THE EMPIRICAL SADDLEPOINT ESTIMATOR

BENJAMIN HOLCBLAT AND FALLAW SOWELL

Abstract. We define a moment-based estimator that maximizes the empirical saddlepoint (ESP) approximation of the distribution of solutions to empirical moment conditions. We call it the ESP estimator. We prove its existence, consistency and asymptotic normality, and we propose novel test statistics. We also show that the ESP estimator corresponds to the MM (method of moments) estimator shrunk toward parameter values with lower implied estimated variance, so it reduces the documented instability of existing moment-based estimators. In the case of just-identified moment conditions, which is the case we focus on, the ESP estimator is different from the MM estimator, unlike the more recent alternatives, such as the empirical-likelihood-type estimators.

Keywords: Empirical Saddlepoint Approximation; Method of Moments; Kullback-Leibler Divergence Criterion; Variance Penalization.

1. Introduction

The saddlepoint (SP) approximation has been developed to approximate distributions. Because of its accuracy, it is regularly used in several fields, such as numerical analysis (e.g., Loader (2000)'s algorithm to approximate binomial distributions, and which is notably used in the statistical software R) and actuarial sciences (e.g., Esscher (1932)'s approximation for distributions tails). In statistics and econometrics, the SP approximation and its empirical version —the empirical saddlepoint (ESP) approximation— have been used to approximate finite-sample distributions (e.g., Daniels 1954, Holly and Phillips 1979, Davison and Hinkley 1988). Standard monographs and introductions about the ESP and the SP approximation for statistics include Field and Ronchetti (1990), Kolassa (1994/2006), Jensen (1995), Goutis and Casella (1999) and Butler (2007).

In the present paper, we propose to use the ESP approximation to define a *point* estimator $\hat{\theta}_T$. We call it the ESP estimator, and denote it with $\hat{\theta}_T$. It maximizes the Ronchetti and Welsh (1994)'s ESP approximation of the distribution of solutions to the empirical moment conditions

$$\frac{1}{T}\sum_{t=1}^{T}\psi\left(X_t,\theta\right)=0 \tag{1}$$

where $\psi(.,.)$ denotes the moment function s.t. $\mathbb{E}[\psi(X_1,\theta_0)] = 0_{m\times 1}$ an m-dimensional vector of zeros, $(X_t)_{t=1}^T$ i.i.d. data, $\theta_0 \in \mathbf{\Theta} \subset \mathbf{R}^m$ the unknown parameter of interest, and

Date: December 29, 2021.

T the sample size. In the present paper, for clarity and simplicity, in line with most of the SP literature, we consider the so-called just-identified case, in which the number of parameters is the same as the number of moment conditions.

The ESP estimator is a moment-based estimator. Since Pearson (1894, 1902)'s method of moment (MM), moment-based estimators have been found useful in a variety of applications (e.g., covariance structure analysis in psychology, and asset pricing in economics). Their two main advantages are (i) they do not require a parametric family of probability distributions for the data so they are less prone to model misspecification, and (ii) they allow complex models for which the likelihood function is intractable.

Nevertheless, the increase use of the MM and its extensions has revealed that they can be unstable and perform poorly in finite samples (e.g., July 1996 special issue of Journal of Business & Economics Statistics). The idea of the ESP estimator $\hat{\theta}_T$ is to improve on the MM estimator θ_T^* as follows — note the difference between $\hat{\theta}_T$, which denotes the ESP estimator, and θ_T^*, which denotes the MM estimator or, equivalently, a solution to the empirical moment conditions (1). By definition, the MM estimate $\theta_T^*(\omega)$ solves the empirical moment conditions (1) evaluated for a given sample $(X_t(\omega))_{t=1}^T$, but it typically does not solve the empirical moment condition for another sample $(X_t(\dot{\omega}))_{t=1}^T$ where $\dot{\omega} \neq \omega$. Thus, we might want an estimate that not only takes into account the empirical moment conditions evaluated for the given sample $(X_t(\omega))_{t=1}^T$, but also for other potential samples. More precisely, we want an estimate that accounts for all possible evaluations of the empirical moment conditions according to their probability of occurrence. This is what the ESP estimate does: The ESP estimate maximizes the ESP approximation to the distribution of the solutions to the empirical moment conditions (Ronchetti and Welsh 1994). If the empirical moment conditions (1) have a unique solution with a continuous distribution, the ESP estimator maximizes the ESP approximation of a probability density function of the solution θ_T^*. We rely on the ESP approximation because simulation and theoretical evidence shows the ESP approximation can be very accurate in small sample (e.g., Davison and Hinkley 1988, Ronchetti and Welsh 1994).

We also show that the ESP estimator corresponds to an MM estimator shrunk toward parameter values with lower implied estimated variance of the solution to the corresponding finite-sample moment conditions. More precisely, we decompose the logarithm of the ESP approximation as the sum of a term, which is maximized at the MM estimator, and a variance penalty, which discounts any parameter value $\dot{\theta} \in \mathbf{\Theta}$ that implies a high estimated variance for the corresponding MM estimator. Under assumptions adapted from the entropy literature, we establish the ESP estimator has the same good asymptotic properties as the MM estimator, so the variance penalization is a finite-sample correction. We also derive the ESP counterparts of the Wald, Lagrange multiplier (LM), analogue likelihood-ratio (ALR) test statistics, as well as another test statistic. Then, we investigate the ESP estimator through Monte-Carlo simulations. We compare its performance with the exponential tilting (ET) estimator, which is equal to the MM estimator in the just-identified case. Results show that the variance penalization of the ESP estimator reduces the finite-sample instability of the ET estimator (or equivalently, of the MM estimator). An empirical application illustrates the gain from this greater stability in terms of inference.

The ESP estimator is not the first proposal to improve on the MM and its extensions. Alternative moment-based approaches have been proposed such as the empirical likelihood approach of Owen (Qin and Lawless 1994), the continuously updating approach (Hansen et al. 1996), the already-mentioned exponential tilting (ET) approach (Kitamura and Stutzer 1997, Imbens et al. 1998), and combinations of the aforementioned approaches (e.g., Schennach 2007). All these approaches yield an estimator closely related to the empirical likelihood estimator, so we call them empirical-likelihood-type estimators. In the just-identified case, when well-defined, all of these empirical-likelihood-type estimators are numerically equal to the original Pearson's MM estimator θ_T^*. Because we focus on the just-identified case, it is sufficient for us to compare the ESP estimator with the MM estimator, or with one of any of these more recent estimators.

In addition to the already cited papers, the present paper, which supersedes the unpublished manuscript Sowell (2009), is related to many other ones. We clarify these relations in Section 5 (p. 15). To the best of our knowledge, none of the prior papers use the SP or the ESP to propose a novel moment-based point estimator. Overall, the present paper (i) brings together the literature on the saddlepoint approximation and the literature on moment-based estimation, and (ii) points out the untapped potential of the ESP to tackle issues faced in moment-based estimation.

Remark 1. Another motivation for the ESP estimator is decision theoretic. The ESP estimator follows from the minimization of the expectation of a loss "function" that equals zero when θ solves the empirical moment conditions and one otherwise by normalization. This motivation is similar to the decision-theoretic justification for the Bayesian maximum a posteriori estimator (e.g., Robert 2007/1994, sec. 4.1.2). As in Bayesian analysis, the choice of other loss functions is possible. The investigation of different loss functions is left for future research.

2. Finite-sample analysis

In the present section, we remind the formula for the ESP approximation, and analyze its finite-sample structure. Then, we decompose the log-ESP into two terms and show that the ESP estimator is a MM estimator shrunk toward parameter values with lower implied estimated variance, so the estimation stability is improved.

2.1. **The ESP approximation.**

Formalizing and generalizing prior works (Davison and Hinkley 1988, Feuerverger 1989, Wang 1990, Young and Daniels 1990), Ronchetti and Welsh (1994) propose the following ESP approximation to estimate the distribution of a solution to the empirical moment conditions (1)

$$\hat{f}_{\theta_T^*}(\theta) := \exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}\left(\frac{T}{2\pi}\right)^{m/2}|\Sigma_T(\theta)|_{\det}^{-\frac{1}{2}} \tag{2}$$

where $|.|_{\det}$ denotes the determinant function, θ_T^* a solution to (1), $\psi_t(.) := \psi(X_t,.)$, and

$$\Sigma_T(\theta) := \left[\sum_{t=1}^{T} w_{t,\theta}\frac{\partial\psi_t(\theta)}{\partial\theta'}\right]^{-1}\left[\sum_{t=1}^{T} w_{t,\theta}\psi_t(\theta)\psi_t(\theta)'\right]\left[\sum_{t=1}^{T} w_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]^{-1}, \tag{3}$$

$$w_{t,\theta} := \frac{\exp\left[\tau_T(\theta)'\psi_t(\theta)\right]}{\sum_{i=1}^{T}\exp\left[\tau_T(\theta)'\psi_i(\theta)\right]}, \tag{4}$$

$$\tau_T(\theta) \text{ such that } \sum_{t=1}^{T}\psi_t(\theta)\frac{\exp\left[\tau_T(\theta)'\psi_t(\theta)\right]}{\sum_{i=1}^{T}\exp\left[\tau_T(\theta)'\psi_i(\theta)\right]}\times\frac{1}{T} = 0. \tag{5}$$

The ESP estimator maximizes the ESP approximation (2), or equivalently, its logarithm, which is given in the upcoming formula (7) apart for terms constant w.r.t. (with respect to) θ. The ESP approximation (2) is the empirical counterpart of the SP approximation of Field (1982). From a computational point of view, the ESP approximation (2) is not complicated. See Section 4.1 and Online Appendix E for more details. The only implicit quantity is $\tau_T(\theta)$, which solves the tilting equation (5), which, in turn, is just the FOC (first-order condition) of the unconstrained convex problem $\min_{\tau\in\mathbf{R}^m}\sum_{t=1}^{T}\mathrm{e}^{\tau'\psi_t(\theta)}$. A full understanding of the ESP approximation (2) arguably requires to work through higher-order asymptotic expansions along the lines of Field (1982). However, direct inspection of the ESP approximation (2) also provides insight for how it incorporates information from the data through two channels.

The first channel is the *ET (exponential tilting) term* $\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}$. In equation (5), for any $\theta\in\boldsymbol{\Theta}$, the terms $\frac{\exp[\tau_T(\theta)'\psi_t(\theta)]}{\sum_{i=1}^{T}\exp[\tau_T(\theta)'\psi_i(\theta)]}$ tilt the empirical probability $1/T$, so the finite-sample moment conditions (5) hold. This tilting determines, through equation (4), the multinomial distribution $(w_{t,\theta})_{t=1}^{T}$ that is the closest to the empirical distribution —in the sense of the Kullback-Leibler divergence criterion— s.t. the finite-sample moment conditions (5) holds: The tilting equation (5) is the FOC w.r.t. (with respect to) τ of the Lagrangian dual problem of the minimization problem

$$\min_{(w_{1,\theta},w_{2,\theta},\cdots,w_{T,\theta})\in]0,1]^T}\sum_{t=1}^{T} w_{t,\theta}\log\left(\frac{w_{t,\theta}}{1/T}\right)$$
$$\text{s.t.}\sum_{t=1}^{T} w_{t,\theta}\psi_t(\theta) = 0 \text{ and } \sum_{t=1}^{T} w_{t,\theta} = 1, \tag{6}$$

where $\sum_{t=1}^{T} w_{t,\theta}\log[w_{t,\theta}/(1/T)]$ is the Kullback-Leibler divergence criterion between the empirical distribution and the multinomial distribution $(w_{t,\theta})_{t=1}^{T}$ with the same support (e.g., Csiszár 1975, Efron 1981). Then, for the given $\theta\in\boldsymbol{\Theta}$, in the ESP approximation

(2), the ET term $\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}$ indicates the extent of the tilting needed to set the finite-sample moment conditions (6) (or equivalently, equation (5)) to zero. The bigger is the tilting of the empirical distribution, the less compatible are the data with θ solving the empirical moment conditions, and the smaller should be the ET term $\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}$. It can be easily seen that $\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}$ reaches its maximum when θ is a solution θ_T^* of the empirical moment conditions (1), i.e., when $\tau_T(\theta_T^*) = 0_{m\times 1}$ and no tilting is needed —For a formal proof, one can follow the same reasoning as in the proof of Lemma 10 on p. 33 (Online Appendix) with the empirical distribution in lieu of $\mathbb{P}$.

In the ESP approximation on equation (2), the second term $\left(\frac{T}{2\pi}\right)^{m/2}$ comes from the multivariate Gaussian distribution that is the leading term of the Edgeworth's asymptotic expansions underlying ESP approximations. However, because it is constant w.r.t. θ, it does not affect the maximization of the ESP approximation, so it is *not* an information channel for the ESP estimator. The remaining term $|\Sigma_T(\theta)|_{\det}^{-\frac{1}{2}}$, which we call the *variance term*, is the second channel through which the ESP approximation incorporates information from data. The variance term discounts the ET term according to the tilted estimated variance of the solution to the finite-sample moment conditions. Under standard assumptions, a consistent estimator of the asymptotic variance of $\sqrt{T}(\theta_T^* - \theta_0)$ is $\Sigma_T(\theta_T^*) := \left[\sum_{t=1}^{T} w_{t,\theta_T^*}\frac{\partial\psi_t(\theta_T^*)}{\partial\theta}\right]^{-1}\left[\sum_{t=1}^{T} w_{t,\theta_T^*}\psi_t(\theta_T^*)\psi_t(\theta_T^*)'\right]\left[\sum_{t=1}^{T} w_{t,\theta_T^*}\frac{\partial\psi_t(\theta_T^*)'}{\partial\theta}\right]^{-1}$. The bigger the variance term is, the less plausible a solution takes exactly this value, and the smaller is $|\Sigma_T(\theta)|_{\det}^{-\frac{1}{2}}$ —note the negative power. Therefore, overall, for a given $\theta \in \mathbf{\Theta}$, the bigger the tilting or the estimated variance, the smaller the ESP approximation, i.e., the estimated probability weight $\hat{f}_{\theta_T^*}(\theta)$ that θ solves the empirical moment conditions (1).

2.2. The ESP estimator as a shrinkage estimator.

As explained in the introduction, the more recent moment-based estimators are numerically equal to the Pearson's MM estimator in the just-identified case. Thus, it is sufficient to compare the ESP estimator with one of them in order to understand the difference between the former and the other proposed moment-based estimators. The ET estimator of Kitamura and Stutzer (1997) and Imbens et al. (1998) is particularly convenient for this purpose. Taking the logarithm of the ESP approximation (2), removing the terms constant w.r.t. θ, and dividing by the sample size T, it can be seen that the ESP estimator $\hat{\theta}_T$ maximizes the objective function

$$\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right] - \frac{1}{2T}\ln|\Sigma_T(\theta)|_{\det}, \tag{7}$$

where $\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]$ is a strictly increasing transformation of the objective function of the ET estimator. Thus, the difference between the ESP estimator and the ET estimators comes only from the log-variance term $-\frac{1}{2T}\ln|\Sigma_T(\theta)|_{\det}$. As explained in Section 2.1, the variance term incorporates additional information, which penalizes parameter values with higher implied estimated variance. More precisely, the variance term discounts any parameter value $\dot{\theta} \in \mathbf{\Theta}$ that implies a high estimated variance for the ET —or equivalently, MM— estimator of $\dot{\theta}$ based on the tilted moment condition $\int_{\mathbf{\Omega}} \psi(X_1(\omega), \dot{\theta}) \mathrm{P}_{\dot{\theta}}(\mathrm{d}\omega) = 0_{m\times 1}$, where $\frac{\mathrm{dP}_{\dot{\theta}}}{\mathrm{d}\mathbb{P}} := \frac{\mathrm{e}^{\tau(\dot{\theta})'\psi(X_1,\dot{\theta})}}{\mathbb{E}[\mathrm{e}^{\tau(\dot{\theta})'\psi(X_1,\dot{\theta})}]}$ with $\mathbb{P}$ the physical probability measure. Thus, the ESP estimator is an ET estimator shrunk toward parameter values with lower implied estimated variance. As the following Proposition 1 shows, it immediately implies a smaller estimated variance for the ESP estimator.

Proposition 1 (Shrunk estimated variance of the ESP estimator)**.** *Assume the existence of an ESP estimator $\hat{\theta}_T$ and an ET estimator θ_T^* s.t. $\hat{\theta}_T$ is different from any ET estimator i.e., $\hat{\theta}_T \notin \arg\max_{\theta\in\mathbf{\Theta}} \ln\left[\frac{1}{T}\sum_{t=1}^{T} \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]$. Then*

$$|\Sigma_T(\hat{\theta}_T)|_{\det} < |\Sigma_T(\theta_T^*)|_{\det}.$$

For more details about Proposition 1, see the online Appendix D. The variance shrinkage is desirable because of the documented instability of existing nonlinear moment-based estimators.

3. Asymptotic properties

In the present section, we investigate the asymptotic properties of the ESP estimator. Good asymptotic properties can be regarded as a minimal requirement for the ESP estimator, which is based on a small-sample asymptotic approximation. While the asymptotic properties of the ESP estimator are standard, their complete proofs, which are in the Online Appendix, require the development of some nonstandard arguments.

3.1. **Existence, consistency and asymptotic normality.**

We require the following assumption to prove the existence and the consistency of the ESP estimator.

Assumption 1. **(a)** *The data $(X_t)_{t=1}^{\infty}$ are a sequence of i.i.d. random vectors of dimension p on the complete probability sample space $(\mathbf{\Omega}, \mathcal{E}, \mathbb{P})$.* **(b)** *Let the moment function $\psi : \mathbf{R}^p \times \mathbf{\Theta}^\epsilon \mapsto \mathbf{R}^m$ be s.t. $\theta \mapsto \psi(X_1, \theta)$ is continuously differentiable $\mathbb{P}$-a.s., and $\forall \theta \in \mathbf{\Theta}^\epsilon$, $x \mapsto \psi(x, \theta)$ is $\mathcal{B}(\mathbf{R}^p)/\mathcal{B}(\mathbf{R}^m)$-measurable, where, for $\epsilon > 0$, $\mathbf{\Theta}^\epsilon$ denotes the ϵ-neighborhood of $\mathbf{\Theta}$, and $\mathcal{B}(\mathbf{R}^p)$ the Borel σ-algebra on $\mathbf{R}^p$.* **(c)** *In the parameter space $\mathbf{\Theta}$, there exists a unique $\theta_0 \in \mathrm{int}(\mathbf{\Theta})$ s.t. $\mathbb{E}[\psi(X_1, \theta_0)] = 0_{m\times 1}$.* **(d)** *Let the parameter space $\mathbf{\Theta} \subset \mathbf{R}^m$ be a compact set, s.t., for all $\theta \in \mathbf{\Theta}$, there exists $\tau(\theta) \in \mathbf{R}^m$ that solves the equation $\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1, \theta)\right] = 0$ for τ.* **(e)** *$\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}^\epsilon} \mathrm{e}^{2\tau'\psi(X_1,\theta)}\right] < \infty$ where $\mathbf{S} := \{(\theta, \tau) : \theta \in \mathbf{\Theta}\ \&\ \tau \in \mathbf{T}(\theta)\}$ and $\mathbf{T}(\theta) := \overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))}$ with $\overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))}$ the closed ball of radius $\epsilon_{\mathbf{T}} > 0$ and center $\tau(\theta)$.* **(f)** *$\mathbb{E}\left[\sup_{\theta\in\mathbf{\Theta}} |\frac{\partial\psi(X_1,\theta)}{\partial\theta'}|^2\right] < \infty$, where $|.|$ denotes the Euclidean norm.* **(g)** *$\mathbb{E}\left[\sup_{\theta\in\mathbf{\Theta}^\epsilon} |\psi(X_1, \theta)\psi(X_1, \theta)'|^2\right] < \infty$.* **(h)** *For all $\theta \in \mathbf{\Theta}$, $\Sigma(\theta) := \left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)}{\partial\theta'}\right]^{-1} \mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1, \theta)\psi(X_1, \theta)'\right]\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]^{-1}$ is invertible.*

We require the following additional assumption to prove the asymptotic normality of the ESP estimator.

Assumption 2. **(a)** *The function $\theta \mapsto \psi(X_1, \theta)$ is three times continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 in $\mathbf{\Theta}$ $\mathbb{P}$-a.s.* **(b)** *There exists a $\mathcal{B}(\mathbf{R}^p)/\mathcal{B}(\mathbf{R})$-measurable function $b(.)$ satisfying $\mathbb{E}\left[\sup_{\theta\in\mathcal{N}} \sup_{\tau\in\mathbf{T}(\theta)} \mathrm{e}^{k_1\tau'\psi(X_1,\theta)} b(X_1)^{k_2}\right] < \infty$ for $k_1 \in [\![1, 2]\!]$ and $k_2 \in [\![1, 4]\!]$ s.t., for all $j \in [\![0, 3]\!]$, $\sup_{\theta\in\mathcal{N}} |\nabla^j \psi(X_1, \theta)| \leqslant b(X_1)$ where $\nabla^j \psi(X_1, \theta)$ denotes a vector of all partial derivatives of $\theta \mapsto \psi(X_1, \theta)$ of order j, and $[\![a, b]\!] := [a, b] \cap \mathbf{Z}$ for all $(a, b) \in \mathbf{R}^2$.*

Assumptions 1 and 2 are stronger than the usual assumptions in the MM literature, but are similar to assumptions used in the entropy literature and related literatures. Assumptions 1 and 2 are essentially adapted from Haberman (1984), Kitamura and Stutzer (1997), and Schennach (2007, Assumption 3). See also Chib et al. (2018) for similar assumptions. Appendix B.1 contains a detailed discussion of Assumptions 1 and 2. Under Assumptions 1 and 2, the following theorem establishes the existence, the strong consistency, and the asymptotic normality of the ESP estimator $\hat{\theta}_T$.

Theorem 1 (Existence, consistency and asymptotic normality)**.** *Under Assumption 1, $\mathbb{P}$-a.s. for T big enough, there exists $\hat{\theta}_T$ s.t.*

(i) *$\mathbb{P}$-a.s. as $T \to \infty$, $\hat{\theta}_T \to \theta_0$; and*
(ii) *under the additional Assumption 2, as $T \to \infty$, $\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{D} \mathcal{N}(0, \Sigma(\theta_0))$.*

where $\Sigma(\theta_0) := \left[\mathbb{E}\frac{\partial\psi(X_1,\theta_0)}{\partial\theta'}\right]^{-1} \mathbb{E}\left[\psi(X_1, \theta_0)\psi(X_1, \theta_0)'\right] \left[\mathbb{E}\frac{\partial\psi(X_1,\theta_0)'}{\partial\theta}\right]^{-1}$, $\xrightarrow{D}$ denotes the convergence in distribution.

