobre o conceito de L-momentos. Agradeço-a pelo incentivo à pesquisa quando este projeto era tão somente uma ideia incipiente e mal-afigurada; e pelo curso por ela ministrado, que certamente mudou minha visão sobre a natureza da inferência estatística.

A professora Cristine Pinto foi minha orientadora de mestrado e, desde então, sempre se dispôs a aconselhar-me com relação a minhas ideias de pesquisa e decisões profissionais. A ela, agradeço a presteza e o tempo dispendido – seus ensinamentos certamente foram essenciais à minha formação.

Ao professor Ricardo Masini, que me ensinou boa parte do que sei sobre inferência estatística, agradeço as discussões sempre estimulantes e o apoio que tem dado à minha trajetória acadêmica. Ao professor Marcelo Fernandes, agradeço as indicações de agendas de pesquisa possíveis decorrentes de meu trabalho e o apoio à minha carreira.

Aos amigos que me acompanham desde a graduação, Ariel, Gabriel, Lucas, Matias, Rodrigo, Vinicius e Victor, e àqueles que vêm desde o mestrado, Gabriel, Leonardo, Nícolas e Victor, agradeço o companheirismo, as discussões acadêmicas e não acadêmicas sempre estimulantes, as parcerias de pesquisa e os momentos de diversão. Sem vocês, certamente esse percurso teria sido muito mais difícil.

Por fim, agradeço meus pais, Ana e Vicente, que sempre me apoiaram em todas as etapas de minha formação; e a minha esposa, Bruna, pelo amor paciente e companhia infatigável durante todas as etapas de meu doutorado – que venham tantas outras jornadas para caminharmos juntos.

Resumo

Luis Antonio Fantozzi Alvarez. **Inferência em modelos paramétricos com muitos L-momentos**. Tese (Doutorado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

L-momentos são valores esperados de combinações lineares de estatísticas de ordem que proveem alternativas robustas aos momentos tradicionais. A estimação de modelos paramétricos por meio da minimização da distância entre L-momentos amostrais e teóricos – um procedimento conhecido na literatura como "método dos L-momentos" – produz estimadores de menor erro quadrático médio que aqueles de máxima verossimilhança em pequenas amostras de diversas distribuições conhecidas. Não obstante, a escolha do número de L-momentos usados na estimação é tipicamente *ad-hoc*: pesquisadores costumeiramente usam o mesmo número de L-momentos que parâmetros, de modo a satisfazer uma condição de ordem para identificação do modelo. Nesta tese, mostra-se que, ao escolher o número de L-momentos apropriadamente e ponderando-os corretamente, é possível construir um estimador que se mostra de menor risco que a abordagem tradicional de L-momentos e que máxima verossimilhança em amostras finitas, e ainda assim se mantém assintoticamente eficiente. Esse resultado é obtido propondo-se um estimador de método "generalizado" de L-momentos e derivando suas propriedades estatísticas num ambiente em que o número de L-momentos varia com o tamanho amostral. Em seguida, propõem-se métodos para selecionar automaticamente o número ótimo de L-momentos em uma dada amostra. Como extensão, mostra-se que uma modificação da abordagem proposta pode ser usada na estimação de modelos semiparamétricos de efeitos de tratamento em experimentos aleatorizados controlados. Essa extensão produz um estimador eficiente e com propriedades computacionais atraentes. Os ganhos associados a essa nova abordagem são ilustrados aplicando a metodologia proposta no contexto de um experimento aleatório conduzido em São Paulo, Brasil. De maneira mais geral, com essa extensão, espera-se introduzir a abordagem baseada em L-momentos como um procedimento atrativo em ambientes em que estimadores de máxima verossimilhança semi/não paramétricos são computacionalmente complicados.

Palavras-chave: L-momentos. Método dos momentos generalizados. Métodos de seleção de hiperparâmetros. Modelos semiparamétricos.

Abstract

Luis Antonio Fantozzi Alvarez. **Inference in parametric models with many L-moments**. Thesis (Doctorate). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2022.

L-moments are expected values of linear combinations of order statistics that provide robust alternatives to traditional moments. The estimation of parametric models by matching sample L-moments – a procedure known as "method of L-moments" – has been shown to outperform maximum likelihood estimation in small samples from popular distributions. The choice of the number of L-moments to be used in estimation remains *ad-hoc*, though: researchers typically set the number of L-moments equal to the number of parameters, as to achieve an order condition for identification. In this thesis, we show that, by properly choosing the number of L-moments and weighting these accordingly, we are able to construct an estimator that outperforms both MLE and the traditional L-moment approach in finite samples, and yet does not suffer from efficiency losses asymptotically. We do so by considering a "generalised" method of L-moments estimator and deriving its asymptotic properties in a framework where the number of L-moments varies with sample size. We then propose methods to automatically select the number of L-moments in a given sample. As an extension, we show that a modification of our approach can be be used in the estimation of semiparametric models of treatment effects in randomised controlled trials (RCTs). This extension produces an efficient estimator with attractive computational properties. We illustrate the usefulness of our approach by applying it to data on an RCT conducted in São Paulo, Brazil. With such extension, we hope more generally to introduce L-moment-based estimation as an attractive procedure in settings where semi- and nonparametric maximum likelihood estimation is computationally complicated.

Keywords: L-moments. Generalised method of moments. Tuning parameter selection methods. Semiparametric models.

List of Symbols

$a \wedge b$	$\min\{a, b\}$
$a \vee b$	$\max\{a, b\}$

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Overview

L-moments, expected values of linear combinations of order statistics, were introduced by Hosking (1990) and have been successfully applied in areas as diverse as computer science (Hosking, 2007; Yang et al., 2021), hydrology (Wang, 1997; Sankarasubramanian and Srinivasan, 1999; Das, 2021; Boulange et al., 2021), meteorology (Wang and Hutson, 2013; Šimková, 2017; Li et al., 2021) and finance (Gourieroux and Jasiak, 2008; Kerstens et al., 2011). By appropriately combining order statistics, L-moments offer robust alternatives to traditional measures of dispersion, skewness and kurtosis. Models fit by matching sample L-moments (a procedure labeled "method of L-moments" by Hosking (1990)) have been shown to outperfom maximum likelihood estimators in small samples from flexible distributions such as generalised extreme value (Hosking et al., 1985; Hosking, 1990), generalised Pareto (Hosking and Wallis, 1987; Broniatowski and Decurninge, 2016), generalised exponential (Gupta and Kundu, 2001) and Kumaraswamy (Dey et al., 2018).

Statistical analyses of L-moment-based parameter estimators rely on a framework where the number of moments is fixed (Hosking, 1990; Broniatowski and Decurninge, 2016). Practitioners often choose the number of L-moments equal to the number of parameters in the model, so as to achieve the order condition for identification. This raises the question of whether overidentifying restrictions, together with the optimal weighting of L-moment conditions, could improve the efficiency of "method of L-moments" estimators, as in the framework of generalized-method-of-moment (GMM) estimation (Hansen, 1982). Another natural question would be how to choose the number of L-moments in finite samples, as it is well-known from GMM theory that increasing the number of moments with a fixed sample size can lead to substantial biases (Newey and Smith, 2004). In the end, one can only ask if, by correctly choosing the number of L-moments and under an appropriate weighting scheme, it may not be possible to construct an estimator that outperforms maximum likelihood estimation in small samples and yet achieves the Cramér-Rao bound assymptotically. Intuitively, the answer appears to be positive, especially if one takes into account that Hosking (1990) shows L-moments characterise distributions with finite first moments.

The goal of this thesis lies in answering the questions outlined in the previous paragraph. Specifically, we propose to study L-moment-based estimation in a context where: (i) the number of L-moments varies with sample size; and (ii) weighting is used in order to optimally account for overidentifying conditions. In this framework, we derive sufficient conditions on the L-moment estimator under which consistency and asymptotic normality can be achieved. We also show that, under iid data and the optimal weighting scheme, the L-moment estimator achieves the Cramér-Rao lower bound. We provide simulation evidence that our L-moment approach outperforms (in a mean-squared error sense) both the conventional L-moment estimator and MLE in smaller samples; and works as well as MLE in larger samples. We then construct methods to automatically select the number of L-moments used in estimation. For that, we rely on higher order expansions of the method-of-L-moment estimator, similarly to the procedure of Donald and Newey (2001) and Donald et al. (2009) in the context of GMM. We use these expansions to find a rule for choosing the number of L-moments so as to minimise the estimated (higher-order) mean-squared error. We also consider an approach based on ℓ_1-regularisation (Luo et al., 2015). With these tools, we hope to introduce a fully automated procedure for estimating parametric density models that improves upon maximum likelihood and the conventional L-moment approach in small samples, and yet does not underperform in larger datasets. Finally, we extend our approach to accommodate the estimation of semiparametric models of treatment effects in randomised controlled trials (RCTs). Our proposal produces an efficient estimator with attractive computational properties. We illustrate the usefulness of this approach by using real data from an RCT conducted in Brazil. With such extension, we hope more generally to introduce L-moment-based estimation as an attractive procedure in settings where semi- and nonparametric maximum likelihood estimation is computationally complicated.

The remainder of this thesis is organised as follows. In the next section, we briefly review L-moments and parameter estimation based on these quantities. Chapter 2 works out the asymptotic properties of our proposed estimator. Chapter 3 proposes methods to select the number of L-moments. Chapter 4 discusses our semiparametric extension and illustrates it with an empirical application. Chapter 5 concludes.

1.2 L-moments: definition and estimation

Consider a scalar random variable Y with distribution function F and finite first moment. For $r \in \mathbb{N}$, Hosking (1990) defines the r-th L-moment as:

$$\lambda_r := \int_0^1 Q_Y(u) P^*_{r-1}(u) du \,, \tag{1.1}$$

where $Q_Y(u) := \inf\{y \in \mathbb{R} : F(y) \geq u\}$ is the quantile function of Y, and $P^*_r(u) = \sum_{k=0}^r (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} u^k$ are shifted Legendre polynomials.[1] Expanding the polynomials and using the quantile representation of a random variable (Billingsley, 2012, Theorem 14.1),

[1] Legendre polynomials are defined by applying the Gramm-Schmidt orthogornalisation process to the polynomials $1, x, x^2, x^3 \ldots$ defined on $[-1, 1]$ (Kreyszig, 1989, p. 176-180). If P_r denotes the r-th Legendre polynomial, shifted Legendre polynomials are related to the standard ones through the affine transformation $P^*_r(u) = P_r(2u - 1)$ (Hosking, 1990).

we arrive at the equivalent expression:

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}[Y_{(r-k):r}] , \tag{1.2}$$

where, $Y_{j:l}$ is the j-th order statistic of a random sample from F with l observations. Equation (1.2) motivates our description of L-moments as the expected value of linear combinations of order statistics. Notice that the first L-moment corresponds to the expected value of Y.

To see how L-moments may offer “robust” alternatives to conventional moments, it is instructive to consider, as in Hosking (1990), the second L-moment. In this case, we have:

$$\lambda_2 = \frac{1}{2}\mathbb{E}[Y_{2:2} - Y_{1:2}] = \frac{1}{2}\int\int (\max\{y_1, y_2\} - \min\{y_1, y_2\})\, F(dy_1)F(dy_2) = \frac{1}{2}\mathbb{E}|Y_1 - Y_2| ,$$

where Y_1 and Y_2 are independent copies of Y. This is a measure of dispersion. Indeed, comparing it with the variance, we have:

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{1}{2}\mathbb{E}[(Y_1 - Y_2)^2] ,$$

from which we note that the variance puts more weight to larger differences.

Next, we discuss sample estimators of L-moments. Let $Z_1, Z_2 \dots Z_T$ be an identically distributed sample of T observations, where each Z_t, $t = 1, \dots , T$, is distributed according to F. A natural estimator of the r-th L-moment is the sample analog of (1.1), i.e.

$$\hat{\lambda}_r = \int_0^1 \hat{Q}_Y(u) P^*_{r-1}(u) du , \tag{1.3}$$

where $\hat{Q}_Y$ is the empirical quantile process:

$$\hat{Q}_Y(u) = Z_{i:T}, \quad \text{if} \quad \frac{i-1}{T} < u \leq \frac{i}{T} ,$$

with $Z_{i:T}$ being the i-th sample order statistic. The estimator given by (1.3) is generally biased (Hosking, 1990; Broniatowski and Decurninge, 2016). When observations $Z_1, Z_2, \dots Z_T$ may be assumed to be independent, researchers thus often resort to an unbiased estimator of λ_r, which is given by an empirical analog of (1.2):

$$\tilde{\lambda}_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \binom{T}{r}^{-1} \sum_{1 \leq i_1, i_2 \leq \dots \leq i_r \leq T} Z_{i_{r-k}:T} . \tag{1.4}$$

In practice, it is not necessary to iterate over all size r subsamples of $Z_1, \dots Z_T$ to compute the sample r-th L-moment through (1.4). Hosking (1990) provides a direct formula that avoids such computation.

We are now ready to discuss the estimation of parametric models based on matching L-moments. Suppose that F belongs to a parametric family of distribution functions $\{F_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ and $F = F_{\theta_0}$ for some $\theta_0 \in \Theta$. Let $l_r(\theta) := \int_0^1 P^*_{r-1}(u)Q(u|\theta)du$ denote the theoretical r-th L-moment, where $Q(\cdot|\theta)$ is the quantile function associated with F_θ. Let $H^L(\theta) := (\lambda_1(\theta), \lambda_2(\theta), \dots, \lambda_L(\theta))'$, and $\hat{H}^L$ be the vector stacking estimators for the first L L-moments (e.g. (1.3) or (1.4)). Researchers then usually estimate θ by solving:

$$H^d(\theta) - \hat{H}^d = 0\,.$$

As discussed in Section 1.1, this procedure has been shown to lead to efficiency gains over maximum likelihood estimation in small samples from several distributions. Nonetheless, the choice of L-moments $L = d$ appears rather *ad-hoc*, as it is based on an order condition for identification. One may then wonder whether increasing the number of L-moments used in estimation – and weighting these properly –, might lead to a more efficient estimator in finite samples. Moreover, if one correctly varies the number of L-moments with sample size, it may be possible to construct an estimator that does not undeperform MLE even asymptotically. The latter appears especially plausible if one considers the result in Hosking (1990), who shows that L-moments characterise a distribution with finite first moment.

In light of the preceding discussion, we propose to analyse the behaviour of the estimator:

$$\hat{\theta} \in \arg\inf_{\theta\in\Theta}(H^L(\theta) - \hat{H}^L)'M^L(H^L(\theta) - \hat{H}^L)\,, \tag{1.5}$$

where L may vary with sample size; and M^L is a (possibly estimated) weighting matrix. In Chapter 2, we work out the asymptotic properties of this estimator in a framework where $T, L \longrightarrow \infty$.[2] We derive sufficient conditions for an asymptotic linear representation of the estimator to hold. We also show that the estimator is asymptotically efficient, in the sense that, under iid data and when optimal weights are used, the variance of the leading term of its asymptotic representation converges to the inverse of the Fisher information matrix. We then conduct a small Monte Carlo exercise which showcases the gains associated with our approach. Specifically, we show that our L-moment approach entails mean-squared error gains over both the conventional L-moment approach and MLE in smaller samples, and performs as well as MLE in larger samples. In light of these results, in Chapter 3 we propose to construct a semiautomatic method of selection of the number of L-moments by working with higher-order expansions of the mean-squared error of the estimator – in a similar fashion to what has already been done in the GMM literature (Donald and Newey, 2001; Donald et al., 2009; Okui, 2009; Abadie et al., 2019). We also consider an approach based on ℓ_1-regularisation borrowed from the GMM literature (Luo et al., 2015). Finally, in Chapter 4, we consider an extension of our methodology that can be used in the estimation of semiparametric models of treatment effects in RCTs.

In the next chapter, we will focus on the case where estimated L-moments are given by

[2] As it will become clear in Chapter 2, our framework also nests the case with fixed L as a special case by properly filling the weighting matrix M^L with zeros.

(1.3). As shown in Hosking (1990), under random sampling and finite second moments, for each $r \in \mathbb{N}$, $\hat{\lambda}_r - \tilde{\lambda}_r = O_p(T^{-1})$, which implies that the estimator in (1.5) using either (1.3) or (1.4) as $\hat{H}^L$ are asymptotically equivalent **when** L **is fixed**. In our framework, where L varies with T, we need a stronger, **joint** result. In particular, to obtain first order equivalence between both approaches, it would suffice that, as $T, L \longrightarrow \infty$, $\left(\sum_{r=1}^{L}(2r-1)|\hat{\lambda}_r - \tilde{\lambda}_r|^2\right)^{\frac{1}{2}} = o_p(T^{-1/2})$. We provide sufficient conditions for this to hold in the lemma below. Importantly, our result does not require random sampling nor finite second moments, implying that, under the conditions of Lemma 1, the L-moment estimator $\tilde{\lambda}_r$ may be used with dependent data or heavy-tailed distributions, provided L grows sufficiently slowly[3] and some law of large numbers holds for the sample mean of $|Z_t|$.

Lemma 1. *Suppose that* $\frac{1}{T}\sum_{t=1}^{T}|Z_t| = O_P(1)$. *If* $L^{10}/T \longrightarrow 0$, *we have that* $\left(\sum_{r=1}^{L}(2r-1)|\hat{\lambda}_r - \tilde{\lambda}_r|^2\right)^{\frac{1}{2}} = o_p(T^{-1/2})$.

Proof. See Appendix A. □

If we assume the conditions in the previous lemma, then the results in Chapter 2 remain valid if the L-moment estimator (1.3) is replaced with (1.4). These results concern the first order asymptotic behaviour of the estimator. The higher-order behaviour will be clearly different, as, for instance, (1.4) is unbiased under random sampling whereas (1.3) is generally not.

We could also consider alternative L-moment estimators where we replace the càglàd quantile estimator in (1.4) by an alternative function. For example, we could use a linear interpolation of the order statistics as the estimator of Q_Y. In this case, first-order equivalence between methods follows from known results on the estimation of the quantile process (Neocleous and Portnoy, 2008).

[3] If we assume random sampling and finite second moments, then it may be possible to weaken the rate requirement by adapting our proof and resorting to law of large number of linear functions of order statistics (Zwet, 1980).

Chapter 2

Asymptotic properties of the method of L-moments estimator with many moments

2.1 Setup

We consider a setting where we have a sample with T identically distributed observations, $Y_1, Y_2 \dots Y_T$, $Y_t \sim F$ for $t = 1, 2 \dots T$, where F belongs to a parametric family $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$; and $F = F_{\theta_0}$ for some $\theta_0 \in \Theta$. We will analyse the behaviour of the estimator:

$$\hat{\theta} \in \arg\inf_{\theta\in\Theta} \sum_{k=1}^{L}\sum_{l=1}^{L}\left(\int_{\underline{p}}^{\bar{p}}\left[\hat{Q}_Y(u) - Q_Y(u|\theta)\right]P_k(u)du\right) w_{k,l}^L \left(\int_{\underline{p}}^{\bar{p}}\left[\hat{Q}_Y(u) - Q_Y(u|\theta)\right]P_l(u)du\right), \quad (2.1)$$

where $\hat{Q}_Y(\cdot)$ is the empirical quantile process:

$$\hat{Q}_Y(u) = Y_{i:T}, \quad \text{if } \frac{i-1}{T} < u \leq \frac{i}{T},$$

with $Y_{i:T}$ being the i-th sample order statistic; $Q_Y(\cdot|\theta)$ is the quantile function associated with F_θ; $\{w_{k,l}^L\}_{1\leq k,l\leq L}$ are a set of (possibly estimated) weights; $\{P_k\}_{1\leq k\leq L}$ are a set of quantile "weighting" functions; and $0 \leq \underline{p} < \bar{p} \leq 1$. This setting encompasses the method-of-L-moment estimation discussed in the previous section, and actually expands upon it. Indeed, by choosing $P_k(u) = \sqrt{2k-1} \cdot P_{k-1}^*(u)$, where $P_k^*(u)$ are the shifted Legendre polynomials on $[0, 1]$, $0 = \underline{p} < \bar{p} = 1$, we have the L-moment-based estimator in (1.5) using (1.3) as an estimator for the L-moments. [1] We leave $\underline{p} < \bar{p}$ fixed throughout.[2] All limits are taken **jointly** with respect to T **and** L.

