Robust Inference for Moment Condition Models without Rational Expectations[*]

Xiaohong Chen[†] Lars Peter Hansen[‡] Peter G. Hansen[§]

This draft: October 18, 2021

Abstract

Applied researchers using structural models under rational expectations (RE) often confront empirical evidence of misspecification. In this paper we consider a generic dynamic model that is posed as a vector of unconditional moment restrictions. We suppose that the model is globally misspecified under RE, and thus empirically flawed in a way that is not econometrically subtle. We relax the RE restriction by allowing subjective beliefs to differ from the data-generating probability (DGP) model while still maintaining that the moment conditions are satisfied under the subjective beliefs of economic agents. We use statistical measures of divergence relative to RE to bound the set of subjective probabilities. This form of misspecification alters econometric identification and inferences in a substantial way, leading us to construct robust confidence sets for various set identified functionals.

JEL Classification: C14, C15, C31, C33, G40

Keywords: Subjective beliefs, bounded rationality, misspecification sets, nonlinear expectation, divergence, Lagrange multipliers, stochastic dual programming, confidence sets

[*] We thank Orazio Attanasio, Stephane Bonhomme, Timothy Christensen, Bo Honoré, Andrey Malenko, Per Mykland, Eric Renault, Tom Sargent, Azeem Shaikh, Ken Singleton and Grace Tsiang for helpful comments and suggestions.

[†] Yale University. Email: xiaohong.chen@yale.edu

[‡] University of Chicago. Email: lhansen@uchicago.edu

[§] **Corresponding author:** New York University. Email: peter.hansen@nyu.edu

1 Introduction

Dynamic models in economics have forward-looking decision makers who form beliefs about their uncertain environment. For instance, investment decisions depend on forecasts about the future, firms speculate about the future demand for their products, and strategic players in a dynamic game setting make conjectures about the other players' actions. One common approach is to assume agents inside the economic model form Rational Expectations (RE). This postulate can be enforced as an equilibrium construct, or as is done in the Generalized Method of Moments (GMM) analysis, imposing that beliefs of economic agents coincide with the Data Generating Process (DGP) as revealed by empirical evidence.[1] For this latter approach, the corresponding optimization and equilibrium conditions lead to moment restrictions that can be estimated and tested via GMM, Generalized Empirical Likelihood (GEL), or other related methods. The expectations used in forming the moment conditions are presumed to coincide with the subjective beliefs of the economic decision makers inside the dynamic economic model.

A substantial body of research has emerged in operations research and statistics that studies optimization in the presence of statistical model uncertainty about the underlying DGP pertinent for the decision problem. This research explores robustness bounds for parameter estimation and statistical decision rules, typically restricting the misspecification to be "local." While related, we address a different problem. In the empirical implementation of RE models, external researchers assume the expectations used in defining the moment conditions are consistent with both the beliefs of the economic agents inside the model and population limit implied by the DGP. While it is common in this literature to impose RE as a simplifying assumption, there are substantive and empirical reasons for relaxing this assumption. An important part of our econometric ambition is to make inferences about the subjective beliefs of economic agents without imposing RE while restricting them not to be excessively distant from the DGP.

Rather than pursue the often illusive or unrevealing goal of finding a correctly specified model under RE, we take a different approach. We suppose that the RE model is misspecified. Moreover, we deliberately shun approaches that assume this misspecification is local. In other words, we do not presume that misspecification vanishes at a rate that is exogenously linked to sample size. While a local approach is common in much of the

[1] For the conceptual underpinnings of the later approach see Hansen (1982), Hansen and Singleton (1982) and Hansen and Richard (1987).

related theoretical literature on estimation and inference, for many applications the belief distortions are direct substantive interests and are not intended to be subtle. By altering the potential subjective beliefs, we allow the population moment conditions to be satisfied, building on a suggestion in Hansen (2014). This approach to misspecification introduces an identification challenge as there is potentially a very large set of subjective beliefs for which the moment conditions will be satisfied. Moreover, once we allow for belief distortions, we lose point identification of the unknown model parameters.

In this paper we consider a generic dynamic model that is posed as a vector of unconditional moment restrictions. We suppose that the conditional moment model is globally misspecified under RE, and thus empirically flawed in a way that is not econometrically subtle. We allow subjective beliefs of economic agents to differ from those implied by the DGP while still maintaining that the moment conditions are satisfied under the subjective beliefs. This form of misspecification alters econometric identification and inferences in a substantial way, leading us to address some new econometric challenges. We replace the RE restriction by a statistical divergence bound on the potentially distorted beliefs relative to the DGP. By limiting the magnitude of the statistical divergence, we avoid excessively large identified sets of potential parameter values. This constraint limits both the set of potential beliefs and the underlying model parameters. In effect, we use statistical divergence as a formal way to "bound irrationality". We relax RE while still embracing the notion that economic agents will not form beliefs that are obviously inconsistent with statistical evidence. We show that some popular statistical divergence measures are problematic for studying distorted beliefs because they are poor at revealing some forms of model misspecification that interest us. Since our interest is in dynamic models with forward-looking economic agents, our methodology can be viewed as a way to (i) extract information on investor beliefs from equilibrium prices and from survey data, and (ii) to provide revealing diagnostics for model builders that embrace specific formulations of belief distortions.

Specifying a probability distribution is equivalent to specifying an associated expectation operator applied to a rich class of measurable functions of underlying random vector. Extending this insight, we represent bounds on a family of probability distributions as a nonlinear expectation operator constructed as follows. For each possible function we minimize the expectation of the function of the random vector (over the set of distributions). Because of the minimization, the resulting expectation operator is nonlinear. Since the class of functions is sufficiently large to include both a function and its negative, this operator gives both upper and lower bounds on the admissible set of expectations. This nonlinear expectation operator is a mathematically convenient way to depict the identified set of probabilities that capture subjective beliefs.[2]

Central to our econometric analysis are estimation of the Lagrange multipliers of the model-implied moment conditions evaluated at the subjective probabilities of the economic agents inside the model. These multipliers inform us how we must reshape the DGP to match the model moment implications. For each fixed parameter satisfying the belief distorted moment restrictions, the multipliers are uniquely identified and can be estimated at the standard root-T rate and is asymptotically normally distributed. However, without assuming unique "pseudo-true" parameter value, the "optimal" multipliers are no-longer point identified. To make inferences about the sets of multipliers and corresponding parameter values, we rely on econometric theory as in Chernozhukov et al. (2007) and Chen et al. (2018). Our results feature the estimation of sets of models with restricted magnitudes of belief distortions. In addition, we care about confidence interval construction for the nonlinear expectation functional for temporally dependent DGP's.

This paper focuses on unconditional moment restrictions, fully specified, dynamic stochastic equilibrium models include conditional moment conditions expressed using the beliefs of economic agents. While the approach we describe here has counterparts to the study of general dynamic conditional moment models, such an econometric extension is beyond the scope of this paper.[3]

1.1 Organization

The rest of the paper is organized as follows. Section 2 presents our model framework. Section 3 bounds agents' beliefs from the RE using the Cressie and Read (1984) family of divergences. Within this family of ϕ-divergences, we illustrate how divergences constructed from convex functions ϕ that are strictly decreasing are problematic for identifying misspecification in moment condition models. This implies that Hellinger and minus log-likelihood divergences are problematic, but relative entropy and quadratic divergences remain applicable for our analysis. In Section 4, using ϕ-divergences, we propose a nonlinear expectation functional for representing restrictions on the subjective expectations subject to model-implied moment conditions and a divergence constraint. We give dual representations that

[2] An analogous nonlinear expectation operator is central to Peng (2004)'s development of a novel control theory designed to confront uncertainty for Brownian motion information structures.

[3] While a general treatment of conditioning information is difficult, it is straightforward to extend our analysis to accommodate discrete conditioning.

makes the nonlinear expectation computationally tractable. As illustrated in Section 5, our approach is not restricted to ϕ-divergences and applies to any convex divergences including the Wasserstein distance between the probability measure that underlies subjective beliefs and probabilities implied by the DGP. We consider two econometric challenges posed by our methods. Section 6 considers inference on minimal ϕ-divergence measure over all probabilities that satisfy moment conditions. This is important because smaller divergence bounds imply an empty constraint set. This section also proposes and justifies confidence sets for the parameter values that attain the minimal divergence. This set is a subset of the parameters of interest consistent with divergence bounds that exceed the minimal threshold. Section 7 studies estimation and inference on nonlinear expectation functionals associated with the family of probabilities satisfying unconditional moment restrictions and a ϕ-divergence constraint. Section 8 briefly concludes.

1.2 Relation to the existing literature

There is a well known and long-standing literature on the important role of subjective beliefs in determining investment and other economic decisions. This literature is too vast to summarize but it includes a variety of models of expectations in addition to rational expectations. More recently, there has been interest in collecting additional data on agents' beliefs and using these often sparse data to estimate parametric/semiparametric models of subjective beliefs. For instance, see Manski (2018), Meeuwis et al. (2018), Bordalo et al. (2020), Bhandari et al. (2019), and Attanasio et al. (2019). We can allow for the incorporation of even limited survey data by adding them into moment restrictions in our framework.

Our approach is loosely related to the GEL literature on estimation and testing of moment restriction models in that both use statistical ϕ-divergence measures. GEL estimates point-identified parameters and probabilities jointly in hopes of improving second-order statistical efficiency over GMM estimates for correctly specified moment restrictions. It presumes the expectation used in representing the moment conditions coincides with DGP (i.e., imposing RE). See, for example, Qin and Lawless (1994), Imbens (1997), Kitamura and Stutzer (1997), Smith (1997), Imbens et al. (1998) and Newey and Smith (2004). There is also a local sensitivity approach in GEL literature, which assumes misspecification is within a root-T shrinking neighborhood around the unique "true" parameter value that satisfies the unconditional moment conditions (under RE). See, e.g., Kitamura et al.

(2013), Bonhomme and Weidner (2018), Armstrong and Kolesár (2018), and Andrews et al. (2020). Unlike these papers, we feature a particular form of global model misspecification for which there is substantive economic interest: agents' subjective beliefs that diverge from the DGP. Given that we entertain subjective beliefs that satisfy moment restrictions but differ from the DGP, we are lead naturally to entertain set identification of both beliefs and model parameters.

Many econometric contributions that entertain global misspecification assume that the pseudo true parameter vector is uniquely determined. For example, Luttmer et al. (1995), Almeida and Garcia (2012), Gagliardini and Ronchetti (2019) and Antoine et al. (2018) use meaningful bounds on pricing errors for asset pricing models to make inferences about their unique pseudo true parameter vectors. This literature, however, does not target misspecification induced by belief distortions. See Hall and Inoue (2003), Ai and Chen (2007), Schennach (2007), Lee (2016), and Hansen and Lee (2021) for a more generic global approach to misspecification in moment restrictions, again featuring uniquely identified pseudo true parameters. Minimizing specification errors measured with statistical divergence is a starting point for us; but we are interested in the implications of larger divergence bounds for both beliefs and parameters while imposing an economic structure to the misspecification. Our approach deliberately puts the notion of a pseudo true parameter vector to the wayside.

In terms of applications of stochastic dual programs and ϕ-divergence balls, our work has similarity to Shapiro (2017), Duchi and Namkoong (2021), Christensen and Connault (2019) and the references therein.[4] None of these papers, however, motivates the misspecification in terms of dynamic settings with a potentially large set of subjective beliefs of the decision makers being modeled. This is what gives rise in our analysis to the time series moment restrictions and the corresponding misspecified parameter sets.

This paper is similarly motivated and complementary to our recent publication Chen et al. (2021). The Chen et al. (2021) contribution features identification only and does not explore estimation and inference. As a consequence, it does not forge connections with the substantial econometrics literature on misspecification that we referenced. On the other hand, the Chen et al. (2021) paper features conditional moment conditions whereas in this paper we presume a finite number of unconditional moment conditions as the starting point.

[4] While Shapiro (2017) and Duchi and Namkoong (2021) constructed confidence intervals assuming unique "true" parameter value, Christensen and Connault (2019) construct bootstrapped confidence set allowing for partial-identification while imposing separable moment conditions.

2 Model Specification

In dynamic economic applications, moment conditions are often justified via an assumption of RE. This assumption equates population expectations with those used by economic agents inside the model. These expectations are therefore presumed to be revealed by the Law of Large Numbers implied by the DGP.

Let $(\Omega, \mathfrak{G}, P)$ denote the underlying probability space and $\mathfrak{I} \subset \mathfrak{G}$ represent information available to an economic agent. The original moment equations under rational expectations are of the form

$$\mathbb{E}\left[f(X,\theta) \mid \mathfrak{I}\right] = 0 \quad \text{for some } \theta \in \Theta.$$

where the vector-valued function f captures the parameter dependence (θ) of either the payoff or the stochastic discount factor along with variables (X) observed by the econometrician and used to construct the payoffs, prices, and the stochastic discount factor. By applying the Law of Iterated Expectations,

$$\mathbb{E}\left[f(X,\theta)\right] = 0 \quad \text{for some } \theta \in \Theta. \tag{1}$$

The vector-valued function f may include scaling by $\mathfrak{I}$ measurable random variables as a device to bring conditioning information through the "back door."

In this paper we allow for agents' beliefs that are revealed by the data to differ from the rational expectations beliefs implied by (infinite) histories of stationary ergodic data. We represent agents' belief by a positive random variable M with a unit conditional expectation. Thus, we consider moment restrictions of the form: for any $\theta \in \Theta$,

$$\mathbb{E}\left[Mf(X,\theta)\right] = 0. \tag{2}$$

The random variable M provides a flexible change in the probability measure, and is sometimes referred to as a Radon-Nikodym derivative or a likelihood ratio. The dependence of M on random variables not in the information captured by $\mathfrak{I}$ defines a relative density that informs how RE are altered by agent beliefs. By changing M, we allow for alternative densities. Notice that we are restricting the implied probability measures to be absolutely continuous with respect to the original probability measure implied by rational expectations. That is, we restrict the agent beliefs so that any event that has probability measure zero under the DGP will continue to have probability zero under this change in distribution. We will, however, allow for agents to assign probability zero to events that actually have positive probability.

For any parameter vector θ in equation (2), there are typically many specifications of beliefs M that will satisfy the model implied moment conditions. Rather than imposing *ad hoc* assumptions to resolve this identification failure, we will characterize the multiplicity by using bounds on statistical divergence. A statistical divergence quantifies how close two probability measures are. In our analysis, one of these probability measures governs the data evolution while the other governs the investment decisions or the equilibrium pricing relations. We define a range of allowable probability measures, and we consider a family of divergences commonly used in the statistics and machine learning literatures.

3 Bounding Beliefs with ϕ-Divergences

In this section we study a family of so-called ϕ-divergences and explore within this family which divergences are most revealing for assessing misspecification in dynamic economic models.[5] For the moment, fix θ in equation (2) and write $f(X, \theta)$ as $f(X)$. Initially we also abstract from the role of conditioning information, but the expectations can be interpreted as being conditioned on a sigma algebra $\mathfrak{I}$ as in our earlier paper, Chen et al. (2021).

3.1 Constructing ϕ-Divergences

Introduce a convex function ϕ defined on $\mathbb{R}^+$ for which $\phi(1) = 0$. As a scale normalization we will assume that $\phi''(1) = 1$. The corresponding divergence of a belief M from the underlying data generation is defined by $\mathbb{E}[\phi(M)]$. By Jensen's inequality, we know that

$$\mathbb{E}[\phi(M)] \geqslant \phi(1) = 0$$

since $\mathbb{E}[M] = 1$. The divergences $\mathbb{E}[\phi(\cdot)]$ are known as ϕ-divergences. Special cases include:

(i) $\phi(m) = -\log m$ (negative log likelihood)

(ii) $\phi(m) = 4\left(1 - \sqrt{m}\right)$ (Hellinger distance)

(iii) $\phi(m) = m \log m$ (relative entropy)

(iv) $\phi(m) = \frac{1}{2}(m^2 - m)$ (Euclidean divergence).

[5] Proofs and supporting analyses for this section are given in appendix A.