Theorem 1 shows the variance penalization $-\frac{1}{2T} \ln |\Sigma_T(\theta)|_{\det}$ vanishes sufficiently quickly asymptotically, so it does not distort the first-order asymptotic of the estimator. In particular, Theorem 1(ii) shows that the ESP estimator reaches the same semiparametric efficiency bound as the MM and the more recent moment-based estimators (Chamberlain 1987, Carrasco and Florens 2014). Parts of the proofs of Theorem 1 are involved, although the proofs strategies follow traditional approaches. The proof of existence follows the Schmetterer and Jennrich approach (Schmetterer 1966 Chap. 5; Jennrich 1969), with an additional complication coming from the implicit nature of the function $\theta \mapsto \tau_T(\theta)$. The proof of Theorem 1(i) (i.e., consistency) follows Wald's approach to consistency (Wald 1949). The usual consistency approach for empirical-likelihood-type estimators (e.g., Newey and Smith 2004, Smith 2011) cannot be easily followed here because of the variance term. Moreover, the later consistency approach does not articulate well with an existence proof. The basic idea of our consistency proof is to show that, $\mathbb{P}$-a.s. for T big enough, the ESP estimator maximizes the LogESP function (7), where, $\mathbb{P}$-a.s. as $T \to \infty$,

$$\sup_{\theta\in\mathbf{\Theta}} \left| \ln\left[\frac{1}{T}\sum_{t=1}^{T} \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right] - \ln \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}] \right| = o(1) \text{ , and}$$
$$\sup_{\theta\in\mathbf{\Theta}} \left| \frac{1}{2T} \ln |\Sigma_T(\theta)|_{\det} \right| = O(T^{-1}). \tag{8}$$

The two main complications w.r.t. the proofs available in the entropy literature are the following. First, the need to ensure that, for T big enough, for all $\theta \in \mathbf{\Theta}$, $|\Sigma_T(\theta)|_{\det}$ is bounded away from zero, so that the LogESP function (7) does not diverge on parts of the parameter space. Second, we establish that the joint parameter space for θ and τ (i.e., $\mathbf{S}$) is a compact set. For the latter purpose, we appropriately modify assumptions from the entropy literature and develop a proof based on set-valued analysis. The proof of Theorem 1(ii) (i.e., asymptotic normality) follows the traditional approach of expanding the FOCs. The two main complications w.r.t. the proofs in the entropy literature are the following. First, instead of expanding the exact FOC $\left.\frac{\partial \ln[\hat{f}_{\theta_T^*}(\theta)]}{\partial\theta}\right|_{\theta=\hat{\theta}_T}$, we expand an approximate FOC combined with the FOC (5) for τ. This is technical because it involves differentiation of the log variance term. Second, we control the asymptotic behaviour of the derivatives that come from the log-variance term $\ln |\Sigma_T(\theta)|_{\det}$. Another shorter proof approach based on an approximate FOC of the first term of the log-ESP is possible. We do not follow it because it would complicate and lengthen the presentation and the proof of the upcoming Theorem 2.

3.2. **More on inference: The trinity+1.**

The ESP estimator provides different ways to test parameter restrictions

$$\mathrm{H}_0 : r(\theta_0) = 0_{q\times 1} \tag{9}$$

where $r : \mathbf{\Theta} \to \mathbf{R}^q$ with $q \in [\![1, \infty[\![$. More precisely, within the ESP framework, there exist the usual trinity of Wald, LM and ALR tests statistics, plus another test statistic, which we call the Tilt test statistic.

In addition to Assumptions 1 and 2, we require the following standard and mild assumption to establish the asymptotic distribution of the Wald, LM, ALR, and Tilt statistics.

Assumption 3 (For the trinity+1). **(a)** *The function $r : \mathbf{\Theta} \to \mathbf{R}^q$ in the null hypothesis* (9) *is continuously differentiable.* **(b)** *The derivative $R(\theta) := \frac{\partial r(\theta)}{\partial\theta'}$ is full rank at θ_0.*

Under Assumptions 1, 2 and 3, the following theorem shows that the Wald, LM ALR, and Tilt statistics asymptotically follow a chi-squared distribution with q degrees of freedom.

Theorem 2 (The trinity+1: Wald, LM, ALR and Tilt tests)**.** *Define* $R(\theta) := \frac{\partial r(\theta)}{\partial \theta'}$*, and the following Wald, LM, ALR and Tilt test statistics*

$$\begin{aligned}
\mathrm{Wald}_T &:= T r(\hat{\theta}_T)'[R(\hat{\theta}_T)\widehat{\Sigma(\theta_0)}_T R(\hat{\theta}_T)']^{-1} r(\hat{\theta}_T) \\
\mathrm{LM}_T &:= T\check{\gamma}_T'[R(\check{\theta}_T)\widehat{\Sigma(\theta_0)}_T R(\check{\theta}_T)']\check{\gamma}_T = \frac{1}{T}\frac{\partial \ln[\hat{f}_{\theta_T^*}(\check{\theta}_T)]}{\partial \theta'}\widehat{\Sigma(\theta_0)}_T \frac{\partial \ln[\hat{f}_{\theta_T^*}(\check{\theta}_T)]}{\partial \theta} \\
\mathrm{ALR}_T &:= 2\{\ln[\hat{f}_{\theta_T^*}(\hat{\theta}_T)] - \ln[\hat{f}_{\theta_T^*}(\check{\theta}_T)]\} \\
\mathrm{Tilt}_T &:= T\tau_T(\check{\theta}_T)'\widehat{V}_T \tau_T(\check{\theta}_T)
\end{aligned}$$

where $\widehat{\Sigma(\theta_0)}_T$ *and* $\widehat{V}_T$ *are symmetric matrices that converge in probability to* $\Sigma(\theta_0)$ *and* $\mathbb{E}[\psi(X_1,\theta_0)\psi(X_1,\theta_0)']$*, respectively; and where* $\check{\gamma}_T$ *and* $\check{\theta}_T$ *respectively denote the Lagrange multiplier and a solution to the maximization of* $\hat{f}_{\theta_T^*}(\theta)$ *w.r.t.* $\theta \in \mathbf{\Theta}$ *under the constraint that* $r(\theta) = 0_{q\times 1}$*, i.e.,* $\check{\theta}_T \in \arg\max_{\theta\in\check{\Theta}} \hat{f}_{\theta_T^*}(\theta)$ *with* $\check{\Theta} := \{\theta \in \Theta : r(\theta) = 0_{q\times 1}\}$ *and* $\check{\gamma}_T$ *s.t.* $\frac{1}{T}\frac{\partial \ln[\hat{f}_{\theta_T^*}(\check{\theta}_T)]}{\partial\theta} + \frac{\partial r(\check{\theta}_T)'}{\partial\theta}\check{\gamma}_T = 0_{m\times 1}$*. Under Assumptions 1, 2 and 3, if the test hypothesis (9) holds, as* $T \to \infty$*,*

$$\mathrm{Wald}_T, \mathrm{LM}_T, \mathrm{ALR}_T, \mathrm{Tilt}_T \xrightarrow{D} \chi_q^2.$$

Theorem 2 can also be used to obtain valid confidence regions by the inversion of the test statistics with $\check{\theta}_T = \theta_0$. Our Wald, LM, ALR and Tilt test statistics share some similarity with the test statistics proposed in Kitamura and Stutzer (1997), Imbens et al. (1998), and Robinson et al. (2003). As explained in Section 2, the difference between our test statistic and aforementioned test statistics come from the variance term, which affects both the objective function and the (possibly constrained) estimator. Thus, an inspection of the formulas for the test statistics shows the LM and ALR test statistics are the most different from their ET counterpart, and that the ALR test statistic exploits more the variance term than the LM test statistic—a derivative of a function contains less information than the function. Moreover, as usual, LM test statistics should be avoided in non linear setting because of local extrema. The proof of Theorem 2 follows the traditional proof strategy for deriving the trinity. The main complications w.r.t. the proofs available in the entropy literature are the same as for the proof of Theorem 1(ii). Note that the uniform convergences (8) of the two parts of the ESP objective function combined with results from the entropy literature do not imply Theorem 2 because the trinity+1 test statistics are scaled by T, and $T\frac{1}{2T}\ln|\Sigma_T(\theta)|_{\det} \neq o(1)$, $\mathbb{P}$-a.s. as $T \to \infty$.

Remark 2. As a referee noted, among Kitamura and Stutzer (1997), Imbens et al. (1998), and Robinson et al. (2003), the latter is unique to establish a relative error of order $O(T^{-1})$ for the SP approximation. A relative error of lower magnitude is especially useful for improving accuracy in distributions tails, which are of particular interest for testing.

4. Examples

In the present section, we further investigate and illustrate the finite-sample properties of the ESP estimator. We focus on the comparison with the ET estimator, as previously noted, (i) in the just-identified case, which is the case addressed in the present paper, the MM estimator and the more recent moment-based estimators are equal to the ET estimator so there is no loss of generality in terms of point estimation, and (ii) the ESP objective function nests the ET objective function, so that the source of the difference between the two is easily understood —it necessarily comes from the variance term (see Section 2.2). In order to provide some finite-sample evidence for the test statistics, we also report their actual rejection probabilities. More information and simulation results are presented in Appendix E.

Remark 3. In addition to our finite-sample analysis of the ESP objective function (Section 2), our derivation of the first-order asymptotic properties (Section 3), our Monte-Carlo simulations and empirical application (present section), another way to shed light on the finite-sample properties of the ESP estimator would be to derive its higher-order asymptotic properties such as its second-order bias (e.g., Rilstone et al. 1996). However, higher-order asymptotic properties of shrinkage estimators are typically complex to study, and have often remained an open problem. The ESP estimator appears to be no exception among shrinkage estimators. Our preliminary derivations yield a long and complicated structure for the second-order bias, from which we struggle to gain insight. The length and the complexity of the second-order bias mainly comes from (i) the derivatives of the variance $|\Sigma_T(\theta)|_{\det}^{-1/2}$; and (ii) the reliance on the exact FOCs instead of approximate FOCs. A mild preview of this complexity can be seen in the proof of asymptotic normality.

4.1. **Numerical example : Monte-Carlo simulations.**

4.1.1. *ET and ESP estimators for the two-parameter Hall and Horowitz model.* We simulate a two-parameter just-identified version of the Hall and Horowitz (1996) model, which has become a standard benchmark to compare the performance of moment-based estimators in statistics (e.g., Schennach 2007, Lô and Ronchetti 2012) and econometrics (e.g., Imbens et al. 1998, Kitamura 2001). This model can be interpreted as a simplified consumption-based asset pricing model where β is the relative risk aversion (RRA) parameter (Gregory et al. 2002). In the simulations, we estimate the two parameters (μ, β) with the moment function

$$\psi_t(\beta,\mu) = \begin{bmatrix} \exp\{\mu - \beta(X_t + Y_t) + 3Y_t\} - 1 \\ Y_t\left(\exp\{\mu - \beta(X_t + Y_t) + 3Y_t\} - 1\right) \end{bmatrix}$$

where $\mu_0 = -.72$, $\beta_0 = 3$, and X_t and Y_t are jointly i.i.d. random variables with distribution $\mathcal{N}(0, .16)$. The parameters are set to the usual values in the literature, see e.g., Hall and Horowitz (1996), Gregory et al. (2002), Schennach (2007) and Lô and Ronchetti (2012). The Hall and Horowitz (1996) model is known to be challenging to estimate because it induces some instability for usual moment-based estimators in small samples. This kind of instability has been observed in several empirical applications.

TABLE 1. **ESP vs. ET estimator for the two-parameter Hall and Horowitz model.**

T		β		μ	
		ET	ESP	ET	ESP
25	MSE	3.6729	0.6644	1.5633	0.2356
	Bias	0.4761	-0.0127	-0.1955	0.1042
	Var.	3.4462	0.6642	1.5250	0.2247
50	MSE	1.5685	0.3359	0.8974	0.1234
	Bias	0.2602	-0.0128	-0.1232	0.0638
	Var.	1.5008	0.3358	0.8822	0.1193
100	MSE	0.6653	0.1589	0.4375	0.0611
	Bias	0.1490	-0.0085	-0.0770	0.0355
	Var.	0.6431	0.1589	0.4316	0.0599
200	MSE	0.2361	0.0823	0.1744	0.0312
	Bias	0.0633	-0.0158	-0.0314	0.0250
	Var.	0.2321	0.0821	0.1734	0.0305
500	MSE	0.0435	0.0322	0.0198	0.0134
	Bias	0.0229	-0.0096	-0.0102	0.0115
	Var.	0.0430	0.0322	0.0197	0.0133
1000	MSE	0.0243	0.0184	0.0115	0.0071
	Bias	0.0137	-0.0038	-0.0060	0.0060
	Var.	0.0241	0.0184	0.0115	0.0071
5000	MSE	0.0040	0.0038	0.0016	0.0015
	Bias	0.0016	-0.0021	-0.0005	0.0021
	Var.	0.0040	0.0038	0.0016	0.0015

Note: The reported statistics are based on 10,000 simulated samples of sample size equal to the indicated T. For ET, the parameter space is restricted to $\beta < 15$ in order to limit the erratic behaviour of the estimator at sample sizes $T = 25$ and 50. No such parameter restriction is imposed for ESP.

Table 1 reports the mean-squarred error (MSE), bias and variance of the ESP and ET estimators for different sample sizes. The MSE of the ESP estimator is always smaller than for the ET estimator, and the differences are notable for small sample sizes. The decomposition of the MSE as the sum of the variance and the squared bias indicates that the variance contributes more than the bias to the reduction of the MSE for the ESP estimator. Note also that Table 1 understates the improvement delivered by the variance penalization of the ESP objective function. We helped the ET estimator (or equivalently, the MM estimator), by restricting its parameter space to $\beta < 15$. Without this parameter restriction, the behaviour of the ET estimator is very unstable for sample sizes below 100. An analysis of the typical shape of the objective functions for small sample size explains this phenomenon. The typical ET objective function has a ridge that follows from around the population parameter values ($\beta_0 = 3$, $\mu_0 = -.72$) towards $(1000, -600)$. The ridgeline is not totally flat, and it often has a gentle downward slope as we move away from the area near the population parameter values. However, regularly, for some simulated samples, the very top of the ridge is extremely far from the population parameter values, so that ET estimates are very far from the population parameter values. This does not happen for the ESP estimator. The variance term of the ESP objective function ensures that the ridge drops sufficiently as we move away from the maximum that is near the population parameter value. Thus, in line with our finite-sample analysis of the ESP objective function (Section 2.2), the ESP estimator is much more stable.

4.1.2. *ET and ESP estimators for a stochastic volatility model.* The stochastic lognormal volatility model has been a competitor for the GARCH model. The system evolves as

$$\ln\left(\sigma_t^2\right) = w + \beta \ln\left(\sigma_t^2\right) + \sigma_u U_t$$

and

$$Y_t = \sigma_t Z_t$$

where U_t and Z_t are jointly i.i.d. random variables with distribution $\mathcal{N}(0,1)$. The model has been used to compare moment-based estimators (e.g., Andersen and Sørensen 1996, Lô and Ronchetti 2012).

TABLE 2. **ESP vs. ET estimator for the two-parameter stochastic volatility model**.

T		w		σ_u	
		ET	ESP	ET	ESP
	MSE	0.0015	0.0016	0.0615	0.0187
25	Bias	-0.0048	-0.0069	-0.1549	-0.0798
	Var.	0.0015	0.0015	0.0375	0.0123
	MSE	0.0010	0.0010	0.0420	0.0158
50	Bias	-0.0009	-0.0022	-0.1214	-0.0709
	Var.	0.0010	0.0010	0.0272	0.0108
	MSE	0.0006	0.0006	0.0226	0.0113
100	Bias	0.0001	-0.0005	-0.0814	-0.0558
	Var.	0.0006	0.0006	0.0160	0.0082
	MSE	3e-04	3e-04	1e-02	7e-03
200	Bias	0.0005	0.0003	-0.0471	-0.0384
	Var.	0.0003	0.0003	0.0078	0.0055

Note: The reported statistics are based on 10,000 simulated samples of sample size equal to the indicated T.

As documented in Andersen and Sørensen (1996), the joint estimation of β, w and σ_u yield numerical convergence problem. Thus, we fix $\beta = .95$ and estimate the two parameters (w, σ_u) with the moment function

$$\psi_t(w, \sigma_u) = \begin{bmatrix} |Y_t| - \sqrt{\frac{2}{\pi}} \exp\left\{\frac{w}{2(1-.95)} + \frac{\sigma_u^2}{8(1-.95^2)}\right\} \\ Y_t^2 - \exp\left\{\frac{w}{(1-.95)} + \frac{\sigma_u^2}{2(1-.95^2)}\right\} \end{bmatrix}.$$

We simulate the model for $(w_0, \sigma_{u,0}) = (-0.368, 0.260)$ in order to match the middle case considered in Andersen and Sørensen (1996).

Table 2 shows that the ESP MSE are either similar to, or smaller than, the ET MSE. For the w parameter, beyond the similarity in terms of MSE, the biases are slightly different: The ESP bias is slightly bigger. For the σ_u parameter, the ESP MSE, bias and variance are always smaller, although the difference is never big. Overall, the results indicate that the ET and the ESP estimators perform similarly, when the first one already performs well.

4.1.3. *Test statistics for the two-parameter Hall and Horowitz model.* In the present section we investigate the finite-sample behavior of the trinity+1. We simulate again the two-parameter Hall and Horowitz model, and study the actual rejection probabilities of the test statistics for the null hypothesis $\mathrm{H}_0 : \beta = 3$ and $\mu = -.72$, i.e., one minus the actual coverage probability. We do not report the actual rejection probability for the LM test, for which the asymptotic results of Theorem 2 provide poor finite sample guidance. As previously mentioned, LM tests are typically unreliable in nonlinear setting because of local extrema. For comparison, we also report the actual rejection probabilities for the ET ALR test statistic.

Table 3 presents the results. The performances of the different test statistics are comparable in terms of actual rejection probabilities, although the ESP ALR_T and the ESP Tilt_T seem to perform slightly better. The closer is the actual rejection probabilities to the nominal size $\alpha = .05$ the more accurate is the asymptotic approximation provided by Theorem 2.

TABLE 3. **Actual rejection probabilities for the two-parameter Hall and Horowitz model.**

T	ESP ALR_T	ET ALR_T	ESP Wald_T	ESP Tilt_T
50	0.1996	0.1964	0.2332	0.2016
100	0.1608	0.1622	0.1831	0.1722
200	0.1265	0.1278	0.1419	0.1361
1000	0.0765	0.0783	0.0839	0.0821
2000	0.0669	0.0681	0.0688	0.0676

Note: Under the null hypothesis $\mathrm{H}_0 : \theta = 3$ and $\mu = -.72$, asymptotically the test statistics follows a chi-square distribution with two degree of freedom. The tests used the critical value with size of $\alpha = .05$. The probabilities are based on 10,000 simulated samples of sample size equal to the indicated T.

4.2. **Empirical example.** In this section, we present an empirical example from asset pricing. In empirical consumption-based asset pricing, the literature has found little common ground about the value of the RRA of the representative agent: In most studies, point estimates from economically similar moment conditions are generally outside of each other's confidence intervals. The present section revisits the estimation of the RRA, which goes back to Hansen and Singleton (1982). The popularity of moment-based estimation in consumption-based asset pricing, and more generally in economics is due to the fact that moment-based estimation does not necessarily require the specification of a family of distributions for the data (e.g. Hansen 2013, sec. 3). Typically, an economic model does not imply such family of distributions, except for tractability reasons. Imposing a family of distributions makes it difficult to disentangle the part of the inference results due to the empirical relevance of the economic model from the part due to these additional restrictions. Under regularity conditions, assuming a distribution corresponds to imposing an infinite number of extra moment restrictions (e.g., Feller 1966/1971, chap. VII, sec. 3).

In order to estimate the RRA θ, we rely on the following moment condition

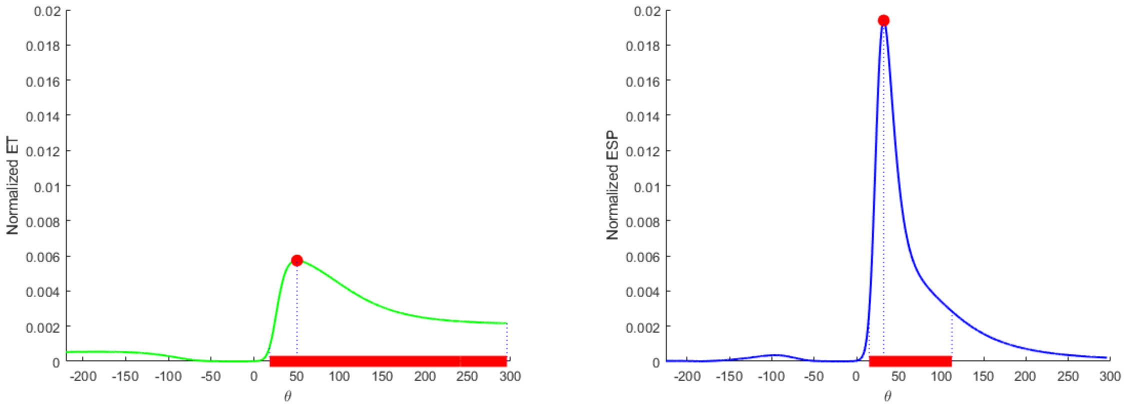
$$\mathbb{E}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0, \tag{10}$$

where $\frac{C_t}{C_{t-1}}$ is the growth consumption and $(R_{m,t}-R_{f,t})$ the market return in excess of the risk-free rate. The moment condition condition (10) and the data, which correspond to standard US data at yearly frequency from Shiller's website spanning from 1890 to 2009, are similar to Julliard and Ghosh (2012). The moment condition (10) has several advantages. Firstly, it is as consistent with Lucas (1978) as with more recent consumption-based asset-pricing models, such as Barro (2006) or Gabaix (2012). In other words, despite its simplicity it also correspond to sophisticated models, and it allows us to obtain estimates that are robust to different variations of consumption-based asset pricing theory. Secondly, without loss of generality, it does not require to estimate the time discount rate, about which there is little debate: The time discount rate of the representative agent is consistently found to be between .9 and 1.