[1] The rescaling of the polynomials by $\sqrt{2k-1}$ is used so they present unit $L^2[0, 1]$-norm and thus constitute an orthonormal sequence (not just orthogonal). Since the weights $w_{k,l}^L$ are unrestricted (up to regularity conditions), this leads to the same estimator as (1.5) under suitable choices of weighting matrices.

[2] In Comment 5 later on, we briefly discuss an extension to sample-size-dependent trimming.

To facilitate analysis, we let $\mathbf{P}^L(u) := (P_1(u), P_2(u) \dots P_L(u))'$; and write W^L for the $L \times L$ matrix with entry $W^L_{i,j} = w^L_{i,j}$. We may then rewrite our estimator in matrix form as:

$$\hat{\theta} \in \arg\inf_{\theta\in\Theta} \left[\int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^L(u)' du\right] W^L \left[\int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta)\right) \mathbf{P}^L(u) du\right].$$

2.2 Consistency

The aim of this section is to derive conditions under which our estimator is consistent.

We impose the following assumptions on our environment. In what follows, we write $Q_Y(\cdot) = Q_Y(\cdot|\theta_0)$.

Assumption 1 (Consistency of empirical quantile process)**.** *The empirical quantile process is uniformly consistent on* $(\underline{p}, \bar{p})$,[3]*, i.e.*

$$\sup_{u\in(\underline{p},\bar{p})} |\hat{Q}_Y(u) - Q_Y(u)| \xrightarrow{P^*} 0\,, \tag{2.2}$$

where $\xrightarrow{P^*}$ *denotes convergence in (outer) probability.*[4]

Assumption 1 is satisfied in a variety of settings under rather weak dependence conditions. For example, if Y_1, $Y_2 \dots Y_T$ are iid and the family $\{F_\theta : \theta \in \Theta\}$ is continuous with a (common) compact support; then (2.2) follows with $\underline{p} = 0$ and $\overline{p} = 1$ (Ahidar-Coutrix and Berthet, 2016, Proposition 2.1.). Cf., *inter alia*, Yoshihara (1995) and Portnoy (1991) for sufficient conditions under dependent observations.

Assumption 2 (Quantile weighting functions)**.** *The functions* $\{P_l : l \in \mathbb{N}\}$ *constitute an orthonormal sequence on* $L^2[0, 1]$.

Assumption 2 above is satisfied by (rescaled) shifted Legendre polynomials, shifted Jacobi polynomials and other weighting functions.

Next, we impose restrictions on the estimated weights. In what follows, we write, for a $c \times d$ matrix A, $\|A\|_2 = \sqrt{\lambda_{\max}(A'A)}$.

Assumption 3 (Estimated weights)**.** *There exists a sequence of nonstochastic symmetric positive semidefinite matrices* Ω^L *such that, as* $T, L \longrightarrow \infty$, $\|W^L - \Omega^L\|_2 = o_{p^*}(1)$; $\|\Omega^L\|_2 = O(1)$.

---

[3] For the main proofs in this section, it would be sufficient to assume convergence in the $L^2[\underline{p}, \bar{p}]$ norm. We state results in the sup norm because convergence results on the quantile process in the literature are usually proved in $L^\infty(\underline{p}, \bar{p})$. See Mason (1984) and Barrio et al. (2005) for results in the L^2 norm; and more recently Kaji (2019) for results in the L^1 norm.

[4] We state our main assumptions and results in outer probability in order to abstract from measurability concerns. We note these results are equivalent to convergence in probability when the appropriate measurability assumptions hold.

Assumption 3 restricts the range of admissible weight matrices. Notice that $W^L = \Omega^L = I_L$ trivially satisfies these assumptions. By the triangle inequality, Assumption 3 implies that $\|W^L\|_2 = O_{P^*}(1)$.

Finally, we introduce our identifiability assumption. For $X \in L^2[0,1]$, let $\|X\|_{L^2[0,1]} = \left(\int_0^1 X(u)^2 du\right)^{\frac{1}{2}}$:

Assumption 4 (Strong identifiability and suprema of L^2 norm of parametric quantiles)**.** *For each $\epsilon > 0$:*

$$\liminf_{L\to\infty} \inf_{\theta\in\Theta:\|\theta-\theta_0\|_2\geq\epsilon} \left[\int_{\underline{p}}^{\bar{p}} (Q_Y(u|\theta) - Q_Y(u|\theta_0))\, \mathbf{P}^L(u)' du\right] \Omega^L \left[\int_{\underline{p}}^{\bar{p}} (Q_Y(u|\theta) - Q_Y(u|\theta_0))\, \mathbf{P}^L(u) du\right] > 0\,.$$

Moreover, we require that $\sup_{\theta\in\Theta}\|Q_Y(\cdot|\theta)\mathbb{1}_{[\underline{p},\bar{p}]}\|_{L^2[0,1]} < \infty$.

The first part of this assumption is closely related to the usual notion of identifiability in parametric distribution models. Indeed, if Θ is compact, $\theta \mapsto \|Q(\cdot|\theta)\|_{L^2[0,1]}$ is continuous, $\underline{p} = 0$, $\bar{p} = 1$, the $\{P_l\}_l$ constitute an orthonormal *basis* in $L^2[0,1]$ and $W^L = \mathbb{I}_L$, then part 1 is equivalent to identifiability of the parametric family $\{F_\theta\}_\theta$.

As for the second part of the assumption, we note that boundedness of the L^2 norm of parametric quantiles uniformly in θ is satisfied in several settings. If the parametric family $\{F_\theta : \theta \in \Theta\}$ has common compact support, then the assumption is trivially satisfied. More generally, if we assume Θ is compact and $Q_Y(u|\theta)$ is jointly continuous and bounded on $[\underline{p}, \bar{p}] \times \Theta$, then the condition follows immediately from Weierstrass' theorem, as in this case: $\sup_{\theta\in\Theta}\|Q_Y(\cdot|\theta)\mathbb{1}_{[\underline{p},\bar{p}]}\|_{L^2[0,1]} \leq \sqrt{\bar{p}-\underline{p}} \cdot \sup_{(\theta,u)\in\Theta\times[\underline{p},\bar{p}]} |Q_Y(u|\theta)| < \infty$.

In what follows, we prove our first result:

Proposition 1. *Suppose Assumptions 1 to 4 hold. Then $\hat{\theta} \xrightarrow{P^*} \theta_0$ as $L, T \longrightarrow \infty$.*

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

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

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

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

$$h_0^L(\theta) := \int_{\underline{p}}^{\bar{p}} (Q_Y(u) - Q_Y(u|\theta))\, \mathbf{P}^L(u) du\,.$$

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:

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

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

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

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

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

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 $\epsilon > 0$:

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

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. □

2.3 Asymptotic linear representation

Following the usual argument in the Generalised Method of Moments literature (Newey and McFadden, 1994), we will consider the case where $h^L(\theta) = \int_{\underline{p}}^{\bar{p}} \left(\hat{Q}_Y(u) - Q_Y(u|\theta) \right) \mathbf{P}^L(u) du$ is differentiable on a neighborhood of θ_0. In this case, whenever $\hat{\theta}$ is within such neighborhood, the estimator satisfies the first order condition:

$$
\nabla_{\theta'} h^L(\hat{\theta})' W^L h^L(\hat{\theta}) = 0 \,. \tag{2.3}
$$

where $\nabla_{\theta'} h^L(\tilde{\theta})$ is the Jacobian of h^L with respect to θ, evaluated at $\tilde{\theta}$. In order to obtain (2.3), it suffices that:

Assumption 5. *There exists an open ball $\mathcal{O}$ in $\mathbb{R}^d$ containing θ_0 such that $\mathcal{O} \subseteq \Theta$ and $Q_Y(u|\theta)$ is differentiable on $\mathcal{O}$, uniformly in $u \in [\underline{p}, \bar{p}]$. Moreover, $\theta \mapsto Q_Y(u|\theta)$ is **continuously** differentiable on $\mathcal{O}$ for each u; and, for each $\theta \in \mathcal{O}$, $\nabla_{\theta'} Q_Y(\cdot|\theta)$ is square integrable on $[\underline{p}, \bar{p}]$.*

Straightforward application of the dominated convergence theorem shows that Assumption 5 implies $h^L(\theta)$ 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), implying that (2.3) holds wpa1.

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:

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

where $\nabla_{\theta'} \widetilde{h^L(\theta)}$ is the $L \times d$ matrix where each line l is equal to $-\int_{\underline{p}}^{\bar{p}} \nabla_{\theta'} Q_Y(u|\tilde{\theta}(u)) P_l(u) du$, 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^L yields:

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

The crucial step now is to work out asymptotic tightness of a normalization of $h^L(\theta_0)$. For that, we will assume weak convergence of the empirical quantile process, which we state below:

Assumption 6. *$\sqrt{T}(\hat{Q}_Y(\cdot) - Q_Y(\cdot))$ converges weakly in $L^\infty(\underline{p}, \bar{p})$ to a zero-mean Gaussian process B with continuous sample paths and covariance kernel Γ.*[5]

Weak convergence of the empirical quantile process has been derived in a variety of settings, ranging from iid data (Vaart, 1998, Corollary 21.5) to nonstationary and weakly dependent observations (Portnoy, 1991). In the iid setting, if the family $\{F_\theta : \theta \in \Theta\}$ is continuously differentiable with strictly positive density f_θ over a (common) compact support; then weak-convergence holds with $\underline{p} = 0$ and $\bar{p} = 1$. In this case, the covariance kernel is $\Gamma(i, j) = \frac{(i \wedge j - ij)}{f_Y(Q_Y(i)) f_Y(Q_Y(j))}$.

Assumption 6 entails that:

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

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

[5] It would be sufficient to assume $\| \sqrt{T}(Q_Y(\cdot) - \hat{Q}_Y(\cdot)) \mathbb{1}_{[\underline{p}, \bar{p}]} \|_{L^2[0,1]}^2 = O_{P^*}(1)$, which is implied by weak convergence in L^2.

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

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 $T \geq N$:

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

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

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

and since d is fixed, we conclude that:

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

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

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

If we add to Assumption 5:

Assumption 7. *$Q_Y(u|\theta)$ is **twice** continuously differentiable on Θ, for each $u \in [\underline{p}, \overline{p}]$. Moreover, $\sup_{\theta \in \Theta} \sup_{u \in [\underline{p},\bar{p}]} \|\nabla_{\theta\theta'} Q_Y(u|\theta)\|_2 < \infty$;*

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

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

where the remainder r^{TL} satisfies $\|r^{TL}\|_2^2 = o_{P^*}(1)$. We would like a condition that allows us to invert $(\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0) + r^{TL})$ with high probability. We state this below:

Assumption 8. *The smallest eigenvalue of $\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0)$ is bounded away from 0, uniformly in L.*

Assumption 8 is similar to the rank condition used in the proof of asymptotic normality of M-estimators (Newey and McFadden, 1994), which is known to be equivalent to a local identification condition under rank-regularity assumptions (Rothenberg, 1971). In our setting, where L varies with sample size, we show in Appendix B that a stronger version of Assumption 4 implies Assumption 8.

Under the condition in Assumption 8, we have, using the Bauer-Fike theorem (Bhatia, 1997, Theorem VIII.3.1), that, wpa 1,

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

and further using that, for invertible matrices A_0 and A

$$\begin{aligned} \|A^{-1}\|_2 &\leq \|A\|_2^{-1} \\ \|A_0^{-1}\|_2 &\leq \|A_0\|_2^{-1} \\ \|A^{-1} - A_0^{-1}\|_2 = \|A^{-1}(A_0 - A)A_0^{-1}\|_2 &\leq \|A^{-1}\|_2 \|(A_0 - A)\|_2 \|A_0^{-1}\|_2 . \end{aligned}$$

We obtain, by applying the triangular inequality:

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

Taking $A_0 = \nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0)$ and $A = \nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0) + r^{TL}$, and using that Assumption 7 implies $\|\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0)\|_2$ is bounded away from zero uniformly in L,[6]

[6] For a positive (semi)definite symmetric matrix, eigenvalues and singular values coincide, thus the spectral

we conclude that:

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

From which we conclude that, wpa 1:

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

The formula above provides an asymptotic linear representation of the L-moment based estimator. We collect this result in the proposition below.

Proposition 2. *Suppose Assumptions 1-8 hold. Then the estimator admits the asymptotic linear representation:*

$$\sqrt{T}(\hat{\theta} - \theta_0) = -(\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0))^{-1} \nabla_{\theta'} h^L(\theta_0)'\Omega^L(\sqrt{T} h^L(\theta_0)) + o_{P^*}(1)\,. \tag{2.4}$$

In the next subsection, we work out an asymptotic approximation to the distribution of the leading term in (2.4). We will consider both a Gaussian approximation and a uniform approximation.

2.4 Asymptotic distribution

Finally, to work out the asymptotic distribution, we will work with approximations to the quantile process $\sqrt{T}(\hat{Q}_Y(\cdot) - Q_Y(\cdot))$. Due to the fact that $L \longrightarrow \infty$, we will require a stronger concept than weak convergence of the quantile process (Assumption 6).[7] We consider two approaches. The first one utilises **a strong approximation concept**. The idea is to construct, in the *same* underlying probability space, a sequence of Brownian bridges that approximates, in the supremum norm, the empirical quantile process. This could then be used for inference based on a Gaussian distribution. The second approach utilises a **Bahadur-Kiefer representation** to the quantile process. The idea is to approximate $\sqrt{T}(\hat{Q}_Y(\cdot) - Q_Y(\cdot))$ to a (transformation) of the empirical process $\sqrt{T}(\hat{F}_Y(Q_Y(\cdot)) - F_Y(Q_Y(\cdot)))$, where $\hat{F}_Y$ is the empirical cummulative distribution function (cdf). Observe that, when F_Y is continuous, $F_Y(Q_Y(u)) = u$, and if in addition F_Y is strictly increasing, one may write $\hat{F}_Y(Q_Y(u)) = \frac{1}{T}\sum_{t=1}^T \mathbb{1}\{Y_t \leq Q_Y(u)\} = \frac{1}{T}\sum_{t=1}^T \mathbb{1}\{F(Y_t) \leq u\} = \frac{1}{T}\sum_{t=1}^T \mathbb{1}\{U_t \leq u\}$, where $U_t := F(Y_t)$ is a Uniform[0,1] random variable. This could be used as a basis for an inferential procedure, at least in the iid case.

norm is bounded below by the smallest eigenvalue.

[7] An alternative, in the iid case, would be to use Cramér-Wold's device directly in equation (2.4) along with a central limit theorem for linear combinations of order statistics (Stigler, 1969, 1974; Norvaiša and Zitikis, 1991). We opt not to pursue this approach, because there does not appear to be many results when generalising to dependent observations (the method of proof uses Hajèk's projections); whereas there is now an extensive literature on strong approximations and Bahadur-Kiefer representations in dependent contexts (e.g. Wu and Zhou (2011)).

2.4.1 Strong approximation to a Gaussian process

In this subsection, we consider a strong approximation to a Gaussian process. We state below a classical result, in the iid context, due to Csorgo and Revesz (1978):

Theorem 1 (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). Then, if the underlying probability space is rich enough, one can define, for each $t \in \mathbb{N}$, a Brownian bridge $\{B_t(u) : u \in [0, 1]\}$ such that, if $\gamma < 2$:

$$\sup_{0<u<1} |\sqrt{T} f(Q_Y(u))(\hat{Q}_Y(u) - Q_Y(u)) - B_T(u)| \stackrel{a.s.}{=} O(T^{-1/2}\log(T))\,, \tag{2.5}$$

and, if $\gamma \geq 2$

$$\sup_{0<u<1} |\sqrt{T} f(Q_Y(u))(\hat{Q}_Y(u) - Q_Y(u)) - B_T(u)| \stackrel{a.s.}{=} O(T^{-1/2}(\log\log T)^{\gamma}(\log T)^{\frac{(1+\epsilon)}{(\gamma-1)}})\,, \tag{2.6}$$

for arbitrary $\epsilon > 0$.

The above theorem is much stronger than the weak convergence of Assumption 6. Indeed, Theorem 1 requires variables to be defined in the same probability space and yields explicit bounds in the sup norm; whereas weak convergence is solely a statement on the convergence of integrals (Vaart and Wellner, 1996). Suppose the approximation (2.5)/(2.6) holds in our context. Let B_T be as in the statement of the theorem, and assume in addition that $\int_{\underline{p}}^{\bar{p}} \frac{1}{f_Y(Q_Y(u))^2} du < \infty$. A simple application of Bessel's inequality then shows that:

$$\sqrt{T}(\hat{\theta} - \theta_0) = -(\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0))^{-1} \nabla_{\theta'} h^L(\theta_0)'\Omega^L \left[\int_{\underline{p}}^{\bar{p}} \frac{B_T(u)}{f_Y(Q_Y(u))} \mathbf{P}^L(u) du\right] + o_{P^*}(1)\,. \tag{2.7}$$

Note that the distribution of the leading term in the right-hand side is known (by Riemann integration, it is Gaussian) up to θ_0. This representation could thus be used as a basis for inference. The validity of such approach can be justified by verifying that the Kolmogorov distance between the distribution of $\sqrt{T}(\hat{\theta} - \theta_0)$ and that of the leading term of the representation goes to zero as T and L increase. We show that this indeed is true later in this section, where bounds in the Kolmogorov distance are obtained as a byproduct of weak convergence.

Next, we reproduce a strong approximation result in the context of dependent observations. The result is due to Fotopoulos and Ahn (1994) and Yu (1996).

Theorem 2 (Fotopoulos and Ahn (1994); Yu (1996))**.** *Let $Y_1, Y_2 \dots Y_T$ be a strictly stationary, α-mixing sequence of random variables, with mixing coefficient satisfying $\alpha(t) = O(t^{-8})$. Let F denote the distribution function of Y_1. Suppose the following* ***Csorgo and Revesz conditions*** *hold:*

a. F is twice differentiable on (a, b), where $-\infty \le a = \sup\{z : F(z) = 0\}$ and $b = \inf\{z : F(z) = 1\} \le \infty$;

b. $\sup_{0<s<1} |f'(Q_Y(s))| < \infty$;

as well as the condition:

c. $\inf_{0<s<1} f(Q_Y(s)) > 0$

Let $\Gamma(s, t) := \mathbb{E}[g_1(s)g_1(t)] + \sum_{n=2}^{\infty}\{\mathbb{E}[g_1(s)g_n(t)] + \mathbb{E}[g_1(t)g_n(s)]\}$, where $g_n(u) := \mathbb{1}\{U_n \le u\} - u$ and $U_n := F(Y_n)$. Then, if the probability space is rich enough, there exists a sequence of Brownian bridges $\{\tilde{B}_n : n \in \mathbb{N}\}$ with covariance kernel Γ and a positive constant $\lambda > 0$ such that:

$$\sup_{0<u<1} |\sqrt{T}(\hat{Q}_Y(u) - Q_Y(u)) - f(Q_Y(u))^{-1}\tilde{B}_T(u)| \overset{a.s.}{=} O((\log T)^{-\lambda}) . \tag{2.8}$$

A similar argument as the previous one then shows that, under the conditions of the theorem above:

$$\sqrt{T}(\hat{\theta} - \theta_0) = -(\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0))^{-1}\nabla_{\theta'} h^L(\theta_0)'\Omega^L \left[\int_{\underline{p}}^{\bar{p}} \frac{\tilde{B}_T(u)}{f_Y(Q_Y(u))}\mathbf{P}^L(u)du\right] + o_{P^*}(1) . \tag{2.9}$$

Differently from the iid case, the distribution of the leading term on the right-hand side is now known up to θ_0 **and the covariance kernel** Γ. The latter could be estimated with a Newey and West (1987) style estimator.

To conclude the discussion, we note that the strong representation (2.7) (resp. (2.9)) allows us to establish asymptotic normality of our estimator. Indeed, let L_T be the leading term of the representation on the right-hand side of (2.7) (resp. (2.9)), and $V_{T,L}$ be its variance. Observe that $V_{T,L}^{-1/2}L_T$ is distributed according to a multivariate standard normal. It then follows by Slutsky's theorem that $V_{T,L}^{-1/2}\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbb{I}_d)$. Since pointwise convergence of cdfs to a continuous cdf implies uniform convergence (Parzen, 1960, page 438), and given that $V_{T,L}^{-1/2}$ is positive definite, we obtain that:

$$\lim_{T\to\infty} \sup_{c\in\mathbb{R}^d} |P[\sqrt{T}(\hat{\theta} - \theta_0) \le c] - P[L_T \le c]| = 0 , \tag{2.10}$$

which justifies our approach to inference based on the distribution of the leading term on the right-hand side of (2.7) (resp. (2.9)).