These four cases are widely used and are nested in the family of ϕ-divergences introduced by Cressie and Read (1984) defined by

$$\phi(m) = \begin{cases} \frac{1}{\eta(1+\eta)} \left[(m)^{1+\eta} - 1\right] & \eta < 0 \\ \frac{1}{\eta(1+\eta)} \left[(m)^{1+\eta} - m\right] & \eta \geqslant 0 \end{cases} \tag{3}$$

For $\eta = -1$ or 0, we can apply L'Hôpital's rule to obtain cases (i) and (iii) respectively. The divergence corresponding to $\eta = -\frac{1}{2}$ is equivalent to the Hellinger distance between probability densities. Empirical likelihood methods use the $\eta = -1$ divergence.[6] Two cases of particular interest to us are $\eta = 0$ and $\eta = 1$. We refer to the divergence for $\eta = 0$ as *relative entropy*. We refer to the $\eta = 1$ case as a quadratic or Euclidean divergence.[7]

3.2 Problematic divergences

For the purposes of misspecification analysis, we show that monotone decreasing divergence functions are problematic. For instance, the Cressie and Read divergences defined by (3) and used in the GEL literature are decreasing whenever $\eta < 0$. Our finding that the empirical likelihood ($\eta = -1$)and Hellinger (the case $\eta = -\frac{1}{2}$) divergences are problematic under model misspecification is noteworthy, as both have been widely used in statistics and econometrics. Our negative conclusion about monotone decreasing divergences leads us to focus on divergences for which $\eta \geqslant 0$ as robust measures of probability distortions.

To understand why monotone decreasing divergences are problematic, we study the corresponding population problem:

Problem 3.1.

$$\underline{\kappa} \doteq \inf_{M>0} \mathbb{E}[\phi(M)]$$

subject to

$$\mathbb{E}[M] = 1$$

$$\mathbb{E}[Mf(X)] = 0.$$

[6] This same divergence is also featured in the analysis of Alvarez and Jermann (2005) in their characterization of the martingale component to stochastic discount factors.

[7] Given our interest is in sets of belief distortions, our method is distinct from those designed for estimation under correct specification. In particular, our motivation and assumptions differ substantially from the literature on GEL methods. The so-called pseudo-true parameter value that is often the centerpiece of misspecification analysis in the econometrics literature plays a tangential role in our analysis as does point identification.

When the constraint set is empty, we adopt the convention that the optimized objective is ∞. We call a model misspecified if

$$\mathbb{E}\left[f(X)\right] \neq 0.$$

For a divergence to be of interest to us, the greatest lower bound on the objective should inform us as to how big of a statistical discrepancy is needed to satisfy equation (2). Therefore the infimum should be strictly positive whenever $\mathbb{E}[f(X)] \neq 0$. Conversely, notice that under correct specification, $\mathbb{E}[f(X)] = 0$, and $M = 1$ is in the constraint set of problem 3.1. By the design of a divergence measure, for $M = 1$ the minimized objective for problem 3.1 is zero.

Theorem 3.2. *Assume that $\phi(m)$ is decreasing in m, $\mathbb{E}[f(X)] \neq 0$, $f(X)$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, and there exists a convex cone $C \subset \mathbb{R}^d$ such that $f(X)$ has strictly positive density on C and $-\mathbb{E}[f(X)] \in int(C)$. Then for any $\kappa > 0$ there exists a belief distortion M such that i) $M > 0$ on supp$[f(X)]$; ii) $\mathbb{E}[M] = 1$; iii) $\mathbb{E}[Mf(X)] = 0$; iv) $\mathbb{E}[\phi(M)] < \kappa$.*

Theorem 3.2 shows dramatically that when the vector $f(X)$ has unbounded support, problem 3.1 can become degenerate. The infimized divergence can be equal to zero even though $\mathbb{E}[f(X)] \neq 0$ so the model is misspecified. In this case the infimum is not attained by any particular M, but can be approximated by sequences that assign small probability to extreme realizations of $f(X)$.[8] We view the assumption of unbounded support as empirically relevant, since moment conditions coming from asset pricing typically have terms that are multiplicative in the returns. Note that gross returns have no *a priori* upper bound, and excess returns have no *a priori* upper or lower bounds. The condition in Theorem 3.2 that $\phi(m)$ is decreasing in m is crucial to the degeneracy. As we noted, this condition is satisfied for the Cressie-Read family whenever $\eta < 0$.

Remark 3.3. *Previously Schennach (2007) demonstrated problematic aspects of empirical likelihood estimators under misspecification. She assumed the existence of a unique pseudo-true parameter value that is additionally consistently estimated by the empirical likelihood estimator computed using the dual problem, but pointed out that such an estimator may fail*

[8] An explicit construction of such sequences is given in appendix A. Heuristically, we perturb the original distribution of $f(X)$ by shifting a very small amount of probability mass into an extreme tail so that the moment condition $\mathbb{E}[Mf(X)] = 0$ is satisfied. These perturbed distributions will converge weakly to the original distribution, and the divergence will approach zero.

to be root-T consistent under model misspecification, where T is the sample size for iid data. In relation to this, we showed that the primal problem may also fail to detect misspecification for any monotone decreasing divergence. This includes the $\eta = -1$ divergence used in empirical likelihood methods. As we emphasized previously, our paper is not concerned with the point identification of pseudo-true parameter values.

For future reference, consider the dual to problem 3.1:

$$\sup_{\lambda,\nu} \inf_{M>0} \mathbb{E}\left[\phi(M) + M\lambda \cdot f(X) + \nu\left(M - 1\right)\right] \tag{4}$$

where λ and ν are Lagrange multipliers. This problem is of interest because it is typically easier to solve than the primal problem, especially when the inner minimization over M has a quasi-analytical solution.

3.3 Relative Entropy Divergence

This section considers the relative entropy divergence (i.e., $\phi(m) = m \log m$ or $\eta = 0$). Among the class of ϕ divergences, relative entropy has same convenient mathematical properties and interpretations.

As known from a variety of sources and reproduced in the appendix, the dual to problem 3.1 with relative entropy divergence is:

Problem 3.4.

$$\sup_{\lambda} - \log \mathbb{E}\left[\exp\left(-\lambda \cdot f(X)\right)\right].$$

In this problem we have maximized over the scalar multiplier ν. The first-order conditions for problem 3.4 are $\mathbb{E}[M^* f(X)] = 0$ where M^* is constructed using

$$M^* = \frac{\exp\left(-\lambda^* \cdot f(X)\right)}{\mathbb{E}\left[\exp\left(-\lambda^* \cdot f(X)\right)\right]} \tag{5}$$

where λ^* is the maximizing choice of λ.

For this candidate M^* to be a valid solution, we must restrict the probability distribution of $f(X)$. Notice that $\psi(\lambda) \equiv \mathbb{E}\left[\exp\left(-\lambda \cdot f(X)\right)\right]$, when viewed as a function of $-\lambda$, is the multivariate moment-generating function for the random vector $f(X)$. We include $+\infty$ as a possible value of ψ in order that it be well defined for all λ. The negative of its logarithm is a concave function, which is the objective for the optimization problem that interests us. A unique solution to the dual problem exists under the following restrictions on this generating function.

Restriction 3.5. *The moment generating function ψ satisfies:*

(i) ψ *is continuous in* λ*;*

(ii) $\lim_{|\lambda| \to \infty} \psi(\lambda) = +\infty$.[9]

A moment generating function is infinitely differentiable in neighborhoods in which it is finite. To satisfy condition (i) of restriction 3.5, we allow for ψ to be infinite as long as it asymptotes to $+\infty$ continuously on its domain. In particular, ψ does not have to be finite for all values of λ. Condition (ii) requires that ψ tends to infinity in all directions. Restriction 3.5 is satisfied when the support sets of the entries of $f(X)$ are not subsets of either the positive real numbers or negative real numbers. Importantly for us, restriction 3.5 allows for $f(X)$ to have unbounded support.

Theorem 3.6. *Suppose that restriction 3.5 is satisfied. Then problem 3.4 has a unique solution* λ^**. Using this* λ^* *to form* M^* *in* (5)*, which satisfies the two constraints imposed in problem 3.1. Thus the optimized objective for both problems (with relative entropy) is*

$$\underline{\kappa} = -\log \mathbb{E} \exp[-\lambda^* \cdot f(X)].$$

4 Bounding expectations

Computing minimal divergences in problem 3.1 is merely a starting point for our analysis. Our primary aim is to construct misspecified sets of expectations, we use $\kappa > \underline{\kappa}$ to bound the divergence of belief misspecification. This structure will allow us to explore belief distortions other than the one implied by minimal divergence. While we represent alternative probability distributions with alternative specifications of the positive random variable M with unit expectation, we find it most useful and revealing to depict bounds on the resulting expectations. Larger values of κ will lead to bigger sets of potential expectations.

Given a measurable function g of X, we consider the following problem:

[9] This condition rules out redundant moment conditions as well as $f(X)$'s which only take on nonnegative or nonpositive values with probability one.

Problem 4.1.

$$\mathbb{K}(g) \doteq \min_{M \geqslant 0} \mathbb{E}\left[Mg(X)\right]$$

subject to the three constraints:

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(M\right)\right] \leqslant \kappa \\
&\mathbb{E}\left[Mf(X)\right] = 0, \\
&\mathbb{E}\left[M\right] = 1.
\end{aligned}$$

As before we can solve this problem using convex duality.[10] The function g could define a moment of an observed variable of particular interest or it could be the product of the stochastic discount factor and an observed payoff to a particular security whose price we seek to bound.

4.1 A nonlinear expectation operator

Formally, we represent the bounds on subjective expectations as a nonlinear expectation operator. While we are potentially interested in more general functions g, we initially focus on the set $\mathcal{B}$ of bounded Borel measurable functions g to be evaluated at alternative realizations of the random vector X. The mapping $\mathbb{K}$ from $\mathcal{B}$ to the real line can be thought of as a "nonlinear expectation," as formalized in the following proposition.

Proposition 4.2. *The mapping $\mathbb{K} : \mathcal{B} \rightarrow \mathbb{R}$ has the following properties*[11]*:*

(i) *if $g_2 \geqslant g_1$, then $\mathbb{K}(g_2) \geqslant \mathbb{K}(g_1)$.*

(ii) *if g constant, then $\mathbb{K}(g) = g$.*

(iii) $\mathbb{K}(\mathsf{r}g) = \mathsf{r}\mathbb{K}(g)$, *for a scalar* $\mathsf{r} \geqslant 0$

(iv) $\mathbb{K}(g_1) + \mathbb{K}(g_2) \leqslant \mathbb{K}(g_1 + g_2)$

[10] There is an extensive literature studying the mathematical structure of more general versions of this problem including more general specifications of entropy. Representatives of this literature include the insightful papers Csiszar and Matus (2012) and Csiszar and Breuer (2018). We find it pedagogically simpler to study the dual problem directly and verify that the solution is constraint feasible rather than to verify regularity conditions in this literature.

[11] The first two of these properties are taken to be the definition of a nonlinear expectation by Peng (2004). Properties (*iii*) and (*iv*) are referred to as "positive homogeneity" and "superadditivity".

All four properties follow from the definition of $\mathbb{K}$. Property (iv) includes an inequality instead of an equality because we compute $\mathbb{K}$ by solving a minimization problem, and the M's that solve this problem can differ depending on g. This nonlinear expectation operator can be extended to more general functions g depending on the application.

Remark 4.3. *While $\mathbb{K}(g)$ gives a* lower *bound on the expectation of $g(X)$, by replacing g with $-g$, we construct an* upper *bound on the expectation of $g(X)$. The upper bound will be given by $-\mathbb{K}(-g)$. The interval*

$$[\mathbb{K}(g), -\mathbb{K}(-g)]$$

captures the set of possible values for the distorted expectation of $g(X)$ consistent with divergence less than or equal to κ.

There is a closely related problem that is often more convenient to work with. We revert back to a minimum discrepancy formulation and augment the constraint set to include expectations of $g(X)$ subject to alternative upper bounds. We then characterize how changing this upper bound alters the divergence objective. Stated formally,

Problem 4.4.

$$\mathbb{L}(\vartheta; g) \doteq \inf_{M>0} \mathbb{E}\left[\phi\left(M\right)\right]$$

subject to:

$$\begin{aligned} \mathbb{E}\left[Mf(X)\right] &= 0, \\ \mathbb{E}\left[Mg(X)\right] &\leqslant \vartheta \\ \mathbb{E}\left[M\right] &= 1. \end{aligned}$$

Notice that $\mathbb{L}(\vartheta; g)$ increases as we decrease ϑ because smaller values of ϑ make the constraint set more limiting. Thus we may decrease ϑ to attain a prespecified value κ used in constructing the nonlinear expectation $\mathbb{K}(g)$.

4.2 Bounding conditional expectations

Consider an event Λ with $\mathsf{P}(\Lambda) = \mathbb{E}[\mathbf{1}_\Lambda] > 0$ where $\mathbf{1}_\Lambda$ is the indicator function for the event Λ. Given a function $g(X)$ of the data X, we can extend our previous arguments to produce a bound on the conditional expectation. Instead of entering $\mathbb{E}\left[Mg(X)\right] \leqslant \vartheta$ as an additional moment condition in problem 4.4, we include

$$\mathbb{E}\left[M\mathbf{1}_\Lambda\left(g(X)-\vartheta\right)\right] \leqslant 0$$

in the constraint set and vary ϑ to attain a divergence target. This moment inequality is essentially equivalent to the conditional moment bound:

$$\frac{\mathbb{E}\left[M\mathbf{1}_\Lambda\left(g(X)\right)\right]}{\mathbb{E}\left[M\mathbf{1}_\Lambda\right]} \leqslant \vartheta$$

provided that the denominator is strictly positive. The left side is recognizable as the conditional expectation of $g(X)$ conditioned on Λ.

4.3 Relative entropy reconsidered

Next we give a dual representation of $\mathbb{K}(g)$ in problem 4.1 for the special case of the relative entropy divergence:[12]

$$\mathbb{K}(g) = \sup_{\xi>0}\max_{\lambda} -\xi \log \mathbb{E}\left[\exp\left(-\frac{1}{\xi}g(X) - \lambda\cdot f(X)\right)\right] - \xi\kappa. \tag{6}$$

Notice that conditioned on ξ, the maximization over λ does not depend on κ because $-\xi\kappa$ is additively separable. This makes it convenient to explore the supremum over λ for each $\xi > 0$. Write:

$$\widehat{\mathbb{K}}(\xi; g) \doteq \sup_{\lambda} -\xi \log \mathbb{E}\left[\exp\left(-\frac{1}{\xi}g(X) - \lambda\cdot f(X)\right)\right]. \tag{7}$$

We deduce ξ and the resulting moment bound by solving:

$$\mathbb{K}(g) = \sup_{\xi\geqslant 0}\left[\widehat{\mathbb{K}}(\xi; g) - \xi\kappa\right]. \tag{8}$$

Remark 4.5. *For sufficiently large values of κ, it is possible the constraint on relative entropy actually does not bind. The additional moment restrictions by themselves limit the family of probabilities, and might do so in ways that restrict the implied entropy of the probabilities. Appendix A gives sufficient conditions under which the relative entropy constraint will bind, and provides examples suggesting that the relative entropy constraint may bind in many cases of interest even for arbitrarily large choices of κ.*

[12] See appendix A for a justification.

By imitating our previous logic for the minimum divergence problem subject to moment conditions, the dual for Problem 4.4 with relative entropy divergence is:

Problem 4.6.

$$\sup_{\rho \geqslant 0, \lambda} - \log \mathbb{E}\left[\exp\left(-\rho g(X) - \lambda \cdot f(X)\right)\right] - \vartheta \rho.$$

The variable ρ is a Lagrange multiplier on the moment restriction involving g.

A natural starting point is to take the solution M^* given in (5) from problem 3.4 and compute

$$\mathsf{u}_g = \mathbb{E}\left[M^* g(X)\right].$$

By setting $\vartheta = \mathsf{u}_g$, the solution to problem 4.6 sets $\rho = 0$ and $\lambda = \lambda^*$. This choice satisfies the first-order conditions. Lowering ϑ will imply a binding constraint:

$$\mathbb{E}\left[M g(X)\right] - \vartheta = 0.$$

Given the binding constraint, we may view problem 4.4 as an extended version of problem 3.1 (for $\eta = 0$) with an additional moment restriction added. This leads us to state following analog to theorem 3.6.