In some of the more recent literature, it has been common to use other moment conditions with a separate parameter for the so-called intertemporal elasticity of substitution, i.e., use Epstein-Zin-Weil preferences (e.g. Epstein and Zin 1991). However, Bommier et al. (2012, 2017) show that such a specification makes the economic interpretation of the parameters difficult. In particular, they show that an increase of the so-called RRA parameter does not yield a behaviour that would be considered more risk averse (Bommier et al. 2012) E.g., All other things being equal, savings can be a decreasing function of the so-called RRA parameter for an agent with Epstein-Zin-Weil preferences (e.g., Bommier et al. 2017, sec. 6). This difficulty of interpretation comes from a violation of the monotonicity axiom according to which an agent does not choose an action if another available action is preferable in every state of the world.

In light of Section 2, we report ET and ESP estimates as well as confidence regions based on the inversion of the ALR test statistics of Theorem 2 (p. 8) with $\check{\theta}_T=\theta_0$. The latter have the advantage to better take into account the whole shape of the objective function than the other confidence regions such as the Wald-based (i.e., t-statistics-based)

Figure 1. **ET vs. ESP inference (1890–2009)**



(A) ET est. and ALR conf. region. (B) ESP est. and ALR conf. region.

Note: The empirical moment condition is $\frac{1}{2009-1889}\sum_{t=1890}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0$, where $R_{m,t}:=$ gross market return, $R_{f,t}:=$risk-free asset gross return, $C_t:=$ consumption, and $\theta:=$relative risk aversion. Normalized ET:$=\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(.)'\psi_t(.)}\right]\right\}/\int_\Theta\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}\mathrm{d}\theta$ and Normalized ESP:$=\hat{f}_{\theta_T^*}(.)/\int_\Theta\hat{f}_{\theta_T^*}(\theta)\mathrm{d}\theta$. The ET and MM estimate are $\hat{\theta}_{\mathrm{ET},T}=\hat{\theta}_{\mathrm{MM},T}=50.3$ (bullet on (A)). The ESP estimate is $\hat{\theta}_{\mathrm{ESP},T}=32.21$ (bullet on (B)). The 95% ET ALR confidence region is $[18.3, 289.0]$ (stripe on (A)). The 95% ESP ALR confidence region is $[15.0, 112.7]$ (stripe on (B)).

confidence regions, which only account for the shape of the objective function in a neighborhood of the estimate through its standard errors.

In Figure 1, (A) and (B) respectively displays the ET term and the ESP approximation. For ease of comparison, the scale is the same, and we normalize both of them so they integrate to one. The normalized ET term is much flatter around its maximum than the normalized ESP approximation. Flatness of the objective function around the estimate has been documented for other existing moment-based estimators, and it has often been regarded as one of the main sources of the instability of the RRA estimates (e.g., Hall 2005, p. 60-64). Figure (B) shows that the normalized ESP is sharp around the ESP estimator. The relative sharpness of the ESP yields sharper confidence regions: The ESP confidence region is less than half its ET counterpart. In light of the variance penalization term in the ESP objective function (Section 2.2 on p. 5) and the shrinkage-like behavior of the ESP estimator in the Monte-Carlo simulations (Section 4.1), the relative sharpness of the ESP inference is not surprising. In Appendix F on p. 39, additional empirical evidences corroborate the increased stability and precision of the ESP estimator w.r.t. the ET estimator (or equivalently, MM estimator).

5. Connection to the literature and further research directions

The present paper demonstrates a previously unknown connection between the SP approximation and moment-based estimation, and hence it is related to many papers in these literatures on top of the ones already cited. Following Daniels (1954), the literature in statistics (e.g., Easton and Ronchetti 1986, Spady 1991, Jensen 1992, Vecchia et al.

2012, Broda and Kan 2015, Fasiolo et al. 2018) and econometrics (e.g., Phillips 1978, Holly and Phillips 1979, Phillips 1982, Lieberman 1994, Aït-Sahalia and Yu 2006) has used the SP (saddlepoint) and ESP approximations to obtain accurate approximations of distributions, especially in the tails. The strand of the SP literature that is closest to our paper derives SP approximations to the distribution of statistics that correspond to solutions of nonlinear estimating equations. The latter strand of literature started with Field (1982) and continued with Skovgaard (1990), Monti and Ronchetti (1993), Imbens (1997), Jensen and Wood (1998), Almudevar et al. (2000), Robinson et al. (2003), and Ronchetti and Trojani (2003), among others. More recently, Czellar and Ronchetti (2010), Ma and Ronchetti (2011), and Lô and Ronchetti (2012), Kundhi and Rilstone (2013, 2015) propose more accurate tests for indirect inference, functional measurement error models, moment condition models, nonlinear estimators and GEL (generalized empirical likelihood) estimators, respectively. To the best of our knowledge, unlike the present paper, none of the prior papers use the SP or the ESP to develop an estimation method that yields a novel moment-based estimator. In ongoing work, we generalize the ESP approximation to the over-identified case and time-dependent data.

The present paper is also related to a large and growing literature on shrinkage estimators. Following Stein (1956)'s example, shrinkage has emerged as a powerful idea to develop more stable estimation methods. Examples include the ridge regression (Hoerland and Kennard 1970), the LASSO regression (Tibshirani 1996), the SCAD penalization (Fan 1997, Fan and Li 2001), and the elastic net penalization (Zou and Hastie 2005). Some of the latter have been adapted and extended to moment-based estimation (e.g., Caner 2009). While the ESP estimator can be regarded as a shrinkage estimator (Section 2.2), it has several particularities. Firstly, unlike the aforementioned shrinkage estimators, the ESP estimator does not require the calibration of tuning parameters, which is often delicate (e.g., Sengupta and Sowell 2020). Secondly, the ESP estimator does not require the user to choose a parameter value. The ESP estimator is *not* shrunk toward a user-chosen shrinkage value, but toward parameter values with lower estimated implied variance. Such a data-driven determination of the shrinkage value is particularly convenient for nonlinear moment-based estimation: While in regression models the choice of a shrinkage value is often easy to justify —e.g., zero, which corresponds to a more parsimonious model—, the choice is often more difficult for nonlinear moment-based estimation. E.g., in the numerical and the empirical example, it is unclear why one would like to shrink the risk-aversion parameter toward zero. Thirdly, the shrinkage nature of the ESP estimator is a consequence of defining an estimator that maximizes the ESP approximation of the finite-sample distribution of solutions to empirical moment conditions. It is not the consequence of the addition of an ad hoc penalization as is sometimes the case for shrinkage estimators.

As hinted in Remark 1 (p. 3), the present paper is additionally related to Bayesian inference. Like several widely-used shrinkage estimators (e.g., Hoerland and Kennard 1970, sec. 6; Tibshirani 1996, sec. 5), the ESP estimator has connections to Bayesian inference. For example, the variance term of the ESP approximation share some similarities with the Jeffreys' prior used in parametric Bayesian inference. The investigation of these connections are left for future research. In a companion paper, we investigate the asymptotic connection between the ESP approximation and Bayesian posterior distributions.

Finally, the present paper is related to the econometric weak instrument literature, which is also motivated by the poor finite-sample stability and performance of usual moment-based estimators (e.g., see Introduction in Stock and Wright 2000, which is the seminal paper of the literature). Despite a common motivation, there are major differences with respect to the present paper. Firstly, by definition, the weak instrument approach requires to assume that the moment conditions depend on the sample size T (Stock and Wright 2000, Assumption C). In several applications (e.g., the empirical example in Section 4.2), this definitional assumption is incompatible with the model of interest. As Hall (2005, p. 296) explains in his standard textbook on moment-based estimation, this definitional assumption is "artificial" in the sense that nobody seems to believe that economic and financial data induce moment conditions depending on the sample size T in this way: This is just a "mathematical device" that is used to derive an asymptotic theory that aims at providing good approximations to finite sample behaviours. See also Stock and Wright (2000, p. 1060-1061 and footnote 3) for a similar justification of the definitional assumption. In contrast, no modification of the moment conditions is required for the ESP estimator. The idea is simply to define an estimator that maximizes an accurate approximation of the finite-sample distribution of the solutions to the empirical moment conditions. The relative stability and sharpness of the ESP objective function in both the numerical and the empirical examples illustrate the usefulness of the idea. Secondly, unlike the present paper, the weak instrument literature does not provide new estimators, but only novel test statistics. Actually, moment-based estimators are generally inconsistent for weakly identified parameters (Stock and Wright 2000, Theorem 1; Guggenberger and Smith 2005, Theorem 2). Thirdly, test statistics derived under the weak instrument assumption induce confidence regions that can be empty (e.g., Stock and Wright 2000, Table IV-VI), and that have infinite length with positive probability (Dufour 1997), so they can be "unreasonable" (Müller and Norets 2016). In contrast to Anderson-Rubin-type statistics used in the weak instrument literature (Stock and Wright 2000), the test statistics of the present paper enjoy the same properties as the traditional trinity statistics, and thus do not yield empty confidence regions. Finally, note that, when, in applications, general test statistics robust to weak instrument appear necessary, it should be possible to extend the present paper for this purpose.

References

Aït-Sahalia, Y. and Yu, J.: 2006, Saddlepoint approximations for continuous-time Markov processes, *Journal of Econometrics* **134**(2), 507–551.

Aliprantis, C. D. and Border, K. C.: 2006/1999, *Infinite Dimensional Analysis. A Hitchhiker's Guide*, third edition edn, Springer.

Almudevar, A., Field, C. and Robinson, J.: 2000, The density of multivariate M-estimates, *The Annals of Statistics* **28**(1), 275–297.

Anatolyev, S. and Gospodinov, N.: 2011, *Methods for estimation and inference in modern econometrics.*, CRC Press.

Andersen, T. G. and Sørensen, B. E.: 1996, GMM estimation of a stochastic volatility model: A Monte-Carlo study, *Journal of Business & Economic Statistics* **14**(3), 328–352.

Andrews, D. W. K.: 1999, Estimation when a parameter is on a boundary, *Econometrica* **67**(6), 1341–1383.

Barro, R. J.: 2006, Rare disasters and asset markets in the twentieth century, *The Quarterly Journal of Economics* **121**(3), 823–866.

Berk, R. H.: 1972, Consistency and asymptotic normality of MLE's for exponential models, *The Annals of Mathematical Statistics* **43**(1), 193–204.

Bommier, A., Chassagnon, A. and Legrand, F.: 2012, Comparative risk aversion : A formal approach with applications to saving behavior, *Journal of Economic Theory* **147**(4), 1614–1641.

Bommier, A., Kochov, A. and Legrand, F.: 2017, On monotone recursive preferences, *Econometrica* **85**(5), 1433–1466.

Broda, S. and Kan, R.: 2015, On distributions of ratios, *Biometrika* **103**(1), 205–218.

Butler, R. W.: 2007, *Saddlepoint Approximations with Applications*, Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Caner, M.: 2009, Lasso-type GMM estimator, *Econometric Theory* **25**(1), 270–290.

Carrasco, M. and Florens, J.-P.: 2014, On the asymptotic efficiency of GMM, *Econometric Theory* **30**(2), 372–406.

Chamberlain, G.: 1987, Asymptotic efficiency in estimation with conditional moment restrictions, *Journal of Econometrics* **34**(3), 305–334.

Chib, S., Shin, M. and Simoni, A.: 2018, Bayesian estimation and comparison of moment condition models, *Journal of the American Statistical Association* **113**, 1656–1668.

Cover, T. M. and Thomas, J. A.: 2006/1991, *Elements of Information Theory*, second edn, John Wiley & Sons.

Csiszár, I.: 1975, I-divergence geometry of probability distributions and minimization problems, *The Annals of Probability* pp. 146–158.

Czellar, V. and Ronchetti, E.: 2010, Accurate and robust tests for indirect inference, *Biometrika* **97**(3), 621–630.

Daniels, H. E.: 1954, Saddlepoint approximations in statistics, *The Annals of Mathematical Statistics* **25**(4), 631–650.

Davidson, J.: 1994, *Stochastic Limit Theory*, Advanced Texts in Econometrics, Oxford University Press. printed in 2002.

Davison, A. C. and Hinkley, D. V.: 1988, Saddlepoint approximations in resampling methods, *Biometrika* **75**(3), 417–431.

Dufour, J.-M.: 1997, Some impossibility theorems in econometrics with applications to structural and dynamic models, *Econometrica* **65**(6), 1365–1387.

Easton, G. S. and Ronchetti, E.: 1986, General saddlepoint approximations with applications to L statistics, *Journal of the American Statistical Association* **81**(394), 420–430.

Efron, B.: 1981, Nonparametric standard errors and confidence intervals, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **9**(2), 139–158.

Epstein, L. G. and Zin, S. E.: 1991, Risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis, *The Journal of Political Economy* **99**(2), 263–286.

Esscher, F.: 1932, On the probability function in the collective theory of risk, *Scandinavian Actuarial Journal* pp. 175–195.

Fan, J.: 1997, Comments on "wavelets in statistics :a review" by A. Antoniadis, *Journal of the Italian Statistical Association* **2**, 131–138.

Fan, J. and Li, R.: 2001, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fasiolo, M., Wood, S. N., Hartig, F. and Bravington, M. V.: 2018, An extended empirical saddlepoint approximation for intractable likelihoods, *Electronic Journal of Statistics* **12**, 1544–1578.

Feller, W.: 1966/1971, *An Introduction to Probability Theory and Its Applications.*, Vol. 2, 2nd edn, Wiley.

Feuerverger, A.: 1989, On the empirical saddlepoint approximation, *Biometrika* **76**(3), 457–464.

Field, C.: 1982, Small sample asymptotic expansions for multivariate M-estimates, *The Annals of Statistics,* **10**(3), 672–689.

Field, C. A. and Ronchetti, E.: 1990, *Small Sample Asymptotics*, Lecture notes-Monograph Series, Institute of Mathematical Statistics.

Gabaix, X.: 2012, Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance, *The Quarterly Journal of Economics* **127**(2), 645–700.

Ghosh, J. K. and Ramamoorthi, R. V.: 2003, *Bayesian Nonparametrics*, Statistics, Springer.

Gourieroux, C. and Monfort, A.: 1995/1989, *Statistics and econometric models*, Vol. 2, Cambridge University Press. Translated from French by Quang Vuong.

Goutis, C. and Casella, G.: 1999, Explaining the saddlepoint approximation, *The American Statistician* **53**(3), 216–224.

Gregory, A. W., Lamarche, J.-F. and Smith, G. W.: 2002, Information-theoretic estimation of preference parameter: macroeconomic applications and simulation evidence, *Journal of Econometrics* **107**, 213–233.

Guggenberger, P. and Smith, R. J.: 2005, Generalized empirical likelihood estimators and tests under partial, weak, and strong identification, *Econometric Theory* **21**, 667–709.

Haberman, S. J.: 1984, Adjustmennt by minimum discriminant information, *The Annals of Statistics* **12**(3), 971–988.

Hall, A. R.: 2005, *Generalized Method of Moments*, Advanced Texts in Econometrics, Oxford University Press.

Hall, P. and Horowitz, J. L.: 1996, Bootstrap critical values for tests based on generalized-method-of-moments estimators, *Econometrica* **64**(4), 891–916.

Hansen, L. P.: 1982, Large sample properties of generalized method of moments estimators, *Econometrica* **50**(4), 1029–1054.

Hansen, L. P.: 2013, Uncertainty outside and inside economic models, *Nobel Prize Lecture*.

Hansen, L. P., Heaton, J. and Yaron, A.: 1996, Finite-sample properties of some alternative GMM estimators, *Journal of Business and Economic Statistics* **14**(3), 262–280.

Hansen, L. P. and Singleton, K. J.: 1982, Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica* **50**(5), 1269–1286.

Hiriart-Urruty, J.-B. and Lemaréchal, C.: 1993/1996, *Convex Analysis and Minimization Algorithms*, Comprehensive Studies in Mathematics. Second corrected printing.

Hoerland, A. E. and Kennard, R. W.: 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1), 55–67.

Holcblat, B.: 2012, *A Classical Moment-Based Inference Framework with Bayesian Properties*, PhD thesis, Carnegie Mellon University.

Holcblat, B. and Sowell, F.: 2019, Complete online appendix to "The ESP estimator". Available online at https://arxiv.org/abs/1905.06977.

Holly, A. and Phillips, P. C. B.: 1979, A saddlepoint approximation to the distribution of the k-class estimator of a coefficient in a simultaneous system, *Econometrica* **47**(6), 1527–1547.

Imbens, G. W.: 1997, One-step estimators for over-identified generalized method of moments models, *The Review of Economic Studies* **64**(3), 359–383.

Imbens, G. W., Spady, R. H. and Johnson, P.: 1998, Information theoretic approaches to inference in moment condition models, *Econometrica* **66**(2), 333–357.

Jennrich, R. I.: 1969, Asymptotic properties of non-linear least squares estimators, *The Annals of Mathematical Statistics* **40**(2), 633–643.

Jensen, J. L.: 1992, The modified signed likelihood statistic and saddlepoint approximations, *Biometrika* **79**(4), 693–703.

Jensen, J. L.: 1995, *Saddlepoint Approximations*, Oxford Statistical Science Series, Oxford University Press.

Jensen, J. L. and Wood, A. T.: 1998, Large deviation and other results for minimum contrast estimators, *Annals of the Institute of Statistical Mathematics* **50**(4), 673–695.

Julliard, C. and Ghosh, A.: 2012, Can rare events explain the equity premium puzzle?, *Review of Financial Studies* **25**(10).

Kallenberg, O.: 2002 (1997), *Foundation of Modern Probability*, Probability and Its Applications, second edn, Springer.

Kitamura, Y.: 2001, Asymptotic optimality of empirical likelihood for testing moment restrictions, *Econometrica* **69**(6), 1661–1672.

Kitamura, Y.: 2007, *Advances in Economics and Econometris*, Cambridge University Press, chapter Empirical Likelihood Methods in Econometrics: Theory and Practice, pp. 174–237.

Kitamura, Y. and Stutzer, M.: 1997, An information-theoretic alternative to generalized method of moments estimation, *Econometrica* **65**(4), 861–874.

Kolassa, J. E.: 1994/2006, *Series Approximation Methods in Statistics*, number 88 in *Lecture Notes in Statistics*, Springer.

Kumagai, S.: 1980, An implicit function theorem: Comment, *Journal of Optimization Theory and Applications* **31**(2), 285–288.

Kundhi, G. and Rilstone, P.: 2013, Edgeworth and saddlepoint expansions for nonlinear estimators, *Econometric Theory* **29**, 1057–1078.

Kundhi, G. and Rilstone, P.: 2015, Saddlepoint expansions for GEL estimators, *Statistical Methods & Applications* **24**, 1–24.

Lieberman, O.: 1994, On the approximation of saddlepoint expansions in statistics, *Econometric Theory* **10**(5), 900–916.

Lô, S. N. and Ronchetti, E.: 2012, Robust small sample accurate inference in moment condition models, *Computational Statistics and Data Analysis* **56**, 3182–3197.

Loader, C.: 2000, Fast and accurate computation of binomial probabilities, *Technical report.* Cited in *R: A Language and Environment for Statistical Computing. Reference Index*, Version 3.5.3 (2019-03-11). Available at https://lists.gnu.org/archive/html/octave-maintainers/2011-09/pdfK0uKOST642.pdf.

Lucas, R. E.: 1978, Asset prices in an exchange economy, *Econometrica* **46**(6), 1429–1445.

Ma, Y. and Ronchetti, E.: 2011, Saddlepoint test in measurement error models, *Journal of the American Statistical Association* **106**(493), 147–156.

Monfort, A.: 1996/1980, *Cours de Probabilités*, "Economie et statistiques avancées", ENSAE et CEPE, third edn, Economica.

Monti, A. C. and Ronchetti, E.: 1993, On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators, *Biometrika* **80**(2).

Müller, U. K. and Norets, A.: 2016, Credibility of confidence sets in nonstandard econometric problems, *Econometrica* **84**(6), 21832213.

Newey, W. K. and McFadden, D. L.: 1994, *Handbook of Econometrics*, Vol. 4, Elsevier Science Publishers, chapter "Large Sample Estimation and Hypothesis Testing", pp. 2113–2247.

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

Pearson, K.: 1894, Contribution to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society* pp. 71–110.

Pearson, K.: 1902, On the systematic fitting of curves to observations and measurements, parts I and II, *Biometrika* **1, 2**(3, 1), 265–303, 1–23.

Phillips, P. C. B.: 1978, Edgeworth and saddlepoint approximations in the first-order non circular autoregression, *Biometrika* **65**(1), 91–98.

Phillips, P. C. B.: 1982, Exact small theory in the simultaneous equations model, *Cowles Foundation Discussion Paper NO. 621* .

Qin, J. and Lawless, J.: 1994, Empirical likelihood and general estimating equations, *The Annals of Statistics* **22**, 300–325.

Rilstone, P., Srivastava, V. K. and Ullah, A.: 1996, The second-order bias and mean squared error of nonlinear estimators, *Journal of Econometrics* **75**(2), 369–395.

Robert, C. P.: 2007/1994, *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*, Texts in Statistics, second edn, Springer.

Robinson, J., Ronchetti, E. and Young, G. A.: 2003, Saddlepoint approximations and tests based on multivariate M-estimates, *Annals of Statistics* **31**(4), 1154–1169.

Ronchetti, E. and Trojani, F.: 2001, Robust inference with GMM estimators, *Journal of Econometrics* **101**, pp. 37–69.

Ronchetti, E. and Trojani, F.: 2003, Saddlepoint approximations and test statistics for accurate inference in overidentified moment conditions models, *Working paper, National Centre of Competence in Research, Financial Valuation and Risk Management* .

Ronchetti, E. and Welsh, A. H.: 1994, Empirical saddlepoint approximations for multivariate M-estimators, *Journal of the Royal Statistical Society.Series B (Methodological),* **56**(2), 313–326.

Rudin, W.: 1953, *Principles of Mathematical Analysis*, 3rd edn, McGraw-Hill.

Schennach, S. M.: 2005, Bayesian exponentially tilted empirical likelihood, *Biometrika* **92**(1), 31–46.

Schennach, S. M.: 2007, Point estimation with exponentially tilted empirical likelihood, *The Annals of Statistics* **35**(2), 634–672.