We collect the main results in this subsection under the corollary below.

Corollary 1. *Suppose Assumptions 1-8 hold. Moreover, suppose a strong approximation condition such as (2.5)/(2.6) or (2.8) is valid; and, in addition, that* $\int_{\underline{p}}^{\bar{p}} \frac{1}{f_Y(Q_Y(u))^2} du < \infty$*. Then the approximation (2.7) (resp. (2.9)) holds. Moreover, we have that* $V_{T,L}^{-1/2}\sqrt{T}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0, \mathbb{I}_d)$ *and that (2.10) holds.*

Comment 1 (Optimal choice of weighting matrix under Gaussian approximation)**.** Under (2.7), the optimal choice of weights that minimises the variance of the leading term is:

$$\Omega_L^* = \mathbb{E}\left[\left(\int_{\underline{p}}^{\bar{p}} \frac{B_T(u)}{f_Y(Q_y(u))}\mathbf{P}^L(u)du\right)\left(\int_{\underline{p}}^{\bar{p}} \frac{B_T(u)}{f_Y(Q_y(U))}\mathbf{P}^L(u)du\right)'\right]^-, \qquad (2.11)$$

where A^- denotes the generalised inverse of a matrix A. This weight can be estimated using a preliminary estimator for θ_0. A similar result holds under (2.9), though in this case one also needs an estimator for the covariance kernel Γ. In Appendix C, we provide an estimator for Ω_L in the iid case when Σ_L are shifted Legendre Polynomials. ▸

Comment 2 (A test statistic for overidentifying restrictions)**.** The strong approximation discussed in this subsection motivates a test statistic for overidentifying restrictions. Suppose $L > d$. We can consider the test-statistic:

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

An analogous statistic exists in the overidentified GMM setting (Newey and McFadden, 1994; Wooldridge, 2010). Under the null that the model is correctly specified (i.e. that there exists $\theta \in \Theta$ such that $Q_Y(\cdot) = Q_Y(\cdot|\theta)$), we can use the results in this section to compute the distribution of this test statistic. Indeed, we observe that, under the null,

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

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 (2.11) 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 $L - d$ degrees of freedom. This approximation will be valid, provided L grows sufficiently slowly so that the chi-squared approximation does not concentrate too fast. To see this, let $\tilde{J}$ denote the distribution of the leading term in the representation above, with $\sqrt{T}h_L(\theta_0)$ replaced by the Gaussian approximating random variable. Let $e_{TL} = J - \tilde{J}$ be the approximation error. A simple argument then shows that, for each $c \in \mathbb{R}$:

$$\mathbb{P}[J \leq c] - \mathbb{P}[\tilde{J} \leq c] \leq \mathbb{P}[c \leq \tilde{J} \leq c + \epsilon] + \mathbb{P}[|e_{TL}| > \epsilon]\,.$$

Using that the distribution of $\tilde{J}$ is known, assuming $k := L - d > 2$, and optimising, one obtains that:

$$\sup_{c\in\mathbb{R}} \left[\mathbb{P}[J \leq c] - \mathbb{P}[\tilde{J} \leq c]\right] \leq \left[D_k\left(\frac{\epsilon e^{\epsilon/(k-2)}}{e^{\epsilon/(k-2)} - 1}\right) - D_k\left(\frac{\epsilon}{e^{\epsilon/(k-2)} - 1}\right)\right] + \mathbb{P}[|e_{TL}| > \epsilon]\,,$$

where D_k is the cdf of a χ^2_k. By the mean-value theorem, we may be bound this term by:

$$\left[D_k\left(\frac{\epsilon e^{\epsilon/(k-2)}}{e^{\epsilon/(k-2)} - 1}\right) - D_k\left(\frac{\epsilon}{e^{\epsilon/(k-2)} - 1}\right)\right] + \mathbb{P}[|e_{TL}| > \epsilon] \leq \frac{\epsilon^{k/2} e^{\epsilon/2}}{(e^{\epsilon/(k-2)} - 1)^{k/2-1} 2^{k/2} \Gamma(k/2)} \exp(-0.5\epsilon/(e^{\epsilon/(k-2)} - 1)) + \mathbb{P}[|e_{TL}| > \epsilon]\,.$$

The approximation error e_{TL} consists of two parts: the error due to linearisation, and the error due to approximating $\sqrt{T} h_L(\theta_0)$ by a Gaussian random variable. In the next chapter, we provide conditions that ensure the first error is $O_P(T^{-1/2})$. The rate of the second error depends on the dependence between observations and the assumptions on the distribution: Theorems 1 and 2 provide rates in the iid and strongly mixing settings. Let b_T denote the rate of the second type of error, and $c_T := T^{-1/2} \vee b_T$. By setting $\epsilon = c_T^{\alpha}$ for some $\alpha < 1$, and restricting the growth of L to ensure that, under such choice of ϵ,

$$\frac{\epsilon^{k/2} e^{\epsilon/2}}{(e^{\epsilon/(k-2)} - 1)^{k/2-1} 2^{k/2} \Gamma(k/2)} \exp(-0.5\epsilon/(e^{\epsilon/(k-2)} - 1)) \longrightarrow 0\,,$$

we ensure $\sup_{c\in\mathbb{R}} \left[\mathbb{P}[J \leq c] - \mathbb{P}[\tilde{J} \leq c]\right] \longrightarrow 0$. A similar argument yields that $\sup_{c\in\mathbb{R}} -\left[\mathbb{P}[J \leq c] - \mathbb{P}[\tilde{J} \leq c]\right] \longrightarrow 0$, implying that $\sup_{c\in\mathbb{R}} \left|\mathbb{P}[J \leq c] - \mathbb{P}[\tilde{J} \leq c]\right| \longrightarrow 0$ and justifying the validity of the chi-squared approximation. ▸

Comment 3 (Sample-size-dependent trimming)**.** It is possible to adapt our assumptions and results to the case where the trimming constants $\underline{p}, \overline{p}$ are deterministic functions of the sample size and, as $T \longrightarrow \infty$, $\underline{p} \longrightarrow 0$ and $\overline{p} \longrightarrow 1$. In particular, Theorem 6 of Csorgo and Revesz (1978) provide uniform strong approximation results for sample quantiles ranging from $[1 - \delta_T, \delta_T]$, where $\delta_T = 25n^{-1} \log\log n$. This could be used as the basis for an inferential theory on a variable-trimming L-moment estimator under weaker assumptions than those in this section. Since, in practice, one would have to provide methods to select the constants $\underline{p}, \overline{p}$ **along with** L, and this is not the focus of this work, we choose not to analyse this extension. We note, however, that, in an asymptotic framework where $\underline{p}$ and $\overline{p}$ are kept fixed, a data-driven method for selecting these constants, for a given L, consists in choosing them so as to minimise an estimate of the variance of the leading term (2.4).[8] ▸

Comment 4 (Inference based on the weighted bootstrap)**.** In Appendix D, we show how one can leverage the strong approximations discussed in this section to conduct inference on the model parameters using the weighted bootstrap. ▸

2.4.2 Bahadur-Kiefer representation

Next, we consider a Bahadur-Kiefer representation. We first state the result of Kiefer, in the iid context, as extended by Csorgo and Revesz (1978).

Theorem 3 (Bahadur-Kiefer, Csorgo and Revesz (1978))**.** *Let $Y_1, Y_2 \dots Y_T$ 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

$$\begin{aligned} \sup_{0<u<1} |f(Q_Y(u))\sqrt{T}(\hat{Q}_Y(u) - Q_Y(u)) - \sqrt{T}(\hat{F}_Y(Q_Y(u)) - F(Q_Y(u))| \overset{a.s.}{=} \\ \overset{a.s.}{=} O(T^{-1/4}(\log T)^{1/2}(\log\log T)^{\frac{1}{4}}) \,. \end{aligned} \tag{2.12}$$

As previously discussed, the result above could be used as the basis for an inferential procedure – as well as for the computation of the optimal weights Ω^L. 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 Equation (2.4) and applying Bessel's inequality, we get:

$$\begin{aligned} \sqrt{T}(\hat{\theta} - \theta_0) = \\ = -(\nabla_{\theta'} h^L(\theta_0)'\Omega^L \nabla_{\theta'} h^L(\theta_0))^{-1} \nabla_{\theta'} h^L(\theta_0)'\Omega^L \left[\int_{\underline{p}}^{\bar{p}} \frac{\sqrt{T}(\hat{F}_Y(Q_Y(u)) - F_Y(Q_Y(u))}{f_Y(Q_Y(u))} \mathbf{P}^L(u) du \right] + o_{P^*}(1) \,, \end{aligned} \tag{2.13}$$

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 (2.13) as a basis for an inferential procedure, as in this case there would be dependence between the $U_t := F_Y(Y_t)$

[8] See Athey et al. (2021) for a discussion of this approach in estimating the mean of a symmetric distribution; and Crump et al. (2009) for a related approach that can be used in estimating treatment effects in observational studies.

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:

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

where $V_{T,L}$ is the variance of the leading term of the Bahadur-Kiefer representation, $\mathbb{E}[\xi_t^{L,T}] = 0$ and $\mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_t^{L,T}\right] = 1$. 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 $V_{T,L}^{-1/2}\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \mathbb{I}_d)$.[9] 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 (2.12) goes to zero, analogously to (2.10). This result justifies our papproach to inference.

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

Corollary 2. *Suppose Assumptions 1-8 hold. Moreover, suppose a Bahadur-Kiefer representation such as* (2.12) *is valid; and that* $\int_{\underline{p}}^{\overline{p}} \frac{1}{f_Y(Q_Y(u))^2}du < \infty$. *Then the approximation* (2.13) *holds. In addition, under the conditions of Theorem 3,* $F_Y(Q_Y(u)) = u$ *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,* $V_{T,L}^{-1/2}\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \mathbb{I}_d)$, *where* $V_{T,L}$ *is the variance of the leading term in* (2.13)*; and a bound analogous to* (2.10) *holds.*

Comment 5 (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 (2.11). This is due to the fact that both the Brownian bridge and empirical distribution process share the same covariance kernel. ▸

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

2.5 Asymptotic efficiency

In this section, 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 the variance of the leading term of a

[9] 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.

first order representation converges to the inverse of the Fisher information matrix of the parametric model.[10] Unless stated otherwise, we work under Assumptions 1-8 and those of Corollary 2. To proceed with the analysis, we introduce the alternative estimator:

$$\tilde{\theta}_T \in \operatorname{argmin}_{\theta\in\Theta} \sum_{i\in\mathcal{G}_T}\sum_{j\in\mathcal{G}_T}(\hat{Q}_Y(i) - Q_Y(i|\theta))\kappa_{i,j}(\hat{Q}_Y(j) - Q_Y(j|\theta)), \tag{2.14}$$

for a grid of G_T points $\mathcal{G}_T = \{g_1, g_2, \dots, g_{G_T}\} \subseteq (0,1)$ 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 previous sections,[11] the estimator in (2.14) admits the following asymptotic linear representation as $T \longrightarrow \infty$ and $G_T \longrightarrow \infty$ at a rate:

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = -(\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\partial Q_{G_T})^{-1}\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\sqrt{T}Q_{G_T} + o_p(1), \tag{2.15}$$

where $Q_{G_T} = Q_{G_T}(\theta_0) = (\hat{Q}_Y(g_1) - Q_Y(g_1|\theta_0), \dots, \hat{Q}_Y(g_T) - Q_Y(g_T|\theta_0))'$; ∂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, we arrive at:

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = -(\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\partial Q_{G_T})^{-1}\partial Q'_{G_T}\boldsymbol{\kappa}_{G_T}\left[\boldsymbol{f}^{-1} * \sqrt{T}F_{G_T}\right] + o_p(1), \tag{2.16}$$

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

For a given $\mathcal{G}_T$, representation (2.16) 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 (2.16) under such 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 $\left\{\frac{1}{G_T+1}, \frac{2}{G_T+1}, \dots, \frac{G_T}{G_T+1}\right\}$, it follows from Lemma C.1. in Firpo et al. (2022) that:

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

where

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

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

---

[10] In the nonidependent case, "efficiency" should be defined as achieving the efficiency bound of the parametric model (Newey, 1990), where we model the marginal distribution of the Y_t, but leave the time series dependence unrestricted (except for regularity conditions). Indeed, in general, our L-moment estimator will be inefficient with respect to the MLE estimator that models the dependency structure between observations. See Carrasco and Florens (2014) for further discussion.

[11] We omit these conditions for brevity, but we note that, using the notation in (2.15), 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.

$$
\begin{aligned}
\left(\mathbb{V}^{*-1}\right)_{d_1,d_2} &= \\
(G_T+1)\sum_{i=2}^{G_T} f_Y(Q_Y(g_i))\partial_{d_1}Q_Y(g_i|\theta_0)&\left[f_Y(Q_Y(g_i)\partial_{d_2}Q_Y(g_i|\theta_0) - f_Y(Q_Y(g_{i-1})\partial_{d_2}Q_Y(g_i|\theta_0)\right] \\
-(G_T+1)\sum_{i=1}^{G_T-1} f_Y(Q_Y(g_i))\partial_{d_1}Q_Y(g_i|\theta_0)&\left[f_Y(Q_Y(g_{i+1})\partial_{d_2}Q_Y(g_i|\theta_0) - f_Y(Q_Y(g_i)\partial_{d_2}Q_Y(g_i|\theta_0)\right] \\
&+(G_T+1)(f_Y(Q_Y(g_1)))^2\partial_{d_1}Q_Y(g_1|\theta_0)\partial_{d_2}Q_Y(g_1|\theta_0) \\
&+(G_T+1)(f_Y(Q_Y(g_{G_T})))^2\partial_{d_1}Q_Y(g_{G_T}|\theta_0)\partial_{d_2}Q_Y(g_{G_T}|\theta_0)\,.
\end{aligned}
\tag{2.17}
$$

Assuming the tail condition:[12]

$$
\lim_{u\to 0}\frac{(f_Y(Q_Y(u)))^2\partial_{d_1}Q_Y(u|\theta_0)\partial_{d_2}Q_Y(u|\theta_0) + (f_Y(Q_Y(1-u)))^2\partial_{d_1}Q_Y(1-u|\theta_0)\partial_{d_2}Q_Y(1-u|\theta_0)}{u} = 0\,, \tag{2.18}
$$

leads to the last term of (2.17) being asymptotically negligible as $T \longrightarrow \infty$. 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:

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

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

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

where $I(\theta) = \mathbb{E}[\nabla_\theta\log(f(Y|\theta))\nabla_{\theta'}\log(f(Y|\theta))]$ is the Fisher information matrix. In words, a version of the estimator (2.19) under optimal weights and an appropriate grid is asymptotically efficient (in the sense defined in the beginning of this section). We summarise the discussion in the lemma below:

Lemma 2. *Consider the estimator* (2.19)*. Assume that representation* (2.16) *holds. If, in addition, the tail condition* (2.18) *and the uniform differentiability condition in the text holds; then the estimator is asymptotically efficient, in the sense that the variance of the leading term in* (2.16) *converges to the inverse of the Fisher information matrix as* $T \longrightarrow \infty$.

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

How does the previous estimator relate to our L-moment estimator? Notice that, if the $\{P_l\}_{l\in\mathbb{N}}$ are orthornormal **bases** 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 $\{Q_Y(\cdot|\theta) \, : \, \theta \in \Theta_0\} \subseteq L^2[0,1]$,[13] 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 an 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 L L-moments and weights $\tilde{\boldsymbol{\kappa}}_L = (\tilde{\kappa}_{i,j})_{i,j=1,\ldots L}$. Denote the estimator by $\check{\theta}_T$, and its objective function by $A^L_T(\theta)$. It can be shown that, for an identifiable parametric family and $\|\tilde{\boldsymbol{\kappa}}\|_2 = O(1)$, $\nabla_\theta A^L_T(\check{\theta}_T) = \nabla_\theta A^{\infty}_T(\check{\theta}_T) + o_{p^*}(T^{-1/2})$. This shows that the estimator admits the same first order representation as (2.16); and from the previous lemma we know the variance of the leading term of this representation converges to $I(\theta_0)^{-1}$. Using the last part of Corollary 2, we can then show that the variance of the leading term of representation (2.13) of $\check{\theta}_T$ will converge to the same limit.[14] But then, since the optimal weights (2.11) minimise the variance of the leading term in (2.13) (recall Comment 5), we conclude that they too must, asymptotically, yield a variance equal to $I(\theta_0)^{-1}$. This shows that the L-moment estimator is efficient, in the sense that the variance of the leading term in (2.13) under optimal weights converges to $I(\theta_0)^{-1}$.

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

Corollary 3. *Suppose the conditions of the previous lemma hold. Suppose the $\{P_l\}_{l\in\mathbb{N}}$ are orthonormal **bases**. Consider the estimator $\check{\theta}_T$ defined in the main text. Suppose that Assumptions 1-8 and those of Corollary 2 hold with $\Omega_L = \Sigma_L = \boldsymbol{\kappa}_L$. We then have that, for an identifiable parametric family:*

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

where $V^*_{T,L}$ is the variance of the leading term of (2.13) under optimal weights, i.e. $V^*_{T,L} = (\nabla_{\theta'} h^L(\theta_0)'\Omega^*_L \nabla_{\theta'} h^L(\theta_0))^{-1}$, where Ω^*_L is given by (2.11).

Comment 7 (Related estimators)**.** Similarly to (2.14), we can show that estimators based on minimising the objective functions:

[13] This is implied by Assumption 4.

[14] Let $\mathbb{V}_T$ denote the variance of the leading term of representation (2.13) of the estimator $\check{\theta}$. By weak convergence (the last part of Corollary 2) and Fatou's lemma, it follows that $\liminf_{T\to\infty} \xi'(\mathbb{V}_T^{-1/2} M_T \mathbb{V}_T - \mathbb{I}_d)\xi \geq 0$ for any $\xi \in \mathbb{R}^d$, where $\lim_T M_T = I(\theta_0)^{-1}$. It then follows that $\lim_{T\to\infty} \mathbb{V}_T = I(\theta_0)^{-1}$.

$$
\begin{aligned}
W^1(\theta) &:= \int_0^1 w(u)(\hat{Q}_Y(u) - Q_Y(u|\theta))^2 du \,, \\
W^2(\theta) &:= \int_0^1 \int_0^1 (\hat{Q}_Y(u) - Q_Y(u|\theta)) w(u,v)(\hat{Q}_Y(v) - Q_Y(v|\theta)) dv du \,,
\end{aligned}
\tag{2.19}
$$

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 method of L-moments estimator under optimal weights will be at least as efficient as estimators based on minimising (2.19). 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. This is indeed verified in the simulations in the next section. ▸

2.6 Monte Carlo exercise

In our exercise, we draw random samples $Y_1, Y_2, \ldots, Y_T$ from a distribution function $F = F_{\theta_0}$ belonging to a parametric family $\{F_\theta : \theta \in \Theta\}$. Following Hosking (1990), we consider the goal of the researcher to be estimating quantiles $Q_Y(\tau)$ of the distribution F_{θ_0} by using a plugin approach: first, the researcher estimates θ_0; then she estimates $Q_Y(\tau)$ by setting $\widehat{Q_Y(\tau)} = Q_Y(\tau|\hat{\theta})$. As in Hosking (1990), we consider $\tau \in \{0.9, 0.99, 0.999\}$. In order to compare the behaviour of alternative procedures in estimating more central quantiles, we also consider the median $\tau = 0.5$. We analyse sample sizes $T \in \{50, 100, 500\}$.

We compare the root mean squared error of four types of method of L-moment estimators under varying choices of L with the root mean squared error obtained were θ_0 to be estimated via MLE. We consider the following estimators: (i) the method of L-moments estimator that uses the càglàd L-moment estimates (1.3) and identity weights (**Càglàd FS**);[15] (ii) a two step-estimator which first estimates (i) and then uses this preliminary estimator[16] to estimate the optimal weighting matrix (2.11), which is then used to reestimate θ_0 (**Càglàd TS**); (iii) the method of L-moments estimator that uses the unbiased L-moment estimates (1.4) and identity weights (**Unbiased FS**); and (iv) the two-step estimator that uses the unbiased L-moment estimator in the first and second steps (**Unbiased TS**). The estimator of the optimal-weighting matrix we use is given in Appendix C.