Theorem 4.7. *Suppose*

i) $\vartheta < \mathsf{u}_g$*;*

ii) restriction 3.5 is satisfied for the random vector: $\begin{bmatrix} g(X) & f(X)' \end{bmatrix}'$.

Then problem 4.6 has a unique solution (ρ^*, λ^*) *for which*

$$M^* = \frac{\exp\left[-\rho^* g(X) - \lambda^* \cdot f(X)\right]}{\mathbb{E}\left[\exp\left[-\rho^* g(X) - \lambda^* \cdot f(X)\right]\right]},$$

this choice of M^* *satisfies* $\mathbb{E}[M^*] = 1$, $\mathbb{E}[M^* f(X)] = 0$, *and* $\mathbb{E}[M^* g(X)] = \vartheta$. *Thus objectives for problems 4.4 (with* $\eta = 0$*) and 4.6 coincide, and the optimized objective is*[13]

$$\mathbb{L}(\vartheta; g) = -\log \mathbb{E}\left[\exp\left(-\rho^* g(X) - \lambda^* \cdot f(X)\right)\right] - \vartheta \rho^*.$$

The relative entropy objective for problem 4.4 increases as we decrease ϑ. For instance, by decreasing ϑ in this way we could hit the relative entropy threshold of problem 4.1.

[13] While ρ^*, λ^*, M^* depend on the choice of ϑ, to simplify notation we leave this dependence implicit.

Both approaches feature the same intermediate problem in which we initially condition on ξ or ρ and optimize over λ. For computational purposes we deduce the implied expectation of $g(X)$ and relative entropy by tracing out both as functions of the scalars ξ or ρ.

4.4 Quadratic Divergence

While the relative entropy ($\eta = 0$) divergence has many nice properties, it imposes restrictions on thinness of tails of the probability distribution of $f(X)$ that may be too severe for some applications.[14] As an alternative, we now consider the quadratic or Euclidean divergence obtained when we set $\eta = 1$. We will not repeat the analysis of alternative bounds. Since a key input is the dual to a divergence bound problem, we will characterize the resulting solution for bounds and leave the extensions to the appendix. We study the counterpart to problem 4.1.

We impose two assumptions to ensure non-degenerate bounds.

Restriction 4.8. *$f(X)$ and $g(X)$ have finite second moments.*

Restriction 4.9. *There exists an $M > 0$ such that $\mathbb{E}[M] = 1$, $\mathbb{E}[Mf(X)] = 0$ and $\frac{1}{2}\mathbb{E}[M^2 - M] \leqslant \kappa$.*

The problem of interest is:

Problem 4.10.

$$\mathbb{Q}(g) \doteq \inf_{M \geqslant 0} \mathbb{E}[Mg(X)]$$

subject to:

$$\begin{aligned} &\frac{1}{2}\mathbb{E}\left[M^2 - M\right] \leqslant \kappa \\ &\mathbb{E}[Mf(X)] = 0 \\ &\mathbb{E}[M] = 1. \end{aligned}$$

We allow M to be zero with positive probability for mathematical convenience. Since there exists an $M > 0$ for which $\mathbb{E}\left[Mf(X)\right] = 0$, we can form a sequence of strictly positive M's with divergences that are arbitrarily close to bound we derive. Solving this problem for alternative bounded g's gives us a nonlinear expectation function $\mathbb{Q}$ satisfying the properties in Proposition 4.2.

[14] For instance, if we specify S as an exponential-affine model of the form $S = \exp(\psi \cdot Z + Z'\Psi W)$ where W is a conditionally Gaussian shock, then restriction 3.5 may be violated.

Problem 4.11.

$$\widehat{\mathbb{Q}}(g) \doteq \sup_{\xi \geqslant 0, \nu, \lambda} -\frac{\xi}{2}\mathbb{E}\left[\left(\left[\frac{1}{2} - \frac{1}{\xi}\left[g(X) + \lambda \cdot f(X) + \nu\right]\right]^+\right)^2\right] - \xi\kappa - \nu.$$

Proposition 4.12. *Assume that restrictions 4.8 and 4.9 hold and that the supremum in problem 4.10 is attained with $\xi^* > 0$. Then $\mathbb{Q}(g) = \widehat{\mathbb{Q}}(g)$. Furthermore, the solution $(\xi^*, \nu^*, \lambda^*)$ to problem 4.11, corresponds to the belief distortion*

$$M^* = \left[\frac{1}{2} - \frac{1}{\xi^*}\left[g(X) + \lambda^* \cdot f(X) + \nu^*\right]\right]^+$$

which satisfies the constraints of problem 4.10 with equality, and attains the infimum, i.e. $\mathbb{E}[M^* g(X)] = \mathbb{Q}(g)$.

Proposition 4.12 follows from theorem 6.7 of Borwein and Lewis (1992). It characterizes the solution to problem 4.10 when the divergence constraint binds. Otherwise, we can obtain the expectation bound by solving problem 4.11 for a fixed sequence of ξ's converging to zero where we maximize with respect to λ and ν given any ξ in this sequence.

5 Wasserstein Distance and Regularization

Our analysis thus far used ϕ-divergences as the relevant notion of statistical discrepancy between probability measures. An alternative measure, which has gained some recent interest in the operations research and machine learning literature, is the Wasserstein distance. Here we show how to extend our analysis to statistical neighborhoods defined in terms of Wasserstein distance.[15]

Consider a joint distribution between two random vectors X and Z. Now fix the marginal distributions while searching over alternative joint distribution to minimize the expectation of $|X - Z|^p$. The pth root of the objective gives a Wasserstein distance between the pre-specified marginal distributions. The resulting optimization problem is recognizable as an optimal transport problem.

For our analysis we combine minimization required for computing the Wasserstein distance with the minimization problems that interest us when deducing the moment bounds. To accomplish this, let Z be statistically independent of X in accordance with the $\mathbb{E}$

[15] See for instance Arjovsky et al. (2017).

expectation. In addition, let M denote a positive random variable used to change the distribution of X conditioned on Z while leaving the distribution for Z intact. This is enforced by restricting $\mathbb{E}\left[M \mid Z\right] = 1$.[16] We then solve:

$$\inf_{M \geqslant 0} \mathbb{E}\left[M \left|X - Z\right|^p\right]$$

where

$$\begin{aligned} \mathbb{E}\left[M f(X)\right] &= 0 \\ \mathbb{E}\left[M g(X)\right] &\leqslant \vartheta \\ \mathbb{E}\left[M \mid Z\right] &= 1 \end{aligned}$$

In principle, this can be solved as a linear programming problem. However, in practice, this linear programming problem is computationally costly.

Cuturi (2013) proposed a tractable regularization of the objective obtained by adding a small relative entropy penalty to the objective function:

$$\mathbb{E}\left[M \left|X - Z\right|^p\right] + \epsilon \mathbb{E}\left[M \log M\right]$$

To characterize the solution to the penalized problem, form the Lagrangian:

$$\max_{\lambda, \rho \geqslant 0}, \max_{\nu \geqslant 0} \min_{M \geqslant 0} \mathbb{E}\left[M\left(|X - Z|^p + \epsilon \log M + \lambda \cdot f(X) + \rho g(X) + \nu\right)\right] - \mathbb{E}\nu - \rho\vartheta$$

where ρ is a nonnegative scalar, λ is a vector of real numbers and ν can depend on Z. We may solve the "inner problem" by first conditioning in Z:

$$\max_{\nu \geqslant 0} \min_{M \geqslant 0} \mathbb{E}\left[M\left(|X - Z|^p + \epsilon \log M + \lambda \cdot f(X) + \rho g(X) + \nu\right) \mid Z\right] - \nu$$

[16] To see that this preserves the marginal distribution over Z, note that by the law of iterated expectations, for any Borel-measurable function ψ of Z we have that

$$\begin{aligned} \mathbb{E}[M\psi(Z)] &= \mathbb{E}[\mathbb{E}[M\psi(Z)|Z]] \\ &= \mathbb{E}[\psi(Z)\mathbb{E}[M|Z]] \\ &= \mathbb{E}[\psi(Z)]. \end{aligned}$$

$$= -\epsilon \log \mathbb{E}\left[\exp\left(-\frac{1}{\epsilon}\left[|X - Z|^p + \lambda \cdot f(X) + \rho g(X)\right]\right)\Big| Z\right]$$

where we use an argument that is entirely similar to an earlier derivation for the relative entropy discrepancy. This leads to solve an outer optimization problem:

$$\max_{\lambda,\rho \geqslant 0} -\epsilon \mathbb{E}\left[\log \mathbb{E}\left[\exp\left(-\frac{1}{\epsilon}\left[|X - Z|^p + \lambda \cdot f(X) + \rho g(X)\right]\right)\Big| Z\right]\right] - \rho\vartheta.$$

Using this approach, we may approximate the solution to the original Wasserstein distance problem by setting ϵ to be sufficiently small.

Remark 5.1. *Regularized Wasserman distances for a given ϵ may be computed efficiently using "Sinkhorn Iterations." See Léger (2020) for a recent formal justification for the convergence of these iterations for the case in which $p = 2$. Notice the relative penalization in our formulation has X and Z independent while the term $\mathbb{E}\,|X - Z|^p$ will be zero only when $Z = X$. Thus even under correct specification, penalized discrepancy will not be zero. This tension gets reduced as we make ϵ arbitrarily small.*

Remark 5.2. *Xie et al. (2019) note some difficulties in implementing the relative entropy regularization for small values of ϵ. They propose an alternative approach whereby the approximation to the minimizing M is obtained through an iterative scheme within which ϵ tends to zero. Moreover, they propose replacing the baseline probability used to measure relative entropy by the one computed in a previous iteration. We cannot directly implement their approach since our optimization problem differs for the reasons that we explained. Nevertheless, their strategy for reducing ϵ as part of the iterations and changing the baseline probability used in the relative entropy penalty may have a tractable counterpart for our problem.*

Remark 5.3. *We posed the Wasserstein distance in terms of the random vector X and its coupled counterpart. But there may be little reason to prefer X to alternative nonlinear transformations of X for our application. The other measures of statistical divergence that we considered do not require such a distinction, and in this sense might be more appropriate for our application.*[17]

[17] Another option would be to replace X by $g(X)$ when forming the Wasserstein criterion. But we prefer to hold fixed the measure of divergence as we change the functions of g of X whose expectations we seek to bound.

6 Minimum divergence estimation and inference with unknown parameters

So far, we have suppressed the parameter dependence and focused solely on the belief distortion. In this section we include an unknown parameter vector θ in a parameter space Θ in the specification of the moment conditions and suggest some large sample approaches to inference. While minimizing the divergence can be thought of as an objective for estimation, we deliberately avoid assuming point identification of "pseudo-true" parameter for θ. This perspective is consistent with our subsequent investigation when we entertain larger divergence bound which will typically leave us with a set of parameter values θ in the corresponding population problem.

In view of the arguments made in the previous sections, we focus only on the ϕ-divergences that are not strictly decreasing. For convenience we use the divergence in the family given by (3) for $\eta \geqslant 0$. With this family, we lever the well known Legendre transforms of these convex functions in formulated tractable dual problems. In addition to the unknown parameters, the Lagrange multipliers are of interest for our analysis because they suggest how one might reshape the distributions to satisfy the moment conditions.[18] While the econometric objective posed using duality is concave in the multipliers for each θ, this shape restriction does not carry over to the global dependence on θ, as θ can enter the moment functions $f(X, \theta)$ nonlinearly and non-separably. While we do not assume that θ is uniquely identified, we do restrict that the multipliers be unique for each $\theta \in \Theta$. We previously presented some sufficient conditions for this uniqueness.

6.1 Extended dual problem

In this section we show how to extend our previous analysis to include parameter uncertainty. We impose the following restrictions on the parameter space and moment conditions.

Assumption 6.1. *(i) Θ is compact with non-empty interior Θ^o; (ii) For each $\theta \in \Theta$, there is no nondegenerate linear combination of the $f(\theta, x)$ that is independent of x for all potential realizations x of X; and (iii) for each $\theta \in \Theta^o$ and each x, $f(x, \theta)$ is continuously differentiable in θ.*

The problem of interest inclusive of the unknown parameter θ is:

[18] This is a substantial development of an original idea suggested by Back and Brown (1993).

Problem 6.2.

$$\underline{\kappa} \doteq \min_{\theta \in \Theta} \mathcal{L}(\theta)$$

where for any fixed $\theta \in \Theta$,

$$\mathcal{L}(\theta) \doteq \inf_{M \geqslant 0} \mathbb{E}\left[\phi(M)\right]$$

subject to:

$$\begin{aligned} &\mathbb{E}\left[M f(X, \theta)\right] = 0, \\ &\mathbb{E}\left[M\right] = 1. \end{aligned}$$

Let $\mu \doteq (\lambda', \nu)'$ denote the composite multipliers for the two sets of constraints. The minimized divergence for a given θ solves the dual problem

$$\mathcal{L}(\theta) = \max_{\mu} \inf_{M \geqslant 0} \ \mathbb{E}\left[\phi(M) - \lambda \cdot f(X, \theta) M - \nu(M - 1)\right].$$

We further simplify this problem by bringing the minimization with respect to M inside the expectation. Let $M(X, \mu, \theta)$ denote the resulting solution, which is given by

$$M(X, \mu, \theta) = \begin{cases} \left(\left[\eta[\lambda \cdot f(X, \theta) + \nu] + \frac{1}{1+\eta}\right]^{+}\right)^{\frac{1}{\eta}}, & \eta > 0 \\ \exp\left[\lambda \cdot f(X, \theta) + \nu - 1\right], & \eta = 0 \end{cases} \tag{9}$$

for almost all X. The notation $[\cdot]^{+}$ denotes a) the number in the square brackets if it is nonnegative and b) zero if the number is negative. Substituting this solution back in to the objective gives the following dual alternative to problem 6.2:

Problem 6.3.

$$\underline{\kappa} \doteq \min_{\theta \in \Theta} \mathcal{L}(\theta)$$

where for any fixed $\theta \in \Theta$,

$$\mathcal{L}(\theta) = \max_{\mu} \ \mathbb{E}\left[F(X, \mu, \theta)\right] \tag{10}$$

where

$$F(x, \mu, \theta) = \begin{cases} -\frac{1}{1+\eta}\left(\left[\eta\lambda \cdot f(x, \theta) + \eta\nu + \frac{1}{1+\eta}\right]^{+}\right)^{\frac{1+\eta}{\eta}} + \nu & \eta > 0 \\ -\exp\left[\lambda \cdot f(x, \theta) + \nu - 1\right] + \nu & \eta = 0 \end{cases}$$

The solutions $\mu^*(\theta)$ to Problem (10) are given by the first-order conditions wrt μ:

$$\mathbb{E}\left[\frac{\partial F}{\partial \mu}(X,\mu,\theta)\right] = \mathbb{E}\begin{bmatrix} -f(X,\theta)M(X,\mu,\theta) \\ 1 - M(X,\mu,\theta) \end{bmatrix} = 0. \tag{11}$$

These first-order conditions imply a function μ^* of θ so that

$$\mathbb{E}\left[\frac{\partial}{\partial \mu}F(X,\mu^*(\theta),\theta)\right] = 0 \quad \text{for all} \quad \theta \in \Theta^o.$$

The corresponding optimized belief distortion as a function of θ using formula (9) is given by

$$M^*(\theta) \doteq M(X,\mu^*(\theta),\theta) \tag{12}$$

By familiar Envelope Theorem argument, we have

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X,\mu^*(\theta),\theta)\right] \quad \text{for all} \quad \theta \in \Theta^o. \tag{13}$$

To provide additional formulas of interest in our supporting analysis, we assume:

Assumption 6.4. *For each θ in Θ^o, (i) the matrix*

$$\mathbf{H}(\mu,\theta) \doteq \mathbb{E}\left(\frac{\partial^2 F}{\partial \mu \partial \mu'}[X,\mu,\theta]\right) \quad \textit{is continuous in a neighborhood of} \quad \mu^*(\theta).$$

(ii) the matrix $\mathbf{H}^*(\theta) \doteq \mathbf{H}(\mu^*(\theta),\theta)$ is negative definite.