Schmetterer, L.: 1966, *Mathematische Statistik*, second edn, Springer.

Sengupta, N. and Sowell, F.: 2020, On the asymptotic distribution of ridge regression estimators using training and test samples, *Econometrics* **8**(39). doi:10.3390/econometrics8040039.

Skovgaard, I. M.: 1990, On the density of minimum contrast estimators, *The Annals of Statistics* **18**(2), 779–789.

Smith, R. J.: 2011, GEL criteria for moment condition models, *Econometric Theory* **27**(6), 1192–1235.

Sowell, F.: 1996, Optimal tests of parameter variation in the generalized method of moments framework, *Econometrica* **64**(5), 1085–1108.

Sowell, F.: 2007, The empirical saddlepoint approximation for GMM estimators, *working paper, Tepper School of Business, Carnegie Mellon University* .

Sowell, F.: 2009, The empirical saddlepoint likelihood estimator applied to two-step GMM, *working paper, Tepper School of Business, Carnegie Mellon University* .

Spady, R. H.: 1991, Saddlepoint approximations for regression models, *Biometrika* **78**(4), 879–889.

Stein, C.: 1956, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. Volume 1: Contributions to the Theory of Statistics, University of California Press, Berkeley, California., pp. 197–206.

Stock, J. H. and Wright, J. H.: 2000, GMM with weak identification, *Econometrica* **68**(5), 1055–1096.

Tibshirani, R.: 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Vecchia, D. L., Ronchetti, E. and Trojani, F.: 2012, Higher-order infinitesimal robustness, *Journal of the American Statistical Association* **107**, 1546–1557.

Wald, A.: 1949, Note on the consistency of the maximum likelihood estimate, *The Annals of Mathematical Statistics* **20**(4), 595–601.

Wang, S.: 1990, Saddlepoint approximations in resampling analysis, *Annals of the Institute of Statistical Mathematics* **42**(1), 115–131.

Young, G. A. and Daniels, H. E.: 1990, Bootstrap bias, *Biometrika* **77**(1), 179–185.

Zou, H. and Hastie, T.: 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**(2), 301–320.

Notes and acknowledgements

Parts of the present paper have previously circulated under the title “The Empirical Saddlepoint Likelihood Estimator Applied to Two-Step GMM” (Sowell 2009). Some proofs of the present paper also borrow technical results from Holcblat (2012). Helpful comments were provided by Patrick Gagliardini (discussant), Philipp Ketz (discussant), Eric Renault, Enrique Sentana, Aman Ullah and seminar/conference participants at Carnegie Mellon University, CFE-CMStatistics 2017, Swiss Finance Institute (EPFL and the University of Lausanne), 10th French Econometrics Conference (Paris School of Economics), at the Econometric Society European Winter Meeting 2018 (University of Naples Federico II), the University of Luxembourg, and Collegio Carlo Alberto (University of Torino, 2nd LTI conference).

ONLINE APPENDIX: The ESP estimator

Benjamin Holcblat and Fallaw Sowell

This appendix mainly consists of a detailed proof of the first part of Theorem 1(i), i.e., existence and consistency of the ESP estimator. The proof relies on set-valued analysis. The high level of details should make the proof more transparent, should ease the use of the intermediary results in further research, and should make clear that the assumptions and the proofs available in the current literature are mathematically insufficient to establish Theorem 1(i). The proofs of the other results (i.e., asymptotic normality of the ESP estimator, and Theorem 2, asymptotic distributions of the Trintiy+1 test statistics) are skipped because, while technical, they rely on extensions of more standard arguments, so the indications in the main text should be sufficient. Nevertheless, the latter proofs are available in Holcblat and Sowell (2019).

In addition to the proofs, this appendix contains a formalization of the variance shrinkage, an additional numerical example, and complementary information regarding the numerical and empirical examples.

Appendix A. Proofs

A.1. Proof of Theorem 1(i) : Existence and consistency.

Core of the proof of Theorem 1i. Under Assumption 1(a)(b) and (d)-(h), by Lemma 1 (p. 25), $\mathbb{P}$-a.s. for T big enough, the ESP approximation and the ESP estimator exist. Moreover, under Assumption 1(a)-(b) and (d)-(h), by Lemma 6iv (p. 30), $\mathbb{P}$-a.s. for T big enough, $|\Sigma_T(\theta)|_{\text{det}} > 0$, for all $\theta \in \boldsymbol{\Theta}$. Thus, we can apply the strictly increasing transformation $x \mapsto \frac{1}{T}[\ln(x) - \frac{m}{2}\ln(\frac{T}{2\pi})]$ to the ESP approximation in equation (2) on p. 4, so that, $\mathbb{P}$-a.s. for T big enough,

$$\begin{aligned} &\hat{\theta}_T \in \arg\max_{\theta\in\boldsymbol{\Theta}} \hat{f}_{\theta_T^*}(\theta) \\ \Leftrightarrow\ &\hat{\theta}_T \in \arg\max_{\theta\in\boldsymbol{\Theta}} \left\{ \ln\left[\frac{1}{T}\sum_{t=1}^{T} \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right] - \frac{1}{2T}\ln|\Sigma_T(\theta)|_{\text{det}} \right\}. \end{aligned} \tag{11}$$

Now, by the triangle inequality,

$$\begin{aligned} &\sup_{\theta\in\boldsymbol{\Theta}} \left| \ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right] - \frac{1}{2T}\ln|\Sigma_T(\theta)|_{\text{det}} - \ln\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}] \right| \\ \leqslant\ &\sup_{\theta\in\boldsymbol{\Theta}} \left| \ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right] - \ln\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}] \right| + \sup_{\theta\in\boldsymbol{\Theta}} \left| \frac{1}{2T}\ln|\Sigma_T(\theta)|_{\text{det}} \right| \\ =\ &o(1)\ \mathbb{P}\text{-a.s. as } T\to\infty \end{aligned} \tag{12}$$

where the last equality follow from Lemma 2iv (p. 25) and Lemma 6v (p. 30) under Assumption 1(a)-(b) and (d)-(h). Thus, regarding $\hat{\theta}_T$, it is now sufficient to check the assumptions of the standard consistency theorem (e.g. Newey and McFadden 1994, pp. 2121-2122 Theorem 2.1, which is also valid in an almost-sure sense). Firstly, under Assumption 1 (a)-(e) and (g)-(h), by Lemma 10iv (p. 33), $\theta \mapsto \ln\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]|$ is uniquely maximized at θ_0, i.e., for all $\theta \in \boldsymbol{\Theta}\setminus\{\theta_0\}$,

$\ln \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}] < \ln \mathbb{E}[\mathrm{e}^{\tau(\theta_0)'\psi(X_1,\theta_0)}] = 0$. Secondly, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28), $\theta \mapsto \ln \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]$ is continuous in $\mathbf{\Theta}$. Finally, by Assumption 1(d), the parameter space $\mathbf{\Theta}$ is compact. □

Lemma 1 (Existence of the ESP approximation and estimator)**.** *Under Assumption 1(a)(b) and (d)-(h),* $\mathbb{P}$*-a.s. for* T *big enough,*

(i) *the ESP approximation* $\hat{f}_{\theta_T^*}(.)$ *exists;*
(ii) $\theta \mapsto \tau_T(\theta)$ *is unique and continuously differentiable in* $\mathbf{\Theta}$*, so that the ESP approximation* $\theta \mapsto \hat{f}_{\theta_T^*}(\theta)$ *is also unique and continuous in* $\mathbf{\Theta}$*;*
(iii) *for all* $\theta \in \mathbf{\Theta}$*, the ESP approximation* $\omega \mapsto \hat{f}_{\theta_T^*}(\theta)$ *is* $\mathcal{E}/\mathcal{B}(\mathbf{R})$*-measurable; and*
(iv) *there exists an ESP estimator* $\hat{\theta}_T \in \arg\max_{\theta\in\mathbf{\Theta}} \hat{f}_{\theta_T^*}(\theta)$ *that is* $\mathcal{E}/\mathcal{B}(\mathbf{R}^m)$*-measurable.*

Proof. The result follows from Lemmas 2 (p. 25), 3 (p. 28) and 6 (p. 30) and standard arguments. For completeness, a detailed proof is provided.

(i) Under Assumption 1(a)(b), (d)-(e)(g) and (h), by Lemma 2ii (p. 25), $\mathbb{P}$-a.s. there exists a $\mathcal{B}(\mathbf{\Theta}) \otimes \mathcal{E}/\mathcal{B}(\mathbf{R}^m)$-measurable function $\tau_T(.)$ s.t., for T big enough, for all $\theta \in \mathbf{\Theta}$, $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\psi_t(\theta) = 0_{m\times 1}$ and $\tau_T(\theta) \in \mathrm{int}[\mathbf{T}(\theta)]$. Moreover, under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28) with $\mathrm{P} = \frac{1}{T}\sum_{t=1}^T \delta_{X_t}$, for all $T \in [\![1,\infty[\![$, for all $(\theta,\tau) \in \mathbf{S}$, $0 < \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$, so that, for all $\theta \in \mathbf{\Theta}$, $0 < \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}$. Thus, the ET term exists. Now, under Assumption 1(a)-(b) and (d)-(h), by Lemma 6iv (p. 30), $\mathbb{P}$-a.s. for T big enough, $\inf_{\theta\in\mathbf{\Theta}} |\Sigma_T(\theta)|_{\det} > 0$, so that the variance term of the ESP approximation exists. Thus, the ESP approximation exists.

(ii) By Assumption 1(b), $\theta \mapsto \psi(X_1,\theta)$ is continuously differentiable in $\mathbf{\Theta}^\epsilon$ $\mathbb{P}$-a.s., so that it is sufficient to show that $\tau_T(.)$ is unique and continuous, which we prove at once with the standard implicit function theorem. Check its assumptions. Firstly, under Assumption 1(a)(b), (d)-(e)(g) and (h), by Lemma 2ii (p. 25), $\mathbb{P}$-a.s. there exists a function $\tau_T(.)$ s.t., for T big enough, for all $\theta \in \mathbf{\Theta}$, $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\psi_t(\theta) = 0_{m\times 1}$ and $\tau_T(\theta) \in \mathrm{int}[\mathbf{T}(\theta)]$. Secondly, for all $\dot{\theta} \in \mathbf{\Theta}$, $\left.\frac{\partial\left[\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}\psi_t(\theta)\right]}{\partial\tau'}\right|_{(\theta,\tau)=(\dot{\theta},\tau_T(\dot{\theta}))} = \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\dot{\theta})}\psi_t(\dot{\theta})\psi_t(\dot{\theta})'$, which is full rank $\mathbb{P}$-a.s. for T big enough for all $\theta \in \mathbf{\Theta}$, because under Assumption 1(a)-(b) and (d)-(h), by Lemma 6iv (p. 30), $\mathbb{P}$-a.s. for T big enough, $\inf_{\theta\in\mathbf{\Theta}} |\Sigma_T(\theta)|_{\det} > 0$. Finally, by Assumption 1(b), $(\theta,\tau) \mapsto \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}\psi_t(\theta)$ is continuously differentiable in $\mathbf{S}^\epsilon$.

(iii) By Assumption 1(b), for all $\theta \in \mathbf{\Theta}$, $x \mapsto \psi(x,\theta)$ is $\mathcal{B}(\mathbf{R}^p)/\mathcal{B}(\mathbf{R}^m)$-measurable. Moreover, under Assumption 1(a)(b), (d)-(e)(g) and (h), by Lemma 2ii (p. 25), $\mathbb{P}$-a.s. $\tau_T(.)$ is a $\mathcal{B}(\mathbf{\Theta}) \otimes \mathcal{E}/\mathcal{B}(\mathbf{R}^m)$-measurable function. Thus, the result follows.

(iv) By Assumption 1(d), $\mathbf{\Theta}$ is compact, so that, by the statements (i)-(iii) of the present lemma, the result follows from the Schmetterer-Jennrich lemma (Schmetterer 1966 Chap. 5 Lemma 3.3; Jennrich 1969 Lemma 2). □

Lemma 2 (Asymptotic limit of the ET term)**.** *Under Assumption 1(a)(b), (d)-(e)(g) and (h),*

(i) $\mathbb{P}$*-a.s. as* $T \to \infty$*,* $\sup_{(\theta,\tau)\in\mathbf{S}} \left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]\right| = o(1)$*, which implies that* $\mathbb{P}$*-a.s. as* $T \to \infty$*,* $\sup_{(\theta,\tau)\in\mathbf{S}} \left|\ln\left[\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}\right] - \ln \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]\right| = o(1)$ *;*

(ii) $\mathbb{P}$-*a.s. there exists a* $\mathcal{B}(\mathbf{\Theta}) \otimes \mathcal{E}/\mathcal{B}(\mathbf{R}^m)$*-measurable function* $\tau_T(.)$ *s.t., for* T *big enough, for all* $\theta \in \mathbf{\Theta}$*,* $\tau_T(\theta) \in \arg\min_{\tau \in \mathbf{R}^m} \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$*,* $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\psi_t(\theta) = 0_{m\times 1}$ *and* $\tau_T(\theta) \in \operatorname{int}[\mathbf{T}(\theta)]$*;*

(iii) $\mathbb{P}$-*a.s. as* $T \to \infty$*,* $\sup_{\theta\in\mathbf{\Theta}} |\tau_T(\theta) - \tau(\theta)| = o(1)$*;*

(iv) $\mathbb{P}$-*a.s. as* $T \to \infty$*,* $\sup_{\theta\in\mathbf{\Theta}} \left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right| = o(1)$*, which implies that* $\mathbb{P}$-*a.s. as* $T \to \infty$*,* $\sup_{\theta\in\mathbf{\Theta}} \left|\ln\left[\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right] - \ln \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right| = o(1)$.

Proof. (i) Under Assumptions 1 (a)-(b)(d)(e)(g) and (h), by Lemma 4iii (p. 28), $\mathbf{S} := \{(\theta,\tau) : \theta \in \mathbf{\Theta} \wedge \tau \in \mathbf{T}(\theta)\}$ is a compact set.[1] Thus, under Assumption 1(a)-(b), (d) (e) and (h), the ULLN (uniform law of large numbers) à la Wald (e.g., Ghosh and Ramamoorthi 2003, pp. 24-25, Theorem 1.3.3) yields the first part of the result. Now, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28), $(\theta,\tau) \mapsto \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ is continuous, so that $\{\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}] : (\theta,\tau) \in \mathbf{S}\}$ is a compact set by Assumption 1(d) —continuous mappings preserve compactness (e.g., Rudin 1953, Theorem 4.14). Moreover, $x \mapsto \ln x$ is continuous, and, under Assumptions 1 (a)(b)(d)(e)(g) and (h), again by Lemma 3 (p. 28), $0 < \inf_{(\theta,\tau)\in\mathbf{S}} \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$. Thus, we can choose an $\eta \in \left]0, \inf_{(\theta,\tau)\in\mathbf{S}} \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]\right[$ s.t. $x \mapsto \ln x$ is uniformly continuous on the closed η-neighborhood of $\{\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}] : (\theta,\tau) \in \mathbf{S}\}$ —continuous mappings on a compact set are uniformly continuous (e.g., Rudin 1953, Theorem 4.19). Then, the second part follows from the first part of the result: By the first part, $\mathbb{P}$-a.s. there exists a $\dot{T} \in [\![1,\infty[\![$ s.t., $\forall T \in [\![\dot{T},\infty[\![$, $\sup_{(\theta,\tau)\in\mathbf{S}} |\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]| < \eta/2$.

(ii)-(iii) Let $\eta \in]0, \epsilon_{\mathbf{T}}]$ be a fixed constant. By Assumption 1(a)(b), $(\theta,\omega) \mapsto \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$ is continuous w.r.t θ and $\mathcal{E}/\mathcal{B}(\mathbf{R})$-measurable w.r.t to ω, so that it is $\mathcal{B}(\mathbf{\Theta}) \otimes \mathcal{E}/\mathcal{B}(\mathbf{R})$-measurable (e.g., Aliprantis and Border 2006/1999, Lemma 4.51). Moreover, under Assumptions 1 (a)-(b)(d)(e)(g) and (h), by Lemma 4ii (p. 28), $\theta \mapsto \mathbf{T}(\theta)$ is a nonempty compact valued measurable correspondence. Then, by a generalization of the Schmetterer-Jennrich lemma (e.g., Aliprantis and Border 2006/1999, Theorem 18.19), we can define a $\mathcal{B}(\mathbf{\Theta}) \otimes \mathcal{E}/\mathcal{B}(\mathbf{R})$-measurable function $\tilde{\tau}_T(\theta)$ s.t., for all $\theta \in \mathbf{\Theta}$, $\tilde{\tau}_T(\theta) \in \arg\min_{\tau\in\mathbf{T}(\theta)} \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$. For the present proof, put $\varepsilon := \inf_{\theta\in\mathbf{\Theta}} \inf_{\tau\in\mathbf{T}(\theta):|\tau-\tau(\theta)|\geqslant\eta} |\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]|$, which is strictly positive[2] by Lemma 5 (p. 29) under Assumptions 1 (a)(b)(d)(e) and (h).[3] Then, by the definition of ε, whenever $\sup_{\theta\in\mathbf{\Theta}} |\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]| < \varepsilon$, then $\sup_{\theta\in\mathbf{\Theta}} |\tilde{\tau}_T(\theta) - \tau(\theta)| \leqslant \eta$. We now show that it is happening $\mathbb{P}$-a.s. as $T \to \infty$. Under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma

[1]Note that, unlike what has been sometimes suggested in the entropy literature, if $\mathbf{T}(\theta)$ is an unspecified compact set, $\{(\theta,\tau) : \theta \in \mathbf{\Theta} \wedge \tau \in \mathbf{T}(\theta)\}$ does not need to be a compact set: $\{(\theta,\tau) : \theta \in \mathbf{\Theta} \wedge \tau \in \mathbf{T}(\theta)\}$ is not a Cartesian product, but the graph of a correspondence. See Lemma 4 (p. 28) for more details. The possible non-compactness of $\mathbf{S}$ under the assumptions used in the entropy literature is one of the several mathematical reasons why, while we try to remain close as much as possible to proofs in the entropy literature (e.g., Schennach 2007), we cannot rely much on them.

[2]The argument requires $\varepsilon > 0$. If $\varepsilon = 0$, then the upcoming inequality (13) is not sufficient to show that $\sup_{\theta\in\mathbf{\Theta}} |\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]| < \varepsilon$.

[3]Strict convexity of $\tau \mapsto \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ and compactness of $\mathbf{\Theta}$ are not sufficient to ensure that $\varepsilon > 0$: We also need the continuity of the value function of the first infimum, which we obtain through Berge's maximum theorem. See Lemma 5 (p. 29).

10 (p. 33), $\tau(\theta) = \arg\min_{\tau\in\mathbf{R}^m} \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$, so that

$$
\begin{aligned}
&\sup_{\theta\in\mathbf{\Theta}} \left|\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right| \\
&= \sup_{\theta\in\mathbf{\Theta}} \left\{\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right\} \\
&\overset{(a)}{=} \sup_{\theta\in\mathbf{\Theta}} \left\{\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tilde{\tau}_T(\theta)'\psi_t(\theta)} + \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tilde{\tau}_T(\theta)'\psi_t(\theta)} - \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau(\theta)'\psi_t(\theta)} \right. \\
&\qquad \left. + \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau(\theta)'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right\} \\
&\overset{(b)}{\leqslant} \sup_{\theta\in\mathbf{\Theta}} \left\{\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tilde{\tau}_T(\theta)'\psi_t(\theta)} + \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau(\theta)'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right\} \\
&\overset{(c)}{\leqslant} \sup_{\theta\in\mathbf{\Theta}} \left|\mathbb{E}[\mathrm{e}^{\tilde{\tau}_T(\theta)'\psi(X_1,\theta)}] - \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tilde{\tau}_T(\theta)'\psi_t(\theta)}\right| + \sup_{\theta\in\mathbf{\Theta}} \left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau(\theta)'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right| \\
&\overset{(d)}{=} o(1) \ \mathbb{P}\text{-a.s. as } T\to\infty.
\end{aligned} \tag{13}
$$

(a) Add and subtract $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tilde{\tau}_T(\theta)'\psi_t(\theta)}$ and $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau(\theta)'\psi_t(\theta)}$. *(b)* Note that, under Assumption 1(d) and (e), by definition, $\tau(\theta)\in\mathbf{T}(\theta)$ and $\tilde{\tau}_T(\theta)\in\arg\min_{\tau\in\mathbf{T}(\theta)}\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$ so that $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tilde{\tau}_T(\theta)'\psi_t(\theta)} - \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau(\theta)'\psi_t(\theta)} \leqslant 0$. *(c)* Triangle inequality w.r.t. the uniform norm. *(d)* Under Assumption 1(d)(e), by definition, for all $\theta\in\mathbf{\Theta}$, $\tau(\theta)\in\mathbf{T}(\theta)$ and $\tilde{\tau}_T(\theta)\in\mathbf{T}(\theta)$ so that the conclusion follows from statement (i).

Inequality (13) implies that $\sup_{\theta\in\mathbf{\Theta}}|\tilde{\tau}_T(\theta)-\tau(\theta)| = o(1)$ $\mathbb{P}$-a.s. as $T\to\infty$. Moreover, by Assumption 1(e), for all $\theta\in\mathbf{\Theta}$, $\mathbf{T}(\theta) = \overline{B_{\epsilon_\mathbf{T}}(\tau(\theta))}$ where $\epsilon_\mathbf{T} > 0$. Thus, $\mathbb{P}$-a.s., for T big enough, for all $\theta\in\mathbf{\Theta}$, $\tilde{\tau}_T(\theta)\in\mathrm{int}[\mathbf{T}(\theta)]$. Now, for all $\theta\in\mathbf{\Theta}$, $\tau\mapsto\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$ is a convex function (Lemma 12i on p. 37 with $\mathrm{P} = \frac{1}{T}\sum_{t=1}^T \delta_{X_t}$ ensures that $\frac{\partial^2[\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}]}{\partial\tau\partial\tau'} = \frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}\psi_t(\theta)\psi_t(\theta)' \geqslant 0$), and the local minimum of a convex function is a global minimum (e.g., Hiriart-Urruty and Lemaréchal 1993/1996, p. 253). Therefore, $\mathbb{P}$-a.s. for T big enough, for all $\theta\in\mathbf{\Theta}$, $\tilde{\tau}_T(\theta)$ minimizes $\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$ not only over $\mathbf{T}(\theta)$, but also over $\mathbf{R}^m$, which means that we can put $\tilde{\tau}_T(\theta) = \tau_T(\theta)$.