---

[15] To be precise, our choice of weights does not coincide with actual identity weights. Given that the coefficients of Legendre polynomials rapidly scale with L – and that this increase generates convergence problems in the numerical optimisation – we work directly with the underlying estimators of the probability-weighted moments $\int_0^1 Q_Y(u)U^r du$ (Landwehr et al., 1979), of which L-moment estimators are linear combinations. When (estimated) optimal weights are used, such approach is without loss, since the optimal weights for L-moments constitute a mere rotation of the optimal weights for probability-weighted moments. In other cases, however, this is not the case: a choice of identity weights when probability-weighted moments are directly targeted coincides with using $\boldsymbol{D}^{-1'}\boldsymbol{D}^{-1}$ as a weighting matrix for L-moments, where $\boldsymbol{D}$ is a matrix which translates the first L probability-weighted moments onto the first L L-moments. For small L, we have experimented with using the "true" L-moment estimator with identity weights, and have obtained the same patterns presented in the text.

[16] This preliminary estimator is computed with $L = d$.

2.6.1 Generalized Extreme value distribution (GEV)

Following Hosking et al. (1985) and Hosking (1990), we consider the family of distributions

$$F_\theta(z) = \begin{cases} \exp\{-[1-\theta_2(x-\theta_1)/\theta_3]^{1/\theta_3}\}, & \theta_3 \neq 0 \\ \exp\{-\exp(-(x-\theta_1)/\theta_2)\}, & \theta_3 = 0 \end{cases},$$

and $\theta_0 = (0, 1, -0.2)'$.

Table 2.1 reports the RMSE of each procedure, divided by the RMSE of the MLE, under the choice of L that achieves the smallest RMSE. Values above 1 indicate the MLE outperforms the estimator in consideration; and values below 1 indicate the estimator outperforms MLE. The value of L that minimises the RMSE is presented under parentheses. Some patterns are worth highlighting. Firstly, the L-moment estimator, under a proper choice of L and (estimated) optimal weights (two-step estimators) is able to outperform MLE in most settings, especially at the tail of the distribution function. Reductions in these settings can be as large as 27%. At the median, two-step L-moment estimators behave similarly to the MLE. The performance of two-step càglàd and unbiased estimators is also quite similar. Secondly, the power of overidentifying restrictions is evident: except in three cases, two-step L-moment estimators never achieve a minimum RMSE at $L = 3$, the number of parameters. The relation between N and L in the two-step Càglàd estimator is also monotonic, except at $\tau = 0.5$. Finally, the role of optimal weights is clear: first step estimators tend to underperform the MLE as the sample size increases. In larger samples, and when optimal weights are not used, the best choice tends to be setting L close to or equal to 3, which reinforces the importance of weighting when overidentifying restrictions are included.

To better understand the patterns in the table, we report in Figure 2.1, the relative RMSE curve for different sample sizes and choices of L. The role of optimal weights is especially striking: first-step estimators usually exhibit an increasing RMSE, as a function of L. In contrast, two-step estimators are able to better control the RMSE across L. It is also interesting to note that the two-step unbiased L-moment estimator behaves poorly when L is close to T. This is related to the result in Lemma 1, which shows that the unbiased L-moment estimator is first-order equivalent to the càglàd estimator when L is small relative to T. When L is large, such equivalence is not guaranteed, and the optimal weighting scheme given by (2.11) need not be the best choice for the unbiased estimator.

2.6.2 Generalized Pareto distribution (GPD)

Following Hosking and Wallis (1987), we consider the family of distributions:

$$F_\theta(z) = \begin{cases} 1-(1-\theta_2 x/\theta_1)^{-1/\theta_2}, & \theta_2 \neq 0 \\ 1-\exp(-x/\theta_1), & \theta_2 = 0 \end{cases},$$

and $\theta_0 = (1, -0.2)'$.

Table 2.2 and Figure 2.2 summarise the results of our simulation. Overall patterns are similar to the ones obtained in the GEV simulations. Importantly, though, estimation of the optimal weighting matrix appears to impact two-step estimators quite negatively in this setup, which leads to the choice of $L = 2$ (i.e. a just-identified estimator) being optimal for TS estimators at most quantiles when $T < 500$. This phenomenon also leads to FS estimators, which do not use estimated weights, being able to perform somewhat better than TS estimators when $T < 500$. When $T = 500$, estimation error of the weighting matrix becomes secondary and TS estimators slightly dominate first-step ones. In all settings, L-moment estimators compare favourably to the MLE.

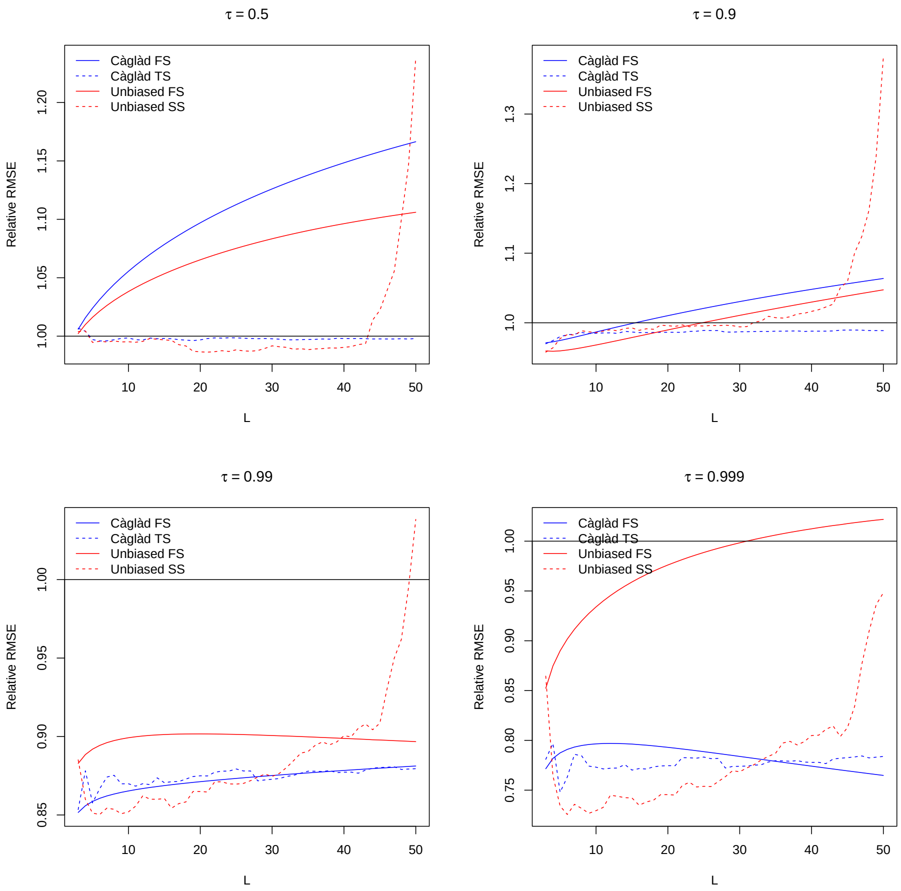
Table 2.1: *GEV : relative RMSE under MSE-minimising choice of L*

	$T = 50$				$T = 100$				$T = 500$			
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$
Càglàd FS	1.006	0.971	0.852	0.765	1.02	0.975	0.911	0.864	1.05	0.999	1.046	1.083
	(3)	(3)	(3)	(50)	(3)	(3)	(3)	(3)	(3)	(5)	(3)	(3)
Càglàd TS	0.996	0.970	0.853	0.748	1.01	0.974	0.908	0.844	1.00	0.995	0.978	0.968
	(6)	(3)	(3)	(5)	(11)	(3)	(11)	(5)	(8)	(97)	(89)	(89)
Unbiased FS	1.002	0.959	0.883	0.852	1.02	0.970	0.930	0.910	1.04	0.998	1.052	1.096
	(3)	(4)	(3)	(3)	(3)	(4)	(3)	(3)	(3)	(6)	(3)	(3)
Unbiased SS	0.986	0.958	0.850	0.726	1.00	0.969	0.897	0.827	1.00	0.993	0.975	0.964
	(21)	(3)	(6)	(6)	(18)	(3)	(18)	(18)	(8)	(79)	(86)	(62)

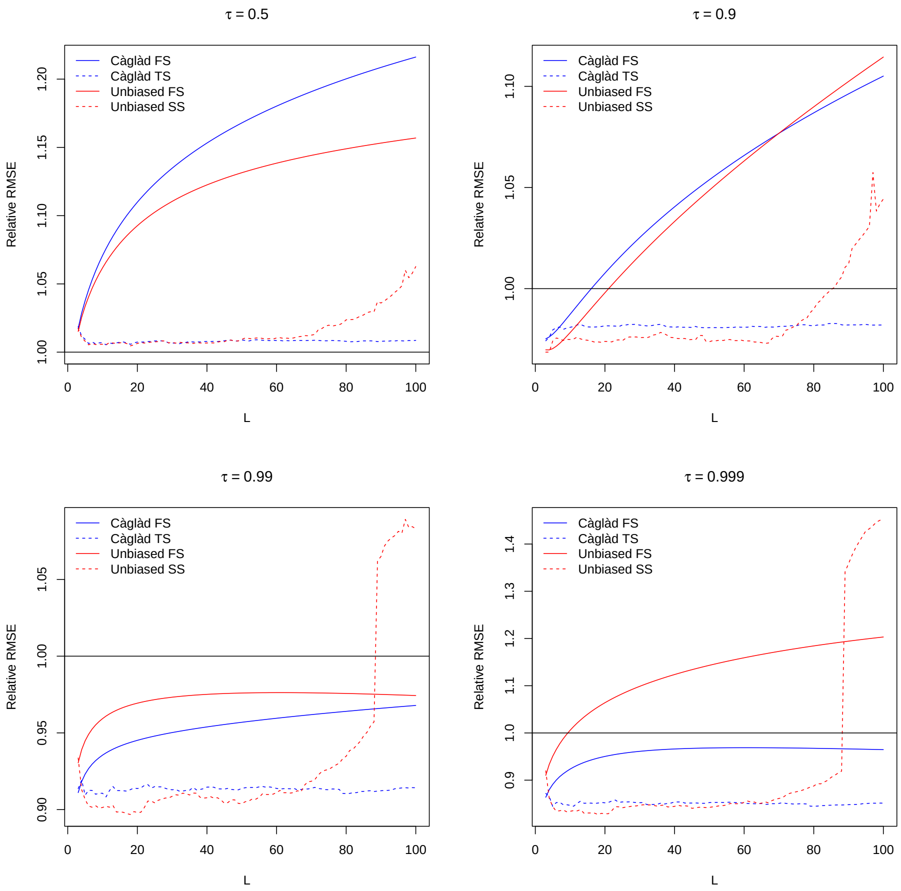
Table 2.2: *GPD : relative RMSE under MSE-minimising choice of L*

	$T = 50$				$T = 100$				$T = 500$			
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.99$	$\tau = 0.999$
Càglàd FS	0.977	0.979	0.797	0.599	0.984	0.986	0.883	0.804	0.996	1.000	0.988	0.985
	(5)	(3)	(50)	(50)	(5)	(3)	(14)	(100)	(3)	(8)	(3)	(3)
Càglàd TS	0.963	0.979	0.818	0.660	0.976	0.986	0.899	0.824	0.996	0.996	0.974	0.962
	(5)	(2)	(3)	(2)	(3)	(2)	(2)	(2)	(11)	(97)	(96)	(100)
Unbiased FS	0.954	0.965	0.828	0.728	0.971	0.979	0.901	0.862	0.993	0.998	0.993	0.999
	(5)	(4)	(12)	(5)	(6)	(4)	(8)	(5)	(3)	(16)	(3)	(2)
Unbiased SS	0.941	0.965	0.839	0.734	0.965	0.980	0.912	0.874	0.989	0.994	0.973	0.970
	(4)	(2)	(2)	(2)	(6)	(2)	(2)	(2)	(58)	(85)	(81)	(81)

(a) $T = 50$


τ = 0.5
τ = 0.9
τ = 0.99
τ = 0.999
Càglàd FS
Càglàd TS
Unbiased FS
Unbiased SS
Relative RMSE
L

(b) $T = 100$


τ = 0.5
τ = 0.9
τ = 0.99
τ = 0.999
Càglàd FS
Càglàd TS
Unbiased FS
Unbiased SS
Relative RMSE
L

(c) $T = 500$

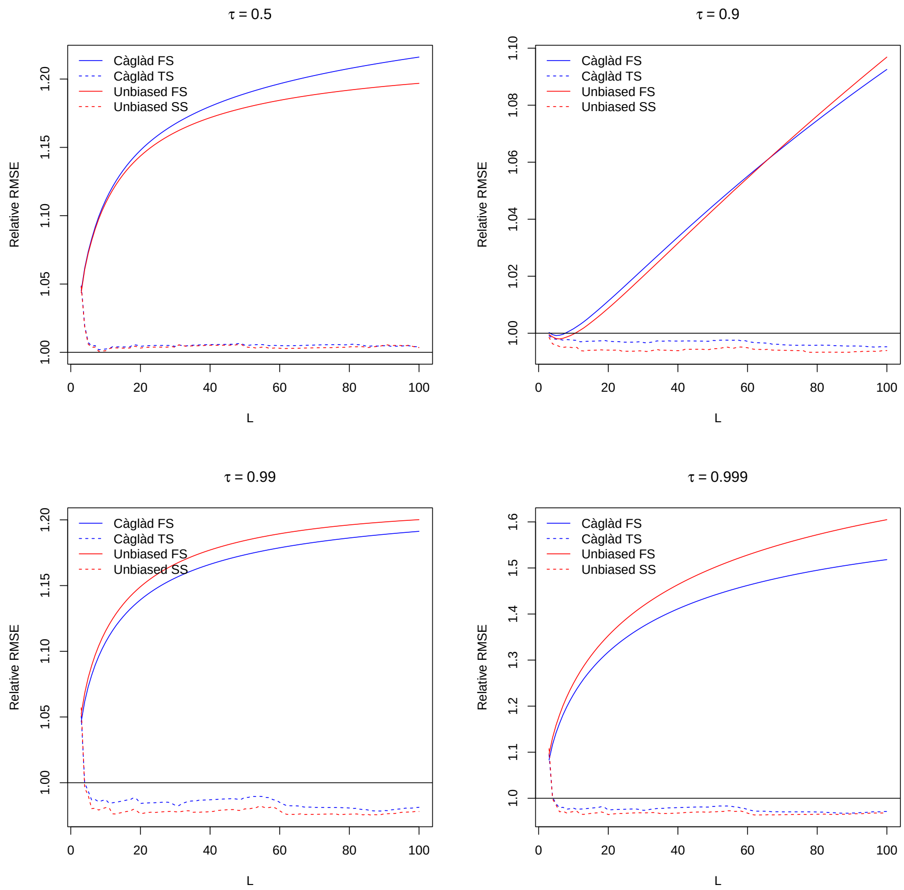
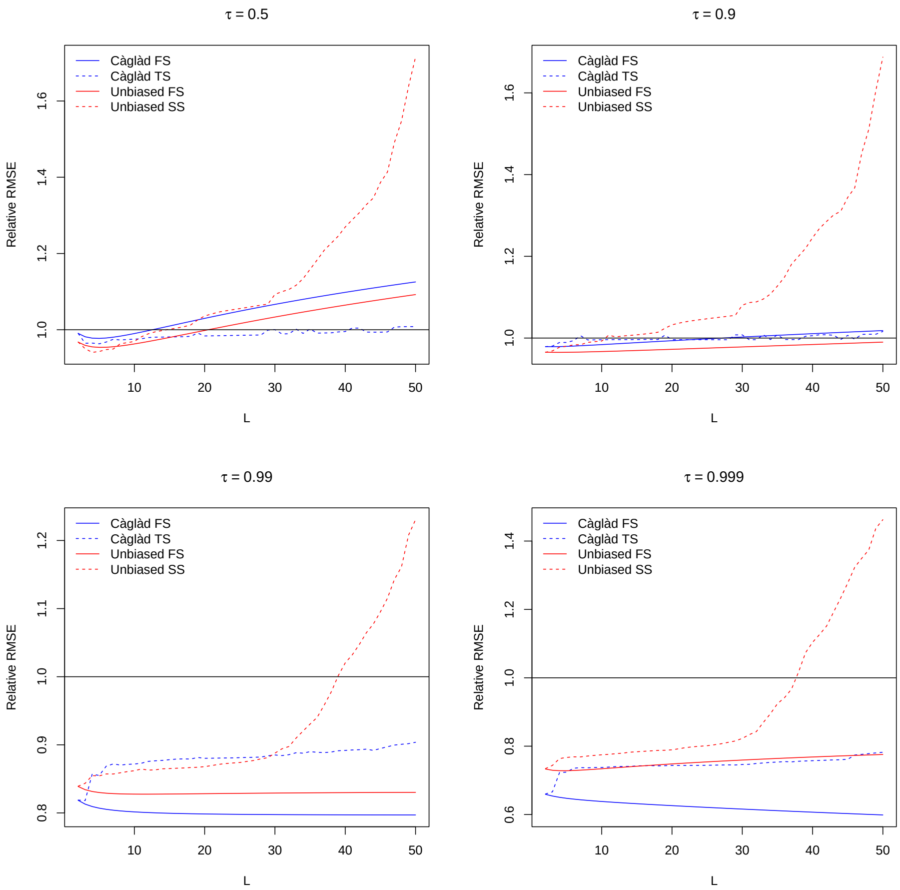
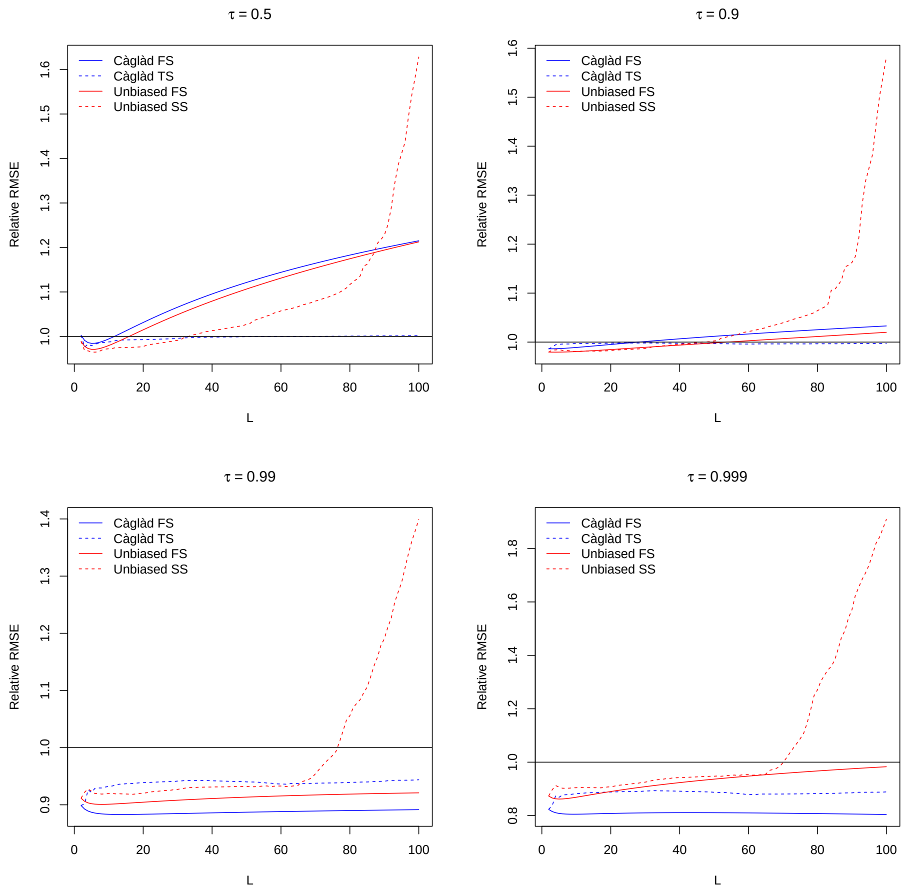


Figure 2.1: *GEV: relative RMSE for different choices of L.*

(a) T = 50


τ = 0.5
τ = 0.9
τ = 0.99
τ = 0.999
Càglàd FS
Càglàd TS
Unbiased FS
Unbiased SS
Relative RMSE
L