Remark 6.5. *This formula for $\mathbf{H}$ is directly applicable for $0 \leqslant \eta < 1$. As we discussed previously, we are also interested in the case in which $\eta = 1$. In this case the F may fail to be twice continuously differentiable for some realizations x of X. We may include this more general case provided that the expectation smooths out the kink points in the first derivative of F.*

By the Implicit Function theorem, we have that $\mu^*(\theta)$ is continuously differentiable in $\theta \in \Theta^o$:

$$\frac{d\mu^*(\theta)}{d\theta} = -\left[\mathbf{H}^*(\theta)\right]^{-1}\mathbb{E}\left(\frac{\partial^2 F}{\partial \mu \partial \theta'}[X,\mu^*(\theta),\theta]\right).$$

Remark 6.6. *For the relative entropy ($\eta = 0$) case, the above results simplify as $\nu^*(\theta)$ could be solved explicitly as a function of $\lambda^*(\theta)$, and the implied minimal divergence belief* *is*

$$M^*(\theta) = \frac{\exp\left[\lambda^*(\theta) \cdot f(X,\theta)\right]}{\mathbb{E}\left(\exp\left[\lambda^*(\theta) \cdot f(X,\theta)\right]\right)}.$$

Furthermore,

$$\mathbb{E}\left[\phi(M^*(\theta))\right] = -\log \mathbb{E}\left(\exp\left[\lambda^*(\theta) \cdot f(X,\theta)\right]\right),$$

and

$$\frac{d\lambda^*(\theta)}{d\theta} = -\left(\mathbb{E}[M^*(\theta) f(X,\theta) f(X,\theta)']\right)^{-1} \mathbb{E}\left[M^*(\theta)(1 + \lambda^*(\theta) \cdot f(X,\theta)) \frac{\partial f(X,\theta)}{\partial \theta}\right].$$

Consider the set

$$\Theta^* \doteq \arg\min_{\theta \in \Theta} \mathcal{L}(\theta),$$

which is the set of minimal ϕ-divergence implied pseudo-true model parameter values. Under correct specification, $\underline{\kappa} = 0$, and $M^*(\theta) = 1$ for $\theta \in \Theta^*$. Our interest is when $\underline{\kappa} > 0$. While Θ^* is assumed to be a singleton in many investigations of misspecification, we deliberately avoid imposing this restriction.

Assumption 6.7. $\Theta^* \subset \Theta^o$.

6.2 Estimation based on the dual problem

6.2.1 Estimation of the Lagrange multipliers conditioned on θ

In this subsection we study the estimation of multipliers conditioned on parameters. The estimation of the multipliers for alternative values of the parameters give us an important intermediate component to our analysis. By conditioning on the parameters, the minimum divergence problem has a nice mathematical structure because the dual objective function is concave in the Lagrange multipliers.

We restrict the DGP as follows.

Assumption 6.8. *The process $\{X_t : t \geqslant 0\}$ is strictly stationary, $\beta - mixing$.*

The estimation of $\mu^*(\theta)$ for each θ is a special case of an M-estimation problem with a concave objective function where the sample counterpart to Problem 6.3 is given as follows:

Problem 6.9.

$$\mathcal{L}_T(\theta) = \max_{\mu} \ \frac{1}{T}\sum_{t=1}^{T} F(X_t, \mu, \theta) = \frac{1}{T}\sum_{t=1}^{T} F(X_t, \mu_T(\theta), \theta)$$

where $\mu_T(\theta)$ is the corresponding estimate for $\mu^*(\theta)$.

This problem fits within the framework analyzed by Haberman (1989), Hjort and Pollard (1993) among others. Since our data is assumed to be $\beta - mixing$, we apply Chen and Shen (1998) for results on time series M-estimation.

We proceed by obtaining a functional central limit approximation for:

$$\sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t, \mu_T(\theta), \theta) - F(X_t, \mu^*(\theta), \theta)\right] + F_T^*(\theta)$$

where

$$F_T^*(\theta) \doteq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t, \mu^*(\theta), \theta) - \mathbb{E}F(X_t, \mu^*(\theta), \theta)\right].$$

Only the second term $F_T^*(\theta)$ contributes to the approximation. To see why, note that since F is concave in μ for each θ, a gradient inequality for such functions implies that

$$0 \leqslant \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[F(X_t, \mu_T(\theta), \theta) - F(X_t, \mu^*(\theta), \theta)\right] \leqslant \left[\mu_T(\theta) - \mu^*(\theta)\right] \cdot h_T^*(\theta),$$

where

$$h_T^*(\theta) \doteq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial F}{\partial \mu}(X_t, \mu^*(\theta), \theta).$$

This leads us to focus on a joint functional central limit approximation for $F_T^*(\theta)$ and h_T^* for making approximate inferences using $\mathcal{L}_T(\theta)$.

Assumption 6.10.

i) $\{F_T^*(\theta) : \theta \in \Theta\}$ is Donsker, converges weakly to a tight Gaussian process $\{\mathcal{G}(\theta) : \theta \in \Theta\}$ with zero mean and covariance function

$$C^*(\theta_1, \theta_2) \doteq \lim_{T\to\infty} Cov\left[F_T^*(\theta_1), F_T^*(\theta_2)\right] = \sum_{j=-\infty}^{\infty} Cov\left[F(X_1, \mu^*(\theta_1), \theta_1), F(X_{1+j}, \mu^*(\theta_2), \theta_2)\right]$$

ii) The process $\{h_T^*(\theta) : \theta \in \Theta\}$ is Donsker, converges weakly to a tight Gaussian process *with zero mean and covariance function*

$$\mathbf{V}^*(\theta) \doteq \lim_{T\to\infty} Var\left[h_T^*(\theta)\right] = \sum_{j=-\infty}^{\infty} Cov\left[\frac{\partial F}{\partial \mu}(X_1, \mu^*(\theta), \theta), \frac{\partial F}{\partial \mu}(X_{1+j}, \mu^*(\theta), \theta)\right]$$

Sufficient conditions for the central limit approximations entail verifying weak convergence for any finite collections of θ's in conjunction a tightness restriction implied by some form of stochastic equicontinuity. Such an approximation may be obtained from more primitive assumptions on the β-mixing coefficients of data-generating process $\{X_t\}$ and restrictions on the functions of X_t and θ. See Doukhan et al. (1995), Dedecker and Louhichi (2002).

Under Assumption 6.10, we obtain the following result:

Result 6.11.

1. *Uniformly over $\theta \in \Theta$, $\sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] = F_T^*(\theta) + o_p(1)$, converges weakly to the Gaussian process $\{\mathcal{G}(\theta) : \theta \in \Theta\}$*

2. *Uniformly in $\theta \in \Theta$,*

$$\sqrt{T}\left[\mu_T(\theta) - \mu^*(\theta)\right] = -\left[\mathbf{H}^*(\theta)\right]^{-1} h_T^*(\theta) + o_p(1),$$

which converges weakly to a normally distributed random vector with mean zero and covariance:

$$\left[\mathbf{H}^*(\theta)\right]^{-1} \mathbf{V}^*(\theta) \left[\mathbf{H}^*(\theta)\right]^{-1}.$$

6.2.2 Estimation of $\underline{\kappa}$

For $\mathcal{L}_T(\theta)$ given in Problem 6.9, a simple plug-in estimator of $\underline{\kappa}$ is given by

$$\underline{\kappa}_T = \min_{\theta\in\Theta} \mathcal{L}_T(\theta),$$

and a corresponding set estimate of Θ^* is given by

$$\Theta_T^* = \{\theta \in \Theta : \mathcal{L}_T(\theta) = \underline{\kappa}_T + o_P(T^{-1})\}.$$

Then under mild conditions including those ensuring the uniqueness of $\mu^*(\theta)$, we obtain a direct extension of Theorem 3.6 of Shapiro (1991) from iid data to stationary $\beta - mixing$

data,

Result 6.12. $\sqrt{T}(\underline{\kappa}_T - \underline{\kappa}) = \min_{\theta \in \Theta^*} \sqrt{T}\left[\mathcal{L}_T(\theta) - \mathcal{L}(\theta)\right] + o_p(1) \rightsquigarrow \min_{\theta \in \Theta^*} \mathcal{G}(\theta)$.

We observe that if Θ^* is a singleton set $\{\theta_0\}$, then $\sqrt{T}(\underline{\kappa}_T - \underline{\kappa}) \rightsquigarrow \mathcal{G}(\theta_0)$, which is a mean zero normal random variable with variance $C^*(\theta_0, \theta_0)$.

6.3 Confidence sets via quasi-posteriors

In devising inferential methods, we target sets of solutions to the first-order conditions. In the case of the unknown parameter θ, this leads us to construct

$$\underline{\Theta} \doteq \left\{\theta \in \Theta^o : \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X, \mu^*(\theta), \theta)\right] = 0\right\},$$

which could be larger than Θ^*. By looking at solutions to first-order conditions (as we do in the construction of $\underline{\Theta}$, we run risk of including too many θ's, which might make our inferences conservative. We now combine the first order conditions for θ with those for μ, leading us to construct:

$$\rho(x, \mu, \theta) = \begin{bmatrix} \frac{\partial F}{\partial \mu}(x, \mu, \theta) \\ \frac{\partial F}{\partial \theta}(x, \mu, \theta) \end{bmatrix},$$

Then the joint first-order conditions with respect to $\beta = (\mu, \theta)$ is:

$$\mathbb{E}[\rho(X, \beta)] = 0. \tag{14}$$

We let $\mathbf{B}$ be the space of admissible parameter values for β. To formulate a tractable inference approach, we view (14) as the moment conditions for a "just-identified" GMM estimation problem.

Remark 6.13. *If $\underline{\Theta}$ is further assumed to be a singleton set $\{\theta_0\}$, then the sample analog of Problem 6.3 will lead to the joint asymptotically normal frequentist estimates for $(\mu^*(\theta_0), \theta_0)$. Several papers, such as Schennach (2007), Broniatowski and Keziou (2012), and Lee (2016), have used these just-identified moment conditions to establish root-T asymptotic normality of their estimators for $(\mu^*(\theta_0), \theta_0)$ jointly for possibly $\mathbb{E}[f(X, \theta)] \neq 0$ with iid data. Almeida and Garcia (2012) establish root-T asymptotic normality for $(\mu^*(\theta_0), \theta_0)$ under stationary strongly mixing data.*

We follow a recently developed approach of Chen et al. (2018) to compute critical values for confidence sets based on a "quasi posterior" designed to allow for set-valued $\underline{\Theta}$. We implement their approach by first following Hansen et al. (1996) we define the continuously-updated GMM criterion function:

$$L_T(\beta) = -\frac{1}{2}\left[\frac{1}{T}\sum_{t=1}^{T}\rho(X_t,\beta)\right]' [\Sigma_T(\beta)]^- \left[\frac{1}{T}\sum_{t=1}^{T}\rho(X_t,\beta)\right]$$

where, for each β,$[\Sigma_T(\beta)]^-$ is the generalized inverse of $\Sigma_T(\beta)$, which is a consistent estimator of $\Sigma(\beta)$:

$$\Sigma(\beta) = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\rho(X_t,\beta)\right)$$

where $Var(\cdot)$ denotes the covariance matrix of the argument in parentheses. Given $L_T(\beta)$, the data $\mathbf{X}$ and a prior Π over $\mathbf{B}$, the quasi-posterior distribution Π_T for β given $\mathbf{X}$ is defined as

$$d\Pi_T(\beta \mid \mathbf{X}) = \frac{\exp[TL_T(\beta)]d\Pi(\beta)}{\int_{\mathbf{B}}\exp[TL_T(\beta)]d\Pi(\beta)}\,. \tag{15}$$

6.3.1 Confidence sets for the unknown parameter vector and multipliers

Consider first inferences about the composite set of parameters

$$\underline{\mathbf{B}} \doteq \{\beta = (\mu^*(\theta),\theta) : \theta \in \underline{\Theta}\}.$$

Draw a sample $\{\beta^1,\ldots,\beta^N\}$ from the quasi-posterior Π_T. Any Monte Carlo sampler could be used. Chen et al. (2018) suggested to use an adaptive sequential Monte Carlo (SMC) algorithm that is known to perform well for drawing from irregular, multi-modal distributions. Construct a confidence set $\mathbf{B}_T^\alpha$ for $\underline{\mathbf{B}}$ such that $\lim_{T\to\infty} Pr(\underline{\mathbf{B}} \subseteq \mathbf{B}_T^\alpha) = \alpha$ as follows:[19]

1. Draw a sample $\{\beta^1,\ldots,\beta^N\}$ from the quasi-posterior distribution Π_T in (15).

2. Calculate the $(1-\alpha)$ quantile of $\{L_T(\beta^1),\ldots,L_T(\beta^N)\}$; call it $\zeta_{T,\alpha}^{mc}$.

[19] This is procedure I in Chen et al. (2018).

3. Our $100\alpha\%$ confidence set for $\underline{\mathbf{B}}$ is then:

$$\mathbf{B}_T^\alpha = \{\beta \in \mathbf{B} : L_T(\beta) \geqslant \zeta_{T,\alpha}^{mc}\} \,. \tag{16}$$

Remark 6.14. *A simple projection-based confidence set for $\underline{\Theta}$ is given by:*

$$\widehat{\underline{\Theta}}_T^\alpha = \{\theta : (\mu, \theta) \in \boldsymbol{B}_T^\alpha \textit{ for some } \mu\} \tag{17}$$

which is a valid $100\alpha\%$ confidence set for $\underline{\Theta}$ whenever $\boldsymbol{B}_T^\alpha$ is a valid $100\alpha\%$ confidence set for $\underline{\boldsymbol{B}}$. However, this approach is known to be conservative:

$$\lim_{T\to\infty} Pr(\underline{\Theta} \subseteq \widehat{\underline{\Theta}}_T^\alpha) > \alpha \ .$$

6.3.2 Confidence sets for θ based on profiled moment

Note that for each fixed θ, we have unique solution $\mu^*(\theta)$ of μ, which satisfies

$$\mathbb{E}\left[\frac{\partial F}{\partial \mu}(X, \mu^*(\theta), \theta)\right] = 0 \quad \forall \theta \in \Theta^o.$$

There is an analogous sample counterpart function, $\mu_T(\theta)$, which is also the solution to Problem 6.9:

$$\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial F}{\partial \mu}(X_t, \mu_T(\theta), \theta)\right] = 0 \quad \forall \theta \in \Theta^o. \tag{18}$$

As we noted previously, by Implicit Function Theorem:

$$\begin{aligned}
\frac{d\mu^*(\theta)}{d\theta} &= -\left(\mathbb{E}\left[\frac{\partial^2 F}{\partial\mu\partial\mu'}(X, \mu^*(\theta), \theta)\right]\right)^{-1} \mathbb{E}\left[\frac{\partial F}{\partial\mu\partial\theta'}(X, \mu^*(\theta), \theta)\right] \\
\frac{d\mu_T(\theta)}{d\theta} &= -\left(\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial^2 F}{\partial\mu\partial\mu'}(X_t, \mu_T(\theta), \theta)\right]\right)^{-1} \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial F}{\partial\mu\partial\theta'}(X_t, \mu_T(\theta), \theta)\right]
\end{aligned}$$

where the second formula is just the finite sample counterpart of the first.