(iv) Addition and subtraction of $\mathbb{E}[\mathrm{e}^{\tau_T(\theta)'\psi(X_1,\theta)}]$, and the triangle inequality yield $\mathbb{P}$-a.s. for T big enough

$$
\begin{aligned}
&\sup_{\theta\in\mathbf{\Theta}} \left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right| \\
&\leqslant \sup_{\theta\in\mathbf{\Theta}} \left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau_T(\theta)'\psi(X_1,\theta)}]\right| + \sup_{\theta\in\mathbf{\Theta}} \left|\mathbb{E}[\mathrm{e}^{\tau_T(\theta)'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\right| \\
&= o(1)\ , \text{ as } T\to\infty,
\end{aligned}
$$

where the explanations for the last equality are as follows. By the statement (i) of the present lemma, $\mathbb{P}$-a.s. as $T\to\infty$, $\sup_{(\theta,\tau)\in\mathbf{S}}\left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)} - \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]\right| = o(1)$. Moreover, by the statement (ii) of the present lemma, $\mathbb{P}$-a.s. for T big enough, $\tau_T(\theta)\in\mathrm{int}[\mathbf{T}(\theta)]$, so that, for all $\theta\in\mathbf{\Theta}$, $(\theta,\tau_T(\theta))\in\mathbf{S}$. Thus, the first supremum is $o(1)$ as $T\to\infty$. Regarding the second supremum, under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28), $(\theta, \tau) \mapsto \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ is continuous in $\mathbf{S}$. Now, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 4iii (p. 28), $\mathbf{S}$ is compact, so that $(\theta, \tau) \mapsto \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ is also uniformly continuous in $\mathbf{S}$ —continuous functions on compact sets are uniformly continuous (e.g., Rudin 1953, Theorem 4.19). Thus, under Assumption 1(a)(b), (d)-(e), (g) and (h), by the statement (iii) of the present lemma, which states that $\sup_{\theta\in\mathbf{\Theta}} |\tau_T(\theta) - \tau(\theta)| = o(1)$ $\mathbb{P}$-a.s. as $T \to \infty$, the second supremum is also $o(1)$ $\mathbb{P}$-a.s. as $T \to \infty$.

The second part of the result follows from the first part as in the proof of the statement (i) of the present lemma. □

Lemma 3. *Let* P *be any probability measure, and* $\mathbb{E}_{\mathrm{P}}$ *denote the expectation under* P. *Under Assumption 1 (a)(b)(d)(e)(g) and (h), if* $\mathbb{E}_{\mathrm{P}}[\sup_{(\theta,\tau)\in\mathbf{S}} \mathrm{e}^{\tau'\psi(X_1,\theta)}] < \infty$, *then* $0 < \inf_{(\theta,\tau)\in\mathbf{S}} \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$, *so that* $0 < \inf_{\theta\in\mathbf{\Theta}} \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]$. *Moreover,* $(\theta, \tau) \mapsto \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ *and* $\theta \mapsto \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]$ *are continuous in* $\mathbf{S}$ *and* $\mathbf{\Theta}$, *respectively. All of these results hold for* $\mathrm{P} = \mathbb{P}$ *under the aforementioned assumptions.*

Proof. Under Assumption 1 (a) and (b), the Lebesgue dominated convergence theorem and the lemma's assumption $\mathbb{E}_{\mathrm{P}}[\sup_{(\theta,\tau)\in\mathbf{S}} \mathrm{e}^{\tau'\psi(X_1,\theta)}] < \infty$ imply that $(\theta, \tau) \mapsto \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ is continuous. Moreover, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 4 (p. 28), $\mathbf{S}$ is compact, and continuous functions over compact sets reach a minimum (e.g., Rudin 1953, Theorem 4.16). Now, if there exist $(\dot{\tau}, \dot{\theta}) \in \mathbf{S}$ s.t. $0 = \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\dot{\tau}'\psi(X_1,\dot{\theta})}]$, then $\mathrm{e}^{\dot{\tau}'\psi(X_1,\dot{\theta})} = 0$ P-a.s. (e.g., Kallenberg 2002 (1997, Lemma 1.24), which is impossible by definition of the exponential function. Thus, $0 < \inf_{(\theta,\tau)\in\mathbf{S}} \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$, so that $0 < \inf_{\theta\in\mathbf{\Theta}} \mathbb{E}_{\mathrm{P}}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]$ because by the definition of $\mathbf{S}$ in Assumption 1(e), for all $\theta \in \mathbf{\Theta}$, $(\theta, \tau(\theta)) \in \mathbf{S}$. Regarding the second part of the result, it immediately follows from the Lebesgue dominated convergence theorem, the lemma's assumption that $\mathbb{E}_{\mathrm{P}}[\sup_{(\theta,\tau)\in\mathbf{S}} \mathrm{e}^{\tau'\psi(X_1,\theta)}] < \infty$, and the continuity of $\tau : \mathbf{\Theta} \to \mathbf{R}^m$ by Lemma 10iii (p. 33) under Assumptions 1 (a)(b)(d)(e)(g) and (h). Regarding the third part of the result, it is sufficient to note that, under Assumption 1 (a)(b), by the Cauchy-Schwarz inequality, $\mathbb{E}[\sup_{(\theta,\tau)\in\mathbf{S}} \mathrm{e}^{\tau'\psi(X_1,\theta)}] \leqslant \mathbb{E}[\sup_{(\theta,\tau)\in\mathbf{S}} \mathrm{e}^{2\tau'\psi(X_1,\theta)}]^{1/2} < \infty$, where the last inequality follows from Assumption 1(e). □

Lemma 4 (Compactness of $\mathbf{S}$). *Under Assumptions 1 (a)(b)(d)(e)(g) and (h),*

(i) *The closure of the* $\epsilon_{\mathbf{T}}$*-neighborhood of* $\tau(\mathbf{\Theta})$ *(i.e.,* $\overline{\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}}}$*) is compact*

(ii) *For all* $\theta \in \mathbf{\Theta}$, *the correspondence* $\theta \mapsto \mathbf{T}(\theta)$ *is nonempty compact-valued and uhc (upper hemi-continuous), and thus measurable;*

(iii) *The set* $\mathbf{S} := \{(\theta, \tau) : \theta \in \mathbf{\Theta} \wedge \tau \in \mathbf{T}(\theta)\}$ *is compact.*

Proof. (i) Under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 10iii (p. 33), $\tau : \mathbf{\Theta} \to \mathbf{R}^m$ is continuous. Moreover, by Assumption 1(d), $\mathbf{\Theta}$ is compact. Thus, $\tau(\mathbf{\Theta})$ is bounded —continuous mappings preserve compactness (e.g., Rudin 1953, Theorem 4.14). Consequently, $\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}} =: \{\tau \in \mathbf{R}^m : \inf_{\tilde{\tau}\in\tau(\mathbf{\Theta})} |\tau - \tilde{\tau}| < \epsilon_{\mathbf{T}}\}$ is bounded, which means that its closure $\overline{\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}}}$ is closed and bounded, i.e., compact.

(ii) Proof that $\mathbf{T}$ *is nonempty and compact valued.* By Assumption 1(d), for all $\theta \in \mathbf{\Theta}$, there exists $\tau(\theta)$ s.t. $\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1, \theta)] = 0$. Thus, for all $\theta \in \mathbf{\Theta}$, $\mathbf{T}(\theta) = \overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))}$ is nonempty. Moreover, by construction, $\overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))}$ is compact, so that it is nonempty compact valued.

Proof that $\mathbf{T}$ *is uhc.* Because $\mathbf{T}$ is compact valued, we can use the sequential characterization of upper hemicontinuity (e.g., Aliprantis and Border 2006/1999, Theorem 17.20). Let $((\theta_n,\tau_n))_{n\in\mathbf{N}} \in (\mathbf{S})^{\mathbf{N}}$ be a sequence s.t., for all $n \in \mathbf{N}$, $\tau_n \in \mathbf{T}(\theta_n)$ and $\theta_n \to \bar{\theta} \in \mathbf{\Theta}$ as $n \to \infty$. By construction, for all $n \in \mathbf{N}$, $\tau_n \in \overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta_n))} \subset \overline{\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}}}$. Moreover, by statement (i), $\overline{\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}}}$ is compact, so that there exists a subsequence $(\tau_{\alpha(n)})_{n\in\mathbf{N}}$ s.t. $\tau_{\alpha(n)} \to \bar{\tau} \in \overline{\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}}}$, as $n \to \infty$. Again, by construction, for all $n \in \mathbf{N}$, $(\theta_n,\tau_n) \in \mathbf{S}$, so that $|\tau_{\alpha(n)} - \tau(\theta_{\alpha(n)})| \leqslant \epsilon_{\mathbf{T}}$. Now, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 10iii, $\tau : \mathbf{\Theta} \to \mathbf{R}^m$ is continuous. Thus, $|\tau_{\alpha(n)} - \tau(\theta_{\alpha(n)})| \to |\bar{\tau} - \tau(\bar{\theta})|$ as $n \to \infty$. Thus, $|\bar{\tau} - \tau(\bar{\theta})| \leqslant \epsilon_{\mathbf{T}}$, which means that $\bar{\tau} \in \overline{B_{\epsilon_{\mathbf{T}}}(\tau(\bar{\theta}))} = \mathbf{T}(\bar{\theta})$.

Proof that $\mathbf{T}$ *is measurable.* Let F be a closed subset of $\mathbf{R}^m$. Then, its complement F^c is an open subset of $\mathbf{R}^m$. Now, a correspondence is uhc iff the upper inverse image of a open set is an open set (e.g., Aliprantis and Border 2006/1999, Lemma 17.4). Thus, by the previous paragraph, $\mathbf{T}^u(F^c) \in \mathcal{B}(\mathbf{\Theta})$, where $\mathbf{T}^u$ denotes the upper inverse of $\mathbf{T}$. Now, denoting the lower inverse of $\mathbf{T}$ with $\mathbf{T}^l$, notice that $\mathbf{T}^u(F^c) = [\mathbf{T}^l(F)]^c$ (e.g., Aliprantis and Border 2006/1999, p. 557), so that $[\mathbf{T}^l(F)]^c \in \mathcal{B}(\mathbf{\Theta})$, which, in turn implies that $\mathbf{T}^l(F) \in \mathcal{B}(\mathbf{\Theta})$ because of the stability of σ-algebras under complementation.

(iii) Note that the compactness of $\mathbf{\Theta}$ and $\mathbf{T}(\theta)$ are not sufficient to ensure the compactness of $\mathbf{S}$ because $\mathbf{S}$ is not a Cartesian product. By the statement (ii) of the present lemma, $\mathbf{T}$ is uhc and closed valued, so that it has a closed graph (e.g., Aliprantis and Border 2006/1999, Theorem 17.10), i.e., $\mathbf{S}$ is closed. Now, by construction, $\mathbf{S}$ is a subset $[\overline{\tau(\mathbf{\Theta})^{\epsilon_{\mathbf{T}}}} \times \mathbf{\Theta}]$, which is compact by statement (i) and Assumption 1(d). Thus, $\mathbf{S}$ is also compact —in metric spaces, closed subsets of compact sets are compact (e.g., Rudin 1953, Theorem 2.35). ☐

Lemma 5. *Under Assumptions 1 (a)(b)(d)(e) and (h),*

(i) *for any constant* $\eta \in]0,\epsilon_{\mathbf{T}}]$*, there exists a continuous value function* $v : \mathbf{\Theta} \to \mathbf{R}_+$ *s.t., for all* $\theta \in \mathbf{\Theta}$*,* $v(\theta) = \inf_{\tau\in\mathbf{T}(\theta):|\tau-\tau(\theta)|\geqslant\eta} |\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]|$*;*

(ii) *for any constant* $\eta \in]0,\epsilon_{\mathbf{T}}]$*,* $0 < \inf_{\theta\in\mathbf{\Theta}} \inf_{\tau\in\mathbf{T}(\theta):|\tau-\tau(\theta)|\geqslant\eta} |\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]|$*.*

Proof. (i) It is a consequence of Berge's maximum theorem (e.g., Aliprantis and Border 2006/1999, Theorem 17.31). Thus, it remains to check its assumptions. For the present proof, define the correspondence $\varphi : \mathbf{\Theta} \twoheadrightarrow \mathbf{R}^m$ s.t. $\varphi(\theta) = \{\tau \in \mathbf{T}(\theta) : |\tau - \tau(\theta)| \geqslant \eta\}$, and the function $f : \mathbf{S} \to \mathbf{R}_+$ s.t. $f(\theta,\tau) = |\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}] - \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]|$.

Proof of the continuity of f*.* Under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28), $(\theta,\tau) \mapsto \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ and $\theta \mapsto \mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]$ are continuous in $\mathbf{S}$ and $\mathbf{\Theta}$, respectively, so that the continuity of f follows immediately.

Proof that φ *is nonempty compact valued.* By the definition of $\mathbf{T}$ in Assumption 1(e), for all $\theta \in \mathbf{\Theta}$, $\mathbf{T}(\theta) = \overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))}$, so that, for any $\eta \in]0,\epsilon_{\mathbf{T}}]$, $\varphi(\theta) = \overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))} \cap \{\tau \in \mathbf{R}^m : \eta \leqslant |\tau - \tau(\theta)|\} \neq \emptyset$, i.e., φ is nonempty valued. Moreover, for all $\theta \in \mathbf{\Theta}$, $\overline{B_{\epsilon_{\mathbf{T}}}(\tau(\theta))}$ is a compact set and $\{\tau \in \mathbf{R}^m : \eta \leqslant |\tau - \tau(\theta)|\}$ is a closed set, so that $\varphi(\theta)$, which is their intersection, is compact (e.g., Rudin 1953, Theorem 2.35 and the following Corollary).

Proof of the upper hemicontinuity of φ*.* Because φ is compact valued, we can use the sequential characterization of upper hemicontinuity (e.g., Aliprantis and Border 2006/1999, Theorem 17.20). Let $((\theta_n,\tau_n))_{n\in\mathbf{N}} \in \mathbf{S}^{\mathbf{N}}$ be a sequence s.t., for all $n \in \mathbf{N}$, $\tau_n \in \varphi(\theta_n)$ and $\theta_n \to \bar{\theta} \in \mathbf{\Theta}$ as $n \to \infty$. Now, under Assumptions 1 (a)(b)(d)(e)(g) and (h), Lemma 4iii (p. 28), $\mathbf{S}$ is a compact set, so that there exists a subsequence $((\theta_{\alpha(n)},\tau_{\alpha(n)}))_{n\in\mathbf{N}}$ s.t. $(\theta_{\alpha(n)},\tau_{\alpha(n)})\to(\bar\theta,\bar\tau)\in\mathbf{S}$, as $n\to\infty$. The definition of $\mathbf{S}$ implies that $\bar\tau\in\mathbf{T}(\bar\theta)$. Thus, it remains to show that $\eta\leqslant|\bar\tau-\tau(\bar\theta)|$ in order to conclude that $\bar\tau\in\varphi(\bar\theta)$. By construction, for all $n\in\mathbf{N}$, $\eta\leqslant|\tau_{\alpha(n)}-\tau(\theta_{\alpha(n)})|$. Moreover, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 10iii (p. 33), $\tau:\mathbf{\Theta}\to\mathbf{R}^m$ is continuous, so that $|\tau_{\alpha(n)}-\tau(\theta_{\alpha(n)})|\to|\bar\tau-\tau(\bar\theta)|$ as $n\to\infty$, which means that $\eta\leqslant|\bar\tau-\tau(\bar\theta)|$.

Proof of the lower hemicontinuity of φ. Use the sequential characterization of the lower hemicontinuity (e.g., Aliprantis and Border 2006/1999, Theorem 17.21). Let $(\theta_n)_{n\in\mathbf{N}}\in\mathbf{\Theta}^{\mathbf{N}}$ be a sequence s.t. $\theta_n\to\bar\theta\in\mathbf{\Theta}$ and $\bar\tau\in\varphi(\bar\theta)$. Define the sequence $(\tau_n)_{n\in\mathbf{N}}$ s.t., for all $n\in\mathbf{N}$, $\tau_n=\tau(\theta_n)+\bar\tau-\tau(\bar\theta)$. By definition of the correspondence φ, for all $n\in\mathbf{N}$, $|\tau_n-\tau(\theta_n)|=|\bar\tau-\tau(\bar\theta)|\in[\eta,\epsilon_{\mathbf{T}}]$, which implies that $\tau_n\in\varphi(\theta_n)$. Moreover, under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 10iii (p. 33), $\tau:\mathbf{\Theta}\to\mathbf{R}^m$ is continuous, so that $\lim_{n\to\infty}\tau_n=\lim_{n\to\infty}\tau(\theta_n)+\bar\tau-\tau(\bar\theta)=\tau(\bar\theta)+\bar\tau-\tau(\bar\theta)=\bar\tau$.

(ii) Under Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 10 (p. 33), for all $\theta\in\mathbf{\Theta}$, $\tau(\theta)$ is the unique minimum of the strictly convex minimization problem $\inf_{\tau\in\mathbf{R}^m}\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$. Thus, for all $\theta\in\mathbf{\Theta}$, $v(\theta)>0$. Moreover, by Assumption 1(d), $\mathbf{\Theta}$ is compact, and by statement (i) of the present lemma, $v(.)$ is continuous. Thus, there exists $\varepsilon_v>0$ s.t. $\min_{\theta\in\mathbf{\Theta}}v(\theta)>\varepsilon_v$ because a continuous function over a compact set reaches a minimum (e.g., Rudin 1953, Theorem 4.16). □

Lemma 6 (Asymptotic limit of the variance term)**.** *Under Assumption 1(a)-(b) and (d)-(h),*

(i) $\mathbb{P}$-*a.s. for T big enough,* $0<\inf_{\theta\in\mathbf{\Theta}}\left|\left|\left[\sum_{t=1}^T\hat w_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]\right|_{\det}\right|$;

(ii) $\mathbb{P}$-*a.s. as $T\to\infty$,* $\sup_{\theta\in\mathbf{\Theta}}\left|\Sigma_T(\theta)-\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|=o(1)$

(iii) $\theta\mapsto\Sigma(\theta)$ *and* $\theta\mapsto\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)$ *are continuous in* $\mathbf{\Theta}$

(iv) $\mathbb{P}$-*a.s. for T big enough,* $\inf_{\theta\in\mathbf{\Theta}}|\Sigma_T(\theta)|_{\det}>0$;

(v) $\mathbb{P}$-*a.s. as $T\to\infty$,* $\sup_{\theta\in\mathbf{\Theta}}\left|\ln|\Sigma_T(\theta)|_{\det}-\ln\left|\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|_{\det}\right|=o(1)$, *so that, for all $\eta>0$, $\mathbb{P}$-a.s. as $T\to\infty$,* $\sup_{\theta\in\mathbf{\Theta}}|\frac{1}{2T^\eta}\ln|\Sigma_T(\theta)|_{\det}|=o(1)$.

Proof. (i) Under Assumption 1(a)-(b) and (d)-(h), by Lemma 7 (p. 32), $\mathbb{P}$-a.s. as $T\to\infty$, $\sup_{\theta\in\mathbf{\Theta}}\left|\left[\frac{1}{T}\sum_{t=1}^T\hat w_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|=o(1)$, so that it is sufficient to check the invertibility of $\frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ for all $\theta\in\mathbf{\Theta}$ and the continuity of $\theta\mapsto\frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ (Lemma 13 on p. 37). Firstly, by Assumption 1(h), for all $\theta\in\mathbf{\Theta}$, $\Sigma(\theta):=\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)}{\partial\theta'}\right]^{-1}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ $\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]^{-1}$ is a positive-definite symmetric matrix, and thus $\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ is invertible. Moreover, under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28), and Assumption 1(e), for all $\theta\in\mathbf{\Theta}$, $0<\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]<\infty$, so that $\frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ is invertible for all $\theta\in\mathbf{\Theta}$. Secondly, under Assumption 1(a)-(b), (e)-(f), by Lemma 8i (p. 33), $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}}|\mathrm{e}^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}|\right]<\infty$, so that the Lebesgue dominated convergence theorem and Assumption 1(b) imply the continuity of $(\theta,\tau)\mapsto\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ in $\mathbf{S}$. Moreover, by definition in Assumption 1(e), for all $\theta\in\mathbf{\Theta}$, $(\tau(\theta),\theta)\in\mathbf{S}$, and under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 10iii (p. 33), $\tau:\mathbf{\Theta}\to\mathbf{R}^m$ is continuous. Thus, $\theta\mapsto\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)}{\partial\theta}'\right]$ is continuous. Then, the continuity of

$\theta \mapsto \frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ follows from Lemma 3 (p. 28) under Assumption 1 (a)(b)(d)(e)(g) and (h).

(ii) On one hand, by definition, $\Sigma(\theta) := \left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)}{\partial\theta'}\right]^{-1}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ $\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]^{-1}$, which is symmetric positive definite by Assumption 1 (h), and $\Sigma_T(\theta) :=$ $\left[\sum_{t=1}^T \hat{w}_{t,\theta}\frac{\partial\psi_t(\theta)}{\partial\theta'}\right]^{-1}\left[\sum_{t=1}^T \hat{w}_{t,\theta}\psi_t(\theta)\psi_t(\theta)'\right]\left[\sum_{t=1}^T \hat{w}_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]^{-1}$, which is well defined $\mathbb{P}$-a.s. for T big enough by the statement (i) of the present lemma. On the other hand, under Assumption 1(a)-(b) and (d)-(h), by Lemma 7iii (p. 32), $\mathbb{P}$-a.s. as $T\to\infty$, $\sup_{\theta\in\mathbf{\Theta}}\left|\left[\frac{1}{T}\sum_{t=1}^T \hat{w}_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right] - \frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right| = o(1)$, and, under Assumptions 1(a)-(b), (d)-(e) and (g)-(h), by Lemma 8 (p. 33), $\mathbb{P}$-a.s. as $T\to\infty$, $\sup_{\theta\in\mathbf{\Theta}}\left|\left[\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\psi_t(\theta)\psi_t(\theta)'\right] - \frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]\right| = o(1)$. Thus, the claim follows from the continuity of the inverse transformation (e.g., Rudin 1953, Theorem 9.8) and the limiting functions, and the compactness of $\mathbf{\Theta}$.