(b) $T = 100$


τ = 0.5
τ = 0.9
τ = 0.99
τ = 0.999
Relative RMSE
L
0
20
40
60
80
100
1.0
1.1
1.2
1.3
1.4
1.5
1.6
0.8
0.9
1.8
Càglàd FS
Càglàd TS
Unbiased FS
Unbiased SS

(c) $T = 500$

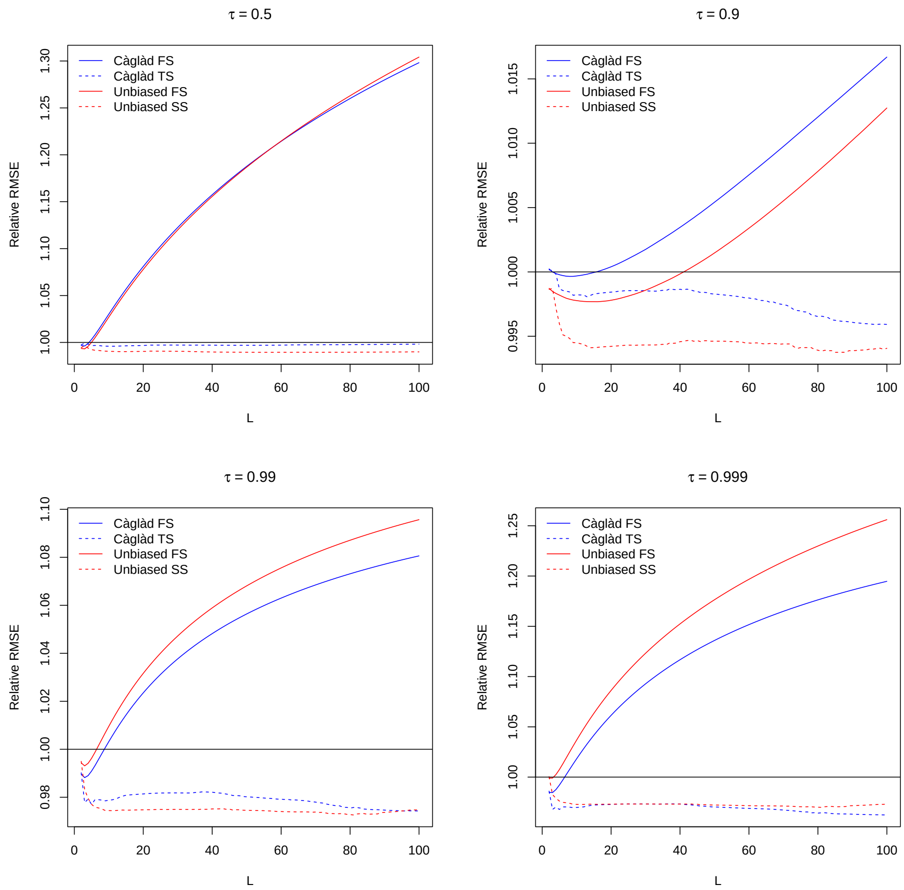


Figure 2.2: *GPD: relative RMSE for different choices of L.*

Chapter 3

Choosing the number of L-moments in estimation

3.1 Overview

This chapter proposes methods to select the number of moments L used in estimation. We consider two approaches. In Section 3.2, we derive a higher-order expansion of the L-moment estimator (2.1). We then propose to choose L by minimising linear combinations of the higher-order mean-squared error obtained from these expansions. A similar approach is considered in the GMM literature by Donald and Newey (2001) – where the goal is to choose the number of instruments in linear instrumental variable models –, and Donald et al. (2009) – where one wishes to choose moment conditions in models defined by conditional moment restrictions (in which case infinitely many restrictions are available). Similarly, Okui (2009) considers the choice of moments in dynamic panel data models; and, more recently, Abadie et al. (2019) use higher order expansions to develop a method of choosing subsamples in linear instrumental variables models with first stage heterogeneity.

In Section 3.3, we consider an approach to selecting L-moments via ℓ_1-regularisation. Following Luo et al. (2015), we note that (2.3) may be written as:

$$A_L h^L(\hat{\theta}) = 0\,,$$

for a $d \times L$ matrix A_L which combines the L-moments linearly into d restrictions. The idea is to estimate A_L using a Lasso penalty. This approach implicitly performs moment selection, as the method yields exact zeros for several entries of A_L. In particular, the quadratic program suggested by Luo et al. (2015) is easily implemented, which constrasts with competing approaches to moment selection using ℓ_1-regularisation in the GMM literature (Cheng and Liao, 2015). We consider using the approach of Luo et al. (2015) adapted to our L-moment setting; and contrast it with the higher-order expansion in the simulation exercise of Section 3.4.

3.2 Higher order expansion of the L-moment estimator

In this section, we derive a higher order expression for the L-moment estimator in Chapter 2. Our goal is to derive a representation of the estimator as follows:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = \Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T} + O_{P^*}(T^{-3/2}), \tag{3.1}$$

for tight sequences of random variables $\Theta_1^T, \Theta_2^T, \Theta_3^T$. Under uniform integrability conditions on Θ_1^T, Θ_2^T, Θ_3^T and the remainder, representation (3.1) allows us to write:

$$\mathbb{E}[T(\hat{\theta}_T - \theta_0)(\hat{\theta}_T - \theta_0)'] = \mathbb{E}\left[\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)\left(\Theta_1^T + \frac{\Theta_2^T}{\sqrt{T}} + \frac{\Theta_3^T}{T}\right)'\right] + O(T^{-3/2}), \tag{3.2}$$

which may be used as a basis for a method of selecting L, provided the expectation on the right-hand side is estimable.[1] The idea would be to choose L 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 (3.1) 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 (3.1) for the L-moment estimation, we assume, in addition to Assumptions 1-8 in the previous chapter, the following conditions.

Assumption 9. *As $T, L \longrightarrow \infty$, $\mathbb{P}[W_L^{-1}$ and Ω_L^{-1} exist$] \longrightarrow 1$. We also assume that, as $T, L \longrightarrow \infty$, $W_L^{-1} = \Omega_L^{-1} + O_p(T^{-1/2})$.*

Assumption 10. *$Q_Y(u|\theta)$ 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 $[\underline{p}, \overline{p}]$, 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 1-10 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 $L \longrightarrow \infty$ and the L-moment structure in our setting.[2]

Proposition 3. *Suppose Assumptions 1-10 in the main text are satisfied. Then (3.1) holds for $\hat{\beta} = \left(\hat{\theta}' \quad -h^L(\hat{\theta})' W^L\right)'$ with the objects as follows:*

---

[1] Even if the uniform integrability conditions that allow us to write (3.2) from (3.1) do not hold, we can posit that our goal is to minimise the MSE of the leading term in (3.1). This is the Nagar (1959) style approach of Rothenberg (1984) and Donald and Newey (2001). We return to this point later on.

[2] 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.

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

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

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

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

Proof. See Appendix E. □

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, $(W^L)^{-1} - (\Omega^L)^{-1}$; (ii) moments of the (recentered) L-moment vector, $\sqrt{T}h^L(\theta_0)$. 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 $\sqrt{T}((W^L)^{-1} - (\Omega^L)^{-1})$, which can then be plugged onto (3.1) to obtain an $O_{P^*}(T^{-3/2})$ expansion in terms of quantities whose moments may be estimated. In particular, if $(\Omega^L)^{-1}$ may be written as a function $M^L(\theta_0)$ (e.g. the optimal weights under a Gaussian approximation) and $(W^L)^{-1} = M^L(\tilde{\theta}_T)$ 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 $M^L(\cdot)$. We state these below:

Lemma 3. *Suppose $(\Omega^L)^{-1} = M^L(\theta_0)$ and that we estimate it by $(W^L)^{-1} = M^L(\tilde{\theta}_T)$, 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^L are three times continuously differentiable on Θ. Let $\partial_{s_1,\ldots,s_k}M^L(\theta)$ be the $L\times L$ matrix with entry (i,j) corresponding to the partial derivative $\partial_{s_1,\ldots s_k}(M^L(\theta))_{i,j}$. Suppose $\sum_{i=1}^d \|\partial_i M^L(\theta_0)\|_2^2 = O(1)$, $\sum_{i=1}^d\sum_{j=1}^d \|\partial_{i,j}M^L(\theta_0)\|_2^2 = O(1)$ and $\sup_{\theta\in\Theta}\sum_{i=1}^d\sum_{j=1}^d\|\partial_{i,j}M^L(\theta)\|_2^2 = O(1)$. Then the estimator satisfies:*

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

where

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

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

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 $\sqrt{T}h^L(\theta_0)$, one may be tempted to use the strong approximations considered in Chapter 2 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., 2021). 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

$$\frac{\|\hat{E}_T + F_T\|_2}{\|\hat{H}_T\|_2} = o_p(1)\,. \tag{3.3}$$

Clearly, the Gaussian approximation to the moments of $\sqrt{T}h^L(\theta_0)$ 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 $\sqrt{T}h^L(\theta_0)$, 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 $\sqrt{T}h^L(\theta_0)$ 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 (e.g. the extension in Chapter 4) either, as in these cases the distribution of the data is not fully specified. Given these limitations, it is important to consider alternative approaches to approximating moments of $\sqrt{T}h^L(\theta_0)$.

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 $\sqrt{T}h^L(\theta_0)$ in terms of estimable terms, which could then be plugged onto (3.1) 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. (2021) have been able to solve this problem by casting quantile estimation as a particular case of quantile regression (Koenker and Bassett, 1978) and working with strong approximations to the objective function. However, due to the nonuniqueness of the minimiser, this approach is only able to deliver an $O_p(T^{-3/4+\gamma})$ remainder for arbitrary $\gamma > 0$. Some of the terms in their bias formula are not estimable either.

A useful "workaround" to the inability of computing tractable higher-order expansion formulae would be to use Phillips (1991)'s informal "shortcut", where one assumes the quantile estimator satisfies a first order condition in terms of the (nonsmooth) subgradient. One then performs a Taylor expansion of this nonsmooth gradient in terms of the Dirac delta function. As explained by Franguridi et al. (2021), this approach abstracts from the multiplicity of solutions in the estimation procedure, as well as from the fact that quantile estimators do not, in general, achieve a subgradient equal to zero. These are two large sources of bias in Franguridi et al.'s analysis. Still, Phillips' heuristic is able to deliver precisely those higher-order bias terms that are estimable – and additional variance-related terms if we keep on expanding. Recently, Lee et al. (2017) (see also Lee et al. (2018)) used the Dirac trick to obtain an $O_p(T^{-3/2})$ expansion of the quantile estimator. We could plug their formula onto (3.1) to obtain a tractable higher-order MSE up until $O_p(T^{-1})$ terms.[3]

Comment 8 (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 that the $\sup_{\theta\in\Theta} \sum_{i=1}^{p} \sum_{j=1}^{p} \|\partial_{i,j} \nabla_{\theta\theta'} g_t(\theta)\|_2 = O(\xi_T)$. A fourth order mean-value expansion then yields:

$$\begin{aligned}\sqrt{T}(g_T(\hat{\theta}_T) - g_T(\theta_0)) = \nabla_{\theta'} g_T(\theta_0)\sqrt{T}(\hat{\theta}_T - \theta_0) + \sqrt{T}(\hat{\theta}_T - \theta_0)'\nabla_{\theta\theta'} g_T(\theta_0)(\hat{\theta}_T - \theta_0) + \\ \sum_{i=1}^{p} \sqrt{T}(\hat{\theta}_{i,T} - \theta_{i,0})(\hat{\theta}_T - \theta_0)\partial_i \nabla_{\theta\theta'} g_T(\theta_0)(\hat{\theta}_T - \theta_0) + \\ \sum_{i=1}^{p}\sum_{j=1}^{p} \sqrt{T}(\hat{\theta}_{i,T} - \theta_{i,0})(\hat{\theta}_{j,T} - \theta_{j,0})(\hat{\theta}_T - \theta_0)\partial_{ij}\nabla_{\theta\theta'} g_T(\tilde{\theta}_T)(\hat{\theta}_T - \theta_0),\end{aligned} \tag{3.4}$$

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:

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

where ψ_T is the order of the sum of the first three terms in the mean-value expansion (3.4). We can then plug (3.2) 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). ▸

3.3 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 introduction, our L-moments estimator can be

[3] To be precise, we should also provide conditions so the Taylor expansion in Lee et al. (2018) is uniform in the quantile being estimated. However, given that their approach is solely an heuristic, we abstract from this problem.

seen as combining the L moments used in estimation into d linear restrictions via the mapping:

$$\hat{A}_L h_L(\hat{\theta}) = 0\,, \tag{3.5}$$

where the combination matrix is estimated as:

$$\hat{A}_L = \nabla_{\theta'} h^L(\hat{\theta})' W^L\,. \tag{3.6}$$

The poor behaviour of the L-moment estimator with large L may be partly attributed to the estimation of (3.6). Indeed, we note that the term $\Theta_T^2/\sqrt{T}$ in the higher order expansion of (3) is closely related to the estimation error of Ω^L and $\nabla_{\theta'} h_L(\theta_0)$; and correlation of these errors with $h^L(\hat{\theta})$ affects the bias due to this term. Suppose $\Xi = (W^L)^{-1}$ exists (as in an estimator of the optimal weighting matrix). Instead of estimating $A_L = \nabla_{\theta'} h^L(\theta_0)'\Omega^L$ by $\hat{A}_L$, Luo et al. (2015) propose to estimate the j-th row of A_l as (adapting their program to our context):

$$\tilde{\lambda}_j \in \operatorname{argmin}_{\lambda \in \mathbb{R}^L} \frac{1}{2}\lambda'\Xi\lambda - \lambda'\nabla_{\theta'} h^L(\tilde{\theta}) e_j + \frac{k}{T}\sum_{l=1}^{L} \nu_l^j \cdot |\lambda_l|\,, \tag{3.7}$$

for penalties $k \geq 0, \nu_l^j \geq 0$, $l = 1, \ldots L$; 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 $\hat{\lambda}_j = \Xi^{-1}\nabla_{\theta'} h^L(\tilde{\theta}) e_j$, which coincides with the j-th row of $\hat{A}_L$. In general, however, the penalties will induce sparsity on the estimated rows, so only a few entries are selected. Importantly, (3.7) can be efficiently estimated by quadratic programming algorithms.[4] The problem is also well-defined even if Ξ, but not Ξ^{-1}, exists.

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

$$\tilde{A}_L h^L(\check{\theta}) = 0\,. \tag{3.8}$$

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 $\tilde{A}_L$, i.e. we use the moments given by the indices $\mathcal{I}_S = \{l \in \{1, 2, \ldots, L\} : \hat{\lambda}_{ij} = 0 \text{ for some } j = 1 \ldots d\}$.

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_L is *approximately sparse* – in the sense that it is well approximated by a sparse matrix at a rate –, then the estimator based on (3.8) is asymptotically efficient under some additional conditions and as $T, L \longrightarrow \infty$. Their result can be adapted to our L-moment context – an extension we pursue in Appendix F. In what follows, we contrast the Lasso approach with the higher-order MSE method in a Monte Carlo exercise.

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

3.4 Monte Carlo exercise

We return to the Monte Carlo exercise in Chapter 2. We consider the behaviour of four estimators: (i) the L-moment estimator with identity weights and $L = d$ (**FS**); (ii) the two-step L-moment estimator with L selected in order to minimise the approximate RMSE of the quantile one wishes to estimate (**TS RMSE**); (iii) the L-moment estimator with optimal weights and Lasso selection (**TS Lasso**); and (iv) the post-Lasso estimator that runs the two-step 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 (1.3). As in Chapter 2, 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 Comment 8, **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 (3.4)), and by working with the expansion of the L-moment estimator up to order T^{-1}. We discard third-order terms from the expansion for computational reasons – doing so allows us to compute the required derivatives automatically using efficient automatic differentiation algorihms.[5] As can be seen from the formulae in comment 8, 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 L. Given the nonlinear nature of the problem, one expects such choices to lead to small values of L. Therefore, in order to somewhat counterbalance the effect of ignoring third order terms terms, we run our selection approach with L ranging from $d+1$ to $T \wedge 100$, i.e. we do not allow the procedure to pick the just-identified choice $L = d$.

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 υ_l^j. Specifically the penalty k follows the rule in Luo et al. (2015). In setting the penalty-specific loading, we observe that Assumption 14 in Appendix F requires that, for each program $j = 1 \ldots d$ and variable $l = 1, \ldots L$, the penalty $\upsilon_l^j k/T$ 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 υ_l^j equal to an estimate of an upper bound to the standard error of the gradient. We estimate this bound by using the Delta-method to compute an approximate variance to the gradient. 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 υ_l^j to be equal to zero for $l = 1, 2, \ldots, d$, i.e. we do not regularise the first d L-moments, so

[5] Specifically, we use R package `autodiffr`, which serves as a wrapper to Julia routines that compute Jacobians and Hessians efficiently using automatic differentiation.

they are effectively "always included" in the second step estimation.

Table 3.1 presents the relative RMSE (vis-à-vis the MLE) of each approach in the GEV Monte Carlo. The average number of selected moments is reported in parentheses (for the FS estimator, this number is always equal to d). The results show that both the SS RMSE and SS Post-Lasso approaches are able to outpeform MLE and the "just-identified" L-moment estimator (FS) in small samples, and, contrary to the FS estimator, still compare favourably to the MLE in larger samples. The SS Lasso performs well at the tails of the distribution in small samples, but performs poorly both in central quantiles and in larger samples. This is due to the large regularisation bias imparted by the "coarse" penalty, which tends to dominate the RMSE in more central quantiles or when the sample size is larger. The Post-Lasso approach is able to attenuate such bias and produce estimators with desirable properties. Importantly, the Post-Lasso is able to compete with the RMSE approach even though it chooses a single set of selected L-moments for all quantiles (recall the Lasso does not directly target the quantiles of interest). This approach also runs much faster than the RMSE approach. All in all, these results lead us to consider the Post-Lasso the best approach in this Monte Carlo exercise, both from a statistical as well as a computational viewpoint.

In Table 3.2, we report similar results for the GPD Monte Carlo. Recall this is a setting where estimation of the weighting matrix bodes ill for two-step estimators in smaller sample sizes, and one-step estimators with identity weights work relatively well. Consequently, the RMSE and post-selection methods are not able to outperform the FS estimator in sample sizes smaller than 500 at several quantiles, though they compare favourably to it. When $T = 500$, both selection methods slightly outperform the FS estimator. We do also note that, similarly to the pattern observed in the GEV Monte Carlo, regularisation biases are large, which leads to the SS Lasso being unappealing. Finally, and as in the GEV Monte Carlo, the Post-Lasso estimator is to be preferred on both computational and statistical grounds.

Table 3.1: *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	1.017	0.962	0.818	0.751	1.030	0.992	0.999	1.005	1.02	1.00	1.08	1.12
	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)
SS RMSE	0.999	0.968	0.806	0.690	1.005	0.972	0.911	0.867	1.01	1.00	1.01	1.02
	(20.492)	(12.13)	(16.116)	(17.01)	(39.394)	(22.85)	(26.84)	(29.636)	(19.07)	(39.212)	(40.296)	(42.994)
SS Lasso	2.373	1.269	0.855	0.776	4.508	2.014	1.958	10.497	4.98	2.02	1.29	1.75
	(7.116)	(7.116)	(7.116)	(7.116)	(7.906)	(7.906)	(7.906)	(7.906)	(8.448)	(8.448)	(8.448)	(8.448)
SS Post-Lasso	1.013	0.962	0.799	0.694	1.016	0.971	0.902	0.844	1.02	1.00	1.02	1.02
	(7.116)	(7.116)	(7.116)	(7.116)	(7.906)	(7.906)	(7.906)	(7.906)	(8.448)	(8.448)	(8.448)	(8.448)

Table 3.2: *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	0.979	0.983	0.792	0.569	1.007	0.996	0.918	0.857	1.019	1.001	0.983	0.982
	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)
SS RMSE	0.968	0.988	0.819	0.604	1.004	0.987	0.952	0.912	1.000	0.995	0.977	0.967
	(10.914)	(8.084)	(9.824)	(11.608)	(80.792)	(10.81)	(99.128)	(99.338)	(99.21)	(98.678)	(99.524)	(99.526)
SS Lasso	> 10	> 10	> 10	> 10	3.860	> 10	> 10	> 10	> 10	> 10	> 10	> 10
	(2.462)	(2.462)	(2.462)	(2.462)	(2.652)	(2.652)	(2.652)	(2.652)	(2.828)	(2.828)	(2.828)	(2.828)
SS Post-Lasso	0.977	0.994	0.840	0.630	1.011	0.981	0.897	0.842	1.013	1.001	0.983	0.980
	(2.462)	(2.462)	(2.462)	(2.462)	(2.652)	(2.652)	(2.652)	(2.652)	(2.828)	(2.828)	(2.828)	(2.828)

Chapter 4

Semiparametric estimation of treatment effects using L-moments

4.1 Setup

In this chapter, we consider an extension of our methodology that can be used in the semiparametric estimation of treatment effects in randomised experiments. Consider a setting where there is an outcome of interest Y and a binary treatment $D \in \{0, 1\}$. For a population of interest, we define the random variables $(Y(0), Y(1))$ as the **potential outcomes** (Imbens and Rubin, 2015), where $Y(d)$ specifies what would occur if a subject randomly drawn from the population is assigned treatment status $D = d$.[1] The distribution of $(Y(0), Y(1))$ reflects the distribution of potential outcomes in the population. The distribution of treatment effects in the population is given by $\tau = Y(1) - Y(0)$; and the average treatment effect is given by $\beta := \mathbb{E}[Y(1) - Y(0)]$. We consider the goal of experimentation to be to conduct inference on β.