Hypothetically, we could follow the previous approach to provide a consistent confidence set for the identified set $\underline{\Theta}$ based on the following moment condition for θ that adjusts for estimation of μ^*:

$$\rho^*(X, \theta) = \frac{\partial F}{\partial \theta}(X, \mu^*(\theta), \theta) + \frac{\partial F}{\partial \mu}(X, \mu^*(\theta), \theta)\frac{d\mu^*(\theta)}{d\theta}.$$

Since μ^* is constructed from an optimization problem given θ, we have:

$$\underline{\Theta} = \left\{\theta \in \Theta^o : \mathbb{E}\left[\frac{\partial F}{\partial \theta}(X, \mu^*(\theta), \theta)\right] = 0\right\} = \{\theta \in \Theta^o : \mathbb{E}[\rho^*(X, \theta)] = 0\}$$

This approach is not feasible because we do not know μ^*, only its consistent estimator μ_T. It suffices, however, to substitute, μ_T for μ^* and proceed analogously. Thus we form a new continuously-updated GMM criterion function:

$$L_T(\theta) = -\frac{1}{2} G_T(\theta)'[\Sigma_T(\theta)]^- G_T(\theta)$$

where for each θ,

$$\begin{aligned} G_T(\theta) =& \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial F}{\partial \theta}(X_t, \mu_T(\theta), \theta) + \frac{\partial F}{\partial \mu}(X_t, \mu_T(\theta), \theta)\frac{d\mu_T(\theta)}{d\theta}\right] \\ =& \frac{1}{T}\sum_{t=1}^{T}\frac{\partial F}{\partial \theta}(X_t, \mu_T(\theta), \theta). \end{aligned}$$

The matrix $[\Sigma_T(\theta)]^-$ is the generalized inverse of $\Sigma_T(\theta)$, which is a consistent estimator of $\Sigma(\theta)$:

$$\Sigma(\theta) = \lim_T Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\rho^*(X_t, \theta)\right).$$

Notice that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial F}{\partial \mu}(X_t, \mu^*(\theta), \theta)\frac{d\mu^*(\theta)}{d\theta}$$

contributes to $\Sigma(\theta)$ and is included to adjust for the estimation of μ^*.

Given $L_T(\theta)$, the data $\mathbf{X} = \{X_t\}_{t=1}^T$ and a prior Π over Θ, the quasi-posterior distribution Π_T for θ is

$$d\Pi_T(\theta \mid \mathbf{X}) = \frac{\exp[TL_T(\theta)]d\Pi(\theta)}{\int_\Theta \exp[TL_T(\theta)]d\Pi(\theta)}. \tag{19}$$

We draw a sample $\{\theta^1, \ldots, \theta^N\}$ from the quasi-posterior Π_T. We seek a CS $\widehat{\underline{\Theta}}_\alpha$ for $\underline{\Theta}$ such that $\lim_{T\to\infty} Pr(\underline{\Theta} \subseteq \widehat{\underline{\Theta}}_\alpha) = \alpha$.

Confidence sets for $\underline{\Theta}$:

1. Draw a sample $\{\theta^1, \ldots, \theta^N\}$ from the quasi-posterior distribution Π_T in (19).

2. Calculate the $(1-\alpha)$ quantile of $\{L_T(\theta^1), \ldots, L_T(\theta^N)\}$; call it $\zeta^{mc}_{T,\alpha}$.

3. Our $100\alpha\%$ confidence set for $\underline{\Theta}$ is then:

$$\widehat{\underline{\Theta}}_\alpha = \{\theta \in \Theta : L_T(\theta) \geqslant \zeta^{mc}_{T,\alpha}\}\,. \tag{20}$$

Remark 6.15. *As Chen et al. (2018) note, confidence sets for individual components of the θ vector can be constructed by minimizing over the remaining parameters and using a chi-square one threshold.*[20]

7 Estimation and Inference on nonlinear expectation functionals

This section presents estimation and inference for the nonlinear expectation functionals we constructed in section 4. We first describe a direct approach which approximates the nonlinear expectation by a finite-sample counterpart. Afterwards, we describe an indirect approach which treats the distorted expectations as a set-identified parameter, and then recover the nonlinear expectation bounds from a robust confidence set for this parameter.

7.1 Bounding expectation functionals using divergence balls

We first extend the previous results by letting κ be a parameter satisfying $\kappa \geqslant \underline{\kappa}$. Using the defintion of $M^*(\theta)$ given in equation (12) we define

$$\underline{\vartheta} \doteq \min_{\theta \in \Theta} \mathbb{E}\left[M^*(\theta) g(X,\theta)\right].$$

Given a real-valued function g of X and θ, we consider the following problem:

Problem 7.1. *Let $\kappa \geqslant \underline{\kappa}$ and $\underline{\vartheta} < \infty$.*

$$\mathbb{K}(\kappa) \doteq \min_{\theta \in \Theta} \mathcal{K}(\theta, \kappa) \leqslant \underline{\vartheta},$$

$$\mathcal{K}(\theta,\kappa) \doteq \inf_{M \geqslant 0} \mathbb{E}\left[M g(X,\theta)\right] \quad \textit{subject to}$$

[20] See their procedure 3 for details.

$$\begin{aligned}&\mathbb{E}\left[Mf(X,\theta)\right]=0,\\&\mathbb{E}\left[M\right]=1,\\&\mathbb{E}\left[\phi(M)\right]\leqslant\kappa.\end{aligned}$$

Similar to the previous section, the dual problem is

$$\mathcal{K}(\theta,\kappa)=\max_{\xi\geqslant 0}\max_{\mu=(\lambda,\nu)}\inf_{M\geqslant 0}\ \mathbb{E}\left[Mg(X,\theta)+\xi(\phi(M)-\kappa)-\lambda\cdot f(X,\theta)M-\nu(M-1)\right].$$

Let $M(X,\xi,\mu,\theta)$ denote the solution to the inner part $\inf_{M\geqslant 0}[\cdot]$ of the above dual problem, which is given by (for $\xi>0$ only)

$$M(X,\xi,\mu,\theta)=\begin{cases}\left(\left[\eta(\xi)^{-1}[\lambda\cdot f(X,\theta)+\nu-g(X,\theta)]+\frac{1}{1+\eta}\right]^{+}\right)^{\frac{1}{\eta}}\ , & \eta>0\\ \exp\left[(\xi)^{-1}[\lambda\cdot f(X,\theta)+\nu-g(X,\theta)]-1\right]\ , & \eta=0\end{cases}\tag{21}$$

for almost all X. Then the dual problem could be re-expressed as:

Problem 7.2. *Let $\kappa\geqslant\underline{\kappa}$. For any fixed $\theta\in\Theta$,*

$$\mathcal{K}(\theta,\kappa)=\max_{\xi>0}\max_{\mu}\ \mathbb{E}\left[F(X,\xi,\mu,\theta,\kappa)\right],$$

where

$$F(x,\xi,\mu,\theta,\kappa)=\begin{cases}-\frac{\xi}{1+\eta}\left(\left[\eta(\xi)^{-1}[\lambda\cdot f(X,\theta)+\nu-g(X,\theta)]+\frac{1}{1+\eta}\right]^{+}\right)^{\frac{1+\eta}{\eta}}+\nu-\xi\kappa\ , & \eta>0\\ -\xi\exp\left[(\xi)^{-1}[\lambda\cdot f(X,\theta)+\nu-g(X,\theta)]-1\right]+\nu-\xi\kappa\ , & \eta=0\end{cases}$$

We have:

$$\begin{aligned}\frac{\partial F}{\partial\mu}(x,\xi,\mu,\theta,\kappa)&=\begin{bmatrix}-f(x,\theta)M(x,\xi,\mu,\theta)\\ 1-M(x,\xi,\mu,\theta)\end{bmatrix},\\ \frac{\partial F}{\partial\xi}(x,\xi,\mu,\theta,\kappa)&=\phi(M(x,\xi,\mu,\theta))-\kappa\end{aligned}$$

The solutions to Problem 7.2 are given by the first-order conditions wrt $(\mu,\xi>0)$:

$$\mathbb{E}\left[\frac{\partial F}{\partial\mu}(X,\xi,\mu,\theta,\kappa)\right]=0,\ \ \mathbb{E}\left[\frac{\partial F}{\partial\xi}(X,\xi,\mu,\theta,\kappa)\right]=0,$$

which is

$$
\begin{aligned}
-\mathbb{E}\left[f(X,\theta)M(X,\xi,\mu,\theta)\right] =& 0 \\
1-\mathbb{E}\left[M(X,\xi,\mu,\theta)\right] =& 0 \\
\mathbb{E}[\phi(M(X,\xi,\mu,\theta))]-\kappa =& 0 \qquad (22)
\end{aligned}
$$

Remark 7.3. *More generally we should allow for $\xi \geqslant 0$ in (22) by replacing the 3rd equation by*

$$\xi \geqslant 0, \quad \xi\left(\mathbb{E}[\phi(M(X,\xi,\mu,\theta))]-\kappa\right)=0.$$

Appendix A gives sufficient conditions under which the relative entropy constraint binds, which rules out $\xi=0$ as an optimal solution.

Let $\mu_0(\theta,\kappa),\xi_0(\theta,\kappa)$ denote the solution to (22), and $M_0(\theta,\kappa)=M(X,\xi_0(\theta,\kappa),\mu_0(\theta,\kappa),\theta)$ given in (21) denote the (g,κ,θ) implied belief. Then Problem 7.1 becomes

$$\mathbb{K}(\kappa) \doteq \min_{\theta\in\Theta}\mathcal{K}(\theta,\kappa)$$

$$\mathcal{K}(\theta,\kappa)=\max_{\xi>0}\max_{\mu}\ \mathbb{E}\left[F(X,\xi,\mu,\theta,\kappa)\right]=\mathbb{E}\left[F(X,\xi_0(\theta,\kappa),\mu_0(\theta,\kappa),\theta,\kappa)\right]=\mathbb{E}\left[M_0(\theta,\kappa)g(X,\theta)\right].$$

We note that

$$
\begin{aligned}
\frac{\partial\mathcal{K}(\theta,\kappa)}{\partial\theta} &= \mathbb{E}\left[\frac{\partial F}{\partial\theta}(X,\xi_0(\theta,\kappa),\mu_0(\theta,\kappa),\theta,\kappa)\right], \\
\frac{\partial\mathcal{K}(\theta,\kappa)}{\partial\kappa} &= -\xi_0(\theta,\kappa)<0.
\end{aligned}
$$

Thus $\mathbb{K}(\kappa)$ increases as κ decreases as long as $\kappa \geqslant \underline{\kappa}$. Moreover,

$$\mathbb{K}(\kappa)=+\infty \ \text{ and } \ \mathcal{K}(\theta,\kappa)=+\infty \quad \text{for all } \ \kappa<\underline{\kappa}.$$

Remark 7.4. *For relative entropy ($\eta=0$) case, Problem 7.1 simplifies to: let $\kappa \geqslant \underline{\kappa}$,*

$$\mathbb{K}(\kappa) \doteq \min_{\theta\in\Theta}\mathcal{K}(\theta,\kappa)$$

$$\mathcal{K}(\theta,\kappa)=\max_{\xi>0}\max_{\lambda}\ -\xi\log\mathbb{E}[\exp\left[(\xi)^{-1}(\lambda\cdot f(X,\theta)-g(X,\theta))\right]]-\xi\kappa=\mathbb{E}\left[M_0(\theta,\kappa)g(X,\theta)\right],$$

with

$$M_0(\theta,\kappa)=\frac{\exp\left[(\xi_0(\theta,\kappa))^{-1}(\lambda_0(\theta)\cdot f(X,\theta)-g(X,\theta))\right]}{\mathbb{E}[\exp\left[(\xi_0(\theta,\kappa))^{-1}(\lambda_0(\theta)\cdot f(X,\theta)-g(X,\theta))\right]]}.$$

Remark 7.5. *In this section we focus on the divergences ϕ (with $\eta \geqslant 0$) and moment functions f and g such that, for any fixed $\kappa \geqslant \underline{\kappa}$ and $\theta \in \Theta$, there is a unique multipliers $(\mu_0(\theta, \kappa), \xi_0(\theta, \kappa))$ and a unique belief $M_0(\theta, \kappa) = M(X, \xi_0(\theta, \kappa), \mu_0(\theta, \kappa), \theta)$ given in (21) solves the dual problem Problem 7.2. For any fixed $\kappa \geqslant \underline{\kappa}$ we denote*

$$\Theta_\kappa \doteq \arg\min_{\theta} \mathcal{K}(\theta, \kappa) = \{\theta \in \Theta : \mathcal{K}(\theta, \kappa) = \mathbb{K}(\kappa)\}$$

as the set of pseudo-true model parameter values that solves Problem 7.1.

In the rest of the section, to simplify notation we drop κ from the definition of $\mathbb{K}(\kappa)$, $\mathcal{K}(\theta, \kappa)$, $F(X, \xi, \mu, \theta, \kappa)$, $(\mu_0(\theta, \kappa), \xi_0(\theta, \kappa))$ and $M_0(\theta, \kappa)$.

7.1.1 Profile M-estimation of Lagrange multipliers

The sample counterpart to Problem 7.1 is Problem 7.6:

Problem 7.6. *For any given $\kappa \geqslant \underline{\kappa}$,*

$$\widehat{\mathbb{K}} \doteq \min_{\theta \in \Theta} \mathcal{K}_T(\theta)$$

$$\mathcal{K}_T(\theta) = \max_{\xi > 0} \max_{\mu} \frac{1}{T} \sum_{t=1}^{T} [F(X_t, \xi, \mu, \theta)] = \frac{1}{T} \sum_{t=1}^{T} [F(X_t, \xi_T(\theta), \mu_T(\theta), \theta)],$$

where $\xi_T(), \mu_T()$ are the corresponding estimates for $\xi_0(), \mu_0()$.

To simplify notation we let $\mu^a = (\xi, \mu')'$. Similar to Problem 6.9, problem 7.6 , is a standard M-estimation problem with concave criterion. Therefore we easily obtain that problem 7.6 provides consistent estimators for $\mu_0^a(\theta)$. We also obtain similar asymptotic results as follows:

$$\sqrt{T}\left[\mathcal{K}_T(\theta) - \mathcal{K}(\theta)\right] = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [F(X_t, \mu_T^a(\theta), \theta) - F(X_t, \mu_0^a(\theta), \theta)] + F_T^a(\theta)$$

where

$$F_T^a(\theta) \doteq \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [F(X_t, \mu_0^a(\theta), \theta) - \mathcal{K}(\theta)].$$

We again show that only the second term $F_T^a(\theta)$ contributes to the approximation. Again due to the concavity of F in μ^a for each θ, a gradient inequality for such functions implies that

$$0 \leqslant \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[F(X_t, \mu_T^a(\theta), \theta) - F(X_t, \mu_0^a(\theta), \theta) \right] \leqslant \left[\mu_T^a(\theta) - \mu_0^a(\theta) \right] \cdot h_T^a(\theta),$$

where

$$h_T^a(\theta) \doteq \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial F}{\partial \mu^a}(X_t, \mu_0^a(\theta), \theta).$$

This leads us to make the following assumption similar to assumption 6.10.

Assumption 7.7.

i) The empirical process $\{[F_T^a(\theta) : \theta \in \Theta\}$ is Donsker, which converges weakly to a tight Gaussian process $\{\mathcal{G}^a(\theta) : \theta \in \Theta\}$ with zero mean and covariance function $C^a(,)$,

$$C^a(\theta_1, \theta_2) \doteq \lim_{T \to \infty} Cov \left[F_T^a(\theta_1), F_T^a(\theta_2) \right] = \sum_{j=-\infty}^{\infty} Cov \left[F(X_1, \mu_0^a(\theta_1), \theta_1), F(X_{1+j}, \mu_0^a(\theta_2), \theta_2) \right]$$

for any $\theta_1, \theta_2 \in \Theta$.

ii) The empirical process $\{h_T^a(\theta) : \theta \in \Theta\}$ is Donsker, which converges weakly to a tight Gaussian process with zero mean and covariance function

$$\mathbf{V}_0^a(\theta_1, \theta_2) \doteq \lim_{T \to \infty} Cov \left[h_T^a(\theta_1), h_T^a(\theta_2) \right] = \sum_{j=-\infty}^{\infty} Cov \left[\frac{\partial F}{\partial \mu^a}(X_1, \mu_0^a(\theta_1), \theta_1), \frac{\partial F}{\partial \mu^a}(X_{1+j}, \mu_0^a(\theta_2), \theta_2) \right]$$

Result 7.8. *Under Assumption 7.7, we obtain the following result:*

1. *Uniformly over $\theta \in \Theta$,*

$$\sqrt{T} \left[\mathcal{K}_T(\theta) - \mathcal{K}(\theta) \right] = F_T^a(\theta) + o_p(1),$$

which converges weakly to the Gaussian process $\{\mathcal{G}^a(\theta) : \theta \in \Theta\}$

2. *Uniformly in $\theta \in \Theta$,*

$$\sqrt{T} \left[\mu_T^a(\theta) - \mu_0^a(\theta) \right] = - \left[\mathbf{H}_0^a(\theta) \right]^{-1} h_T^a(\theta) + o_p(1),$$

which converges weakly to a tight Gaussian process with mean zero and covariance function:

$$\left[\mathbf{H}_0^a(\theta_1) \right]^{-1} \mathbf{V}_0^a \left(\theta_1, \theta_2 \right) \left[\mathbf{H}_0^a \left(\theta_2 \right) \right]^{-1}.$$

7.1.2 Estimation of nonlinear expectation functional

Under mild conditions, a slight extension of theorem 3.6 of Shapiro (1991) from iid data to stationary $\beta - mixing$ data, we have

Result 7.9. $\sqrt{T}(\widehat{\mathbb{K}} - \mathbb{K}) = \min_{\theta \in \Theta_\kappa} \sqrt{T}\left[\mathcal{K}_T(\theta^a) - \mathcal{K}(\theta^a)\right] + o_p(1) \rightsquigarrow \min_{\theta \in \Theta_\kappa} \mathcal{G}^a(\theta)$.