(iii) Under Assumption 1(a)-(b), (e)-(g), by Lemma 7i (p. 32) and 8 (p. 33), $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}}|\mathrm{e}^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}|\right] < \infty$ and $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}}|\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'|\right] < \infty$, so that, by the Lebesgue dominated convergence theorem and Assumption 1(b), $(\theta,\tau)\mapsto\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ and $(\theta,\tau)\mapsto\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ are continuous in $\mathbf{S}$. Moreover, by definition in Assumption 1(e), for all $\theta\in\mathbf{\Theta}$, $(\theta,\tau(\theta))\in\mathbf{S}$, and under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 10iii (p. 33), $\tau:\mathbf{\Theta}\to\mathbf{R}^m$ is continuous. Thus $\theta\mapsto\Sigma(\theta)$ is continuous, which is the first result. Under Assumption 1 (a)(b)(d)(e)(g) and (h), the second result follows from Lemma 3 (p. 28), which states that $\theta\mapsto\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]$ is also continuous.

(iv) By construction, $\Sigma_T(\theta)$ is a symmetric positive semi-definite matrix (Lemma 12i on p. 37 with $\mathrm{P} = \frac{1}{T}\sum_{t=1}^T\delta_{X_t}$), so that $|\Sigma_T(\theta)|_{\det}\geqslant 0$. Thus, by the statement (ii) and (iii) of present lemma, it is sufficient to check the invertibility of $\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)$ for all $\theta\in\mathbf{\Theta}$ (Lemma 13 on p. 37). By Assumption 1 (h), for all $\theta\in\mathbf{\Theta}$, $\Sigma(\theta) := \left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)}{\partial\theta'}\right]^{-1}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]\left[\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]^{-1}$ is a positive-definite symmetric matrix, and thus a fortiori invertible. Moreover, under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28), and Assumption 1(e), for all $\theta\in\mathbf{\Theta}$, $0<\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]<\infty$, so that it is also invertible.

(v) Under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28) with $\mathrm{P} = \sum_{t=1}^T\delta_{X_t}$ and by the statement (iv) of the present lemma, $\mathbb{P}$-a.s. for T big enough, $\ln|\Sigma_T(\theta)|_{\det}$ is well-defined in $\mathbf{\Theta}$. Similarly, under Assumption 1 (a)(b)(d)(e)(g) and (h), by Lemma 3 (p. 28) and Assumption 1 (h), $\ln\left|\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|_{\det}$ is well-defined in $\mathbf{\Theta}$. Then, the first part of the result follows from the statement (ii) of the present lemma. Regarding the second part, by the triangle inequality, $\mathbb{P}$-a.s. as $T\to\infty$,

$$
\begin{aligned}
&\frac{1}{T^\eta}\sup_{\theta\in\mathbf{\Theta}}|\ln|\Sigma_T(\theta)|_{\det}| \\
&\leqslant \frac{1}{T^\eta}\sup_{\theta\in\mathbf{\Theta}}\left|\ln\left[|\Sigma_T(\theta)|_{\det}\right] - \ln\left[\left|\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|_{\det}\right]\right| + \frac{1}{T^\eta}\sup_{\theta\in\mathbf{\Theta}}\left|\ln\left[\left|\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|_{\det}\right]\right| \\
&= o(1)
\end{aligned}
$$

where the explanations of the last equality are as follows. Under Assumption 1(a)-(b) and (d)-(h), by the statement (iii) of the present lemma $\theta \mapsto \mathbb{E}[e^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)$ is continuous in $\mathbf{\Theta}$, which is a compact set by Assumption 1(d). Now, continuous functions over compact sets are bounded (e.g., Rudin 1953, Theorem 4.16), so that $\sup_{\theta\in\mathbf{\Theta}}\left|\ln\left[\left|\mathbb{E}[e^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|_{\det}\right]\right|$ is bounded, which, in turn, implies that $\frac{1}{T^\eta}\sup_{\theta\in\mathbf{\Theta}}\left|\ln\left[\left|\mathbb{E}[e^{\tau(\theta)'\psi(X_1,\theta)}]\Sigma(\theta)\right|_{\det}\right]\right| = o(1)$, as $T\to\infty$. Now the last equality follows from the statement (iv) of the present lemma. □

Lemma 7. *Under Assumptions 1(a)-(b) and (e)-(f),*

(i) $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}}|e^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}|\right]<\infty$;

(ii) *under additional Assumption 1(d)(g) and (h), $\mathbb{P}$-a.s. as $T\to\infty$,*
$\sup_{(\theta,\tau)\in\mathbf{S}}\left|\left[\frac{1}{T}\sum_{t=1}^T e^{\tau'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\mathbb{E}\left[e^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|=o(1)$, *so that*
$\sup_{\theta\in\mathbf{\Theta}}\left|\left[\frac{1}{T}\sum_{t=1}^T e^{\tau_T(\theta)'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\mathbb{E}\left[e^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|=o(1)$; *and*

(iii) *under additional Assumption 1(d)(g) and (h), $\mathbb{P}$-a.s. as $T\to\infty$,*
$\sup_{\theta\in\mathbf{\Theta}}\left|\left[\sum_{t=1}^T \hat{w}_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\frac{1}{\mathbb{E}[e^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[e^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|=o(1)$

Proof. (i) It follows from the Cauchy-Schwarz inequality and other standard inequalities.

(ii) By the triangle inequality, as $T\to\infty$ $\mathbb{P}$-a.s.,

$$\begin{aligned}
&\sup_{\theta\in\mathbf{\Theta}}\left|\left[\frac{1}{T}\sum_{t=1}^T e^{\tau_T(\theta)'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\mathbb{E}\left[e^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|\\
&\leqslant \sup_{\theta\in\mathbf{\Theta}}\left|\left[\frac{1}{T}\sum_{t=1}^T e^{\tau_T(\theta)'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\mathbb{E}\left[e^{\tau_T(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|\\
&\qquad+\sup_{\theta\in\mathbf{\Theta}}\left|\mathbb{E}\left[e^{\tau_T(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]-\mathbb{E}\left[e^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|\\
&= o(1)
\end{aligned}$$

where the explanations for the last equality are as follows. Regarding the first supremum, under Assumptions 1 (a)-(b)(d)(e)(g) and (h), by Lemma 4iii (p. 28), $\mathbf{S}:=\{(\theta,\tau):\theta\in\mathbf{\Theta}\wedge\tau\in\mathbf{T}(\theta)\}$ is a compact set, so that Assumptions 1(a)-(b), the statement (i) of the present lemma and the ULLN (uniform law of large numbers) à la Wald (e.g., Ghosh and Ramamoorthi 2003, pp. 24-25, Theorem 1.3.3) imply that, $\mathbb{P}$-a.s. as $T\to\infty$,

$$\sup_{(\theta,\tau)\in\mathbf{S}}\left|\left[\frac{1}{T}\sum_{t=1}^T e^{\tau'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}\right]-\mathbb{E}\left[e^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]\right|=o(1).$$

Now, by Assumption 1(e), for all $\theta\in\mathbf{\Theta}$, $\tau(\theta)\in\mathbf{T}(\theta)$, and under Assumption 1(a)(b), (d)-(e), (g) and (h), by Lemma 2ii (p. 25), $\mathbb{P}$-a.s. for T big enough, for all $\theta\in\mathbf{\Theta}$, $\tau_T(\theta)\in\mathbf{T}(\theta)$. Moreover, under Assumption 1(a)(b), (d)-(e), (g) and (h), by Lemma 2iii (p. 25), $\sup_{\theta\in\mathbf{\Theta}}|\tau_T(\theta)-\tau(\theta)|=o(1)$ $\mathbb{P}$-a.s. as $T\to\infty$. Thus, the first supremum is $o(1)$, i.e., $\sup_{\theta\in\mathbf{\Theta}}|\frac{1}{T}\sum_{t=1}^T e^{\tau_T(\theta)'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}-\mathbb{E}e^{\tau_T(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}|=o(1)$, as $T\to\infty$ $\mathbb{P}$-a.s. Regarding the second supremum, by Assumption 1(b), $(\theta,\tau)\mapsto e^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}$ is continuous. Moreover under Assumptions 1(a)-(b), and (e)-(f), by the statement (i) of the present lemma, $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}}|e^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}|\right]<\infty$. Thus, by the Lebesgue dominated convergence theorem and Assumption 1(b), $(\theta,\tau)\mapsto\mathbb{E}\left[e^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ is also continuous in $\mathbf{S}$. Now, under

Assumptions 1 (a)(b)(d)(e)(g) and (h), by Lemma 4iii (p. 28), $\mathbf{S}$ is compact, so that $(\theta,\tau)\mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right]$ is uniformly continuous in $\mathbf{S}$ —continuous functions on compact sets are uniformly continuous (e.g., Rudin 1953, Theorem 4.19). Thus, under Assumption 1(a)(b), (d)-(e), (g) and (h), by Lemma 2iii (p. 25), which states that $\sup_{\theta\in\mathbf{\Theta}}|\tau_T(\theta)-\tau(\theta)|=o(1)$ $\mathbb{P}$-a.s. as $T\to\infty$, the second supremum is also $o(1)$ $\mathbb{P}$-a.s. as $T\to\infty$.

(iii) Under Assumptions 1 (a)(b)(d)(e)(g) and (h), Lemma 3 (p. 28) yields $0<\inf_{(\theta,\tau)\in\mathbf{S}}\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}$ with $\mathrm{P}=\frac{1}{T}\sum_{t=1}^T\delta_{X_t}$, and $0<\inf_{(\theta,\tau)\in\mathbf{S}}\mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$ with $\mathrm{P}=\mathbb{P}$. Consequently, under Assumption 1(a)(b), (d)-(f), (g) and (h), by Lemma 2iii and iv (p. 25) and the statement (ii) of the present lemma, as $T\to\infty$, $\mathbb{P}$-a.s., uniformly w.r.t. θ

$$\begin{aligned}\sum_{t=1}^T \hat{w}_{t,\theta}\frac{\partial\psi_t(\theta)'}{\partial\theta} &= \frac{1}{\frac{1}{T}\sum_{i=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_i(\theta)}}\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\frac{\partial\psi_t(\theta)'}{\partial\theta}\\ &\to \frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)'}{\partial\theta}\right].\end{aligned}$$

□

Lemma 8. *Under Assumptions 1(a)-(b), (e) and (g),*

(i) $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}^\epsilon}|\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'|\right]<\infty$

(ii) *under additional Assumption 1(d) and (h), $\mathbb{P}$-a.s. as $T\to\infty$,*
$\sup_{(\theta,\tau)\in\mathbf{S}}\left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau'\psi_t(\theta)}\psi_t(\theta)\psi_t(\theta)'-\mathbb{E}\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right|=o(1)$, *so that*
$\sup_{\theta\in\mathbf{\Theta}}\left|\frac{1}{T}\sum_{t=1}^T \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\psi_t(\theta)\psi_t(\theta)'-\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right|=o(1)$

(iii) *under additional Assumption 1(d)(f) and (h), $\mathbb{P}$-a.s. as $T\to\infty$,*
$\sup_{\theta\in\mathbf{\Theta}}\left|\sum_{t=1}^T \hat{w}_{t,\theta}\psi_t(\theta)\psi_t(\theta)'-\frac{1}{\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}]}\mathbb{E}\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right|=o(1).$

Proof. The proof is the same as for Lemma 7 with $\psi(X_1,\theta)\psi(X_1,\theta)'$ and $\psi_t(\theta)\psi_t(\theta)'$ in lieu of $\frac{\partial\psi(X_1,\theta)'}{\partial\theta}$ and $\frac{\partial\psi_t(\theta)'}{\partial\theta}$, respectively. □

Lemma 9. *Under Assumptions 1(a)(b)(g),*

(i) $\mathbb{E}\left[\sup_{\theta\in\mathbf{\Theta}^\epsilon}|\psi(X_1,\theta)|^4\right]<\infty$, *so that* $\mathbb{E}\left[\sup_{\theta\in\mathbf{\Theta}^\epsilon}|\psi(X_1,\theta)|^2\right]<\infty$; *and*

(ii) *under additional Assumption 1(e),* $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}^\epsilon}|\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)|\right]<\infty$

Proof. It follows from standard inequalities, such as the Cauchy-Schwarz inequality and the Jensen's inequality. □

Lemma 10 (Implicit function $\tau(.)$)**.** *Under Assumption 1 (a)(b)(e)(g) and (h),*

(i) *for all $\theta\in\mathbf{\Theta}$, $\tau\mapsto\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\right]$ is a strictly convex function s.t.* $\frac{\partial\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\right]}{\partial\tau}=\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]$;

(ii) *under additional Assumption 1(d), for all $\theta\in\mathbf{\Theta}$, there exists a unique $\tau(\theta)$ such that* $\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\psi(X_1,\theta)\right]=0$; *and*

(iii) *under additional Assumption 1(d), $\tau:\mathbf{\Theta}\to\mathbf{R}^m$ is continuous; and*

(iv) *under additional Assumption 1(c) and (d), for all $\theta\in\mathbf{\Theta}\setminus\{\theta_0\}$,* $\mathbb{E}\left[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}\right]<\mathbb{E}\left[\mathrm{e}^{\tau(\theta_0)'\psi(X_1,\theta_0)}\right]=1$ *where $\tau(\theta_0)=0_{m\times1}$.*

Proof. *(i)* Under Assumption 1(a) and (b), by the Cauchy-Schwarz inequality, $\mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}^\epsilon} \mathrm{e}^{\tau'\psi(X_1,\theta)}\right] \leqslant \mathbb{E}\left[\sup_{(\theta,\tau)\in\mathbf{S}^\epsilon} \mathrm{e}^{2\tau'\psi(X_1,\theta)}\right]^{1/2}$, which is finite by Assumption 1(e). Now, by Assumption 1(e), for all $\dot\theta \in \mathbf{\Theta}$, $\tau(\dot\theta) \in \operatorname{int}[\mathbf{T}(\dot\theta)]$. Then, by a standard result on Laplace's transform (e.g., Monfort 1996/1980, Theorems 3 on p. 183), $\tau \mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\dot\theta)}\right]$ is C^∞ in a neighborhood of $\tau(\dot\theta)$, and $\tau \mapsto \frac{\partial \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\dot\theta)}\right]}{\partial\tau} = \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\dot\theta)}\psi(X_1,\dot\theta)\right]$ and $\tau \mapsto \frac{\partial^2 \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\dot\theta)}\right]}{\partial\tau\partial\tau'} = \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\dot\theta)}\psi(X_1,\dot\theta)\psi(X_1,\dot\theta)'\right]$. Moreover, under Assumptions 1(a)-(b), (e) and (g), Assumption 1(h) implies that, $\mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\dot\theta)}\psi(X_1,\dot\theta)\psi(X_1,\dot\theta)'\right]$ is a symmetric positive-definite matrix because a well-defined covariance matrix is invertible iff it is invertible under an equivalent probability measure (Lemma 12 and Corollary 2i on p. 37).

(ii) Assumption 1(d) ensures existence, while the statement (i) of the present lemma ensures that $\tau(\theta)$ is the solution of a strictly convex problem, so that it is unique.

(iii) Note that, under our assumptions, an application of the standard implicit function (e.g., Rudin 1953, Theorem 9.28) is not directly possible as it requires $(\theta,\tau) \mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]$ to be continuously differentiable in $\mathbf{S}^\epsilon$, which, in turn, typically requires to uniformly bound the derivative of the latter in $\mathbf{S}^\epsilon$ (e.g., Davidson 1994, Theorem 9.31). Thus, we apply the sufficiency part of Kumagai's implicit function theorem (Kumagai 1980). Check its assumptions. Firstly, under Assumptions 1(a)(b)(e) and (g), by Lemma 9ii (p. 33) and the Lebesgue dominated convergence theorem, $(\theta,\tau) \mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]$ is continuous in $\mathbf{S}^\epsilon$, i.e., in an open neighborhood of every $(\theta,\tau) \in \mathbf{S}$. Secondly, by the inverse function theorem applied to $\tau \mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]$ (e.g., Rudin 1953, Theorem 9.24), for all $\theta \in \mathbf{\Theta}^\epsilon$, $\tau \mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]$ is locally one-to-one:[4] As explained in the proof of (i), under Assumption 1(a)(b)(e) and (h), $\tau \mapsto \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]$ is continuously differentiable and, under Assumption 1(a)(b)(e)(g) and (h), for all $\theta \in \mathbf{\Theta}$, $\frac{\partial \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\right]}{\partial\tau'} = \mathbb{E}\left[\mathrm{e}^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ is invertible, so that the assumptions of the inverse function theorem are valid.

(iv) By the statements (i) and (ii) of the present lemma, for all $\theta \in \mathbf{\Theta}$, for all $\tau \neq \tau(\theta)$, $\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}] < \mathbb{E}[\mathrm{e}^{\tau'\psi(X_1,\theta)}]$. Now, for all $\theta \in \mathbf{\Theta}\setminus\{\theta_0\}$, $\tau(\theta) \neq 0_{m\times 1}$: If there existed $\dot\theta \in \mathbf{\Theta}\setminus\{\theta_0\}$ s.t. $\tau(\dot\theta) = 0_{m\times1}$, then $0 = \mathbb{E}[\mathrm{e}^{\tau(\dot\theta)'\psi(X_1,\dot\theta)}\psi(X_1,\dot\theta)] = \mathbb{E}[\psi(X_1,\dot\theta)]$, which would contradict Assumption 1(c). Thus, for all $\theta \in \mathbf{\Theta}\setminus\{\theta_0\}$, $\mathbb{E}[\mathrm{e}^{\tau(\theta)'\psi(X_1,\theta)}] < \mathbb{E}[\mathrm{e}^{0_{1\times m}\psi(X_1,\theta)}] = 1$. Then, the result follows by the statement (ii) of the present lemma because $0_{m\times1} = \mathbb{E}[\psi(X_1,\theta_0)] = \mathbb{E}[\mathrm{e}^{0_{1\times m}\psi(X_1,\theta)}\psi(X_1,\theta_0)]$. □

A.2. Implications of Assumption 1(h).

Lemma 11. *Let $(\mathbf{\Omega}_A,\mathcal{A})$ be a measurable space, $Z : \mathbf{\Omega} \to \mathbf{R}^k$ be a k-dimensional random vectors with $k \in [\![1,\infty[\![$ and P and Q two probability measures on $(\mathbf{\Omega}_A,\mathcal{A})$. Denote the expectation and the variance under P with $\mathbb{E}_\mathrm{P}$ and $\mathbb{V}_\mathrm{P}$, respectively.*

[4]Here it is necessary to work in an ϵ-neighborhood of $\mathbf{\Theta}$ in order to satisfy the assumption of Kumagai's implicit function theorem (Kumagai 1980). The standard implicit function theorem would also require the existence of open neighborhoods around the parameter values at which the function is zero.

(i) *For all* $\tau \in \mathbf{R}^k$, $\mathbb{E}_{\mathrm{P}}\left(\mathrm{e}^{\tau' Z} Z Z'\right) \geqslant 0$, *it is a positive semi-definite symmetric matrix.*

(ii) *If* $\mathrm{P} \sim \mathrm{Q}$ *(i.e., they are equivalent),* $\mathbb{E}_{\mathrm{P}}(|ZZ'|) < \infty$ *and* $\mathbb{E}_{\mathrm{Q}}(|ZZ'|) < \infty$, *then*

$$\mathbb{E}_{\mathrm{P}}(ZZ') \text{ invertible} \Leftrightarrow \mathbb{E}_{\mathrm{Q}}(ZZ') \text{ invertible}$$

Corollary 1 (Implication of Assumption 1(h))**.** *Under Assumptions 1(a)-(b), (e) and (g), Assumption 1(h) implies that, for all* $(\theta, \tau) \in \mathbf{S}$, $\mathbb{E}\left[\mathrm{e}^{\tau' \psi(X_1,\theta)} \psi(X_1, \theta) \psi(X_1, \theta)'\right]$ *is a positive definite symmetric matrix.*

Appendix B. On the assumptions

B.1. **Discussion.** Assumptions 1 and 2 are mainly adapted from the entropy literature. Assumption 1(a) ensures the basic requirement for inference, that is, data contain different pieces of information (independence) about the same phenomenon (identically distributed). The conditions "independence and identically distributed" are much stronger than needed, and can be relaxed to allow for time dependence along the lines of Kitamura and Stutzer (1997). We restrain ourself to the i.i.d. case for brevity and clarity. Assumption 1(a) also requires completeness of the probability space so that we can define functions only on a probability-one subset of $\mathbf{\Omega}$ without generating potential measurability complications. The completeness of the probability space is without significant loss of generality (e.g., Kallenberg 2002 (1997, p. 13), and it is often implicitly or explicitly required in the literature.

Assumption 1(b) mainly requires standard regularity conditions for the moment function $\psi(.,.)$. The existence of the estimator relies on such regularity conditions. An alternative would be to rely on empirical process theory, but it seems here inappropriate as the implicit nature of the definition of the ESP approximation requires smooth functions. We require Assumption 1(b), as well as some of the following assumptions, to hold in an ϵ-neighborhood of the parameter space $\mathbf{\Theta}$, so that we can deal with its boundary $\partial\mathbf{\Theta}$ in the same way as with its interior. In particular, it ensures that $\Sigma(\theta)$ is invertible for $\theta \in \partial\mathbf{\Theta}$ under probability measures equivalent to $\mathbb{P}$ (Corollary 2ii on p. 37), and it allows to apply an implicit function theorem to $\tau(\theta)$, also for $\theta \in \partial\mathbf{\Theta}$ (Lemma 10 on p. 33). For the latter reason, the entropy literature often appears to also (implicitly) assume that assumptions hold in an ϵ-neighborhood of the parameter space. In applications, this is often innocuous as the boundary of the parameter space is often loosely specified. However, in some specific situations, which we rule out, this may be problematic (e.g., Andrews 1999, and references therein).