In randomised experiments, a random sample of N individuals is drawn from the population; and treatment is assigned randomly to N_1 individuals, independently from their potential outcomes. Denoting by D_i the treatment status indicator of the i-th individual in the sample, we have that the observed outcome is $Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$, where, by the sampling assumption, $(Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} (Y(0), Y(1))$. The researcher observes a resulting sample $\{(D_i, Y_i)\}_{i=1}^N$. The **fundamental problem of causal inference** is that the researcher does not observe both potential outcomes $(Y_i(0), Y_i(1))$ simultaneously, so individual effects $\tau_i = Y_i(1) - Y_i(0)$ are generally unidentifiable. It is possible however, to estimate the average treatment effect by:

[1] Underlying this definition of potential outcome is a non-interference assumption, which postulates that the treatment status of an individual depends only on her assignment, and not on the treatment status of the remaining individuals in the population. The analysis of experimental (and observational) studies under different patterns of interference – where there can be spillovers from one unit's status to the other –, is an active topic of research in Statistics. See Sävje et al. (2021) and references therein.

$$\hat{\beta} = \frac{\sum_{i=1}^{N} D_i Y_i}{N_1} - \frac{\sum_{i=1}^{N}(1 - D_i) Y_i}{N_0}, \tag{4.1}$$

i.e. a comparison of means between treatment and control units. Under random assignment of the treatment, this estimator is unbiased for β, and its variance is given by $\mathbb{V}[Y(0)]/N_0 + \mathbb{V}[Y(1)]/N_1$.

Is $\hat{\beta}$ the "best" we can do in randomised experiments? If the distribution of potential outcomes in the population is left unspecified (except for regularity conditions), then the answer is **yes**, in the sense that the estimator achieves the semiparametric efficiency bound (Newey, 1990) of the problem as $N_1, N_0 \longrightarrow \infty$. However, as argued by Athey et al. (2021), in several contexts, this result is not enough. Especially in experimental settings deployed by big tech companies, treatment effects are small, and yet **economically significant**. These settings are characterised by heavy-tailed distributions of the outcome distribution, which leads to the variance of estimates based on (4.1) being quite large. The combination of small effects and large variances leads to low power in detecting effects based on (4.1), which renders the estimator quite unappealing.[2]

In light of the preceding observations, Athey et al. (2021) propose estimating β using semiparametric methods. Their idea is to leave the distribution (quantile function) of potential outcomes in the absence of treatment, $F_{Y(0)}$ ($Q_{Y(0)}$), unspecified; and to parametrise the distribution in the treatment group as:

$$Q_{Y(1)}(u) = G(Q_{Y(0)}(u); \theta_0), \tag{4.2}$$

where $\theta_0 \in \Theta \subseteq \mathbb{R}^d$. The authors focus on constructing efficient estimators, which achieve the semiparametric efficiency bound of the problem. We propose to analyse their environment through the lenses of L-moment-based estimation, which was shown to work well in finite samples. In particular, we propose to estimate (4.2) using our L-moment estimator, where we plug a nonparametric estimator of $Q_{Y(0)}$ (the quantile function in the control group) on (4.2) and estimate θ by minimising (2.1).

The plug-in approach will lead to a larger variance than in the case where $F_{Y(0)}$ is known. Indeed, proceeding similarly as in previous sections, we are able to write, under differentiability conditions and as $L, N_1 \wedge N_0 \longrightarrow \infty$, $(N_1 \wedge N_0)/N_j \longrightarrow c_j$, $j \in \{0, 1\}$:

[2] Indeed, as exemplified in Athey et al. (2021,page 1): "For example, Lewis and Rao (2015) analyze challenges with statistical power in experiments designed to measure digital advertising effectiveness. They discuss a hypothetical experiment where the average expenditure per potential customer is \$7, with a standard deviation of \$75, and where an average treatment effect of \$0.35 (0.005 of a standard deviation) would be substantial in the sense of being highly profitable for the company. In that example, an experiment with power for a treatment effect of \$0.35 equal to 80%, and a significance level for the test of means of 0.05, would require a sample size of 1.4 million customers. As a result confidence intervals for the average effect of an advertisement are likely to include zero even if the average effect were substantively important, even with large sample sizes, e.g., over a hundred thousand units."

$$\sqrt{N_0 \wedge N_1}(\hat{\theta}^{\text{plug}} - \theta_0) =$$
$$-(\nabla_{\theta'} h_1^L(\theta_0)'\Omega^L \nabla_{\theta'} h_1^L(\theta_0))^{-1}\nabla_{\theta'} h_1^L(\theta_0)'\Omega^L \Bigg[\sqrt{c_1}\sqrt{N_1}\left(\int_{\underline{p}}^{\overline{p}} (\hat{Q}_{Y(1)}(u) - Q_{Y(1)}(u))\boldsymbol{P}^L(u)du\right) -$$
$$\sqrt{c_0}\sqrt{N_0}\left(\int_{\underline{p}}^{\overline{p}} \partial_q G(Q_{Y(0)}(u);\theta_0)(\hat{Q}_{Y(0)}(u) - Q_{Y(0)}(u))\boldsymbol{P}^L(u)du\right)\Bigg] + o_P(1)\,, \tag{4.3}$$

where $h_1^L(\theta) = \int_{\underline{p}}^{\overline{p}}(\hat{Q}_{Y(1)}(u) - G(Q_{Y(0)}(u);\theta))\boldsymbol{P}^L(u)du$; and $\hat{Q}_{Y(d)}$, $d \in \{0,1\}$ denotes the empirical quantile function in the group with $D = d$. The first term in the above equation refers to the asymptotic linear representation that would have been obtained were $Q_{Y(0)}$ known. The additional term is due to nonparametric estimation of $Q_{Y(0)}$. This term will inflate the variance of $\hat{\theta}^{\text{plug}}$, vis-à-vis the case where $Q_{Y(0)}(u)$ is known.[3]

In spite of the additional term in the asymptotic linear representation of the estimator, we are able to show that our L-moment estimator, under optimal weights, a choice of functions $\{P_l\}$ that constitute orthonormal bases and $\underline{p} = 0 < 1 = \overline{p}$, is efficient, in the sense that it attains the semiparametric efficiency bound derived in Lemma 1 of Athey et al. (2021). A proof of this fact is sketched in Appendix G and proceeds similarly as the proof of the parametric case in Section 2.5. Optimal weights can be computed using a Gaussian or Bahadur-Kiefer approximation and the asymptotic linear representation (4.3). Interestingly, our estimator is asymptotically efficient without further corrections,[4] which contrasts with the estimators proposed by Athey et al. (2021), which require estimating the efficient influence function via cross-fitting to correct a first-step estimator. We also expect our estimator to work well in practice, given the Monte Carlo simulations in previous sections. Finally, we note that our plug-in approach is computationally efficient for flexible parametrisations of treatment effects. Indeed, if we consider:

$$G(Q_{Y(0)}(u), u; \theta) = \sum_{j=0}^{K} \theta_j u^j + Q_{Y(0)}(u)\,, \tag{4.4}$$

then the plugin estimator solves a quadratic program, which can be computed effi-

[3] In the text, we consider a parametrisation where $Q_{Y(1)}(u)$ depends on $Q_{Y(0)}$ solely through $Q_{Y(0)}(u)$. This coincides, up to notation, with the parametrisation in Athey et al. (2021). We could consider more general forms of dependence of $Q_{Y(1)}(u)$ on $Q_{Y(0)}$ by working with Gâteaux derivatives (Newey, 1994).

[4] The intuition for this result is that, similarly to the discussion in Section 2.5, a "plug-in" L-moment estimator is asymptotically equivalent to a semiparametric maximum likelihood estimator; and, in the model of Athey et al. (2021), it can be further shown that the semiparametric MLE is asymptotically efficient (Newey, 1994, p. 1357-1358).

ciently.[5,6,7] In this specification, *quantile treatment effects*, which measure how the distributions of treated and untreated potential outcomes differ in the population, are given by $q(u) = Q_{Y(1)}(u) - Q_{Y(0)}(u) = \sum_{j=0}^{K} \theta_j u^j$. The average treatment effect is given by $\beta = \int_0^1 (Q_{Y(1)}(u) - Q_{Y(0)}(u))du = \sum_{j=0}^{K} \frac{\theta_j}{j+1}$.

4.2 Empirical application

We illustrate the usefulness of our approach by using data from a randomised experiment in the municipality São Paulo. In 2018, a large ride-hailing company in Brazil randomised discounts to a subset of its users. The goal of the experiment was to understand whether short-run incentives to modal integration could alter long-run behavior via **learning effects**. Specifically, the platform asked for information from some of its active users and, among the respondents, randomised discounts to trips starting or ending at a subway/train station during two weeks between the end of November and beginning of December 2018. Users were randomised into three regimes: (i) a control group (which was not informed about the experiment); (ii) eligibility to two 20% discounts per day, limited to a total of 10 BRL discount per car ride; and (iii) eligibility to two 50% discounts per day, limited to 10 BRL per ride. The discounts were announced the day they started. We have access to the number of rides taken by each individual in each group on a biweekly basis in the fortnights leading up to, during and after the treatment. We also know how many of these rides started or ended in a subway-train station.

Classical consumer theory, where consumers are assumed to have full knowledge

---

[5] Notice that, for $K > 0$, specification (4.4) is actually not nested in the model (4.2) of Athey et al. (2021). Indeed, for $K > 0$, the u-quantile of $Q_{Y(1)}$ is allowed to depend directly on u, whereas in (4.2) that is not the case. It is straightforward, however, to show that the plug-in L-moment estimator remains consistent in model (4.4); and that the asymptotic linear representation (4.3) also holds for (4.4) if one replaces $G_{Y(0)}(Q_{Y(0)}(u); \theta_0)$ with $G_{Y(0)}(Q_{Y(0)}(u), u; \theta_0)$ in the formula. It should be noted, however, that the efficiency bound computed by Athey et al. (2021) for model (4.2) does not immediately extend to (4.4) when $K > 0$. Indeed, when $K > 0$, the tangent set of model (4.4) has a different structure than the one of (4.2). We thus cannot *a priori* guarantee that the plug-in L-moment estimator is statistically efficient for (4.4) when $K > 0$. We leave exploration of efficiency in this case, and particularly the efficiency bound calculation, for future research. We note, however, that, even if we cannot ensure statistical efficiency of our estimation approach when $K > 0$, in our empirical application our approach is able to provide smaller standard errors than the difference in means estimator even when $K > 0$. Moreover, in light of the linear representation (4.3), our approach also provides us with an specification test of the parametrisation (4.4), by working with the semiparametric version of the J-statistic in Comment 2.

[6] Observe that, since $G(Q_{Y(0)}(u), u; \theta)$ is a quantile function, we expect it to be non-decreasing. In some settings (e.g. Gourieroux and Jasiak (2008)), it is desirable that the *estimated* quantile function preserve monotonicity. In our context, we may impose monotonicity in the estimation of (4.4) by including linear restrictions in the quadratic program that ensure the estimated quantile function is monotonic. Altrnatively, we may estimate the quantile function without restrictions, and monotonise it by using the rearrangement procedure in Chernozhukov et al. (2009).

[7] Another advantage of our estimation method under this polynomial specification is that the optimal weighting matrix can be computed without resorting to a first-step estimator of θ_0. Indeed, since $\partial_q G(Q_Y(0)(u), u; \theta_0) = 1$, an estimator of the optimal weighting matrix may be computed by estimating densities in the treatment and control groups nonparametrically and using the empirical quantile functions in each group. In our empirical application, we use the estimator of Cattaneo et al. (2020) to estimate the densities in each group nonparametrically.

of the goods available to them, predicts two types of effects related to giving discounts: first, one would expect a **substitution** effect, whereby users reduce consumption in other goods and increase consumption in the good affected by the discount. In our setting, such effect is expected to increase bimodal rides in the two weeks during which the discount is available, and reduce the number of "unimodal" rides in the same period. One would also expect a reduction of both types of rides in the weeks after the discount, due to intertemporal substitution (consumers shift consumption between periods). Nonetheless, given that transportation may be taken to be a necessary good, we would expect such effects to be quite small. On the other hand, standard theory also predicts **an income effect**: given that a consumer's basket is less costly, there is more available income to spend. This effect is expected to increase consumption in both types of rides both during and after treatment.

Finally, a third possible effect, which is not predicted by standard consumer theory, would be **a learning effect**. This effect hinges on the assumption that consumers do not fully know *ex-ante* the goods available to them. In our setting, where discounts are randomised among active users of the platform, this effect would be expected to occur among users that do not use public transportation. Inasmuch as the discounts lead them to learn more about this service, one would expect a long-term change in bimodal rides, even after the discount is over. Note that such change could be either positive or negative. We would also expect an oppositely-signed, long-term effect, on "unimodal" rides, as users shift from (to) "unimodal" rides to (from) "bimodal" ones.

In practice, a mix of the three types of effects is expected to occur. Since they act in counteracting directions, one would expect the overall effect to be small. In light of that, and given the heavy tailed nature of the distribution – the average number of rides by fortnight in the dataset, even after excluding those pairs of user/fortnights where no rides were taken, is 3.64 rides, whereas the maximum is 57 rides –, a simple comparison of means is expected to perform poorly. We thus propose to analyse the experiment using the methodology introduced in this chapter.

To perform our analysis, we restrict our sample to users who completed at least one ride between the start of 2018 and up until the day before treatment was in place. With that, we expect to estimate effects among users who already know the car-riding services (so learning effects should not act towards increasing "unimodal" rides). This leaves us with $N_0 = 1,291$ control units, $N_{1,20\%} = 1,381$ units in the 20% discount treatment arm, and $N_{1,50\%} = 1,319$ in the 50% treatment arm. We analyse the effect of each treatment on bimodal and "unimodal" (total minus bimodal) rides, in the fortnights leading up to, on, and after treatment was in place. Since users were only alerted of the treatment the day it started, we expect no effects in the fortnight leading up to treatment. We contrast a simple comparison of means between treatment and control arms with parametrisation (4.4). We consider specifications with $K = 0, 1, 2, 3$ for "unimodal" rides. The "bimodal" outcome is lightly-tailed, so the difference in means tends to perform well. Indeed, Table 4.1 reports treatment effects estimates from differences in means for bimodal rides, in the fortnight prior to ($t = -1$), during ($t = 0$) and after ($t > 0$) treatment was in place. Standard errors are reported in parentheses. We observe a strong and significant increase in bimodal rides on the fortnight during which treatment is in place, and no effects prior or after that. As expected, the 50% discount leads to a larger increase in bimodal rides than the 20%

discount.

Tables 4.2 and 4.3 report results for unimodal rides using the difference in means estimator and the parametric specifications (4.3) with K ranging from 0 to 3. In using the L-moment estimator, we set $L = 10$. We also report the p-value from the specification test in Comment 2. A few patterns stand out: first, the parametric specifications allow us to obtain substantial reductions in standard errors (vis-à-vis the difference in means estimator), especially with $K = 0$ and $K = 1$.[8] This gain in precision allows us to make tighter inferences on treatment effects. Specifically, and discarding specifications for which the J-test rejects the null, we note that:

1. For the 20%, systematic effects on unimodal rides do not appear to exist – estimated effects in the low-K specifications that survive the J-test are small, (relatively) precisely estimated and statistically insignificant.
2. For the 50% discount, there is strong evidence of a **decrease** in unimodal rides in $t = 1$ and $t = 3$. There is also some evidence of a decrease in rides in $t = -1$ when using values of $K > 0$, which would suggest some anticipation effect, even though the discount was announced at $t = 0$. However, it should be noted that this effect at $t = -1$ does not survive in the constant treatment effect specification ($K = 0$), which also estimates effects much more precisely than other specifications. Similarly, there is some evidence of negative effects at $t = 5$ when using $K > 0$, though this effect does not survive the constant treatment effect specification.

All in all, the results appear compatible with the 50% discount producing large intertemporal substitution effects on unimodal rides. The fact that bimodal rides in other periods are not affected could be explained by the assumption that bimodal rides may be considered a necessary good, with low degree of downward substitutability. Negative effects on unimodal rides persist over a month after the discount is over.[9] However, it is interesting to note that there does not appear to be a contemporaneous substitution effect. This could be explained by the income effect counterbalancing the substitution effect at $t = 0$. Such large income effect in the contemporaneous period is expected to occur in settings where the return to savings is low. Finally, note that we do not find evidence of learning effects towards bimodal rides.

Even though our results could be driven by standard substitution effects – coupled with a downward substitutability constraint on bimodal rides, which may be interpreted as reflecting the "necessary" character of this good –, one could also interpret them as being driven by a setting where users, by substantially increasing bimodal rides, learn about the quality of public transportation, and, being positively surprised, shift consumption away from unimodal rides towards purely public transportation trips, whilst keeping bimodal

[8] For values of K larger than 1, in some periods the standard error of the parametric estimator exceeds that of the difference in means. This is due to the fact that L is kept fixed. By increasing L, we are able to decrease the estimated standard error, though potentially at the cost of finite-sample bias. An interesting topic for future research would be to adapt the selection methods in Chapter 3 to the semiparametric setting, so as to produce methods to automatically select L with good statistical properties.

[9] The fact that we dot not encounter significant effects on rides at $t = 2$, but at $t = 1$ and $t = 3$, could be justified by the fact that this fortnight corresponds to the last week of 2018 and the first week of 2019, when unimodal rides may be assumed difficult to substitute due to the holidays/end-of-year celebration.

rides unchanged. In Appendix H, we introduce a learning model that is able to produce these results. This model may be able to better explain the persistence of strong and negative effects on unimodal rides, which may be incompatible with *a priori* reasonable values for the intertemporal elasticity of substitution of rides, depending on the relative price of bimodal and unimodal rides. As future research, it would be interesting to obtain access to the monetary values of each ride, so that we could better understand which explanation is more appropriate. One would also hope to further explore the model in Appendix H, so as to derive further testable implications on our available data.