For any finite sample it is obvious that $\mathbb{E}[\widehat{\mathbb{K}}] \leqslant \mathbb{K}$ but $\mathbb{E}[\widehat{\mathbb{K}}]$ increases as T increases. If $\Theta_\kappa = \{\theta_0\}$ is a singleton, then $\sqrt{T}(\widehat{\mathbb{K}} - \mathbb{K}) \rightsquigarrow \mathcal{G}^a(\theta_0)$, which is a mean zero normal random variable with variance $C^a(\theta_0, \theta_0)$.

When Θ_κ is not a singleton, as shown in Shapiro (1991) and many other papers in the stochastic programming literature, the optimal value (in our case $\mathbb{K}$) is Hadamard directional differentiable, and hence the functional Delta theorem still applies (also see, e.g., Proposition 1 of Römisch (2004)). In the stochastic programming literature, there are at least four popular approaches for confidence sets construction that we could apply for the optimal value $\mathbb{K}$: i) Monte Carlo simulations, ii) non-asymptotic large deviation bounds (see, e.g., chapter 5 of Shapiro et al. (2014)), iii) subsampling, and iv) the extended bootstrap (see, e.g., Eichhorn and Römisch (2007)). To the best of our knowledge, most papers on DRO using convex divergence have simply assumed that Θ_κ is a singleton when constructing confidence intervals (see, e.g., Shapiro (2017), Duchi and Namkoong (2021)).

In the econometrics literature, both the subsampling approach (see, e.g., Chernozhukov et al. (2007) and Romano and Shaikh (2010)) and various modified bootstrap approach (see, e.g., Fang and Santos (2019), Hong and Li (2018), Christensen and Connault (2019)) have been used to construct confidence intervals for Hadamard directional differentiable functionals such as $\mathbb{K}$ for iid data. We note that both approaches lead to less powerful inference when Θ_κ is a singleton. We could extend any of these existing approaches to our framework with β-mixing weakly dependent time series data. See, for example, Appendix B for constructing confidence interval for $\mathbb{K}$ via the numerical delta method of Hong and Li (2018) + weighted bootstrap for time series data. However, this approach seems computationally more demanding than the confidence bands for ϑ we propose in Subsection 7.2.

7.2 Frontier belief distortion under expectation bounds

Here we briefly sketch an alternate approach to computing expectations bounds that agrees in a population sense with our nonlinear expectations approach. We treat the distorted expectation $\mathbb{E}[Mg(X,\theta)]$ as a parameter ϑ and describe how to construct identification-robust confidence sets for ϑ using an analogous procedure to that described in Section 6.

We note that Problem 7.1 can be written equivalently as follows:

Problem 7.10. *Let $\kappa \geqslant \underline{\kappa}$.*

$$\mathbb{K}(\kappa) \doteq \min_{\theta \in \Theta} \mathcal{K}(\theta, \kappa),$$

$$\mathcal{K}(\theta, \kappa) \doteq \inf_{\vartheta, M \geqslant 0} \vartheta \quad \textit{subject to}$$

$$\begin{aligned}
&\mathbb{E}\left[Mf(X,\theta)\right] = 0,\\
&\mathbb{E}\left[M\right] = 1,\\
&\mathbb{E}\left[\phi(M)\right] \leqslant \kappa,\\
&\mathbb{E}\left[Mg(X,\theta)\right] \leqslant \vartheta.
\end{aligned}$$

In particular, the dual problem is

$$\max_{\xi>0} \max_{\mu} \max_{\lambda_g \geqslant 0} \inf_{\vartheta} \inf_{M \geqslant 0} \; \mathbb{E}\left[\vartheta + \lambda_g(Mg(X,\theta) - \vartheta) + \xi(\phi(M) - \kappa) - \lambda \cdot f(X,\theta)M - \nu(M-1)\right],$$

where the optimization part $\max_{\lambda_g \geqslant 0} \inf_{\vartheta}[\cdot]$ can be solved in closed-form as

$$\lambda_g^* = 1, \quad \vartheta^* = \mathbb{E}\left[Mg(X,\theta)\right].$$

Note in particular that the new parameter ϑ is equal to the distorted conditional expectation at the optimum. We use this insight to motivate the alternate formulation below.

For any $M^*(\theta)$ given in equation (12) we define

$$\overline{\vartheta} \doteq \max_{\theta \in \Theta} \mathbb{E}\left[M^*(\theta) g(X,\theta)\right].$$

Problem 7.11. *Assume $\vartheta < \overline{\vartheta}$.*

$$\mathbb{L} \doteq \min_{\theta \in \Theta, \vartheta} \mathcal{L}(\theta, \vartheta) > \underline{\kappa}$$

where for fixed (θ, ϑ),

$$\mathcal{L}(\theta; \vartheta) \doteq \inf_{M \geqslant 0} \mathbb{E}\left[\phi(M)\right]$$

subject to:

$$\begin{aligned}
&\mathbb{E}\left[Mf(X,\theta)\right] = 0, \\
&\mathbb{E}\left[M\right] = 1, \\
&\mathbb{E}\left[Mg(X,\theta)\right] = \vartheta.
\end{aligned}$$

Denote $\theta^a \doteq (\theta, \vartheta)$ and

$$f^a(x, \theta^a) \doteq (f(x,\theta), g(x,\theta) - \vartheta).$$

We slightly strengthen Assumption 6.1(ii)(iii) to hold for θ^a and $f^a(x, \theta^a)$.

Then Problem 7.11 can be estimated exactly the same way as we did for Problem 6.2, with θ^a, $f^a(x, \theta^a)$ replacing θ, $f(x, \theta)$ respectively. In particular, confidence sets for $\theta^a \doteq (\theta, \vartheta)$ can be constructed in ways analogous to those described in section 6.

8 Discussion and Conclusion

Generalized empirical likelihood (GEL) methods typically aim at efficient estimation of a unique parameter vector for which moment conditions hold under data-generating process. Under misspecification and some regularity conditions, such methods will sometimes consistently estimate a pseudo true parameter. The implied empirical probabilities will approximate a minimally divergent probability measure under which the moment conditions hold. In our subjective belief framework, this approach will produce model parameters and implied beliefs which are consistent with model-implied moments and have a minimal divergence relative to rational expectations.

Several papers (including a suggestion in Hansen (2014) and the analysis in Ghosh and Roussellet (2020)) treat the minimal divergent beliefs and corresponding model parameter (assuming point-identified) as the target of estimation. In contrast, we view the distorted probability recovered in this way merely as a plausible measure of subjective beliefs, but we do not view it as the only plausible measure of subjective beliefs. Additionally, we do not necessarily view an identified parameter vector associated with the minimal divergent beliefs as the only parameter value of interest. We take a more eclectic approach because we do not see why the subjective beliefs of market participants must appear to the econometrician to have minimal divergence relative to rational expectations. Instead we consider it more fruitful to characterize and bound sets of plausible beliefs and model parameters consistent with certain levels of divergence from rational expectations (i.e. misspecification sets) and perform sensitivity analysis with respect to the level of divergence. We therefore view the methods developed in Chen et al. (2021) and in this paper as more appropriate to analyze implied subjective beliefs of economic agents than existing GEL methods.

In this paper, we assume that a generic dynamic model of finite-dimensional unconditional moment restrictions is misspecified under rational expectations, but is valid under agents' subjective beliefs. We devise econometrics that embrace model misspecification induced by a form of bounded irrationality. We implement this approach through the use of a statistical measure of divergence between the subjective beliefs of economic agents and the data generating process expectations implies bounds on agent's expectations. We are naturally led to replace point identification by set identification of both the subjective beliefs and the parameters of the moment restrictions. We represent the econometric relations of interest through a nonlinear expectation functional and derive its dual representation. Finally, we present several estimation and confidence set construction for the nonlinear expectation functional.

Our recently published paper Chen et al. (2021) uses a similar perspective to explore identification using conditional moment restrictions and dynamic counterparts to the divergence measures we consider in this paper. As an important future challenge, we plan to extend the econometric methods in this paper to apply to the population characterizations given in Chen et al. (2021).

References

Ai, Chunrong and Xiaohong Chen. 2007. Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Differing Conditioning Variables. *Journal of Econometrics* 141 (1):5–43.

Almeida, Caio and Rene Garcia. 2012. Assessing Misspecified Asset Pricing Models with Empirical Likelihood Estimators. *Journal of Econometrics* 170 (2):519 – 537.

Alvarez, Fernando and Urban J. Jermann. 2005. Using Asset Prices to Measure the Persistence of the Marginal Utility of Wealth. *Econometrica* 73 (6):1977–2016.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro. 2020. On the Informativeness of Descriptive Statistics for Structural Estimates. *Econometrica* 88 (6):2231–2258.

Antoine, Bertille, Kevin Proulx, and Eric Renault. 2018. Pseudo-True SDFs in Conditional Asset Pricing Models*. *Journal of Financial Econometrics* .

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, 214–223. PMLR.

Armstrong, Timothy B and Michal Kolesár. 2018. Sensitivity analysis using approximate moment condition models. *arXiv preprint arXiv:1808.07387* .

Attanasio, Orazio, Flávio Cunha, and Pamela Jervis. 2019. Subjective Parental Beliefs. Their Measurement and Role. Tech. rep., National Bureau of Economic Research.

Back, Kerry and David P. Brown. 1993. Implied Probabilities in GMM Estimators. *Econometrica* 61 (4):971–975.

Bhandari, Anmol, Jaroslav Borovicka, and Paul Ho. 2019. Survey Data and Subjective Beliefs in Business Cycle Models. Tech. rep., Federal Reserve Bank of Richmond Working Papers.

Bonhomme, Stéphane and Martin Weidner. 2018. Minimizing sensitivity to model misspecification. *arXiv preprint arXiv:1807.02161* .

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. 2020. Over-Reaction in Macroeconomic Expectations. *American Economic Review* forthcoming.

Borwein, Jonathan M. and Adrian S. Lewis. 1992. Partially Finite Convex Programming, Part II: Explicit Lattice Models. *Mathematical Programming* 57 (1-3):49–83.

Broniatowski, Michel and Amor Keziou. 2012. Divergences and Duality for Estimation and Test Under Moment Condition Models. *Journal of Statistical Planning and Inference* 142 (9):2554 – 2573.

Chen, Xiaohong and Xiaotong Shen. 1998. Sieve extremum estimates for weakly dependent data. *Econometrica* 289–314.

Chen, Xiaohong, Tim Christensen, and Elie Tamer. 2018. Monte Carlo Confidence Sets for Identified Sets. *Econometrica* 86 (6):1965–2018.

Chen, Xiaohong, Lars Peter Hansen, and Peter G. Hansen. 2021. Robust identification of investor beliefs. *Proceedings of the National Academy of Sciences* 117 (52):33130–33140.

Chernozhukov, V., H. Hong, and E. Tamer. 2007. Estimation and Confidence Regions for Parameter Sets in Econometric Models. *Econometrica* 75 (5):1243–1284.

Christensen, Timothy and Benjamin Connault. 2019. Counterfactual sensitivity and robustness. *arXiv preprint arXiv:1904.00989* .

Cressie, Noel and Timothy R. C. Read. 1984. Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)* 46 (3):440–464.

Csiszar, I. and Thomas Breuer. 2018. Expected Value Minimization in Information Theoretic Multiple Priors Models. *IEEE Transactions on Information Theory* 64 (6):3957–3974.

Csiszar, I. and F. Matus. 2012. Generalized Minimizers of Convex Integral Functionals, Bregman Distance, Pythagorean Identities. *Kybernetika* 48 (4):637–689.

Cuturi, Marco. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2292–2300.

Dedecker, Jérôme and Sana Louhichi. 2002. Maximal Inequalities and Empirical Central Limit Theorems. In *Empirical Process Techniques for Dependent Data*, 137–159. Springer.

Doukhan, Paul, Pascal Massart, and Emmanuel Rio. 1995. Invariance Principles for Absolutely Regular Empirical Processes. In *Annales de l'IHP Probabilités et statistiques*, vol. 31, 393–427.

Duchi, John C and Hongseok Namkoong. 2021. Learning Models with Uniform Performance via Distributionally Robust Optimization. *The Annals of Statistics* 49 (3):1378–1406.

Eichhorn, Andreas and Werner Römisch. 2007. Stochastic integer programming: Limit theorems and confidence intervals. *Mathematics of Operations Research* 32 (1):118–135.

Ekeland, I. and R. Témam. 1999. *Convex Analysis and Variational Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Fang, Zheng and Andres Santos. 2019. Inference on directionally differentiable functions. *The Review of Economic Studies* 86 (1):377–412.

Gagliardini, Patrick and Diego Ronchetti. 2019. Comparing Asset Pricing Models by the Conditional Hansen-Jagannathan Distance*. *Journal of Financial Econometrics* online.

Ghosh, Anisha and Guillaume Roussellet. 2020. Identifying Beliefs from Asset Prices. Tech. rep., SSRN Working paper.

Haberman, Shelby J. 1989. Concavity and Estimation. *Ann. Statist.* 17 (4):1631–1661.

Hall, Alastair R. and Atsushi Inoue. 2003. The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models. *Journal of Econometrics* 114 (2):361–394.

Hansen, Bruce E and Seojeong Lee. 2021. Inference for Iterated GMM Under Misspecification. *Econometrica* 89 (3):1419–1477.

Hansen, Lars Peter. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50 (4):1029–1054.

———. 2014. Nobel Lecture: Uncertainty Outside and Inside Economic Models. *Journal of Political Economy* 122 (5):945 – 987.

Hansen, Lars Peter and Scott F. Richard. 1987. The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. *Econometrica* 55 (3):587–613.

Hansen, Lars Peter and Kenneth J. Singleton. 1982. Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica* 50 (5):1269–1286.

Hansen, Lars Peter, John C. Heaton, and Amir Yaron. 1996. Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics* 14 (3):262–280.

Hjort, Nils Lid and David Pollard. 1993. Asymptotics for Minimisers of Convex Processes. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420* .

Hong, Han and Jessie Li. 2018. The numerical delta method. *Journal of Econometrics* 206 (2):379–394.

Imbens, Guido W. 1997. One-Step Estimators for Over-Identified Generalized Method of Moments Models. *The Review of Economic Studies* 64 (3):359–383.

Imbens, Guido W., Richard H. Spady, and Phillip Johnson. 1998. Information Theoretic Approaches to Inference in Moment Condition Models. *Econometrica* 66 (2):333–357.

Kitamura, Yuichi and Michael Stutzer. 1997. An Information-Theoretic Alternative to Generalized Method of Moments Estimation. *Econometrica* 65 (4):861–874.

Kitamura, Yuichi, Taisuke Otsu, and Kirill Evdokimov. 2013. Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica* 81 (3):1185–1201.

Lachout, Petr. 2005. Stochastic Optimisation: Sensitivity and Delta Theorem. *PAMM: Proceedings in Applied Mathematics and Mechanics* 5 (1):725–726.