Assumption 1(c) requires global identification, which is a necessary condition to prove the consistency of an estimator. If we were interested in the ESP approximation instead of its maximizer (i.e., the ESP estimator), global identification could be relaxed as Holcblat (2012) and a companion paper show. Assumption 1(c) also requires equality between the dimension of the parameter space and the number of moment conditions, i.e., just-identified moment conditions. We impose the latter for mainly three reasons. Firstly, it appears reasonable to investigate the ESP estimator in the just-identified case before moving to the more complicated over-identified case. Secondly, the just-identified case makes clear the difference between the ESP estimator and the existing alternatives, which are all equal in this case (see Section 2.2). Thirdly, this is a standard assumption in the saddlepoint literature. However, note that (i) this assumption is less restrictive than it seems at first sight because, in the linear case, over-identified moment conditions correspond to just-identified moment conditions through the FOCs (Imbens 1997, Ronchetti and Trojani 2001) or the Sowell (1996)'s reparameterization thereof (Sowell 2007, 2009, Czellar and Ronchetti 2010), and, in the nonlinear case, we can transform over-identified estimating equations into just-identified estimating equations through an extension of the parameter space (Holcblat (2012) who follows Newey and McFadden (1994, p. 2232)); (iii) ongoing work show how to generalize the ESP approximation to over-identified moment conditions without transformation.

Assumption 1(d) requires the compactness of the parameter space $\mathbf{\Theta}$, and the existence of a solution $\tau(\theta) \in \mathbf{R}^m$ that solves the equation $\mathbb{E}\left[e^{\tau' \psi(X_1,\theta)}\psi(X_1,\theta)\right] = 0$, for all $\theta \in \mathbf{\Theta}$. Schennach (2005) also makes this assumption. Compactness of the parameter space is a convenient standard mathematical assumption that is often relevant in practice. A computer can only handle a bounded parameter space —finite memory of a computer. Regarding the existence of $\tau(\theta)$, it is necessary to ensure the asymptotic existence of the ESP approximation. From a theoretical point of view, the existence of $\tau(\theta)$ looks like a reasonable assumption: If, for some $\theta \in \mathbf{\Theta}$, $0_{m\times 1}$ is outside the convex hull of the support of $\psi(X_1, \theta)$, there is not such a solution $\tau(\theta)$, which also means that θ cannot be θ_0, so that it should be excluded from the parameter space. However, the existence of $\tau(\theta)$ might be difficult to check in practice. A way to get around this assumption is to (i) assume the existence of $\tau(\theta)$ only in a neighbohood of θ_0; and (ii) to set the ESP approximation to zero for the θ values that do not have a solution to the finite-sample moment conditions (5). Holcblat (2012) follows such an approach. We do not follow such an approach because it significantly complicates the proofs and the presentation.

Assumptions 1(e) and 2(b) rule out fat-tailed distributions. More precisely, they require the existence of exponential moments. They are necessary to apply the the ULLN (uniform law of large numbers) à la Wald (e.g., Ghosh and Ramamoorthi 2003, pp. 24-25, Theorem 1.3.3) to components of the ESP approximation. Assumptions 1(e) and 2(b) are stronger than the moment existence assumption in Hansen (1982), but they are a common type of assumptions in the entropy literature (e.g., Haberman 1984, Kitamura and Stutzer 1997, Schennach 2007), the saddlepoint literature (e.g., Almudevar et al. 2000) and the literature on exponential models (e.g., Berk 1972). In particular, Assumptions 1(d) and 2(b) are a convenient variant of Assumptions 3.4 and 3.5 in Schennach (2007). Both in Schennach (2007) and in the present paper, the successful estimation of the Hall and Horowitz model, which does not satisfy Assumptions 1(e) and 2(b), suggests that the latter can be relaxed.[5] In practice, Assumptions 1(e) and 2(b) are not as strong as it may appear because observable quantities have finite support (finite memory of computers), which, in turn, implies that they have all finite moments. Moreover, in the case in which unboundedness is a concern (e.g., moment conditions derived from a likelihood), Ronchetti and Trojani (2001) provide a way to bound moment functions.

Assumptions 1(f) and (g) play the same role as Assumptions 1(e) and 2(b), although they are less stringent. Assumption 1(h) requires the invertibility of the asymptotic variance of standard

[5]For information, we checked that our simulation results (i.e., Table 1) are qualitatively the same for a truncated Hall and Horowitz model, in which Assumption 1(e) and 2(b) hold. The robustness of the results is not surprising. The violation of Assumption 1(e) and 2(b) should —if anything— put the ESP estimator at disadvantage w.r.t. the MM estimator, which is known to not require Assumption 1(e) and 2(b).

estimators (scaled by $\sqrt{T}$) of any solution to the tilted moment condition. In the present paper, this assumption has two main roles. Firstly, it ensures that the determinant term $|\Sigma_T(\theta)|_{\det}^{-\frac{1}{2}}$ in the ESP approximation (2) does not explode, asymptotically. Secondly, it ensures the positive definiteness of the symmetric matrix $\mathbb{E}\left[e^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ for all $(\theta,\tau)\in\mathbf{S}$, so that the $\min_{\tau\in\mathbf{R}^m}\mathbb{E}[e^{\tau'\psi(X_1,\theta)}]$ is a strictly convex problem, which, in turn, implies the unicity of its solution $\tau(\theta)$. In the setup of the present paper, Assumption 1(g) is equivalent to the invertibility of $\mathbb{E}\left[e^{\tau(\theta)'\psi(X_1,\theta)}\frac{\partial\psi(X_1,\theta)}{\partial\theta'}\right]$ and $\mathbb{E}\left[\psi(X_1,\theta)\psi(X_1,\theta)'\right]$, for all $\theta\in\mathbf{\Theta}$ (Lemma 12 on p. 37 with $\mathrm{P}=\mathbb{P}$ and $\frac{\mathrm{dQ}}{\mathrm{dP}}=\frac{1}{e^{\tau(\theta)'\psi(X_1,\theta)}}$). In this way, it is stronger than the Assumption 4 in Kitamura and Stutzer (1997), but it is close to Stock and Wright (2000, Assumption C). Note that Schennach (2007) also implicitly assumes that $\mathbb{E}\left[e^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ is full rank for all $(\theta,\tau)\in\mathbf{S}$, because Schennach (2007, p. 649) regards $\tau(\theta)$ as a solution to a strictly convex problem (e.g., Hiriart-Urruty and Lemaréchal 1993/1996, chap. 4, Theorem 4.3.1). Assumption 1(g) should often hold because the set of singular matrices has zero Lebesgue measure in the space of square matrices.[6]

B.2. Implications of Assumption 1(h).

Lemma 12. *Let $(\mathbf{\Omega}_A,\mathcal{A})$ be a measurable space, $Z:\mathbf{\Omega}\to\mathbf{R}^k$ be a k-dimensional random vectors with $k\in[\![1,\infty[\![$ and P and Q two probability measures on $(\mathbf{\Omega}_A,\mathcal{A})$. Denote the expectation and the variance under P with $\mathbb{E}_{\mathrm{P}}$ and $\mathbb{V}_{\mathrm{P}}$, respectively.*

(i) *For all $\tau\in\mathbf{R}^k$, $\mathbb{E}_{\mathrm{P}}\left(e^{\tau'Z}ZZ'\right)\geqslant 0$, it is a positive semi-definite symmetric matrix.*
(ii) *If $\mathrm{P}\sim\mathrm{Q}$ (i.e., they are equivalent), $\mathbb{E}_{\mathrm{P}}(|ZZ'|)<\infty$ and $\mathbb{E}_{\mathrm{Q}}(|ZZ'|)<\infty$, then*

$$\mathbb{E}_{\mathrm{P}}(ZZ')\ \textit{invertible}\Leftrightarrow\mathbb{E}_{\mathrm{Q}}(ZZ')\ \textit{invertible}$$

Corollary 2 (Implication of Assumption 1(h))**.** *Under Assumptions 1(a)-(b), (e) and (g), Assumption 1(h) implies that, for all $(\theta,\tau)\in\mathbf{S}$, $\mathbb{E}\left[e^{\tau'\psi(X_1,\theta)}\psi(X_1,\theta)\psi(X_1,\theta)'\right]$ is a positive definite symmetric matrix.*

Appendix C. Remaining technical results

Lemma 13 (Asymptotic invertibility of a sequence of matrix functions)**.** *Let $A(\gamma)$ be a family of invertible matrices indexed by $\gamma\in\mathbf{\Gamma}$ s.t. $\gamma\mapsto A(\gamma)$ is continuous, and where $\mathbf{\Gamma}$ is a compact subset of a Euclidean space. Let $(A_T(\gamma))_{T\in[\![1,\infty[\![}$ be a sequence of square matrices. If, as $T\to\infty$, $\sup_{\gamma\in\mathbf{\Gamma}}|A_T(\gamma)-A(\gamma)|\to 0$, then there exist a constant $\varepsilon_A>0$ and $T_A\in[\![1,\infty[\![$ s.t. for all $T\in[\![T_A,\infty[\![$, for all $\gamma\in\mathbf{\Gamma}$, $||A_T(\gamma)|_{\det}|\geqslant\varepsilon_A$.*

Lemma 14 (Asymptotic positivity and definiteness of matrices)**.** *Let $(A_T)_{T\geqslant 1}$ a sequence of square matrices converging to A as $T\to\infty$.*[7] *Then, if $(A_T)_{T\geqslant 1}$ is a sequence of symmetric matrices and A is a positive-definite matrix (p-d.m), then there exists $\dot{T}\in[\![1,\infty[\![$ such that $T\geqslant\dot{T}$ implies A_T is p-d.m.*

[6]The set of singular matrices corresponds to the set of zeros of the determinant, which is nonzero polynomial in several variables. Moreover, by induction over the number of variables with the fundamental theorem of algebra for the base step, a nonzero polynomials has a finite number of zeros.
[7]Note that we do not need to specify the norm as all norms are equivalent in finite-dimensional spaces.

Appendix D. On the shrinkage

In this appendix, we formalize that the ESP estimator $\hat{\theta}_T$ corresponds to an ET estimator θ_T^* (or equivalently a MM estimator) shrunk toward parameter values with lower implied estimated variance. More precisely, we provide a proof of Proposition 1 (p. 6), which corresponds to the upcoming Proposition 2 put in context.

Definition 1 (Loewner partial order, Partial ordering of real positive semi-definite symmetric matrices)**.** *A real positive semi-definite symmetric matrix A of size m is said to be larger than another real symmetric matrix B of size m if one of the following equivalent conditions holds*

(a) $x'Bx \leqslant x'Ax$*, for all* $x \in \mathbf{R}^m$*;*

(b) *for all* $i \in [\![1, m]\!]$, $\lambda_{B,i} \leqslant \lambda_{A,i}$ *where* $(\lambda_{A,1}, \lambda_{A,2}, \cdots, \lambda_{A,m})$ *and* $(\lambda_{B,1}, \lambda_{B,2}, \cdots, \lambda_{B,m})$ *are the ordered spectra of A and B s.t.* $0 \leqslant \lambda_{A,1} \leqslant \lambda_{A,2} \leqslant \cdots \leqslant \lambda_{A,m}$ *and* $0 \leqslant \lambda_{B,1} \leqslant \lambda_{B,2} \leqslant \cdots \leqslant \lambda_{B,m}$.

Lemma 15 (Compability of the Loewner partial order with determinant)**.** *Under Assumption 1(a)(b) and (d)-(h), $\mathbb{P}$-a.s. for T big enough, for all $(\dot{\theta}, \ddot{\theta}) \in \mathbf{\Theta}^2$, if the real positive semi-definite symmetric matrix $\Sigma_T(\dot{\theta})$ is larger than $\Sigma_T(\ddot{\theta})$, then $|\Sigma_T(\ddot{\theta})|_{\det} \leqslant |\Sigma_T(\dot{\theta})|_{\det}$.*

Proof. Under Assumption 1(a)(b) and (d)-(h), by Lemma 6 (p. 25) of the paper, $\mathbb{P}$-a.s. for T big enough, $\Sigma_T(\theta)$ is well-defined on $\mathbf{\Theta}$. Then, the results follows from standard bilinear algebra (e.g., Gourieroux and Monfort 1995/1989, Corollary A.2). □

Proposition 2 (Smaller estimated variance of the ESP estimator)**.** *Under Assumption 1(a)(b) and (d)-(h), $\mathbb{P}$-a.s. for T big enough,*

$$|\Sigma_T(\hat{\theta}_T)|_{\det} < |\Sigma_T(\theta_T^*)|_{\det},$$

where $\hat{\theta}_T$ and θ_T^* respectively denote an ESP and an ET estimator s.t. $\hat{\theta}_T$ is different from any ET estimator, i.e., $\hat{\theta}_T \notin \arg\max_{\theta \in \mathbf{\Theta}} \ln \left[\frac{1}{T} \sum_{t=1}^{T} \mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]$.

Proof. Under Assumption 1(a)(b) and (d)-(h), by Lemma 1i and iv (p. 20) $\mathbb{P}$-a.s. for T big enough, the ESP estimator $\hat{\theta}_T$ and the ESP approximation $\hat{f}_{\theta_T^*}(\theta)$ exist, which in turn implies that the ET estimator θ_T^* also exists. Then, the result immediately follows from the structure of the ESP objective function. □

Remark 4. (i) The penalty $-\frac{1}{2T} \ln |\Sigma_T(\theta)|_{\det}$ can also be analyzed in terms of entropy (e.g., Cover and Thomas 2006/1991, Theorem 8.4.1). However, we refrain from such an analysis in the paper to avoid any confusion with the relative entropy (i.e., Kullback-Leibler Divergence Criterion) minimization used to compute the multinomial distribution $(w_{t,\theta})_{t=1}^T$. (ii) The estimated variance $|\Sigma_T(\hat{\theta}_T)|_{\det}$ of the ESP estimator is strictly smaller than the estimated variance of the ET estimator $|\Sigma_T(\theta_T^*)|_{\det}$ w.r.t. the order induced by the determinant, while it is smaller or equivalent w.r.t. the Loewner partial order.

Appendix E. More on numerical examples

E.1. **Computational details.** The simulations were performed in R. Each model parameterization is simulated 10,000 times. The robustness of the simulation results was checked with different optimization routines, starting values and tolerance parameter values. The computational complexity of the ESP estimator is comparable to the ET estimator, which is usually considered the easiest empirical-likelihood-type estimator to compute (e.g., Anatolyev and Gospodinov 2011, Section 2.2.6). The tilting equation (5) is the same for the ET estimator and ESP estimators. Once it is solved, the ESP estimator just requires the additional evaluation of the variance term. There are no intense additional calculations.

The estimation for a single sample is typically performed in less than a few seconds. The calculations were done on a 24 CPU cores of a Dell server with 4 AMD Opteron 8425 HE processors running at 2.1 GHz. In accordance with the empirical likelihood literature (e.g., Kitamura 2007, Section 8.1; Schennach 2007, Theorem 1) parameter values where the objective functions are undefined are considered inadmissible for the parameter estimates. We numerically checked that the ET and MM estimates are the same even for the small sample sizes $T = 25$ and 50. We numerically checked that the reported statistics have a converging behaviour as we increase the number of simulated samples to 10,000.

E.2. **One-parameter Hall and Horowitz model.** We simulate the one-parameter just identified Hall and Horowitz model, using only the second moment condition

$$\mathbb{E}\left[Y_t\Big(\exp\left\{\mu_0-\beta\left(X_t+Y_t\right)+3Y_t\right\}-1\Big)\right]$$

i.e., we assume that $\mu_0 = -.72$ is known. The first moment condition cannot be used because it has two solutions. The reported statistics are based on 10,000 simulated samples.

Table 4 (p. 40) displays MSE, bias and variance of the ESP and ET estimators. The results are similar to the results for the two-parameter Hall and Horowitz model: Compare with Table 1 on p. 11. In particular, again we help the ET estimators by restricting its parameter space. Table 5 (p. 40) displays the actual rejection probabilities of the ESP test statistics for the null hypothesis $\mathrm{H}_0 : \beta = 3$. The Tilt test appears to perform best in terms of actual rejection probability. We do not report the LM test for the reasons previously mentioned in Section 4.1.3. Table 6 (p. 41) reports the average length for the confidence intervals deduced from the inversion of the test statistics. Both ALR and Wald confidence intervals are significantly shorter than Tilt confidence intervals. Thus, the ALR confidence intervals appear as a good compromise in terms of average length and actual rejection probabilities.

Appendix F. Additional empirical evidence

Table 7 (p. 42) is the same as Figure 1 (p. 15) with the additional Table 7 Figures (A). The latter clearly shows that the normalized ESP is relatively sharp around the ESP estimator. Table 9 (p. 44) is the counterpart of Table 7 (p. 42) for the 1930-2009 data set. The 95% ET ALR confidence region is based on the inversion of the ALR ET statistic $2T\left[\mathrm{LogET}(\hat{\theta}_T)-\mathrm{LogET}(\theta_0)\right]$ $= 2T\mathrm{LogET}(\theta_0) \to \chi^2_1$ (Kitamura and Stutzer 1997, Theorem 4 with $K = 0$ and $\mathrm{H}_0 : \theta = \theta_0$), where $\mathrm{LogET}(\theta) := \ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]$ and $\mathrm{LogET}(\hat{\theta}_T) = \ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\hat{\theta}_T)'\psi_t(\hat{\theta}_T)}\right] = 0$ because, in the just-identified case, $\frac{1}{T}\sum_{t=1}^{T}\psi_t(\hat{\theta}_T) = 0$ so that $\tau_T(\hat{\theta}_T) = 0_{m\times 1}$. The ET and ESP support correspond to the parameter values $\theta \in \mathbf{\Theta}$ for which there exists a solution $\tau_T(\theta)$ to the equation (5) on p. 4. Table 9 confirms the findings of Table 7 (p. 42) in Section 4.2: The ESP

TABLE 4. **ESP vs. ET estimator for the just-identified one-parameter Hall and Horowitz model.**

T		β ET	β ESP
25	MSE	3.3394	0.9231
	Bias	0.4264	0.0514
	Var.	3.1576	0.9204
50	MSE	1.7245	0.4687
	Bias	0.2730	0.0454
	Var.	1.6499	0.4666
100	MSE	0.6897	0.1978
	Bias	0.1480	0.0240
	Var.	0.6678	0.1973
200	MSE	0.2059	0.0893
	Bias	0.0617	0.0036
	Var.	0.2021	0.0893
500	MSE	0.0514	0.0367
	Bias	0.0259	0.0027
	Var.	0.0507	0.0367
1000	MSE	0.0292	0.0189
	Bias	0.0126	-0.0006
	Var.	0.0291	0.0189
5000	MSE	0.0041	0.0039
	Bias	0.0027	0.0002
	Var.	0.0041	0.0039

Note: We estimate β using the moment condition $\mathbb{E}[Y_t\,(\,\exp\{\mu_0-\beta\,(X_t+Y_t)+3Y_t\}-1)]$, i.e., we assume that $\mu_0=-.72$ is known. The reported statistics are based on 10,000 simulated samples of sample size equal to the indicated T. For ET, the parameter space is restricted to $-5<\beta<15$ in order to limit the erratic behaviour of the estimator at sample sizes $T=25$ and 50. No such parameter restriction is imposed for ESP.

TABLE 5. **Actual rejection probabilities for the ESP test statistics with the just-identified one-parameter Hall and Horowitz model.**

T	ALR_T	Wald_T	Tilt_T
25	0.1030	0.1447	0.0603
50	0.0863	0.1108	0.0548
100	0.0713	0.0905	0.0493
200	0.0670	0.0794	0.0500
500	0.0628	0.0656	0.0529
1000	0.0595	0.0608	0.0512
5000	0.0557	0.0558	0.0526

Note: Under the null hypothesis $H_0:\beta=3$, asymptotically the test statistics follow a chi-square distribution with one degree of freedom. The tests used the critical value with size of $\alpha=.05$. The probabilities are based on 10,000 simulated samples of sample size equal to the indicated T.

is sharper than the ET around its maximum, so that the ESP confidence region is also shorter. Note also that the ESP estimate is almost the same as for the data set 1890-2009. These results

TABLE 6. **Average length of the confidence intervals for the ESP test statistics with the one-parameter Hall and Horowitz model.**

T	ALR_T	Wald_T	Tilt_T
25	3.7284	3.3339	6.2758
50	2.7146	2.4709	4.9978
100	1.9617	1.8795	3.3012
200	1.4161	1.5448	1.9116
500	0.9556	1.3938	0.8967
1000	0.7811	1.5013	0.5717

Note: The parameter β is estimated with the moment condition $\mathbb{E}[Y_t\,(\,\exp\{-.72-\beta\,(X_t+Y_t)+3Y_t\}-1)]$. Under the null hypothesis $H_0 : \beta = 3$, asymptotically the test statistics follow a chi-square distribution with one degree of freedom. The test statistics are inverted to obtain confidence intervals. The confidence intervals use the critical value with size of $\alpha = .05$. The lengths are based on 10,000 simulated samples of sample size equal to the indicated T.

are in line with the ESP shrinkage-like behaviour documented in the Monte-Carlo simulations of the section 4.1.