Table 4.1: *Bimodal rides*

20% discount

$t=-1$		$t=0$		$t=1$		$t=2$		$t=3$		$t=4$		$t=5$		$t=6$	
Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean
0.1084	-0.0100	0.1325	0.0471	0.1154	0.0157	0.0434	0.0037	0.1030	-0.0053	0.1038	-0.0046	0.0813	0.0258	0.1131	0.0086
(0.0140)	(0.0189)	(0.0181)	(0.0269)	(0.0154)	(0.0231)	(0.0070)	(0.0098)	(0.0149)	(0.0212)	(0.0151)	(0.0197)	(0.0132)	(0.0222)	(0.0145)	(0.0218)

50% discount

$t=-1$		$t=0$		$t=1$		$t=2$		$t=3$		$t=4$		$t=5$		$t=6$	
Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean	Mean control	Diff mean
0.1084	-0.0061	0.1325	0.0950	0.1154	-0.0100	0.0434	0.0105	0.1030	-0.0090	0.1038	-0.0090	0.0813	0.0134	0.1131	0.0021
(0.0140)	(0.0189)	(0.0181)	(0.0281)	(0.0154)	(0.0210)	(0.0070)	(0.0120)	(0.0149)	(0.0200)	(0.0151)	(0.0199)	(0.0132)	(0.0193)	(0.0145)	(0.0206)

Table 4.2: *Unimodal rides: 20% discount*

	Mean control	Diff Means	$t=-1$ $K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9899	-0.0377	0.0202	-0.1683	-0.1192	-0.1312
Std. Error	(0.0777)	(0.1148)	(0.0420)	(0.0987)	(0.1083)	(0.1091)
pval J-test			0.1139	0.2800	0.2841	0.2561
			$t=0$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1472	0.0961	0.0475	0.1121	0.1018	0.1015
Std. Error	(0.0860)	(0.1257)	(0.0455)	(0.1000)	(0.1146)	(0.1146)
pval J-test			0.8334	0.8110	0.7267	0.6181
			$t=1$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1526	-0.0519	0.0229	0.0802	0.0667	0.0693
Std. Error	(0.0900)	(0.1258)	(0.0437)	(0.1013)	(0.1125)	(0.1129)
pval J-test			0.3235	0.2687	0.1959	0.1340
			$t=2$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.2208	0.0979	0.0301	0.2108	0.1386	0.1527
Std. Error	(0.0601)	(0.0853)	(0.0404)	(0.0913)	(0.1005)	(0.1020)
pval J-test			0.3187	0.6993	0.9210	0.9251
			$t=3$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.8737	0.0046	-0.0045	0.0316	0.0284	0.0286
Std. Error	(0.0895)	(0.1253)	(0.0521)	(0.1133)	(0.1278)	(0.1288)
pval J-test			0.9995	0.9990	0.9969	0.9906
			$t=4$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9535	-0.0136	0.0648	0.1932	0.1424	0.1546
Std. Error	(0.0883)	(0.1267)	(0.0491)	(0.1152)	(0.1250)	(0.1273)
pval J-test			0.0137	0.0135	0.0113	0.0065
			$t=5$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	1.9620	-0.0707	-0.0141	-0.0917	-0.0660	-0.0744
Std. Error	(0.0916)	(0.1299)	(0.0654)	(0.1270)	(0.1336)	(0.1349)
pval J-test			0.9800	0.9802	0.9770	0.9631
			$t=6$			
	Mean control	Diff Means	$K=0$	$K=1$	$K=2$	$K=3$
Estimate	2.1309	-0.0201	-0.0120	-0.0980	-0.0609	-0.0851
Std. Error	(0.0977)	(0.1358)	(0.0506)	(0.1256)	(0.1313)	(0.1351)
pval J-test			0.9641	0.9642	0.9820	0.9883

Table 4.3: *Unimodal rides: 50% discount*

			$t = -1$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	1.9899	-0.1507	-0.0323	-0.2035	-0.1804	-0.1735
Std. Error	(0.0777)	(0.1067)	(0.0382)	(0.0973)	(0.1021)	(0.1035)
pval J-test			0.7931	0.9868	0.9902	0.9829
			$t = 0$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	2.1472	-0.0554	0.0040	-0.1070	-0.0558	-0.0655
Std. Error	(0.0860)	(0.1190)	(0.0419)	(0.1066)	(0.1130)	(0.1135)
pval J-test			0.8439	0.8909	0.9723	0.9872
			$t = 1$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	2.1526	-0.2049	-0.0487	-0.3115	-0.2299	-0.2475
Std. Error	(0.0900)	(0.1259)	(0.0426)	(0.1048)	(0.1108)	(0.1124)
pval J-test			0.0658	0.3854	0.8441	0.8642
			$t = 2$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	1.2208	-0.0767	-0.0118	-0.0997	-0.0838	-0.0762
Std. Error	(0.0601)	(0.0839)	(0.0395)	(0.0865)	(0.0966)	(0.0982)
pval J-test			0.9864	0.9984	0.9971	0.9957
			$t = 3$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	1.8737	-0.2960	-0.0803	-0.3768	-0.2032	-0.2172
Std. Error	(0.0895)	(0.1177)	(0.0506)	(0.1039)	(0.1245)	(0.1247)
pval J-test			0.0004	0.0107	0.0611	0.1419
			$t = 4$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	1.9535	-0.1613	0.0074	-0.0931	-0.0772	-0.0700
Std. Error	(0.0883)	(0.1230)	(0.0468)	(0.1113)	(0.1201)	(0.1223)
pval J-test			0.6302	0.6387	0.5455	0.4397
			$t = 5$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	1.9620	-0.2357	-0.0290	-0.3059	-0.2555	-0.2520
Std. Error	(0.0916)	(0.1257)	(0.0641)	(0.1253)	(0.1307)	(0.1326)
pval J-test			0.4859	0.9848	1.0000	1.0000
			$t = 6$			
	Mean control	Diff Means	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Estimate	2.1309	-0.2295	-0.0423	-0.2098	-0.1480	-0.1716
Std. Error	(0.0977)	(0.1304)	(0.0530)	(0.1293)	(0.1366)	(0.1396)
pval J-test			0.6826	0.8051	0.9210	0.9270

Chapter 5

Concluding remarks

This thesis considered the estimation of parametric models using a “generalised” method of L-moments procedure, which extends the approach introduced in Hosking (1990), whereby a d-dimensional parametric model for a distribution function is fit by matching the first d L-moments. We have shown that, by appropriately choosing the number of L-moments and under an appropriate weighting scheme, we are able to construct an estimator that is able to outperform maximum likelihood estimation (and Hosking’s original approach) in small samples from popular distributions, and yet does not suffer from efficiency losses in larger samples. We have developed tools to automatically select the number of L-moments used in estimation, and have shown the usefulness of such approach in Monte Carlo simulations.

In Chapter 4, we have extended the L-moment approach to the estimation of semiparametric models of treatment effects in randomised experiments. We have shown that the L-moment estimator is both computationally attractive and asymptotically efficient. We have illustrated our approach using data from a Brazilian RCT.

The extension of the L-moments approach to other semi- and nonparametric settings appears to be a promising venue of future research. The L-moment approach appears especially well-suited to problems where semi- and nonparametric maximum likelihood estimation is computationally complicated. We intend to analyse such extensions in future research.

Appendix A

Proof of Lemma 1

Proof. Observe that the rescaled vector of càglàd L-moments $\hat{\lambda}_r$ may be written as:

$$\begin{pmatrix} \hat{\lambda}_1 \\ \sqrt{3} \cdot \hat{\lambda}_2 \\ \vdots \\ \sqrt{2L-1} \cdot \hat{\lambda}_L \end{pmatrix} = C\hat{M} \,,$$

where C is a $L \times L$ matrix with entry $C_{ij} = \frac{\sqrt{2i-1}}{j}(-1)^{i-j}\binom{i-1}{j-1}\binom{i+j-2}{j-1}$ if $j \leq i$ and zero otherwise; and $\hat{M}$ is the $L \times 1$ vector with r-th entry equal to:

$$\hat{M}_j = \sum_{t=1}^{T} \left[\left(\frac{t}{T}\right)^j - \left(\frac{t-1}{T}\right)^j \right] Z_{t:T} \,.$$

Similarly, it follows from Landwehr et al. (1979) that:

$$\begin{pmatrix} \tilde{\lambda}_1 \\ \sqrt{3} \cdot \tilde{\lambda}_2 \\ \vdots \\ \sqrt{2L-1} \cdot \tilde{\lambda}_L \end{pmatrix} = C\tilde{M} \,,$$

where the j-th entry o $\tilde{M}$ is given by:

$$\tilde{M}_j = \frac{1}{\binom{T}{j}} \sum_{t=1}^{T} \binom{t-1}{j-1} Z_{t:T} \,.$$

By the properties of Legendre polynomials, it follows that $\|C\|_2 = O(L)$, where $\|\cdot\|_2$ denotes the spectral norm. Therefore, it suffices to prove that $\|\hat{M} - \tilde{M}\|_2 = o_P(L^{-1}T^{-1/2})$.

Observe that $\hat{M}_1 = \tilde{M}_1$. Consider $j > 1$. In this case, we may write:

$$\hat{M}_j - \tilde{M}_j = \sum_{t=1}^{j-1} \left[\left(\frac{t}{T}\right)^j - \left(\frac{t-1}{T}\right)^j \right] Z_{t:T} + \sum_{t=j}^{T} \left[\left(\frac{t}{T}\right)^j - \left(\frac{t-1}{T}\right)^j - \frac{j \prod_{s=1}^{j-1}(t-s)}{\prod_{s=1}^{j}(T+1-s)} \right] Z_{t:T} = A_j + B_j .$$

We deal with each term separately. We first observe that, by the mean-value theorem:

$$|A_j| \leq \frac{(j-1)j}{T} \left(\frac{j-1}{T}\right)^{j-1} |Z_{(j-1):T}| ,$$

from which one obtains that:

$$L\sqrt{T} \sum_{j=2}^{L} |A_j| \leq (L-1)L^2 \sqrt{T} \left(\frac{L-1}{T}\right)^{L-1} \frac{\sum_{t=1}^{T} |Z_t|}{T} .$$

Now, by Assumption, $\frac{\sum_{t=1}^{T}|Z_t|}{T}$. Next, by noticing that $(L-1)L^2\sqrt{T}\left(\frac{L-1}{T}\right)^{L-1} = O\left(\left(\frac{L^2}{\sqrt{T}}\right)^L\right) = o(1)$, it follows that $\sqrt{T}\sum_{j=2}^{L}|A_j| = o(1)$, as desired.

As for the second term, we observe that, for $j > 1$ and $t \geq j$, applying the mean-value theorem twice.

$$\left[\left(\frac{t}{T}\right)^j - \left(\frac{t-1}{T}\right)^j - \frac{j \prod_{s=1}^{j-1}(t-s)}{\prod_{s=1}^{j}(T+1-s)} \right] \leq \frac{j}{T}\left(\frac{t}{T}\right)^{j-1} \left[1 - \left(1 - \frac{j-1}{t}\right)^{j-1} \right] \leq \frac{j}{T}\left(\frac{t}{T}\right)^{j-1} \frac{(j-1)^2}{t} \leq \frac{j(j-1)^2}{T^2} ,$$

and, similarly, we can extract a lower bound:

$$\left[\left(\frac{t}{T}\right)^j - \left(\frac{t-1}{T}\right)^j - \frac{j \prod_{s=1}^{j-1}(t-s)}{\prod_{s=1}^{j}(T+1-s)} \right] \geq \frac{j}{T}\left(\frac{t-1}{T}\right)^{j-1} \left[1 - \frac{1}{\left(1-\frac{j-1}{T}\right)^j} \right] \geq -\frac{j}{T}\left(\frac{t-1}{T}\right)^{j-1} \frac{j}{(1-\frac{j-1}{T})} \frac{(j-1)}{T} \geq -C\frac{j^3}{T^2} ,$$

for some constant $C > 0$. Combining both bounds, we obtain that:

$$L\sqrt{T} \sum_{j=2}^{L} |B_j| \leq C \frac{L^5}{\sqrt{T}} \frac{\sum_{t=1}^{T} |Z_t|}{T} ,$$

which proves the desired result. □

Appendix B

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:

Assumption 11. *There exists $C > 0$ and $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that, for every $L \in \mathbb{N}$ and $\epsilon > 0$:*

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

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

It is clear that Assumption 11 implies Assumption 4. Perhaps less obviously, Assumption 11 implies Assumption 8 under conditions that allow differentiability under the integral sign (Assumption 5).

Proposition 4. *Suppose Assumption 5 holds. Then Assumption 11 implies Assumption 8.*

Proof. Suppose Assumption 11 holds. Fix $\iota \in \mathbb{R}^d$, $\|\iota\|_2 = 1$. We then have that, by Assumption 11:

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

Taking limits yields that:

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

Now, since $\nabla_{\theta'}h^L(\theta_0)'\Omega^L\nabla_{\theta'}h^L(\theta_0)$ is symmetric and real, it admits an eigedecomposition $P_L\Lambda_L P_L'$, where $P_L'P_L = I_d$ and $\Lambda_L = \text{diag}(\lambda_{1L}, \lambda_{2L}\ldots\lambda_{dL})$, with $\lambda_{1L}\leq\lambda_{2L}\ldots\leq\lambda_{dL}$ being the eigenvalues of $\nabla_{\theta'}h^L(\theta_0)'\Omega^L\nabla_{\theta'}h^L(\theta_0)$. This in turn implies that:

$$\lambda_{1L} = \min_{x : \|x\|_2 = 1} x' \Lambda_L x = \min_{u : \|u\|_2 = 1} (P_L u)' \Lambda_L (P_L u) \geq C > 0 \,,$$

which proves the desired result. $\square$

Appendix C

Calculations for optimal weighting matrix in the iid case

Consider the optimal weighting matrix as in (2.11). We focus on the case where $0 = \underline{p} < \overline{p} = 1$ and the data is iid. Note that we may write:

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

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

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

Since standard L-moments consist of a choice of weighting functions $\mathbf{P}^L$ 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:

$$W_{r,s} = \int_0^1 \int_0^1 \frac{(U \wedge V - UV)}{f_{\theta_0}(Q_y(U))f_{\theta_0}(Q_y(V))} U^r V^s dU dV,$$

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

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

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 D

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 (2.4). In the main text, we have shown that, under a Gaussian approximation, the term $\sqrt{T}h^L(\theta_0)$ can be approximated by the integral of a Brownian bridge. Consider, now, the alternative process:

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

where $\check{Q}_Y(u)$ is the quantile function associated with the 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 3 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 3 in Alvarez-Andrade and Bouzebda (2013) indicates that $\left(\int_{\underline{p}}^{\bar{p}} \sqrt{T}(\check{Q}_Y(u) - \hat{Q}_Y(u))\mathbf{P}^L(u)du \right)$ is strongly approximated by the integral of a Brownian bridge. Importantly, this strong approximation is **identically distributed** to the strong approximation of the term $\sqrt{T}h^L(\theta_0)$ obtained in the main text. This motivates the use of the weighted bootstrap to conduct inference in our setting: given consistent estimators of θ_0 an Ω_L, we can simulate the distribution of A_T by generating random draws of the Z_t and computing the quantile function of the generated weighted cdf.

If the requirements on the density discussed in the previous paragraph are deemed too strong, we note that we are able to obtain a strong approximation to the weighted bootstrap using only the assumptions in Theorem 3 if we consider Theorem 3.1 of Gu and Ghosal (2008). Nonetheless, it should be noted their results are restricted to the Bayesian bootstrap.

Appendix E

Proof of Proposition 3

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

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

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

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

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

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

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

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

$$m(\beta) := \begin{pmatrix} -\nabla_{\theta'} h^L(\theta)'\lambda \\ -h^L(\theta) - (W^L)^{-1}\lambda \end{pmatrix}\,.$$

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

$$\begin{aligned} 0 = m(\hat{\beta}) = m(\beta_0) + M(\hat{\beta} - \beta_0) + \frac{1}{2}\sum_j (\hat{\beta}_j - \beta_{j0})\partial_j M(\hat{\beta} - \beta_0) + \\ + \frac{1}{6}\sum_{i,j}(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\partial_{i,j} M(\hat{\beta} - \beta_0) + \frac{1}{24}\sum_{g,i,j}(\hat{\beta}_g - \beta_{g0})(\hat{\beta}_i - \beta_{i0})(\hat{\beta}_j - \beta_{j0})\widetilde{\partial_{g,i,j} M}(\hat{\beta} - \beta_0)\,, \end{aligned}$$

where $M = \nabla_{\beta'} m(\beta_0)$ and $\partial_j M$ is the $(d+L)\times(d+L)$ 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 $(d+L)\times(d+L)$ matrix with entry (l,k) equal to $\frac{\partial m^l(\theta_0)}{\partial\beta_i\partial\beta_j\partial\beta_k}$; and $\widetilde{\partial_{g,i,j}M}$ is a $(d+L)\times(d+L)$ 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:

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

Letting:

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

Then by Assumption 9, $\|M - M_0\|_2 = O_p(T^{-1/2})$. Moreover, note that:

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

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

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

Our first goal is to show that, on $\hat{\theta}_T \in \mathcal{O}$ and existence of $(W^L)^{-1}$ and $(\Omega^L)^{-1}$, $(\hat{\beta}-\beta_0) = -M_0^{-1}m(\beta_0) + O_p(T^{-1})$. 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 $\|\Omega_L^*\|_2 = O(1)$ (Assumption 8), $\Omega_L = O(1)$ (Assumption 3), $\|\nabla_{\theta'}h^L(\theta_0)\|_2^2 \leq \mathrm{tr}(\nabla_{\theta'}h^L(\theta_0)'\nabla_{\theta'}h^L(\theta_0)) \leq \sum_{s=1}^p \int_{\underline{p}}^{\bar{p}} |\partial_{\theta_s}Q_Y(u|\theta_0)|^2 du < \infty$ (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 (E.1) 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:

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

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:

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

where we used that $\sum_{j=1}^{p+L} \|\partial_j M_j\|_2^2 = O(1)$, which follows from Bessel's inequality and Assumption Assumption 10.

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

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

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 9, 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 10 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 $(W^L)^{-1}$ and $(\Omega^L)^{-1}$:

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

Plugging this expression back onto (E.1) 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 (3.1). 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 $(W^L)^{-1}$ and $(\Omega^L)^{-1}$ exist. By Assumption 9 and $\theta_0 \in \mathcal{O}$, we know that $\mathbb{1}_{I_T} \xrightarrow{p} 1$ and $\mathbb{1}_{I_T^\complement}\sqrt{T}(\hat{\theta}_T-\theta_0) = o_p(1)$.[1] To show the rates derived from (E.1) 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 $\mathbb{P}[I_T^\complement] \longrightarrow 0$, which proves the desired result. Using that $\mathbb{1}_{I_T^\complement} = o_p(T^{-3/2})$, we can write

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

which proves the result. □

[1] 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.

Appendix F

Conditions for validity of Lasso approach

This appendix presents sufficient conditions for the validity of the Lasso approach described in the main text. 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_L. We state these assumptions below. In what follows, define, for $v \in \mathbb{R}^n$, $\|v\|_0 := \#\{j : v_j \neq 0\}$.

Assumption 12 (Approximate sparsity of combination matrix)**.** *For* $j = 1, \ldots, d$, *let* $\lambda_j^* := A_L' e_j$. *We assume that, for each* j, *there exist constants* K_j^l, K_j^u *and a sequence of vectors* $\bar{\lambda}_d \in \mathbb{R}^L$ *such that, as* $T, L \longrightarrow \infty$*:*

1. $\|\bar{\lambda}_j\|_0 = s_T$
2. $\|\bar{A}_L - A_L\|_2 = o(1)$, *where* $\bar{A}_L = \begin{bmatrix} \bar{\lambda}_1 & \bar{\lambda}_2^* & \ldots & \bar{\lambda}_d^* \end{bmatrix}'$.

Assumption 13 (Sparse eigenvalue and spectral norm condition)**.** *Let:*

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

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

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

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

Assumption 14 (Penalties)**.** *The penalties satisfy:*

1. *For a sequence α_T converging to zero such that $\alpha_T L \longrightarrow \infty$:*

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

where $k = (1+\epsilon)\sqrt{T\Phi^{-1}(1-\frac{\alpha_T}{4Ld})}$ 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:*

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

Under Assumptions 12-14, 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 $1-\epsilon_T$

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

where $\hat{\lambda}_j$ denotes the solution to program (3.7).