Lee, Seojeong. 2016. Asymptotic Refinements of a Misspecification-Robust Bootstrap for GEL estimators. *Journal of Econometrics* 192:86–104.

Léger, Flavien. 2020. A Gradient Descent Perspective on Sinkhorn. *Applied Mathematics and Optimization* online.

Luttmer, Erzo, Lars P. Hansen, and John Heaton. 1995. Econometric Evaluation of Asset Pricing Models. *Review of Financial Studies* 8:237–274.

Ma, Shuangge and Michael R Kosorok. 2005. Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* 96 (1):190–217.

Manski, Charles F. 2018. Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise. *NBER Macroeconomics Annual* 32 (1):411–471.

Meeuwis, Maarten, Jonathan A Parker, Antoinette Schoar, and Duncan I Simester. 2018. Belief disagreement and portfolio choice. Tech. rep., National Bureau of Economic Research.

Newey, Whitney K. and Richard J. Smith. 2004. Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* 72 (1):219–255.

Peng, Shige. 2004. Nonlinear Expectations, Nonlinear Evaluations and Risk Measures. In *Stochastic Methods in Finance: Lecture Notes in Mathematics*, edited by M. Frittelli and W. Runggaldier, 165–253. Berlin, Heidelberg: Springer Berlin Heidelberg.

Qin, Jin and Jerry Lawless. 1994. Empirical Likelihood and General Estimating Equations. *Ann. Statist.* 22 (1):300–325.

Romano, Joseph P and Azeem M Shaikh. 2010. Inference for the identified set in partially identified econometric models. *Econometrica* 78 (1):169–211.

Römisch, Werner. 2004. Delta Method, Infinite Dimensional. *Encyclopedia of Statistical Sciences* 3.

Schennach, Susanne M. 2007. Point Estimation with Exponentially Tilted Empirical Likelihood. *Annals of Statistics* 35 (2):634–672.

Shapiro, Alexander. 1991. Asymptotic Analysis of Stochastic Programs. *Annals of Operation Research* 30:169—-186.

———. 2017. Distributionally Robust Stochastic Programming. *SIAM Journal on Optimization* 27 (4):2258–2275.

Shapiro, Alexander, Darinka Dentcheva, and Andrzej Ruszczyński. 2014. *Lectures on stochastic programming: modeling and theory*. SIAM.

Smith, Richard. 1997. Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation. *Economic Journal* 107:503–519.

Xie, Yujia, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2019. A fast proximal point method for computing exact Wasserstein distance. In *35th Conference on Uncertainty in Artificial Intelligence, UAI 2019*.

Appendix

A Proofs and Derivations for Section 2

A.1 Proof of Theorem 3.2

Construct a sequence $\pi_j \searrow 0$ such that $\pi_j < \frac{1}{2}$ for all j. Then choose $r_j \in \mathbb{R}^d$ such that

$$(1-\pi_j)\mathbb{E}[f(X)] + \pi_j r_j = 0$$

i.e.

$$r_j = -\left(\frac{1-\pi_j}{\pi_j}\right)\mathbb{E}[f(X)]$$

Let $B(r,\epsilon)$ denote an open ball with center r and radius ϵ. Since $-\mathbb{E}[f(X)] \in \operatorname{int}(C)$ there exists an $\epsilon > 0$ such that the open ball $B(-\mathbb{E}[f(X)],\epsilon) \subset C$. Since C is a cone and $\pi_j < \frac{1}{2}$ it follows that $B(r_j,\epsilon) \subset C$. Write $v(\epsilon) = \text{vol}[B(0,\epsilon)] > 0$.[21] Now, construct a sequence of belief distortions M_j as follows:

$$M_j(x) = (1-\pi_j) + \pi_j \frac{1}{v(\epsilon)h_0[f(x)]}\mathbf{1}\{f(x) \in B(r_j,\epsilon)\}$$

where $h_0(y)$ is the density of the random variable $Y = f(X)$ under the objective probability measure P. By construction, we have that for all $j \in \mathbb{N}$

- $M_j > 0$
- $\mathbb{E}[M_j] = 1$
- $\mathbb{E}[M_j f(X)] = 0$.

Additionally note that $M_j \geqslant (1-\pi_j)$ with probability one. Since $\phi(\cdot)$ is decreasing, we have that $\phi(M_j) \leqslant \phi(1-\pi_j)$ with probability one. By continuity, $\phi(1-\pi_j) \to \phi(1) = 0$. By monotonicity of expectations we see that

$$0 \leqslant \mathbb{E}[\phi(M_j)] \leqslant \mathbb{E}[\phi((1-\pi_j))] = \phi(1-\pi_j) \to 0.$$

The statement follows immediately. □

[21] Here we use the definition $\text{vol}(S) = \int \mathbf{1}(y \in S)dy$.

A.2 Proof of Theorem 3.6

The negative of a log moment generating function is strictly concave. Conditions (i) and (ii) guarantee that the function ψ is continuous and coercive. It follows from (Ekeland and Témam, 1999, Proposition 1.2, Ch. II.1, p.35) that the supremum in Problem 3.1 with relative entropy divergence is attained uniquely at vector we denote λ^*. Since ψ is differentiable, λ^* is determined uniquely by solving the first-order conditions. Moreover, from known results about moment generating functions we may differentiate inside the expectation to conclude that the first-order conditions with respect to λ imply

$$\mathbb{E}\left[\frac{\exp(\lambda^* \cdot f(X))}{\mathbb{E}[\exp(\lambda^* \cdot f(X))]} f(X)\right] = \mathbb{E}[M^* f(X)] = 0.$$

This can be seen directly via the dominated convergence theorem. Thus M^* is feasible for Problem 3.1.

To verify that M^* solves Problem 3.1, note that for any other $M \geqslant 0$ with $\mathbb{E}[M] = 1$,

$$\mathbb{E}[M(\log M - \log M^*)] \geqslant 0,$$

and thus

$$\mathbb{E}[M \log M] \geqslant \mathbb{E}[M \log M^*].$$

The first expression is nonnegative because it is the entropy of M relative to M^*.[22] Compute

$$\mathbb{E}[M \log M^*] = \mathbb{E}[M(\lambda^* \cdot f(X))] - \log \mathbb{E}\left[\exp\left(\lambda^* \cdot f(X)\right)\right].$$

Thus if $\mathbb{E}[Mf(X)] = 0$,

$$\mathbb{E}[M \log M^*] = -\log \mathbb{E}\left[\exp\left(\lambda^* \cdot f(X)\right)\right].$$

We conclude that

$$\inf_{\mathsf{B}} \mathbb{E}[M \log M] \geqslant -\log \mathbb{E}\left[\exp\left(\lambda^* \cdot f(X)\right)\right]$$

where $\mathsf{B} = \{M \in L^1(\Omega, \mathfrak{G}, P) : \mathbb{E}[M] = 1, \mathbb{E}[Mf(X)] = 0\}$. Furthermore, the right-hand side is attained by setting $M = M^*$ and that other $M \in \mathsf{B}$ that attains the infimum is equal to M^* with probability one. □

[22] Formally $\mathbb{E}[M(\log M - \log M^*)] = \mathbb{E}[M^* \phi(M/M^*)]$ with $\phi(x) = x \log x$, so the expectation is nonnegative by Jensen's inequality.

A.3 Derivation of equation (6)

By standard duality arguments, the dual formulation of problem 4.1 is the saddlepoint equation

$$\sup_{\xi>0,\lambda,\nu} \inf_{M\geqslant 0} \mathbb{E}\left[Mg(X) + \xi(M\log M - \kappa) + \lambda \cdot Mf(X) + \nu(M-1)\right] \tag{23}$$

where ξ, λ and ν are Lagrange multipliers. Since the objective function is separable in M, we minimize

$$Mg(X) + \xi(M\log M - \kappa) + \lambda \cdot Mf(X) + \nu(M-1)$$

with respect to M. The first-order condition is

$$g(X) + \xi + \xi\log M + \lambda \cdot f(X) + \nu = 0.$$

Thus,

$$M = \frac{\exp\left(-\frac{1}{\xi}\left[g(X) + \lambda \cdot f(X)\right]\right)}{\mathbb{E}\left[\exp\left(-\frac{1}{\xi}\left[g(X) + \lambda \cdot f(X)\right]\right)\right]}.$$

Substituting back into equation (23) gives equation (6).

We can connect these results to our earlier analysis of dual Problem **??** by defining an alternative expectation $\widehat{\mathbb{E}}$ using a relative density:

$$\frac{\exp\left[-\frac{1}{\xi}g(X)\right]}{\mathbb{E}\exp\left[-\frac{1}{\xi}g(X)\right]}$$

Then write the objective as

$$\widehat{\mathbb{K}}(\xi; g) \doteq \sup_{\lambda} -\xi\log\widehat{\mathbb{E}}\exp\left[-\lambda \cdot f(X)\right] - \xi\log\mathbb{E}\exp\left[-\frac{1}{\xi}g(X)\right].$$

Since the last term does not depend on λ, we may appeal to Theorem 3.6 for the existence of a solution where Restriction 3.5 is imposed under the change of measure.[23]

[23] For computational purposes, there may be no reason to use the change of measure.

A.4 When will the relative entropy constraint bind?

We first give a high-level sufficient condition under which the relative entropy constraint in problem 4.1 binds. Write

$$\mathbb{K}(g;\xi) = \max_{\lambda} -\xi \log \mathbb{E}\left[\exp\left(-\frac{1}{\xi}g(X) + \lambda \cdot f(X)\right)\right] - \xi\kappa.$$

Let $\lambda(g;\xi)$ denote the maximizer in the definition of $\mathbb{K}(g,\xi)$, and define

$$\begin{aligned} M_1(g;\xi) &= \frac{\exp\left[-\frac{1}{\xi}g(X)\right]}{\mathbb{E}\left(\exp\left[-\frac{1}{\xi}g(X)\right]\right)} \\ M_2(g;\xi) &= \frac{\exp\left[-\frac{1}{\xi}g(X) + \lambda(\xi)f(X)\right]}{\mathbb{E}\left(\exp\left[-\frac{1}{\xi}g(X) + \lambda(\xi)f(X)\right]\right)} \end{aligned}$$

Restriction A.1.

$$\lim_{\xi\downarrow 0} \mathbb{E}\left[M_1(g;\xi)g(X)\right] - \mathbb{E}\left[M_2(g;\xi)g(X)\right] > 0$$

Proposition A.2. *Under restriction A.1,*

$$\lim_{\xi\downarrow 0} \frac{\partial}{\partial\xi}\mathbb{K}(g;\xi) = \infty$$

and therefore the relative entropy constraint in problem 4.1 binds for any value of $\kappa > \overline{\kappa}$.

Proof. An application of the Envelope Theorem gives that

$$\begin{aligned} \frac{\partial}{\partial\xi}\mathbb{K}(g;\xi) &= -\log \mathbb{E}\left(\exp\left[-\frac{1}{\xi}g(X) + \lambda(g;\xi)\cdot f(X)\right]\right) - \frac{1}{\xi}\mathbb{E}[M_2(g;\xi)g(X)] - \kappa \\ &= \frac{1}{\xi}\mathbb{H}(g;\xi) - \kappa \end{aligned}$$

where

$$\mathbb{H}(g;\xi) = -\xi\log\mathbb{E}\left(\exp\left[-\frac{1}{\xi}g(X) + \lambda(g;\xi)f(X)\right]\right) - \mathbb{E}[M_2(g;\xi)g(X)].$$

Applying L'Hôpital's rule, we see that

$$\lim_{\xi \downarrow 0} \mathbb{H}(g;\xi) = \lim_{\xi \downarrow 0} \mathbb{E}\left[M_1(g;\xi)g(X)\right] - \mathbb{E}\left[M_2(g;\xi)g(X)\right] > 0.$$

The result follows. □

Restriction A.1 is difficult to verify in practice. To make things more concrete, we give two somewhat general examples under which the relative entropy constraint will bind.

Example A.3 establishes that the relative entropy constraint will bind in problem 4.1 whenever the target random variable $g(X)$ has a lower bound $\underline{g}$ with arbitrarily small probability near that bound.

Example A.3. *For simplicity, omit the moment condition* $\mathbb{E}[Mf(X)] = 0$*. Suppose that*

(i) $\text{ess inf}[g(X)] = \underline{g} > -\infty$*,*

(ii) $\lim_{\epsilon \to 0} \mathsf{P}\left\{g(X) \leqslant \underline{g} + \epsilon\right\} = 0$*,*

Then for any $\kappa > 0$*, the relative entropy constraint in Problem 4.1 will bind.*

Example A.3 rules out indicator functions for the choice of g. Bounding such functions may be of interest if the econometrician wishes to consider bounds on distorted probabilities. We consider a version that allows for these in example A.4

Example A.4. *We consider a scalar moment condition with a support condition and consider bounds on indicator functions of the moment function. Suppose*

(i) $f(X)$ *is a scalar random variable;*

(ii) $\text{ess sup}(f(X)) = \mathsf{u} < \infty$*,*

(iii) $\lim_{\epsilon \to 0} \mathsf{P}\left\{f(X) \geqslant \mathsf{u} - \epsilon\right\} = 0$*.*

(iv) $g(X) = \mathbf{1}_{\{f(X) \geqslant -\mathsf{r}\}}$ *for* $\mathsf{r} > 0$*;*

Then for any $\kappa > 0$*, the relative entropy constraint in Problem 4.1 will bind.*

The statement that the relative entropy constraint binds for any $\kappa > 0$ in examples A.3 and A.4 follows immediately from lemmas A.5 and A.6 respectively. These two examples suggest that the relative entropy constraint will bind in many cases of interest even for arbitrarily large choices of κ.

A.5 Auxiliary results

Lemma A.5. *Let* $\underline{g} = \text{ess inf } g(X)$ *and assume that*

$$\lim_{\epsilon \to 0} \mathsf{P}\left\{g(X) \leqslant \underline{g} + \epsilon\right\} = 0.$$

Then for any $\kappa > 0$*, there exists a constant* $\zeta > \underline{g}$ *such that for any belief distortion* M *satisfying* $M \geqslant 0$*,* $\mathbb{E}[M] = 1$*, and* $\mathbb{E}[Mg(X)] \leqslant \zeta$*, we must have that* $\mathbb{E}[M \log M] > \kappa$*.*

Proof:

Write

$$h(\epsilon) = \mathsf{P}\left\{g(X) \leqslant \underline{g} + \epsilon\right\}$$

and observe that $h(\epsilon) > 0$ and $h(\epsilon) \to 0$ as $\epsilon \to 0$. Define an event $A(\epsilon)$ by

$$A(\epsilon) = \left\{g(X) \leqslant \underline{g} + \epsilon\right\}$$

Now, let $\zeta = \underline{g} + \frac{\epsilon}{2}$. Then for any M satisfying the constraints, we have that

$$\begin{aligned}
\underline{g} + \frac{\epsilon}{2} &\geqslant \mathbb{E}[Mg(X)] \\
&= \mathbb{E}\left[Mg(X)\mathbf{1}_{A(\epsilon)}\right] + \mathbb{E}\left[Mg(X)\mathbf{1}_{A(\epsilon)^c}\right] \\
&\geqslant \underline{g}\ \mathbb{E}\left[M\mathbf{1}_{A(\epsilon)}\right] + (\underline{g} + \epsilon)\mathbb{E}\left[M\mathbf{1}_{A(\epsilon)^c}\right] \\
&\geqslant \underline{g} + \epsilon\mathbb{E}\left[M\mathbf{1}_{A(\epsilon)^c}\right] \\
&= \underline{g} + \epsilon\left(1 - Q(\epsilon; M)\right)
\end{aligned}$$

where $Q(\epsilon; M) = \mathbb{E}\left[M\mathbf{1}_{A(\epsilon)}\right]$. Rearranging, we obtain the bound

$$\frac{1}{2} \geqslant 1 - Q(\epsilon)$$

which simplifies to

$$Q(\epsilon) \geqslant \frac{1}{2}.$$

It follows that

$$\mathbb{E}\left[M | A(\epsilon)\right] = \frac{\mathbb{E}[M\mathbf{1}_{A(\epsilon)}]}{\mathbb{E}\left[\mathbf{1}_{A(\epsilon)}\right]} = \frac{Q(\epsilon)}{h(\epsilon)} \geqslant \frac{1}{2h(\epsilon)}.$$