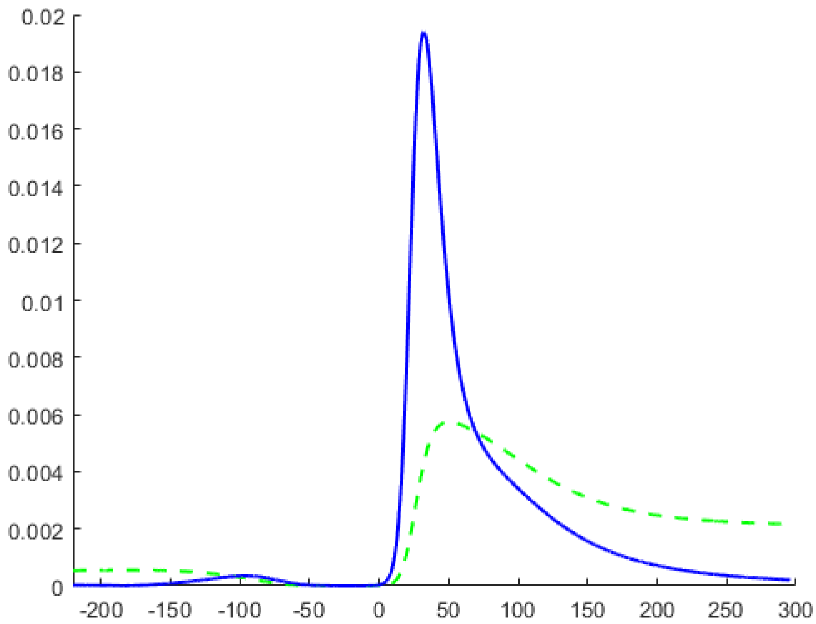
Tables 8 (43) and 10 (45) report the trinity+1-based 95% confidence intervals of the ESP and ET. Except for the ALR-based and the LM-based confidence regions, the theoretical comparison of ESP and ET in Section 2 does not provide much indication about whether ESP-based confidence regions should be better. While the ALR-based and the LM-based confidence regions account for the global shape of the objective function, the Wald-based and ET-based confidence regions only account for local features of the objective function, and thus they do not really take advantage of the variance term. In line with the analysis of the variance term, ESP LM-based and ALR-based confidence regions appear sharper than their ET counterparts. Nevertheless, as it is often the case with nonlinear model, LM-based confidence regions, which rely on the distance from zero of the derivative of the objective function, are contaminated by local extrema so they are less reliable than ALR-based confidence regions. Moreover, LM-based confidence regions rely on the derivative of the log-ESP approximation —instead of the logESP itself— so they incorporate less information than ALR-based confidence regions. ET and ESP ET-based confidence regions only differ through the estimation of $\mathbb{E}[\psi(X_1,\theta_0)\psi(X_1,\theta_0)]$ so they are very close. In the same vein, ET and ESP Wald-based confidence regions are similar

Tables 11 (p. 45) and 13 (p. 46) report the MM estimates and the confidence regions based on the inversion of the MM ALR test statistic $T\left[Q_{\text{MM},T}(\theta_0)-Q_{\text{MM},T}(\hat\theta_{\text{MM},T})\right]=TQ_{\text{MM},T}(\theta_0)\xrightarrow{D}\chi^2_1$, as $T\to\infty$, (e.g., Newey and McFadden 1994, Theorem 9.2), where $Q_{\text{MM},T}(\theta):=\left[\frac{1}{T}\sum_{t=1}^T\psi_t(\theta)\right]'$ $\times\left[\frac{1}{T}\sum_{t=1}^T\psi_t(\hat\theta_{\text{MM},T})\psi_t(\hat\theta_{\text{MM},T})'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^T\psi_t(\theta)\right]$ and $Q_{\text{MM},T}(\hat\theta_{\text{MM},T})=0$ because $\frac{1}{T}\sum_{t=1}^T\psi_t(\hat\theta_T)=0$ in the just-identified case. The MM objective function is sharper around its minimum for the 1930-2009 data set than for the 1890-2009. However, the former sharpness appears misleading as it yields a confidence region that does not include the MM estimate of the 1890-2009 data set.

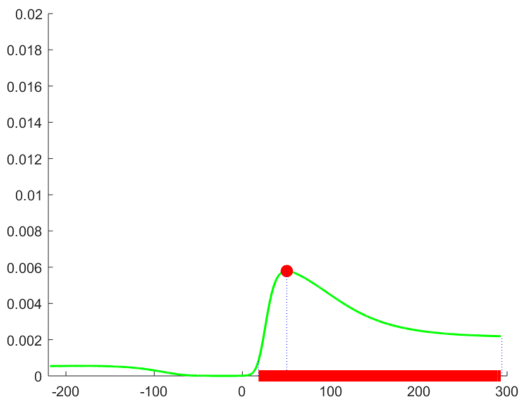
Tables 12 (p. 46) and 14 (p. 47) report the CU (continuously updating) MM estimates and the confidence regions based on the inversion of the CU ALR test statistic $T\left[Q_{\text{CU},T}(\theta_0)-Q_{\text{CU},T}(\hat\theta_{\text{MM},T})\right]=TQ_{\text{CU},T}(\theta_0)\to\chi^2_1$, as $T\to\infty$, where

Table 7. **ET vs. ESP inference (1890–2009)**

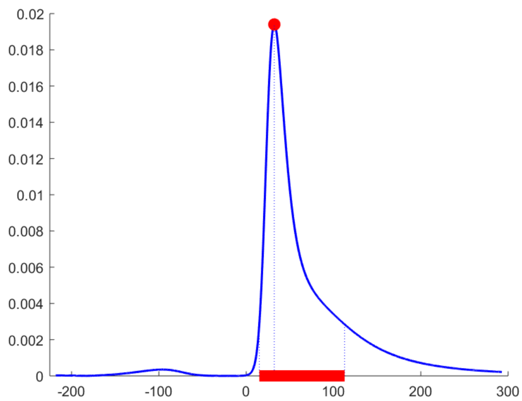
Empirical moment condition: $\frac{1}{2009-1889}\sum_{t=1890}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0$, where $R_{m,t}:=$ gross market return, $R_{f,t}$:=risk-free asset gross return, $C_t:=$ consumption, and θ :=relative risk aversion;
Normalized ET:=$\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(.)'\psi_t(.)}\right]\right\}/\int_\Theta\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}\mathrm{d}\theta$;
Normalized ESP:=$\hat{f}_{\theta_T^*}(.)/\int_\Theta\hat{f}_{\theta_T^*}(\theta)\mathrm{d}\theta$;
$\hat{\theta}_{\mathrm{ET},T}=\hat{\theta}_{\mathrm{MM},T}=50.3$ (bullet) and $\hat{\theta}_{\mathrm{ESP},T}=32.21$ (bullet);
ET and ESP support $=[-218.2,289.0]$; 95% ET ALR conf. region=$[18.3,289.0]$ (stripe); 95% ESP ALR conf. region=$[15.0,112.7]$ (stripe).



(A) Normalized ET (dashed green) vs. normalized ESP (dark blue).



(A) ET est. and ALR conf. region.



(B) ESP est. and ALR conf. region.

Note: The boundaries of the ESP and ET are indicative. The exact boundaries depend on the optimization routine.

Table 8. **ET vs. ESP 95% confidence regions (1890–2009)**

	ET	ESP
Wald conf. region	$[-26.9, 127.4]$	$[-76.2, 140.6]$
LM conf. region	$[-218.2, -83.6] \cup [-25.6, -14.6]$ $\cup[32.1, 289.0]$	$[-218.2, -202.1] \cup [-135.9, -79.6]$ $\cup[28.6, 38.4] \cup [58.0, 289.0]$
ALR conf. region	$[18.3, 289.0]$	$[15.0, 112.7]$
Tilt conf. region	$[-218.2, -65.8] \cup [34.1, 300.0]$	$[-218.2, -48.4] \cup [25.0, 300.0]$

$Q_{\mathrm{CU},T}(\theta) := \left[\frac{1}{T}\sum_{t=1}^{T}\psi_t(\theta)\right]' \left[\frac{1}{T}\sum_{t=1}^{T}\psi_t(\theta)\psi_t(\theta)'\right]^{-1} \left[\frac{1}{T}\sum_{t=1}^{T}\psi_t(\theta)\right]$ and $Q_{\mathrm{CU},T}(\hat{\theta}_{\mathrm{CU},T}) = 0$ because $\frac{1}{T}\sum_{t=1}^{T}\psi_t(\hat{\theta}_T) = 0$ in the just-identified case. In the just-identified case, which is the case addressed in the present paper, such confidence regions correspond to the S-sets, which were proposed by Stock and Wright (2000) —following Hansen et al. (1996, (c) Constrained-Minimized)— as a solution to the flatness of GMM objective functions. As previously documented in the literature (e.g., Hansen et al. 1996), CU GMM objective functions tend to be flat and low in the tails. Thus, the CU ALR confidence regions (and S-sets in the just-identified case) are huge, and hardly informative.

University of Luxembourg, FDEF, 6 Rue Richard Coudenhove-Kalergi, L-1359, Luxembourg.

E-mail address: Benjamin.Holcblat@uni.lu

Carnegie Mellon University, Tepper School of Business, 5000 Forbes Ave. Pittsburgh, PA, USA

E-mail address: fs0v@andrew.cmu.edu

TABLE 9. **ET vs. ESP inference (1930–2009)**

Empirical moment condition: $\frac{1}{2009-1889}\sum_{t=1890}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0$, where $R_{m,t}:=$ gross market return, $R_{f,t}:=$risk-free asset gross return, $C_t:=$ consumption, and θ :=relative risk aversion;

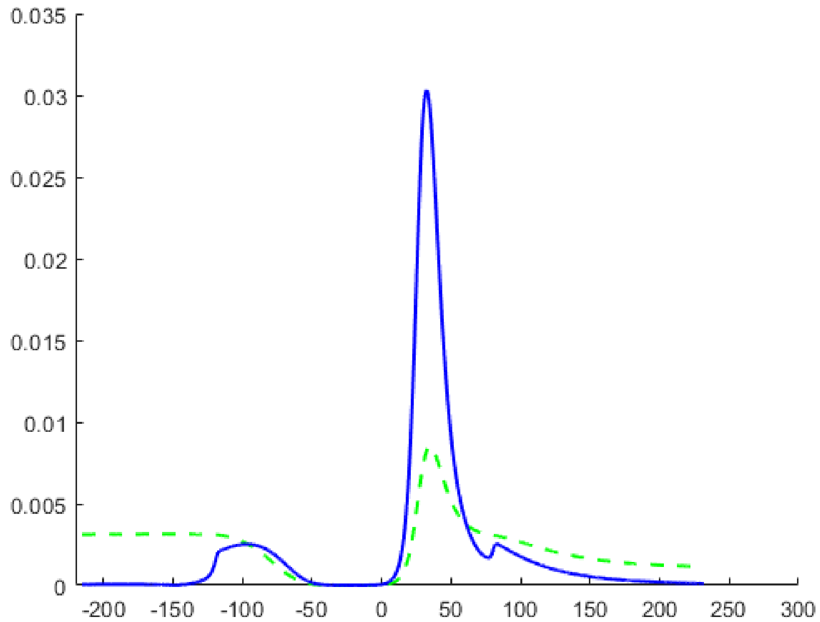
Normalized ET:=$\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(.)'\psi_t(.)}\right]\right\}/\int_\Theta\exp\left\{T\ln\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{e}^{\tau_T(\theta)'\psi_t(\theta)}\right]\right\}\mathrm{d}\theta$;

Normalized ESP:=$\hat{f}_{\theta_T^*}(.)/\int_\Theta\hat{f}_{\theta_T^*}(\theta)\mathrm{d}\theta$;

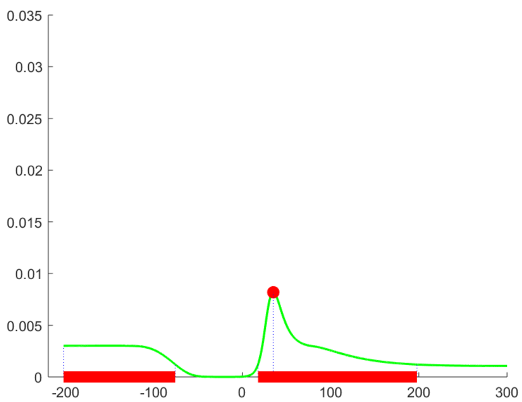
$\hat{\theta}_{\mathrm{ET},T}=35.0$ (bullet) and $\hat{\theta}_{\mathrm{ESP},T}=32.5$ (bullet); ET and ESP support= $[-202.8, 813.3]$

95% ET ALR conf. region=$[-202.8,-76.0]\cup[17.7,197.8]$ (stripe);

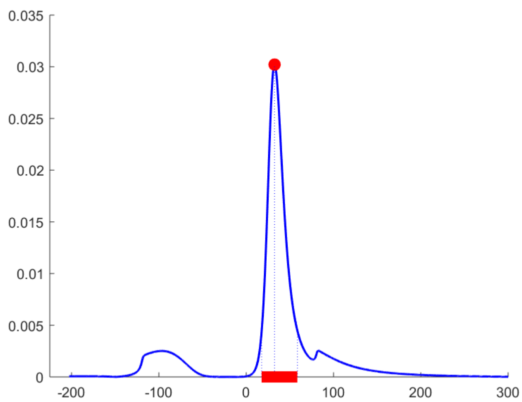
95% ESP ALR conf. region=$[17.7, 58.7]$ (stripe).



(A) Normalized ET (dashed green) vs. normalized ESP (dark blue).



(B) ET est. and ALR conf. region.



(C) ESP est. and ALR conf. region.

Note: The boundaries of the ESP and ET are indicative. The exact boundaries depend on the optimization routine.

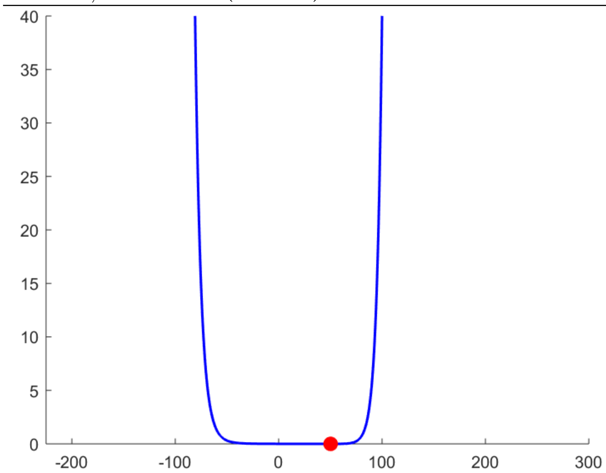
Table 10. **ET vs. ESP** 95% **confidence regions (1930–2009)**

	ET	ESP
Wald conf. region	$[14.8, 55.1]$	$[11.6, 53.3]$
LM conf. region	$[-215.2, -5.2] \cup [20.0, 813.3]$	$[-215.2, -153.6] \cup [-141.8, -122.7]$ $\cup[-118.8, -47.1] \cup [22.34, 813.3]$
ALR conf. region	$[-202.8, -76.0] \cup [17.7, 197.8]$	$[17.7, 58.7]$
ET conf. region	$[-202.8, -45.8] \cup [23.1, 813.3]$	$[-202.8, -45.8] \cup [23.1, 813.3]$

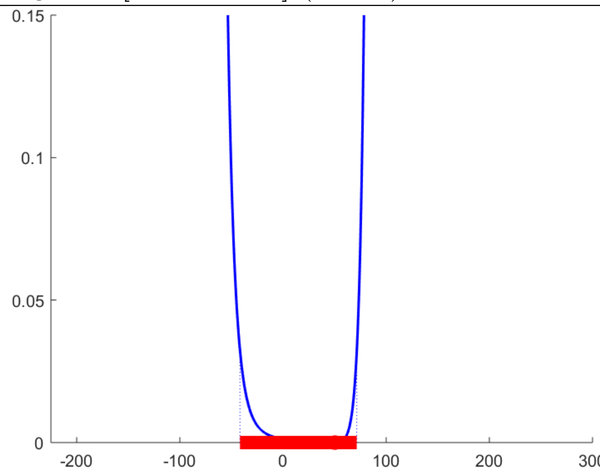
Table 11. **MM inference (1890–2009)**

Empirical moment condition: $\frac{1}{2009-1889}\sum_{t=1890}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right] = 0$, where $R_{m,t}$:= gross market return, $R_{f,t}$:=risk-free asset gross return, C_t := consumption, and θ :=relative risk aversion.

$\hat{\theta}_{\mathrm{GMM},T} = 50.3$ (bullet); 95% ALR confidence region= $[-41.7, 71.5]$ (stripe).



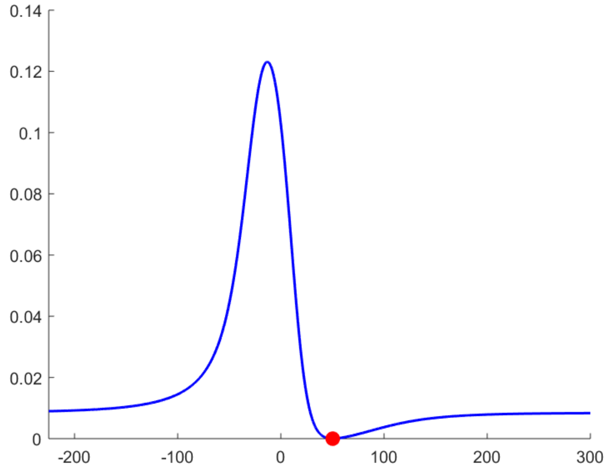
(A) MM objective function and point estimate.



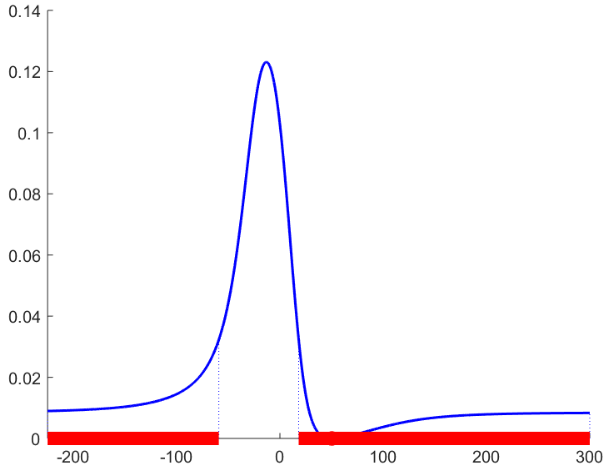
(A zoom) MM obj. function ALR conf. region.

TABLE 12. **Continuously updated (CU) GMM inference (1890–2009)**

Empirical moment condition: $\frac{1}{2009-1890}\sum_{t=1890}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0$, where
$R_{m,t}:=$ gross market return, $R_{f,t}:=$risk-free asset gross return, $C_t:=$ consumption, and $\theta:=$relative risk aversion.
$\hat{\theta}_T^{\mathrm{CU}}=50.3$ (bullet); 95% ALR confidence region (and S-set)$=]\ldots,-59.1]\cup[18.2,\ldots[$ (stripe).
Rk: We constrain the numerical search for point estimate to discard large values of θ.



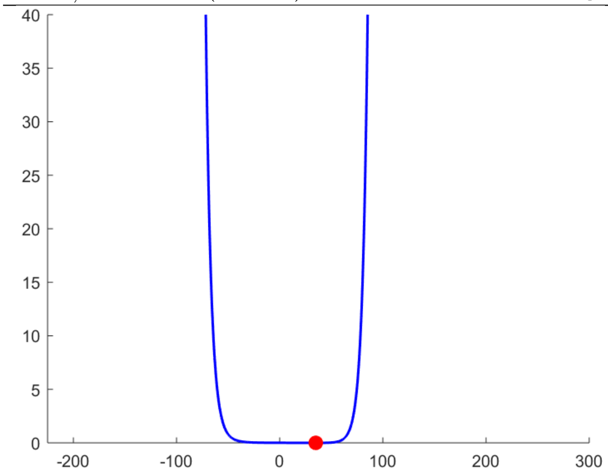
(A) Objective function and point estimate.



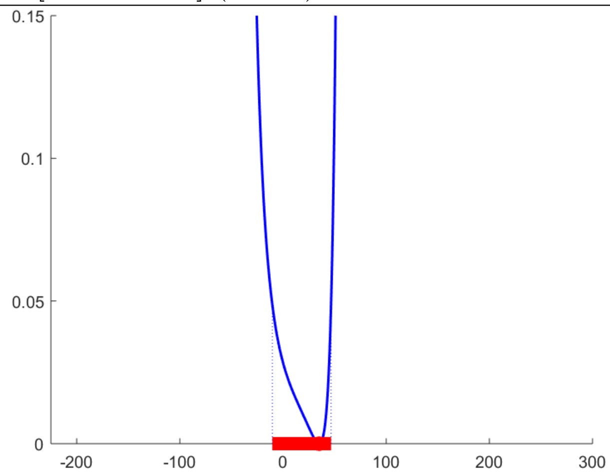
(B) Truncated ALR conf. region (and S-set).

TABLE 13. **MM inference (1930-2009)**

Empirical moment condition: $\frac{1}{2009-1930}\sum_{t=1930}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0$, where
$R_{m,t}:=$ gross market return, $R_{f,t}:=$risk-free asset gross return, $C_t:=$ consumption, $\theta:=$relative risk aversion.
$\hat{\theta}_{\mathrm{MM},T}=35.0$ (bullet), ALR confidence region$=[-10.4,46.5]$ (stripe)



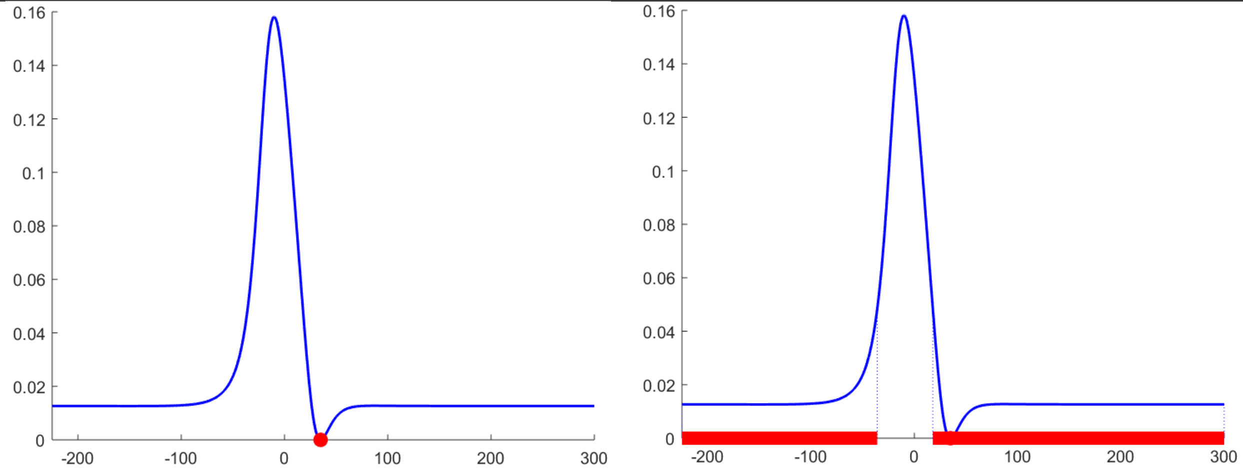
(A) MM objective function and point estimate.



(A zoom) Objective function and point estimate.

TABLE 14. **Continuously updated (CU) GMM inference (1930–2009)**

Empirical moment condition: $\frac{1}{2009-1890}\sum_{t=1890}^{2009}\left[\left(\frac{C_t}{C_{t-1}}\right)^{-\theta}(R_{m,t}-R_{f,t})\right]=0$, where
$R_{m,t}$:= gross market return, $R_{f,t}$:=risk-free asset gross return, C_t := consumption,
θ :=relative risk aversion.
$\hat{\theta}_T^{\mathrm{CU}} = 35.0$ (bullet); ALR confidence region (and S-set) $=]\ldots, -35.8] \cup [17.9, \ldots[$ (stripe).
Rk: We constrain the numerical search for point estimate to discard large values of θ.



(A) Objective function and point estimate. (B) Truncated ALR conf. region (and S-set).