The bound in (F.1), together with Assumption 12, implies that:

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

If we assume that $\frac{s_T^2\log(L)}{T} \longrightarrow 0$, then $\|\tilde{A}_L - A_L\|_2 = o_p(1)$, 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 Chapter 2. To see this, we observe that the estimator $\hat{\theta}^{\text{selected}}$ solves $R(\hat{\theta}^{\text{selected}}) = 0$, where:

$$R(\theta) = \tilde{A}_L h^L(\theta)\,.$$

If we define the population objective as $R_0(\theta) = A_L[\int_{\underline{p}}^{\overline{p}}(\hat{Q}_Y(u) - Q_Y(u|\theta)\boldsymbol{P}^L(u)]du$, 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 G

Efficiency of semiparametric L-moment estimator

Let $N = N_0 + N_1$ denote the sample size. Consider the alternative (unfeasible) estimator:

$$\check{\theta} \in \operatorname{argmin}_{\theta\in\Theta} \sum_{i\in\mathcal{S}_N}\sum_{j\in\mathcal{S}_N}\Big(\frac{1}{\sqrt{p_1}}(\hat{Q}_{Y(1)}(i) - Q_{Y(1)}(\theta)) + \frac{1}{\sqrt{p_0}}\partial_q G \cdot (\hat{Q}_{Y(0)}(i) - Q_Y(0))\Big) \times \kappa_{i,j} \times \left(\frac{1}{\sqrt{p_1}}(\hat{Q}_{Y(1)}(j) - Q_{Y(1)}(\theta)) + \frac{1}{\sqrt{p_0}}\partial_q G \cdot (\hat{Q}_{Y(0)}(j) - Q_Y(0))\right),$$

for a grid of S_N points $\mathcal{S}_N = \{s_1, s_2, \ldots, s_{S_N}\} \subseteq (0,1)$ and weights $\kappa_{i,j}$, $i, j \in \mathcal{S}_N$. Here, $Q_{Y(1)}(\theta) = G(Q_{Y(0)}(u), \theta)$ and $\partial_q G = \partial_q G(Q_{Y(0)}(u), \theta_0)$. We also set $p_l = \lim \frac{N_l}{N}$, $l \in \{0,1\}$, and we assume $p_l \in (0,1)$. Under some conditions, and as $N, S_N \longrightarrow \infty$, the estimator has asymptotic linear representation as follows:

$$\sqrt{N}(\check{\theta} - \theta_0) = -(\partial_{\theta'} G'_{S_N} \boldsymbol{\kappa}_{S_N} \partial_{\theta'} G_{S_N})^{-1} \partial_{\theta'} G'_{S_N} \boldsymbol{\kappa}_{G_N} \left[\frac{1}{\sqrt{p_1}} \boldsymbol{f}_{Y(1)}{}^{-1} * \sqrt{N_1} F_{Y(1),S_N} + \frac{\boldsymbol{\partial_q G}}{\sqrt{p_0}} * \boldsymbol{f}_{Y(0)}{}^{-1} * \sqrt{N_0} F_{Y(0),S_N}\right] + o_p(1),$$

where, as in Section 2.5, $*$ denotes entry-by-entry multiplication, $\boldsymbol{f}_{Y(d)}{}^{-1} = \left(\frac{1}{f_{Y(d)}(Q_{Y(d)}(s_1)} \quad \frac{1}{f_{Y(d)}(Q_{Y(d)}(s_2))} \quad \cdots \quad \frac{1}{f_{Y(d)}(Q_{Y(d)}(s_{S_N}))}\right)'$ and $\boldsymbol{\partial_q G}$ is similarly defined. First, we observe that:

$$\frac{1}{f_{Y(1)}(Q_{Y(1)}(v))} = Q'_{Y(1)}(v) = \partial_q G(Q_{Y(0)}(v); \theta_0) \cdot \frac{1}{f_{Y(0)}(Q_{Y(0)}(v))},$$

which implies, by taking $\mathcal{S}_N = \left\{\frac{1}{S_N+1}, \frac{2}{S_N+1}, \ldots, \frac{S_N}{S_N+1}\right\}$ and applying Lemma C.1. in Firpo et al. (2022), that we are able to show that the variance of the estimator under optimal weights is:

$$\mathbb{V}^* = ((\partial G_{S_N} * (\mathbf{1}_d' \otimes \boldsymbol{f}_{Y(1)}))' \Sigma_{S_N}^{-1} (\partial G_{S_N} * (\mathbf{1}_d' \otimes \boldsymbol{f}_{Y(1)})))^{-1} ,$$

where

$$(\Sigma_{S_N}^{-1})_{s_i,s_j} = p_0 p_1 [\mathbb{1}_{\{s_i=s_j\}} 2(S_N+1) - (\mathbb{1}_{\{s_i=s_{j+1}\}} + \mathbb{1}_{\{s_i=s_{j-1}\}})(S_N+1)] .$$

Proceeding similarly as in Section 2.5, we obtain that:

$$\lim_{N\to\infty} \left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = p_0 p_1 \int_0^1 \left.\frac{dH_{d_1}(v)}{dv}\right|_{v=u} \left.\frac{dH_{d_2}(v)}{dv}\right|_{v=u} du ,$$

with $H_d(u) = f_{Y(1)}(Q_{Y(1)}(v))\partial_{\theta_d} G(Q_{Y(0)}(v), \theta_0)$. Next, proceeding similarly as in the proof of Section 2.5, we conclude that:

$$\left(\mathbb{V}^{*-1}\right)_{d_1,d_2} = p_0 p_1 (I(\theta_0))_{d_1,d_2} ,$$

where $I(\theta_0)$ is the Fisher information matrix of the parametric model $\theta \mapsto f_{Y(1)}(y|\theta)$ that assumes $Q_{Y(0)}$ known. It then follows by Lemma 1 of Athey et al. (2021) that the estimator is asymptotically efficient, as it achieves the efficiency bound derived by the authors.

To conclude, we note that, when the $\{P_l\}_l$ form orthonormal **bases**, the estimator $\check{\theta}$ corresponds to a method-of L-moments estimator that uses **infinitely** many moments. We are then able to show, by reasoning similarly as in Section 2.5, that the estimator described in Chapter 4, which uses a finite but increasing number of L-moments, is also efficient, since under an appropriate choice of weights this estimator is asymptotically equivalent to $\check{\theta}$.

Appendix H

A learning model on the demand for public transportation

In this section, we introduce a simple model which is able to rationalise the findings in Chapter 4. The model consists of a standard intertemporal choice problem with one additional ingredient: a learning mechanism, whereby increased demand for bimodal rides increases knowledge of the quality of public transportation in the future. The model produces two effects: first, a discount in an initial period may increase demand for bimodal rides in the same period due to a *learning bequest*, whereby an agent increases demand so as to have better information on the quality of the service in the future. Moreover, for a given configuration of parameters, bimodal rides need not change by much, even if public transportation is revealed to be of good quality: in contrast, unimodal rides decrease when this occurs. The two effects combined may be better able to rationalise the strong and long-running negative substitution effects reported in the main text for unimodal rides, and the insignificant changes encountered in bimodal rides.

We consider a two-period problem ($t \in \{0, 1\}$) where an agent has to choose her share of consumption on unimodal rides (u), bimodal rides (b) and public transportation (c). Her instantaneous preferences at each period are given by:

$$S(u_t, b_t, c_t) = u_t^\alpha + \gamma A^\phi b_t^\alpha + A c_t^\alpha \,,$$

for positive constants α, γ, ϕ and $\alpha < 1$. Observe that the parameter A governs the relative gain of choosing public transportation over unimodal rides. It also affects the relative gain of bimodal rides through the term A^ϕ.[1] We assume the agent does not observe A, but has a Gaussian prior over $\log(A)$, i.e. the agent assumes $\log(A) \sim N(\mu_0, \sigma_0^2)$. Consumption of bimodal rides and public transportation in period 0 produces an unbiased signal of the quality of rides, where the informativeness of the signal is inversely related to the amount consumed in period 0. Specifically, we assume that, afeter consuming (b_0, c_0) units in period zero, the agent observes a signal:

[1] The term A^ϕ could be interpreted as a reduced form for the agent's problem of deciding the composition of bimodal rides between public transportation and car trips according to a Cobb-Douglas production function, where A enters the production function as a public-transportation-augmenting factor.

$$Y|\log(A) \sim N\left(\log(A), \frac{1}{h_1(b_0, c_0)}\right),$$

for an increasing, everywhere positive and differentiable mapping h_1. We assume that the agent maximises expected utility, given the information available to her. Specifically, and taking c as the numéraire, at period 0, the agent solves:

$$\begin{aligned} &\max_{u_0, b_0, c_0 \geq 0} \mathbb{E}[S(u_0, b_0, c_o) + \beta S(u_1(y,d), b_1(y,d), c_1(y,s))] \\ s.t. \quad & d = (1+r)[w_0 - p_{0,b}b_0 - p_{0,u}u_0 - c_0], \end{aligned} \tag{H.1}$$

where β is the discount factor, d is the savings from period zero to one, w_0 is the income in period 0, and r is the interest rate. The random variables $u_1(y,d), b_1(y,d), c_1(y,s)$ are the solutions to the second period problem, i.e.:

$$\begin{aligned} &\max_{u_1, b_1, c_1 \geq 0} \mathbb{E}[S(u_1, b_1, c_1)|Y] \\ s.t. \quad & d + w_1 = p_{1,b}b_1 + p_{1,u}u_1 + c_1, \end{aligned} \tag{H.2}$$

We note that the model above nests a standard intertemporal choice problem with full knowledge of the quality of the service if we set $\sigma_0^2 = 0$.

We begin by solving the model by backward induction. First, by known results on conjugate priors, we have that:

$$\log(A)|Y \sim N\left(\frac{h_0\mu_0 + h_1(b_0, x_0)Y}{h_0 + h_1(b_0, c_0)}, (h_0 + h_1(b_0, c_0))^{-1}\right),$$

where $h_0 = 1/\sigma_0^2$ is the precision of the prior. It then folows, by the properties of the lognormal distribution that:

$$\mathbb{E}[A|Y] = \exp\left(\frac{h_0\mu_0 + h_1(b_0, c_0)Y}{h_0 + h_1(b_0, c_0)} + \frac{(h_0 + h_1(b_0, c_0))^{-1}}{2}\right), \tag{H.3}$$

and, similarly:

$$\mathbb{E}[A^\delta|Y] = (\mathbb{E}[A|Y])^\delta \exp\left(\delta(\delta - 1)\frac{(h_0 + h_1(b_0, c_0))^{-1}}{2}\right) = (\mathbb{E}[A|Y])^\delta \psi_\delta(b_0, c_0). \tag{H.4}$$

Next, we note that the first-order conditions on the second period problem entail:

$$
\begin{aligned}
u_1 &= \left(\frac{1}{p_{1,u}\mathbb{E}[A|Y]} \right)^{\frac{1}{1-\alpha}} c_1 \,, \\
b_1 &= \left(\frac{\gamma \psi_\delta(b_0, c_0)}{p_{1,b}\mathbb{E}[A|Y]^{1-\delta}} \right)^{\frac{1}{1-\alpha}} c_1 \,, \\
c_1 &= \frac{d + w_1}{p_{1,u}^{-\alpha/(1-\alpha)}\mathbb{E}[A|Y]^{-1/(1-\alpha)} + p_{1,b}^{-\alpha/(1-\alpha)} \left(\frac{\gamma \psi_\delta(b_0,c_0)}{p_{1,b}\mathbb{E}[A|Y]^{1-\delta}} \right)^{\frac{1}{1-\alpha}} + 1} \,.
\end{aligned}
\tag{H.5}
$$

A few properties show up in this optimisation. First, if $\delta < 1$, then, all else equal, an increase in the signal Y unambiguously increases c_1 and necessarily decreases u_1. As for the effect on bimodal rides, its sign depends on relative prices and the region $\mathbb{E}[A|Y]$ lies. Specifically, it is immediate from the above expression that:

$$
\frac{db_1}{dY} \propto p_{1,u}^{-\alpha/(1-\alpha)} \delta \mathbb{E}[A|Y]^{(\alpha-\delta-1)/(1-\alpha)} - (1-\delta)\mathbb{E}[A|Y]^{(\alpha-\delta)/(1-\alpha)} \,. \tag{H.6}
$$

Observe that, for large enough values of $p_{1,u}$, the effect is negative. Alternatively, if $\delta \approx 1$, the effect is positive. Clearly, there exist combination of parameter-signals where the effect is arbitrarily small.

As for understanding the learning bequest, it is useful to consider some candidate choice $(\tilde{b}_0, \tilde{c}_0)$ which leads to a signal with precision $\tilde{h}_0$; and an alternative consumption bundle which produces a signal with precision $\check{h}_0 < \tilde{h}_0$. In this case, it can be shown that, from the viewpoint of the posterior distribution, observing a signal with precision $\check{h}_0$ is equivalent to observing **two independent** signals: one with precision $\check{h}_0$, and an additional signal with another precision. It then follows by iterated expectations and the properties of maximisation that, from the viewpoint of the agent in period 0, choosing a more precise signal necessarily leads to higher expected utility in period 1.[2] In this way, a discount on bimodal rides acts in period 0 as a subsidy to the learning mechanism, and we would expect an increase in bimodal rides at period 0 due to this reason (in addition to income and substitution effects).

[2] Alternatively, this result can be derived by applying Blackwell's informativeness theorem (de Oliveira, 2018).

References

Abadie, A., J. Gu, and S. Shen (2019): "Instrumental Variable Estimation with First Stage Heterogeneity," Tech. rep., MIT Economics Department.

Ahidar-Coutrix, A. and P. Berthet (2016): "Convergence of Multivariate Quantile Surfaces," *arXiv:1607.02604*.

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., P. J. Bickel, A. Chen, G. W. Imbens, and M. Pollmann (2021): "Semiparametric Estimation of Treatment Effects in Randomized Experiments," *arXiv:2109.02603*.

Barrio, E. D., E. Giné, and F. Utzet (2005): "Asymptotics for L2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances," *Bernoulli*, 11, 131 – 189.

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, 2369–2429.

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

Billingsley, P. (2012): *Probability and Measure*, Wiley.

Boulange, J., N. Hanasaki, D. Yamazaki, and Y. Pokhrel (2021): "Role of dams in reducing global flood exposure under climate change," *Nature communications*, 12, 1–7.

Broniatowski, M. and A. Decurninge (2016): "Estimation for models defined by conditions on their L-moments," *IEEE Transactions on Information Theory*, 62, 5181–5198.

Carrasco, M. and J.-P. Florens (2014): "On the Asymptotic Efficiency of GMM," *Econometric Theory*, 30, 372–406.

Cattaneo, M. D., M. Jansson, and X. Ma (2020): "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 115, 1449–1455.

Cheng, X. and Z. Liao (2015): "Select the valid and relevant moments: An information-based LASSO for GMM with many moments," *Journal of Econometrics*, 186, 443–464.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009): "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, 96, 559–575.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Csorgo, M. and P. Revesz (1978): "Strong Approximations of the Quantile Process," *Annals of Statistics*, 6, 882–894.

Das, S. (2021): "Extreme rainfall estimation at ungauged locations: information that needs to be included in low-lying monsoon climate regions like Bangladesh," *Journal of Hydrology*, 601, 126616.

de Oliveira, H. (2018): "Blackwell's informativeness theorem using diagrams," *Games and Economic Behavior*, 109, 126–131.

Dey, S., J. Mazucheli, and S. Nadarajah (2018): "Kumaraswamy distribution: different methods of estimation," *Computational and Applied Mathematics*, 37, 2094–2111.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009): "Choosing instrumental variables in conditional moment restriction models," *Journal of Econometrics*, 152, 28–36, recent Adavances in Nonparametric and Semiparametric Econometrics: A Volume Honouring Peter M. Robinson.

Donald, S. G. and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

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

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

Fotopoulos, S. and S. Ahn (1994): "Strong approximation of the quantile processes and its applications under strong mixing properties," *Journal of Multivariate Analysis*, 51, 17–45.

Franguridi, G., B. Gafarov, and K. Wuthrich (2021): "Conditional quantile estimators: A small sample theory," *arXiv:2011.03073*.

Gourieroux, C. and J. Jasiak (2008): "Dynamic quantile models," *Journal of Econometrics*, 147, 198–205.

Gu, J. and S. Ghosal (2008): "Strong approximations for resample quantile processes and application to ROC methodology," *Journal of Nonparametric Statistics*, 20, 229–240.

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

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hosking, J. R. (2007): "Some theory and practical uses of trimmed L-moments," *Journal of Statistical Planning and Inference*, 137, 3024–3039.

Hosking, J. R. M. (1990): "L-Moments: Analysis and Estimation of Distributions Using

Linear Combinations of Order Statistics," *Journal of the Royal Statistical Society: Series B (Methodological)*, 52, 105–124.

Hosking, J. R. M. and J. R. Wallis (1987): "Parameter and Quantile Estimation for the Generalized Pareto Distribution," *Technometrics*, 29, 339–349.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985): "Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments," *Technometrics*, 27, 251–261.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kaji, T. (2019): "Asymptotic Theory of L-Statistics and Integrable Empirical Processes," *arXiv:1910.07572.*

Kerstens, K., A. Mounir, and I. V. De Woestyne (2011): "Non-parametric frontier estimates of mutual fund performance using C- and L-moments: Some specification tests," *Journal of Banking and Finance*, 35, 1190–1201.

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

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

Kulik, R. (2007): "Bahadur–Kiefer theory for sample quantiles of weakly dependent linear processes," *Bernoulli*, 13, 1071–1090.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis (1979): "Probability weighted moments compared with some traditional techniques in estimating Gumbel Parameters and quantiles," *Water Resources Research*, 15, 1055–1064.

Lee, T.-H., A. Ullah, and H. Wang (2017): "The Second-order Bias and MSE of Quantile Estimators," *Unpublished manuscript.*

——— (2018): "The second-order bias of quantile estimators," *Economics Letters*, 173, 143–147.

Li, C., F. Zwiers, X. Zhang, G. Li, Y. Sun, and M. Wehner (2021): "Changes in annual extremes of daily temperature and precipitation in CMIP6 models," *Journal of Climate*, 34, 3441–3460.

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

Mason, D. M. (1984): "Weak Convergence of the Weighted Empirical Quantile Process in $L^2(0, 1)$," *Annals of Probability*, 12, 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, 575–595.

Neocleous, T. and S. Portnoy (2008): "On monotonicity of regression quantile functions," *Statistics & Probability Letters*, 78, 1226–1229.

Newey, W. K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. K. and D. McFadden (1994): "Chapter 36 Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Elsevier, vol. 4, 2111 – 2245.

Newey, W. K. and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Norvaiša, R. and R. Zitikis (1991): "Asymptotic behaviour of linear combinations of functions of order statistics," *Journal of Statistical Planning and Inference*, 28, 305 – 317.

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

Parzen, E. (1960): *Modern probability theory and its applications*, Wiley.

Phillips, P. C. B. (1991): "A Shortcut to LAD Estimator Asymptotics," *Econometric Theory*, 7, 450–463.

Portnoy, S. (1991): "Asymptotic behavior of regression quantiles in non-stationary, dependent cases," *Journal of Multivariate Analysis*, 38, 100 – 113.

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. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577–91.

——— (1984): "Chapter 15 Approximating the distributions of econometric estimators and test statistics," in *Handbook of Econometrics*, Elsevier, vol. 2, 881–935.

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

Sankarasubramanian, A. and K. Srinivasan (1999): "Investigation and comparison of sampling properties of L-moments and conventional moments," *Journal of Hydrology*, 218, 13–34.

Šimková, T. (2017): "Statistical inference based on L-moments," *Statistika*, 97, 44–58.

Stigler, S. M. (1969): "Linear Functions of Order Statistics," *Annals of Mathematical Statistics*, 40, 770–788.

——— (1974): "Linear Functions of Order Statistics with Smooth Weight Functions," *Annals of Statistics*, 2, 676–693.

Sävje, F., P. M. Aronow, and M. G. Hudgens (2021): "Average treatment effects in the presence of unknown interference," *Annals of Statistics*, 49, 673 – 701.

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

Vaart, A. W. v. d. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Vaart, A. W. V. d. and J. A. Wellner (1996): *Weak convergence and empirical processes*, Springer.

Wang, D. and A. D. Hutson (2013): "Joint confidence region estimation of L-moment ratios with an extension to right censored data," *Journal of Applied Statistics*, 40, 368–379.

Wang, Q. J. (1997): "LH moments for statistical analysis of extreme events," *Water Resources Research*, 33, 2841–2848.

Wooldridge, J. (2010): *Econometric analysis of cross section and panel data*, Cambridge, Mass: MIT Press.

Wu, W. B. and Z. Zhou (2011): "Gaussian approximations for non-stationary multiple time series," *Statistica Sinica*, 21, 1397–1413.

Yang, J., Z. Bian, J. Liu, B. Jiang, W. Lu, X. Gao, and H. Song (2021): "No-Reference Quality Assessment for Screen Content Images Using Visual Edge Model and AdaBoosting Neural Network," *IEEE Transactions on Image Processing*, 30, 6801–6814.

Yoshihara, K. (1995): "The Bahadur representation of sample quantiles for sequences of strongly mixing random variables," *Statistics & Probability Letters*, 24, 299 – 304.

Yu, H. (1996): "A note on strong approximation for quantile processes of strong mixing sequences," *Statistics & Probability Letters*, 30, 1 – 7.

Zwet, W. R. V. (1980): "A Strong Law for Linear Functions of Order Statistics," *Annals of Probability*, 8, 986 – 990.