Additionally, since $M \geqslant 0$ we have the trivial inequality

$$\mathbb{E}\left[M|A(\epsilon)^c\right] \geqslant 0.$$

Now, let $\mathcal{F}(\epsilon)$ denote the σ-algebra generated by the event $A(\epsilon)$. Applying Jensen's inequality conditional on $\mathcal{F}(\epsilon)$ to the relative entropy, we obtain

$$\begin{aligned}
\mathbb{E}[M \log M] &\geqslant \mathbb{E}\left[\mathbb{E}[M|\mathcal{F}(\epsilon)] \log\left(\mathbb{E}[M|\mathcal{F}(\epsilon)]\right)\right] \\
&= h(\epsilon)\frac{Q(\epsilon)}{h(\epsilon)} \log\left[\frac{Q(\epsilon)}{h(\epsilon)}\right] + [1 - h(\epsilon)]\left(-\frac{1}{e}\right) \\
&\geqslant \frac{1}{2} \log\left[\frac{1}{2h(\epsilon)}\right] - \frac{1}{e}
\end{aligned}$$

where the second term comes from the fact that the function $\phi(m) = m \log m$ is bounded from below by $-e^{-1}$. Choosing ϵ sufficiently small so that the lower bound exceeds κ gives the desired result. □

Lemma A.6. *Let $f(X)$ be a scalar random variable. Assume that $M \geqslant 0$, $\mathbb{E}[M] = 1$, $\mathbb{E}[Mf(X)] = 0$ and that $\mathsf{P}\left\{f(X) \leqslant \mathsf{u}\right\} = 1$. Then for any $\mathsf{r} > 0$*

$$\mathbb{E}[M\mathbf{1}(f(X) \leqslant -r)] \leqslant \frac{\mathsf{u}}{\mathsf{u} + \mathsf{r}}$$

Proof.

$$\begin{aligned}
0 &= \mathbb{E}[Mf(X)] \\
&= \mathbb{E}\left[Mf(X)\mathbf{1}_{\{f(X) \leqslant -\mathsf{r}\}}\right] + \mathbb{E}\left[Mf(X)\mathbf{1}_{\{f(X) > -\mathsf{r}\}}\right] \\
&\leqslant -\mathsf{r}\mathbb{E}\left[M\mathbf{1}_{\{f(X) \leqslant -\mathsf{r}\}}\right] + \mathsf{u}\mathbb{E}\left[M\mathbf{1}_{\{f(X) > -\mathsf{r}\}}\right] \\
&\leqslant -(\mathsf{u} + \mathsf{r})\mathbb{E}\left[M\mathbf{1}_{\{f(X) \leqslant -\mathsf{r} + \mathsf{u}\}}\right].
\end{aligned}$$

Rearranging gives the desired result. □

Note that this upper bound is sharp so long as X has strictly positive density near $\overline{x}$ and $-r$. It can be approximated by letting M approach a two-point distribution with a point mass at $\overline{x}$ with probability $\pi = \frac{\overline{x}}{\overline{x}+r}$ and a point mass at $-r$ with probability $1 - \pi = \frac{r}{\overline{x}+r}$.

Lemma A.7. *Let $\mathsf{u} =$ ess sup $f(X)$ and assume that*

$$\lim_{\epsilon \to 0} \mathsf{P}(f(X) \geqslant \mathsf{u} - \epsilon) = 0$$

Then for any $\kappa > 0$ *and* $\mathsf{r} > 0$ *such that* $\mathsf{P}\{f(X) \leqslant -\mathsf{r}\} > 0$*, there exists a constant* $\delta > 0$ *such that for any belief distortion* M *satisfying* $M \geqslant 0$*,* $\mathbb{E}[M] = 1$*,* $\mathbb{E}[Mf(X)] = 0$ *and*

$$\mathbb{E}[M\mathbf{1}_{\{f(X)\leqslant -\mathsf{r}\}}] \geqslant \frac{\mathsf{u}}{\mathsf{u}+\mathsf{r}} - \delta,$$

we must have that $\mathbb{E}[M \log M] > \kappa$*.*

Proof. Write

$$h(\epsilon) = P(f(X) \geqslant \mathsf{u} - \epsilon)$$

and observe that $h(\epsilon) > 0$ and $h(\epsilon) \to 0$ as $\epsilon \to 0$.

Now, take $\epsilon \in (0, \mathsf{u} + \mathsf{r})$ and define the following events

$$\begin{aligned} A &= \{f(X) \leqslant -\mathsf{r}\} \\ B(\epsilon) &= \{-\mathsf{r} < f(X) < \mathsf{u} - \epsilon\} \\ S(\epsilon) &= \{f(X) \geqslant \mathsf{u} - \epsilon\}. \end{aligned}$$

Observe that A, $B(\epsilon)$ and $S(\epsilon)$ are mutually exclusive. Using the fact that $\mathbf{1}_{B(\epsilon)} = 1 - \mathbf{1}_A - \mathbf{1}_{S(\epsilon)}$ with probability one, we obtain

$$\begin{aligned} 0 &= \mathbb{E}[Mf(X)] \\ &= \mathbb{E}[Mf(X)\mathbf{1}_A] + \mathbb{E}[Mf(X)\mathbf{1}_{B(\epsilon)}] + \mathbb{E}[Mf(X)\mathbf{1}_{S(\epsilon)}] \\ &\leqslant -\mathsf{r}\mathbb{E}[M\mathbf{1}_A] + (\mathsf{u} - \epsilon)\mathbb{E}[M\mathbf{1}_{B(\epsilon)}] + \mathsf{u}\mathbb{E}[M\mathbf{1}_{S(\epsilon)}] \\ &= -\mathsf{r}\mathbb{E}[M\mathbf{1}_A] + (\mathsf{u} - \epsilon)\mathbb{E}[M(1 - \mathbf{1}_A - \mathbf{1}_{S(\epsilon)})] + \mathsf{u}\mathbb{E}[M\mathbf{1}_{S(\epsilon)}] \\ &\leqslant (\mathsf{u} - \epsilon) - (\mathsf{u} + \mathsf{r} - \epsilon)\mathbb{E}[M\mathbf{1}_A] + \epsilon\mathbb{E}[M\mathbf{1}_{S(\epsilon)}]. \end{aligned}$$

Rearranging, we obtain the lower bound

$$\mathbb{E}[M\mathbf{1}_{S(\epsilon)}] \geqslant \frac{(\mathsf{u} + \mathsf{r} - \epsilon)}{\epsilon}\left(\mathbb{E}[M\mathbf{1}_A] - \frac{\mathsf{u} - \epsilon}{(\mathsf{u} + \mathsf{r} - \epsilon)}\right)$$

Now, for any M such that

$$\mathbb{E}[M\mathbf{1}_A] \geqslant \frac{\mathsf{u}}{\mathsf{u}+\mathsf{r}} - \frac{\epsilon}{2}\frac{\mathsf{r}}{(\mathsf{u}+\mathsf{r})(\mathsf{u}+\mathsf{r}-\epsilon)}$$

we have that

$$
\begin{aligned}
\mathbb{E}[M\mathbf{1}_{S(\epsilon)}] &\geqslant \frac{(\mathsf{u}+\mathsf{r}-\epsilon)}{\epsilon}\left(\frac{\mathsf{u}}{\mathsf{u}+\mathsf{r}}-\frac{\epsilon}{2}\frac{\mathsf{r}}{(\mathsf{u}+\mathsf{r})(\mathsf{u}+\mathsf{r}-\epsilon)}-\frac{\mathsf{u}-\epsilon}{(\mathsf{u}+\mathsf{r}-\epsilon)}\right)\\
&\geqslant \frac{(\mathsf{u}+\mathsf{r}-\epsilon)}{\epsilon}\left(\frac{\epsilon}{2}\frac{\mathsf{r}}{(\mathsf{u}+\mathsf{r})(\mathsf{u}+\mathsf{r}-\epsilon)}\right)\\
&\geqslant \frac{1}{2}\frac{r}{\mathsf{u}+\mathsf{r}}
\end{aligned}
$$

It follows that

$$
\mathbb{E}[M|S(\epsilon)] = \frac{\mathbb{E}[M\mathbf{1}_{S(\epsilon)}]}{\mathbb{E}[\mathbf{1}_{S(\epsilon)}]} \geqslant \frac{1}{2h(\epsilon)}\frac{\mathsf{r}}{\mathsf{u}+r}.
$$

Now, let $\mathcal{F}(\epsilon)$ denote the σ-algebra generated by the event $S(\epsilon)$. Applying Jensen's inequality conditional on $\mathcal{F}(\epsilon)$ to the function $\phi(m) = m\log m$, we obtain

$$
\begin{aligned}
\mathbb{E}[M\log M] &\geqslant \mathbb{E}\left[\mathbb{E}[M|\mathcal{F}(\epsilon)]\log\left(\mathbb{E}[M|\mathcal{F}(\epsilon)\right)\right]\\
&\geqslant h(\epsilon)\frac{\mathbb{E}[M\mathbf{1}_{S(\epsilon)}]}{h(\epsilon)}\log\left(\frac{\mathbb{E}[M\mathbf{1}_{S(\epsilon)}]}{h(\epsilon)}\right)+(1-h(\epsilon))\left(-\frac{1}{e}\right)\\
&\geqslant \frac{1}{2}\frac{\mathsf{r}}{\mathsf{u}+\mathsf{r}}\log\left(\frac{1}{2h(\epsilon)}\frac{\mathsf{r}}{\mathsf{u}+\mathsf{r}}\right)-\frac{1}{e}.
\end{aligned}
$$

Note that we have used the inequality $x\log x \geqslant -\frac{1}{e}$ for all $x \in \mathbb{R}$. Now choosing ϵ sufficiently small so that the lower bound exceeds κ gives the desired result. □

B Confidence sets for $\mathbb{K}(\kappa)$ via numerical weighted bootstrap

We let $\{W_t\}_{t=1}^T$ be a positive correlated random vector that is independent of the original time series data $\{X_t\}_{t=1}^T$ and satisfies the following assumption.

Assumption B.1. *(i) $\{W_t\}_{t=1}^T$ is strictly stationary and independent of data $\{X_t\}_{t=1}^T$;*
(ii) $E[W_t] = 1$, $E[W_t)^3] < \infty$, $Cov(W_t, W_{t+j}) = \omega(j/J)$, where $\omega()$ is a positive symmetric kernel function with J be the lag truncation parameter.

For example, we could let $\{W_t\}_{t=1}^T$ be positively correlated with exp(1) as the marginal distribution. iid bootstrap weight with mean one and variance one (e.g., exp(1) distribution. If the data $\{X_t\}_{t=1}^T$ were iid, a popular natural choice of $\{W_t\}_{t=1}^T$ is iid exp(1) random variables, which is also called Bayesian bootstrap.

For the sake of completeness, we let $\theta^a = (\theta, \kappa)$ for any fixed $\kappa \geqslant \underline{\kappa}$. We define the weighed bootstrap analog of $\mathcal{K}_T(\theta^a)$ as follows:

Problem B.2.

$$\mathcal{K}_T^B(\theta^a) = \max_{\mu^a} \frac{1}{T}\sum_{t=1}^{T} W_t F(X_t, \mu^a, \theta^a) = \frac{1}{T}\sum_{t=1}^{T} W_t F(X_t, \mu_T^B(\theta^a), \theta^a)$$

where $\mu_T^B(\theta^a)$ is the bootstrap analog of $\mu_T^a(\theta^a)$.

Following Ma and Kosorok (2005) we can obtain the following results:

Result B.3. *Under Assumptions 6.10 and B.1, we have:*

1. *Conditional on data, $\{\sqrt{T}\left[\mathcal{K}_T^B(\theta^a) - \mathcal{K}_T(\theta^a)\right] : \theta^a\}$ converges weakly to a tight mean zero Gaussian process $\{\mathcal{G}^a(\theta^a) : \theta^a\}$.*
2. *Conditional on data, and uniformly in θ^a, $\sqrt{T}\left[\mu_T^B(\theta^a) - \mu_T^a(\theta^a)\right]$ converges weakly to a normally distributed random vector with mean zero and covariance:*

$$\left[\mathbf{H}_0^a(\theta^a)\right]^{-1} \mathbf{V}_0^a(\theta^a) \left[\mathbf{H}_0^a(\theta^a)\right]^{-1}.$$

The following is intuitively why the numerical weighted bootstrap will work asymptotically.

$$\sqrt{T}\left[\mathcal{K}_T^B(\theta^a) - \mathcal{K}_T(\theta^a)\right] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[W_t F(X_t, \mu_T^B(\theta^a), \theta^a) - W_t F(X_t, \mu_T^a(\theta^a), \theta^a)\right] + F_T^B(\theta^a)$$

where

$$F_T^B(\theta^a) \doteq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[(W_t - 1) F(X_t, \mu_T^a(\theta^a), \theta^a)\right].$$

Then, again by concavity of F in μ^a we have

$$0 \leqslant \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[W_t F(X_t, \mu_T^B(\theta^a), \theta^a) - W_t F(X_t, \mu_T^a(\theta^a), \theta^a)\right] \leqslant \left[\mu_T^B(\theta^a) - \mu_T^a(\theta^a)\right] \cdot h_T^B(\theta^a),$$

where

$$h_T^B(\theta^a) \doteq \frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t \frac{\partial F}{\partial \mu}(X_t, \mu_T^a(\theta^a), \theta^a).$$

Let $\mathbf{X} = \{X_t\}_{t=1}^T$ denote the data, and $E[\cdot|\mathbf{X}]$ denote conditioning on the data. We note that

$$E[h_T^B(\theta^a)|\mathbf{X}] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial F}{\partial \mu}(X_t, \mu_T^a(\theta^a), \theta^a) = 0 \ , \quad \text{by definition of} \quad \mu_T^a(\theta^a),$$

and

$$Var[h_T^B(\theta^a)|\mathbf{X}] = \sum_{j=-(T-1)}^{T-1}\omega(j/J)V_{T,j}(\theta^a),$$

$$V_{T,j}(\theta^a) = \frac{1}{T}\sum_{t=1}^{T-j}\frac{\partial F}{\partial \mu}(X_t, \mu_T^a(\theta^a), \theta^a)\left[\frac{\partial F}{\partial \mu}(X_{t+j}, \mu_T^a(\theta^a), \theta^a)\right]'.$$

Similarly we have:

$$E[F_T^B(\theta^a)|\mathbf{X}] = 0$$

and

$$Cov[F_T^B(\theta_1^a), F_T^B(\theta_2^a)|\mathbf{X}] = \sum_{j=-(T-1)}^{T-1}\omega(j/J)C_{T,j}(\theta_1^a, \theta_2^a),$$

$$C_{T,j}(\theta_1^a, \theta_2^a) = \frac{1}{T}\sum_{t=1}^{T-j}F(X_t, \mu_T^a(\theta_1^a), \theta_1^a)F(X_{t+j}, \mu_T^a(\theta_2^a), \theta_2^a).$$

We can now apply the numerical directional delta method of Hong and Li (2018) to estimate the limiting distribution of $\sqrt{T}(\widehat{\mathbb{K}}(\kappa) - \mathbb{K}(\kappa))$ by that of the ratio

$$\mathbb{D}_T^B \doteq \frac{\min_{\theta\in\Theta}\left(\mathcal{K}_T(\theta^a) + \epsilon_T\sqrt{T}[\mathcal{K}_T^B(\theta^a) - \mathcal{K}_T(\theta^a)]\right) - \min_{\theta\in\Theta}\mathcal{K}_T(\theta^a)}{\epsilon_T}$$

for $\epsilon_T = o(1)$ and $\epsilon_T\sqrt{T} \to \infty$.

By Theorem 2.2 of Lachout (2005) or Theorem 3.1 of Hong and Li (2018), we have:

Result B.4. *Let $\epsilon_T = o(1)$ and $\epsilon_T\sqrt{T} \to \infty$. Conditional on data, $\mathbb{D}_T^B \rightsquigarrow \min_{\theta\in\Theta_\kappa}\mathcal{G}^a(\theta, \kappa)$.*